Enabling lipidomics though bioinformatics and network biology

Ferran Casbas Pinto

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

September 2018

Abstract

Human disease and medical research have advanced greatly, thanks to developments in 'omics' techniques. Lipidomics focuses on the study of lipid profiles in biological samples. It is a growing area of research and development, especially with a view to finding clinical biomarkers and uncovering novel cellular mechanisms. However, the analysis and interpretation of raw data have been slow and disconnected from other information resources. Improvements to analysis time and novel capability can be enabled through the use of Bioinformatics. This thesis outlines three areas in which it has been applied.

In the first, a network, linking lipid molecules to the reactions in which they participate and the associated enzymes, can be queried using a set of web-enabled scripts. This software identifies and ranks the enzymes potentially causing the observed perturbations (between conditions) in lipid profiles and also generates a network specific to the perturbations. In turn, these enzymes can be used as queries to other databases to identify potential gene-expresion and protein-interaction regulators. This tool has been queried with a polycystic ovary syndrome dataset, revealing an important role for Phospholipase D1. The results suggest the potential involvement of the transcription factors MAZ, NFAT and STAT5B in the disease.

The second area focuses on oxylipins, which are signalling molecules involved in inflammation and other physiological responses and are the subject of targeted lipidomic studies. An interface for converting lipid perturbations to a colour-coded lipidomic network has been developed and database searches with the enzymes concerned revealed that they form membrane-bound multi-protein complexes of both enzymes and specific oxylipin receptors. Furthermore, transcriptional factors for certain enzymes have been identified, which suggest potential regulatory mechanisms. This software was used to study 5 datasets from investigations into Osteo-Arthritis, and the results are consistent with both differential regulation of EPHX2 and CYP2B6, and thromboxane levels usually (but not exclusively) following those of prostaglandins.

The third area concerns a set of solutions to the most common challenges in lipidomics-data analysis. These scripts can also be used for metabolomics data in general. The specific software includes the design of a method to pair C13-marked molecular species, data

reformatting, improved lipid identification, and a study of lipid-adduct formation with recommendations on how to identify and compensate for their presence.

The outcome of this work has been much reduced data-analysis times and new tools to understand the biological significance of changes in lipid profiles.

Acknowledgments

I am really grateful for the help, support and patience of both of my supervisors, Professor Charlie Hodgman and Professor David Barrett during the PhD. I must thank also all the help from many researchers from the CAB department, Dr Salah Abdelrazig, Dr Srinivas Ravipati, Catharine Ortori and Dr Petros Pousinis. They have helped me to understanding the principal mechanisms of HPLC-MS and offered me advice on how to tackle analytical problems.

A special thanks to colleagues Rupert Norman and Sergey Evseev, they have helped me on basic things as English but also in major things like taking breaks from overwhelming coding afternoons.

Lastly, I would like to thank my family, my dad Jose Luis Casbas, my mom Ester Pinto my brother Guillem Casbas, for their love, encouragement, and continuous support. Finally, I would like to thank the University of Nottingham for my PhD scholarship.

Abbreviations

Lp(a)	Lipoprotein(a)
12-HETE	12-Hydroxyeicosatetraenoic acid
12-HpETE	Arachidonic acid 12-hydroperoxide
13-HODE	13-Hydroxyoctadecadienoic acid
15-HETE	15-Hydroxyeicosatetraenoic acid
5-HETE	5-Hydroxyeicosatetraenoic acid
5-HpETE	Arachidonic acid 5-hydroperoxide
AA	Arachidonic acid
AAC	Aminoglycoside-acetylating
ADP	Adenosine di-phosphate
ALOX5	Arachidonate 5-lipoxygenase
ATP	Adenosine tri-Phosphate
CAv1	Caveolin 1
CBR1	Carbonyl Reductase 1
CE	Cholesterol esters
CEPT1	Ethanolamine phosphotransferase 1
Cer	Ceramide
CerP	Ceramide phosphate
CHPT1	Choline phosphotransferase
CL	Cardiolipins
CM	Chylomicron
CoA	Coenzyme A
COX	Cyclooxygenase
CYP	Cytochrome P450
CYP51A1	Cytochrome P450 family 51 subfamily A member 1
DG	Diglycerides
DHCR24	24-Dehydrocholesterol reductase
DHETs	Dihydroyeicosatrienoic acids
DMSO	Dimethylsulfoxide
E	Enzyme
EBP	EBP, Cholestenol delta-isomerase
EETs	Epoxyeicosatrienoic acids
ELAVL1	ELAV Like RNA Binding Protein 1
ER	Endoplasmic reticulum
ES	Enzyme substrate complex
ESI	Electrospray ionisation
FA	Fatty acids
FADH2	Flavin adenine dinucleotide
FAM213B	Family with sequence similarity 213 Member B
FDFT1	Farnesyl-diphosphate farnesyltransferase 1
FDPS	Farnesyl diphosphate synthase
FDPS	Farnesyl diphosphate synthase
G3P	Glycerol-3-phosphate
GC	Gas chromatography
GL	Glycerolipids
HDL	High-density lipoprotein
HDMB	Human metabolome database
HETEs	Hydroxyeicosatetraenoic acids
HMGCR	3-Hydroxy-3-methylglutaryl-CoA Reductase
HMGCS	3-Hydroxy-3-methylglutaryl-CoA Synthase
HODEs	Hydroxyloctadecadienoic acids
HPLC	High-performance liquid chromatography
HSD17B7	Hydroxysteroid 17-beta dehydrogenase 7
KEGG	Kyoto Encyclopedia of Genes and Genomes

lcb	Long-chain-base
LC-MS	Liquid chromatography coupled to mass spectrometry
LDL	Low density lipoprotein
LOX	Lipoxygenase
LPCAT	Lysophosphatidylcholine acyltransferase
LSS	Lanosterol synthase
LTA4	Leukotriene A4
LTA4H	Leukotriene A4 hydrolase
LTC4	Leukotriene C4
m/z	Mass to charge ratio
MG	Monoglycerides
MIA	Intra-articular injection of monosodium iodoacetate
MNX	Meniscal transection
MOGAT3	Monoacylglycerol O-acyltransferase 3
MS	Mass spectrometry
MSMO1	Methylsterol monooxygenase 1
MVD	Diphosphate Decarboxylase
MVK	Mevalonate kinase
NADH	Nicotinamide adenine dinucleotide
NUCB1	Nucleobindin 1
OPLS-DA	Orthogonal partial least squares discriminant analysis
P	Product
PC	Phosphocholines
PCOS	Polycystic ovary syndrome
PE	Phosphoethanolamines
PG	Glycerophosphoglycerols
PGD2	Prostaglandin D2
PGE2	Prostaglandin E2
PGI2	Prostaglandin I2
PGF2α	Prostaglandin F2 alpha
PHP	Hypertext Pre-processor
PI	Phosphoinositol
Pld1	Phospholipase D1
PLS	Partial least squares
PMVK	Phosphomevalonate kinase
PS	Phosphoserine
PTGIS	Prostaglandin I2 synthase
PUFA	Polyunsaturated fatty acid
S	Substrate
SC5D	Sterol-C5-desaturase
Sgms1	Sphingomyelin synthase 1
SM	Sphingomyelins
Smpd1	Sphingomyelin phosphodiesterase
SP	Sphingolipids
SQLE	Squalene epoxidase
TG	Triglycerides
TLC	Thin layer chromatography
TM7SF2	Transmembrane 7 superfamily member 2
TXA2	Thromboxane A2
VLDL	Very low-density lipoprotein

Table of content

Abstract 2
Acknowledgments 4
Abbreviations 5
1. Introduction 11
1.1 Lipidomics and lipids 12
1.1.1 Lipidomics 12
1.1.2 Lipids 13
1.1.3 Lipid Biochemistry 22
1.2 The lipidomic pipeline 23
1.2.1 Steps leading to data analysis 23
1.2.2 Lipid quantification 23
1.2.3 Data analysis 24
1.2.4 Identification 25
1.2.5 Post pipeline analysis 25
1.3 Computer support for lipidomics. 26
1.4 Issues facing lipidomics analysis 27
1.5 Network biology and bioinformatics 27
1.6 Structure of this thesis 28
2. Metabolic and regulatory interpretation of untargeted lipidomic data 30
2.1 Introduction 31
2.2 A methodology for elucidating regulatory mechanisms leading to changes in lipid profiles 32
2.3 Supplementary information 36
2.3.1 Manual curation 36
2.3.2 Lipidomic network properties 37
2.3.3 Worked Example 40
2.3.4. Search for potential regulators 46
2.3.5. Advanced techniques 47

2.4 Scoring 48
2.5 Testing the methodology 53
2.5.1 Output from random input data 53
2.5.2 Error tolerance 55
2.5.3 Artificial data test 57
2.6 Web usage statistics 62
2.7 Discussion 62
3. The oxylipins family 64
3.1 Background 65
3.1.1 Oxylipins 65
3.1.2 Targeted lipidomics 68
3.1.3 Why integrated network analysis is needed 69
3.1.4 Aims and objectives 70
3.2 Literature review 70
3.3 Pro-/-anti-inflammation score 71
3.4 Integrated network 72
3.5 Methodology 78
3.5.1 Description of data importation 78
3.5.2 Setting the colouring function to aid data visualisation and interpretation 80
3.6 Application of the method to existing oxylipin databases from clinical and pre-clinical inflammatory studies 81
3.6.1 Osteoarthritis 81
3.6.2 Application to oxylipin data from a surgical model of osteoarthritis 84
3.6.3 Application to oxylipin data from a chemical model of osteoarthritis 87
3.6.4 Application to an oxylipin dataset from a study in a rat model to assess the effect of diet on OA progression 90
3.6.5 Application of the method to a clinical study of OA in a patient cohort (Nottingham study) 93

3.6.6 Application of the method to a clinical study of OA in a patient cohort (PROOF study) 99
3.7 Conclusions 99
4. Software tools for improving lipidomic data analysis 104
4.1 Introduction 106
4.2 Lipid identification 107
4.3 Adducts on ESI 110
4.4 Adduct characterization 112
4.4.1 Adducts separated by retention time 113
4.4.2 Adduct proportions within lipid families 115
4.4.3 Adduct distribution in fatty-acid subcategories 119
4.5 Correction of ^{13}C marked species 121
4.6 Data management and arrangement 123
4.7 Conclusions 124
5. Conclusions and Future work 126
5.1 Integration of lipidomic data into systems biology 127
5.2 Osteoarthritis potential biomarkers and mediators 128
5.3 Advancements on lipidomic and metabolomic data processing. 129
5.4 Future work 130
Bibliography 131
Appendix 145
Appendix chapter 1 145
A. Lipid Biochemistry 145
Fatty acids 145
Phospholipid biosynthesis 149
Eicosanoid pathway 150
Cholesterol and derivatives pathway 152
Signalling pathways 153
B. Post data analysis lipidomic pipeline 154

Experimental design 154
Sample collection 154
Lipid extraction 155
Appendix chapter 2 156
A. VIP score vs other scores 156
B. Random Scored data old script 157
C. Summary of ROC curve analysis 158
Appendix chapter 3 162
A. List of all the oxylipins 162
B. Excel for inflamation score calculation 166
C. Colour-Layout excel sheet 167
Appendix chapter 4 168
A. Adduct summation 168
B. Scripts 171
a. Adduct-cleaner.py 171
b. List-mass-C13-check-ACT.py 174
c. Excelprocesing-2.0.py 176
d. Mass-LipidMAPS-Checking.py 177
e. Mass-LipidMAPS-Cleaner.py 178
f. Graph-RT.py 179
g. Graph-Family-2.py 180
h. Graph-FA-separator.py 181
C. C13 Correction results 182

1. Introduction

1.1 Lipidomics and lipids

1.1.1 Lipidomics

Lipidomics is a subset of metabolomics that aims to identify and quantify a large number of lipids and has different branches.

The approach of cellular lipidomics uses whole cell lysates of cultured cell lines to observe how the lipidome is remodelled under various physiological conditions (Garcia-Canaveras et al. 2017). For example, fatty-acid elongation enzymes are induced when monocytes differentiate into macrophages (Wallner et al. 2014). It is also possible to do organelle-specific lipidomics, to gain an insight into organellar function. A comprehensive lipidomics analysis of subcellular organelles during macrophage activation found that some changes occur in all organelles, whereas others are organelle specific (Andreyev et al. 2010).

Lipidomic analysis of plasma is more commonly used for diagnosis. It is known that fasting, feeding, genetic background and other physiological conditions alter the plasma-lipid composition. The lipids on the plasma are carried by lipoproteins (Feingold et al. 2000), which are split into seven classes: chylomicron (CM), CM remnant, very low-density lipoprotein (VLDL), low density lipoprotein (LDL), LDL remnant, lipoprotein(a) (Lp(a)), and high-density lipoprotein (HDL). The balance in the lipoproteins has been used for the past 50 years to diagnose cardiovascular disease (CVD), by comparing LDL against HDL. Additional parameters have been investigated (Serna et al. 2015) to be used with the HDL and LDL information to better distinguish CVD groups.

Because of the crucial role of lipids in metabolism, the use of lipids as biomarkers for therapeutic targets has been on the rise (Pyragius et al. 2013). The monitoring of levels of individual types of lipids has been able to assess disease progression (Hobbs et al. 2004). The different lipid species can be easily monitored in many biological fluids, like blood, urine, tears and amniotic fluid, most of them are non-invasive procedures (Stegemann et al. 2011). The lipids present in these biofluids are optimal biomarker candidates owing to their normally easy availability. There are several examples of successful candidate lipid biomarkers. The quantification of 24-hydroxycholesterol has been shown to correlate with coronary artery

diseases (Vaya et al. 2001). Another study (Anand et al. 2016) presents 23 possible lipid biomarkers for preeclampsia. These candidate biomarkers might improve clinical diagnosis without invasive procedures.

The use of lipids as therapeutic tools is also something that has been evolving in the field of lipidomics. Lipid homeostasis is crucial to prevent metabolic diseases, and, since lipids are not encoded by the genome, one of the ways to control lipid homeostasis is by targeting enzymes that metabolize lipids. For example, statin drugs block a step-in cholesterol synthesis resulting in lower plasma cholesterol level to treat hypercholesterolemia (Schonewille et al., 2016). Another option is to use lipids themselves as a therapeutic tool, because some are mediators of human inflammation. Essential FAs have been used on rheumatoid arthritis to promote favorable outcomes (Zurier et al. 1996). Ocular surface diseases have been improved thanks to lipid base therapy (Lim et al. 2015).

1.1.2 Lipids

Lipids are compounds with a very diverse chemical composition. A fundamental characteristic of these molecules is polarity and hydrophobicity, which makes them soluble in mainly nonpolar solvents. These properties make their selective extraction from biological tissue easier. This, combined with the ready availability of samples (such as blood or urine), encourages the study of these biomolecules.

Lipids have a range of different biological functions:

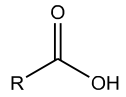
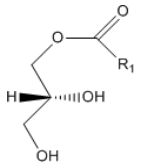
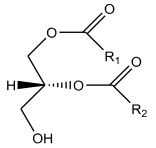
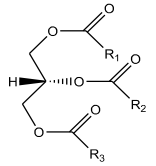
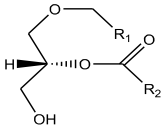
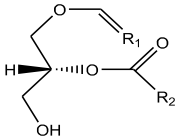
- Structural: They are the main component in cell membranes, some lipids act in the membrane as cell recognition tools.
- Energy storage: Lipid triglycerides have a high energy per mol when compared to glucose or other biomolecular classes.
- Regulatory: Some lipids like steroid hormones or oxylipins (which include the prostaglandins) are responsible for the activation and regulation of physiological processes, including inflammation and pain.

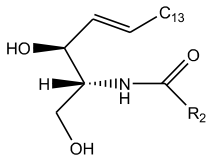
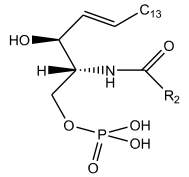
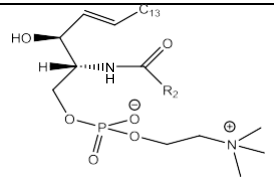
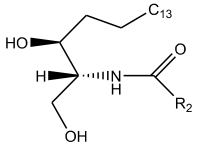
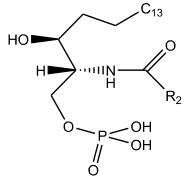
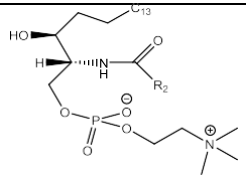
The control and tracking of some lipids are a day-to-day task. The control of cholesterol and triglycerides is a step done in many blood tests. The control of these through the diet is highly

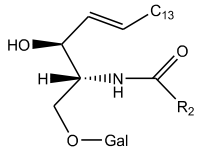
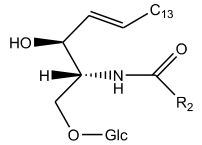
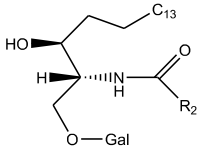
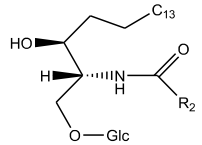
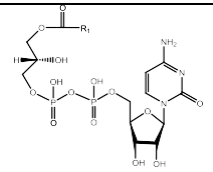
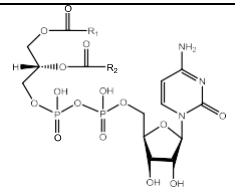
encouraged. If not controlled, they can lead to atherosclerosis (M and KS, 2016). Cholesterol can build up in the arteries, clogging them and making them less flexible. Blood does not flow as well through stiff arteries, so the heart must work harder to push blood through them. As plaque builds up in the arteries, over time it can lead to heart disease.

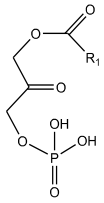
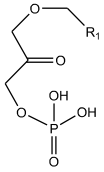
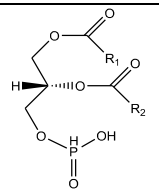
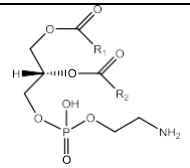
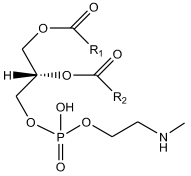
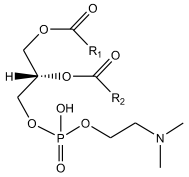
Lipids are not classified by their function. In 2005, the LIPID MAPS consortium, the International Lipid Classification, and National Nomenclature Committee decided that lipids will be divided into eight categories (Fahy et al. 2007), and each category would be divided into subclasses. By 2018, there are more than 40000 structures in LIPID MAPS. Table1.1 below presents the general structure of the different lipid species and their families. It can be appreciated the many different forms most of them can present, also it is important to take in consideration that where the R appears many fatty acids can be connected, from length of 1 up to 26 or even longer (Quehenberger et al., 2010).

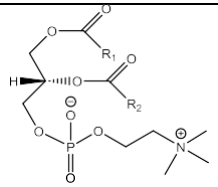
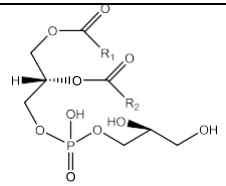
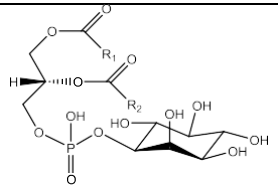
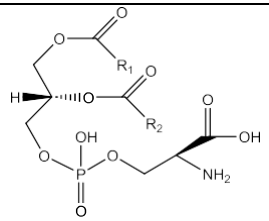
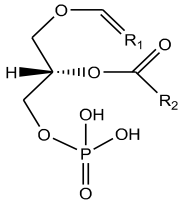
Table 1.1 Lipid categories

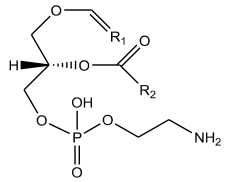
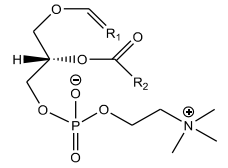
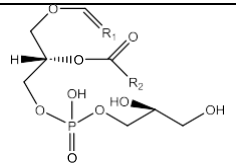
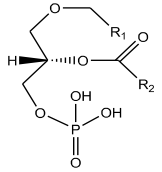
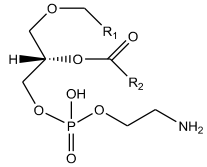
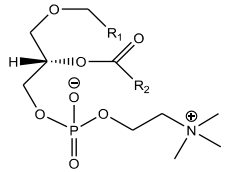
Abbreviation	General structure	Subclasses	URL
FATTY ACYLS [FA]			
FA		Fatty Acids and Conjugates [FA01]	LMFA0101
GLYCEROLIPIDS [GL]			
MG		Monoacylglycerols [GL01]	LMGL01010001
DG		Diacylglycerols [GL02]	LMGL02010002
TG		Triacyclglycerols [GL03]	LMGL03010007
DG(O-		Diradylglycerols [GL02]	LMGL02020001
DG(P-		Diradylglycerols [GL02]	LMGL02040001

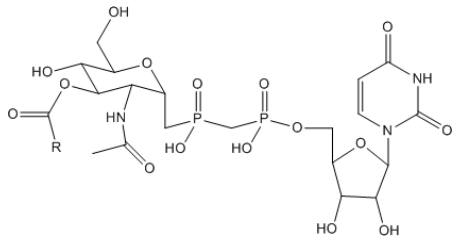
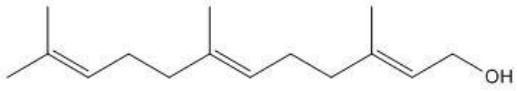
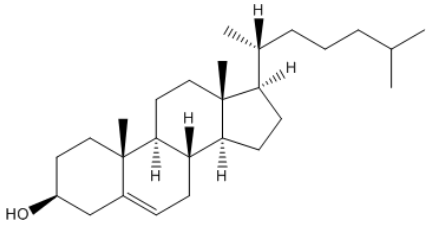
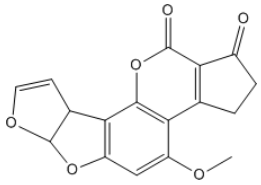
SPHINGOLIPIDS [SP]			
Cer		Ceramides [SP02]	LMSP02010001
CerP		Sphingolipids [SP]	LMSP02050001
SM		Phosphosphingolipids [SP03]	LMSP03010001
DH Cer		Ceramides [SP02]	LMSP02020001
DH CerP		Ceramides [SP02]	LMSP020500A0
DH SM		Phosphosphingolipids [SP03]	LMSP03010020

Gal Cer		Neutral glycosphingolipids [SP05]	LMSP0501AC09
Glc Cer		Neutral glycosphingolipids [SP05]	LMSP0501AA11
DH Gal Cer		Neutral glycosphingolipids [SP05]	LMSP0501AC11
DH Glc Cer		Neutral glycosphingolipids [SP05]	LMSP0501AA23
GLYCEROPHOSPHOLIPIDS [GP]			
CDP-MG		CDP-Glycerols [GP13]	LMGP13050000
CPD-DG		CDP-Glycerols [GP13]	LMGP13010001

Glycerone-Phosphate		Glycerophospholipids [GP]	LMGP00000063
Glycerone-Phosphate(O-		Glycerophospholipids [GP]	Not found
PA		Glycerophosphates [GP10]	LMGP10010028
PE		Glycerophosphoethanolamines [GP02]	LMGP02010021
PE-(Me)		Glycerophospholipids [GP]	LMGP02010344
PE-$(Me)_2$		Glycerophospholipids [GP]	LMGP02010317

PC		Glycerophosphocholines [GP01]	LMGP01010006
PG		Glycerophosphoglycerols [GP04]	LMGP04010027
PI		Glycerophosphoinositols [GP06]	LMGP06010008
PS		Glycerophosphoserines [GP03]	LMGP03010001
PA(P-		Glycerophosphates [GP10]	LMGP10030001

PE(P-		Glycerophosphoethanolamines [GP02]	LMGP02030014
PC(P-		Glycerophosphocholines [GP01]	LMGP01030004
PGP		Glycerophosphoglycerols [GP04]	LMGP04030001
PA(O-		Glycerophosphates [GP10]	LMGP10020002
PE(O-		Glycerophosphoethanolamines [GP02]	LMGP02020006
PC(O-		Glycerophosphocholines [GP01]	LMGP01020084

SACCHAROLIPIDS [SL]			
Common Name		Acylaminosugars [SL01]	LMSL01010001
PRENOL LIPIDS [PR]			
Common Name		Isoprenoids [PR01]	LMPR0103010001
STEROL LIPIDS [ST]			
Common Name		Sterols [ST01]	LMST01010001
POLYKETIDES [PK]			
Common Name		Aflatoxins and related substances [PK10]	LMPK10000006

While they have been classed according to their structure, lipids are unevenly distributed within cells (see Figure 1.1). The quantity of lipids is also quite different, over 60% of the lipids can be found on the cellular membrane (Agmon et al. 2017).

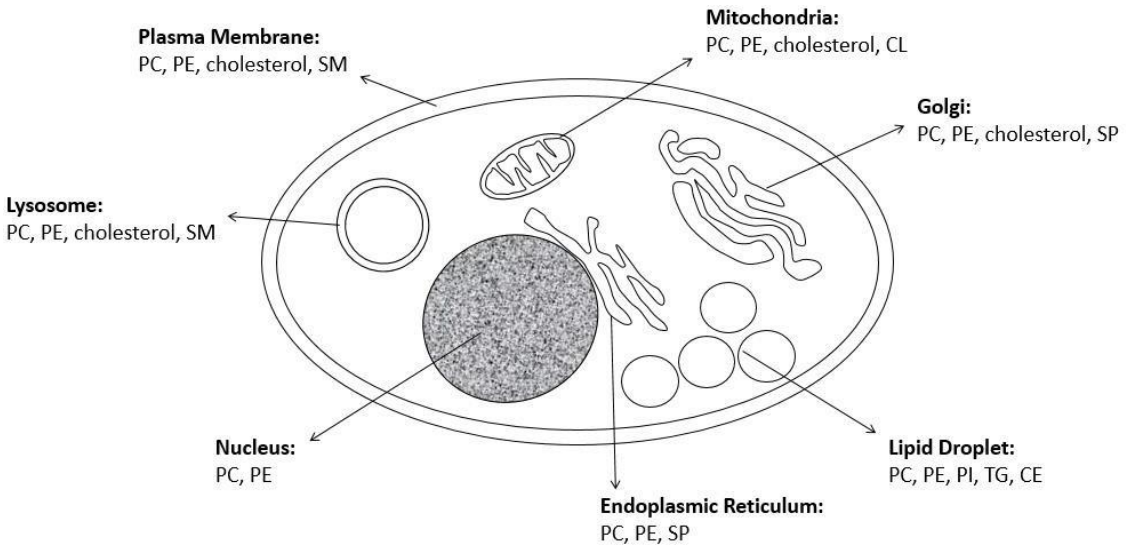


Figure 1.1: Lipid distribution in Eukaryotic cells. The figure presents the different most common lipid species across the different organelles of a eukaryotic cell

1.1.3 Lipid Biochemistry

The biochemical conversions among lipids are generally carried by low specificity enzymes, which involve the polar sections of the molecules, but paying little attention to fatty-acyl chain length or saturation of the nonpolar section. On the other hand, the size of those chains is critical for transportation and in the synthesis of bioactive lipids. Another property of the lipid metabolism is the irreversibility of their enzymes, a reverse reaction usually has to be carried out by another enzyme (Gurr, Harwood and Frayn, 2002).

More detailed information about the common pathway, the enzyme that catalyse them and the way that lipids are transported can be found in Appendix chapter 1 A.

1.2 The lipidomic pipeline

The main steps of a lipidomic experiment are shown in Figure 1.2 and described in greater detail below.

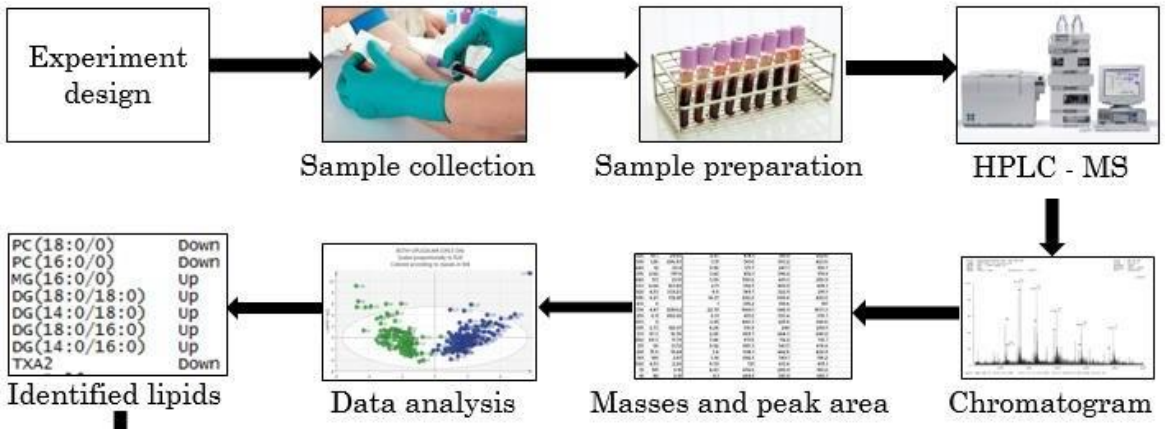


Figure 1.2: Lipidomic pipeline. The figure presents the main steps taken on an untargeted lipidomic experiment.

1.2.1 Steps leading to data analysis

The first steps in the lipidomic pipeline are the experimental design, sample collection and lipid extraction. Since these steps aren't the topic of this thesis all the information of what these techniques are and how are normally performed can be found on Appendix chapter 1 B.

1.2.2 Lipid quantification

In MS analysis, individual lipid species may be quantified by comparing the peak intensity or peak area with one or more internal standards. This step is employed in targeted methodologies. Where there is a limited number of compounds, it may be feasible to include heavy isotope or otherwise chemically-labelled lipids as internal standards for absolute quantification (Wang et al. 2016). This strategy is not useful for untargeted analysis since it would be unfeasibly expensive to add an internal standard for each peak, however one technique has been to include a set of exogenous compounds that are representative of each lipid class of interest. For example, Avanti Polar Lipids provides commercially available cocktails of lipid standards such as the SPLASH® (https://avantilipids.com/product/330707).

Even with internal standards, the length and degree of saturation of the molecules has a strong effect on its relative ionization (Duncan et al. 2008). In view of this, the output data are usually reported as relative percent abundances or fold change.

1.2.3 Data analysis

Peak identification and quantification can be done manually or by using software platforms. Manual curation is only recommended for small number of analytes and samples. There are over 10 software platforms able to do the data analysis (Tsugawa et al. 2017). In their simplest form, they should be able to identify and quantify peaks and correct for the equivalent ^{13}C forms. More details about this can be found in section 4.5. It is important to take into consideration the mode of detection: MS is able to detect in positive and in negative mode. The external electric field of an MS-ESI induces a separation of cations from negative electrolyte ions in the infusate while differential ionization of molecular species that possess differential electrical propensities can be induced in either the positive- or negative-ion mode during the electrospray ionization process. Some families and adducts prefer one mode over the other. Electrospray ionization can have the same lipids ionized in both modes; because, when the analytes in the infused solution do not carry separable charge(s) but possess intrinsic dipoles, these compounds can be induced to interact with small cation(s) or anion(s) available in the matrix to yield adduct ions in the positive- or negative-ion mode, respectively, in a high electric field (Han et al. 2006).

After processing the chromatographic output, the multicomponent data must be analysed with statistical methods to find significant differences between samples, molecules or time differences. This can be done with published software like SimcaP (Wu et al. 2010) and GraphPad (GraphPad Software, La Jolla California USA, www.graphpad.com), but also in spreadsheet software. The most common methods are OPLS-DA for untargeted lipidomic analysis and univariate analysis for targeted analysis.

1.2.4 Identification

Lipid identification is one of the most difficult steps in untargeted lipidomic studies. It is so significant that many publications only publish mass and peak-area values.

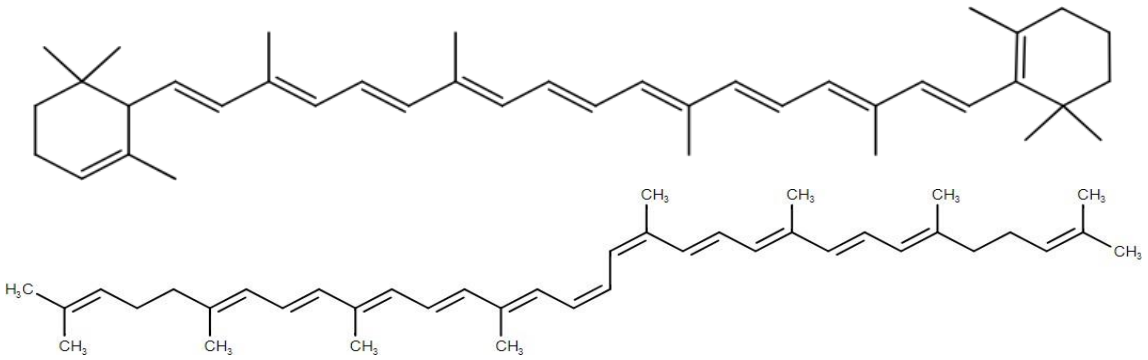


Figure 1.3: Chemical structure of βcarotene and lycopene. They have identical mass but different structure.

The main problem is that lipids contain a huge number of conformational isomers. For this reason, a high precision of the mass does not always help identify the molecule. Just a change one double bond position completely changes the molecular species, and in more complex molecules like PC the change of the chains being (18:0/16:0) to (16:0/18:0) would be unnoticeable. This also goes across lipid classes, for example PC(30:0) and PE(33:0) have identical masses. In some cases, identification can be achieved using prior knowledge, such as retention time or the unlikeliness of a particular molecular species being present in the sample. The safest option for untargeted analysis is to carry out $MS^{(n)}$. This technique detects the masses of fragments from individual species to validate their identification.

1.2.5 Post pipeline analysis

Up to this point is what classic lipidomic pipeline has been. Now we move into is one of the least covered aspects in lipidomics. Some attempts to fill it have been pathway tools (Karp et al. 2002) or Escher (King et al. 2015). These platforms have a limited number of reactions and are aimed at general metabolomics, omitting most of lipid metabolism. This is going to be the main focus of our chapter 2.

1.3 Computer support for lipidomics.

Many of the steps of the lipidomic pipeline use software. For example, for all the HPLCMS data management, Progenesis QI (http://www.nonlinear.com/progenesis/qi/) can do the peak picking, retention time alignment, convolution and deconvolution of the data, and normalization. In targeted analysis, the most common tools in the School of Pharmacy have been SIEVE (http://pmcore.med.hku.hk/index.php/instrumentation/bioinformatics-software) IDEOM (http://mzmatch.sourceforge.net/ideom.php) and LIPIDSEARCH (https://www.thermofisher.com/order/catalog/product/IQLAAEGABSFAPCMBFK), which have difficulty in dealing with identification of lipid species and adduct management.

The recent review (Tsugawa et al. 2017) goes into greater detail on the software tools and databases which have been developed for LC–MS or LC–MS/MS based untargeted lipidomics: MS-DIAL (Tsugawa et al. 2015), OpenMS (Röst et al. 2016), and XCMS family (XCMS (Smith et al. 2006), CAMERA (Kuhl et al. 2012), XCMS2 (Benton et al. 2008)). All these programs use the retention time (RT), precursor m/z, isotopic ratio, and MS/MS spectrum against a reference compound of MoNA (http://mona.fiehnlab.ucdavis.edu) or MassBank (Horai et al. 2010) spectral databases. There are other tools even more specialized on lipidomic identification: LipidFinder (O'Connor et al. 2010) and LOBSTAHS (Collins et al. 2016) They provide solution for isotope-, adduct-and other artefacts filtering in conventional LC solvent conditions, but also to deal with the annotation of oxylipins and oxidized lipids. The recent publication of this software enhances the feeling that this field is still developing.

The main database resources for lipidomic analysis include LIPIDMAPS, KEGG, and HMDB:

- LIPID Metabolites and Pathways Strategy (LIPID MAPS®) is a multi-institutional effort created in 2003 to identify and quantify all of the major, and many minor, lipid species in mammalian cells, using a systems biology approach and sophisticated mass spectrometers, as well as to quantify the changes in these species in response to perturbations (Fahy et al., 2007).
- KEGG (Kanehisa et al, 2017; Kanehisa et al, 2016; Kanehisa et al, 2000) is a database resource for understanding the higher-level functions of biological systems (such as the cell, organism and ecosystem) implicit in molecular-level information, especially

large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.

- The Human Metabolome Database (HMDB) (Wishart et al, 2007; Wishart et al, 2009; Wishart et al, 2013; Wishart et al, 2018) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body. It is intended to be used for applications in metabolomics, clinical chemistry, biomarker discovery and general education.

1.4 Issues facing lipidomics analysis

The first issue of lipidomics is on identification. Many software tools have been developed, but there is still a great number of false positives and 40-60% of unassigned peaks (Tsugawa et al. 2015). Their manual identification is a tedious task with no easy solution. This becomes even more complex when combined with adducts present in the ESI source.

Some of the data analysis steps on lipidomics are slow because they haven't been automated. This leaves some tasks to be done manually, which can be tedious. This has a second effect, normally lipidomic study output consists only of a shortened list of the top most changing lipids to find biomarkers. This leaves out a lot of potential information on other changing lipids that could present a more general trend in lipid metabolism as a whole. The final issue is the low amount of support for lipid-pathway analysis. Most tools are prepared for metabolomics due the complex nature of identifying lipids.

1.5 Network biology and bioinformatics

For a long time, biology has focused on the study of biomolecules. But it is not possible to describe biological systems with only the individual properties of such molecules. The interactions between the different biological components can be defined by a set of physical laws and is called systems biology. This discipline attempts to screen the behaviour of multiple interactions, including in some cases the environment required for some interactions to take place.

A biological network is any network applied to biological systems, such as animal species linked to form a food web, or proteins and metabolites to form a metabolic network (Proulx et al. 2005). The analysis of biological networks with respect to human diseases has led to the field of network medicine (Barabási et al. 2011, Habibi et al. 2014). There is a branch of maths, called graph theory, that is focussed on the study of the networks. Graphs are being used to model many types of relations and processes in physical, biological (Mashaghi et al. 2004), social (De, Sagar et al. 2012; Lakshmipraveena et al. 2017) and information systems (de Moor, 1997). These graphs are defined by nodes and edges. In biomolecular networks, nodes normally represent the biomolecules and the edges the interactions between them. Edges can be directed (for example inhibition of an enzyme by another a biomolecule, conversion of one biomolecule to another), or undirected (describing the proximity of two enzymes, the co-expression of two genes).

Many tools exist for visually exploring biological networks including well-known examples, such as Cytoscape, VisANT (http://visant.bu.edu/), Pathway Studio (https://www.pathwaystudio.com/) , Patika (http://www.cs.bilkent.edu.tr/~patikaweb/) and METExplore (https://metexplore.toulouse.inra.fr/index.html/) and have played a key role in the development of integrative biology, systems biology and integrative bioinformatics. The trend in the development of these tools is to go beyond 'static' representations of interactions, towards a more dynamic model of cellular processes through the incorporation of gene expression data, subcellular localization information and time- dependent behaviour (Suderman et al. 2007).

The study of diseases and biological mechanisms using networks has been increasing. Thanks to the creation of network models, such as genetic interaction networks and physical interaction networks, that have been constructed to study complex disease (Hu et al. 2016).

1.6 Structure of this thesis

The study of lipids in diagnosis and disease prevention with the current lipidomic pipeline has been able to provide great advances in medical research (Perrotti et al. 2016). However, there is a great gap on the steps to follow when the lipids have been properly identified and characterized. Given the past history of bioinformatics and systems biology, their application to lipidomics should improve the efficiency of the lipidomics pipeline and enhance

understanding of the processes regulating changes in lipid profiles. Hence the aims of this project are:

- Apply network biology to lipidomics to elucidate potential causes of lipid perturbations.
- Apply network biology to the lipid pathway of the pain and inflammation, specifically the study of oxylipins in OA.
- Develop bioinformatic tools to speed up the data analysis step of the lipidomic analysis pipeline.

In this thesis, the next chapter describes a way to extend the analysis pipeline to elucidate potential mechanisms to explain observed perturbations in global lipidomics data. It takes the form of the published paper on the methodology and includes further discussion of its application. Chapter 3 covers work on a more focused study of the oxylipins, with particular emphasis on lipidomics in Osteoarthritis and inflammation. Chapter four outlines new software tools that aim to improve the throughput of the lipidomic pipeline and includes work on adduct identification. Chapter 5 discusses the extent to which the thesis aims have been met, the discoveries made and potential follow-on work that could be carried out.

2. Metabolic and regulatory interpretation of untargeted lipidomic data

2.1 Introduction

This chapter describes the development of a methodology to extend the lipidomic analysis pipeline, which aids biological interpretation of differences in lipidomic profiles. It brings the lipidomic analysis closer to a biological and medical understanding of the observed changes. The method involves the creation of a network of lipid metabolism which is used to identify enzymes catalysing the reactions linked to perturbed lipids. These enzymes can then be used as a reference to published regulatory networks. The next section of this chapter consists of the published paper on this work, followed by its supplementary information. The subsequent sections present greater detail of the enzyme-node scoring system, validation tests, web-usage statistics and a final discussion

2.2 A methodology for elucidating regulatory mechanisms leading to changes in lipid profiles

The following paper is the work of my supervisor and I in the creation of the software tool draft of the text and design of the experiments. The experimental data was obtained from Dr Srini Ravipati (University of Nottingham), supervisory work was carried by Professor Charlie Hodgman and Professor David Barrett.

Metabolomics (2017) 13:81
DOI 10.1007/s11306-017-1214-y

SOFTWARE/DATABASE

A methodology for elucidating regulatory mechanisms leading to changes in lipid profiles

Ferran Casbas Pinto[1,2] · Srinivarao Ravipati[2] · David A. Barrett[2] · T. Charles Hodgman[1]

Received: 6 December 2016 / Accepted: 3 May 2017 / Published online: 29 May 2017 © The Author(s) 2017. This article is an open access publication

Abstract
Introduction It is difficult to elucidate the metabolic and regulatory factors causing lipidome perturbations.
Objectives This work simplifies this process.
Methods A method has been developed to query an online holistic lipid metabolic network (of 7923 metabolites) to extract the pathways that connect the input list of lipids.
Results The output enables pathway visualisation and the querying of other databases to identify potential regulators. When used to a study a plasma lipidome dataset of polycys-tic ovary syndrome, 14 enzymes were identified, of which 3 are linked to ELAVL1—an mRNA stabiliser.
Conclusion This method provides a simplified approach to identifying potential regulators causing lipid-profile perturbations.

1 Introduction

Recent advances in analytical methods, especially mass spectrometry, have shown that complex lipid profiles can be influenced by many factors and have the potential to improve our understanding of the pathophysiology underly-ing disease processes (references listed in Supplementary Information Section 2). However, the latter task is made difficult as a result of the complex chemical nature of lipids and their ubiquitous presence in multiple roles in biological tissues. In addition, the identification of many lipid molec-ular species is complicated by the high number of structural isomers corresponding to a given lipid mass. Although some low-throughput mass spectrometry techniques lead to precisely defined molecular structures (Nguyen et al. 2015), there is still more technical development needed before this can be applied in a routine way to global lipi-domics. Pathways of lipid metabolism are both complex and often not fully understood: some reactions are not well characterized; some metabolites are not endogenously syn-thesised but enter through the diet (Holmes et al. 2008); and many enzymes have low substrate specificity, allow-ing them to accept a huge range of fatty acids and alcohols as substrates (Gibellini and Smith 2010). In addition, most biological and clinical scientists refer to lipids by com-mon and often inconsistent names, rather than a systematic naming scheme, which can lead to confusion. These fac-tors together make it difficult to automate the construction and analysis of metabolic pathways and networks, and a basic requirement for progress is to define a mechanism for resolving synonyms and one-to-many relationships in lipid identities (Horne et al. 2004).

Most recent lipidomics papers conclude their study with a list of the lipids that were found to be significantly dif-ferent between two groups, for example between diseased and healthy tissue or between treated and control samples. However, there is a small number of examples that go one step further. For example, one can create and test a predic-tive model (Floegel et al. 2013), apply advanced statistics to classify profiles into high or low risk of suffering a disease (Rhee et al. 2011), manually construct a network between known disease-associated genes, signal-transduction

Electronic supplementary material The online version of this article (doi:10.1007/s11306-017-1214-y) contains supplementary material, which is available to authorized users.

*\ T. Charles Hodgman
\ Charlie.Hodgman@nottingham.ac.uk

1\ School of Biosciences, University of Nottingham, Sutton Bonington LE12 5RD, UK

2\ Centre for Analytical Bioscience, School of Pharmacy, University of Nottingham, Nottingham NG7 2RD, UK

pathways and the perturbed levels of specific lipids (Wang-Sattler et al. 2012), or colour-code pathway or network maps from databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Cottret et al. 2010; Titz et al. 2015).

In this work, we propose a method of systematically converting lists of perturbed lipids to integrate them with metabolic network which permits visualisation of the affected pathways. This process involves four steps, the first two involving the construction of a network (that links enzymes and hence genes to reactions between specific lipids) and a thesaurus (linking generic lipid identifiers to a list of possi-ble specific metabolites). The second two steps involve the implementation of algorithms to generate the subset of the total network that corresponds to the input profile of perturbed lipids, and to rank enzyme nodes in terms of their significance to the perturbation. As a proof of concept, we apply this method to a published lipidomics dataset .

2 Methods

2.1 Data sources

Source data for creating the holistic human lipidomic network included LIPID MAPS (Sud et al. 2007), KEGG (Kanehisa 1996), HMDB (Wishart et al. 2013) and other papers cited in Supplementary Information Section 3.

2.2 Network construction

The lipid metabolic network was created using Python scripts applied to data from the above sources, plus manual curation, and has been stored in a mySQL database. All scripts can be obtained from https://github.com/Ferran-Casbas/Lipid-Pathway. The network contains a metabo-lite node for each specific lipid, but does not distinguish between the fatty-acyl groups at the sn-1 and sn-2 chain positions to limit the network size and complexity and because analytical resolution of sn-1/sn-2 sidechains is rarely achieved across the whole lipidome. Other nodes may refer to individual reactions, and the enzymes (specified by their Human Genome Identifier) catalysing them. Further details are in the Supplementary Information Section 4.

2.3 Subset generation

The whole process is depicted in Fig. 1. The input used is a list of lipids whose concentrations are significantly higher or lower between two groups, determined for example after univariate or multivariate data analysis of a lipidom-ics dataset and the structural confirmation of the chemical identities of the changed lipids. Lipid names should conform to the standard in LIPID MAPS (http://www.lipidmaps.org/data/classification/lipid_cns.html#N).

The pre-ferred level of identification is to include the lipid class and the numbers of carbons and double bonds in each side chain. For example, this approach can include ambiguously identified molecules such as DG(36:2) which indicates a diglyceride with two side chains totalling 36 carbons with 2 double bonds. The fold change or significance score for the metabolites (often the VIP scores from Partial Least Squares analysis) can also be input and used for meta-bolic prioritisation. Alternatively, +1 and −1 can be used to imply the direction of change in lipid level between one group and another.

PHP scripts then use the list to produce a subgraph (sub-network). The first phase of the analysis works through the list to identify specific, ambiguous and unrecognisable terms. The software reports on this and recommends what might be done to clarify the unrecognised terms. Once the lipids of interest have been adequately associated with network nodes, a sub-network is generated by selecting from the lipidomic network these nodes and their two nearest neighbours, so that the reactions and enzyme/genes linking these lipids are also included. If the input lipids have fold change or significance scores, the output sub-network contains *Relative Importance* scores, which indicate how much influence the enzyme nodes have on producing the observed lipid-profile perturbation in light of the network topology. This can be particularly valuable when larger sub-networks are returned. More information on this is in Supplementary Information Section 6.

The above scripts are available through a web interface, which can be found at: http://lipidnetwork.nottingham.ac.uk/. The web tool outputs two files which can be imported into a network visualization software, such as Cytoscape (Doerks et al. 2002), to see the metabolic pathways involved in the profile changes. Another option is to open the node-info file in a spreadsheet to retrieve the list of enzymes, which can then be used as a query to other network tools, such as the BioGRID (Stark et al. 2006) or genemania, to identify potential regulators.

3 Results anddiscussion

This work describes a novel method for identifying enzymes and potential regulators related to measured changes in the lipid profile obtained from mass spectrometry analysis. The method includes the development of a holistic lipidomic network that comprises 127 enzymes, 13,934 reactions and 7561 metabolites. The latter are from 16 lipid subclasses: 3124 glycerolipids, 3575 glycerophospholipids, 290 sphingolipids, 126 fatty acyls (including eicosansoids, fatty amides, esters and their conjugates), and 446 miscellaneous lipids. An image of the complete

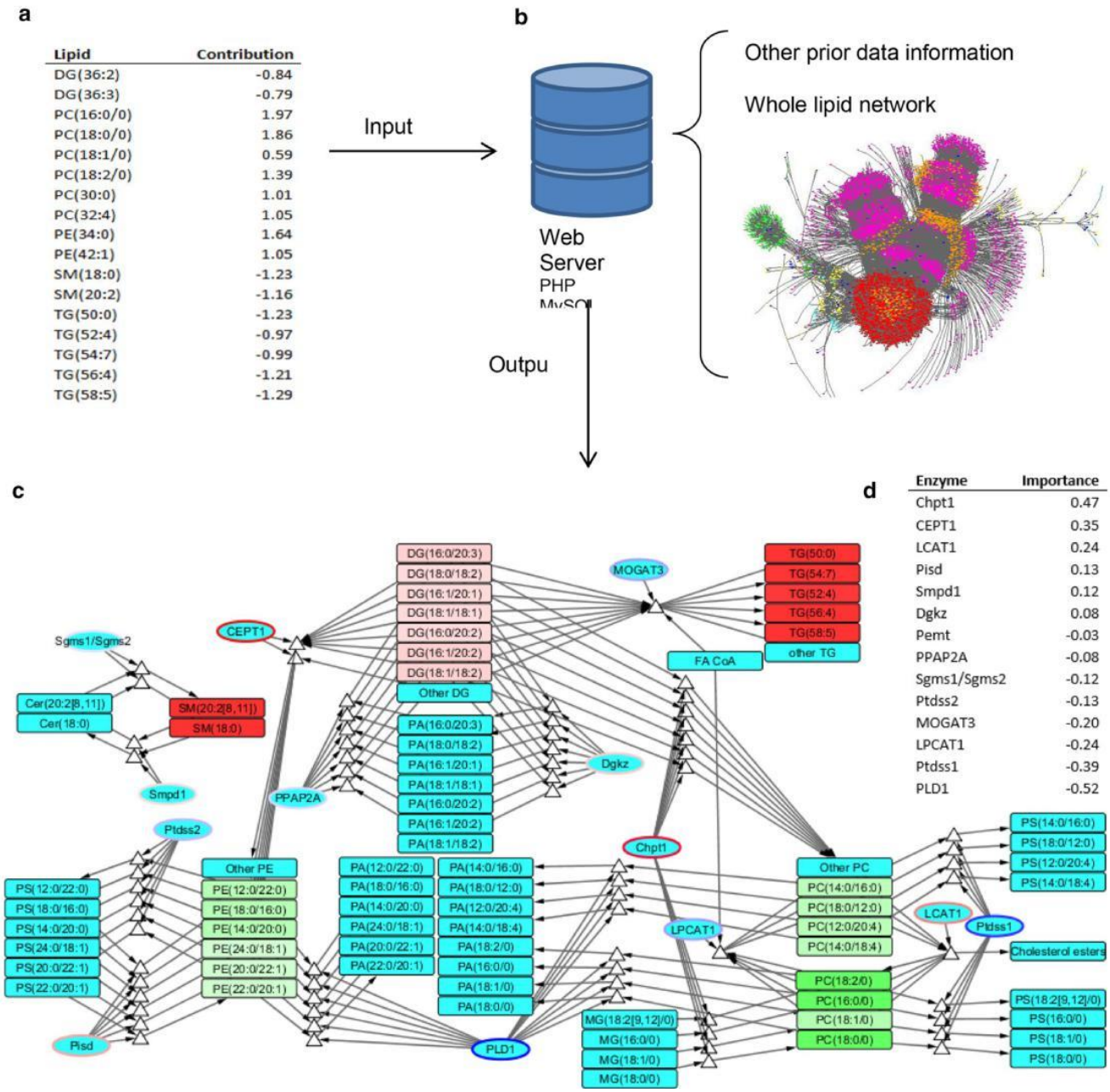


Fig. 1 The Lipid network method. **a** Input list of lipids with OPLS-DA model scores. **b** Depicts the web server that contains scripts and the lipidomic network, in which node colours *green, dark blue, orange, red* and *white* respectively correspond to glycerophospholipids, sphingolipids, diglycerides, enzymes and specific reactions. **c** Derived subset of lipid metabolism linking the input perturbed metabolites; metabolite, reaction and enzyme nodes are respectively represented by *rectangles, triangles* and *ellipses*; input metabolite nodes are respectively coloured a *brighter shade* or *green* or *red* depending on the extent to which their levels have been elevated or decreased, or *blue* if they correspond to nodes introduced during the subset retrieval; the borders of enzyme nodes are *shaded blue* or *red* respectively to depict the extent of positive or negative Relative Importance Score (the sign corresponding to the increase or decrease in enzyme activity). **d** The list of retrieved enzymes ranked by Relative Importance

network can be seen in Supplementary Fig. 1, and further details of the network's properties are described in Supplementary Tables 1–3. The Supplementary Information Section 9 outlines extensions to this method that can be arried out by suitably experienced bioinformaticians.

To show the power and utility of the method, it has been applied to a published dataset that contains lipid profile for women with polycystic ovary syndrome (PCOS) compared

with matched controls (2). The contribution score of the lipids found in significantly different concentration was assigned from an OPLS-DA analysis (see Fig. 1). The output .sif and node-attributes files were imported into Cytoscape and the nodes manually stacked to aid visual interpretation of the network. This allows us to view the enzymes that have a higher impact on the profile.

In this example, 14 enzymes have been retrieved falling into two separate components, with the highest in absolute value of Relative Importance scores being for PLD1 (a phosphatidylcholine-specific phospholipase) and Chpt1 (a choline phosphotransferase). This suggests that the changes in plasma lipid profile associated with PCOS may be domi-nated by a reduction in Chpt1 activity and/or an increase in PLD1 activity. The former is responsible for changes in DG and PC, and PLD1 is responsible for changes in PE and PC. Since there is equilibrium between the PA and DG species, a simple hydrolysis would convert one into the other, the perturbation of Pld1 could by itself explain the majority of observed changes.

The list of enzymes provided by this metabolic insight can be used to query other biomolecular network tools to identify potential common regulators of these lipids. 11 of them were present in BIOGRID, 4 linked by ELAVL1 and 2 by ILF3, which both protect target mRNAs from degradation (see Supplementary Fig. S7). Querying GeneMania revealed consistent patterns for regulation of Chpt1 and Pisd by microRNAs recognising the target sequence TGC TTGCT, and Dgkz, Ppap2a and Smpd1 by the transcription factors MAZ, NFAT and STAT5B (Supplementary Fig. S8). Unfortunately, both regulators have an enzyme whose activity goes in the opposite direction to the rest, so it is difficult to assert any strong hypotheses at this stage.

4 Conclusions

To conclude, the methodology presented here provides a way to move from tables of perturbed lipid levels to a visualization of the metabolic network that could cause these changes and thence to potential regulators. It also provides a metric (*Importance*) to rank the positive or negative contribution of each enzyme to produce the changes seen.

Acknowledgements FCP, SR and DB acknowledge funding from EU FP7 Grant 305815 (D-BOARD).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Open Access This article is distributed under the terms of the Crea-tive Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Cottret, L., Wildridge, D., Vinson, F., Barrett, M. P., Charles, H., Sagot, M. F., et al. (2010). MetExplore: A web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Research, 38*, W132–W137. doi:10.1093/nar/gkq312.

Doerks, T., Copley, R. R., Schultz, J., Ponting, C. P., & Bork, P. (2002). Systematic identification of novel protein domain families associated with nuclear functions. *Genome Research, 12*, 47–56.

Floegel, A., Stefan, N., Yu, Z., Mühlenbruch, K., Drogan, D., Joost, H.-G., et al. (2013). Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes, 62*, 639–648. doi:10.2337/db12-0495.

Gibellini, F., & Smith, T. K. (2010). The Kennedy pathway—de novo synthesis of phosphatidylethanolamine and phosphatidylcholine. *IUBMB Life, 62*, 414–428. doi:10.1002/iub.337.

Holmes, E., Loo, R. L., Stamler, J., Bictash, M., I. K. S. Yap, Chan, Q., et al. (2008). Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature, 453*, 396–400. doi:10.1038/nature06882.

Horne, A. B., Hodgman, T. C., Spence, H. D., & Dalby, A. R. (2004). Constructing an enzyme-centric view of metabolism. *Bioinformatics, 20*, 2050–2055. doi:10.1093/bioinformatics/bth199.

Kanehisa, M. (1996). Toward pathway engineering: A new database of genetic and molecular pathways. *Science & Technology Japan, 59*, 34–38.

Nguyen, T., Aparicio, M., & Saleh, M. A. (2015). Accurate mass GC/LC-quadrupole time of flight mass spectrometry analysis of fatty acids and triacylglycerols of spicy fruits from the Apiaceae family. *Molecules, 20*, 21421–21432. doi:10.3390/molecules201219779.

Rhee, E. P., Cheng, S., Larson, M. G., Walford, G. A., Lewis, G. D., McCabe, E., et al. (2011). Lipid profiling identifies a triacylglyc-erol signature of insulin resistance and improves diabetes predic-tion in humans. *The Journal of Clinical Investigation, 121*, 1402– 1411. doi:10.1172/JCI44442.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: A general repository for interaction datasets. *Nucleic Acids Research, 34*, D535–D539. doi:10.1093/ nar/gkj109.

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E. A., Glass, C. K., et al. (2007). LMSD: LIPID MAPS structure database. *Nucleic Acids Research, 35*, D527–D532. doi:10.1093/nar/gkl838.

Titz, B., Boué, S., Phillips, B., Talikka, M., Vihervaara, T., Schneider, T., et al. (2015). Effects of cigarette smoke, cessation and switching to two heat-not-burn tobacco products on lung lipid metabolism in C57BL/6 and Apoe-/- mice—An integrative systems toxicol-ogy analysis. *Toxicological Sciences, 149*, 441–457. doi:10.1093/ toxsci/kfv244.

Wang-Sattler, R., Yu, Z., Herder, C., Messias, A. C., Floegel, A., He, Y., et al. (2012). Novel biomarkers for pre-diabetes identified by metabolomics. *Molecular Systems Biology, 8*, 615. doi:10.1038/msb.2012.43.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., et al. (2013). HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research, 41*(D1), D801–D807.

2.3 Supplementary information

2.3.1 Manual curation

Before using the python script to generate the global lipid network, a network skeleton is required. A skeleton network consisting of the generic reactions associated with enzymes of lipid metabolism was obtained from Lipid Maps in conjunction with publications of lipid metabolism. The python script, named Network-Multiplicator.py, was then used to generate the global lipid network by creating reactions specific to all the alternative substrate molecules and associating these reaction nodes with the enzyme, catalyzing them. Two reactions have been excluded from the lipidomic network: Lysophosphatidylcholine Acyltransferase 1 (LPCAT1) and lecithin-cholesterol acyltransferase (LCAT1), which are both involved in PC-LysoPC conversion. The reason for this is that they add so many links that many irrelevant PC nodes are returned. To compensate for this, the subset-extraction algorithm adds these reactions back in when any LysoPC and PC is found in the subnetwork generated. In this way, the program still reveals the potential role of these enzymes in producing the observed lipidome perturbations.

2.3.2 Lipidomic network properties

Table 2.1: Number of lipids and families in the total network. The count was done using the some excel string separation to isolate the characters for each family.

Families	**Subfamily**	**Count**
Fatty acyls **(126)**	Fatty amides	44
	Eicosanoids	36
	Fatty esters	23
	Fatty acid conjugates	23
Glycerolipids **(3124)**	MG	44
	DG	1056
	TG	2024
Glycerophospholipids **(3575)**	PC	825
	PS	275
	PA	825
	PE	825
	PI	275
	PG	550
Sphingolipids **(290)**	Ceramides	88
	SM	44
	Other sphingolipids	158
Others lipid metabolites **(446)**		446
Total sum of species		**7561**

The network contains other nodes apart from the presented in table 2.1, around 18,000 reaction nodes and approximately 120 enzymes.

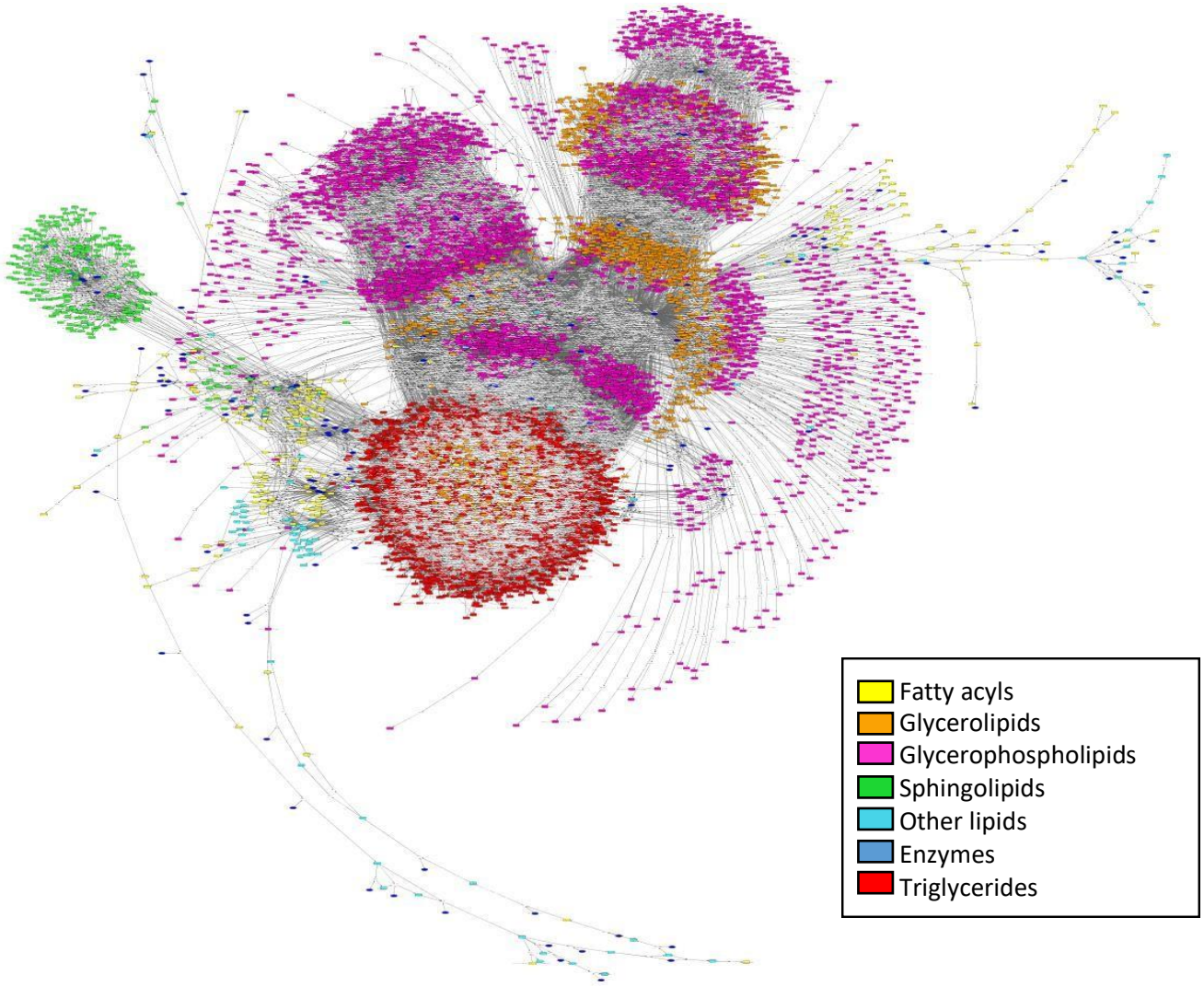


Figure 2.1: The whole network of human lipid metabolism. The box denotes the colour scheme of nodes, excluding reaction nodes coloured white and hence not visible. The links between enzyme, lipid and reaction nodes are coloured grey. The nodes and edges have been positioned using the "organic layout" algorithm within Cytoscape. There is a clear separation of sphingolipids (in green) from other lipid classes. The mono- and di-glycerides (orange) form precursors for various phospholipids and other lipid types.

Tables 2.2 and 2.3 respectively present the top lipid and enzyme nodes ranked by vertex degree, along with associated network properties: stress (the number of the shortest paths between all nodes that cross the specified node), degree (the number of adjacent neighbours of a node) and betweenness (the node stress divided by the number of alternative paths of the same length). The final column states their ranking according to stress.

Table 2.2: Lipid node metrics, presenting the values sorted by stress.

Name	Stress	Degree	Betweenness	Rank Stress
5Z,8Z,11Z,14Z-eicosatetranoyl CoA	1.93E+10	363	0.022	23
Hexadecanoyl CoA	1.32E+10	363	0.013	30
Octadecanoyl CoA	1.32E+10	363	0.013	29
Oleoyl CoA	1.29E+10	363	0.012	33
Docosanoyl CoA	1.30E+10	362	0.013	32
7Z,10Z,13Z,16Z-Docosatetraenoyl CoA	1.28E+10	362	0.012	34
13Z-Docosenoyl CoA	1.28E+10	362	0.012	35
8Z,11Z-eicosadienoyl CoA	1.25E+10	362	0.012	39
Linoleoyl CoA	1.28E+10	361	0.015	36
4Z,7Z,10Z,13Z,16Z,19Z-Docosahexaenoyl CoA	1.19E+10	361	0.012	45

Table 2.3: Enzyme node metrics same structure as table 2.2.

Name	Stress	Degree	Betweenness	Rank Stress
MOGAT3	8.79E+10	7200	0.379	2
CEPT1	2.68E+11	1500	0.278	1
PEMT	2.9E+10	924	0.055	19
PPAP2A	2.24E+09	600	0.036	92
PLD1	5.97E+10	600	0.026	8
AGPAT4	5.32E+09	600	0.019	61
UG7	1.23E+09	600	0.008	242
CDS1	6.54E+09	324	0.025	57
CDIPT	1.64E+09	324	0.010	116
CHPT1	1.65E+09	324	0.006	115

Most of the top-ranked metabolites in Table 2.2 are associated with the later stages of fatty-acyl chain elongation as well as the formation of membrane lipids. This is reasonable since these are central to lipid metabolism. Their stress rankings are high but not remarkably so.

In the case of the enzymes, however, the vertex degrees and betweenness are much higher, most notably regarding MOGAT3 (involved in triglyceride formation) and CEPT1 (PE formation from diglycerides). Given that these enzymes can carry out many different reactions, it is possible that they will appear more often in perturbed lipid profiles; something that should be taken into consideration when drawing conclusions about the processes producing the observed profile changes.

Table 2.4: List of the genes present on the lipid human pathway More information of this enzymes can be found in the excel. These are all protein encoding genes so the name and is shared with the enzyme.

ACACA	ALOX15B	CYP1A2	EBP	FDPS	PLA1	PTGS1	SPTLC1	BCO1
ACACB	ALOX5	CYP2	ELOVL1	FNTA	PLA2G4A	PTGS2	SPTLC2	COX2
ACAT2	ASAH1	CYP2C	ELOVL2	FNTB	PLA2G4B	RCE1	SQLE	CYCS
ACOT2	ASAH2	CYP2C8	ELOVL3	FVT1	PLA2G5	SC4MOL	TBXAS1	CYP26
ACOX1	BACAT	CYP2E	ELOVL5	GGPS1	PLD1	SC5D	TM7SF2	CYP27A1
ACS	BDH2	CYP2J	ELOVL6	GNPAT	PLD4	SCD1	UGCG	CYP39A1
ACSL1	CBR1	CYP2J2	EPHX2	GPD1L	PMVK	SCD2	UGT8A	CYP3A
ACSL3	CDIPT	CYP2U	ETNK1	GPD2	PNPLA3	SCD3	ACOT8	CYP46A1
ACSL4	CDS1	CYP4F3	FAAH	LTC4S	PPAP2A	SGMS1	ACOX2	CYP4A11
AGPAT4	CEPT1	CYP51	FAAH2	PAM	PPAP2B	SGMS2	AKR1C4	CYP7A1
AGPAT6	CERK	CYT C	FABD	PCYT1A	PPAP2C	SGPL1	AKR1D1	CYP7B1
AGPS	CH25H	DEGS1	FADS1	PCYT2	PTDSS1	SMPD1	ALDH1A1	CYP8B1
AKR1C3	CHKB	DEGS2	FADS2	PDSS1	PTDSS2	SOAT1	ALDH1A2	DGAT
ALOX12	CHPT1	DGKZ	FAM213B	PDSS2	PTGDS	SOAT2	AMACR	DHRS4
ALOX12B	CPT1	DHCR24	FASN	PEMT	PTGES	SPHK1	AOX1	EPT1
ALOX15	CPT2	DHCR7	FDFT1	PISD	PTGIS	SPHK2	BAAT	GGT1
LASS6	SPHK2	LSS	UGT8	MVK	CERS3	RETSAT	CERS6	SLC27A5
LPL	SPTLC1	MOGAT3	CERS1	DHDDS	CERS4	RPE65	SOAT2	SOAT1
LRAT	SPTLC2	MVD	CERS2	PNPLA4	CERS5	SCP2	SPHK1	
GPX6	HMGCL	HSD17B4	HMGCR	HSD17B7	IDI1	LASS2	LASS4	
GPAM	HPGDS	HSD3B7	HMGCS1	NSDHL	LASS1	LASS3	LASS5	

More information about these genes, link to NCBI description and more found in https://github.com/Ferran-Casbas/Lipid-Network, even for the UG (unidentified Genes), those are genes that have not been yet identified in human, but it is existence is known in other species.

2.3.3 Worked Example

A list of lipids with significantly different concentrations in patients with Poly-Cystic Ovary Syndrome (PCOS) vs controls has been published (Haoula et al., 2014). Please observe Table

2.5. They are divided into those that are perfectly identified (where Lipid Map IDs have been assigned) and imperfectly identified (where the word 'General' has been inserted). The Contribution values are the output from the OPLS-DA model which are used to identify the perturbed metabolites. In this instance, negative numbers refer to elevated levels in the patients. This data is submitted into the Lipid Network Creator web form at http://lipidnetwork.nottingham.ac.uk/ as shown in Figure 2.2. The Scored box should be ticked, then the Create Network button is clicked.

Note, SM(20:2[8,11]) is a precisely defined metabolite. Although it does not have a Lipid Maps ID, it can be placed on the list with lipids with definitive identities.

Table 2.5: Perturbed lipid concentrations in PCOS patients for human plasma. The table also present their lipid maps ID and their contribution values obtained in the OPLS-DA analysis

Lipid	Lipid Maps ID	Contribution
PC(16:0/0)	LMGP01050018	1.97
PC(18:0/0)	LMGP01050026	1.86
PC(18:1/0)	LMGP01050029	0.59
PC(18:2/0)	LMGP01050035	1.39
SM(18:0)	LMSP03010001	-1.23
SM(20:2[8,11])	-	-1.16
DG(36:2)	General	-0.84
DG(36:3)	General	-0.79
PC(30:0)	General	1.01
PC(32:4)	General	1.05
PE(34:0)	General	1.64
PE(42:1)	General	1.05
TG(50:0)	General	-1.23
TG(52:4)	General	-0.97
TG(54:7)	General	-0.99
TG(56:4)	General	-1.21
TG(58:5)	General	-1.29

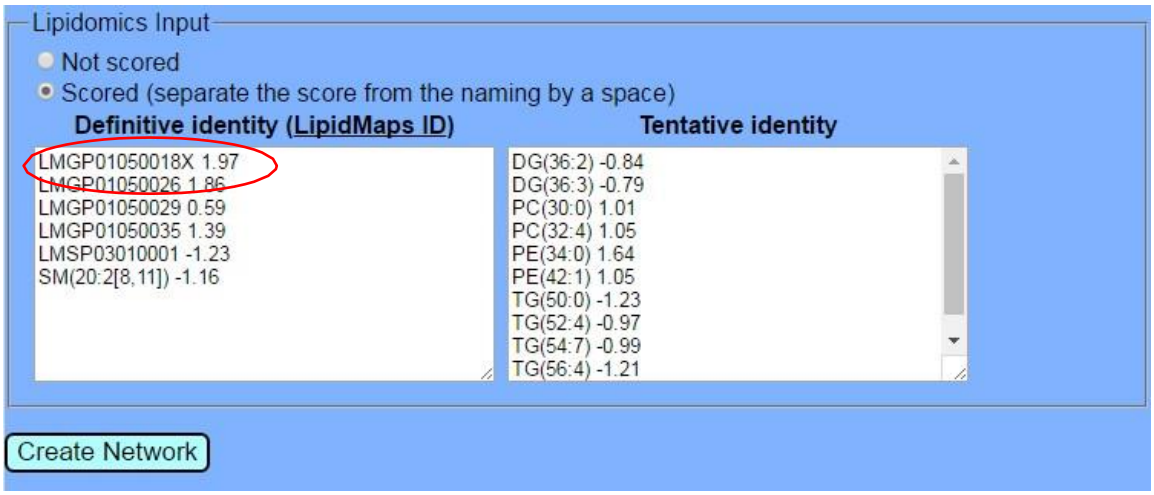


Figure 2.2: Screenshot of the Lipid Network Creator input boxes. Because Contribution scores are available, the "Scored" tick box has been selected. There is a typo mistake marked with a red ellipse. N.B. a single white space should separate lipid names/IDs from their scores.

The web tool returns a page containing a table of metabolites with their associated node names from the Lipid Network (see Figure 2.3). Unassigned lipids, shown in red text, are separated out for the user to check and amend until all the input data are in the correct format. It is possible that a lipid is not in the network, in which case we encourage the user to send an email with details, so that it might be included.

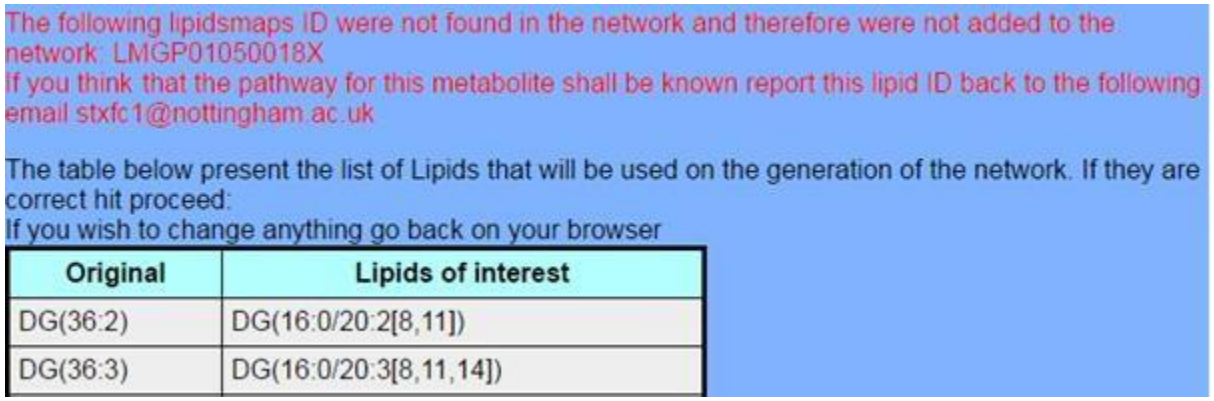


Figure 2.3: Lipid checking page. Red text indicates that there are issues with the input names.

When all is correct, the 'Submit' button is clicked again. It may take several minutes for the results page to appear. This contains the option to download a Zip file that contains the network file (in .sif format), node attributes (as a text file), instructions for network visualisation using Cytoscape (in README.docx, also see Figure 2.4). These files can also be viewed directly through the web browser via the relevant hyperlinks. The downloaded files

should be extracted and the instructions in the Read-me file followed. This leads to a network resembling that of Figure 2.5.

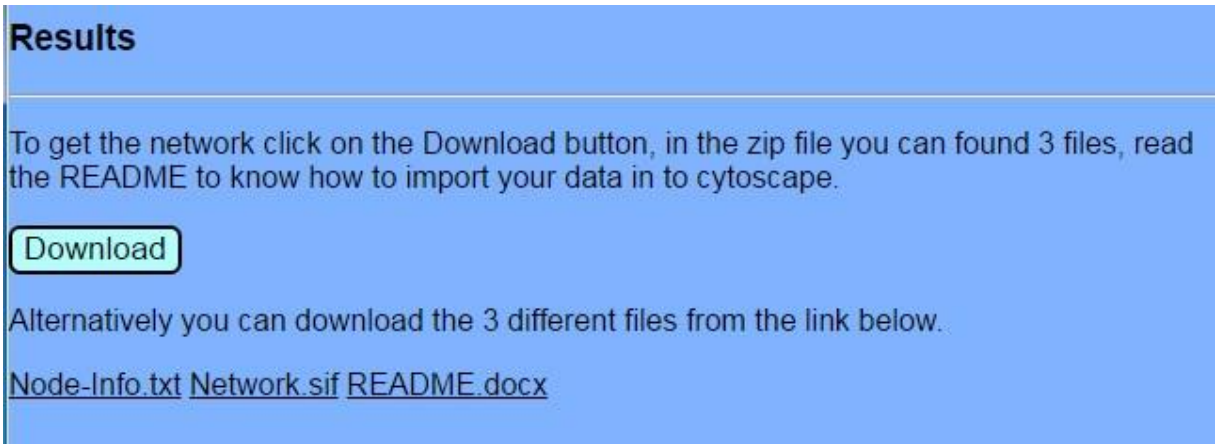


Figure 2.4: Lipid Network Creator results page

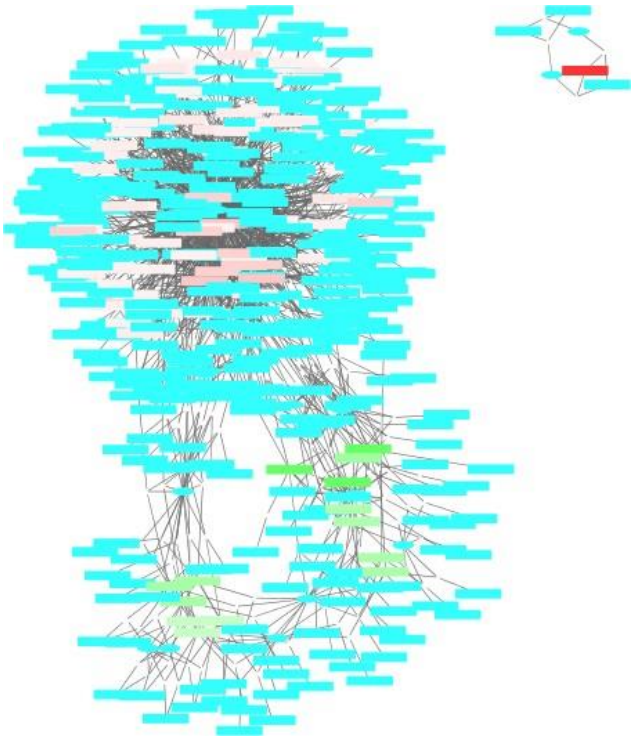
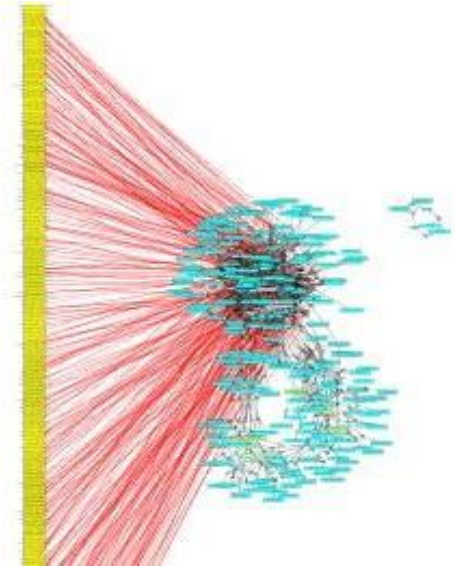


Figure 2.5: Initial Cytoscape layout. Metabolite, reaction and enzyme nodes are respectively represented by rectangles, triangles and ellipses; input metabolite nodes are respectively coloured a brighter shade or green or red depending on the extent to which their levels have been elevated or decreased, or blue if they correspond to nodes introduced during the subset retrieval. The borders of enzyme nodes are shaded blue or red respectively to depict the extent of positive or negative Importance score (the sign corresponding to the increase or decrease in enzyme activity).

Then, using the “align and distribute tool” found in the layout menu in Cytoscape, all the lipid and reaction nodes from the same family are grouped together. For example, the triglyceride nodes can be selected by typing “tg” in the search box and pressing enter. This will result in the TG nodes changing to yellow. From the tool panel, click on one of the vertical stack options to align the nodes which can then be moved to one side of the network (see Figure 2.6 A). For convenience, tall stacks can be reordered into rectangular blocks of nodes for ease of movement later. This process is then repeated for other metabolite families (See Figure 2.6 B).

A



B

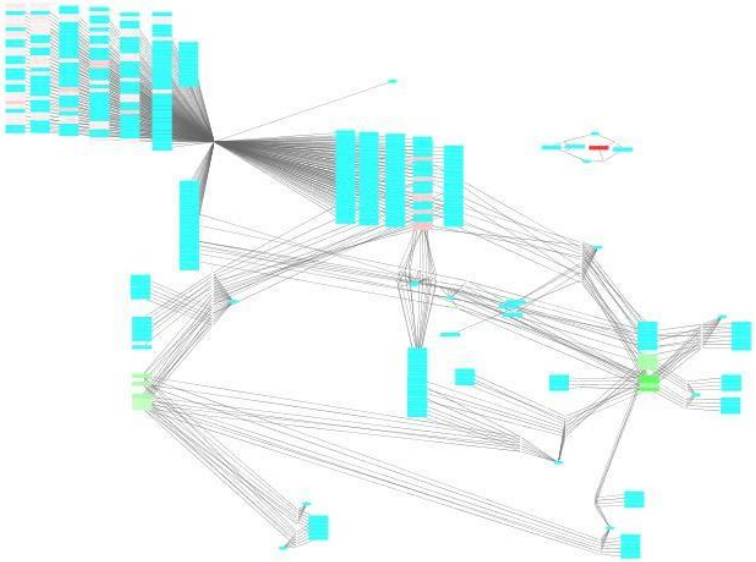


Figure 2.6: TG stack. A. The network from Figure S6 has had the triglycerides selected, stacked and moved to the left. The TG nodes are coloured yellow and their associated edges red. Otherwise the colour scheme is the same as for Figure 2.5. B. The network after further stacking of metabolites.

The network now contains nodes corresponding to metabolites whose concentrations are not thought to vary as well as many reaction nodes for each enzyme. These can be grouped together and collapsed into single nodes. From a given family of metabolite nodes, the blue nodes can be selected, then a right mouse click (or ctrl + mouse-pad click on Apple machines) provides a menu in which the nodes are first defined as a group. The same steps can then be used to collapse these nodes into a single “group node”, named, for example, “other TGs”. This node in the table panel should then be assigned the Type “Metabolite”, and the Basic and Importance scores of 0.

Sets of common reaction nodes can be grouped and collapsed in the same way. The main difference is that the Type of node should be redefined as “Reaction Node”. These steps eventually lead the network layout to something easier for the casual reader to understand (see Figure 2.7).

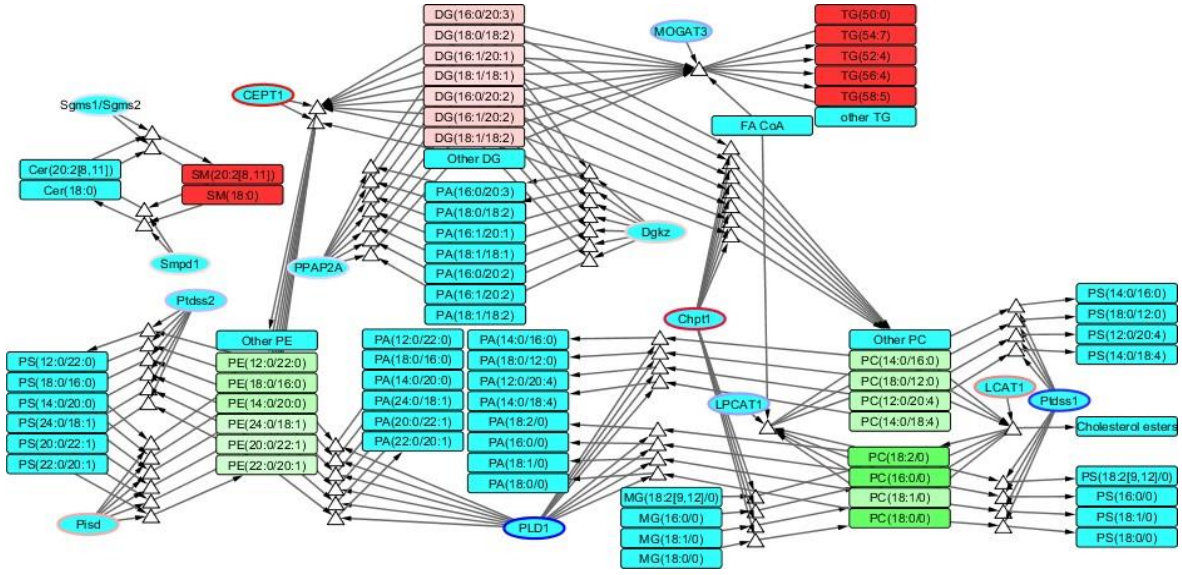


Figure 2.7: Final network layout. Same colour patterns as S4. Also, the shape of the node indicates its type, being rounded is an enzyme, square metabolite and triangles are Reaction nodes. The outer colour in the enzymes indicates their importance value, meaning that red is expected under activated enzymes and blue are active enzymes.

2.3.4. Search for potential regulators

The NodeAttributes&Score file can be filtered to find the enzymes/genes of importance. These IDs can then be used to find potential common regulators by first be pasting them into a txt file with all letters converted to upper case. This provides a set for querying interaction and regulatory databases.

The BIOGRID is an excellent catalogue of protein complexes and regulatory interactions. For the readers' convenience, a subset of BIOGRID (containing only the immediate neighbours of the lipid enzymes) has been generated. This network can be opened in Cytoscape, the nodes with Importance scores selected (using the submenu Select=>Nodes=>From ID List file…. and selecting the text file for the enzyme list), selecting first neighbours (using the submenu Select=>Nodes=>First Neighbours of Selected Nodes=>Undirected), and creating a daughter menu. Rearranging these results in Figure 2.8 Note, MOGAT3 is not present in the release of BIOGRID used here.

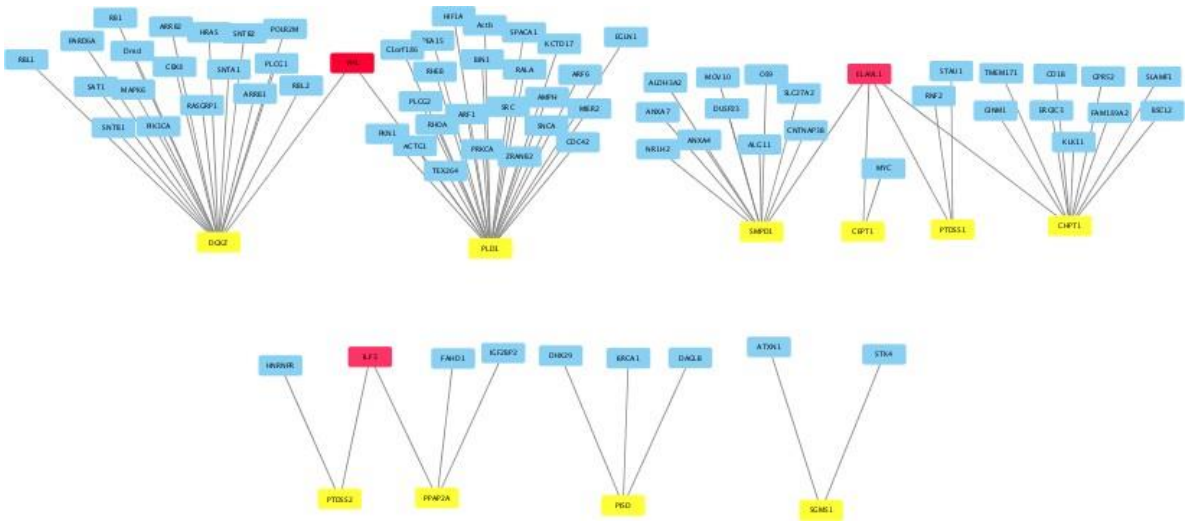


Figure 2.8: First neighbours from the human BIOGRID. The PCOS query nodes are shown in yellow. Blue and red nodes correspond to proteins in BIOGRID which interact with 1 and >1 PCOS nodes respectively. The latter could represent co-regulators. The edges denote interactions catalogued in BIOGRID.

The figure shows that PLD1 interacts with more proteins than any other, which might be expected of an enzyme involved in many different reactions. ELAVL1 is an mRNA-stabilising

protein shown to bind to the transcripts for CHPT1, CEPT1, SMPD1 and PTDSS1 (Abdelmohsen et al. 2009). It could play a role here by stabilising the transcripts for the first three enzymes whose activity goes up. Why the activity of PTDSS1 goes down remains unclear. Likewise, ILF3 is a co-regulator of PTDSS2 and PPAP2A (Havugimana et al. 2012), again at the mRNA level, but their activities also go in opposite directions.

Other interaction databases can also be queried. GeneMANIA is one, which includes data on gene and microRNA-mediated regulation and can be accessed via Cytoscape. The gene-regulatory subset for the PCOS gene list is shown in Figure 2.9. This suggests possible roles for MAZ-Q6, NFAT-04 and STAT5B transcription factors, with the latter two possibly explaining the increased activity of SMPD1, DGKZ, and DGKG, a non-catalytic sub-unit of the DGK complex. Moreover, there is a possible role for microRNA regulation by MIR15, 19, 424 or 497. A decrease in their levels could contribute to the elevated activities of CEPT1 and PISD.

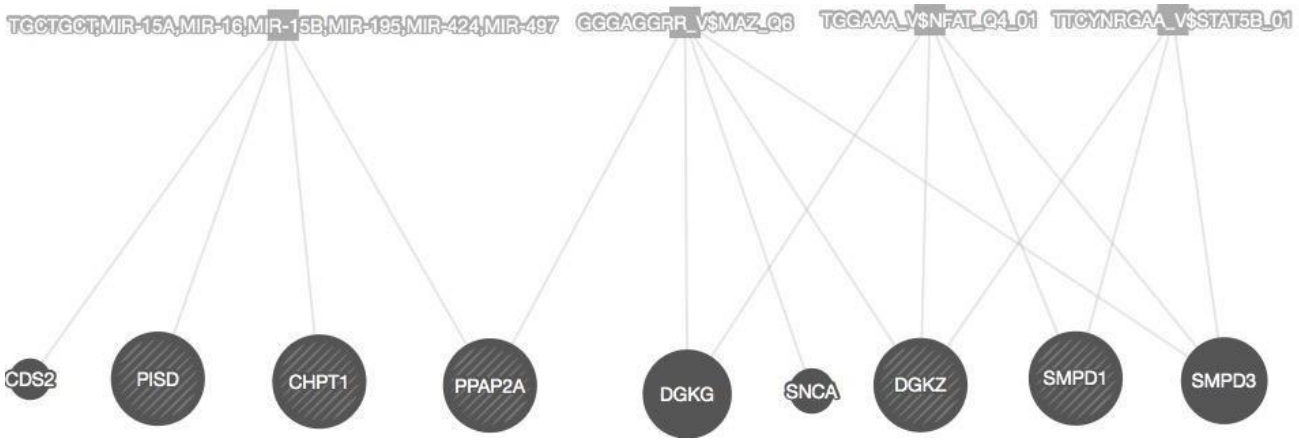


Figure 2.9: First gene-regulatory and miRNA neighbours from GeneMANIA of the PCOS Important enzymes.

2.3.5. Advanced techniques

There are several further ways in which an experienced bioinformatician can use the holistic lipidomic network generated here. The first would consist of merging this network with others, such as BIOGRID, so that metabolic and regulatory pathways can be viewed simultaneously. Although a powerful approach, the resulting subnetworks can still be too large to interpret easily. A second way concerns looking beyond the immediate enzyme neighbours of perturbed metabolites to the next adjacent enzymes, as this enables potential synthesis and degradation pathways to be more clearly seen, but again the subnetworks

become much larger. Where gene expression data (or other omics) are available, they can be used in a third way to colour code subnetworks to see if gene expression changes can account for the observed lipid profile changes.

2.4 Scoring

One of the most complex parts of the global lipidomic methodology is the script that scores the enzyme nodes. This sub-section will describe the problems that arose and their solutions.

The problem faced during some of the first datasets tested was that some networks were large (more than 2000 nodes) due to the large number of input lipid species. These networks were difficult to visualize, making it impossible to detect by eye the enzymes that play a major role. To solve this, different network metrics were investigated, such as centrality, betweenness, stress, vertex degree and average path length. Intuitively, stress might be the most useful, as it is the value that indicates, for a node, the number of shortest paths crossing it. However, this metric was less than satisfactory, as it overly emphasised enzymes linked to ambiguously defined metabolites (see Figure 2.10).

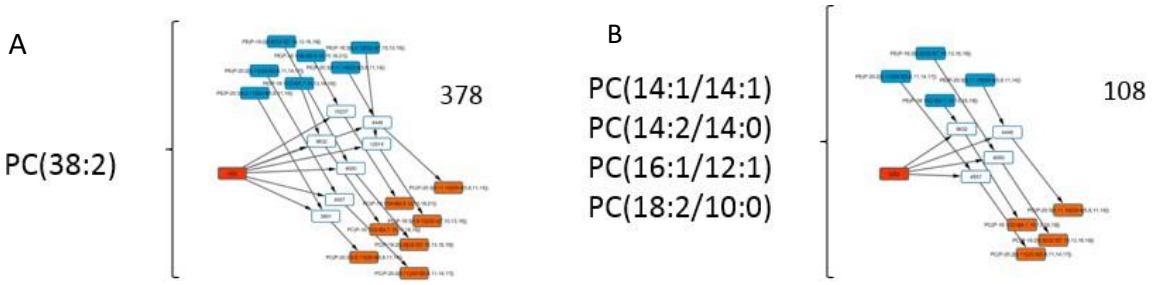


Figure 2.10: Effect of metabolite-term specificity on stress values. Blue, white, orange and red nodes respectively correspond to substrates, reactions, products and the enzyme catalysing the different chemical reactions. Arrows indicate the linkage between the four types of node, with enzymes having an edge to every reaction that they catalyse. (a) represents an ambiguously defined phosphatidylcholine, consisting of 38 carbon atoms and 2 double bonds; while (b) shows the extracted subgraph when 4 precisely defined metabolites (having the same ambiguous definition) are used instead. The numbers on the right of each panel are the stress values as calculated in Cytoscape (Doerks et al. 2012).

In this example, the input PC(38:2) for the left network is a generic node. The software determines all the possible lipids that fit that definition, finds them in the network and then links the associated reactions and enzymes. In the right-hand panel, the input provided was 4 identified lipids. An ambiguously defined metabolite results in more nodes being found than might be present in the biological sample, but they are given the same weighting as the precisely defined metabolites. Clearly, an alternative network metric is needed. It must not only take this ambiguity into account, but also the size of the perturbation as output from the statistical analysis. This can also be positive or negative depending on whether the metabolite levels are elevated or diminished.

The score for enzyme nodes should be a function of the scores for the lipid nodes. From multivariate analysis of the output lipid profiles, suitable values include VIP (Validity Indicator Profile), Contribution (from OPLS-SA) and Welch's t test score. The latter scores might have to be multiplied by -1 to indicate where their levels have decreased. Figure 2.11 shows that, apart from a +/- sign, t test and VIP scores are roughly equivalent. On the other hand, other scores do not fit this well; observe Appendix 2A.

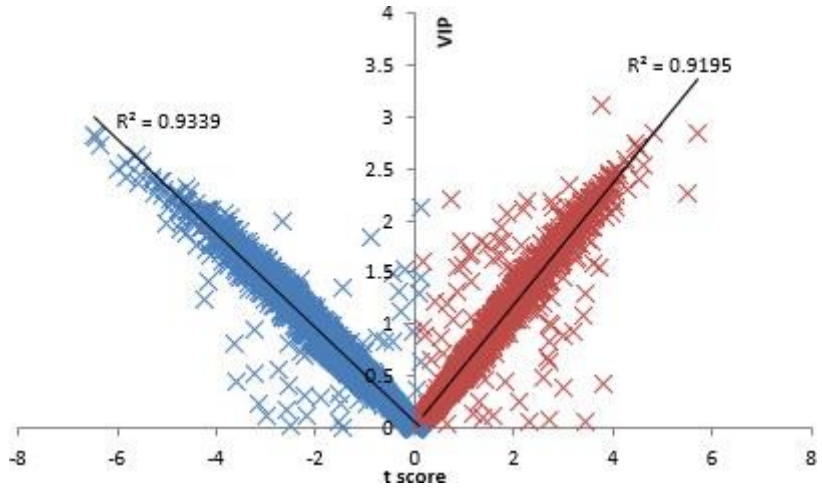


Figure 2.11: Correlation between VIP and Welch's t test scores. The VIP and t scores for ~4000 lipids from a human plasma study on Osteoarthritis have been plotted using blue crosses.

A lipid found at a higher concentration could be due to either an activation of the enzyme(s) catalysing its synthesis, or inhibition/repression of the enzymes, catalysing its conversion to other molecules. The inverse applies to lipids found at lower concentrations. Hence, the values for a given metabolite can be passed to its nearest enzymes (via its reaction nodes).

Enzymes linked either to major perturbations of a single metabolite or minor perturbations of a larger number of metabolites end up with larger values and are of greater interest. In this work, these scores have been defined as *Importance*.

However, simply adding up the values of every lipid binding to an enzyme will result in values that vary between experiments simply because the number of input lipids is different. Hence, *Relative Importance*, which divides these input values by the number of input lipids, provides a metric that is independent of the number of molecules obtained from the lipidomic analysis. Relative Importance values are usually <1. An enzyme would only acquire a score of 1 if it explained all the perturbed lipid levels, i.e. no other enzymes were involved. The algorithm was initially successful, but one infrequent issue emerged with scoring, for which there was not an easy solution. Figure 2.12 presents 3 working examples and one that causes a problem.

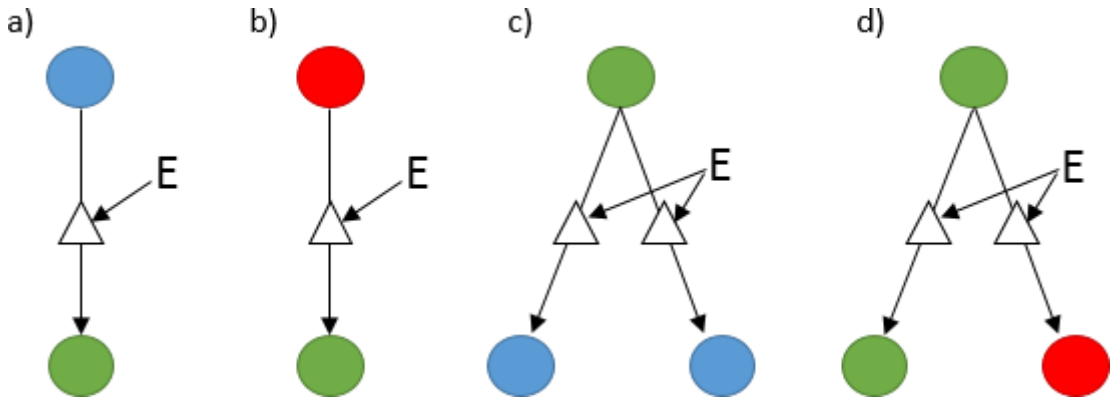


Figure 2.12: Small network cases causing an issue with enzyme scoring. Circles are lipids, triangles are reaction nodes, and the E symbolizes enzymes. In blue the neutral nodes, no score, in red nodes with -0.1 score and in green the nodes with score +0.1. MOGAT 3 is the perfect example of the case d.

A change in the level of a reaction product may reflect the activity of the enzyme(s) producing it. Thus, the score for a product node is simply added to the value for a reaction node. A change in level of a substrate node, however, is the inverse. For example, a lower substrate score indicates a more active reaction. Hence substrate scores are subtracted from reaction scores – subtracting a negative score results in addition. In Figure 2.12a, the green reaction-product node has a positive score, so its reaction (and subsequently, enzyme) score is

incremented by 0.1. For case (b), both substrate and product nodes have perturbed levels. The former is lower, and the latter is higher, leading to a reaction score of 0.2

Case (c) is slightly more complex. The script keeps a list of the lipid nodes that have already been used to score a reaction node. So, in this case, both reaction nodes get scored -0.1. When the score of the enzyme is calculated, it reveals that both nodes have been scored by the same active lipid. The program, therefore, automatically rejects the second node as it would be a duplicate, making the final score of the enzyme -0.1.

However, an issue arises in case (d). The left-hand reaction node gets a score of 0, whereas the right-hand reaction node gets a score of -0.2. Being the sum of the reaction scores, the enzyme score becomes -0.2. However, the desired value should be +0.1 (for the left-hand reaction product) – 0.1 (for the elevated substrate) -0.1 (for the right-hand reaction product) = -0.1. For this reason, a completely different algorithm had to be designed for calculating *Importance* and *Relative Importance*.

The final scoring method starts by dividing the input lipid scores by the number of perturbed lipids, to give relative values. Then, for each metabolite, the global network is queried to find the reaction nodes and thence enzymes to which it is linked. The set of enzymes with their reaction nodes are stored in a dictionary. Figure 2.13 shows the way to calculate Relative Importance scores. The algorithm has been implemented in PHP, because a web platform is much easier to use, more accessible and does not require download of any kind. This can be found in the GitHub, presented in the publication.

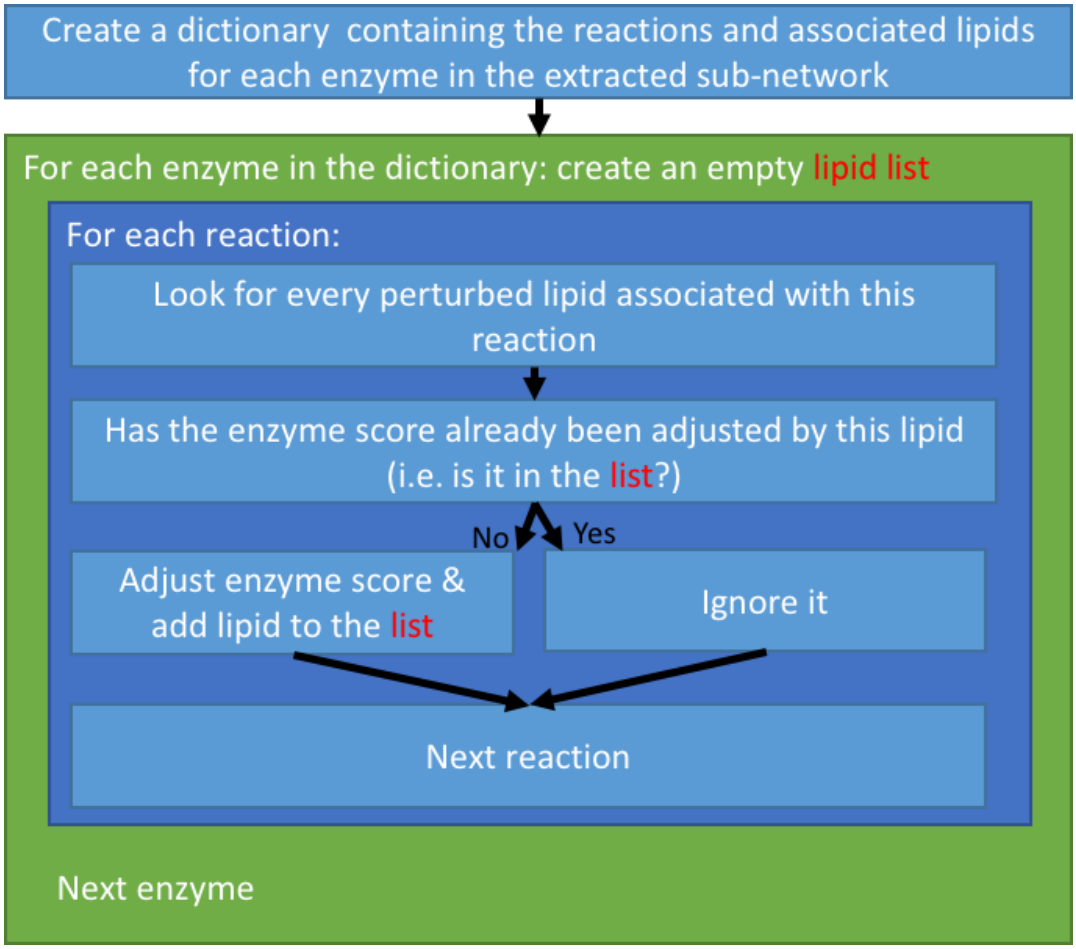


Figure 2.13: The Relative Importance scoring function. The white text on a pale blue background outlines the steps executed by the program, with arrows indicating the order in which they are carried out. The lipid list is denoted by red text, and the nested loops for each enzyme and reaction are respectively shown by green and blue boxes.

For the first enzyme in the dictionary, an empty lipid list is created. For the first reaction, its substrate molecules are taken in turn. If the metabolite is already in the lipid list, the next substrate molecule is considered, otherwise the lipid score is subtracted from the enzyme score (because a higher substrate score indicates less enzyme activity) and the metabolite name is added to the lipid list. Next, the reaction product molecules are taken in turn. If the metabolite is already in the lipid list, the next product molecule is considered, otherwise the lipid score is added to the enzyme score (because a higher product score indicates more enzyme activity) and the metabolite name is added to the lipid list.

At this point the next reaction is taken, and the above steps repeated. Once all the reactions for the specified enzyme have been processed, the enzyme name and its Relative Importance

score are stored in one array, and the enzyme, reaction and metabolite IDs as nodes and their associated edges (for laying out subnetwork) in a separate array. The above steps are repeated for each of the remaining enzymes in the dictionary. Finally, the output consists of the content of the latter two arrays. Enzymes with elevated concentrations of substrates and products will have low Relative Importance, as it does not seem to be the responsible of observed metabolite increases on both sides of its reaction. They can be considered as intermediates in a pathway, rather than a controlling step.

2.5 Testing the methodology

Before this network-analysis approach can be recommended, it should be tested to determine its ability to produce appropriate results. This has been done in three ways:

1) Random sets of input data, this should result in the appearance of enzymes whose frequency only reflect the network topology
2) Introducing increasingly large errors into an actual input dataset, reveals the amount of tolerance allowed before different enzymes are prioritised
3) The software should correctly identify the enzymes affected in an artificially perturbed dataset

These are studied in turn in the next three subsections.

2.5.1 Output from random input data

Random data were applied in two different ways – with and without scores associated with the metabolites. Being the simpler case, the latter was carried out first; 12 different sets of 20 random metabolites were used individually to query the lipidomic network. Using 20 lipids is the regular number on lipidomic experiments.

Table 2.5 shows the enzymes that appear in the output subnetworks and the frequency that they appear over the 12 experiments. It is interesting to note that some enzymes always appear, due to their central position in the network, Table 2.5. These results confirm that random data give subnetworks largely consistent with the overall topology. They also

highlight that care must be taken with real input data. Since the enzymes listed in Table 2.6 often occur by chance, they might not be so relevant to any study.

Table 2.6: Frequency of enzymes appearing in subnetworks from random input

Enzyme	Frequency	Vertex degree	UG7	11	600
MOGAT3	12	7200	CIPT	11	324
CEPT1	12	1500	PLD1	10	600
PPAP2A	12	600	Chpt1	10	320
PEMT	11	920	UG2	10	280

After the unscored experiment, the next experiment to do was the scored one. This experiment revealed some interesting information that allowed us to detect a problem in the first scoring methodology. The information for that experiment can be found on the Appendix 2.B

Random score data allows the calculation of relative-importance scores that might arise at random, and as a result, provide a threshold above which enzyme nodes can be significant. This time, 12 sets of randomly chosen metabolites were randomly assigned scores of +1 or -1, and enzymes counted when their relative importance went beyond three different thresholds (+/-0.2, +/-0.25 and +/-0.35). The results can be seen in Tables 2.8 and 2.9, that no enzymes had a score beyond +/-0.35. Using scores reduced the number of enzymes found at random, and these were restricted again to those with high vertex degrees. At a threshold of +/-0.25, only MOGAT3 and PEMT appeared on a small number of occasions. Hence, this is an appropriate cut-off value for finding relevant enzymes in real datasets.

Table 2.7: Frequency of enzymes with relative importance of 0.2 or more.

MOGAT3	4	33%
CEPT1	2	17%
PEMT	1	8%
UG1	1	8%

Table 2.8: Frequency of enzymes with relative importance of 0.25 or more.

MOGAT3	2	16%
PEMT	1	8%

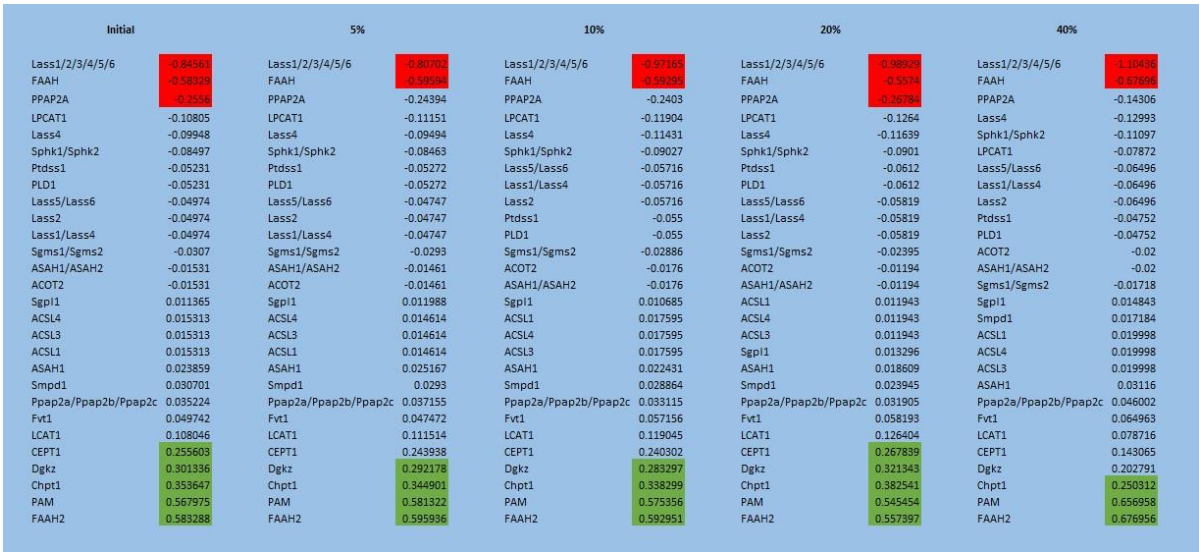
2.5.2 Error tolerance

In this test, different levels of error have been applied to datasets. The results were compared for each level of error with the initial output data to see at what point they completely differ. The IPF analysis has been used in the first instance, because it is one of the largest datasets. The enzymes with a Relative importance score > 0.25 will be compared with the results of applying a 5%, 10%, 20% and 40% error on the input metabolites scores.

To generate the datasets with error, a random value was generated either to add or subtract the % that was required for each sample. For example, for the 20% error, the formula on the excel spread sheet would be: newvalue = oldvalue(1+rndbetween(-20,20)/100)

Results

To start the number of significant genes changing (relative importance score higher than 0.25) was counted. This process was done for all the levels of error and it is presented on figure 2.14. The image shows how in the initial conditions 3 genes were found significantly under expressed and 6 genes were found significantly over expressed. After 5% error, they become 2 and 5, already changing the number of significant genes. Still the genes keep the order they were set on. When the error gets to 40%, genes start to drastically switch places.



Initial		5%		10%		20%		40%	
Lass1/2/3/4/5/6	-0.84561	Lass1/2/3/4/5/6	-0.80702	Lass1/2/3/4/5/6	-0.97165	Lass1/2/3/4/5/6	-0.98929	Lass1/2/3/4/5/6	-1.10436
FAAH	-0.58329	FAAH	-0.59594	FAAH	-0.59295	FAAH	-0.5574	FAAH	-0.67696
PPAP2A	-0.2556	PPAP2A	-0.24394	PPAP2A	-0.2403	PPAP2A	-0.26784	PPAP2A	-0.14306
LPCAT1	-0.10805	LPCAT1	-0.11151	LPCAT1	-0.11904	LPCAT1	-0.1264	Lass4	-0.12993
Lass4	-0.09948	Lass4	-0.09494	Lass4	-0.11431	Lass4	-0.11639	Sphk1/Sphk2	-0.11097
Sphk1/Sphk2	-0.08497	Sphk1/Sphk2	-0.08463	Sphk1/Sphk2	-0.09027	Sphk1/Sphk2	-0.0901	LPCAT1	-0.07872
Ptdss1	-0.05231	Ptdss1	-0.05272	Lass5/Lass6	-0.05716	Ptdss1	-0.0612	Lass5/Lass6	-0.06496
PLD1	-0.05231	PLD1	-0.05272	Lass1/Lass4	-0.05716	PLD1	-0.0612	Lass1/Lass4	-0.06496
Lass5/Lass6	-0.04974	Lass5/Lass6	-0.04747	Lass2	-0.05716	Lass5/Lass6	-0.05819	Lass2	-0.06496
Lass2	-0.04974	Lass2	-0.04747	Ptdss1	-0.055	Lass1/Lass4	-0.05819	Ptdss1	-0.04752
Lass1/Lass4	-0.04974	Lass1/Lass4	-0.04747	PLD1	-0.055	Lass2	-0.05819	PLD1	-0.04752
Sgms1/Sgms2	-0.0307	Sgms1/Sgms2	-0.0293	Sgms1/Sgms2	-0.02886	Sgms1/Sgms2	-0.02395	ACOT2	-0.02
ASAH1/ASAH2	-0.01531	ASAH1/ASAH2	-0.01461	ACOT2	-0.0176	ACOT2	-0.01194	ASAH1/ASAH2	-0.02
ACOT2	-0.01531	ACOT2	-0.01461	ASAH1/ASAH2	-0.0176	ASAH1/ASAH2	-0.01194	Sgms1/Sgms2	-0.01718
Sgpl1	0.011365	Sgpl1	0.011988	Sgpl1	0.010685	ACSL1	0.011943	Sgpl1	0.014843
ACSL4	0.015313	ACSL4	0.014614	ACSL1	0.017595	ACSL4	0.011943	Smpd1	0.017184
ACSL3	0.015313	ACSL3	0.014614	ACSL4	0.017595	ACSL3	0.011943	ACSL1	0.019998
ACSL1	0.015313	ACSL1	0.014614	ACSL3	0.017595	Sgpl1	0.013296	ACSL4	0.019998
ASAH1	0.023859	ASAH1	0.025167	ASAH1	0.022431	ASAH1	0.018609	ACSL3	0.019998
Smpd1	0.030701	Smpd1	0.0293	Smpd1	0.028864	Smpd1	0.023945	ASAH1	0.03116
Ppap2a/Ppap2b/Ppap2c	0.035224	Ppap2a/Ppap2b/Ppap2c	0.037155	Ppap2a/Ppap2b/Ppap2c	0.033115	Ppap2a/Ppap2b/Ppap2c	0.031905	Ppap2a/Ppap2b/Ppap2c	0.046002
Fvt1	0.049742	Fvt1	0.047472	Fvt1	0.057156	Fvt1	0.058193	Fvt1	0.064963
LCAT1	0.108046	LCAT1	0.111514	LCAT1	0.119045	LCAT1	0.126404	LCAT1	0.078716
CEPT1	0.255603	CEPT1	0.243938	CEPT1	0.240302	CEPT1	0.267839	CEPT1	0.143065
Dgkz	0.301336	Dgkz	0.292178	Dgkz	0.283297	Dgkz	0.321343	Dgkz	0.202791
Chpt1	0.353647	Chpt1	0.344901	Chpt1	0.338299	Chpt1	0.382541	Chpt1	0.250312
PAM	0.567975	PAM	0.581322	PAM	0.575356	PAM	0.545454	PAM	0.656958
FAAH2	0.583288	FAAH2	0.595936	FAAH2	0.592951	FAAH2	0.557397	FAAH2	0.676956

Figure 2.14: The image presents the list of all the genes present on the network generated. They are sorted according t the absolute score. In red and green we have the ones that present a score higher than 5 in absolute

value. The different columns are the results with different levels of error applied. It can be appreciated that the number of red and green changes from one to the other.

From the analysis, it has been concluded that more than 20% error in the score of the input metabolites starts to greatly affect the dataset, switching the order of some enzymes. Also, that depending on the threshold applied (relative importance cut off value) the results changed the number of significant enzymes changes.

The test will be repeated with at least 10 replicates using 5%, 10% and 20% error. 'The test will try to determinate if the amount of error we can deal with could be improved by increasing the relative importance cut off.

For each replicate, ten in total, the number of enzymes that moved from one category to the other have been counted. The categories being: significant or non-significant. The significance is set by their relative importance value. The threshold tested have been 0.2, 0.3, 0.4 and 0.55 on relative score. The table below presents the results of how many enzymes were found for each condition changing when compared to the unperturbed dataset, under the same important score threshold over the ten repetitions under those conditions.

Table 2.9: The table presents the results of how many enzymes stayed the same from their group when different error was applied to the data (out of the total 80), in brackets the % of how many stayed the same. The table also shows the effect of moving the threshold used to define the significance.

		Importance score threshold			
		0.2	**0.3**	**0.4**	**0.55**
ERROR	**5%**	74(92.5%)	78(97.5%)	79(98.75%)	78(97.5%)
	10%	75(93.7%)	76(95%)	77(96.25%)	72(90%)
	20%	72(90%)	75(93.75%)	74(92.5%)	71(88.75%)

The first observation was expected: lower error presents a smaller number of enzymes misclassified for any given cut off; the smaller the error, the less enzymes being misplaced.

To answer the question if it was possible to increase the score threshold to fight error; the answer is that it can be done to a certain degree. Setting it too high or too low means that the cut off will result in missing matches or getting false positives. These experiments have also confirmed that the convenient value for a relative score in future experiments is 0.25.

2.5.3 Artificial data test

The idea of this test is to generate an artificial lipid-profile dataset corresponding to the knockdown of a specific enzyme, and then look to see if the methodology finds that enzyme. To obtain this dataset, the control group from the PCOS experiment has been used as starting point, a set of calculated modifications has been applied to represent the enzyme knock down and some randomness added to correspond to the observed variability in metabolite profiles. To see the effect of the decrease in enzyme activity, Michaelis-Menten equations have been used, because they are the recognised approximation to enzyme kinetics (Kuby, 1991).

The generic enzymatic equation can be described as:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} E + P \qquad [1]$$

The first step is to combine the enzyme and the substrate to create the enzyme-substrate complex. The second step is the generation of the final product plus the enzyme. The formula of the concentration over time for this reaction is the presented below:

$$\frac{d[P]}{dt} = k_2[ES]^m - k_{-2}[E]^n[P]^p \qquad [2]$$

The enzymatic reactions are usually of order one we can eliminate the exponents. Also, assuming the general assumption of “steady-state kinetics” it can be assumed that the concentration of substrate complexed with enzyme over time keeps constant. The third assumption that is taken is the final product has a low affinity for the final product, which is the typical case.

$$\frac{d[P]}{dt} = k_2[ES] \qquad [3]$$

This is not the most optimal form of the equation since the concentration of [ES] can be hard to calculate. Using the first reaction we can modify the equation to the form:

$$\frac{d[P]}{dt} = \frac{k_2[E_{total}][S]}{K_m + [S]} \quad [4]$$

This is the classic development of the Michaelis-Menten equation. The main problem we face is that our reactions to the product can be converted to another product, so it is necessary to add that reaction in the equation. We added the conversion of the product P into another molecular species P_2 with the catalysing enzyme E_2. The Michaelis-Menten formula would look like:

$$\frac{d[P]}{dt} = \frac{k_{E_1}[E_1][S]}{K_{mE_1} + [S]} - \frac{[P_2]K_{mE_2}}{[E_2]k_{E_1} + [P_2]} \quad [5]$$

If the concentration of P_2 is very small in comparison with P, the subtracting term in the equation can be approximated to 0 and therefore ignored. Since the concentration of different lipid families is highly variable, it might be possible to find reactions where this approximation can be used.

The typical concentration of lipid families in humans is known (Quehenberger et al. 2011, Ridgway, 2016). From there it is possible to appreciate that the concentration of PE compared to PC and DG species is minimal. If we use PE and PC as P2 and P species, the last assumption becomes true. Hence, the reaction most suitable for the study is CHPT1 because it catalyzes the conversion of DG into PC, and since the next product of PC is the conversion to PE, a minimal specie, we can use equation 5 without the subtraction term.

The reaction that we have chosen to knockdown is the one catalyzed by CHPT1. Regarding the conversion of DG to PC, according to the deductions of the Michaelis-Menten formula, it can be assumed that the concentration of PC will be decreased by 50% and DG will increase by the same amount when the enzyme concentration is lowered by 50%.

Still, this is hypothetical and might depend on the half-life time of PE or other equilibrium like complexed lipids etc. This is the best approximation achieved since we do not have a dataset with genomic alterations its lipidomic profile.

The first step is to localize the DG and PC from over 6000 identified metabolites in the controls dataset. For this task, we have used a ready designed script that is able to look for masses in lipid maps and return their possible ID (Section 4.4). Every potential DG and PC has been treated with the appropriate change. However just applying this transformation would not be correct, since it would be easy for the program to track back those changes. Some randomness has also been applied to all the dataset to make the new data even more different than the initial state.

Several options of random error were used, PCA was used to observe the results of each attempt and verify that the new artificial data was different than the controls.

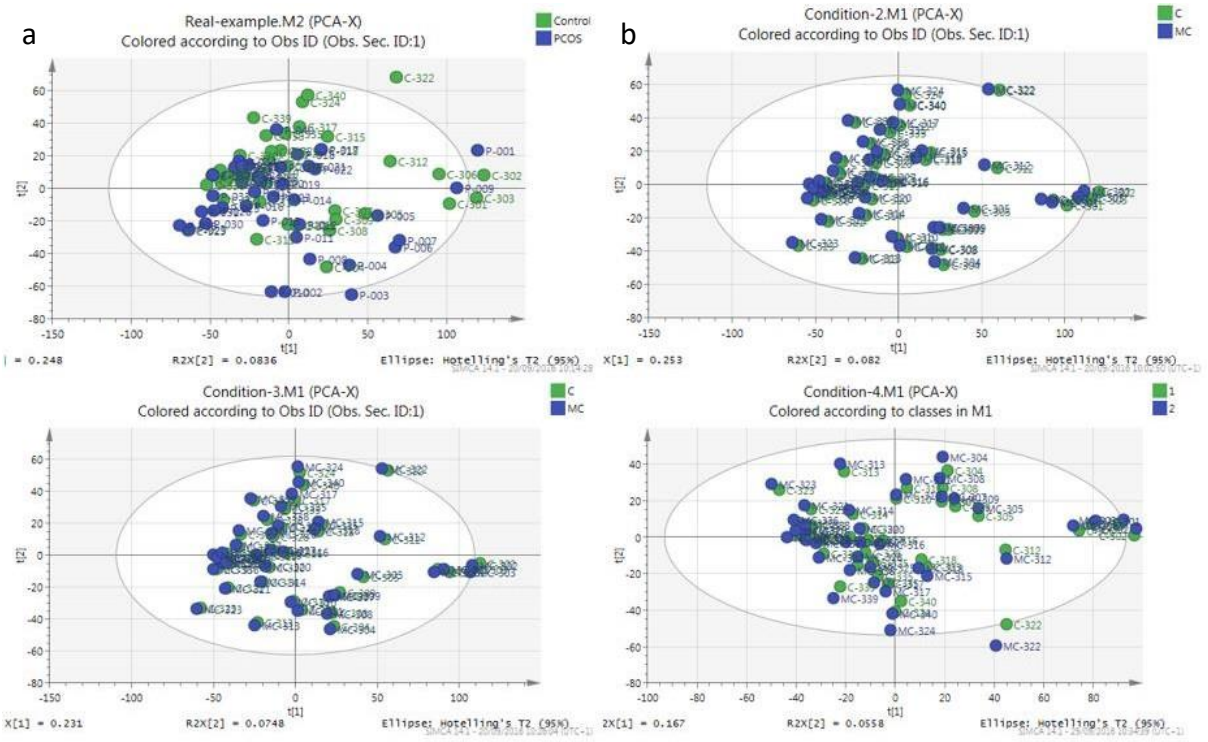


Figure 2.15: PCA plots of artificially perturbed lipid profiles. Panels (a) – (d) respectively represent the comparison of; in a controls vs PCOS, the other three b c and d are the different second, third and fourth conditions. Plots were generated using SIMCA (Umetrics), and controls are shown in green, while disease/perturbed controls are shown in blue.

The different conditions that we have tried are:

Condition 1: Fixed 10% error in all metabolites

Condition 2: Variable 0-30% error in all metabolites. Same error on lipids of the same mass.

Condition 3: Variable 0-40% error in all metabolites. All errors different.

Condition4: Variable 0-90% error in all metabolites.

All these conditions were applied also with the 50% increase of DG and the 50% decrease of PC. The condition 4 lead us to a PCA that did not present such strong clustering between the original and the modified sample like in the other examples.

Results

We proceed to do the OPLS-DA of condition 4 against control and get the list of the lipids that were found with a higher contribution score. The list of lipids was selected with two different ways. The selection of the top 10 lipids on contribution score (absolute value), and the top 5 in contributing to positive and top 5 contributing to negative score. Those lipids were identified and then feed in to the script to see what genes where picked up to be changing.

The top 10 lipids on absolute value were identified and the list is provided below. They are all PC, therefore when they are analyzed to see the enzyme that should be perturbed, the list of enzymes is all the enzymes around PC with their intensities of 1 or -1. The reason for being less than 10 metabolites is that some of the PC perturbed had an even number of C and are therefore not in the network.

Lipid	Contribution
PC(38:3)	-3.67843
PC(38:6)	-3.59013
PC(36:4)	-3.42994
PC(38:5)	-3.14275
PC(40:6)	-3.0113
PC(36:5)	-2.96857
PC(34:2)	-2.93355

Enzyme	R. Importance
Ptdss1	1
PLD1	1
PEMT	-1
Chpt1	-1

The top 5 positive and top 5 negative lipids in contribution scores have detected, 5 PC and 5 DG. After the analysis, the genes with the highest score is CHPT1 on top. Like in the other experiments, some of the lipids had an even number of FA chains or were duplicates.

Lipid	Contribution
PC(38:3)	-3.67843
PC(38:6)	-3.59013
PC(36:4)	-3.42994
DG(44:10)	2.3881
DG(34:0)	2.48981
DG(40:6)	2.70711

Enzyme	R. Importance
Chpt1	-1
PEMT	-0.58514
DGKZ	-0.41486
CEPT1	-0.41486
MOGAT3	-0.41486
PPAP2A	0.414856
Ptdss1	0.585144
PLD1	0.585144

The results have shown us that even after all the error added, the OPLSDA analysis is still able to pick up these PC and DG that we had excited, and therefore catch Chpt1 as the main responsible for the perturbation. Another resulting observation from this is that it is good practice to get the lipids changing from both ends; the ones that are going up and the ones that are going down.

2.6 Web usage statistics

The web usage statistics was recovered using Google Analytics. It is a freemium web analytics service, offered by Google, that tracks and reports website traffic. It allows you to detect who has gone into the website, from where, how long they have stayed as well as many other functions.

The total number of visitors that opened the website has been registered at 183 visitors of which 112 used the tool; that means they advanced to the next page, either introducing their own data or using the provided example data. The distribution of the different visitors is presented in figure 2.16. Nine visitors were recurring users, coming back more than once.

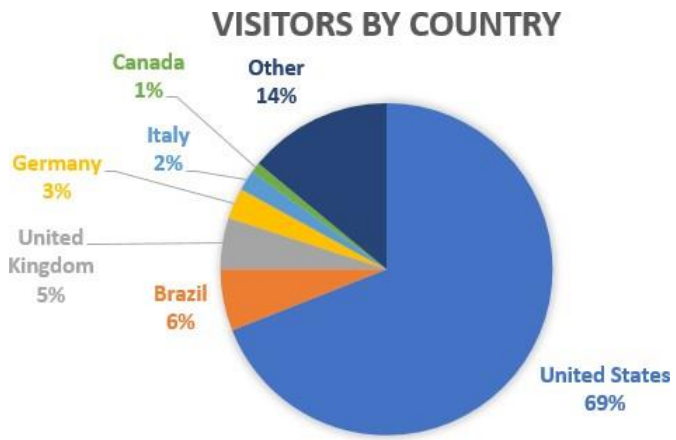


Figure 2.16: Visitors by country. The pie chart presents the total amount of visitors for each country the other is a distribution of over many other countries with less than 3 visitors

Top visitors distributed by city are: Atlanta 65, New York 22 and San Francisco 18. The average time of the users was 2 minutes, some of the longest connections lasted for 44 minutes.

2.7 Discussion

The methodology developed in this chapter has shown that it is possible to convert tables of perturbed lipid species into maps of the lipid metabolism. The main genes involved are scored and the results are then formatted in an easy-to-read way.

From the number of visitors, it is possible to see that this is a topic of interest, however the small number of those returning is probably due the low appeal of the website. Also, the time required to load the website from outside the UK is larger in comparison to the time it takes for the server to interact.

This methodology has been exposed to a series of tests that have improved the methodology and proven its effectiveness. Some tests have even provided an understanding of what the best way is to use the tool. It allowed us to set a score cut off to determine significant enzymes; to know how much error we should allow or how much data we should have, and that sometimes is important to consider both sides of the story, not only what is being increased, but also what is being decreased.

Some attempts to take the methodology output and use it to replace the current biomarker finding models have been done with non-remarkable results (more information in Appendix 3D). The most successful ways to use the information from the methodology are the ones described in the sub section 2.3.6.

The work on this chapter is usable for the majority of the lipidomic analysis. However, some experiments are done for a set of specific lipids, target methodology. While this analysis can be used for the targeted lipidomics methodology. Since the is a huge amount of targeted analysis in the arachidonic acid cascade and their importance in inflammation and pain, a specific methodology and study for that pathway will be done in the next chapter.

3. The oxylipins family

3.1 Background

This chapter focuses on the other major branch of lipidomics, targeted methodology, specifically for the inflammatory oxilipins. This methodology can quantify the lipids and has no problems of identification thanks to the use of internal standards. Run times are normally shorter than in global lipidomics and the data analysis much simpler.

The methodology has been used for the study of membrane structure (Yamamoto et al. 2017), but most commonly to find the effects that lipids have on animal disorders (Hijioka et al. 2016). The latter is what this chapter is going to attempt, to discover the hidden mechanism of oxylipins and try to find if there is an explanation for the observed profile in OA (Osteoarthritis). This involves both making a lipidomic network specifically for oxylipin synthesis and investigating potential protein interactions/associations and gene regulation of the enzymes concerned.

3.1.1 Aims and objectives

- Review literature on newly-discovered inflammatory mediators and create a lipidomic network for the oxylipins.
- Create an inflammation score calculator based on literature reviews and data mining.
- The generation of regulatory networks for the oxylipin-associated enzymes.
- Apply the new data analysis tool to existing targeted lipidomics data sets from pre-clinical and clinical studies to discover if oxylipin perturbations observed in OA and other conditions can be explained by known regulatory interactions.

3.1.2 Oxylipins

Oxylipins are a group of lipids that come from the oxidation or further conversion of polyunsaturated fatty acids. They play a major role, in a hormone-like manner, in signaling in many biological processes. In mammalian their main role is to be the regulator of cellular homeostasis and immune responses (Smith and Murphy, 2016). Oxylipins are presented in two different forms: the free form (these have been characterized to act locally and have a short life span) and the esterified form (Welte, 2015). This group has been found in the liver and adipose tissue; their function is not yet known. Hypothesis suggest its purpose is to

change the properties of membranes. Their three main enzyme pathways for the generation of oxylipins are though cyclooxygenase (COX), lipoxygenase (LOX) and cytochrome P450 (CYP).

- Cyclooxygenases: in these pathways, PUFAs are converted in to prostanoids (PG and thromboxanes). Prostanoids are characterized by their interaction with their most potent ligand, however cross reactivity does exist.
- Lipoxygenase: the pathway catalyses the formation of hydroxyl fatty acids. There are multiple LOX enzymes and they are classified and named by the structural location where they create the hydroperoxidation. Hydroxy fatty acids can be further catalysed into their keto species. The combination of different LOX and other CYP or COX enzymes can lead to the formation of di- or tri- hydroxy FA.
- Cytochrome P450: The enzymes that catalyse this pathway come from a big family of enzymes that are known for their hydroxylation and epoxidation activity. The oxylipins of the family mediate their effects via receptors, meaning they are able to penetrate the cell and act in an intracellular level.

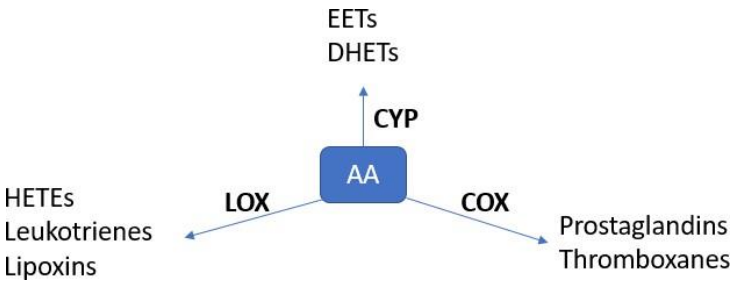


Figure 3.1: General eicosanoid pathway. The image presents at the centre the arachidonic acid, and its three main pathways.

<u>Prostaglandins in inflammation:</u>

Prostaglandins may function in both the promotion of inflammation and resolution. This inflammation normally affects joints, skin and eyes, and corticosteroids are used to inhibit its effects (Devlin, n.d.). There are four main active prostaglandins: PGE2, PGI2, PGD2 and PGF2α. During inflammation, concentration and profile of this lipids changes dramatically. The production of prostaglandins depends on the activity of their synthetases (COX). There are

two isoforms of these enzymes: COX1 and COX2. They both contribute to the generation of the homeostatic levels and inflammation. However, COX1 is closer related to homeostatic levels, while COX2 is more related to the inflammation stimuli.

PGE2

This is the most abundant PG in the body and it is synthesized from PGH2 by PGE synthase. It is an important mediator in many biological processes, for example: PGE2 is important in inflammation, contributing in multiple steps. It increases the blood flow to the tissue by arterial dilation (Funk 2001). It also generates pain by action on the peripheral sensory neurons. When PGE and their derivatives are deleted or inhibited, it has been observed that the inflammatory response has diminished (Kamei et al. 2004). PGE2 acts locally, by binding to cognate receptors like EP3 and EP4 being the most abundant and the ones with the highest affinity (Trebino et al. 2003). They can also have anti-inflammatory action, but it is typically only observed in allergic or immune inflammation, and they are usually balanced by pro-inflammatory actions of other prostaglandins. The pro inflammation action of PGE2 contributes to the regulation of cytokine expression profile of dendritic cells and bias T cell differentiation towards Th1 and Th2 (Egan, Lawson et al. 2004).

PGI2

PGI2 is one of the most important prostaglandins that regulates cardiovascular homeostasis. It is generated by the conversion of PGH2 by cytochrome P450. Once synthesized, it is released by vascular muscle cells and endothelial progenitor cells (Katusic, Santhanam et al. 2012. PGI2 gets rapidly converted by non-enzymatic processes to an inactive hydrolysis product. It is a potent vasodilator, inhibitor of palate aggregation, leukocyte adhesion and VSMC proliferation (Noda, Kariura et al. 2007). PGI2 is attached to IP receptors expressed in the kidney, liver, lungs, platelets, heart and aorta (Libby et al. 1988). Apart from its role in the cardiovascular effects, it is an important mediator of the oedema and is quickly generated in inflammation cells. In damaged tissue, it is rapidly generated and can be found in high concentrations in inflammatory milieus (Bombardieri et al. 1981).

PGD2

PGD2 is synthesised in the central nervous system and peripheral tissues (Jowsey et al. 2001). Its main function is in inflammation and homeostatic capacity (Jowsey et al. 2001). In

the brain, it is associated with the regulation of sleep and other CNS activities which include pain perception (Urade and Hayaishi 2011). In peripheral tissues, it is produced by mast cells and leucocytes (Arima and Fukuda 2011). The pro-inflammation effects of PGD2 appear mediated by the receptors DP1 (expressed on dendritic cells) and DP2/CRTH2. The prostaglandin is highly correlated with lungs (Hammad et al. 2007), allergies and atherosclerosis (Buja et al. 1983).

PGF2α

PGF2α is synthesized from PGH2 via PGF synthase and acts via the receptor FP. When the receptor is activated, it couples with Gq protein to elevate the intracellular concentration of free calcium. FP is the least selective of the prostaglandins receptors, and other species have been proven active with a lower affinity. $PGF_{2\alpha}$ has been detected to play an important role in the female reproductive system (ovulation, luteolysis, contraction of uterine smooth muscle and initiation of parturition)(Sugimoto et al. 1997), but it also has a significant role in contraction of arteries (Tesarova et al. 2013), myocardial dysfunction(Takayama et al. 2005), brain injury (Saleem et al. 2009) and pain(Kunori et al. 2009). The administration of this product leads to acute inflammation (Sugimoto et al. 1997).

Thromboxane

TXA2 is a very short life metabolite, only lasting 30 seconds, and is produced by activated platelets. It is synthesized from PGH_2 by TXS and is non-enzymatically degraded into biologically inactive TXB2. It is mediated by receptor TP which couples with Gq, $G_{12/13}$ that can regulate several factors (Barton 2011). It has the potential role as a pro-inflammation mediator in experiments of septic shock models with deletion of TP (Yamada et al. 2003).

3.1.3 Targeted lipidomics

Targeted lipidomics is the analysis of a set of lipids of interest. The detection and quantification of oxylipins is of great interest because of their important role in many diseases. Due to their very low levels in biological matrices, analytical methods that present high sensitivity and specificity are required for their quantification.

Liquid chromatography mass spectrometry (LC-MS) is the predominant technique used in oxylipin profiling. This is due to technological advances in MS ionization techniques, such as electrospray ionization (ESI), and column chemistry (sub-2 µm particles) which introduced the

use of ultra-high-pressure liquid chromatography (UHPLC), enabling the analysis of oxylipins with great resolution in the MS. Therefore, LC-MS/MS is the method of choice for quantification of oxylipins in a wide variety of biological matrices (Newman et al. 2002, Masoodi and Nicolaou 2006, Schmelzer et al. 2006, Deems et al. 2007). The main LC-MS/MS techniques,

principles and applications in eicosanoid analysis are described and reviewed here (Lundstrom et al. 2009).

Furthermore, there is an increased interest in omega-3 polyunsaturated fatty acids (n-3 PUFAs) and omega-6 polyunsaturated fatty acids (n-6 PUFAs), which exhibit both anti-inflammatory and pro-inflammatory effects, playing a "double-edged sword" role in OA. To this end, Baker *et al.* demonstrated that plasma levels of n-3 and n-6 PUFAs, which are influenced by diet, were detected by GC-MS and are related to altered structural findings in knees of OA patients (Baker et al. 2012).

Considering the above, oxylipin profiling of biofluids from OA studies monitored by MS methods has gained significant interest over the last few years. This approach has the potential to provide a deeper insight into the role of this complex network of oxylipins in OA pathophysiology, and to find and validate potential lipid biomarkers that will aid in early diagnosis of OA.

3.1.4 Why integrated network analysis is needed

The study of pain and inflammation is one of the widest and most extended fields on medicine. The resolution of inflammation is a common practice (taking an ibuprofen) and has been discovered thanks to the study of inhibitors to stop the inflammatory response. The control of the inflammation in chronical diseases however, is more complex. For this reason, it is necessary not only to understand the inhibitors for some key enzymes or the biomarkers involved in a disease, but also study why they happen and how.

Targeted lipidomic study for oxylipins has been done since 2003 (Lee et al. 2003), and the most common output are the potential biomarkers of the study, new methodology have taken step towards detecting particularly complicated lipids (Masoodi et al. 2008) or screening regulatory lipids (Buré et al. 2012).

While screening the profile of target metabolites is a useful and powerful method to characterize changes in the metabolic process (Griffiths et al. 2010; Lu et al. 2008), it is extremely hard to find a targeted lipidomic analysis that goes deeper into the pathway for the study of a disease. Most of the publications just revalidate further and further the biomarkers found with complex analytical techniques. This can be found in recent publications (Halade et al. 2018; Chen et al. 2017) and older target lipidomics (Wakelam et al. 2007). The problem might start with the first definition of lipidomics, where the full focus of it was on the problems in dealing with and analyzing such complex biomolecules (Spener et al. 2003).

What is needed for this technique to succeed is a methodology to combine lipidomic and pathway information. Additionally, since the Arachidonic pathway is small, in comparison to other metabolic routes, more layers of information can be added: protein-protein interaction, gene co-expression, regulation by transcription factors etc.

3.2 Literature review

Due to the lacking of an entire oxylipins pathway in the public domain, a new one was crafted. The information from KEGG was used as the main skeleton, it contains the three main pathways of COX, LOX and CYP. For other oxylipin species (resolvins, maresins and protectins), it was necessary to look for more recent publications (Ashton et al. 2018; Serhan, 2014; Li et al. 2016; Creelman et al. 2002; Mozurkewich et al. 2016) as well as additional reactions which were found on LIPIDMAPS.

The final pathway for the oxylipin family contains a total of 30 enzymes presented in table 3.1, and a total of 63 metabolites (table can be found on Appendix 3.A).

Table 3.1: List of all the enzymes present of the oxylipins pathway. The ones that have an * synthetize homodimer proteins.

AKR1C3	CYP2C	GPX6
ALOX12	CYP2C8	HPGDS
ALOX12B	CYP2E	LTA4H
ALOX15	PTGS1	LTC4S
ALOX15B	CYP2J	PTGDS
ALOX5	CYP2J2	PTGES
CBR1	CYP4F3	PTGIS
COX2 *	EPHX2	PTGS2*
CYP1A2	FAM213B	TBXAS1
CYP2	GGT1	

3.3 Pro-/-anti-inflammation score

A key feature of the oxylipins family members is their role in inflammation and anti-inflammation. They are widely discussed in section 3.2, where it was interesting to be able to give a profile of a case study in which you can know with a quick and broad calculation, if it was a pro inflammation or anti-inflammation result.

For this reason, the data of over 800 papers, compacted on several reviews (Serhan, 2005; Brouwers et al., 2015; Mika et al. 2016; Ricciottiet al. 2011), was collected containing experiments proving the inflammatory and anti-inflammatory effect of some of the oxylipins species. This data was collected in a spreadsheet format that can be used to import an oxylipin profile from LC-MS biofluid analysis and subsequently to calculate an inflammatory or anti-inflammatory score based on the level and type of the oxylipins found.

The excel sheet oxylipin-Inflamation-score.xlsx contains in the INPUT tab in the first sheet, only the cells in white should be edited so the excel does not malfunction. It is recommended to import the data from the Cytoscape map, but manual introduction is also supported. The results are displayed in the last sheet (Inflammation score), some additional information must be added, like the total number of identified metabolites and the Significance level set (there

is a recommended value on top). The excel sheet contains all the instructions and is full of annotations on how to do the full procedure and what things should be taken into consideration. The excel sheet example is presented in the Appendix 3.B

The value of inflammation that appears is a calculation of all the pro inflammatory oxylipins against all the anti-inflammatory oxylipins. Positive values indicate pro-inflammation and negative values indicate anti-inflammation behavior of the experiment against controls. The value indicates approximately the number of pro or anti-inflammation species. A value of -1.6 means that there are approximately 1 and a half species working towards anti-inflammation, while a value of 3.7 would mean that around 4 species are pro inflammatory.

3.4 Integrated network

So far, the only pre-added information in the pathways is the enzymes involved in the reaction. Sometimes this information is enough to find a problem with them, however for many cases another layer of information might be needed to discover what is happening.

All the enzyme information has been introduced in BioGRID and in GeneMANIA, with the idea to obtain what interaction these enzymes and their corresponding enzymes experience. BioGRID is an interaction repository with data compiled through comprehensive curation, containing over 1.5 million protein and gene interactions. A list of all the interactions presented can be found later in this section. GeneMANIA is a public database that finds related genes to a set of input genes. It then detects physical interactions, co-expression, co-localization, transcription factors, protein domain similarities and genetic interactions.

The different coding genes for the list of enzymes was introduced to GeneMANIA, and the results can be seen in Figure 3.2. The figure presents the different transcription factors (TF) surrounding the genes, which makes it easy for us to pick up the ones that are being regulated. For example, the two genes PTGS2 and PTGES are both regulated by the TF NFKB (Merico et al., 2010), and it is important to keep in mind that these two genes are the starting point of the whole PG synthesis. The other TF, HNF1α, CEBPβ and FREAC2 (known as FOXF2) are all known to be involved in the inflammatory response.

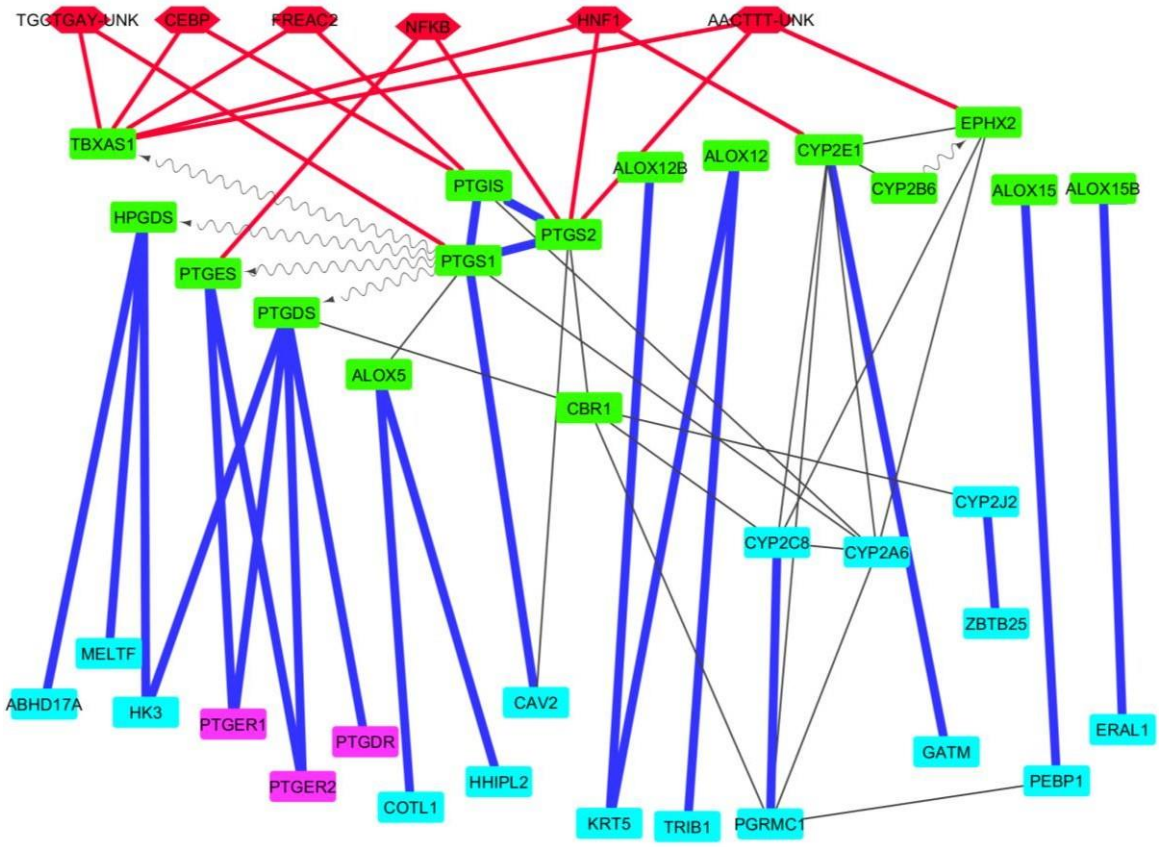


Figure 3.2: GeneMANIA output of the Oxylipin enzymes. This layout has been generated using cytoscape. Rectangular and hexagonal nodes respectively correspond to proteins and transcription factors, which are coloured red. The enzyme nodes are coloured green, purple or blue when they are enzymes in the oxylipin network, prostaglandin receptors or other proteins respectively. The red, blue, black and wavy lines respectively correspond to the gene-regulation, protein-protein interaction, intracellular co-localisation and metabolite intermediates. The latter specifically correspond to PGH2 (product of PTGS1/2 that is the substrate of 4 other enzymes) and EETs (product of CYP2B6 that are substrates for EPHX2). GeneMania also identified a link with the OLF1 transcription factor but has not been included here because it is specific to olfactory neurons, and hence irrelevant to this study.

The first thing that can be observed is that the enzyme TBXAS1 involved the synthesis of thromboxane is highly regulated, more than the others, all the TF found except NFKB are involved in that regulation. Another interesting observation is the fact that most of the regulations occurs for the first products of the AA derivate.

Another interesting thing that can be picked up by GeneMANIA is the fact that the receptors are bound to the catalyzing enzymes. For example, PTGES appears to be physically interacting with PTGER2 and PTGER1 (receptors for the PTGE molecule) the same thing happens with PTGDS and PTGDR. In the case of the PTGES, something even more interesting can be appreciated; if we take a closer look we can see that it contains a positive feedback loop,

something typical for stress response mechanism. Let's say that the loop starts at NFKB, then NFKB transcription factors activate PTGES transcription, this activation will lead to an increase in the production of PTGE, the lipid would bind to the receptor generating an increase the generation of the TF NFKB.

Working with BioGRID it was possible to find multiple interactions between the enzymes of our new oxylipin pathway; the Figure below presents all of them with a Prefuse Force direct layout. All the proteins that did not find any connection were manually removed.

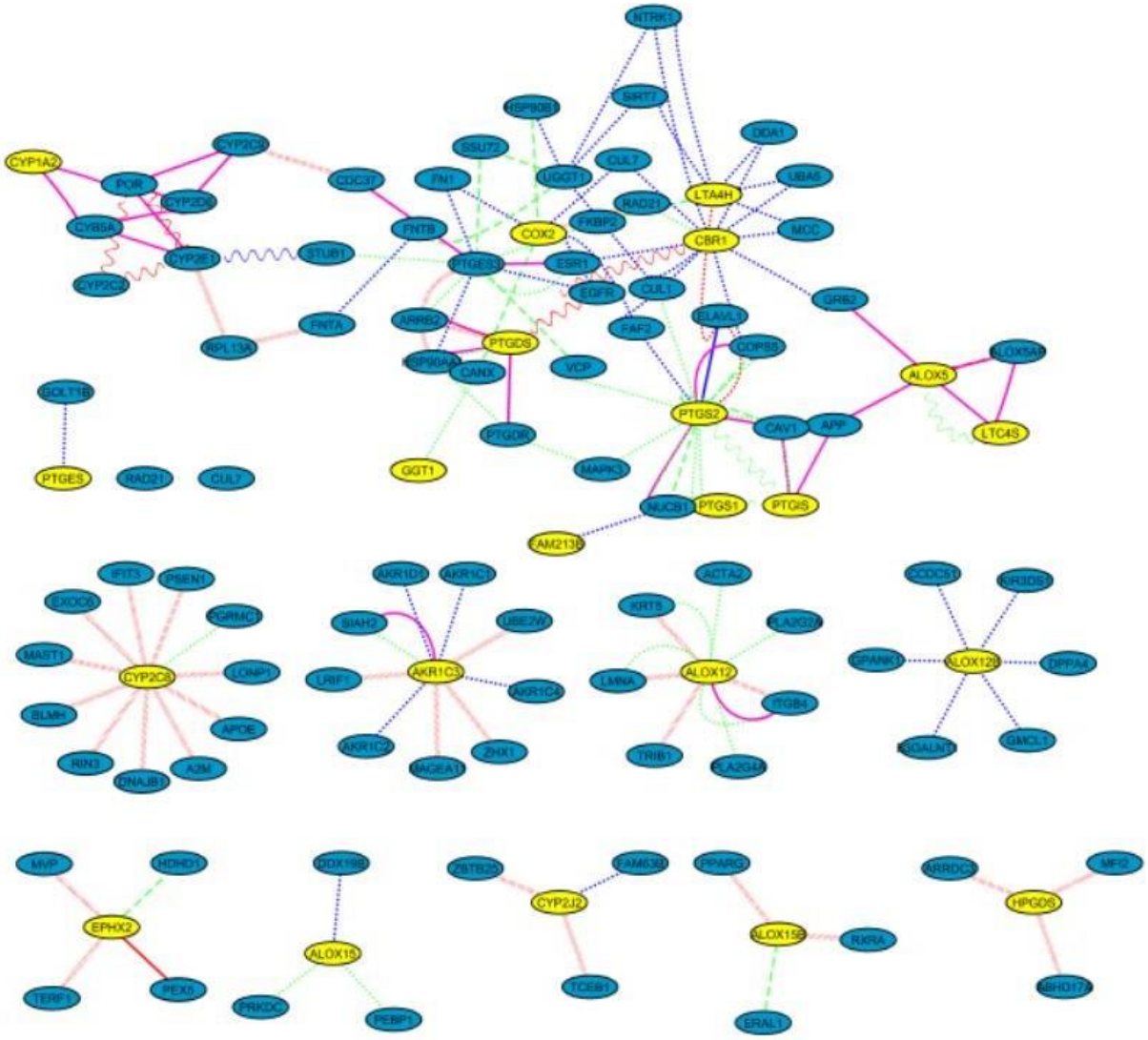


Figure 3.3: Prefuse Force Directed Layout of the nodes obtained from BioGRID. The initial list of enzymes is present in yellow, in blue the other enzymes catch by the program. The different edges have colour and line style based on the type of interaction. Blue dotted- Affinity capture-MS, Red dotted- Affinity capture-RNA, Green doted - Affinity capture-Western, Red dashed - Co-localization, Green sinewave – FRET and pink plain line - Reconstituted complex.

The data was edited, to make it easier to read the different types of interactions where painted in different colours and styles. A list of what they are will be described below to help follow the further discussion.

- **Affinity capture-MS:** An interaction is inferred when a bait protein is affinity captured from cell extracts by either polyclonal antibody or epitope tag and the associated interaction partner is identified by mass spectrometric methods. (Hein et al., 2015). They are present with a blue dotted line.
- **Affinity capture-RNA:** An interaction is inferred when a bait protein is affinity captured from cell extracts by either polyclonal antibody or epitope tag and associated RNA species identified by Northern blot, RT-PCR, affinity labelling, sequencing, or microarray analysis. They are present with a red dotted line.
- **Affinity capture-Western:** An interaction is inferred when a bait protein is affinity captured from cell extracts by either polyclonal antibody or epitope tag and the associated interaction partner identified by Western blot with a specific polyclonal antibody or second epitope tag. This category is also used if an interacting protein is visualized directly by dye stain or radioactivity. Note that this differs from any co-purification experiment involving affinity capture in that the co-purification experiment involves at least one extra purification step to get rid of potential contaminating proteins. (Ghetu et al., 2008). They are present with a green dotted line.
- **Co-localization:** Interaction inferred from two proteins that co-localize in the cell by indirect immunofluorescence only when in addition, if one gene is deleted, the other protein becomes miss-localized. Also includes co-dependent association of proteins with promoter DNA in chromatin immunoprecipitation experiments. (Hatzi et al., 2013). They are present with a red dashed line.
- **FRET:** An interaction is inferred when proximity of interaction partners is detected by fluorescence resonance energy transfer between pairs of fluorophore-labelled molecules, such as occurs between CFP (donor) and YFP (acceptor) fusion proteins. They are present with a green sinewave line.
- **Reconstituted complex:** An interaction is detected between two proteins in vitro. (Hatzi et al., 2013)/ They are present with a pink normal line.

Looking at a section of all the BioGRID data, it was possible to build a set of complexes. By observing what enzymes of interest where close to what others a possible enzyme complex has been designed.

This enzyme complex is formed in three sections, the purple section is involved in the biosynthesis of leukotrienes, the blue section is involved in the biosynthesis of the initial prostaglandins and resolving precursors. In orange the big final big complex connects since it has enzymes involved in the COX and in the LOX pathway. It is also interesting to see that it was not possible to find connections to the CYP pathway.

To present clearer how some of the enzymes interact with the pathway the reactions have been added and a full picture can be seen on Figure 3.4. This network allow for some hypothesis of how the mechanism might work:

NFKB activates PTGS2 causing more AA to convert to prostaglandins (PGH, PGF and PGE) and increasing metabolites downstream of TBXAS1. If PGH is being produced, then CEBP or FREAC2 might act to convert PGH to prostacyclin - and potentially metabolites downstream of TBXAS1. The production of LTA and LTC from AA is independent of regulation by these TFs.

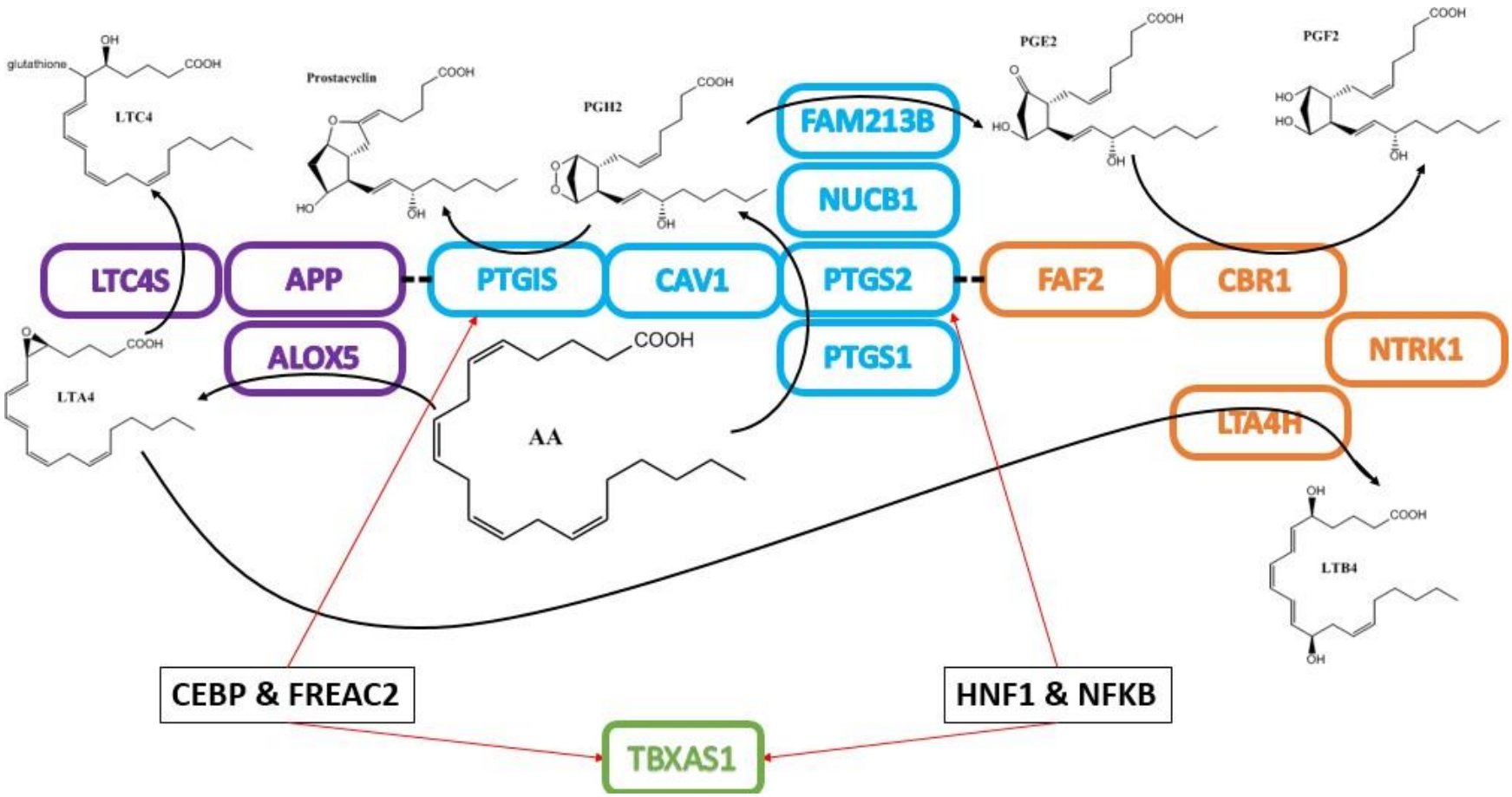


Figure 3.4: Combined data of enzyme complexes, transcription factors and pathway information on the oxylipin pathway. The TF interactions are presented with a redline, the black lines are for pathways. The rounded squares are for different enzymes, some of them involved in the oxylipin pathway some of them not, the colours do not serve other purpose than to highlight different pathway branches, they are a whole enzyme cluster.

3.5 Methodology

Here are a list of the instructions for a biologist/analyst user of the spreadsheet we have created.

3.5.1 Description of data importation

The data processing has been done using the Excel sheet from Appendix C. It is important that the list of oxylipins used has the t score attached to each lipid identity (the t score is described in chapter 2 section 2.4). Also, it is important to introduce the species with the same naming as in the Cytoscape layout (Check appendix A for naming). The data must be entered with the format name on top and t score below, the data will get automatically transposed. Then the values of p and DF (degrees of freedom) should be set by the user, for p it is recommended to use 0.05 and DF is the total number of oxylipins analyzed +1.

A function is used to calculate the HSV (Hue Saturation Value) of the colours. This function is formed of two lines that determinate the Saturation depending on the t score. The only value that the function modifies is Saturation, as it relates to the intensity of the colour.

The excel uses the functions T.INV() and some conditionals to calculate all the colouring that must be done in a column called Saturation. Now it is necessary to do a small copy and paste (paste as values) and sort the data. This will be helpful for the next steps. The three columns under copy must be copied and pasted in the columns next to it, then they should be sorted by Importance. After sorting the two first columns, they should be saved in to a txt file (Make sure to copy the two headers shared name and t-score). Do not close the excel sheet as it will be used.

A copy of the Cytoscape layout Eicosanoids-Plantilla-Complete.cys must be done in the working directory. It is recommended to rename the .cys file. Using Cytoscape execute the .cys file, then at the top select File > Import > Table > File… and select the txt were the sorted column data was saved, press ok and all the nodes on the network that have been identified will become red. Now go to the Style tab, look for the etiquette ‘Fill colour’ and click on the

small triangle, a screenshot of the location is presented on Figure 3.5 for better clarity. Now go entity by entity clicking the triple dot button.

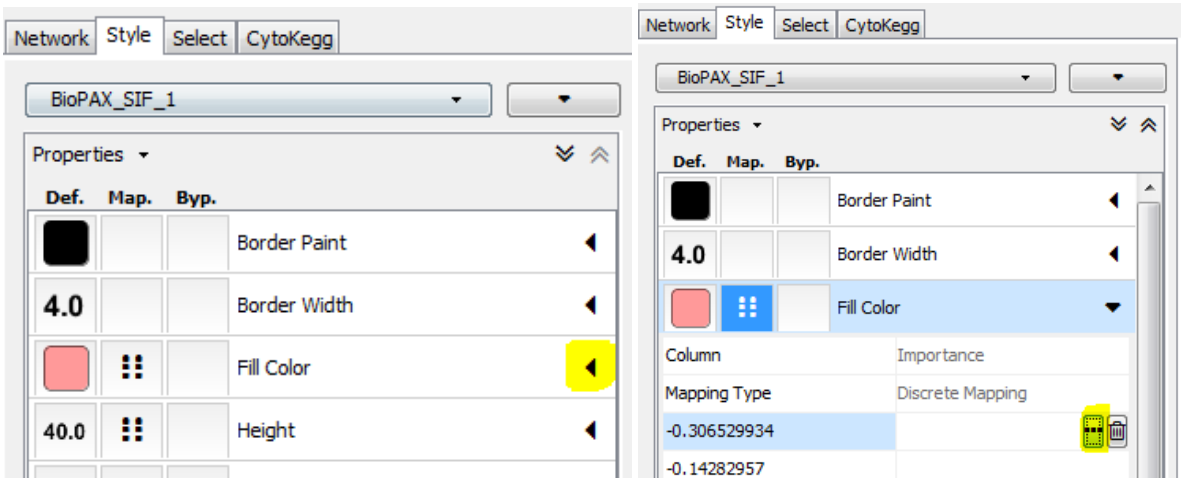


Figure 3.5: Screenshot of the Cytoscape Style tab. The image on the left present the default vision, the image on the right presents the vision once the small triangle has been clicked the. There are two small yellow highlights to show were the user should click.

In the new panel, navigate to the tab 'HSV' (Figure 3.6). For the negative values of the t score, the Hue must be set to 0, then the value of saturation introduced is the one that appears on the excel sheet. Here is where the sorting becomes helpful, as it is as easy as following the table. Upon reaching positive values of the t score, set the Hue to 140 and again, just change the saturation for each entry. If the value of saturation is bigger than 100, set it to 100. The parameters of Value and Transparency never have to be touched.

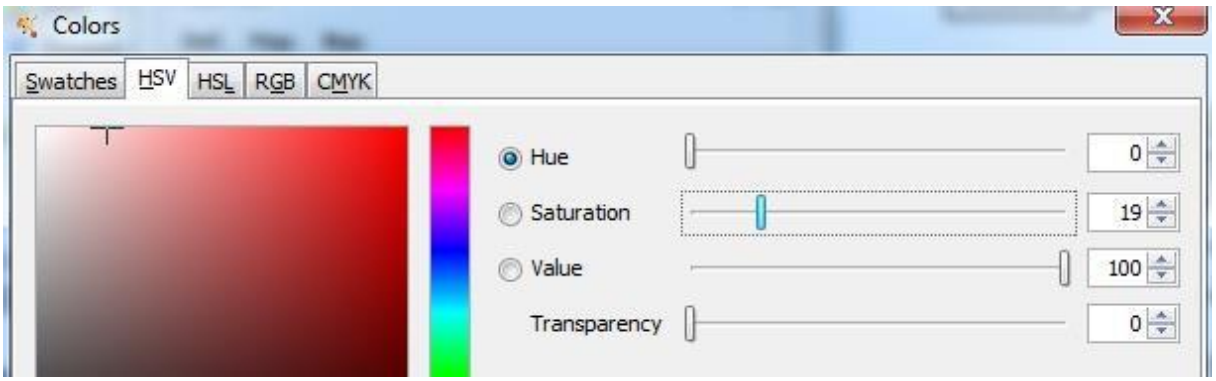


Figure 3.6: Screenshot of the Colours panel that appears in Cytoscape. The parameters presented here are for an entry that had -0.902 t score.in an experiment was p was set to 0.005 and had a DF of 30.

In the final heat map, the nodes will get colored in grey if they have not been identified and appear white if the species don't show any significant changes (t score is practically 0). Finally, the species that have high concentration on the control groups are coloured in green (t score +), the opposite effect goes in red (t score -). The more significant the score, the brighter they are on the map.

3.5.2 Setting the colouring function to aid data visualisation and interpretation

The coloring function is formed of two curves, a lower slope for all the values higher than the p>0.05 and a higher slope on brightness for those under the 0.05. At a p < 0.005 the colour intensity no longer changes, being 100 for all the rest. The brightness curve is used for the green and for the red nodes alike as it is calculated with the absolute value of the t score to calculate the Saturation.

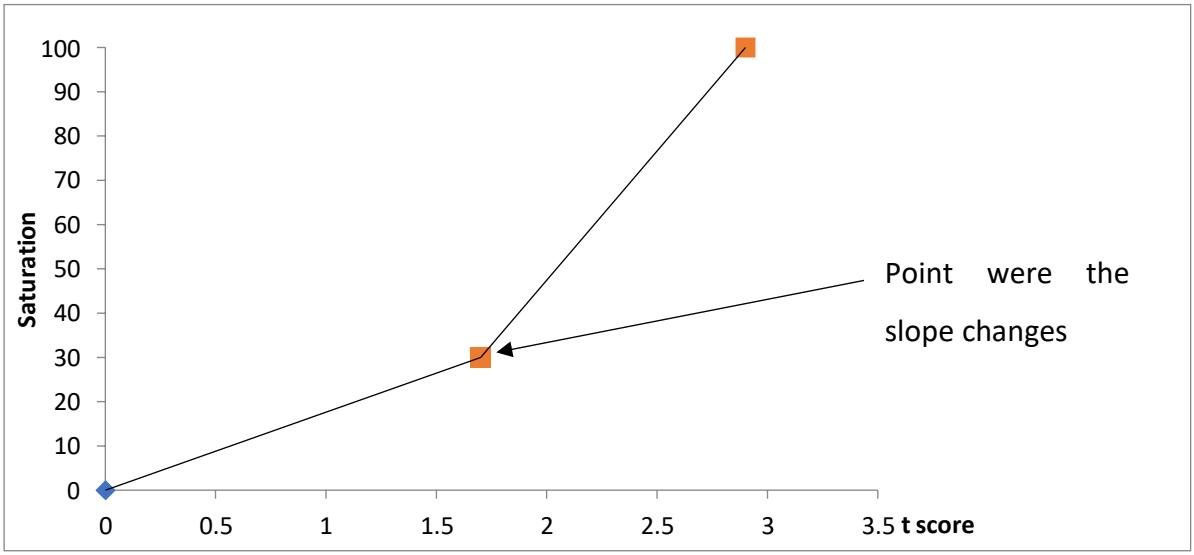


Figure 3.7: Plot of the colouring function. The colouring function present two curves of different slope that they change when the difference becomes significant.

The reason to use the slopes with the gradient for the colouring is because when compared to a differential colouring (colouring only if significant) the species that are almost significantly would have the same colour as the ones with no change at all. Additionally, the different slightly significant and the ones 10 times more significant would present the same colour.

3.6 Application of the method to existing oxylipin databases from clinical and pre-clinical inflammatory studies

In the following section, different experiments for OA targeted analysis will be presented and discussed. In this experiment the sample collection and some of the analytical work has been done by other PhD or post docs, when it is the case it will be specified. But as a general rule the work of the thesis is on the data analysis and the generation of the different pathways.

The analytical method used for all clinical examples is the HPLC system Shimadzu series 10AD VP LC system (Shimadzu, USA). The HPLC column used was ACE C18 (150 x 2.1 mm, 3 µm) with a guard column and Security Guard cartridges ACE 3 C18 (150 x 2.1 mm column). Mobile phase A was 0.02% formic acid in Milli-Q water, and mobile phase B was 0.02% formic acid in methanol/acetonitrile (1:4, v/v). Gradient started with 45% B, gradually increased to 100% by 12.5 min, then kept at 100% for 5.5 min and re-equilibrated for 2 min. Total run time was 20 min. The flow rate was 300 µL /min and the samples were kept at 4°C. The mass spectrometer system used was an Applied Biosystems MDS SCIEX 4000 Q-TRAP hybrid triple-quadrupole–linear ion trap mass spectrometer (Applied Biosystems/MDS Sciex, USA) equipped with an ESI interface. Multiple reaction monitoring (MRM) of all analytes detected in negative mode using specific precursor and product m/z ratios, and their retention times (RT) (Table 2-1). Quantification was performed using Analyst 1.4.1 software. The identification of individual oxylipins in biofluids was confirmed by comparing their LC retention times to that of each standard and their respective MRM transitions.

3.6.1 Osteoarthritis

Osteoarthritis (OA) is the most common form of joint disease, which is characterized by joint failure as its outcome. Latest studies suggested that OA affects up to 250 million people (Litwic et al. 2013). Osteoarthritis is a slow progressive disease and its main characteristics are cartilage degradation, synovial inflammation, joint space narrowing, osteophyte formation, difficulty with range of motion, and pain (Mathiessen et al. 2017). OA is a

multifactorial disease, and it has been shown that it is the result of systemic and local factors within the joint (Zhang and Jordan 2010). Although aging is the main cause of OA, other factors like obesity, a history of joint trauma, genetics, inflammation, heritable metabolic disorders, and previous rheumatoid arthritis also contribute to OA progression (Adatia et al. 2012; Goldring et al. 2011).

Current methods used to diagnose OA are radiographic (X-Ray), the "gold standard" (Hayashi et al. 2012), measuring the joint space width, and clinical symptoms, usually pain and difficulty in motion. However, radiography has a lack in sensitivity and precision and detects the disease only in its latest stages (Hunter et al. 2012). Additionally, MRI despite being a promising technique to monitor changes in cartilage is still not so commonly used as X-rays, mainly due to cost and limited availability (Hayashi et al. 2012). An alternative approach to OA detection is to measure biochemical markers in blood, urine or synovial fluid, which could reflect changes in joint structure and subsequently monitor disease progression (Lotz et al. 2013). Therefore, there is a pressing need for alternative biomarkers of OA that could detect the disease in its early stages where progression of disease is still responsive to pharmacotherapy.

As stated above, currently there is no effective drug treatment for OA (De Ceuninck et al. 2011). Patients are treated for symptoms with anti-inflammatory drugs and analgesics. Thus, discovery of biomarkers for OA will aid in early diagnosis of this slowly progressive disease, initiate personalized medicine, and will contribute in the identification of new drug targets. Although there has been an increasing number of publications about OA biomarkers, there is still not a biomarker that can be user to routine OA detection (Figure 3.8).

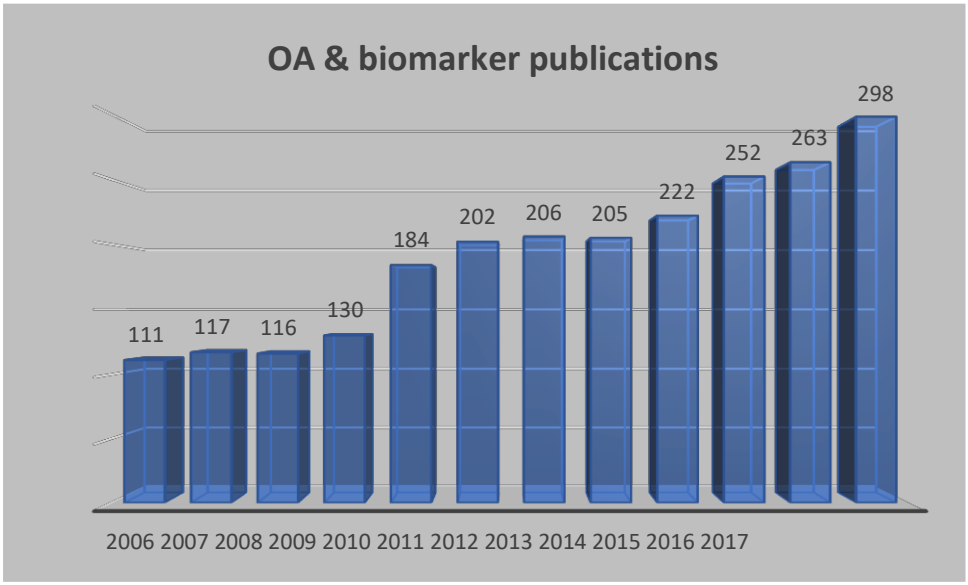


Figure 3.8: Number of papers in PubMed database found when the words "biomarkers" and "osteoarthritis" were searched. An increasing trend in publications is observed. (Kraus et al., 2011; Mobasheri, 2012; Rousseau et al., 2007)

Therefore, an alternative approach is needed for OA research. The new omics technologies (genomics, metabolomics and proteomics) have emerged as alternative methods for the discovery of biomarkers with success.

3.6.2 Application to oxylipin data from a surgical model of osteoarthritis

Goals of the original study:

To assess whether changes in oxylipin levels in different tissues from the MNX rat model of OA measured by a targeted LC-MS/MS method can be used to differentiate using univariate statistical methods. To build oxylipin pathways of MNX rat model in plasma, based on their associated enzymes for metabolism, with an aim to reveal oxylipin changes in MNX-treated rats

Study Design:

The MNX model of OA is an established rodent model of OA because the pathological changes induced by MNX surgery share many similarities with human OA (Mapp et al. 2010, Mapp et al. 2013, Ashraf et al. 2014, Nwosu et al. 2016)**,** reviewed here (Teeple et al. 2013)**.** The model was applied to rats and compared with controls.

The rat MNX study was performed in the laboratory of Professor Victoria Chapman (School of Life Sciences, University of Nottingham). Sample collection, sample preparation, and analysis by (LC-MS/MS) was done by Dr Petros Pousinis (University of Nottingham).

Samples preparation:

A total of 20 samples (10 controls and 10 MNX) with the MNX model of OA to rats was collected combining the three different tissues examined (spinal cord, DRGs and Plasma). Oxylipins were extracted from samples based on a modified method, previously developed by (Zhang et al., 2007). A set of deuterated internal standards (PGF2a-EA-d4, AA-d8, PGD2-d4, 15-HETE-d8) were added to each sample or blank sample (methanol), along with a mixture solvent comprising. Samples were homogenized in microcentrifuge tubes with the addition of the precipitation solvent, followed by a slow vortex stage and centrifugation. The supernatants were transferred to glass tubes and diluted by the addition of 3 ml water. The diluted supernatants were loaded to the Strata-X polymeric SPE column. Finally, oxylipins eluted from the column with ethyl acetate were dried in a centrifugal evaporator. The samples were reconstituted in 100 µL of

ethanol:water (50:50 v/v) transferred to dark brown auto sampler vials and stored at -80 °C until LC-MS/MS analysis.

Dorsal and ventral spinal cords, DRG tissues, and synovial fluid (obtained in a filter paper) from each sample were weighed. The sample preparation steps were the same as for plasma, with the exception that the slow vortex stage (10 min) was replaced by a vigorous shaking, using ice cold (-20 °C) pre-chilled metallic beads, in the Retsch MM301 Mixer Mill (Retsch, Haan, Germany) for 3 min at a frequency of 25 Hz.

Data analysis:

A statistical analysis was performed with unpaired Welch's t-test for the controls vs MNX model rats. The t scores were used to generate the pathway as described in subsection 3.6.2.

Pathway analysis:

Figure 3.8 reveals a shift in AA to a predominant metabolism via CYP450 (CYP) leading to increased levels of 14,15-DHET, 8,9-DHET and 5,6-DHET in the plasma of MNX rats. By contrast, there was a decrease in COX metabolism in MNX rats, since levels of COX related lipids (TBX2) in plasma were increased in control compared to MNX operated rats. These findings are probably connected to the role of DHET as a proinflammatory (Pillarisetti and Khanna 2012). The changes of other lipid species are quite low, suggesting that this model does not massively alter the oxylipin profile.

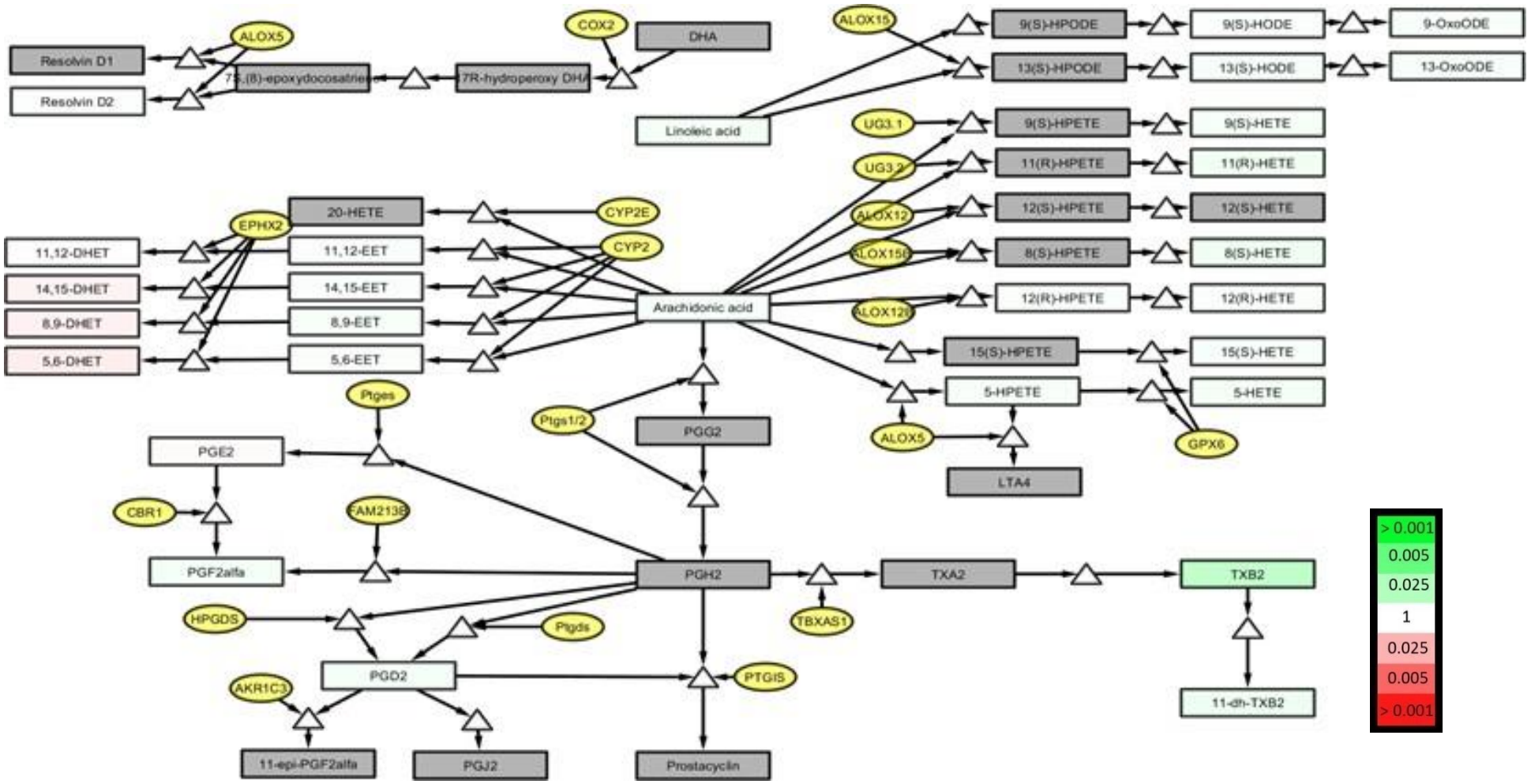


Figure 3.9 The COX, LOX, and CYP oxylipin pathways in MNX rat model of OA in plasma measured by a targeted LC-MS/MS method. Metabolites (oxylipins) are represented as squares, the genes as circles, and the reaction nodes as triangles. Highlighted: in grey, the metabolites not detected; in red, metabolites with high concentration in MNX; in green metabolites with high concentration in sham. The colour intensity indicates how strong the significance is (t values). The smaller the t value, the bigger the significance. COX, cyclooxygenase; LOX, lipoxygenase; CYP, P450 enzyme.

3.6.3 Application to oxylipin data from a chemical model of osteoarthritis

For the following experiments of this sections the data analysis and all the lab work was carried by the other PhD they will be is acknowledge in the corresponding sections.

Goals of the original study:

To assess whether changes in oxylipins levels in different tissues from the MIA rat model of OA measured by a targeted LC-MS/MS method can be used to differentiate using univariate statistical methods. To build oxylipins pathways of MIA rat model in plasma based on their associated enzymes for metabolism, with an aim to reveal oxylipins changes in MIA-treated rats

Study Design:

The monosodium iodoacetate model (MIA) of OA is an established rodent model of OA that pain researchers most often use (Fernihough et al. 2004, Mapp et al. 2013, Nwosu et al. 2016, Xu, Nwosu et al. 2016), reviewed here (Marker and Pomonis 2012). It is believed that the mechanism of MIA induction of OA is that MIA injection inhibits glycolysis and is toxic for chondrocytes, resulting in apoptosis and cell death (Smith et al. 2008).

The rat MIA study was performed in the laboratory of Professor Victoria Chapman (School of Life Sciences, University of Nottingham). Sample collection, sample preparation, and analysis by (LC-MS/MS) was done by Dr Petros Pousinis (University of Nottingham).

Samples collected:

A total of 22 (10 controls 12 MIA models) samples with the MIA model of OA to rats was collected combining the three different tissues examined (spinal cord, DRGs and Plasma). The sample preparation for this is the same that on the MNX model.

Data analysis:

Then a statistical analysis was performed with unpaired Welch's t-test for the controls vs MNX model rats. The t scores were used to generate the pathway as described in section 3.6.2.

Pathway analysis:

After building a pathway map of oxylipins using Cytoscape software, Figure 3.9 reveals a shift in AA to a predominant metabolism via the enzymes CYP450 (CYP) and 12/15-LOX in the plasma of MIA model. Like in the MNX model, the DHET species (proinflammation), are highly present on the MIA rats. This effect is greater on MNX than MIA model. Additionally, the HETE species also appear to be significant on MNX model, something that was not happening on MNX. HODE are higher in controls, as seen in the MIA.

It is safe to conclude that the MIA rodent model creates a stronger effect than the MNX, both present an increase on the DHET species and a slightly lowering on the HODE species.

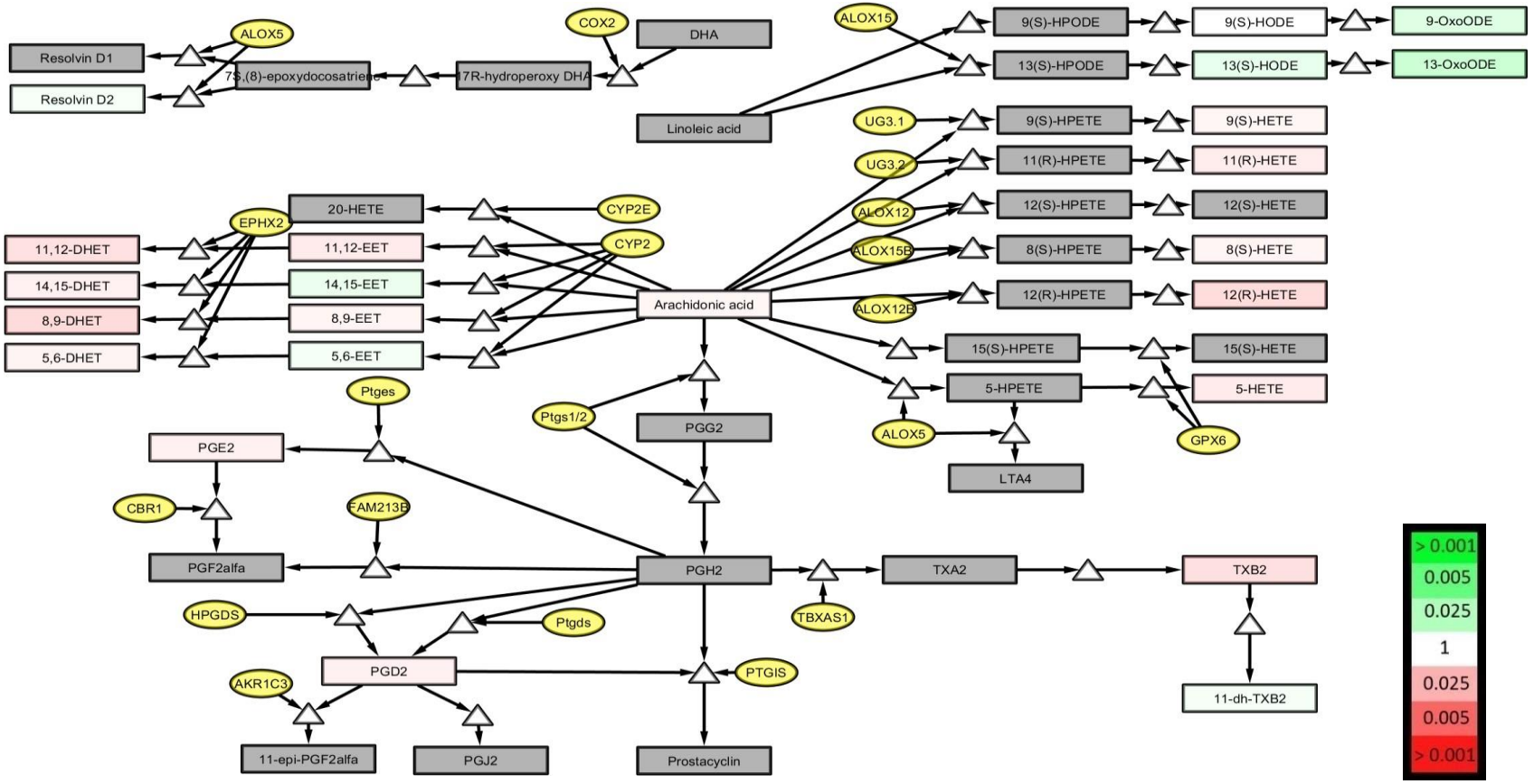


Figure 3.10 The COX, LOX, and CYP oxylipin network in MIA rat model of OA in plasma. Metabolites (oxylipins), measured by a quantitative targeted LC-MS/MS method, are represented as squares, the genes as circles, and the reaction nodes as triangles. Highlighted: in grey, the metabolites not detected; in red, metabolites with high concentration in MIA; in green metabolites with high concentration in saline. The colour intensity indicates how strong the significance is (t values). The smaller the t value, the bigger the significance. COX, cyclooxygenase; LOX, lipoxygenase; CYP, P450 enzyme.

3.6.4 Application to an oxylipin dataset from a study in a rat model to assess the effect of diet on OA progression

Goals of the original study:

To quantify oxylipin levels in plasma and synovial fluid from rats with experimentally induced metabolic osteoarthritis to investigate the potential role of oxylipins as a marker in the disease process of early osteoarthritis.

Study Design:

A target analytical protocol was designed to identify 31 lipid-derived inflammatory mediators' in fasted plasma and synovial fluid. The rats had undergone a local cartilage damage procedure and a high-fat diet to induce a metabolic dysregulation and simulate an early OA.

Samples collected:

A total of 40 rats were randomly divided in two groups. Twenty of were fed a high-fat diet (HFD; 60% of the kcal contained fat) while the others received a normal diet (9% of the kcal contained fat). After 12 weeks, cartilage damage was induced on the femoral condyles, in one knee joint according to the rat groove model (Visser *et al*., 2017) in 14 rats of each group. The remaining 6 were used as control groups for HFD and normal diet. Plasma and synovial fluid samples were obtained from all four groups. The sample preparation is the one described on the paper (He et al., 2015).

The rat study was performed in the laboratory of Professor Harrie Weinans (University of Utrecht, Netherlands). Sample collection, sample preparation, and analysis by (LC-MS/MS) was done by Dr Srini Ravipati (University of Nottingham).

Data analysis:

First a principal component analysis (PCA) was performed on all analytes from synovial tissue and plasma together. To evaluate the effect of HF diet and groove surgery on the normalised average data, for each cluster of lipids, a two-sided independent samples t-test was performed.

Pathway analysis:

A series of t-test were done for the group controls vs HFD and controls vs the groove surgery. Then the results were used to generate the two pathways presented in the Figure 3.10. The figure A is the comparison of the diets and the figure B is the comparison of the groove surgery.

It is observed that the effect of the diet is greater than the surgery, many of the oxylipins have higher concentration compared to controls. When looking at the surgery only 3 oxylipins appear to be significantly changing. Luckily for us 15-HETE and 8,9-DHET do not change from diet so if the groove surgery is a good approximation of OA they would be excellent biomarkers.

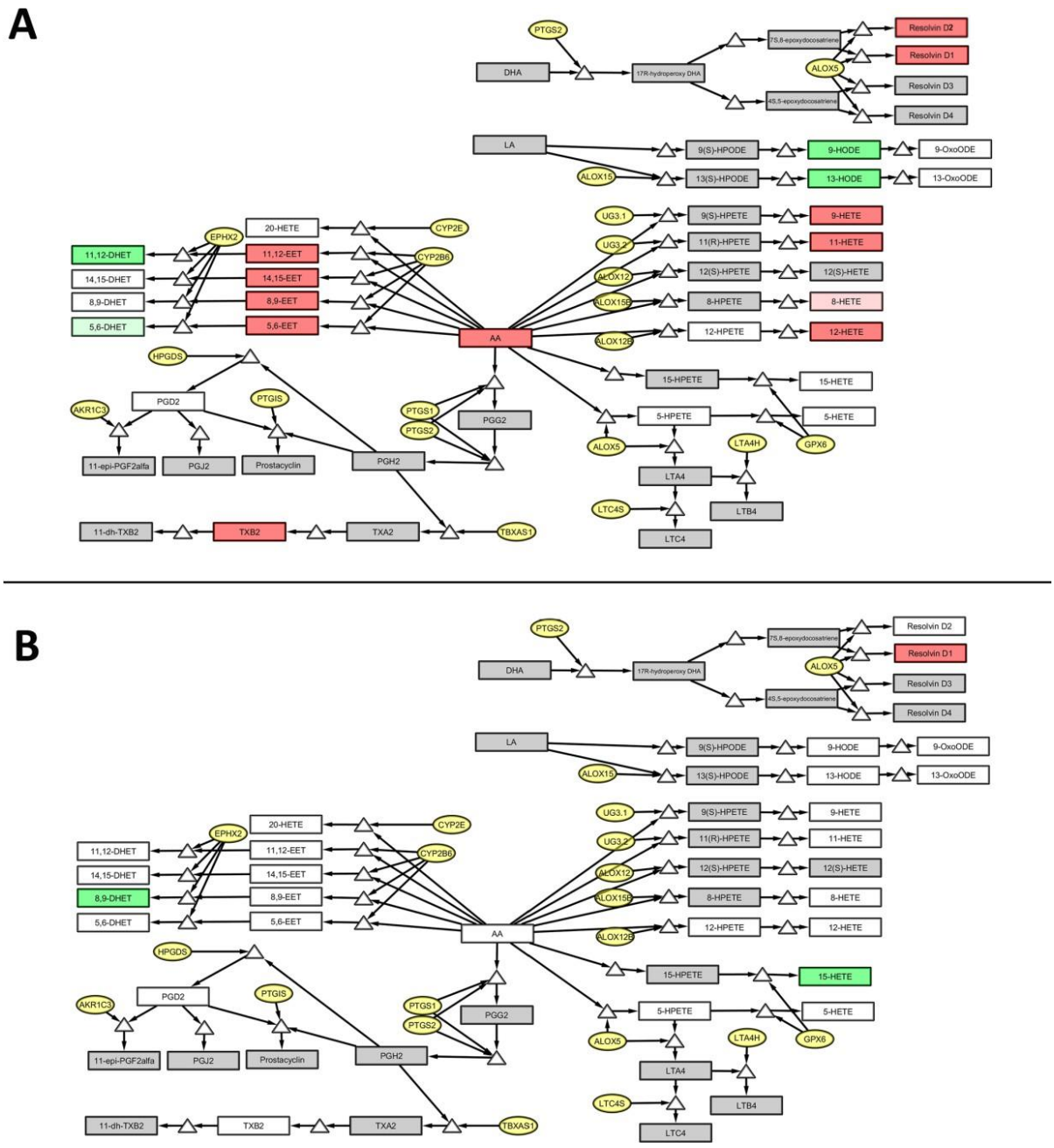


Figure 3.11: Network representing a selection of the most important studied oxylipins in the high fat diet experiment (A) and in the groove study (B), mapped onto the main arachidonic acid (AA), docosahexaenoic acid (DHA) and linoleic acid (LA) metabolic pathways, corresponding to the observed changes in t-test. The green boxes represent statistically significant higher values in control situation compared to the HF diet (A) or groove surgery (B), and red boxes for statistically significant higher oxylipins values in the HF diet (A) or groove surgery (B) compared to their control situation.

3.6.5 Application of the method to a clinical study of OA in a patient cohort (Nottingham study)

Unlike in the past studies were all the data analysis and data processing were already done and was comparing one condition over two datasets, this study contains 4 datasets with over two conditions. For this reason, all the data analysis treatment will be described.

Goals of the original study:

To identify changes in plasma, urine and synovial fluid lipid profiles associated with OA, which may have the potential to be a biomarker for the disease. To perform a follow up study 18 months later and collect plasma to identify potential biomarkers of progression.

Study Design:

The clinical study was undertaken by Professor M Doherty (School of Medicine, University of Nottingham) and was passed by the appropriate University and hospital Ethics Committees. Individuals with established OA and age-matched controls from the same area were recruited. Pain assessment and X-rays were performed in some participants.

Samples collected:

Synovial fluid, serum and urine were collected at baseline. A second clinical assessment was carried out 18-24 months later and serum was collected at a second-time point (n~ 110), 57 of them were controls (healthy individuals) and 53 suffered osteoarthritis. This enabled longitudinal analyses of the role of metabolomics biomarkers on radiographic and pain progression in OA.

Sample preparation and analysis by (LC-MS/MS) was done by Dr Srini Ravipati (University of Nottingham).

New data analysis:

For all the samples a similar targeted lipidomic analysis (of oxylipins) was done. The intention is to compare only the lipid species in common between the first and the follow-up studies.

The analysis looked at the global picture of this data, using PCA analysis of the 4 groups, OC (old controls), NC (new controls), OO (old Osteoarthritis), NO (new Osteoarthritis). Then this was followed by t-test comparison of age and the paired comparison of the groups.

Results I: PCA and general data overview.

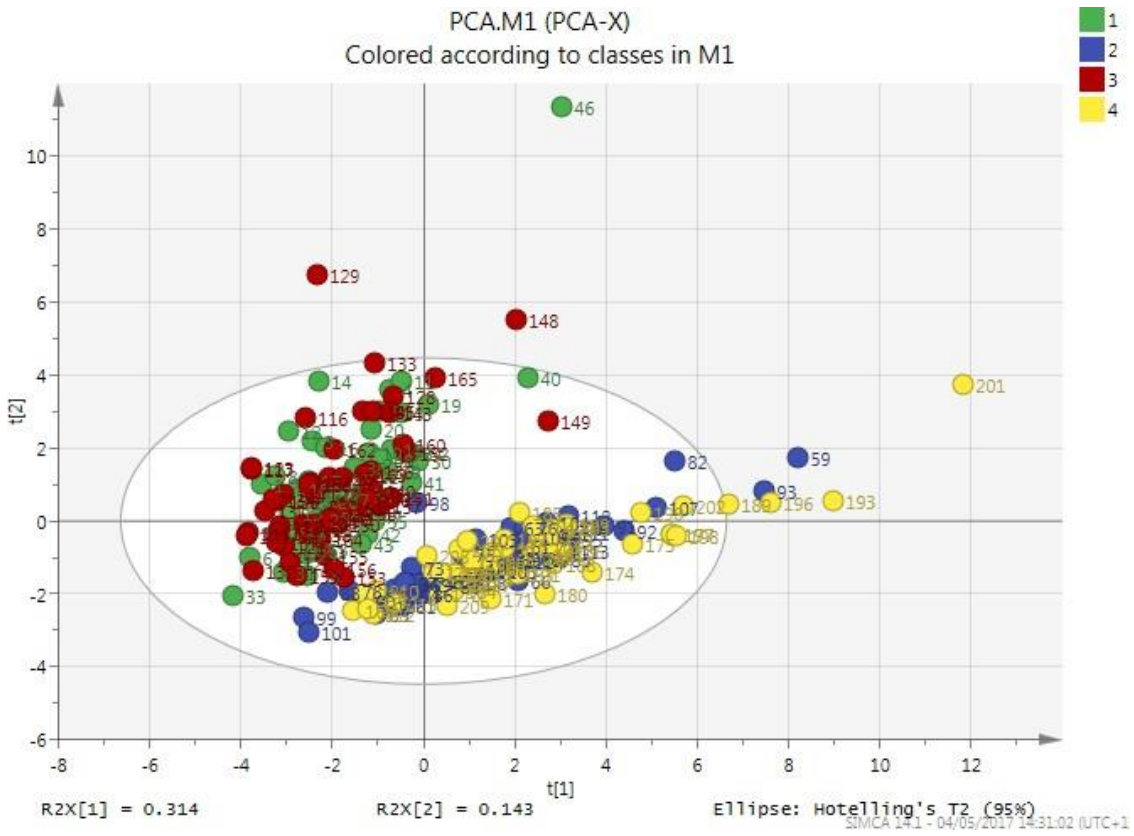


Figure 3.12: PCA plot for the follow study data. In green the OC, in red OO, in blue NC and yellow NO.

PCA shows that while the separation by age is extremely clear (red and green vs blue and yellow) the separation by controls and OA is non-existent. This is expected as ageing is known to have a great effect on lipid profiles (Yu et al. 2012). Another highly likely possibility due to the separation is that this is the effect of the storing of the samples.

This great effect on lipid concentrations due to ageing encourages the use of paired t-test as an analytical method. The option of merging the groups of controls and OA should be completely discarded since the effect of ageing would play a giant role on the standard deviation.

Comparing the profile by age

The first effect studied will be ageing, not of interest itself, but it can help on the final interpretation of the results.

A t-test comparison of the groups only separating by age has been done and a list of the lipids that are significantly different after Bonferroni correction has been done (see table 3.2).

Table 3.2: Means targeted lipid concentrations. The different columns are the mean (M) of Old and New samples and the p value of the t test. The mean is the area under the curve.

Lipid	**M Old**	**M New**	**p**
5-HETE	310.2	553.1	<0.005
8-HETE	99.4	173.3	<0.005
11-HETE	110.9	440.0	<0.005
12-HETE	2465.2	2603.2	22.33
15-HETE	216.1	185.6	5.306
16-HETE	123.5	115.0	1.840
19-HETE	703.0	651.8	7.647
20-HETE	814.3	506.3	<0.005
8,9-EET	110.1	177.1	<0.005
11,12-EET	118.8	447.5	<0.005
14,15-EET	600.8	1449.5	<0.005
5,6-DHET	120.2	109.4	<0.005
8,9-DHET	115.9	145.0	0.007
11,12-DHET	426.8	349.9	13.37
14,15-DHET	45.6	368.2	<0.005
9-HODE	7.7	3.8	<0.005
13-HODE	5.1	7.7	<0.005
9-OxoODE	5.4	10.7	<0.005
17-HDoHE	16.8	19.3	0.025
TXB2	2074.2	640.1	<0.005
AA	14.4	0.9	<0.005
LTB4	1774.6	103.1	<0.005

A total of 17 oxylipins appear to be significantly changing. Some of them having drastic changes as shown by their p values. These metabolites, even if they have a change in OA vs Controls, should be taken carefully because they may not be good candidates for biomarkers. An optimal biomarker should not be substantially affected by age. Another observation is that most of the changes are showing an increase in their concentration.

Paired comparison

The data can be paired according to patient ID, as the same patients have been tested (2 years apart). The paired t-test is performed with FDR correction, a histogram is generated of a (Figure 3.13) to check if the data presented a normal distribution

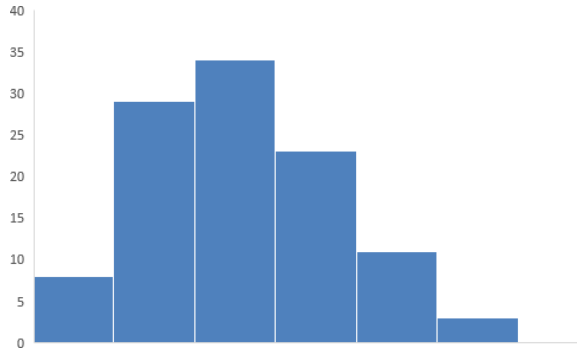
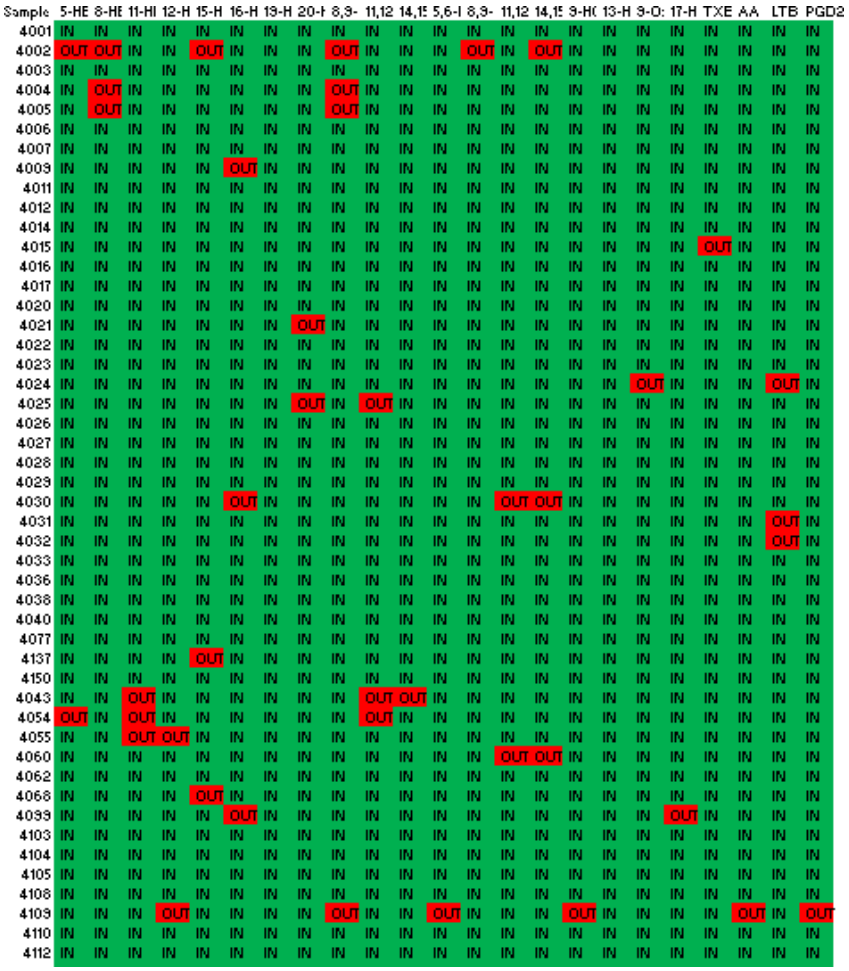


Figure 3.13: Histogram of 8-HETE. This histogram presents a somewhat normal distribution, suggesting that the data is good to work using univariate analysis and no more processing is needed.

Another important thing to be aware if the existence of outliers. If a sample or a lipid specie contains a high number of outliers, it is good practice to remove them. The image 3.14 shows the outliers on each sample for each lipid. The only sample with more than 5 outliers is 4109 and 4002, still it will be kept for this analysis, since it represents less than 40%.



Sample	5-HE	8-HE	11-HI	12-H	15-H	16-H	13-H	20-H	8,9-	11,12	14,15	5,6-I	8,9-	11,12	14,15	9-HO	13-H	9-O:	17-H	TXE	AA	LTB	PGD2
4001	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4002	OUT	OUT	IN	IN	OUT	IN	IN	IN	OUT	IN	IN	IN	OUT	IN	OUT	IN	IN	IN	IN	IN	IN	IN	IN
4003	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4004	IN	OUT	IN	IN	IN	IN	IN	IN	OUT	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4005	IN	OUT	IN	IN	IN	IN	IN	IN	OUT	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4006	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4007	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4009	IN	IN	IN	IN	IN	OUT	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4011	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4012	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4014	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4015	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	OUT	IN	IN	IN	IN
4016	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4017	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4020	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4021	IN	IN	IN	IN	IN	IN	IN	OUT	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4022	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4023	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4024	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	OUT	IN	IN	IN	OUT	IN
4025	IN	IN	IN	IN	IN	IN	IN	OUT	IN	OUT	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4026	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4027	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4028	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4029	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4030	IN	IN	IN	IN	IN	OUT	IN	IN	IN	IN	IN	IN	IN	OUT	OUT	IN	IN	IN	IN	IN	IN	IN	IN
4031	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	OUT	IN
4032	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	OUT	IN
4033	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4036	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4038	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4040	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4077	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4137	IN	IN	IN	IN	OUT	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4150	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4043	IN	IN	OUT	IN	IN	IN	IN	IN	IN	OUT	OUT	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4054	OUT	IN	OUT	IN	IN	IN	IN	IN	IN	OUT	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4055	IN	IN	OUT	OUT	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4060	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	OUT	OUT	IN	IN	IN	IN	IN	IN	IN	IN
4062	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4068	IN	IN	IN	IN	OUT	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4039	IN	IN	IN	IN	IN	OUT	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	OUT	IN	IN	IN	IN	IN
4103	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4104	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4105	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4108	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4109	IN	IN	IN	OUT	IN	IN	IN	OUT	IN	IN	OUT	IN	IN	IN	OUT	IN	IN	IN	IN	OUT	IN	OUT	
4110	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
4112	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN

Figure 3.14: The figure presents all the data points on the analysis, they have been introduced to a box whisker graph the ones on red are the possible outliers.

After this filtering, 2 oxylipins appear to be significant: 11,12-EET and 11-HETE. Both present an increase on their concentration, but both are also affected by age, making them not suitable biomarkers.

In the paired comparison, it is possible to see that over 70% of the oxylipins are increasing their concentration, the only family that is showing to have a lower concentration the HODE derivatives.

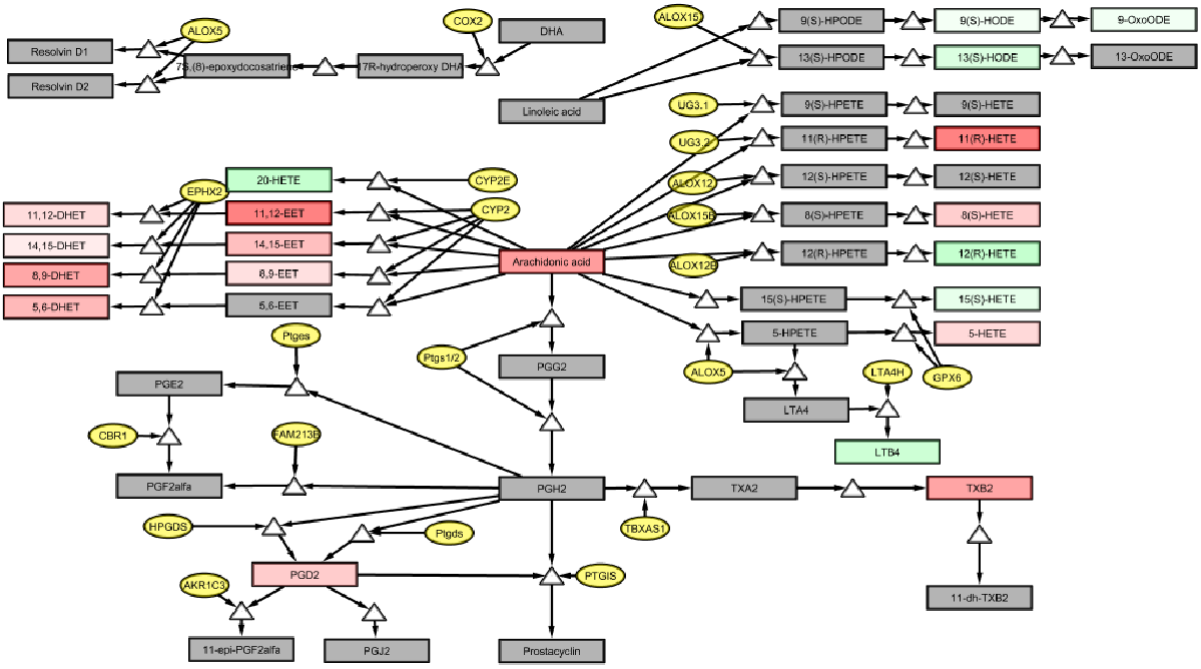


Figure 3.15: Network of Arachidonic acid derivatives. The lipid species of the paired comparison coloured in red or green if they are respectively in higher or lower concentration in OA compared to Controls with grey denoting no change. Enzyme nodes are in yellow.

Conclusions

Two oxylipin species 11-HETE and 11,12-EET have been found to be significantly different in the paired comparison of OA vs controls. However, these species are also affected by age, making them a non-reliable biomarker.

The lower concentration of the HODE species was already reported in the rat MNX and MIA models, which is an interesting effect since HODE is usually linked to inflamed tissues. Another common effect shared is the high concentration of DHET.

3.6.6 Application of the method to a clinical study of OA in a patient cohort (PROOF study)

Rationale / Goals of the original study:

For this PROOF study, 407 women between 50 and 60 years with a BMI ≥ 27 kg/m2 at baseline were recruited. At baseline, all women were free of clinical knee osteoarthritis (OA), but at high risk for future development of knee OA due to their sex, age and body weight status. Data collection was integrated in the 6.5-year follow-up measurements of the PROOF study.

Study Design:

At the 6.5-year follow-up measurements, a fasted serum sample was collected from a random selection of 74 participants. Collection, processing and storage was done according to the standardized protocols provided by the University of Nottingham. After finalizing all data collection, all samples were distributed among the WP-partners and further analyzed at their facilities. Samples were analyzed for general lipid profiles and targeted oxylipins method.

Samples collected: 57 controls and 17 with knee pain. The study was performed at Erasmus MC, Rotterdam. Sample preparation and analysis by (LC-MS/MS) was done by Dr Srini Ravipati (University of Nottingham).

New data analysis:

The data was analyzed with multivariate analysis (PCA and OPLS-DA) and univariate analysis (t test). While all data was used for the multivariate, only the oxylipin species where studied on the univariate. KL-groups for the PROOF study were analyzed. For subjects free of radiologic knee OA (that is KL < 2), the KL_LT variable is zero.

Results

Data overview

The data contains 74 samples, from those, 57 are control and 17 present the knee pain. From these samples, a total of 295 lipids are quantified. These lipids are from many families (Oxylipin, FFA, LPC, PC, DG, TG, SM, Cer).

Univariate analysis

All the oxylipins were analyzed using a t test, three of them came to be significantly different after the FDR corrections. Results are displayed in the following image 3.16.

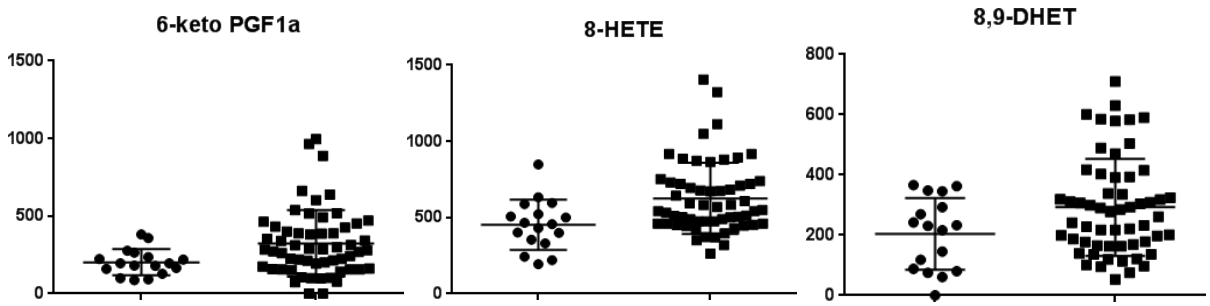


Figure 3.16: T-test plots for the three significant lipid species, analysis performed with prism.

An outlier study was done for these three lipid species: 8-HETE contains 45% of the data as outliers, 6-keto PGF1a contains 34% as outliers, while 8,9-DHET contains only 12%. For this reason, it has been decided to only accept 8,9-DHET as a true significantly changing lipid. The fold change of this specie from the DHET family is -0.3.

We repeated the analysis using a log transformation of the data to make it more normally distributed. However, the results are still the same, the same three lipids remain significant, 6-keto PGF1a and 8-HETE with more than 25% outliers.

Multivariate analysis

For this analysis, all 295 lipids were used. The first step was doing a PCA. Even though it didn't give any good separation, it seems like there are some outliers like sample 280, 301 and 82. Those outliers were identified because they appeared far away from all the data points. These are removed for the incoming OPLSDA analysis.

An important condition when using multivariate methods like OPLDSA is that the data should be normally distributed. The best way to check if the data is normally distributed is to plot a histogram of a random sample Figure 3.16. It was possible to observe that the data wasn't normally distributed, and to correct it, it was decided that the log transformation should be applied.

Still even with the normally distributed data, the dataset does not present the optimal conditions for an OPLS-DA as it contains a great disproportion of the two group sizes. However, the analysis was still done, the results of the first two components are shown below in Figure 3.17. Since it was not possible to see separation with the first OPLSDA, the model was over fitted, and the separation is achieved (Q2 = -0.23).

This is not the optimal solution Q2 values should be bigger than 1.5 we can conclude that from the multivariate analysis it was not possible to obtain a lipid that shows the difference of the two groups.

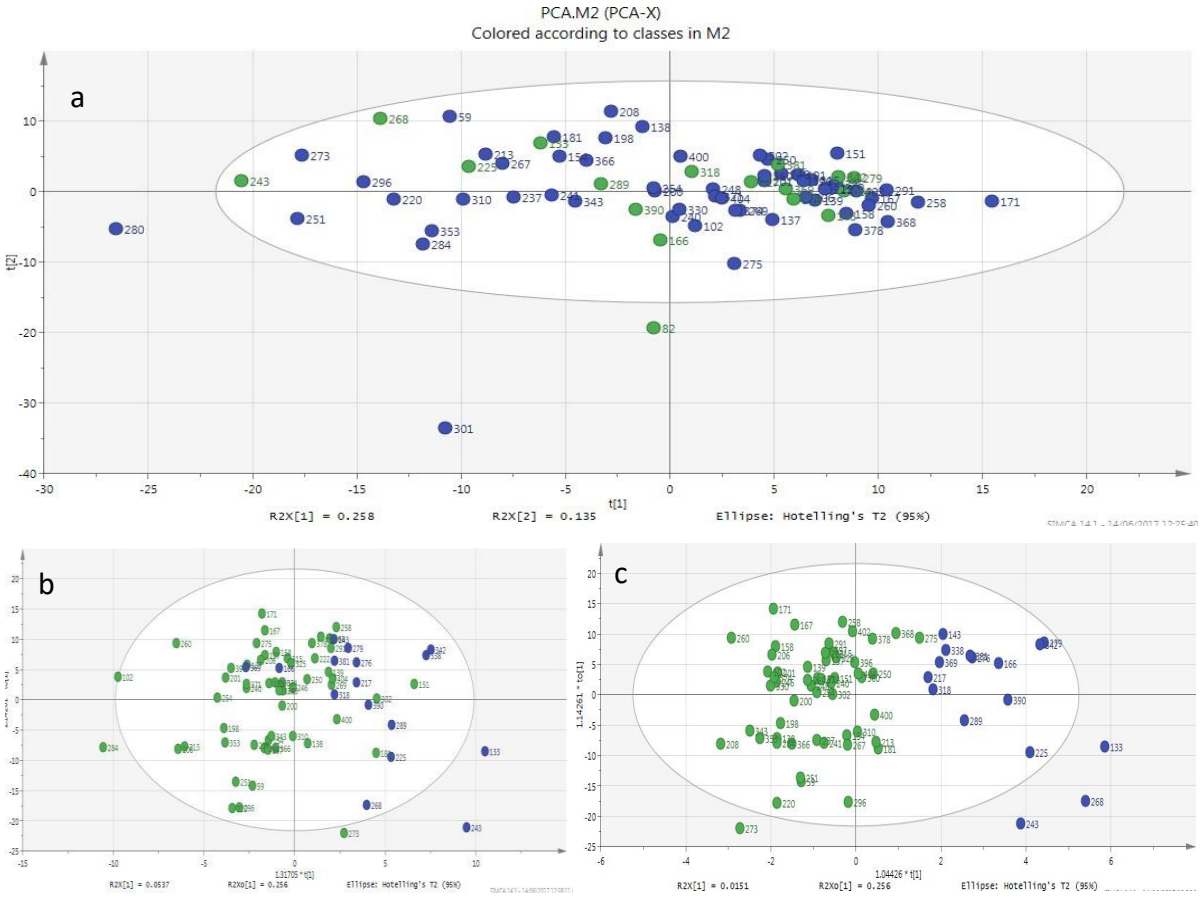


Figure 3.17: The figure presents at the top a PCA plot of all the data together(a), then two OPLS-DA where the outliers have been removed, (b) directly the OPLDS and (c) the overfitted model.

Pathway analysis:

The last option was to map the results in the oxylipin pathway. The Figure 3.18 present the results, most of the lipid derivatives from AA are found higher, the only specie that does not work in that direction is the 20-HETE, that shows a decrease in its concentration. This reaction is carried by CYP2E1, also carries out the metabolism of endogenous fatty acids such as the ω-1 hydroxylation of fatty acids, involving it in important signaling pathways that may link it to diabetes and obesity (Porubsky et el, 2008).

It is important to annotate that the HoDE derivates are found again in a non-significant lower concentration, this is a consistent trend over the other experiments.

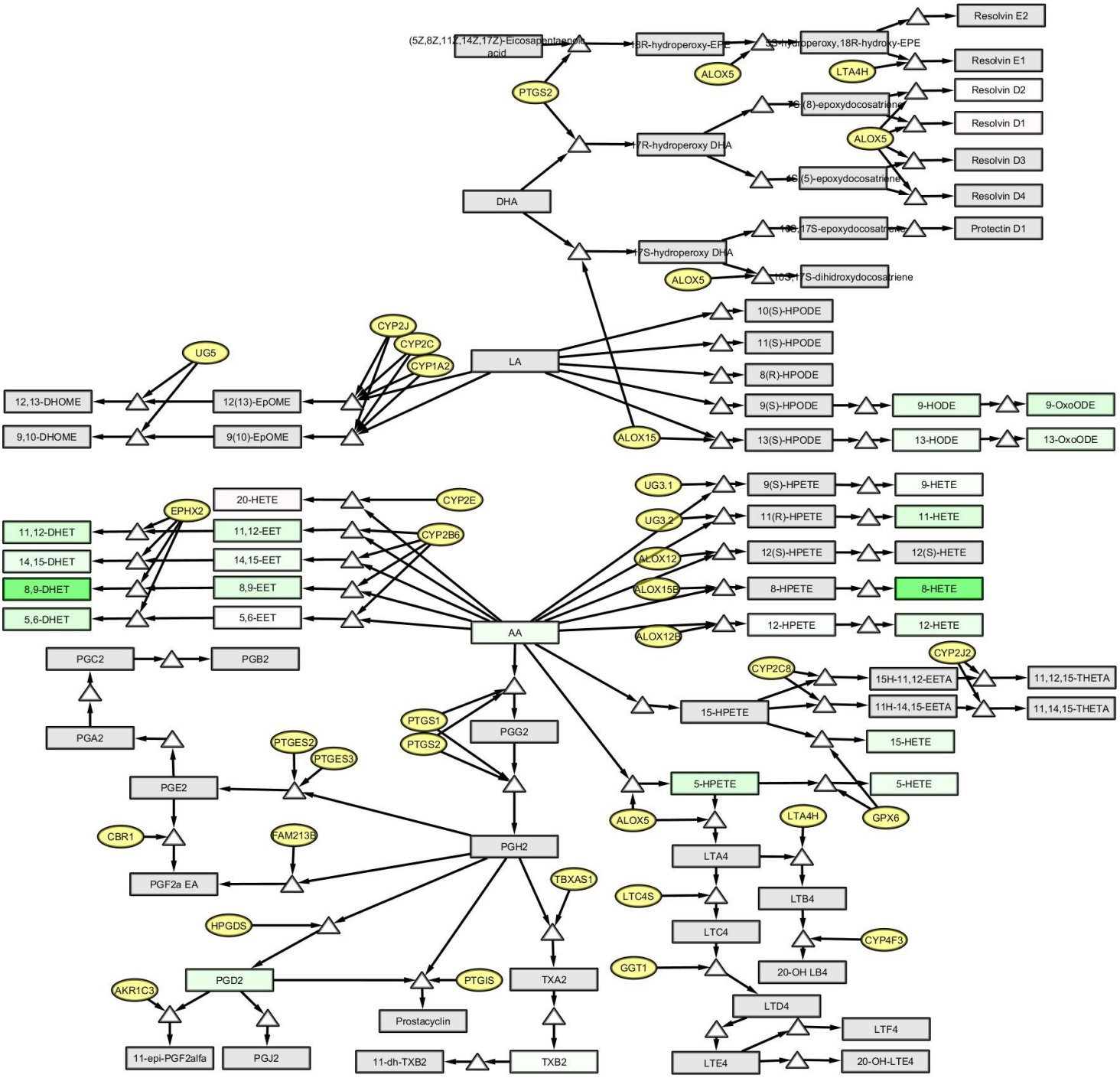


Figure 3.18: Heat map for PROOF study. The lipid species of the paired comparison coloured in red or green if they are respectively in higher or lower concentration in OA compared to Controls with grey denoting no change. Enzyme nodes are in yellow.

3.7 Conclusions

Reviewing the literature of the different oxylipin pathways and published databases has allowed us to create a complete pathway containing many oxylipins. This literature review has also allowed us to generate an inflammation score calculator. While being a worthwhile exercise, it does not provide enough information on its own owing to the diverse effects of oxylipins - a single value for pro- or anti-inflammation is just not that useful to clinicians (Victoria Chapman, Dept of Medicine, University of Nottingham, personal communication).

The integrated networks from GeneMANIA and BioGRID have provided some excellent results. For example, the oxylipin enzymes appear to form co-localised multiprotein complexes, reducing the time needed to generate the signaling molecules from their precursors. In this study, NFKB, a transcription factor, seems to be regulating the generation of prostaglandins and thromboxanes. The enzymes producing PGD2 and especially PGE2 are directly linked to the target receptors of these molecules, causing in the latter case an amplification of the pro-inflammatory response

A colouring methodology for the oxylipin network has been developed, which has allowed us to visualize the changes in the different OA experiments, and compare them easily in different conditions. This heat map method shows both the significant changes and smaller changes that otherwise might not be reported. The results were consistent with some of the hypotheses about the regulation of oxylipins.

The HF diet experiment is especially interesting, Figure 3.11 B, because this is a scenario where AA is being forced into the oxylipin network. It clearly shows that DHET levels drop substantially while their precursors the EETs are elevated. The latter are formed by CYP2B6, for which no transcriptional regulator was identified, and the elevated levels could reflect higher levels of AA and/or a reduced activity of EPHX2 which could be due to regulation by the unknown TF that binds to the motif AACTTT. TXs and PGs are derived from PGH2 and their levels generally change in the same direction, though again the HF diet experiment is an exception. In that case, the TX level is noticeably increased, possibly reflecting activation by one of the other TFs that regulate TBXAS1 expression.

4. Software tools for improving lipidomic data analysis

4.1 Introduction

In mass spectrometry-based lipidomic analysis, there are many steps; from sample extraction to lab work, statistical analysis and interpretation. As presented in the introductory chapter, all the sections after the lab work have the potential to be improved using bioinformatics. This chapter presents the different steps, and how it has been possible to improve them. In some cases, the software developed has been a help for the some of these steps (shortening the time data took to process), as well as being used for a better understanding of lipidomic analysis itself.

These programs were generated to alleviate the struggle that some of the tasks presented when done manually. The programs have been produced after rigorous discussions and in collaboration with the users, to make sure they fit the needs and work with the options they require.

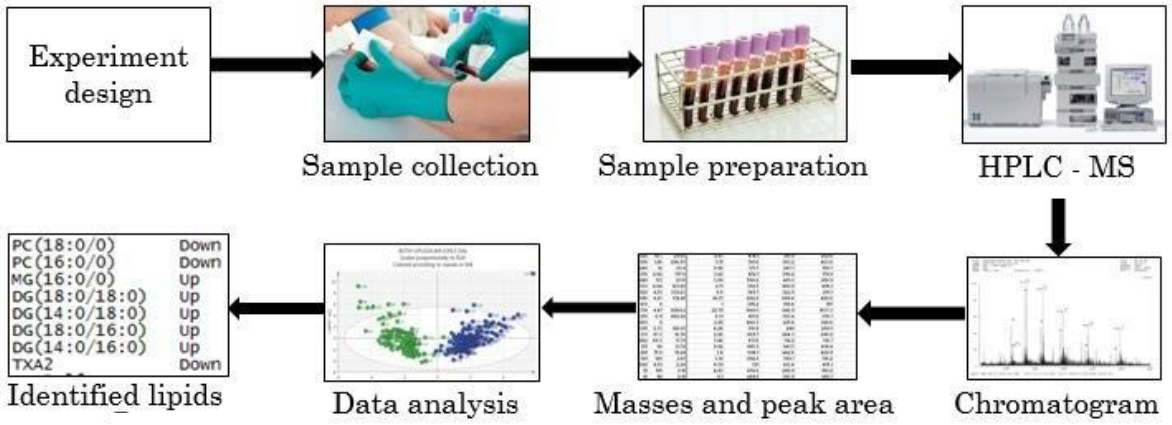
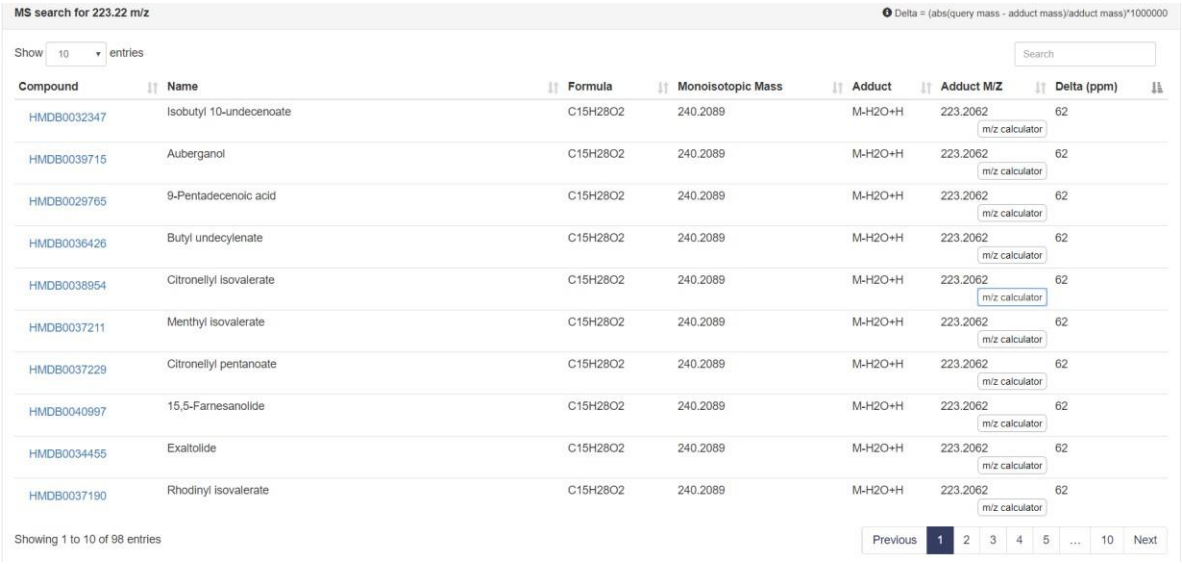


Figure 4.1: Lipidomic analysis pipeline. The scheme presents the different steps generally undertaken on a lipidomic experiment.

4.2 Lipid identification

One of the hardest steps in the lipidomic analysis is the assignment of molecular identities to the observed masses. To perform this task there are many available tools online, including HMDB (Wishart et al, 2018), Metlin (Smith et al, 2005) and LIPIDMAPS (Sud et al, 2006). The problem is that these online identification tools are slow, because they must be used manually, taking around two minutes to perform an identification of one metabolite. Some of them do not allow bulk identification or to have a broader band of possible adducts.

The manual identification in HMDB starts by selecting the adducts and the mode of detection. It does allow to introduce multiple masses at once. Then it would display the output like in the image below 4.2:



MS search for 223.22 m/z

Delta = (abs(query mass - adduct mass)/adduct mass)*1000000

Show 10 entries

Search

Compound	Name	Formula	Monoisotopic Mass	Adduct	Adduct M/Z	Delta (ppm)
HMDB0032347	Isobutyl 10-undecenoate	C15H28O2	240.2089	M-H2O+H	223.2062 m/z calculator	62
HMDB0039715	Auberganol	C15H28O2	240.2089	M-H2O+H	223.2062 m/z calculator	62
HMDB0029765	9-Pentadecenoic acid	C15H28O2	240.2089	M-H2O+H	223.2062 m/z calculator	62
HMDB0036426	Butyl undecylenate	C15H28O2	240.2089	M-H2O+H	223.2062 m/z calculator	62
HMDB0038954	Citronellyl isovalerate	C15H28O2	240.2089	M-H2O+H	223.2062 m/z calculator	62
HMDB0037211	Menthyl isovalerate	C15H28O2	240.2089	M-H2O+H	223.2062 m/z calculator	62
HMDB0037229	Citronellyl pentanoate	C15H28O2	240.2089	M-H2O+H	223.2062 m/z calculator	62
HMDB0040997	15,5-Farnesanolide	C15H28O2	240.2089	M-H2O+H	223.2062 m/z calculator	62
HMDB0034455	Exaltolide	C15H28O2	240.2089	M-H2O+H	223.2062 m/z calculator	62
HMDB0037190	Rhodinyl isovalerate	C15H28O2	240.2089	M-H2O+H	223.2062 m/z calculator	62

Showing 1 to 10 of 98 entries

Previous 1 2 3 4 5 ... 10 Next

Figure 4.2: Output for HMDB on the identification on positive mode of an m/z of 223.22 with the common adducts.

The first problem faced is that this database contains lipids and metabolites, therefore, most of the identifications are not correct. Another difficulty is the repetition of multiple isomers, this makes visualization tremendously hard.

When using LIPIDMAPS to identify lipids, the first problem is gone as only lipids are found. However, it does not allow the introduction of multiple masses nor multiple adducts.

For these reasons, it was decided that a program should be created (Mass-LipidMAPS-Checking.py) that uses the LIPIDMAPS dataset to perform the identification of lipids. The whole script can be found in Appendix 4.B.d. It runs only on Ubuntu, because some of the libraries required for the program to run are not yet available in Windows operating systems.

The first step requires the user to enter a list of the adducts of interest and ppm range. The program then inputs a list of the mode, masses and retention times, searches in LIPIDMAPS for each mass using all the possible adducts. Finally, it prints out a plain text file with all the possible masses that match known molecules along with their different adducts (see Figure 4.3).

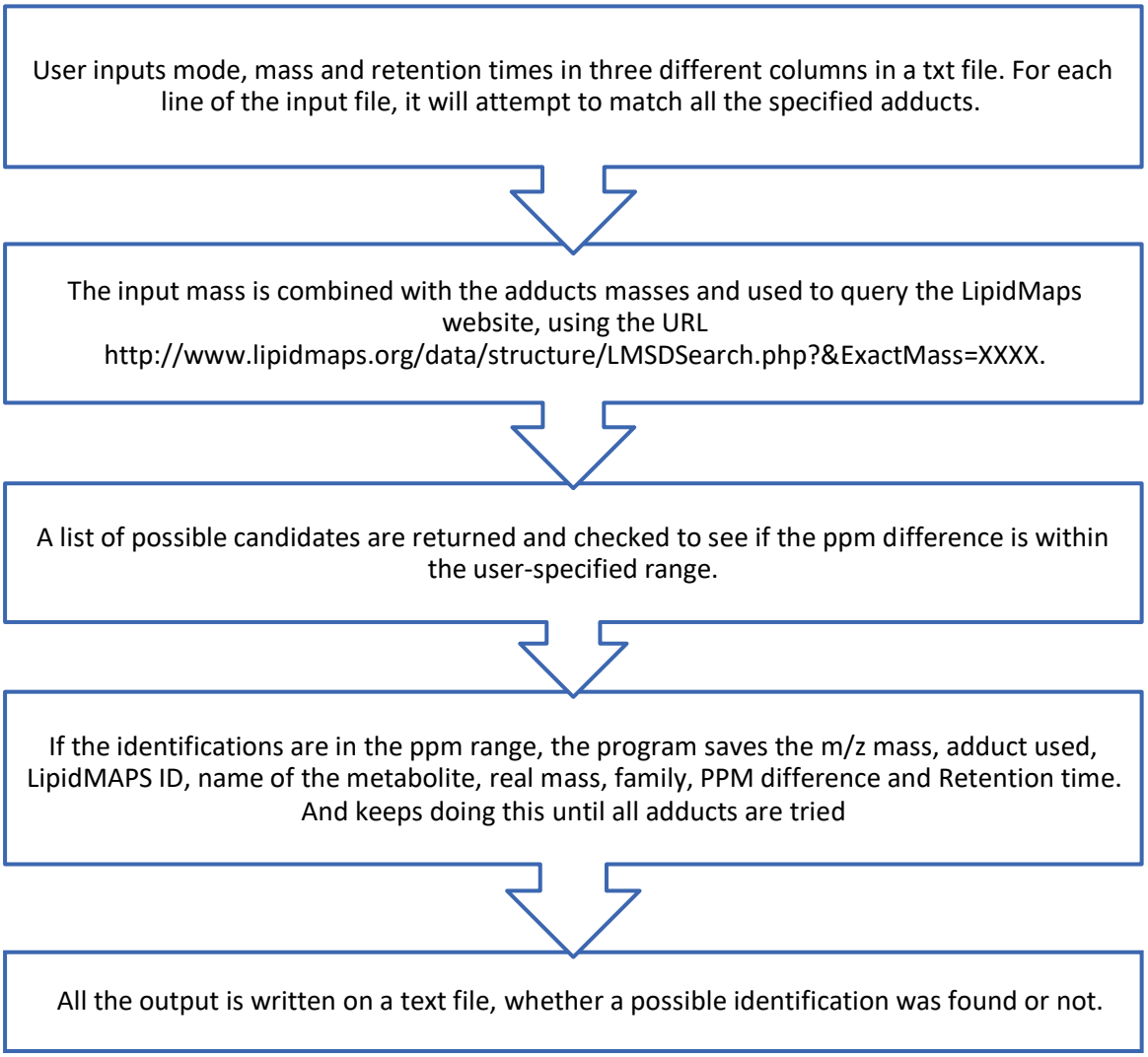
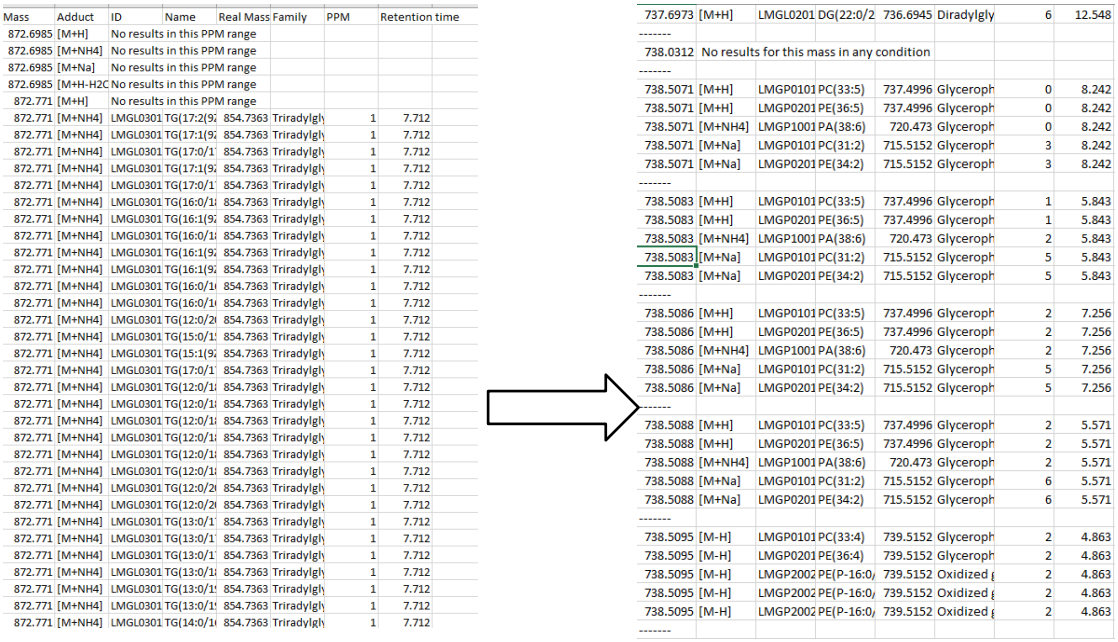


Figure 4.3: Flow chart of the python identification program. The program will get the mass and apply the different adduct possibilities, then it will look in the website for matches. When one is found the data is saved.

A second program, (*Mass-LipidMAPS-Cleaner.py*) (see Appendix 4.B.e), filters and gets rid of non-useful information (see Figure 4.4). By non-useful information it is understood the multiple times that a mass appears with multiple adducts when none of them found. The program provides all the possible identification in the PPM range. The user should still work out which is the most probable identification when there are two or more. However, the data is much more compact and collapsed; the different subfamilies are grouped together, clearer separators are added and when a mass is impossible to find with any adduct, only one line appears explaining so.



Mass	Adduct	ID	Name	Real Mass	Family	PPM	Retention time
872.6985	[M+H]	No results in this PPM range					
872.6985	[M+NH4]	No results in this PPM range					
872.6985	[M+Na]	No results in this PPM range					
872.6985	[M+H-H2C	No results in this PPM range					
872.771	[M+H]	No results in this PPM range					
872.771	[M+NH4]	LMGL0301	TG(17:2(9(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(17:1(9(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(17:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(17:1(9(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(17:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(16:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(16:1(9(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(16:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(16:1(9(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(16:1(9(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(16:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(16:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(12:0/2(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(15:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(15:1(9(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(17:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(12:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(12:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(12:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(12:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(12:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(12:0/2(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(12:0/2(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(13:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(13:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(13:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(13:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(13:0/1(	854.7363	Triradylgly	1	7.712
872.771	[M+NH4]	LMGL0301	TG(14:0/1(	854.7363	Triradylgly	1	7.712

737.6973	[M+H]	LMGL0201	DG(22:0/2	736.6945	Diradylgly	6	12.548

738.0312	No results for this mass in any condition						

738.5071	[M+H]	LMGP0101	PC(33:5)	737.4996	Glycerophl	0	8.242
738.5071	[M+H]	LMGP0201	PE(36:5)	737.4996	Glycerophl	0	8.242
738.5071	[M+NH4]	LMGP1001	PA(38:6)	720.473	Glycerophl	0	8.242
738.5071	[M+Na]	LMGP0101	PC(31:2)	715.5152	Glycerophl	3	8.242
738.5071	[M+Na]	LMGP0201	PE(34:2)	715.5152	Glycerophl	3	8.242

738.5083	[M+H]	LMGP0101	PC(33:5)	737.4996	Glycerophl	1	5.843
738.5083	[M+H]	LMGP0201	PE(36:5)	737.4996	Glycerophl	1	5.843
738.5083	[M+NH4]	LMGP1001	PA(38:6)	720.473	Glycerophl	2	5.843
738.5083	[M+Na]	LMGP0101	PC(31:2)	715.5152	Glycerophl	5	5.843
738.5083	[M+Na]	LMGP0201	PE(34:2)	715.5152	Glycerophl	5	5.843

738.5086	[M+H]	LMGP0101	PC(33:5)	737.4996	Glycerophl	2	7.256
738.5086	[M+H]	LMGP0201	PE(36:5)	737.4996	Glycerophl	2	7.256
738.5086	[M+NH4]	LMGP1001	PA(38:6)	720.473	Glycerophl	2	7.256
738.5086	[M+Na]	LMGP0101	PC(31:2)	715.5152	Glycerophl	5	7.256
738.5086	[M+Na]	LMGP0201	PE(34:2)	715.5152	Glycerophl	5	7.256

738.5088	[M+H]	LMGP0101	PC(33:5)	737.4996	Glycerophl	2	5.571
738.5088	[M+H]	LMGP0201	PE(36:5)	737.4996	Glycerophl	2	5.571
738.5088	[M+NH4]	LMGP1001	PA(38:6)	720.473	Glycerophl	2	5.571
738.5088	[M+Na]	LMGP0101	PC(31:2)	715.5152	Glycerophl	6	5.571
738.5088	[M+Na]	LMGP0201	PE(34:2)	715.5152	Glycerophl	6	5.571

738.5095	[M-H]	LMGP0101	PC(33:4)	739.5152	Glycerophl	2	4.863
738.5095	[M-H]	LMGP0201	PE(36:4)	739.5152	Glycerophl	2	4.863
738.5095	[M-H]	LMGP2002	PE(P-16:0/	739.5152	Oxidized g	2	4.863
738.5095	[M-H]	LMGP2002	PE(P-16:0/	739.5152	Oxidized g	2	4.863
738.5095	[M-H]	LMGP2002	PE(P-16:0/	739.5152	Oxidized g	2	4.863

Figure 4.4: In the left we have the output of Mass-LipidMAPS-Checking.py, in the right we have the transformation of the same dataset when the Mass-LipidMAPS-Cleaner.py is used. Is possible to appreciate that the different metabolites get clearly separated by the , also that when no identification is found with any adduct the data gets compacted. Another useful tool is that the generalised form of the lipids gets compressed to the general notation, while in the left there might be several TG forms in the right only the general will be written TG(68:4) as an example.

This application has been used for many datasets in global lipidomics, of which:

- Brain vs muscle tissue: A set of lipidomics analyses were done on brain tissue and muscle tissue from mice. These experiments were done by the PhD student Amelia Pollard. The identification process for those studies was done using the identification software.

- Mice vs Bat plasma: These experiments were focussed on observing the difference in these two species. While mouse models are often used, they have a short life span. Bats on the other hand have a lifespan closer to humans. In these experiments, the lipid species were again identified by this program.
- Seven OA datasets: analyses for a lot of these experiments have been done. Human plasma samples as well as spinal cords and plasma samples from mice in varying conditions are a few examples.

In all the experiments, the identification program tried to get all the possible candidates for each mass in a quick manner. The program has proven very helpful in performing the identification of a large number of lipid species in much shorter time than usual (around 2 hours, 5 seconds for each metabolite to identify). It is recommended for when there are more than 20 species to identify.

4.3 Adducts on ESI

Electrospray is an especially soft ionization method, capable of generating molecular ions without fragmentation from biological molecules in aqueous solutions (Mikkelsen et al. 2016). A hollow stainless-steel needle outputs continuously the sample into the chamber, an electric field in the needle disperses the liquid by coulombic forces into a fine spray of droplets. These droplets migrate, and their solvent removed by the drying flowing gas (typically N_2) and moves to the typical MS. This method of detection has an effect that complicates identification, namely the formation of adducts.

The term “adduct ion” is a popular term among liquid chromatography/mass spectrometry (LC/MS) users to describe ions formed by adduction of alkali metal ions to an analyte molecule in positive ion analysis. However, the well-informed user should be aware that “adduct ion” correctly refers to any ion formed by adduction of an ionic species to a molecule. Therefore, the common protonated molecule or $[M+H]^+$ is also properly called an adduct ion (https://blog.sepscience.com/separationscience/author/separation-science-ms-solutions).

A list of the most common adducts was found in (Huang et al. 1999). and can be found in Table 4.3. They modify the mass/charge ratio (m/z) of a molecule making identification

harder. Several software packages already implement ways to identify and take them into account. 2M and 3M means that the peak corresponds to a mass of the molecule that is two- or three-fold higher, maybe the adduct couples multiple molecules together. There have also been mases of M+N+H, where two different molecules combine, making the identification of M and N almost impossible.

The adducts come from the sample, or the mobile phase. It is imperative to not look for adducts that are not possible. If the sample is a biological, it is possible to have Na or K but an adducts such as M+DMSO+H should not be recorded unless the mobile phase has included the solvent DMSO. It is not normal for small molecules to become multi-charged adducts, M+3H or M-2H. The multimeric adducts, 2M or 3M should appear at the same retention time because they are formed on from the ionization of their M. In Table 4.1, most adducts are infrequent, and those in bold font are detected by commercial software.

Table 4.1: Monoisotopic exact masses of molecular ion adducts often observed in ESI mass spectra. The species in bond are more commonly present

Positive mode	**Positive mode**	**Negative mode**
M+3H	**M+H**	M-3H
M+2H+Na	**M+NH4**	M-2H
M+H+2Na	**M+Na**	**M-H2O-H**
M+3Na	**M+CH3OH+H**	**M-H**
M+2H	M+K	M+Na-2H
M+H+NH4	M+ACN+H	M+Cl
M+H+Na	M+2Na-H	M+K-2H
M+H+K	M+IsoProp+H	M+FA-H
M+ACN+2H	M+ACN+Na	M+Hac-H
M+2Na	M+2K+H	M+Br
M+2ACN+2H	M+DMSO+H	M+TFA-H
M+3ACN+2H	M+2ACN+H	2M-H
M+IsoProp+Na+H	2M+Na	2M+FA-H
2M+H	2M+3H2O+2H	2M+Hac-H
2M+NH4	2M+K	3M-H
2M+ACN+H	2M+ACN+Na	

4.4 Adduct characterization

One of the first steps in the analysis of mass spectrometry data is adduct summation. However, this task, even when performed in perfect conditions, there are still adducts present (Appendix 4.B.a). The different metabolites in a sample can be present along with many adducts; for this reason, many studies have been gathering information on which adducts can be found for different species and conditions (Keller et al. 2008; Reis et al. 2003; Sleno et al. 2007). During this step of the lipidomic pipeline, all adducts from the same species are combined. This process is normally done by software like Progenesis, which allows the user to select the types of adducts, then combining them to only get peaks with the proton adduct (M+H or M-H). This process is extremely common and can be quite critical.

Adducts have been used recently in a practical manner in experiments. Feeding a sample with additives and enhancing the generation of certain adducts has proven useful for the detection of low concentration species. It has also been used for low stability compounds, such as sugars or explosives (species that were not possible to analyze by traditional ESI/MS) (Ghosh et al. 2015; Breitbach et al. 2016). In the environment of global lipidomics, adducts are undesirable. For that reason, automatic software removal is preferable. However, after using the software, peaks corresponding to other adducts, especially sodium and ammonia, are still present.

In this section, the focus is on the different adduct concentrations and proportions that appear in lipidomic experiments. Using the identification software, the standard format of initial lipidomic identification is shown in Figure 4.5. Although human-readable, it does not lend itself to large-scale analysis of adduct types. Hence, the *Graph* family of python programs have been developed to parse useful information out of this file format.

```
-------104.99329[illegible]4      [M+H]   No results in this PPM range
104.9932934     [M+NH4] No results in this PPM range
104.9932934     [M+Na]  Nothing Found
104.9932934     [M+H-H2O]       No results in this PPM range
-------107.086454      [M+H]   No results in this PPM range
107.086454      [M+NH4] Nothing Found
107.086454      [M+Na]  No results in this PPM range
107.086454      [M+H-H2O]       LMFA06000034    2,4-octadienal  124.0888        Fatty aldehydes [FA06]  2       3.134
107.086454      [M+H-H2O]       LMFA06000035    5,7-octadienal  124.0888        Fatty aldehydes [FA06]  2       3.134
107.086454      [M+H-H2O]       LMFA06000121    2,4-Dimethyl-2E,4E-hexadienal   124.0888        Fatty aldehydes [FA06]  2       3.134
107.086454      [M+H-H2O]       LMFA06000128    2E,7-Octadienal 124.0888        Fatty aldehydes [FA06]  2       3.134
107.086454      [M+H-H2O]       LMFA06000131    2E,6E-Octadienal        124.0888        Fatty aldehydes [FA06]  2       3.134
107.086454      [M+H-H2O]       LMFA06000132    2E,4Z-Octadienal        124.0888        Fatty aldehydes [FA06]  2       3.134
107.086454      [M+H-H2O]       LMFA06000133    2E,6Z-Octadienal        124.0888        Fatty aldehydes [FA06]  2       3.134
107.086454      [M+H-H2O]       LMFA12000017    3E,5E-Octadien-2-one    124.0888        Oxygenated hydrocarbons [FA12]  2       3.134
107.086454      [M+H-H2O]       LMFA12000038    6-Methyl-3E,5-heptadien-2-one   124.0888        Oxygenated hydrocarbons [FA12]  2       3.134
```

Figure 4.5: Example of the format presented by the output of the identification software. This file with this formatting is the one that is used as input for the Graph-Family programs. In the image, two peaks are presented; the first with an m/z of 104.99 has no results found for any of the possible adducts. The second, with a mass of 107.08, has no results for the M+NH4 and M+Na adducts but has 9 possible identifications for the M+H-H2O. The general information is separated by tabs, and the information that is presented is; mass in the chromatogram, adduct, lipid maps ID, general name, mass of the metabolite without the adduct, family, ppm difference, and RT.

4.4.1 Adducts separated by retention time

The program, Graph-RT.py, summarizes the useful information from the identification documents (See Figure 4.7). In its simplest version, Graphs-RT.py only saves the information of retention time and the proportion of adducts for each m/z. The source code for this program can be found in Appendix 4.B.f. It starts by saving a dictionary of the adducts studied. Then it goes through the data line by line. If any adduct information is read, it is saved in a dictionary. Upon reaching the " " separator, the dictionary is then converted into the % format with the dicMOD function, then printed and emptied, ready for the next mass. The dicMOD function converts the frequency of adducts into the proportion (this is the % format). In excel, the output format takes the form shown in Figure 4.6.

Mass	RT	[M-H2O-H]	[M-H]	[M+Na]	[M+Ac-H]	[M+H]	[M+H-H2O]	[M+NH4]
115.0763	0.06	0	0	0	0	0.375	0.625	0
121.1021	0.06	0	0	0	0	1	0	0
193.1227	0.06	0	0	0	0	0.5	0.5	0
203.1798	0.06	0	0	0	0	0.888888889	0.111111111	0
211.1333	0.06	0	0	0	0	0.5	0.5	0
221.154	0.06	0	0	0	0	0.666666667	0.333333333	0
227.2373	0.06	0	0	0	0	0	1	0
239.2373	0.06	0	0	0	0	0.527272727	0.472727273	0
255.2323	0.06	0	0	0	0	0.322580645	0.677419355	0

Figure 4.6: Excel import of the output generated by the Graphs-RT.py. The columns correspond to the mass of the peak, retention time and then, all the common adducts. The adduct values indicate the fraction of the total

adducts in that peak. These values can be later multiplied by the peak area to find out how much adduct was present.

This analysis focused on observing the effect of retention time on the adducts found (Figure 4.7). Proportions of [M+H-H2O] adducts decrease over time, going from ~ 30% to 15%. For a given retention time, this provides a guide as to which adducts may be present in each peak. Likewise, the [M-H] adducts are low at the beginning (< 5% of them) and slowly increasing. While this observation does not guarantee that [M-H] adducts should never be picked over [M+Ac-H], it would make sense to do for retention times shorter than 6-8 minutes. Although useful, this analysis is tremendously dependent on the methodology. A change in any experimental parameter (such as solvent, column, flux or temperature) may change the proportion of adducts, meaning that it is necessary to redo this analysis.

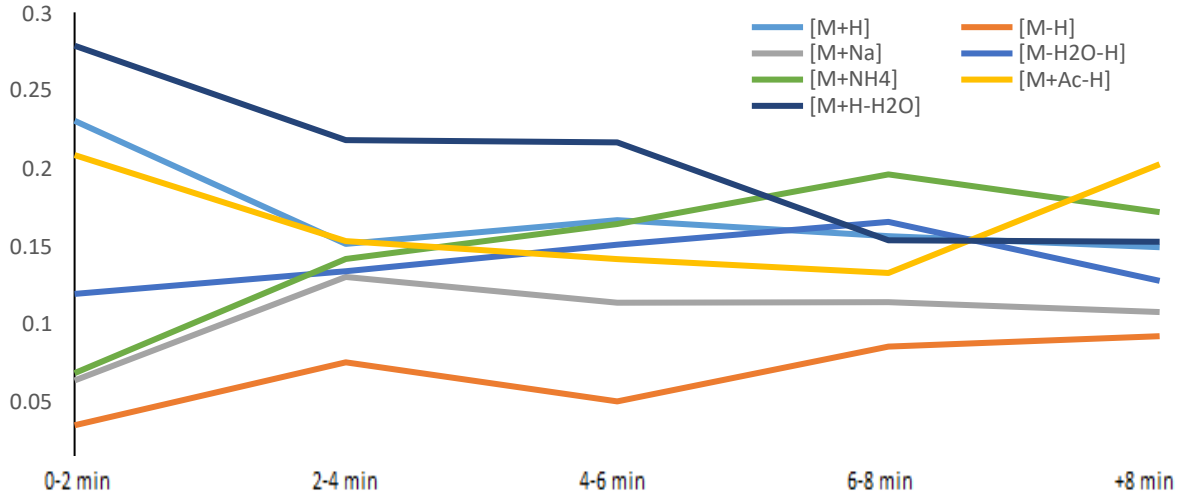


Figure 4.7: The graph presents the relative fraction of the different adducts over time. There are 7 different types of adducts represented by each colour. In the lower axis, the different time brackets used to pack the information are presented.

4.4.2 Adduct proportions within lipid families

LIPIDMAPS defines the following families: Fatty Acyls [FA], Glycerolipids [GL], Glycerophospholipids [GP], Sphingolipids [SP], Sterol Lipids [ST]; Prenol Lipids [PR], Saccharolipids [SL] and Polyketides [PK].

Graphs-Family-2.py generates a combined list, as shown in Figure 4.8. The source code for this program can be found in Appendix 4.B.g. It also inputs the intensities for every m/z, which are calculated by taking the average of the peak areas in the control samples. The program generates a dictionary of the adducts and another for the lipid families. Some exceptions have been made as some molecular species do not belong to a family, so they have been put into a family labelled NI (Not identified). Prior to output, the adduct dictionary is again processed by the dicMOD function. The data still requires further processing, specifically, multiplying the adduct fractions by their intensity and assigning a category, which is the most frequent lipid family. In the case of a tie, the family that has a higher concentration in human lipid will be selected. All this has been done using excel functions and manually.

Mass	Intensity	[M+H-H2C	[M+H]	[M-H+Na/	[M-H2O-H	[M-H]	[M+Ac-H]	[M+Na]	[M-H+3Na	[M+NH4]	[M-H+2Na	PR	NI	GL	FA	SL	SP	GP	PK	ST
259.2054	26.30058	1	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0
294.2426	1261.009	0	0	0	0	0	0	0	0	1	0	0	0	0	50	0	0	0	0	0
294.2428	480.0365	0	0	0	0	0	0	0	0	1	0	0	0	0	50	0	0	0	0	0
186.1488	152.9736	0	0	0	0	0	0	0	0	1	0	0	0	0	45	0	0	0	0	0
249.182	272.8024	0	0	0	0	0	0	0.931818	0	0.068182	0	0	0	0	44	0	0	0	0	0
338.2664	3890.878	0	0	0	0	0	0	0.06383	0	0.93617	0	0	0	0	44	0	3	0	0	0
179.1429	406.2716	1	0	0	0	0	0	0	0	0	0	0	0	0	42	0	0	0	0	0
214.1801	42535.67	0	0	0	0	0	0	0	0	1	0	0	0	0	42	0	0	0	0	0
270.2426	1674.111	0.047619	0	0	0	0	0	0	0	0.952381	0	0	0	0	41	0	1	0	0	0
336.2517	460.5243	0	0	0	0	0	0	0.025	0	0.975	0	0	0	0	40	0	0	0	0	0
370.2561	636.991	0	0	0	0	0	0	0	0	1	0	0	0	0	40	0	0	0	0	0
337.2403	442.7474	1	0	0	0	0	0	0	0	0	0	0	0	0	38	0	0	0	0	0
546.2764	678.1158	0	0	0	0	0	0	0	0.736842	0	0.263158	0	0	0	38	0	0	0	0	0
546.2766	260.674	0	0	0	0	0	0	0	0.736842	0	0.263158	0	0	0	38	0	0	0	0	0

Figure 4.8: Example output of the Graphs-Family-2.py program. The columns respectively represent the mass of the peak, its average peak area for all controls, the fraction of each the common adducts and a count of all the different family species that may be present for that mass. For example, for the mass of 259.2054, there are potentially 50 different FAs and they only form the [M+H-H2O] adduct.

To get a more general view of the human lipid-family and adduct distribution, the program has been used on three different datasets: the studies on Idiopathic pulmonary fibrosis (IPF), Lung cancer and PCOS. Additionally, we have added [M-H+XNaAc] adducts, where X goes from 1 to 3. All the studies looked at human plasma, however each used a different analytical

methodology. This includes a difference in run time: IPF = 4 mins, lung cancer = 20 mins, and PCOS = 14 minutes. The results are summarized in Figure 4.9.

The first thing to note is the family proportions for each experiment (Bar chart in Figure 4.10). For the most part, phospholipids (GP) are more abundant than Glycerolipids (GL), followed by sphingolipids (SP), sterol lipids (ST) and fatty acids (FA), with only trace amounts of the other families. A major exception to this is that IPF has a greater proportion of FA and SP than the other studies. This could be due to greater utilization of FA in cancer and PCOS samples to form GP and GL, with PCOS also using more GL to produce GPs (caused by a difference in the extraction method).

In terms of the proportions of adducts, the IPF study again differs from the other two. The most abundant adduct is [M+H], but in IPF this amounts to about a third of the adducts, whereas it is >50% in cancer and PCOS. The cancer methodology resulted in [M+NH4] making up a third of the adducts. The IPF study, however, yielded [M-H], [M+H-H2O] and [M+Ac-H] totaling >50% of adducts, whereas these were <20% for the other studies. This could be due to the different distribution of lipid families observed, or the much shorter run time in the IPF study.

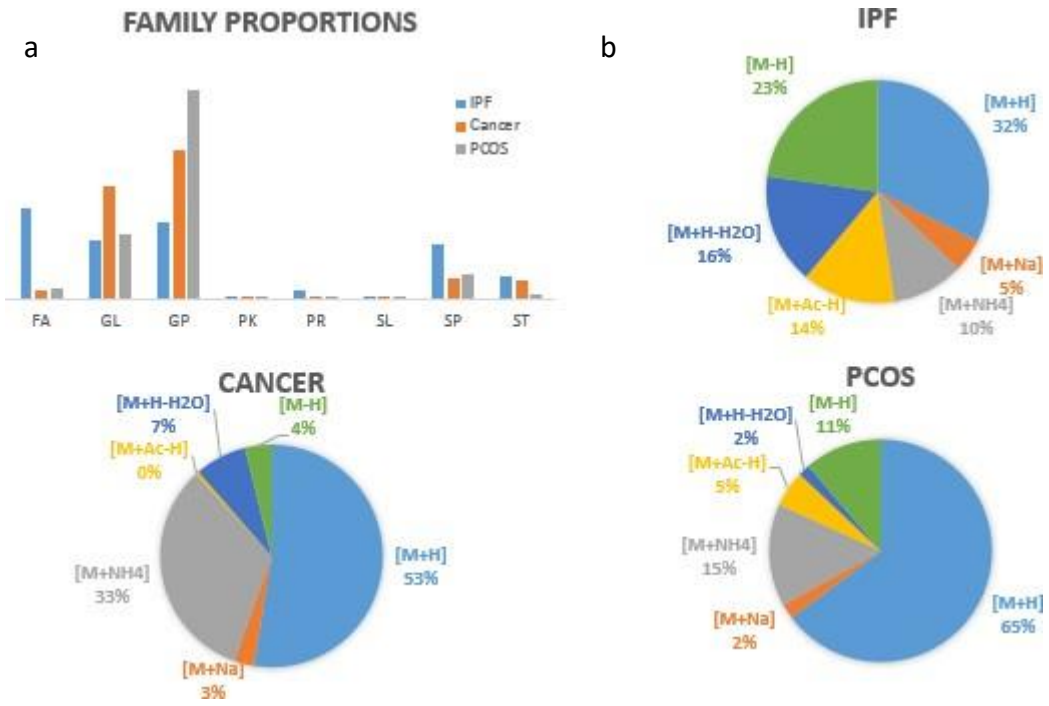


Figure 4.9: Analysis of lipid families and adducts. Panel (a) shows a bar chart of the proportions of lipid family in each experiment, using the abbreviations defined in the main text. Panels (b)-(d) contain pie charts showing the percentages of the different colour-coded adducts, respectively for IPF, cancer and PCOS.

To test the hypothesis that the types of adducts might be influenced by the lipid families present, the data for the three experiments was re-analyzed to produce Table 4.2, separating the proportion of adducts within their lipid families. It contains each experiment as a set rows for the most frequent 5 families of lipids. The other three families were not included due to their very low occurrence.

Table 4.2.: Adduct proportions for each lipid family in each study. The highlighted cells are discussed in the main text.

		[M+H]	[M+Na]	[M+NH4]	[M+Ac-H]	[M+H-H2O]	[M-H]	[M-H+NaAc]	[M-H+2NaAc]	H+3NaAc]	[M-H2O-H]
IPF	FA	51%	7%	19%	1%	7%	16%	0%	0%	0%	0%
	GL	17%	2%	20%	42%	12%	2%	2%	0%	1%	3%
	GP	14%	3%	1%	10%	3%	67%	1%	1%	0%	0%
	SP	40%	3%	0%	13%	34%	4%	2%	1%	0%	1%
	ST	38%	8%	3%	2%	49%	1%	0%	0%	0%	0%
Lung Cancer	FA	63%	6%	2%	1%	2%	25%	0%	0%	0%	0%
	GL	5%	3%	83%	0%	8%	0%	0%	0%	0%	0%
	GP	89%	2%	3%	0%	1%	5%	1%	0%	0%	0%
	SP	77%	2%	5%	4%	1%	9%	0%	2%	0%	0%
	ST	16%	5%	17%	0%	62%	0%	0%	0%	0%	0%
PCOS	FA	38%	3%	2%	5%	4%	45%	2%	0%	0%	1%
	GL	17%	5%	61%	11%	4%	1%	1%	0%	0%	1%
	GP	82%	1%	1%	2%	0%	13%	0%	0%	0%	0%
	SP	60%	9%	8%	16%	1%	4%	1%	0%	0%	2%
	ST	11%	2%	59%	3%	23%	1%	0%	0%	0%	0%

The adducts [M-H+XNaAc] and [M-H2O-H], shown in orange, have hardly been detected. These adducts are not worth the time to add them into the identification process since it will only add run time.

The sodium adducts, shown in blue, are comparatively rare but were still detected in all families and studies. For this reason, we should not remove them from the lipid-identification process. However, if needing to choose between [M+H] and [M+Na], the logical option would be M+H due to its more regular presence.

The Sterol Lipids (ST) in all studies showed a strong preference for [M+H-H2O], shown in dark green. This rule can be helpful in the identification process and provides an interesting point for further study.

Overall, the profiles of the Lung cancer and PCOS studies appeared to be more alike than IPF. The Glycerolipids of the two studies (in pale green) showed a strong preference for [M+NH4], but comparatively little for [M+H]. The Sterol Lipids in PCOS also showed a stronger preference for [M+NH4]. However, the phospholipids (GP) in cancer and PCOS, shown in yellow, most often bound to [M+H], while in IPF seemed to prefer [M-H]. In purple, it is observed that the [M-H] normally appears on FA; this can help in deciding whether an M-H adduct is more correct over other options.

These experiments have generated a set of rules for identification. It is typical in the identification process that two or more possible identifications appear possible, sometimes even members of the same family, but others completely different species with different adducts. The use of these rules can aid in making the decision without going to the MS-MS analysis. For example, in the case of a mass that has two possible identifications; of a [M-H] GL and an [M+H-H_2O] ST, the second one is clearly more likely.

4.4.3 Adduct distribution in fatty-acid subcategories

The 13 subcategories of FA were investigated further, by dividing them into three groups: **positive** (including Fatty Acids and Conjugates [FA01], Octadecanoids [FA02], Eicosanoids [FA03], Docosanoids [FA04] and Fatty ethers [FA10]), **negative** (including Fatty amides [FA08] and Fatty nitriles [FA09]), and **neutral** (including Fatty alcohols [FA05], Fatty aldehydes [FA06], Fatty esters [FA07], Oxygenated hydrocarbons [FA12], Fatty acyl glycosides [FA13] and Other Fatty Acyls [FA00]). This was achieved using the program Graph-FA-separator.py (see Appendix 4.B.h).

This program creates 2 dictionaries; one for adducts and one for FA subclasses. Like the previous one (where manual curation on excel is needed), the intensities are/were considered and (for this case only) it is/was also reported and plotted when all identities were from the same family. The output was visualized using excel (see Figure 4.10). The expected outcome was that negative fatty acids would only be found in negative mode, positive fatty acids in positive mode and neutral in both.

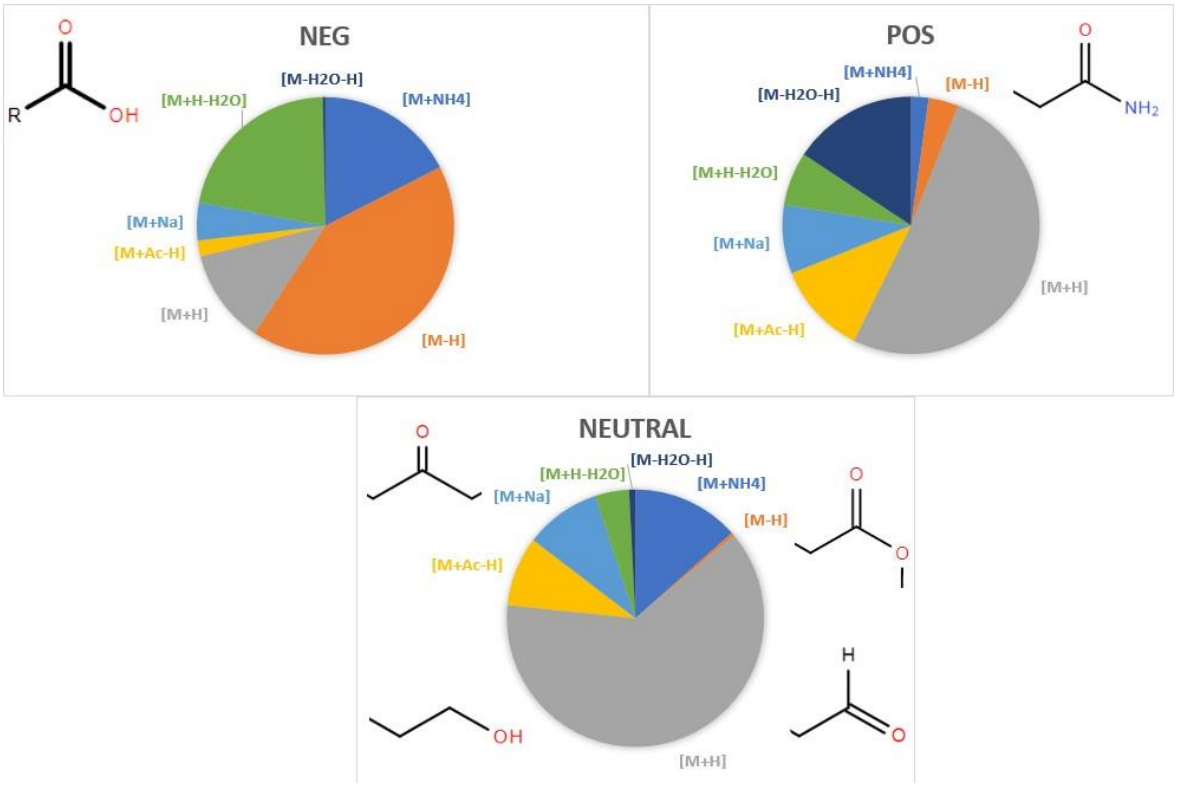


Figure 4.10: Adduct distribution in the different fatty acid subcategories. The three pie charts present the proportion of adducts for positive FA negative and neutral. The molecules help visualize the type of fatty acids in each category.

The results, while fitting the hypothesis, do not follow the rules perfectly as it is possible to see positive adducts on the negative fatty acids and vice versa. The most impressive effect has been on neutral fatty acids; almost all of them have been identified in positive mode.

4.5 Correction of ^{13}C marked species

In recent years, there has been an increase in the use of ^{13}C labelling for data normalization (Bueschl et al. 2013; Schatschneider et al. 2018). The work described in this section has been done with Sergey Evseev (PhD student, Pharmacy, University of Nottingham) and is soon to be published. This section describes the algorithm that made one of the critical steps possible.

The aim of the project is to use uniformly labelled bacteria as a source of internal standards for normalizing metabolomes in human and animal samples (untargeted metabolomics). This technique will allow for the correction of data variation between analytical replicates and reduce the inaccuracies in the data, caused by ion-source fragmentation and matrix effects. While this technique tries to cover many metabolites, it is not possible to cover 100% of them because it can be secured a 100% labelling (O'Maille et al., 2008).

The program is needed because the pre-processing tool IDEOM does not provide correct, full ^{13}C normalization. The software is not able to find the ^{13}C peaks to normalize, and so this task would have had to be done manually. To manually go over list of 1000 peaks to select the appropriate ones for over 500 metabolites would take weeks. For this reason, a python program was done to perform such task.

The normalization process consists of; for a metabolite, the different peak areas for each sample are divided by the peak area of the same C13-labelled metabolite, this should lead to lower RSD on the samples and therefore more rigorous and reproducible results.

The program (List-mass-c13-check-ACT.py in Appendix 4.B.b) has been developed, it takes the output from IDEOM and mzMatch (output from the chromatography). It checks for ^{13}C peaks that match in the user-specified ppm and retention time ranges for original peaks that must be normalized. More information on how this is done can be found in (Figure 4.2).

Users must edit the script directly to specify first the ranges of ppm, retention time (RT), and then the number of data columns. Normally the methods used are coelution (low RT bracket and high ppm range) and stick (small RT and 8ppm). The IDEOM and mzMatch spreadsheets must be in a specific format (described in the appendix) and retention times must be in the

same units, normally minutes. This python program requires the library, *xlrd*, to read excel spreadsheets and the common python libraries *re* and *numpy*. The program output is readable in Excel. On typical datasets, the program ran to completion in less than 40 minutes.

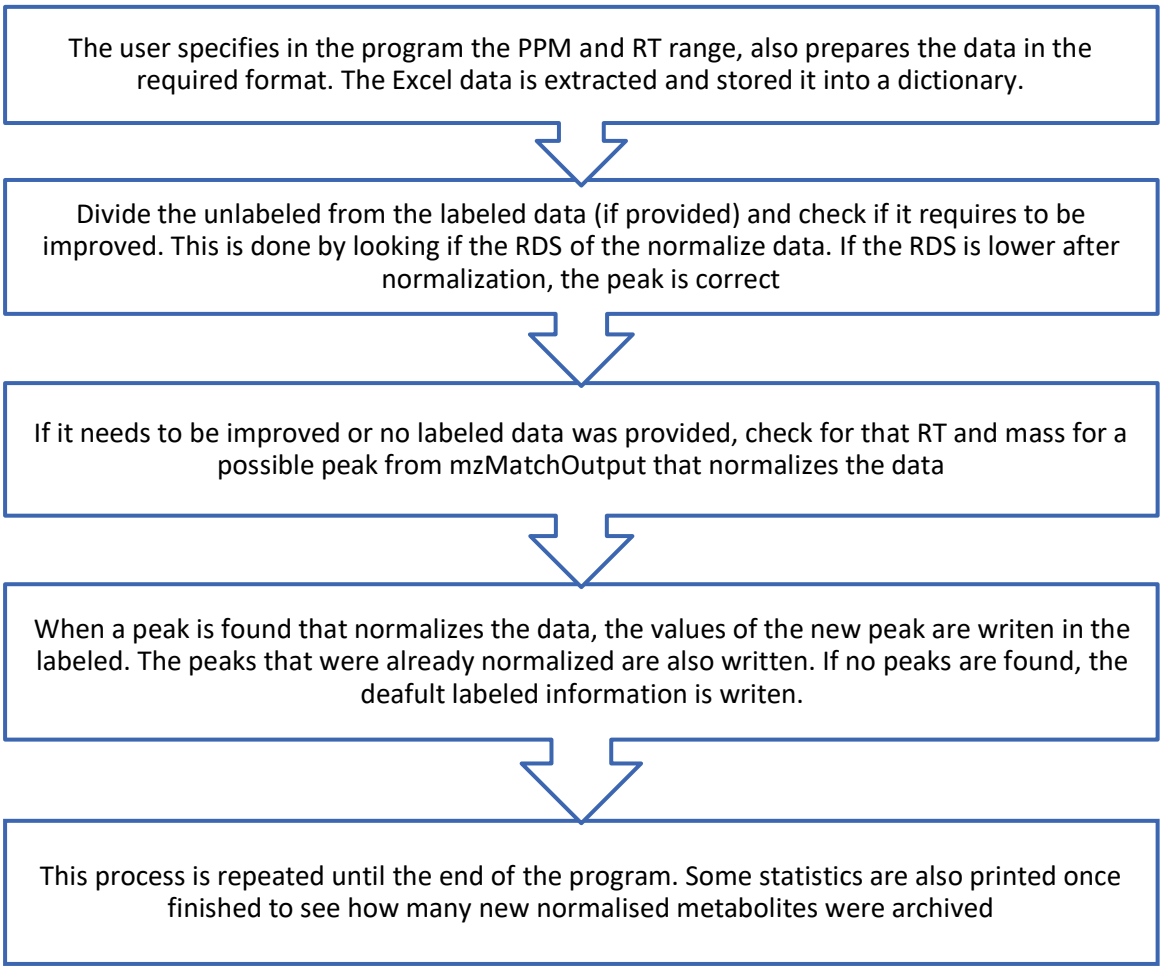


Figure 4.11: Flow chart of List-mass-C13-check-ACT.py. The flow chart describes the main steps taken by the program.

The results and discussion about the experiment can be found in the Appendix chapter 3 C.

4.6 Data management and arrangement

Some commercial software does not allow for too much data output manipulation. A complication regarding this matter is the output from the quantitative part of Xcalibur software by Thermo (the MS machine software manufacturer).

This software saves the information for each metabolite in a single excel sheet. For runs of over 100 metabolites where only one of the columns is of interest (normally the peak area), it can be tedious and time consuming to pick them one by one and paste them into the same sheet.

Despite its wide range of functions, no easy solution is provided within Excel and the current version of Xcalivur does not allow for any data output format change. Hence, the python program Excelprocessing-2.0.py have been developed to simplify the data format manipulation. It does a task that could take up to 2 hours in a few minutes. The script can be found in Appendix 4.B.c.

The program has a small interface, so no minimal knowledge of python is required to use it. The input is the excel sheet which needs to be reformatted, and the output is a table of the compiled data. The user simply selects the column of interest to read. The program has been used by at least 10 lab co-workers and, since it is quite easy to use, it has been shown to other Xcalibur users. Informal feedback from users indicate that the program is useful and simple to use. A training seminar has also been delivered on how to use the tool.

4.7 Conclusions

The aim of the work described in this chapter has been to improve the efficiency of lipidomics pipelines. This has been achieved through the development of a set of python programs, which have been used multiple times by more than one user. Table 4.3 shows the time reductions achieved using these programs, which reduce typical analysis times.

Table 4.3: Bottleneck reduction. The table summarizes the results of the different programs.

Pipeline step	**Typical process time**	**Process time (with this software)**
Correction of C13 species	40 hours	2 hours
Data reformatting	~2 hours	~30 seconds
Lipid Identification	3 minutes per metabolite	~5 seconds per metabolite

The general trend of lipidomics has been focusing on producing short lists of lipids instead of a whole report of all the ones that change. Mainly because some of the steps would take more time, these tools allow to manage bigger data easier. This chapter also describes the results of study protocols on the types of adduct observed.

The work done with the adducts in section 4.5 has provided us with some key conclusions. The study has given some novel information and help in the tasks of the lipidomic pipeline.

In the work done during the retention time experiment, it has been observed that the different adduct concentrations are changed across the retention time. This variation is majorly caused by the change of solvent, temperature and other chromatographic conditions. For this reason, the information obtained is hardly usable for other experiments in different conditions. However, it can be used if the method is fixed and used frequently.

The work done with the family separation has generated a key set of identification rules. Since the work has been done for three different experiments with three completely different conditions, it is safe to assume that the rules generated are correct in most lipidomic experiments. The list of these rules can be found on the Subsection 4.5.2.

The work done on the FA separated by polarity has verified the common assumption that positive FA should be found in positive mode, and that negative FA should be found in negative mode. Still, with some exceptions, in both cases it is not an absolute truth.

All this adduct work has quantification; something that has not been done before. Other papers focus mainly on the identification and characterization of a set of adducts, identifying the exact fragment m/z but not quantifying, like the detection of DMPO (Reis et al. 2003; Samuel et al. 2003). They are interested in adducts formed in the reaction of their study.

The other most common citation about adduct formation in HPLC-MS are papers using vague information about the adducts found (Nobilis et al., 2004); not quantifying, solely informing about the formation of them. Depending on the mobile-phase composition, other adducts may be observed as well, for example, ammonium acetate additive favors the formation of [M+NH4] ions. Rarely, adducts with mobile-phase molecules are found, such as [M+H+acetonitrile] or [M+H+ methanol] (Holčapek, Kolářová and Nobilis, 2008).

5. Conclusions and Future work

The aims of the thesis were to apply network biology to lipidomics to elucidate potential causes of lipid perturbations, apply network biology to the lipid pathway of the pain and inflammation, specifically the study of oxylipins in OA and develop bioinformatics tools to speed up the data analysis step of the lipidomic analysis pipeline. Significant progress has been made in these areas and a detailed discussion of how these challenges were faced is in the following three Sections.

5.1 Integration of lipidomic data into systems biology

In this thesis, a methodology to convert standard lipidomic output data into systems biology data has been crafted. This methodology has been used for the study of a set of potential biomarkers in Polycystic ovary syndrome, Osteoarthritis, rat and bat lipidomic data and some others. It has been implemented in an HTML platform for the usage of other institutions and personnel. Most of the methodology can be extrapolated to metabolomics and used in many other data sets.

The generated lipid *homo sapiens* network is a potential tool to be used, since the pathways were never connected due to their structural complexity; the novel network offers the possibility to visualize and study the whole picture. Lipid pathways found are presented as always separated, some of them with the different lengths of fatty acids taken into consideration, with others only having the description of the polar section of the lipids. This is due to the low specificity of the enzymes that carry most of these reactions. The integration of lipidomic data under this generically notated pathway is almost impossible.

The network generated using python algorithms has a node for every lipid with the exclusion of the non-identifiable by MS; these are the ones that interchange sn1 and sn2 fatty acid chains. One of the biggest challenges has been that the network has required us to move away from the normal functions and properties of graph theory to create a scoring method of our own. The typical scoring properties of nodes would not allow us to integrate quantitative data from lipidomics. This new scoring system can be used for other networks with similar structures, since this method has already been validated on the standard bioinformatics tests. This published methodology has proven useful as it can be on the usage reported on the subsection 2.6.

5.2 Osteoarthritis potential biomarkers and mediators

Osteoarthritis is associated generally to ageing, obesity and mechanical stress on the articular cartilage. However, there has been an emerging amount of evidence (Masuko et al. 2009; Villalvilla et al., 2013a; Korotkova et al. 2014; Brouwers et al. 2015; de Munter et al. 2016) that highlights the role of a broad spectrum of lipid mediators, including fatty acids, sphingolipids, and oxylipins in cartilage degradation in OA. Because OA is a disease affecting the bone, muscles, synovial fluid, joint and cartilage, they can suffer lipid dysregulation (Lago et al., 2007; Conde et al., 2015; Sun et al., 2016). Some of the most key lipids are the mediators of pain and inflammation and for this reason studies on oxylipins are so frequent.

Thanks to the collaboration with several OA projects, a set of OA lipidomic data has been collected, with targeted and untargeted analyses of humans and rodents. All these data have provided a chance to craft a view of the oxylipin network as a whole. This approach can be used in any other disease associated to lipid metabolism, and it allows for a quick visualization of the results. The OA datasets were useful for supporting hypotheses about transcriptional regulation of different oxylipin components. PG and TX levels are normally following the same trend, consistent with them being downstream of PGH2, though the high-fat diet experiment suggests TBXAS1 is separately regulated, perhaps by one of the other TFs shown in Figure 3.2.

The high-fat diet experiment suggests that the enzymes producing DHETs and EETs also appear to have a completely different regulators. In vitro and animal model studies have already shown that the EETs possess anti-inflammatory activity that is directed toward reducing, resolving, and limiting the damage caused by inflammation (Tacconelli et al. 2014; Thomson et al. 2012). However, the work here suggests that DHETs have greater importance, because they present a more consistent change across all inflammation datasets.

5.3 Advancements on lipidomic and metabolomic data processing.

The data analysis of the lipidomic pipeline is a complex, unregulated step. There is a lot of software that can do the same task in a slightly different manner, there is no standard normalization method (normally the most convenient is picked) and there is no agreed method or set of instructions. Many publications use for global lipidomics PLS (Perrotti et al. 2016), OPL-DA or even machine learning tools (Acharjee et al. 2016) to get the significant list of lipids. For this reason, many of the tasks have been done manually or with low automatization, these tests require a lot of double checking and parameter editing.

In this thesis in chapter 4, a set of programs to speed up and help with some of the steps have been created. Some programs aim to assist in the most general spec of dealing with MS data, and they are applicable to other omics techniques that use MS such as metabolomics and proteomics.

The lack of a good ^{13}C normalization tool for metabolomics has required the creation of a python program to try to pair some of the peaks and normalize them. This tool does the manual task that the software IDEOM does not complete. While IDEOM is excellent on the identification of the metabolites, it is not capable of picking out the ^{13}C normalization peaks. This is a tool which specifically deals with the IDEOM data and has been designed in this manner, however the algorithm can be extrapolated for other environments.

Lipid identification is another problematic subject discussed in this thesis, the creation of a tool to perform this process on a large, 'in bulk' scale has resulted in some interesting studies. The work done on adducts, their identification along the retention time and lipid families with quantification, has created an opportunity to explore and justify some of the most common generalizations.

5.4 Future work

The power of bioinformatic tools is to guide science, especially in proposing hypotheses for laboratory experiments. In order to increase the quality and the power of this work, I feel more lipidomic data should be generated. The first experiment would be a global lipidomic study of a set of rodents comparing the effect of gene knockdowns of genes encoding enzymes in the lipidomic network (Table 2.4). Such data would be the best for validating how good the methodology described in in Chapter 2 is and verify the ability to track enzymatic alterations from lipidomic profiles. The results in Chapter 3 suggest that something is happening to EPHX2 activity, and there is a general trend of low concentrations of HODEs. This is an area ripe for further laboratory investigation, as are experiments to test hypotheses about the genetic regulation of enzymes in the oxylipin network.

In the analysis of adducts, three different analytical methods were investigated, and the results picked up a general trend. These conclusions should be verified with more experiments using similar analytical conditions. This could lead to specific experimental protocols that create useful adduct patterns or at least minimise their impact during data analysis.

On the computational side, an interesting idea would be a full integration of the lipidomic network with other gene/enzyme interaction networks. This would require modifying the methodology described in Chapter 2 to distinguish between lipidomic-network edges and those involved in altering enzyme activity by transcription factors and other regulators.

An excellent improvement for the Software developed in Chapter 2 would be to provide a visualization tool like cytoscape. This would allow user to see immediate feedback without having to go through the steps of importing to cytoscape etc. This will also protect the instructions on how to create the visualization of the network from future cytoscape updates.

Bibliography

Abdelmohsen, K., Srikantan, S., Yang, X., Lal, A., Kim, H., Kuwano, Y., Galban, S., Becker, K., Kamara, D., de Cabo, R. and Gorospe, M. (2009). Ubiquitin-mediated proteolysis of HuR by heat shock. The EMBO Journal, 28(9), pp.1271-1282.

Acharjee, A., Ament, Z., West, J., Stanley, E. and Griffin, J. (2016). Integration of metabolomics, lipidomics and clinical data using a machine learning method. BMC Bioinformatics, 17(S15).

Agmon E, Stockwell BR (2017) Lipid homeostasis and regulated cell death. Curr Opin Chem Biol 39:83–89

Anand S, Young S, Esplin MS, Peaden B, Tolley HD, Porter TF et al (2016) Detection and confirmation of serum lipid biomarkers for preeclampsia using direct infusion mass spectrometry. J Lipid Res 57(4):687–696

Andreyev AY, Fahy E, Guan Z, Kelly S, Li X, McDonald JG et al (2010) Subcellular organelle lipidomics in TLR-4-activated macrophages. J Lipid Res 51(9):2785–2797

Arima, M. and T. Fukuda (2011). Prostaglandin D(2) and T(H)2 inflammation in the pathogenesis of bronchial asthma. Korean J Intern Med 26(1): 8-18.

Ashraf S, Mapp PI, Burston J, Bennett AJ, Chapman V, Walsh DA (2014). Augmented pain behavioural responses to intra-articular injection of nerve growth factor in two animal models of osteoarthritis. Ann Rheum Dis 73(9): 1710-1718.

Assfalg, M., I. Bertini, D. Colangiuli, C. Luchinat, H. Schafer, B. Schutz, et al. (2008). Evidence of different metabolic phenotypes in humans. Proc Natl Acad Sci U S A 105, 1420-4

B. Wang, C. Nolte, S. Aoki, H. Kettenmann and K. Wada (2007). Neuroprotective role of bradykinin because of the attenuation of pro-inflammatory cytokine release from activated microglia. J Neurochem 101(2): 397-410.

Baker KR, Matthan NR, Lichtenstein AH, Niu J, Guermazi A, Roemer F, et al. (2012). Association of plasma n-6 and n-3 polyunsaturated fatty acids with synovitis in the knee: the MOST study. Osteoarthritis Cartilage 20(5): 382-387.

Balgoma D, Checa A, Sar DG, Snowden S, Wheelock CE (2013). Quantitative metabolic profiling of lipid mediators. Mol Nutr Food Res 57(8): 1359-1377.

Barabási, A., Gulbahce, N. and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature Reviews Genetics, 12(1), pp.56-68.

Barton, M. (2011). The discovery of endothelium-dependent contraction: the legacy of Paul M. Vanhoutte. Pharmacol Res 63(6): 455-462.

Bennett A. J. , D. A. Walsh and V. Chapman (2016). The anti-NGF antibody muMab 911 both prevents and reverses pain behaviour and subchondral osteoclast numbers in a rat model of osteoarthritis pain. Osteoarthritis Cartilage 24(9): 1587-1595.

Benton H.P., D.M. Wong, S.A. Trauger, G. Siuzdak, (2008) XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. Anal. Chem. 80 6382–6389.

Bhardwaj A, Northington FJ, Carhuapoma JR, et al. P-450 epoxygenase and NO synthase inhibitors reduce cerebral blood flow response to N-methyl-Daspartate. Am J Physiol Heart Circ Physiol 2000; 279:H1616-H1624.

Bligh EG, Dyer WJ (1959) A rapid method of total lipid extraction and purification. Can J Biochem Physiol 37(8):911–917

Bombardieri, S., P. Cattani, G. Ciabattoni, O. Di Munno, G. Pasero, C. Patrono, E. Pinca and F. Pugliese (1981). THE SYNOVIAL PROSTAGLANDIN SYSTEM IN CHRONIC INFLAMMATORY ARTHRITIS: DIFFERENTIAL EFFECTS OF STEROIDAL AND NONSTEROIDAL ANTI-INFLAMMATORY DRUGS. British Journal of Pharmacology 73(4): 893-901.

Brantley, M., Edelstein, S., King, J., Plotzke, M., Apte, R., Kymes, S. and Shiels, A. (2008). Association of complement factor H and LOC387715 genotypes with response of exudative age-related macular degeneration to photodynamic therapy. Eye, 23(3), pp.626-631.

Brash AR, Ingram CD, Harris TM. 1987. Analysis of a specific oxygenation reaction of soybean lipoxygenase-1 with fatty acids esterified in phospholipids. Biochemistry 26:5465–71

Breitbach ZS, Berthod A, Huang K, Armstrong DW. Mass spectrometric detection of trace anions: The evolution of paired-ion electrospray ionization (PIESI). Mass Spectrometry Reviews. 2016;35(2):201-18.

Brouwers H, von Hegedus J, Toes R, Kloppenburg M, Ioan-Facsinay A (2015). Lipid mediators of inflammation in rheumatoid arthritis and osteoarthritis. Best Pract Res Clin Rheumatol 29(6): 741-755.

Brouwers, H., von Hegedus, J., Toes, R., Kloppenburg, M. and Ioan-Facsinay, A. (2015). Lipid mediators of inflammation in rheumatoid arthritis and osteoarthritis. Best Practice & Research Clinical Rheumatology, 29(6), pp.741-755.

Bueschl, C., Kluger, B., Lemmens, M., Adam, G., Wiesenberger, G., Maschietto, V., Marocco, A., Strauss, J., Bödi, S., Thallinger, G., Krska, R. and Schuhmacher, R. (2013). A novel stable isotope labelling assisted workflow for improved untargeted LC–HRMS based metabolomics research. Metabolomics, 10(4), pp.754-769.

Buja, L. M., T. Kita, J. L. Goldstein, Y. Watanabe and M. S. Brown (1983). Cellular pathology of progressive atherosclerosis in the WHHL rabbit. An animal model of familial hypercholesterolemia. Arteriosclerosis, Thrombosis, and Vascular Biology 3(1): 87-101.

Buré, C., Ayciriex, S., Testet, E. and Schmitter, J. (2012). A single run LC-MS/MS method for phospholipidomics. Analytical and Bioanalytical Chemistry, 405(1), pp.203-213.

Cai X, Li R (2016) Concurrent profiling of polar metabolites and lipids in human plasma using HILIC-FTMS. Sci Rep 6:36490

Chen, C., Lin, P., Tsai, M. and Lee, H. (2017). Targeted lipidomics profiling of acute arsenic exposure in mice serum by on-line solid-phase extraction stable-isotope dilution liquid chromatography–tandem mass spectrometry. Archives of Toxicology, 91(9), pp.3079-3091.

Colas, R. A., A. W. Ashton, S. Mukherjee, J. Dalli, O. B. Akide-Ndunge, H. Huang, M. S. Desruisseaux, F. Guan, L. A. Jelicks, F. Matos dos Santos, J. Nagajyothi, M. Zingman, J. Reyes, L. M. Weiss, C. N. Serhan and H. B. Tanowitz (2018). Trypanosoma cruzi produces the specialized proresolving mediators Resolvin D1, Resolvin D5 and Resolvin E2. Infection and Immunity.

Collins J.R. , B.R. Edwards, H.F. Fredricks, B.A.S. Van Mooy, (2016) LOBSTAHS: an adductbased lipidomics strategy for discovery and identification of oxidative stress biomarkers, Anal. Chem. 88 7154–7162.

Conde J, Scotece M, Abella V, Gomez R, Lopez V, Villar R, et al. (2015). Identification of novel adipokines in the joint. Differential expression in healthy and osteoarthritis tissues. PLoS One 10(4): e0123601.

Cordeiro, F. B., T. R. Cataldi, K. J. Perkel, L. do Vale Teixeira da Costa, R. C. Rochetti, J. Stevanato, et al. (2015). Lipidomics analysis of follicular fluid by ESI-MS reveals potential biomarkers for ovarian endometriosis. Journal of Assisted Reproduction and Genetics

Creelman, R. and Mulpuri, R. (2002). The Oxylipin Pathway in Arabidopsis. The Arabidopsis Book, 1, p.e0012.

Dallmann R, Viola AU, Tarokh L, Cajochen C, Brown SA. The human circadian metabolome. Proceedings of the National Academy of Sciences of the United States of America. 2012 Feb 14;109(7)

Dallmann, R., A. U. Viola, L. Tarokh, C. Cajochen, S. A. Brown (2012). The human circadian metabolome. Proc Natl Acad Sci U S A 109, 2625-9

de Moor, A. (1997). Applying conceptual graph theory to the user-driven specification of network information systems. Conceptual Structures: Fulfilling Peirce's Dream, pp.536-550.

de Munter W, van der Kraan PM, van den Berg WB, van Lent PL (2016). High systemic levels of low-density lipoprotein cholesterol: fuel to the flames in inflammatory osteoarthritis? Rheumatology (Oxford) 55(1): 16-24.

De, Sagar S. & Dehuri, Satchidananda & Wang, Gi-Nam. (2012). Machine Learning for Social Network Analysis: A Systematic Literature Review. The IUP Journal of Information Technology. 8. 30-51.

Deems, R., M. W. Buczynski, R. Bowers-Gentry, R. Harkewicz and E. A. Dennis (2007). Detection and quantitation of eicosanoids via high performance liquid chromatography-electrospray ionization-mass spectrometry. Methods Enzymol 432: 59-82.

Devlin, T. (n.d.). Textbook of biochemistry. Hoboken, NJ: Wiley-Liss

Domingues A. P. (2003). Detection and characterization by mass spectrometry of radical adducts produced by linoleic acid oxidation. Journal of the American Society for Mass Spectrometry, 14(11), pp.1250-1261.

Dominiczak MH, Caslake MJ: Apolipoproteins: metabolic role and clinical biochemistry applications, Ann Clin Biochem 2011; 48:498-515.

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences of the United States of America. 2007 Feb 6;104(6)

Duncan, M., Roder, H. and Hunsucker, S. (2008). Quantitative matrix-assisted laser desorption/ionization mass spectrometry. Briefings in Functional Genomics and Proteomics, 7(5), pp.355-370.

Ecker, J. and Liebisch, G. (2014). Application of stable isotopes to investigate the metabolism of fatty acids, glycerophospholipid and sphingolipid species. Progress in Lipid Research, 54, pp.14-31.

Egan, K. M., J. A. Lawson, S. Fries, B. Koller, D. J. Rader, E. M. Smyth and G. A. FitzGerald (2004). COX-2-Derived Prostacyclin Confers Atheroprotection on Female Mice. Science 306(5703): 1954-1957.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews Genetics. 2010 Jun;11(6)

Fahy, E., Cotter, D., Byrnes, R., Sud, M., Maer, A., Li, J., Nadeau, D., Zhau, Y. and Subramaniam, S. (2007). Bioinformatics for Lipidomics. Methods in Enzymology, pp.247-273.

Fan, F. and Roman, R. (2017). Effect of Cytochrome P450 Metabolites of Arachidonic Acid in Nephrology. Journal of the American Society of Nephrology, 28(10), pp.2845-2855.

Farooqui, A., Horrocks, L. and Farooqui, T. (2000). Glycerophospholipids in brain: their metabolism, incorporation into membranes, functions, and involvement in neurological disorders. Chemistry and Physics of Lipids, 106(1), pp.1-29.

Fauland, A., Köfeler, H., Trötzmüller, M., Knopf, A., Hartler, J., Eberl, A., Chitraju, C., Lankmayr, E. and Spener, F. (2011). A comprehensive method for lipid profiling by liquid chromatography-ion cyclotron resonance mass spectrometry. Journal of Lipid Research, 52(12), pp.2314-2322.

Feingold KR, Grunfeld C (2000) Introduction to lipids and lipoproteins. In: De Groot LJ, Chrousos G, Dungan K, Feingold KR, Grossman A, Hershman JM et al (eds) Endotext. MDText.com, Inc., South Dartmouth

Feldstein, A., Lopez, R., Tamimi, T., Yerian, L., Chung, Y., Berk, M., Zhang, R., McIntyre, T. and Hazen, S. (2010). Mass spectrometric profiling of oxidized lipid products in human nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. Journal of Lipid Research, 51(10), pp.3046-3054.

Fernihough J, Gentry C, Malcangio M, Fox A, Rediske J, Pellas T, et al. (2004). Pain related behaviour in two models of osteoarthritis in the rat knee. Pain 112(1-2): 83-93.

Feussner I, K¨uhn H. 2000. Application of lipoxygenases and related enzymes for the preparation of oxygenated lipids. In Enzymes in Lipid Modification, ed. UT Bornscheuer, pp. 309–36. Weinheim, Ger.: Wiley

Fahy, E., Sud, M., Cotter, D. and Subramaniam, S. (2007). LIPID MAPS online tools for lipid research. Nucleic Acids Research, 35(Web Server), pp.W606-W612.

Fleming, I., Fisslthaler, B., Popp, R., Kiss, L., Potente, M., Harder, D. and Busse, R. (1999). Nature, 401(6752), pp.493-497.

Folch J, Lees M, Sloane-Stanley GH (1957) A simple method for the isolation and purification of total lipides from animal tissues. J Biol Chem 226(1):497–509

Funk, C. D. (2001). Prostaglandins and Leukotrienes: Advances in Eicosanoid Biology. Science 294(5548): 1871-1875.

Futerman, A. and Riezman, H. (2005). The ins and outs of sphingolipid synthesis. Trends in Cell Biology, 15(6), pp.312-318.

Futerman, A.H. and Hannun, Y.A. (2004) The complex life of simple sphingolipids. EMBO Rep. 5, 777–782

Garcia-Canaveras JC, Peris-Diaz MD, Alcoriza-Balaguer MI, Cerdan-Calero M, Donato MT, Lahoz A (2017) A lipidomic cell-based assay for studying drug-induced phospholipidosis and steatosis. Electrophoresis 38:2331–2340

Ghetu, A., Corcoran, C., Cerchietti, L., Bardwell, V., Melnick, A. and Privé, G. (2008). Structure of a BCOR Corepressor Peptide in Complex with the BCL6 BTB Domain Dimer. Molecular Cell, 29(3), pp.384-391.

Ghosh, B. and Jones, A. (2015). Dependence of negative-mode electrospray ionization response factors on mobile phase composition and molecular structure for newly-authenticated neutral acylsucrose metabolites. The Analyst, 140(19), pp.6522-6531.

Gieger C. , N. Soranzo, J. Heinrich, M. Standl, E. Thiering, K. Mittelstraß, H.-E. Wichmann, A. Peters, K. Suhre, Y. Li, J. Adamski, T. D. Spector, T. Illig and R. Wang-Sattler (2012). Human serum metabolic profiles are age dependent. Aging Cell 11(6): 960-967.

Giera M, Ioan-Facsinay A, Toes R, Gao F, Dalli J, Deelder AM, et al. (2012). Lipid and lipid mediator profiling of human synovial fluid in rheumatoid arthritis patients by means of LC-MS/MS. Biochim Biophys Acta 1821(11): 1415-1424.

Gierman LM, Wopereis S, van El B, Verheij ER, Werff-van der Vat BJ, Bastiaansen-Jenniskens YM, et al. (2013). Metabolic profiling reveals differences in concentrations of oxylipins and fatty acids secreted by the infrapatellar fat pad of donors with end-stage osteoarthritis and normal donors. Arthritis Rheum 65(10): 2606-2614.

Gillmor SA,Villasenor A, Fletterick R, Sigal E, BrownerMF. (1998). The structure of mammalian 15-lipoxygenase reveals similarity to the lipases and the determinants of substrate specificity. Nat. Struct. Biol. 4:1003–9

Gimeno, R. (2007). Fatty acid transport proteins. Current Opinion in Lipidology, 18(3), pp.271-276.

Griffiths, W., Koal, T., Wang, Y., Kohl, M., Enot, D. and Deigner, H. (2010). Targeted Metabolomics for Biomarker Discovery. Angewandte Chemie International Edition, 49(32), pp.5426-5445.

Gu, R., Wei, Y., Jiang, H., Balazy, M. and Wang, W. (2001). Role of 20-HETE in mediating the effect of dietary K intake on the apical K channels in the mTAL. American Journal of Physiology-Renal Physiology, 280(2), pp.F223-F230.

Gurr, M., Harwood, J. and Frayn, K. (2002). Lipid biochemistry. Malden, Mass.: Blackwell Science.

Habibi, I., Emamian, E. and Abdi, A. (2014). Advanced Fault Diagnosis Methods in Molecular Networks. PLoS ONE, 9(10), p.e108830.

Halade, G., Dorbane, A., Ingle, K., Kain, V., Schmitter, J. and Rhourri-Frih, B. (2018). Comprehensive targeted and non-targeted lipidomics analyses in failing and non-failing heart. Analytical and Bioanalytical Chemistry, 410(7), pp.1965-1976.

Halcox, J., Narayanan, S., Cramer-Joyce, L., Mincemoyer, R. and Quyyumi, A. (2001). Characterization of endothelium-derived hyperpolarizing factor in the human forearm microcirculation. *American Journal of Physiology-Heart and Circulatory Physiology*, 280(6), pp.H2470-H2477.

Hall A., Kou K., Chen Z., Pietka T., Kumar M., Korenblat K., Lee K., Ahn K., Fabbrini E., Klein S., Goodwin, B. and Finck, B. (2012). Evidence for regulated monoacylglycerol acyltransferase expression and activity in human liver. Journal of Lipid Research, 53(5), pp.990-999.

Hamberg M., Samuelsson B. 1967. On the specificitiy of the oxygenation of unsaturated fatty acids catalyzed by soybean lipoxidase. J. Biol. Chem. 242:5329–35

Hammad, H., M. Kool, T. Soullie, S. Narumiya, F. Trottein, H. C. Hoogsteden and B. N. Lambrecht (2007). Activation of the D prostanoid 1 receptor suppresses asthma by modulation of lung dendritic cell function and induction of regulatory T cells. J Exp Med 204(2): 357-367.

Haoula, Z., S. Ravipati, D. Stekel, C. Ortori, C. Hodgman, C. Daykin, et al. (2015). Lipidomic analysis of plasma samples from women with polycystic ovary syndrome. Metabolomics 11, 657-666

Hara, A. (2005). Augmented Cardiac Hypertrophy in Response to Pressure Overload in Mice Lacking the Prostaglandin I2 Receptor. Circulation, 112(1), pp.84-92.

Hartler, J., Tharakan, R., Kofeler, H., Graham, D. and Thallinger, G. (2012). Bioinformatics tools and challenges in structural analysis of lipidomics MS/MS data. Briefings in Bioinformatics, 14(3), pp.375-390.

Hashmi, S., Y. Wang, D. S. Suman, R. S. Parhar, K. Collison, W. Conca, et al. (2015). Human cancer: is it linked to dysfunctional lipid metabolism? Biochim Biophys Acta 1850, 352-64

Hatzi, K., Jiang, Y., Huang, C., Garrett-Bakelman, F., Gearhart, M., Giannopoulou, E., Zumbo, P., Kirouac, K., Bhaskara, S., Polo, J., Kormaksson, M., MacKerell, A., Xue, F., Mason, C., Hiebert, S., Prive, G., Cerchietti, L., Bardwell, V., Elemento, O. and Melnick, A. (2013). A Hybrid Mechanism of Action for

BCL6 in B Cells Defined by Formation of Functionally Distinct Complexes at Enhancers and Promoters. Cell Reports, 4(3), pp.578-588.

Haughey, N. J., L. B. Tovar-y-Romo, V. V. Bandaru (2011). Roles for biological membranes in regulating human immunodeficiency virus replication and progress in the development of HIV therapeutics that target lipid metabolism. J Neuroimmune Pharmacol 6, 284-95

Havugimana M., Giannopoulou, E., Zumbo, P., Kirouac, K., Bhaskara, S., Polo, J., Kormaksson, M., MacKerell (2012) A census of human soluble protein complexes. PMC3477804,

He M., van Wijk, E., Berger, R., Wang, M., Strassburg, K., Schoeman, J., Vreeken, R., van Wietmarschen, H., Harms, A., Kobayashi, M., Hankemeier, T. and van der Greef, J. (2015). Collagen Induced Arthritis in DBA/1J Mice Associates with Oxylipin Changes in Plasma. Mediators of Inflammation, 2015, pp.1-11.

Hermansson, M., Hokynar, K. and Somerharju, P. (2011). Mechanisms of glycerophospholipid homeostasis in mammalian cells. Progress in Lipid Research, 50(3), pp.240-257.

Hijioka, M., Anan, J., Ishibashi, H., Kurauchi, Y., Hisatsune, A., Seki, T., Koga, T., Yokomizo, T., Shimizu, T. and Katsuki, H. (2016). Inhibition of Leukotriene B 4 Action Mitigates Intracerebral Hemorrhage-Associated Pathological Events in Mice. Journal of Pharmacology and Experimental Therapeutics, 360(3), pp.399-408.

Hiley, C. R., P. M. Hoi (2007). Oleamide: A Fatty Acid Amide Signaling Molecule in the Cardiovascular System? Cardiovascular Drug Reviews 25, 46-60

Hobbs, F. (2004). Cardiovascular disease: different strategies for primary and secondary prevention. Heart, 90(10), pp.1217-1223.

Holčapek, M., Kolářová, L. and Nobilis, M. (2008). High-performance liquid chromatography–tandem mass spectrometry in the identification and determination of phase I and phase II drug metabolites. Analytical and Bioanalytical Chemistry, 391(1), pp.59-78., N., Toyoda, Y., Gak, I., Weisswange, I., Mansfeld, J., Buchholz, F., Hyman, A. and Mann, M. (2015). A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. Cell

Holmes, E., R. L. Loo, J. Stamler, M. Bictash, I. K. Yap, Q. Chan, et al. (2008). Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 453, 396-400

Horai H., M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, et al., (2010) MassBank: a public repository for sharing mass spectral data for life sciences, J. Mass Spectrom. 45 703–714.

Hu J., Thomas, C. and Brunak, S. (2016). Network biology concepts in complex disease comorbidities. Nature Reviews Genetics, 17(10), pp.615-629.

Huang N., Siegel M.M., Kruppa G.H., Laukien F.H., J. Am. Soc. Mass. Spectrom. 1999, 10. Automation of a Fourier transform ion cyclotron resonance mass spectrometer for acquisition, analysis, and e-mailing of high-resolution exact-mass electrospray ionization mass spectral data

Hyötyläinen, T. and Orešič, M. (2014). Systems biology strategies to study lipidomes in health and disease. Progress in Lipid Research, 55, pp.43-60.

Illig, T., C. Gieger, G. Zhai, W. Romisch-Margl, R. Wang-Sattler, C. Prehn, et al. (2010). A genome-wide perspective of genetic variation in human metabolism. Nat Genet 42, 137-41 7

Vachek J., Z. Krupickova-Kasalova and J. Malik (2013). Endotelial activation and flow-mediated vasodilation in young patients with breast cancer. Neoplasma 60(6): 690-697.

Jewell A. , J., Lanza, T., Lin, L., Hagmann, W., Evans, D. and Kumar, S. (2003). Addressing the metabolic activation potential of new leads in drug discovery: a case study using ion trap mass spectrometry and tritium labeling techniques. Journal of Mass Spectrometry, 38(2), pp.211-221.

Jowsey, I. R., A. M. Thomson, J. U. Flanagan, P. R. Murdock, G. B. Moore, D. J. Meyer, G. J. Murphy, S.

A. Smith and J. D. Hayes (2001). Mammalian class Sigma glutathione S-transferases: catalytic properties and tissue-specific expression of human and rat GSH-dependent prostaglandin D2 synthases. Biochemical Journal 359(Pt 3): 507-516.

Takayama K. , S. Kuriyama, T. Taniguchi, T. Shiokoshi, Y. Ohsaki, K. Kikuchi, S. Narumiya and F. Ushikubi (2003). Thromboxane A2 regulates vascular tone via its inhibitory effect on the expression of inducible nitric oxide synthase. Circulation 108(19): 2381-2386.

Kamei, D., Yamakawa, K., Takegoshi, Y., Mikami-Nakanishi, M., Nakatani, Y., Oh-ishi, S., Yasui, H., Azuma, Y., Hirasawa, N., Ohuchi, K., Kawaguchi, H., Ishikawa, Y., Ishii, T., Uematsu, S., Akira, S., Murakami, M. and Kudo, I. (2004). Reduced Pain Hypersensitivity and Inflammation in Mice Lacking Microsomal Prostaglandin E Synthase-1. Journal of Biological Chemistry, 279(32), pp.33684-33695.

Kamel, K. and Halperin, M. (2017). Ketoacidosis. Fluid, Electrolyte and Acid-Base Physiology, pp.99-139.

Kanehisa, Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.; (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353-D361.

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

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M.; (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457-D462.

Katusic, Z. S., A. V. Santhanam and T. He (2012). Vascular effects of prostacyclin: does activation of PPARdelta play a role? Trends Pharmacol Sci 33(10): 559-564.

Keller, B., Sui, J., Young, A. and Whittal, R. (2008). Interferences and contaminants encountered in modern mass spectrometry. Analytica Chimica Acta, 627(1), pp.71-81.

Kenny, L. C., D. I. Broadhurst, W. Dunn, M. Brown, R. A. North, L. McCowan, et al. (2010). Robust early pregnancy prediction of later preeclampsia using metabolomic biomarkers. Hypertension 56, 741-9

Korotkova M, Jakobsson PJ (2014). Persisting eicosanoid pathways in rheumatic diseases. Nat Rev Rheumatol 10(4): 229-241.

Kramell R, Miersch O, Atzorn R, Parthier B, Wasternack C. (2000). Octadecanoidderived alteration of gene expression and the Oxylipin Signature in stressed barley leaves. Implications for different signalling pathways. Plant Physiol. 123:177– 88

Krug, S., G. Kastenmuller, F. Stuckler, M. J. Rist, T. Skurk, M. Sailer, et al. (2012). The dynamic range of the human metabolome revealed by challenges. FASEB J 26, 2607-19

Kuhl C. , R. Tautenhahn, C. Böttcher, T.R. Larson, S. Neumann, (2012) CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets, Anal. Chem. 84 283–289.

Kunori, S., S. Matsumura, T. Mabuchi, S. Tatsumi, Y. Sugimoto, T. Minami and S. Ito (2009). Involvement of prostaglandin F 2 alpha receptor in ATP-induced mechanical allodynia. Neuroscience 163(1): 362-371.

Lago F, Dieguez C, Gomez-Reino J, Gualillo O (2007). Adipokines as emerging mediators of immune response and inflammation. Nat Clin Pract Rheumatol 3(12): 716-724.

Lakshmipraveena, T & , n.v.muthulakshmi. (2017). A Review of Graph Based Algorithms in Social Media Data Analytics.

Lee, S., Williams, M., DuBois, R. and Blair, I. (2003). Targeted lipidomics using electron capture atmospheric pressure chemical ionization mass spectrometry. Rapid Communications in Mass Spectrometry, 17(19), pp.2168-2176.

Li, R., Wang, Y., Ma, Z., Ma, M., Wang, D., Xie, G., Yin, Y., Zhang, P. and Tao, K. (2016). Maresin 1 Mitigates Inflammatory Response and Protects Mice from Sepsis. Mediators of Inflammation, 2016, pp.1-9.

Libby, P., S. J. Warner and G. B. Friedman (1988). Interleukin 1: a mitogen for human vascular smooth muscle cells that induces the release of growth-inhibitory prostanoids. J Clin Invest 81(2): 487-498.

Lim A, Wenk MR, Tong L (2015) Lipid-based therapy for ocular surface inflammation and disease. Trends Mol Med 21(12):736–748

Lima-Garcia JF, Dutra RC, da Silva K, Motta EM, Campos MM, Calixto JB (2011). The precursor of resolvin D series and aspirin-triggered resolvin D1 display anti-hyperalgesic properties in adjuvant-induced arthritis in rats. Br J Pharmacol 164(2): 278-293.

Litwic, A., Edwards, M., Dennison, E. and Cooper, C. (2013). Epidemiology and burden of osteoarthritis. British Medical Bulletin, 105(1), pp.185-199.

Liu, Q., J. Zhang (2014). Lipid metabolism in Alzheimer's disease. Neurosci Bull 30, 331-45

Longo, N., Frigeni, M. and Pasquali, M. (2016). Carnitine transport and fatty acid oxidation. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1863(10), pp.2422-2435.

Lu, W., Bennett, B. and Rabinowitz, J. (2008). Analytical strategies for LC–MS-based targeted metabolomics. Journal of Chromatography B, 871(2), pp.236-242.

Lundstrom, S. L., F. L. D'Alexandri, K. Nithipatikom, J. Z. Haeggstrom, A. M. Wheelock and C. E. Wheelock (2009). HPLC/MS/MS-based approaches for detection and quantification of eicosanoids. Methods Mol Biol 579: 161-187.

Lydic TA, Busik JV, Reid GE (2014) A monophasic extraction strategy for the simultaneous lipidome analysis of polar and nonpolar retina lipids. J Lipid Res 55(8):1797–1809

M, A. and KS, Y. (2016). Pathogenesis of Atherosclerosis A Review. Medical & Clinical Reviews, 2(3).

Manolio TA. (2010) Genomewide Association Studies and Assessment of the Risk of Disease. New England Journal of Medicine. ;363(2)

Mapp PI, Sagar DR, Ashraf S, Burston JJ, Suri S, Chapman V, et al. (2013). Differences in structural and pain phenotypes in the sodium monoiodoacetate and meniscal transection models of osteoarthritis. Osteoarthritis Cartilage 21(9): 1336-1345.

Mapp PI, Walsh DA, Bowyer J, Maciewicz RA (2010). Effects of a metalloproteinase inhibitor on osteochondral angiogenesis, chondropathy and pain behavior in a rat model of osteoarthritis. Osteoarthritis Cartilage. 18(4): 593-600.

Marker, C. L. and J. D. Pomonis (2012). The monosodium iodoacetate model of osteoarthritis pain in the rat. Methods Mol Biol 851: 239-248.

Mashaghi, A., Ramezanpour, A. and Karimipour, V. (2004). Investigation of a protein complex network. The European Physical Journal B, 41(1), pp.113-121.

Masoodi, M. and A. Nicolaou (2006). Lipidomic analysis of twenty-seven prostanoids and isoprostanes by liquid chromatography/electrospray tandem mass spectrometry. Rapid Commun Mass Spectrom 20(20): 3023-3029.

Masoodi, M., Mir, A., Petasis, N., Serhan, C. and Nicolaou, A. (2008). Simultaneous lipidomic analysis of three families of bioactive lipid mediators leukotrienes, resolvins, protectins and related hydroxy-fatty acids by liquid chromatography/electrospray ionisation tandem mass spectrometry. Rapid Communications in Mass Spectrometry, 22(2), pp.75-83.

Masuko K, Murata M, Suematsu N, Okamoto K, Yudoh K, Nakamura H, et al. (2009). A metabolic aspect of osteoarthritis: lipid as a possible contributor to the pathogenesis of cartilage degradation. Clin Exp Rheumatol 27(2): 347-353.

Mathiessen, A. and Conaghan, P. (2017). Synovitis in osteoarthritis: current understanding with therapeutic implications. Arthritis Research & Therapy, 19(1).

Matsui K, Nishioka M, Ikeyoshi M, Matsumura Y, Mori T, Kajiwara T. 1998. Cucumber root lipoxygenase can act on acyl groups in phosphatidylcholine. Biochim. Biophys. Acta 1390:8–20

Maxfield FR, Van Meer G: Cholesterol, the central lipid of mammalian cells. Curr Opin Cell Biol 2010; 22:422-429

Men, L., Pi, Z., Zhou, Y., Wei, M., Liu, Y., Song, F. and Liu, Z. (2017). Urine metabolomics of high-fat diet induced obesity using UHPLC-Q-TOF-MS. Journal of Pharmaceutical and Biomedical Analysis, 132, pp.258-266.

Merico, D., Isserlin, R., Stueker, O., Emili, A. and Bader, G. (2010). Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. PLoS ONE, 5(11), p.e13984.

Mika, A. and Sledzinski, T. (2016). Alterations of specific lipid groups in serum of obese humans: a review. Obesity Reviews, 18(2), pp.247-272.

Moessinger, C., K. Klizaite, A. Steinhagen, J. Philippou-Massier, A. Shevchenko, M. Hoch, et al. (2014). Two different pathways of phosphatidylcholine synthesis, the Kennedy Pathway and the Lands Cycle

Mohamed-Hussein ZA, Harun S. Construction of a polycystic ovarian syndrome (PCOS) pathway based on the interactions of PCOS-related proteins retrieved from bibliomic data. Theoretical biology & medical modelling. 2009;6

Moreno, C. (2001). Abnormal pressure-natriuresis in hypertension: role of cytochrome P450 metabolites of arachidonic acid. American Journal of Hypertension, 14(11), pp.S90-S97.

Mozurkewich, E., Greenwood, M., Clinton, C., Berman, D., Romero, V., Djuric, Z., Qualls, C. and Gronert, K. (2016). Pathway Markers for Pro-resolving Lipid Mediators in Maternal and Umbilical Cord Blood: A Secondary Analysis of the Mothers, Omega-3, and Mental Health Study. Frontiers in Pharmacology, 07.

Hirasawa N., K. Ohuchi, H. Kawaguchi, Y. Ishikawa, T. Ishii, S. Uematsu, S. Akira, M. Murakami and I. Kudo (2004). Reduced pain hypersensitivity and inflammation in mice lacking microsomal prostaglandin e synthase-1. J Biol Chem 279(32): 33684-33695.

Newman, J. W., T. Watanabe and B. D. Hammock (2002). The simultaneous quantification of cytochrome P450 dependent linoleate and arachidonate metabolites in urine by HPLC-MS/MS. J Lipid Res 43(9): 1563-1578.

Nobilis, M., Kopecký, J., Květina, J., Svoboda, Z., Pour, M., Kuneš, J., Holčapek, M. and Kolářová, L. (2003). Comparative biotransformation and disposition studies of nabumetone in humans and minipigs using high-performance liquid chromatography with ultraviolet, fluorescence and mass spectrometric detection. Journal of Pharmaceutical and Biomedical Analysis, 32(4-5), pp.641-656.

Noda, M., Kariura, Y., Pannasch, U., Nishikawa, K., Wang, L., Seike, T., Ifuku, M., Kosai, Y., Wang, B., Nolte, C., Aoki, S., Kettenmann, H. and Wada, K. (2007). Neuroprotective role of bradykinin because of the attenuation of pro-inflammatory cytokine release from activated microglia. Journal of Neurochemistry, 101(2), pp.397-410.

Node, K. (1999). Anti-inflammatory Properties of Cytochrome P450 Epoxygenase-Derived Eicosanoids. Science, 285(5431), pp.1276-1279.

Nwosu, L. N., P. I. Mapp, V. Chapman and D. A. Walsh (2016). Blocking the tropomyosin receptor kinase A (TrkA) receptor inhibits pain behaviour in two rat models of osteoarthritis. Ann Rheum Dis 75(6): 1246-1254.

O'Connor A., C.J. Brasher, D.A. Slatter, S.W. Meckelmann, J.I. Hawksworth, S.M. Allen, et al., (2017) LipidFinder, a computational workflow for discovery of lipids identifies new eicosanoid-phosphoinositides in platelets, JCI Insight 2 e91634.

O'Maille, G., Go, E., Hoang, L., Want, E., Smith, C., O'Maille, P., NordstrÖm, A., Morita, H., Qin, C., Uritboonthai, W., Apon, J., Moore, R., Garrett, J. and Siuzdak, G. (2008). Metabolomics relative quantitation with mass spectrometry using chemical derivatization and isotope labeling. Spectroscopy, 22(5), pp.327-343.

Perrotti, F., Rosa, C., Cicalini, I., Sacchetta, P., Del Boccio, P., Genovesi, D. and Pieragostino, D. (2016). Advances in Lipidomics for Cancer Biomarkers Discovery. International Journal of Molecular Sciences, 17(12), p.1992.

Pillarisetti, S. and I. Khanna (2012). Targeting soluble epoxide hydrolase for inflammation and pain - an overview of pharmacology and the inhibitors. Inflamm Allergy Drug Targets 11(2): 143-158.

Pomposiello, S., Carroll, M., Falck, J. and McGiff, J. (2001). Epoxyeicosatrienoic Acid–Mediated Renal Vasodilation to Arachidonic Acid Is Enhanced in SHR. Hypertension, 37(3), pp.887-893.

Porubsky, P. R., Meneely, K. M., & Scott, E. E. (2008). Structures of Human Cytochrome P-450 2E1: INSIGHTS INTO THE BINDING OF INHIBITORS AND BOTH SMALL MOLECULAR WEIGHT AND FATTY ACID SUBSTRATES. The Journal of Biological Chemistry 283(48), 33698–33707.

Proulx, S., Promislow, D. and Phillips, P. (2005). Network thinking in ecology and evolution. Trends in Ecology & Evolution, 20(6), pp.345-353.

Pyragius, C., Fuller, M., Ricciardelli, C. and Oehler, M. (2013). Aberrant Lipid Metabolism: An Emerging Diagnostic and Therapeutic Target in Ovarian Cancer. International Journal of Molecular Sciences, 14(4), pp.7742-7756.

Quehenberger, O., Armando, A., Brown, A., Milne, S., Myers, D., Merrill, A., Bandyopadhyay, S., Jones, K., Kelly, S., Shaner, R., Sullards, C., Wang, E., Murphy, R., Barkley, R., Leiker, T., Raetz, C., Guan, Z., Laird, G., Six, D., Russell, D., McDonald, J., Subramaniam, S., Fahy, E. and Dennis, E. (2010). Lipidomics reveals a remarkable diversity of lipids in human plasma,. Journal of Lipid Research, 51(11), pp.3299-3305.

Quehenberger, O. and Dennis, E. (2011). The Human Plasma Lipidome. New England Journal of Medicine, 365(19), pp.1812-1823.

Quilley, J. and McGiff, J. (2000). Is EDHF an epoxyeicosatrienoic acid?. Trends in Pharmacological Sciences, 21(4), pp.121-124.

Reis A, Rudnitskaya A, Blackburn GJ, Fauzi NM, Pitt AR, Spickett CM (2013) A comparison of five lipid extraction solvent systems for lipidomic studies of human LDL. J Lipid Res 54(7):1812–1824

Reis, A., Rosário, M., Domingues, M., Amado, F., Ferrer-Correia, A. and Domingues, P. (2003). Detection and characterization by mass spectrometry of radical adducts produced by linoleic acid oxidation. Journal of the American Society for Mass Spectrometry, 14(11), pp.1250-1261.

Ricciotti, E. and FitzGerald, G. (2011). Prostaglandins and Inflammation. Arteriosclerosis, Thrombosis, and Vascular Biology, 31(5), pp.986-1000.

Ridgway, N. and McLeod, R. (2016). Biochemistry of lipids, lipoproteins and membranes. Amsterdam: Elsevier.

Rosenfeld, G. and Loose, D. (2014). Pharmacology. Baltimore, MD: Lippincott Williams & Wilkins.

Röst H.L., T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, et al., (2016) OpenMS: a flexible

open-source software platform for mass spectrometry data analysis, Nat. Methods 13 741–748.

Sakai, J. and Rawson, R. (2001). The sterol regulatory element-binding protein pathway: control of lipid homeostasis through regulated intracellular transport. Current Opinion in Lipidology, 12(3), pp.261-266.

Saleem, S., A. S. Ahmad, T. Maruyama, S. Narumiya and S. Dore (2009). PGF(2alpha) FP receptor contributes to brain damage following transient focal brain ischemia. Neurotox Res 15(1): 62-70.

Schatschneider, S., Abdelrazig, S., Safo, L., Henstra, A., Millat, T., Kim, D., Winzer, K., Minton, N. and Barrett, D. (2018). Quantitative Isotope-Dilution High-Resolution-Mass-Spectrometry Analysis of Multiple Intracellular Metabolites in Clostridium autoethanogenum with Uniformly 13C-Labeled Standards Derived from Spirulina. Analytical Chemistry, 90(7), pp.4470-4477.

Schonewille, M., Freark de Boer, J., Mele, L., Wolters, H., Bloks, V., Wolters, J., Kuivenhoven, J., Tietge, U., Brufau, G. and Groen, A. (2016). Statins increase hepatic cholesterol synthesis and stimulate fecal cholesterol elimination in mice. Journal of Lipid Research, 57(8), pp.1455-1464.

Schmelzer, K. R., A. M. Wheelock, K. Dettmer, D. Morin and B. D. Hammock (2006). The role of inflammatory mediators in the synergistic toxicity of ozone and 1-nitronaphthalene in rat airways. Environ Health Perspect 114(9): 1354-1360.

Schulz, H. (1991). Beta oxidation of fatty acids. Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism, 1081(2), pp.109-120.

Schwudke, D., Schuhmann, K., Herzog, R., Bornstein, S. and Shevchenko, A. (2011). Shotgun Lipidomics on High Resolution Mass Spectrometers. Cold Spring Harbor Perspectives in Biology, 3(9), pp.a004614-a004614.

Serhan, C. (2005). Mediator lipidomics. Prostaglandins & Other Lipid Mediators, 77(1-4), pp.4-14.

Serhan, C. N. (2014). Pro-resolving lipid mediators are leads for resolution physiology. Nature 510: 92.

Serna J, Garcia-Seisdedos D, Alcazar A, Lasuncion MA, Busto R, Pastor O (2015) Quantitative lipidomic analysis of plasma and plasma lipoproteins using MALDI-TOF mass spectrometry. Chem Phys Lipids 189:7–18

Shaham, O., R. Wei, T. J. Wang, C. Ricciardi, G. D. Lewis, R. S. Vasan, et al. (2008). Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. Mol Syst Biol 4, 214

Shimizu, T., Radmark, O. and Samuelsson, B. (1984). Enzyme with dual lipoxygenase activities catalyzes leukotriene A4 synthesis from arachidonic acid. Proceedings of the National Academy of Sciences, 81(3), pp.689-693.

Simopoulos, A. (2016). An Increase in the Omega-6/Omega-3 Fatty Acid Ratio Increases the Risk for Obesity. Nutrients, 8(3), p.128.

Sleno, L., Varesio, E. and Hopfgartner, G. (2007). Determining protein adducts of fipexide: mass spectrometry based assay for confirming the involvement of its reactive metabolite in covalent binding. Rapid Communications in Mass Spectrometry, 21(24), pp.4149-4157.

Sloane DL, Leung R, Barnett J, Craik CS, Sigal E. (1995). Conversion of human 15-lipoxygenase to an efficient 12-lipoxygenase: the side-chain geometry of amino acids 417 and 418 determine positional specificity. Protein Eng. 8:275–82

Smith C.A. , E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 779–787.

Smith C.A., O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G. (2005) METLIN: a metabolite mass spectral database. Ther Drug Moni. ; 27:747-51.

Smith, M. M., M. A. Cake, P. Ghosh, A. Schiavinato, R. A. Read and C. B. Little (2008). Significant synovial pathology in a meniscectomy model of osteoarthritis: modification by intra-articular hyaluronan therapy. Rheumatology (Oxford) 47(8): 1172-1178

Smith, W. and Murphy, R. (2016). The Eicosanoids. Biochemistry of Lipids, Lipoproteins and

Membranes, pp.259-296.

Spener, F., Lagarde, M., Géloên, A. and Record, M. (2003). Editorial: What is lipidomics?. European Journal of Lipid Science and Technology, 105(9), pp.481-482.

Stegemann C, Drozdov I, Shalhoub J, Humphries J, Ladroue C, Didangelos A et al (2011) Comparative lipidomics profiling of human atherosclerotic plaques. Circ Cardiovasc Genet 4(3):232–242

Stevens, T., Berk, M., Lopez, R., Chung, Y., Zhang, R., Parsi, M., Bronner, M. and Feldstein, A. (2012). Lipidomic Profiling of Serum and Pancreatic Fluid in Chronic Pancreatitis. Pancreas, 41(4), pp.518-522.

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E., Glass, C., Merrill, A., Murphy, R., Raetz, C., Russell, D. and Subramaniam, S. (2007). LMSD: LIPID MAPS structure database. Nucleic Acids Research, 35(Database), pp.D527-D532.

Suderman, M. and Hallett, M. (2007). Tools for visually exploring biological networks. Bioinformatics, 23(20), pp.2651-2659.

Sugimoto, Y. (1997). Failure of Parturition in Mice Lacking the Prostaglandin F Receptor. Science, 277(5326), pp.681-683.

Sun AR, Friis T, Sekar S, Crawford R, Xiao Y, Prasadam I (2016). Is Synovial Macrophage Activation the Inflammatory Link Between Obesity and Osteoarthritis? Curr Rheumatol Rep 18(9): 57.

Svoboda M., Z. and Květina, J. (2004). Identification and determination of phase II nabumetone metabolites by high-performance liquid chromatography with photodiode array and mass spectrometric detection. Journal of Chromatography A, 1031(1-2), pp.229-236.

Tacconelli, S. and Patrignani, P. (2014). Inside epoxyeicosatrienoic acids and cardiovascular disease. Frontiers in Pharmacology, 5.

Takayama, K., Yuhki, K., Ono, K., Fujino, T., Hara, A., Yamada, T., Kuriyama, S., Karibe, H., Okada, Y., Takahata, O., Taniguchi, T., Iijima, T., Iwasaki, H., Narumiya, S. and Ushikubi, F. (2005). Thromboxane A2 and prostaglandin F2α mediate inflammatory tachycardia. Nature Medicine, 11(5), pp.562-566.

Teeple E, Jay GD, Elsaid KA, Fleming BC (2013). Animal models of osteoarthritis: challenges of model selection and analysis. AAPS J 15(2): 438-446.

Tesarova, P., Kalousova, M., Zima, T., Suchanek, M., Malikova, I., Kvasnicka, J., Duskova, D., Tesar, V., Vachek, J., Krupickova-Kasalova, Z. and Malik, J. (2014). Endotelial activation and flow-mediated vasodilation in young patients with breast cancer. Neoplasma, 60(06), pp.690-697.

Thomson, S., Askari, A. and Bishop-Bailey, D. (2012). Anti-Inflammatory Effects of Epoxyeicosatrienoic Acids. International Journal of Vascular Medicine, 2012, pp.1-7.

Toke, D. A., C. E. Martin (1996). Isolation and Characterization of a Gene Affecting Fatty Acid Elongation in Saccharomyces cerevisiae. Journal of Biological Chemistry 271, 18413-18422

Trebino, C., Stock, J., Gibbons, C., Naiman, B., Wachtmann, T., Umland, J., Pandher, K., Lapointe, J., Saha, S., Roach, M., Carter, D., Thomas, N., Durtschi, B., McNeish, J., Hambor, J., Jakobsson, P., Carty, T., Perez, J. and Audoly, L. (2003). Impaired inflammatory and pain responses in mice lacking an inducible prostaglandin E synthase. Proceedings of the National Academy of Sciences, 100(15), pp.9044-9049.

Tsugawa H. , T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, et al., (2015) MS-DIAL: dataindependent MS/MS deconvolution for comprehensive metabolome analysis, Nat. Methods 12 523–526.

Tsugawa, H., Ikeda, K. and Arita, M. (2017). The importance of bioinformatics for connecting data-driven lipidomics and biological insights. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids, 1862(8), pp.762-765.

Urade, Y. and O. Hayaishi (2011). Prostaglandin D2 and sleep/wake regulation. Sleep Med Rev 15(6): 411-418., T. Tanaka, M. Katsuyama, K.-y. Hasumoto, T. Murata, M. Hirata, F. Ushikubi, M. Negishi, A.

Ichikawa and S. Narumiya (1997). Failure of Parturition in Mice Lacking the Prostaglandin F Receptor. Science 277(5326): 681-683.

Vaya J, Aviram M, Mahmood S, Hayek T, Grenadir E, Hoffman A et al (2001) Selective distribution of oxysterols in atherosclerotic lesions and human plasma lipoproteins. Free Radic Res 34(5):485–497

Venkataraman, K. and Futerman, A.H. (2000) Ceramide as a second messenger: sticky solutions to sticky problems. Trends Cell Biol. 10, 408–412

Villalvilla A, Gomez R, Largo R, Herrero-Beaumont G (2013a). Lipid transport and metabolism in healthy and osteoarthritic cartilage. Int J Mol Sci 14(10): 20793-20808.

Visser HM, Weinans H, Coeleveld K, van Rijen MH, Lafeber FP, Mastbergen, et al (2017) Groove model of tibia-femoral osteoarthritis in the rat. J Orthop Res.; 35(3):496-505.

Wakelam, M., Pettitt, T. and Postle, A. (2007). Lipidomic Analysis of Signaling Pathways. Methods in Enzymology, pp.233-246.

Wakil, S., Stoops, J. and Joshi, V. (1983). Fatty Acid Synthesis and its Regulation. Annual Review of Biochemistry, 52(1), pp.537-579.

Wallner S, Grandl M, Konovalova T, Sigruner A, Kopf T, Peer M et al (2014) Monocyte to macrophage differentiation goes along with modulation of the plasmalogen pattern through transcriptional regulation. PLoS ONE 9(4):e94102

Wang, M., Wang, C. and Han, X. (2016). Selection of internal standards for accurate quantification of complex lipid species in biological extracts by electrospray ionization mass spectrometry-What, how and why?. Mass Spectrometry Reviews, 36(6), pp.693-714.

Welte, M. (2015). Expanding Roles for Lipid Droplets. Current Biology, 25(11), pp. R470-R481.

Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, et al., (2018) HMDB 4.0 — The Human Metabolome Database for 2018. Nucleic Acids Res. Jan 4;46

Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, et al., (2013) HMDB 3.0 — The Human Metabolome Database in 2013. Nucleic Acids Res. Jan 1;41

Wishart DS, Knox C, Guo AC, et al., (2009) HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res. 37

Wishart DS, Tzur D, Knox C, et al., (2007) HMDB: The Human Metabolome Database. Nucleic Acids Res. Jan;35

Wu Z., Li D., Meng J., Wang H. (2010) Introduction to SIMCA-P and Its Application. In: Esposito Vinzi V., Chin W., Henseler J., Wang H. (eds) Handbook of Partial Least Squares. Springer Handbooks of Computational Statistics. Springer, Berlin, Heidelberg, and Dennis E., Journal of Lipid Research 50, S9-S14 (2009)

Xu L, Nwosu LN, Burston JJ, Millns PJ, Sagar DR, Mapp PI, et al. (2016). The anti-NGF antibody muMab 911 both prevents and reverses pain behaviour and subchondral osteoclast numbers in a rat model of osteoarthritis pain. Osteoarthritis Cartilage 24(9): 1587-1595.

Xu, J., A. M. Casas-Ferreira, Y. Ma, A. Sen, M. Kim, P. Proitsi, et al. (2015). Lipidomics comparing DCD and DBD liver allografts uncovers lysophospholipids elevated in recipients undergoing early allograft dysfunction. Sci Rep 5, 17737

Xue, X., Feng, C., Hixson, S., Johnstone, K., Anderson, D., Parrish, C. and Rise, M. (2014). Characterization of the fatty acyl elongase (elovl) gene family, and hepatic elovl and delta-6 fatty acyl desaturase transcript expression and fatty acid responses to diets containing camelina oil in Atlantic cod (Gadus morhua). Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology, 175, pp.9-22.

Yamamoto, K., Miki, Y., Sato, H., Murase, R., Taketomi, Y. and Murakami, M. (2017). Secreted Phospholipase A2 Specificity on Natural Membrane Phospholipids. Methods in Enzymology, pp.101-117.

Yu, Z., Zhai, G., Singmann, P., He, Y., Xu, T., Prehn, C., Römisch-Margl, W., Lattka, E., Gieger, C., Soranzo, N., Heinrich, J., Standl, M., Thiering, E., Mittelstraß, K., Wichmann, H., Peters, A., Suhre, K., Li, Y., Adamski, J., Spector, T., Illig, T. and Wang-Sattler, R. (2012). Human serum metabolic profiles are age dependent. Aging Cell, 11(6), pp.960-967.

Zachary A. King, Andreas Dräger, Ali Ebrahim, Nikolaus Sonnenschein, Nathan E. Lewis, and Bernhard O. Palsson (2015) Escher: A web application for building, sharing, and embedding data-rich visualizations of biological pathways, PLOS Computational Biology 11(8): e1004321.

Zurier RB, Rossetti RG, Jacobson EW, DeMarco DM, Liu NY, Temming JE et al (1996) Gamma-linolenic acid treatment of rheumatoid arthritis. A randomized, placebo-controlled trial. Arthritis Rheum 39(11):1808–1817

Appendix

Appendix chapter 1

A. Lipid Biochemistry

Fatty acids

They are one of the most fundamental lipid structures. They normally bind to other molecules to make more complex structures. In this subsection, the catabolism, transportation and synthesis of fatty acids will be explained. Most fatty acids are ingested in the form of TG, which are the main lipid class for fatty acid storage and held in adipose tissue. The fatty acids can be released from TGs by MOGAT3 and released into the blood stream, attached to albumin, where they are transported to cells for their metabolism. FA catabolism is completely oxidative. It takes place inside the mitochondria or peroxisomes and is catabolized in a series of cycles of β-oxidation.

FATTY ACID TRANSPORT

Fatty acids are normally not presented in their free form, since they could dissolve cell membranes. For this reason, in blood they are always attached to albumin. In the cytosol they are normally attached to a set of transport proteins. But the first step that the fatty acid chains must undergo to be transported is the activation with CoenzymeA by a thio-ester bond (Gimeno et al. 2007). Their mechanism of transport from the cytosol to the mitochondria or peroxisomes depends on their chain length, see Table 1.

Table 1: FA catabolism and transport mechanisms in the size classification column in brackets appear the number of carbons a fatty acid has to contain to consider it in that category.

Size classification	Site of catabolism	Transport mechanism
Short chain (2-4)	Mitochondria	Diffusion
Medium chain (4-12)	Mitochondria	Diffusion

Long chain (12-20)	Mitochondria	Carnitine cycle
Very long chain (>20)	Peroxisomes	Unknown

The carnitine cycle is a reversible mechanism of transport, initially the FA-CoA gets in contact with the outer membrane enzyme CPT1, this replaces the big CoA termination with carnitine, a much smaller terminal group. Then acyl carnitine translocase moves the FA-carnitine to the inside of the mitochondria, where CPT2 reverses the reaction to regenerate FA-CoA, the substrate for β-oxidation (Longo et al. 2016).

FATTY ACID OXIDATION

The oxidation of fatty acids is a series of steps that oxidise carbon β on the FA chain, from that the name β-oxidation. The carbons on the fatty acids are named alpha for the closest to the carbonyl group, then beta, gamma, etc. Each round of the cycle oxidises the beta carbon. This shortens the fatty acid chain by two carbon molecules and produces an equivalent of 1 $FADH_2$, 1NADH and an acetyl-CoA, Figure 1.4.

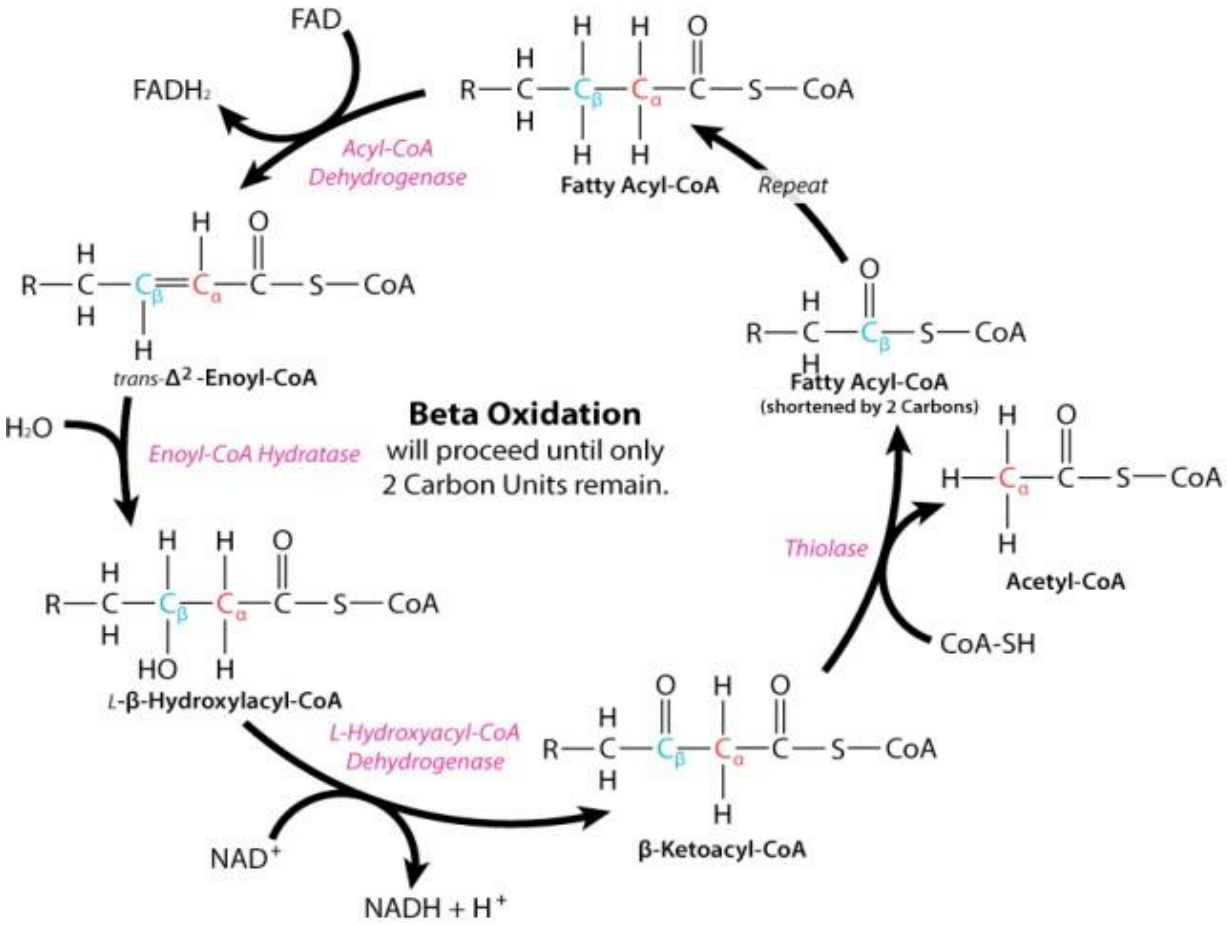


Figure 1.4: Beta oxidation in mitochondria. This figure is taken from (Schulz et al. 1991.)

The first step is the creation of a double bond between β and α, by the Acyl CoA dehydrogenase. This fatty acid now is docks with the Enoyl CoA hydratase and a molecule of water to generate a hydroxy group. The latter is further oxidised to a keto group by hydroxyacyl CoA dehydrogenase. Finally, by the effect of thiolase, an acetyl CoA is released and the new chain with two carbons less is created. This cycle continues until only acetyl or succinyl CoA are left, depending on whether the number of carbons in the chain are even or odd respectively. In the special case of unsaturated fatty acids, the first step does not take place and the energy obtained is fewer.

In peroxisomes, the β-oxidation of long and very long chains is like that of mitochondria. The main difference is in the acyl CoA dehydrogenase which uses O_2, and for this reason the final product of the oxidation is a molecule of H_2O_2. However, this makes the steps less energetically effective.

FATTY ACID BIOSYNTHESIS

Most of the FA required by the human body come from the diet. However, de novo FA synthesis is possible and is done by the repetitive addition of acetyl-CoA (Wakil et al. 1983). The first phase is the generation of a Malonyl-CoA by the enzyme Acetyl-CoA Carboxylase. This process is extremely regulated, since is the starting point for the whole synthesis. It is allosterically inhibited by fatty acids like palmitoyl-CoA and activated by high levels of Acetyl-CoA (Kamel et al. 2017). The second phase is elongation and involves three enzymes. Transacylases respectively convert Acetyl-CoA to Acetyl-ACP and Malonyl-CoA to Malonyl-ACP, where ACP is the abbreviation for Acyl Carrier Protein. The third enzyme is the complex Fatty Acid Synthase, which sequentially adds (and then reduces) Malonyl-ACP after it has covalently bound first to Acetyl-ACP and then to the subsequent growing fatty-acyl-ACP. Synthesis stops when the chain length reaches 16 carbons (i.e. palmitate). The net reaction for this operation is:

$8Ac\text{-}CoA + 7ATP + 14\ NADPH + 14H^{+} \Rightarrow palmitate + 14\ NADP^{+} + 8\ CoA + 6\ H_2O + 7ADP + 7\ CO_2 + 7Pi$

This palmitate can now be a substrate for synthesis of longer chain fatty acid by the ELOVL enzyme family (Xue et al., 2014). Different parts of the body tend to have different preferences for the elongation of fatty acids. Fatty acids that need to be transported or stored are normally converted into their TG form. To do this a unit of glycerol get activated to glycerol-3-P by Glycerol Kinase, which then has fatty-acyl groups added (from their CoA derivatives) by Glycerol-3-phosphate Acyl Transferase (GPAT) to form a mono-glyceride, 1-acylglycerol-3-phosphate O-acyltransferase (AGPAT) or Monoacylglycerol O-acyltransferase 3 (MOGAT3) to form a di-glyceride, and Di-glyceride Acyl Transferase (DGAT) to form a tri-glyceride (Hall et al., 2012).

Fatty acids can also be desaturated, creating double bonds in the carbon chains. These steps are done by the Stearoyl-CoA Desaturase (SCD) enzyme family. However, the human body cannot add saturation in every location, for this reason some essential fatty acids must come from the diet. These essential fatty acids are named ω-3 and ω-6, because double bonds

respectively occur at 3 and 6 carbons from the methyl terminus of the fatty-acyl group. They too can be elongated and further saturated to create compound like the arachidonic acid (Simopoulos, 2016).

Phospholipid biosynthesis

Phospholipids are the main components of the lipid membranes. They are composed of four main domains: a skeleton, a fatty acid, a phosphate and an alcohol. The skeleton is the central part of the phospholipid, where the fatty acids and the phosphate are attached. The two possible skeletons are glycerol and sphingosine, generating the glycerophospholipid and the sphingolipid families respectively. In the phosphoglycerates, glycerol is esterified to fatty acids on carbon 1 or carbon 2 or both, while phosphate is attached to the C3 position. One of the hydroxy groups of the phosphate can be esterified with alcohols, such as choline, ethanolamine, serine or inositol to produce more complex derivatives.

The phosphoglycerates are normally being characterized for their structural function on cell membranes. However, they also have functions in other cellular processes such as signal induction and transport. Regarding signalling, they provide the precursors for prostaglandins and other leukotrienes (Ecker et al. 2014). It is their specific distribution and catabolism that enables them carry out the biological-response processes listed above (Farooquiet al. 2000). Their role as a storage centre for secondary messengers in the membrane is also a contributing factor to their ability to act as transporters (Farooquiet al. 2000). They also influence protein function. For example, they are important constituents of lipoproteins (soluble proteins that transport fat in the blood), and hence affect their metabolism and function (Hermansson et al. 2011).

Sphingolipids (SLs) consist of three main structural elements. The basic building block is the sphingoid long-chain base (lcb), normally a sphingosine, a sphinganine (dihydrosphingosine) or a 4-hydroxysphinganine (phytosphingosine). A fatty acid is attached to carbon-2 (C-2) of the lcb via an amide bond, yielding ceramide, and attachment of hydrophilic head groups to the OH-group at C-1 yields complex SLs. The head group can be a sugar, in the case of glycosphingolipids (GSLs), or phosphorylcholine in the case of sphingomyelin (SM). At least five different lcbs are known in mammalian cells, 20 species of fatty acid (varying in chain

length, degree of saturation, and degree of hydroxylation) can be attached to the lcb, and hundreds of different carbohydrate structures have been described in GSLs (Futerman and Riezman, 2005). The possible relevance of this complexity is discussed elsewhere (Futerman et al. 2004), but a central question in SL cell biology is how the synthesis of these different SL species are regulated. Their main involvement is in lipid membranes, however their role in signalling pathway has been uncovered recently, for example in insulin activation and apoptosis (Dominiczak et al, 2011).

Unlike synthesis of the other major membrane lipids, sterols and glycerolipids, which takes place mainly in the ER, the initial steps of SL synthesis take place in the ER but later steps take place for the most part in the Golgi apparatus (Futerman and Riezman, 2005). This implies significant trans-bilayer and inter-bilayer movement of SLs during the biosynthetic process, and, as these lipids do not appear, overall, to spontaneously undergo such movements (Venkataraman et al. 2000), facilitated mechanisms are probably involved in their transport.

Eicosanoid pathway

This pathway is the one responsible for the generation of prostaglandins and thromboxanes. It starts from arachidonic acid, obtained from diet or by FA metabolism. The eicosanoids are biomolecules of short half life time in plasma (0.5 – 5 min) normally catabolized in the lungs. The metabolising pathway has three main 3 branches, linked to the families of enzymes (Rosenfeld and Loose, 2014):

1. Cytochrome P450 pathway: The P450 pathway has two main routes, the 20-hydroxyeicosatetraenoic acid (20-HETE) and the regio- and stereo-specific epoxyeicosatrienoic acids (EETs) with their corresponding dihydro-epoxyeicosatriencoic acids (DHETs) The latter are mainly found in the kidneys and relate to blood-pressure regulation. EETs have been found to influence a variety of organ functions, the most important being vascular effects (Quilley et al. 2000, Fleming et al. 1999). They also affect cerebral (Bhardwaj et al. 2000), renal (Pomposiello et al. 2001), pulmonary (Yaghi et al. 2001) and peripheral (Halcox et al. 2001) functions. Additionally, anti-inflammatory properties for these species have been reported (Node et al. 1999). The EETs can be further metabolized by either

soluble epoxide hydrolases (sEH) to form dihydroxyeicosatrienoic acids (DHETs), further oxidation by P450 or esterification to become glycero-phospholipids (GPs).

The other family of AA derivatives is the 20-HETE, containing hydroxylation of terminal and sub terminal positions. The enzymes CYP4A and CYP4F have been characterized as the main isoforms involved in this pathway. These species also play an important role in the vascular activity, as a vasoconstrictor, throughout the body (Moreno et al. 2001), but other properties have also been studied, such as its renal effect (Gu et al. 2001).

2. Cyclooxygenase pathway: This pathway is closely related to inflammation and the target for most anti-inflammatory drugs. It is the start point for the generation of prostaglandins, thromboxanes and other derivatives. There are two isoforms of the COX enzyme, COX1 and COX2. They both contribute on the generation of the homeostatic levels and inflammation but in different manner. Some of the most common known products of this pathway are PGI_2, PGE_2, TXA_2, and LTC_4.
3. Lipoxygenase pathway: The formation of hydroperoxide lipids is mainly regulated by enzymes of the LOX family. LOXs constitute a large gene family of nonheme iron containing fatty acids dioxygenases. They catalyse regio- and stereo-specific deoxygenation of polyunsaturated fatty acids (PUFAs) containing a (1Z,14Z)-pentadiene system (Feussner et al, 2000). They are versatile catalysts and can act as dioxygenases (HambergM et al, 1967), hydroperoxidases (Kramell et al, 2000), or leukotriene synthases (Shimizu et al, 1984). For mammalian LOXs, a space-related hypothesis of site specificity has been established (Gillmor et al. 1998, Sloane 1995), namely that the depth of the enzyme's substrate-binding pocket determines the site of hydrogen removal and the positional specificity of oxygen insertion. Most LOXs prefer free fatty acid substrate than the esterified one although some exceptions have been found in plant species (Brash et al. 1987, Matsui et al. 1998).

   There are six known human LOX genes: ALOX5, ALOX15, ALOX15B, ALOX12, ALOX12B and ALOXE3. One of the most central is ALOX5, which generates 5-HpETE, 5-HETE and LTA4, that can be further converted to LTB4 and LTC4 combining with the COX pathway. It can act in series with other lipoxygenases to generate Resolvins. ALOX15 has also a lot of important products, from the 13-HODE to 15-HETE and 12-HpETE. This enzyme usually shares substrates with the ALOX12, which generates 12-HETE and

its precursor 12-HpETE. ALOX15B is a more specific version of ALOX15 and is not able to generate 12-HpETE. ALOX12B generates 12R-HETE, and its mutations are associated with skin disease. Two lipoxygenases can act in series to create di or tri-hydroxy products.

A more in-depth explanation about some of this species can be found in chapter 4 where their relation to inflammation and pain is described.

Cholesterol and derivatives pathway

Cholesterol is an essential component for the structure and the function of lipid membranes. It is also a precursor of many important compound like bile acids, hormones and vitamin D. Cholesterol increases the fluidity of the membranes, and can be found between the layers of phospholipids. The concentration of the free form is rather low.

Cholesterol biosynthesis starts with Ac-CoA. Two units of Ac-CoA combine with the enzymes 3-Hydroxy-3-Methylglutaryl-CoA Synthase (HMGCS) and 3-Hydroxy-3-Methylglutaryl-CoA Reductase (HMGCR) converting them to Mevalonate. After two phosphorylation steps, catalysed by Mevalonate Kinase (MVK) and Phosphomevalonate Kinase (PMVK), the Mevalonate is converted to Mevalonate-5PP. This molecule is converted by Mevalonate Diphosphate Decarboxylase (MVD) to Isoprenyl-PP and sequentially by Farnesyl Diphosphate Synthase (FDPS) to Farnesyl-PP. These last two molecules can be converted to many other biologically significant species.

From Farnesyl-PP, the enzyme Farnesyl-Diphosphate Farnesyltransferase 1 (FDFT1) produces Presqualene-PP and then Squalene. This latter is further metabolized by Squalene Epoxidase (SQLE) and Lanosterol Synthase (LSS) to generate Lanosterol. The next steps to convert it to Zymosterol are still unclear, but the enzymes CYP51A1, TM7SF2, MSMO1, and HSD17B7 are known to be involved. Zymosterol now has many paths to get to Cholesterol, mainly using DHCR24 to hydrogenate the terminal double bond and then enzymes like EBP and SC5D arrange the configuration to cholesterol.

The conversion to mevalonate, the first step in cholesterol synthesis, is the most regulated step. Mevalonate is exclusive to this pathway and is not involved in any other (Sakai et al, 2001). The biological cholesterol synthesis represents a 50% of the total cholesterol intake of a regular diet (Maxfield et al, 2010).

Bile acids are one of the main derivatives from cholesterol, they are structures of 24 carbons; 3 carbons are lost from the cholesterol main ring. They are synthesized on the liver and the most important primary bile acids are taurocholic acid, glycocholic acid, taurochenodeoxycholic acid and glycochenodeoxycholic acid. They are the main components of the bile and help in the digestion of oils and dietary fats. In the intestine, these acids can be further processed to secondary bile acids, which act like detergents and help in lipid absorption.

The other main products from cholesterol are the steroid hormones. There are three main families:

- Corticosteroids: these have 21 carbon atoms and they are generated in the cortex of vertebrates. They are involved in a wide range of physiological processes, like stress, immune response, regulation of inflammation, and carbohydrate metabolism (Nurssey et al. 2001)
- Androgens: Hormones of 19 carbons, they are the regulators in male vertebrates. They are responsible for the development of male sex organs.
- Oestrogens: Hormones of 18 carbons, they are the main regulators of the female reproductive system.

All steroid hormones act by binding to a ligand in the cell nucleus. They are slowly removed in urine. The last important cholesterol derivative is vitamin D3, which oversees the regulation of calcium for the human body.

Signalling pathways

The increasing appreciation of the central roles played by lipid signalling pathways has dispelled the misconception that lipids are inert structural components that are involved solely in keeping a cell intact (Wakelam et al. 2007). Advances in our understanding of cell-

signalling pathways have identified lipids that act to regulate the functions of several proteins either by controlling enzyme activity directly, or by localizing proteins to intracellular compartments where they perform a specialized role. These lipid-binding domains (e.g., PH, PX, FYVE) have been found in many proteins, and considerable detail is recorded of the structural basis of lipid-protein interaction. Additional lipid-binding domains exist, which remain less well characterized (e.g., those that bind phosphatidic acid [PA] or ceramide). However, the important regulatory roles that these lipids play and the pathways involving these messengers are increasingly appreciated.

B. Post data analysis lipidomic pipeline

Experimental design

Experimental design is the planning stage needed to describe or explain the variation of information under controlled conditions to confirm or deny the hypothesis of the study. This is normally the step of the selection of the sample sources, which tissues and why, sample size, detection method and data analysis. This is the most critical part of the experiment, so these decisions should be planned in advance. Otherwise, hundreds of hours can be wasted with no credible results.

The typical lipidomics samples are blood, urine, sweat and synovial fluid. Some of them are easy to obtain but some are more invasive. For this reason, if we are looking for potential biomarkers, the blood is a convenient candidate. Some lipid species in the urine samples are normally affected greatly by the diet and hard to control in medical research (Men et al. 2017). All the sources have their pros and cons. The other important decision is the methodology, whether a targeted or untargeted approach should be used. For biomarker discovery, the approach should be untargeted as it is unknown what molecular species has the potential to be a biomarker. On the other hand, when a potential biomarker may be part of a known pathway, for example pain and inflammation, the targeted approach leads much better results (Halade et al. 2018).

Sample collection

This step is normally done in hospital or prepared facilities, to secure hygiene and non-contamination. The samples are generally labelled anonymously and randomized, so when

they get to the lab only the medical team know the relation between patients and sample number.

Lipid extraction

Lipids are separated from nucleic acids and proteins using organic solvents. There are many techniques to perform the extraction, among the most common are the Folch method or the Bligh and Dyer procedure(Bligh et al. 1959, Folch et al. 1957). Both methods partition lipids into an organic phase by using different ratios of chloroform/methanol solvents: Folch method (chloroform: methanol = 2:1) and Bligh and Dyer method (chloroform: methanol = 1:2). Depending on the hydrophobicity and the abundance of lipids, several alternative solvents such as hexane, isopropanol, ethyl acetate, and tert-butyl methyl ether can be used, with similar extraction efficiency for major lipid classes but with varying effectiveness for the recovery of low abundance lipid classes (Reis et al. 2013). To avoid the loss of more polar lipid species though partial or full partitioning to the aqueous phase during biphasic extraction methods, monophasic extraction methods have been developed, which include variants of the above methanol/chloroform/ water mixtures or methanol, acetonitrile, and acetone (Lydic et al. 2014, Cai et al. 2016).

After extraction, the lipids are often separated by various chromatographic methods such as thin layer chromatography (TLC), gas chromatography (GC), and high-performance liquid chromatography (HPLC), or lipids can be directly infused to a mass spectrometer for detection without separation, shotgun methods (Fauland et al. 2011). Since the separation by any chromatographic method normally has many steps and has a long time-requirement, shotgun methods have been used on many occasions (Schwudke et al. 2011). However, this methodology has a much narrower range of detection, <1000 compared to the 10.000-100.000 of other techniques and cannot distinguish isobaric lipids. Isobaric lipids are molecules with the same nominal mass (sum of the masses of the most abundant isotopes) but differing exact masses. Another tool of lipidomics detection is NMR

Appendix chapter 2

A. VIP score vs other scores

The plot bellows show the correlation between VIP and other scores used to compare data over two groups. The correlation between VIP and t value is excellent, between contribution and VIP is not linear, but still correlates. Finally, with foldchange there is no correlation at all.

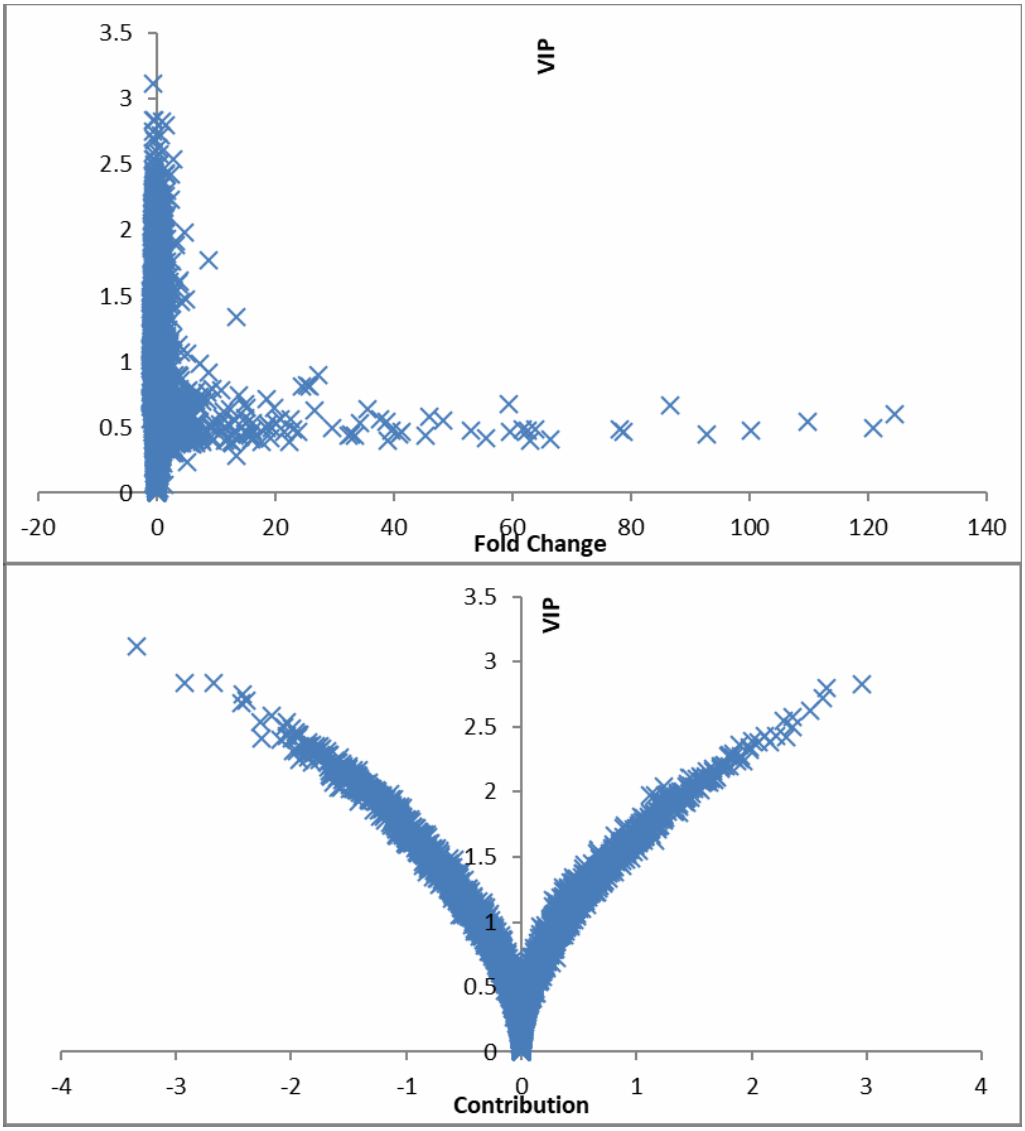


Figure 2.17: The bullet plots above present the correlation of VIP and other parameters used to sort lipids on multivariance analysis. It can be seen that fold change does not correlate with VIP and Contribution has a non linear correlation.

B. Random Scored data old script

This test is the responsible for the detection of the error in the scoring script subsection 2.4. It was performed with the old scoring script. Like in the 2.5.1 experiment, 20 randomly selected metabolites are introduced into the methodology. With the difference that they are now randomly scored, either 1 or -1, also random. This action was done for 12 repetitions.

In this test, the analysis becomes more interesting as it is possible to observe the changes of moving the cut off, which decides when an enzyme becomes significant. To analyse the results the genes that had an absolute score higher than 0.2 were added up on a list. This has also been done for the ones with an absolute score of 0.25.

Table 2.110: List of the genes with an absolute score higher than 0.2 for the 12 experiments.

MOGAT3	9	75%
CEPT1	2	17%
PLD1	2	17%
Agpat4/Agpat6	2	17%
UG1	1	8%
PISD	1	8%
PTDSS2	1	8%

Table 2.121: List of the genes with an absolute score higher than 0.25 for the 12 experiments.

MOGAT3	8	67%
CEPT1	1	8%
Agpat4/Agpat6	1	8%

The scored data shows more interesting results than the unscored one. First, the gene lists generated are much shorter and with lower frequencies. This is what we expect from the program, random data should not give any significance on any gene.

Secondly this experiment has helped us set a threshold to the score of genes. It can be appreciated that higher than 0.25 is a good option to set when we want to see what genes are significant.

Finally, there is the special case of MOGAT3, this gene appears way too much. While is true that there is a huge pool of TG that are closely related to this gene we expected to not be appearing significant in all experiments. This last point helped us by highlighting a problem in the scored method solved as explained in section 2.4.

C. Summary of ROC curve analysis

Receiver operating characteristic curve (ROC), is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is modified. The curve is the result of plotting true positive rate (TPR) against false positive rate (FPR) when the threshold is being modified. This technique is going to be used to compare some models of global lipidomic data.

In lipidomic analysis experiments, the final step is usually the creation of a list of the significantly changing lipids. However, this method might not be good enough since it has been proven that using these individual lipids as biomarkers is not successful in all cases. A test will be trailed to see whether it is possible to generate a better model by using our pathway tool results.

To perform this analysis, we have used the PCOS dataset (40 controls and 40 patients). First, an OPLS-DA model using SIMCA was done to the PCOS dataset, obtaining separation between the controls and the patients (Figure 2.17) and a list of the lipids that contributed to their separation.

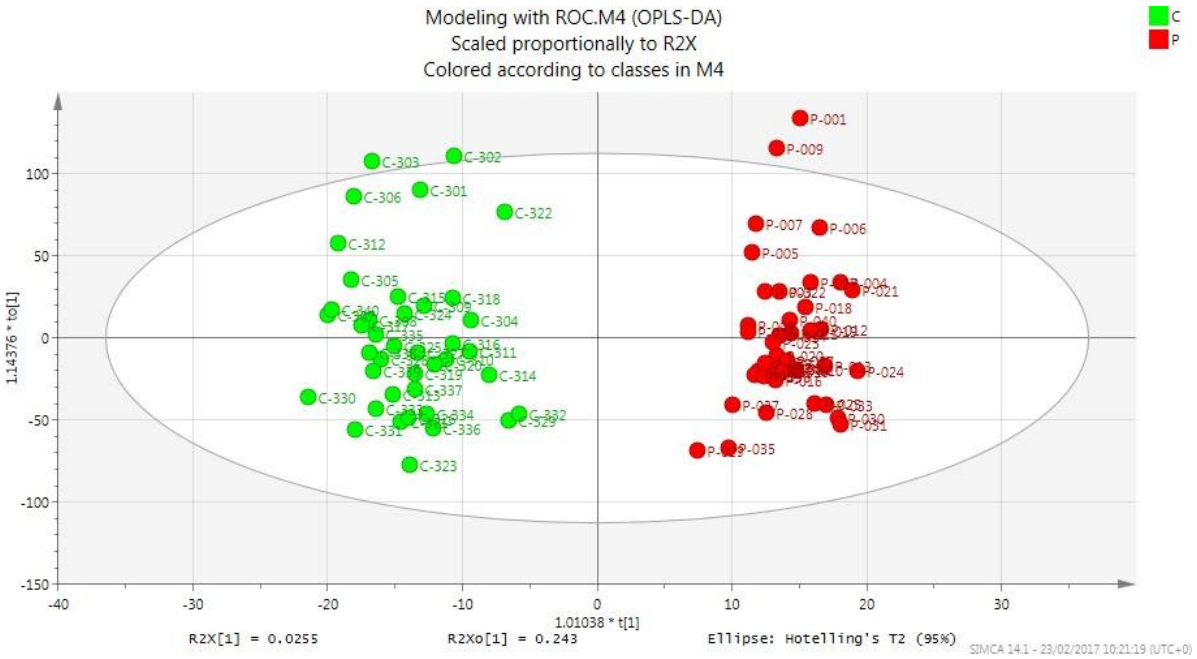


Figure 2.35: ROC model of PCOS dataset, control samples are in green colour and patient samples are in red.

From the analysis, the best three VIP score lipids are taken and used to perform a ROC curve along with a random lipid to observe the effect that choosing the lipid wisely can have.

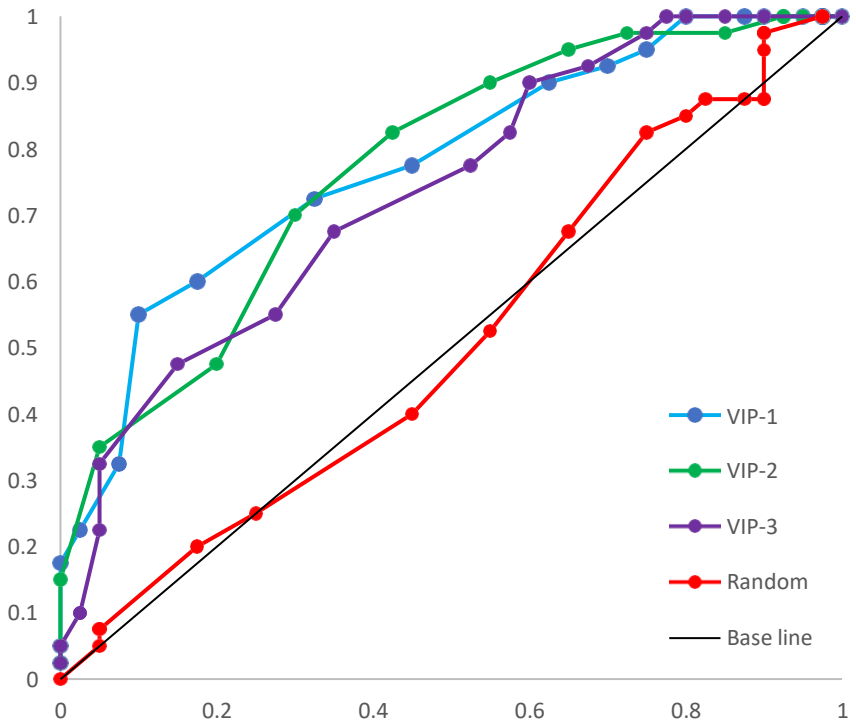


Figure 2.36: ROC curve for the top 3 VIP score lipids and a random lipid. The VIP lipids are generated from the OPLS-DA analysis of the PCOS data.

It is possible to observe that the curves from the VIP lipids have better classification potential than the random sample, although they are still far from perfect. The objective is to compare these results with a model generated from the pathway analysis. For that experiment, as discussed in the paper, the pattern detected was that PCOS patients present higher concentration of PC and lower concentrations of DG when compared to controls. Since the raw data presents more than 8000 peaks corresponding to possible lipids, it is impossible to manually check all peaks.

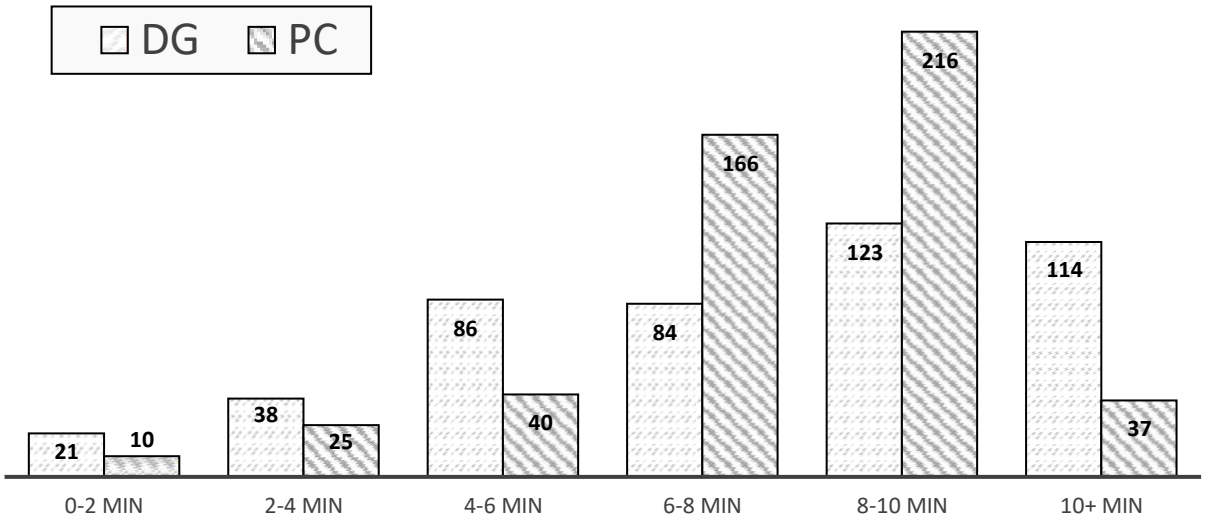


Figure 2.19: Histogram of the possible identified DG and PC separated by their retention time. The number in the bars shows the number of identifications in that time bracket

To identify the peaks that could belong to PC and DG we have used the python program section 2.5 with a slight modification. We identified all the peaks whose masses could be either PC or DG with an error of 6 ppm mass difference. Still the list of the possible PC and DG was too big and with many possible misidentifications, so to reduce this number we plotted separately PC and DG in a histogram divided by their retention time.

It is known from previous experiments that in this chromatographic methodology, DG are found between 4 and 8 minutes, and PC are found in the bracket of 6 to 10 minutes. This information has been used to differentiate the identifications with the highest probability of being accurate. The refined data still has 170 DG and 382 PC. The peaks containing more than 20% of undetected values (appears as a 0) and peaks with a CV bigger than 0.7 were removed. This leaves us with 46 DG and 274 PC, possibly still with some misidentifications.

The next step is the generation of a model, imposing that high PC contribute to stronger to PCOS and low DG to the opposite effect (presented in Figure X as Clean comparison). A second comparison is also done, a more ideal and perhaps 'cheated' comparison, where some parameters were adjusted to improve the model. The number of 0 allowed was greatly reduced and the lipid data was used as univariate.

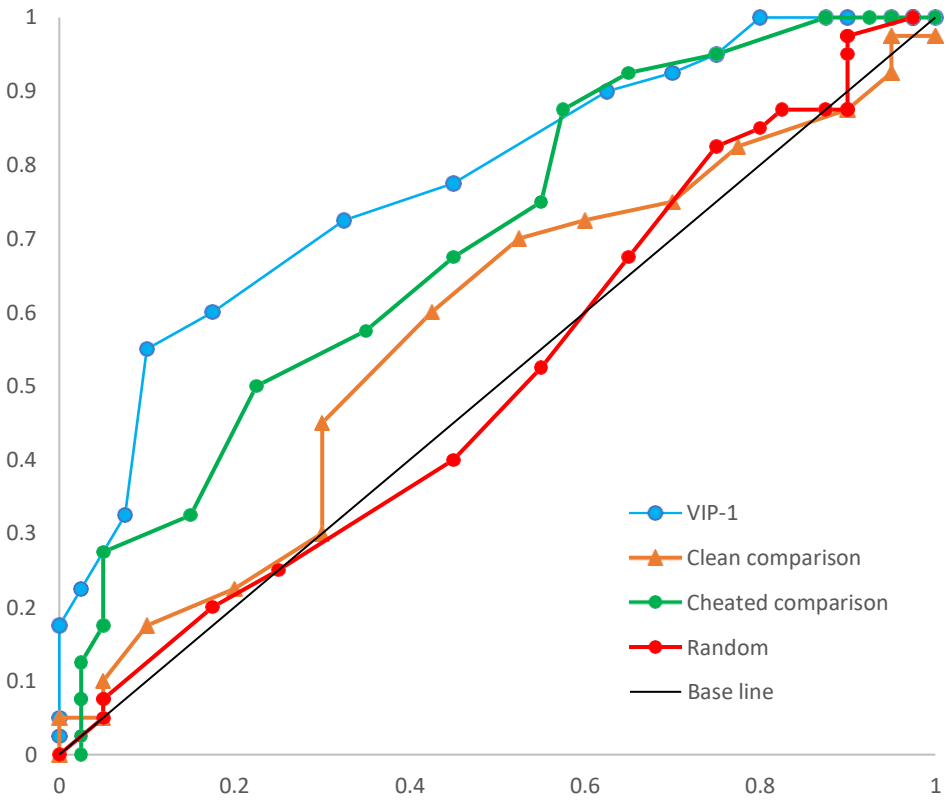


Figure 2.37: ROC curve comparing the models and a random sample. In blue, the lipid with highest VIP score from the OPLS-DA of the PCOS original data is presented. The green and pink lines are the ones generated by the model created using the network information. Finally, the red is the random lipid selected from the data.

Conclusions

While the experiment does not seem to present any significant improvement over the VIP scores traditional method there are a couple of points worth discussing. First of all, the identification was not perfect, as non PC or DG may have gone into the analysis. With perfect identification the results might be better. Second the procedure to obtain the chromatographic data required for the new model will probably be easier than the procedure to obtain the data required for the VIP model.

Appendix chapter 3

A. List of all the oxylipins

Table 3.3: List of oxylipins in the pathway with their ChEBI ID and the full IUPAC naming

Name on the network	ChEBI ID	Full Name (IUPAC Name)
Oleic acid	16196	(9Z)-octadec-9-enoic acid
Linoleic acid	17351	(9Z,12Z)-octadeca-9,12-dienoic acid
Arachidonic acid	15843	(5Z,8Z,11Z,14Z)-icosa-5,8,11,14-tetraenoic acid
EPA	28364	(5Z,8Z,11Z,14Z,17Z)-icosa-5,8,11,14,17-pentaenoic acid
DHA	28125	(4Z,7Z,10Z,13Z,16Z,19Z)-docosa-4,7,10,13,16,19-hexaenoic acid
Oleamide	116314	(9Z)-octadec-9-enamide
Oleoyl CoA (18:1)	15534	3'-phosphoadenosine 5'-{3-[(3R)-3-hydroxy-2,2-dimethyl-4-({3-[(2-{[(9Z)-octadec-9-enoyl]sulfanyl}ethyl)amino]-3-oxopropyl}amino)-4-oxobutyl] dihydrogen diphosphate}
9(S)-HODE	34496	(9S,10E,12Z)-9-hydroxyoctadeca-10,12-dienoic acid
9-OxoODE	72842	(10E,12Z)-9-oxooctadeca-10,12-dienoic acid
13-OxoODE	68947	(9Z,11E)-13-oxooctadeca-9,11-dienoic acid
13(S)-HODE	34154	(9Z,11E,13S)-13-hydroxyoctadeca-9,11-dienoic acid
9(S)-HPODE	34498	(9S,10E,12Z)-9-hydroperoxyoctadeca-10,12-dienoic acid
13(S)-HPODE	15655	(9Z,11E,13S)-13-hydroperoxyoctadeca-9,11-dienoic acid

PGE2	15551	(5Z,13E,15S)-11α,15-dihydroxy-9-oxoprosta-5,13-dien-1-oic acid
LTA4	15651	(5S,6S,7E,9E,11Z,14Z)-5,6-epoxyicosa-7,9,11,14-tetraenoic acid
5-HPETE	91268	(6E,8Z,11Z,14Z)-5-hydroperoxyicosa-6,8,11,14-tetraenoic acid
5-HETE	60943	(6E,8Z,11Z,14Z)-5-hydroxyicosa-6,8,11,14-tetraenoic acid
LTC4	16978	L-γ-glutamyl-S-[(1R,2E,4E,6Z,9Z)-1-[(1S)-4-carboxy-1-hydroxybutyl]pentadeca-2,4,6,9-tetraen-1-yl]-L-cysteinylglycine
PGD2	15555	(5Z,13E,15S)-9α,15-dihydroxy-11-oxoprosta-5,13-dien-1-oic acid
8(S)-HPETE	34487	(5Z,8S,9E,11Z,14Z)-8-hydroperoxyicosa-5,9,11,14-tetraenoic acid
12(S)-HPETE	15626	(5Z,8Z,10E,12S,14Z)-12-hydroperoxyicosa-5,8,10,14-tetraenoic acid
9(S)-HPETE	34497	(5Z,7E,9S,11Z,14Z)-9-hydroperoxyicosa-5,7,11,14-tetraenoic acid
11(R)-HPETE	34127	(5Z,8Z,11R,12E,14Z)-11-hydroperoxyicosa-5,8,12,14-tetraenoic acid
20-HETE	34306	(5Z,8Z,11Z,14Z)-20-hydroxyicosa-5,8,11,14-tetraenoic acid
15(S)-HPETE	15628	(5Z,8Z,11Z,13E,15S)-15-hydroperoxyicosa-5,8,11,13-tetraenoic acid
15(S)-HETE	15558	(5Z,8Z,11Z,13E,15S)-15-hydroxyicosa-5,8,11,13-tetraenoic acid
PGG2	27647	(5Z,9S,11R,13E,15S)-15-hydroperoxy-9,11-epidioxyprosta-5,13-dien-1-oic acid
PGH2	15554	(5Z,13E,15S)-9α,11α-epidioxy-15-hydroxyprosta-5,13-dienoic acid
PGF2alfa	15553	(5Z,13E,15S)-9α,11α,15-trihydroxyprosta-5,13-dien-1-oic acid
11-epi-PGF2alfa	27595	(Z)-7-[(1R,2R,3S,5S)-3,5-dihydroxy-2-[(E,3S)-3-hydroxyoct-1-enyl]cyclopentyl]hept-5-enoic acid
TXA2	15627	(5Z,13E,15S)-9α,11α-epoxy-15-hydroxythromboxa-5,13-dien-1-oic acid
20-OH LB4	28700	(5S,6R,7E,9E,11Z,14Z)-6-(cystein-S-yl)-5,20-dihydroxyicosa-7,9,11,14-tetraenoic acid

LTD4	28666	S-{(1R,2E,4E,6Z,9Z)-1-[(1S)-4-carboxy-1-hydroxybutyl]pentadeca-2,4,6,9-tetraen-1-yl}-L-cysteinylglycine
LTE4	15650	(5S,6R,7E,9E,11Z,14Z)-6-(cystein-S-yl)-5-hydroxyicosa-7,9,11,14-tetraenoic acid
LTB4	15647	(5S,6Z,8E,10E,12R,14Z)-5,12-dihydroxyicosa-6,8,10,14-tetraenoic acid
14,15-EET	34157	(5Z,8Z,11Z)-13-(3-pentyloxiran-2-yl)trideca-5,8,11-trienoic acid
14,15-DHET	63966	(5Z,8Z,11Z)-14,15-dihydroxyicosa-5,8,11-trienoic acid
11,12-EET	34130	(5Z,8Z)-10-{3-[(2Z)-oct-2-en-1-yl]oxiran-2-yl}deca-5,8-dienoic acid
11,12-DHET	63969	(5Z,8Z,14Z)-11,12-dihydroxyicosa-5,8,14-trienoic acid
8,9-EET	34490	(5Z)-7-{3-[(2Z,5Z)-undeca-2,5-dien-1-yl]oxiran-2-yl}hept-5-enoic acid
8,9-DHET	63970	(5Z,11Z,14Z)-8,9-dihydroxyicosa-5,11,14-trienoic acid
5,6-EET	34450	4-{3-[(2Z,5Z,8Z)-tetradeca-2,5,8-trien-1-yl]oxiran-2-yl}butanoic acid
5,6-DHET	63974	(8Z,11Z,14Z)-5,6-dihydroxyicosa-8,11,14-trienoic acid
8(S)-HETE	34486	(5Z,8S,9E,11Z,14Z)-8-hydroxyicosa-5,9,11,14-tetraenoic acid
12(S)-HETE	34146	(5Z,8Z,10E,12S,14Z)-12-hydroxyicosa-5,8,10,14-tetraenoic acid
9(S)-HETE	34495	(5Z,7E,9S,11Z,14Z)-9-hydroxyicosa-5,7,11,14-tetraenoic acid
11(R)-HETE	34126	(5Z,8Z,11R,12E,14Z)-11-hydroxyicosa-5,8,12,14-tetraenoic acid
TXB2	28728	(5Z,13E,15S)-9α,11,15-trihydroxythromboxa-5,13-dien-1-oic acid
Resolvine E1	81559	(5S,6Z,8E,10E,12R,14Z,16E,18R)-5,12,18-trihydroxyicosa-6,8,10,14,16-pentaenoic acid
Resolvine D2	81565	Not found
Resolvine D1	81564	(4Z,7S,9E,11E,13Z,15E,17S,19Z)-7,8,17-trihydroxydocosa-4,9,11,13,15,19-hexaenoic acid

Resolvine E2	15551	(5Z,13E,15S)-11α,15-dihydroxy-9-oxoprosta-5,13-dien-1-oic acid
18HpETE	81562	5S-Hydroperoxy-18R-hydroxy-6E,8Z,11Z,14Z,16E-eicosapentaenoic acid
5,6-epoxy,18R-HEPE	81561	5,6-Epoxy,18R-hydroxy-7E,9E,11Z,14Z,16E-eicosapentaenoic acid
16,17-epoxy,17R-HDHE	-	Not found
17R-HPDHE	-	Not found
4,5-epoxy,17R-HDHE	-	Not found
7,8-epoxy,17R-HDHE	-	Not found
13S,14S-epoxy-maresin	131958	(4Z,7Z,9E,11E)-12-{(2S,3S)-3-[(2Z,5Z)-octa-2,5-dien-1-yl]oxiran-2-yl}dodeca-4,7,9,11-tetraenoate
Protectin D1		(4Z,7Z,10R,11E,13E,15Z,17S,19Z)-10,17-Dihydroxy-4,7,11,13,15,19-docosahexaenoic acid
Resolvine D4		(4S,6E,8E,10E,13E,15Z,17S,19Z)-4,5,17-trihydroxydocosa-6,8,10,13,15,19-hexaenoic acid
Resolvine D3		4S,11R,17S-trihydroxy-5Z,7E,9E,13Z,15E,19Z-docosahexaenoic acid
Maresin 1		7R,14S-dihydroxy-4Z,8E,10E,12Z,16Z,19Z-docosahexaenoic acid

B. Excel for inflamation score calculation

lipid	score	INF	INF score	signif
(5Z,8Z,11Z,14Z,17Z)-Eicosapentaenoic acid	0			N
10(S)-HPODE	-1.31059			N
10S,17S-dihidroxydocosatriene	0			N
11(R)-HPETE	0.032092			N
11(S)-HPODE	-0.54633			N
11,12,15-THETA	0	p		N
11,12-DHET	-0.65338			N
11,12-EET	-0.87439			N
11,14,15-THETA	-0.6312	p		N
11-dh-TXB2	0			N
11-epi-PGF2alfa	-1.51779			N
11H-14,15-EETA	-2	p	-1.25	Y
11-HETE	1.3	p		N
12(13)-EpOME	1.1			N
12(S)-HETE	1.298359	p		N
12(S)-HPETE	0.4			N
12,13-DHOME	-0.66958			N
12-HETE	1.545331	p		N
12-HPETE	-0.83574			N
13(S)-HPODE	1.820067			Y
13-HODE	-1.91168	a	1.194801	Y
13-OxoODE	0.836224	a		N
14,15-DHET	0.146833			N
14,15-EET	-0.1115			N
15H-11,12-EETA	-0.39914	p		N
15-HETE	1.470149	p		N
15-HPETE	0			N
16S,17S-epoxydocosatriene	-1.23327			N
17R-hydroperoxy DHA	0			N
17S-hydroperoxy DHA	0.418429			N
18R-hydroperoxy-EPE	-1.64883			Y
20-HETE	-0.44232			N
20-OH LB4	0.199451			N
20-OH-LTE4	1.379085			N

INPUT

Number of identified metabolites	50
Recommended significance level	1.676551
Significance level	1.6

RESULTS

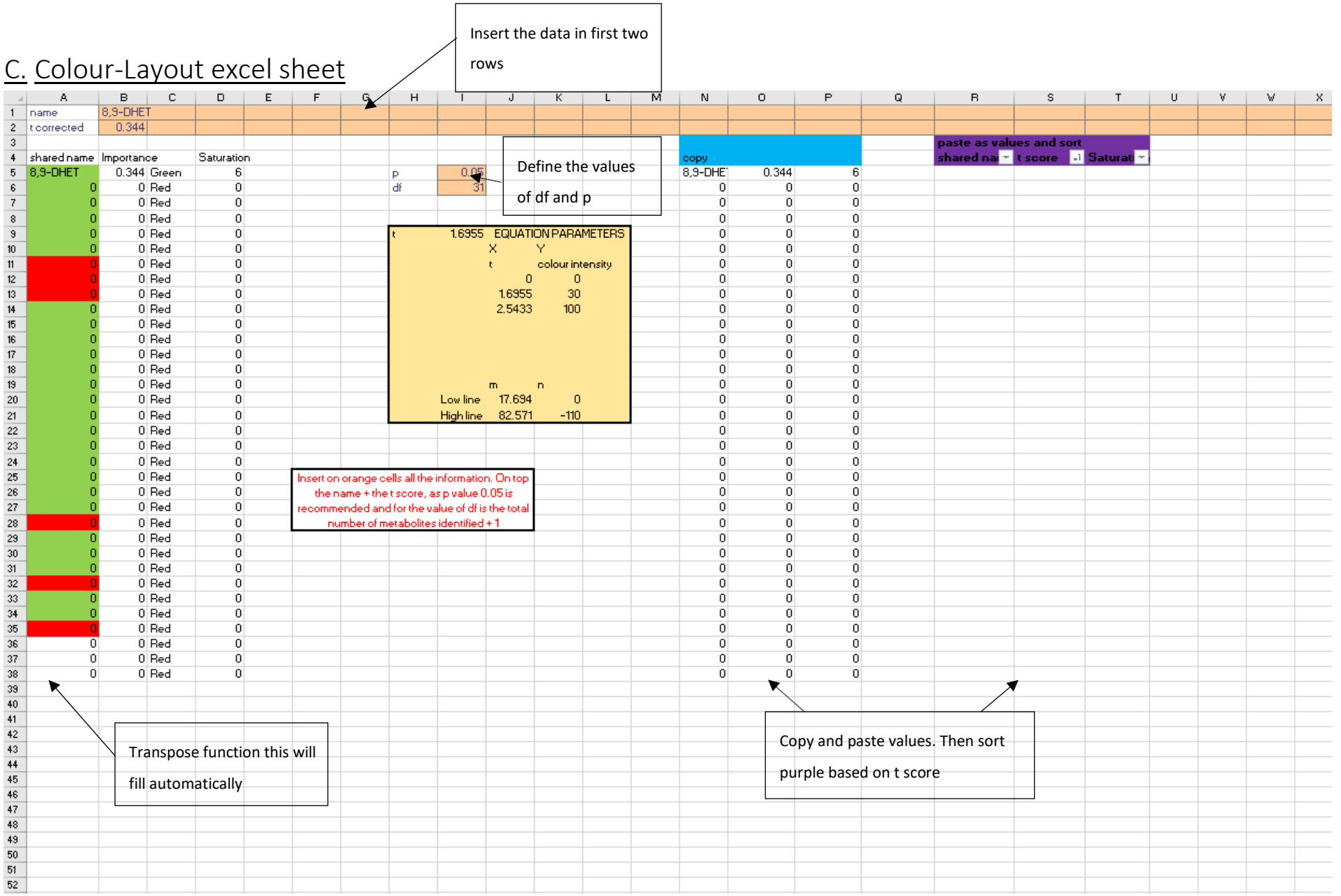
Inflamation score	-1.44427
Number of siginifican species	9

Set the number of metabolites identified in your experiment. Now introduce the significance level that wants to be used. Is recommended to go a bit lower than the one calculated by the formula atv the top. You can check the number of significant species that are considered to decide whether you want more or not.

To read the **INFLAMATION SCORE** the value that apears is an aproximation of the number of the species that are pro inflamation agains anti inflamation. The level of significance of species is taken in consideration so high concentration of proinflamatory species afect more than lower levels of them. Positive values indicate PROINFLAMATION and negatives values indicate ANTIINFLAMATION behavieour of the treatment agains controls.

Excel can be found in github with all other documents https://github.com/Ferran-Casbas/Lipid-Network

C. Colour-Layout excel sheet



Appendix chapter 4

A. Adduct summation

The conversion of chromatogram to mass and peak area to tabulated data is the essential first step and is routinely carried out by commercial software, such as Progenesis. Any software will compare the mass difference of adjacent peaks in retention time to see if they are adducting and whether they should be put together. In multiple analyses, it has been possible to spot adducts not properly grouped, making the later identification task longer and more tedious.

For these reasons, the program, Adduct-cleaner.py, has been developed to provide another layer for summing (with the unmodified molecule) and also removing adduct peaks. This, if done successfully, will help the identification process, meaning less species to identify, and it will assist in the generation of models, as it will have less overfitting.

The software combines peaks found in the same retention time bracket if they have the *m/z* difference of an adduct. It uses the Progenesis data output to begin, as all the peak area integration is needed. Figure 4.12 gives an overview of how the program works.

Although the script is ~200 lines of compact code, its algorithm is particularly complicated, containing many functions and triple loops. The basic idea is that it will read a peak, look for all peaks with the same retention time, compare the masses (depending on user preference, usually within a 5-8 ppm range), and if there is a match, it will add both and remove data from the peak of the adduct. The code can be found on Appendix 4.B.a. The program takes around 4 hours to run on datasets of ~7000 peaks.

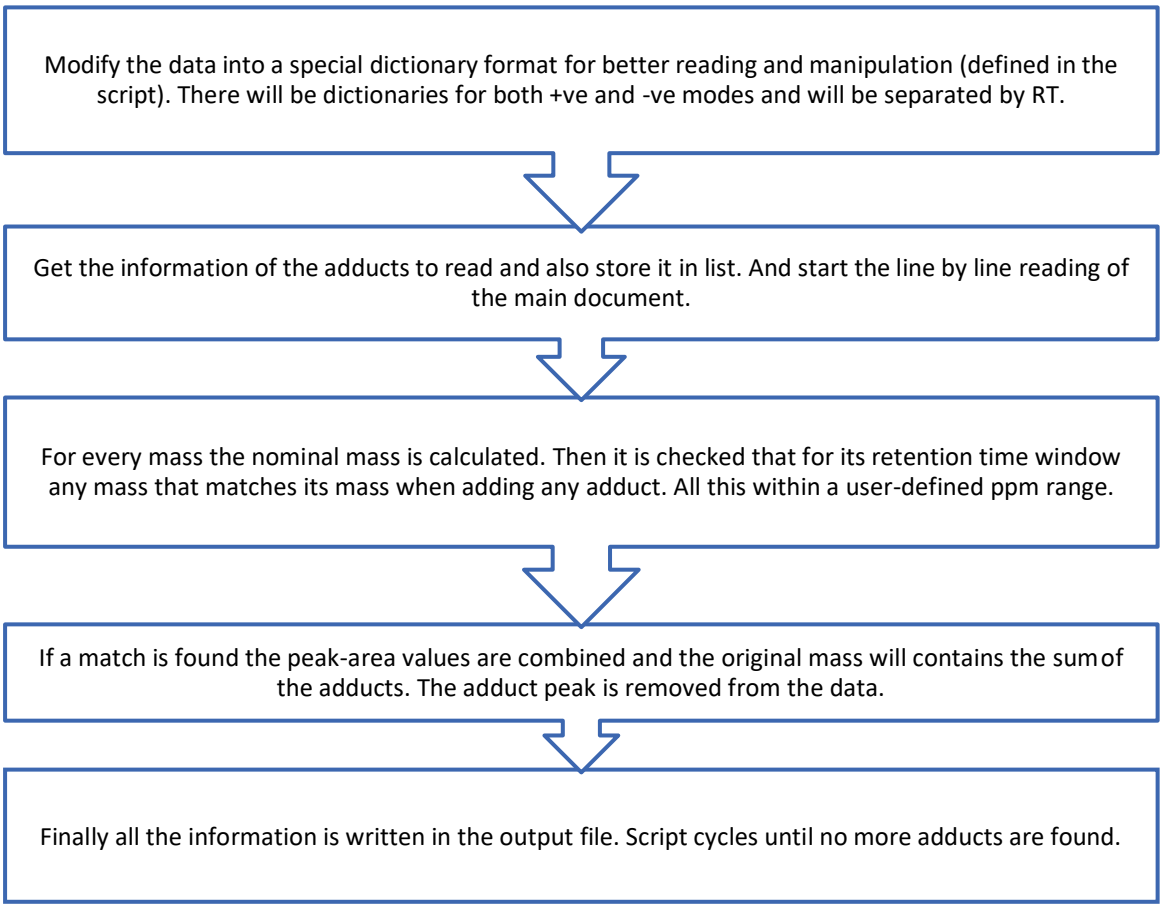


Figure 4.38: Flow chart of Adduct-cleaner.py. The flow chart broadly describes the main steps taken by the program. The program has three inputs: the list of adducts for positive mode, for negative mode and the full excel sheet data with the format described in Appendix 4.B.a.

The program has been tested on three metabolomic datasets (provided by Dr Salah Abdelrazig, School of Pharmacy, Nottingham). The comparison between the input data and the script has been done by checking the Q2 values of the OPLSDA for each dataset, which is a guide of the quality of the OPLSDA model. The bigger the value, the better the model is, thus the motivation of using the program should be to improve the model. A summary of the results can be found in the Table 4.1. The three datasets are:

- Clinical urinary biomarkers of Osteoarthritis (OA): Using the same methodology applied for the malaria and the FOG studies
- Malaria: urine samples, human: A Case Study Based in Ethiopia to Investigate the Effect of Malaria on Urinary Metabolic Profiles and for Biomarker Discovery. The objective was to use LC-MS to identify potential urinary biomarkers of malaria

(Plasmodium falciparum). The specific objectives required to fulfil this aim are as follows:

- FOG: urine samples, human, dietary study. The Influence of Dietary Oligosaccharides on Gut Bacterial Metabolites Detected in Urine. The objective was to use the LC-MS method to explore the changes in the urinary metabolic profiles of healthy subjects treated with a low FODMAP and oligofructose diet.

Table 4.13: Effect of using Lipid-Adduct-cleaner.py on Progenesis datasets.

Experiment		Before treating	After treating	Difference (%)
OA Biomarkers	Number of peaks	7405	6935	-6%
Malaria urine	Number of peaks	2855	2679	-6%
FOG	Number of peaks	4248	3980	-6%

The results reveal an obvious and expected decrease on the number of peaks. That by itself is a good thing because it simplifies the identification task. The Q2 are improved in two cases and slightly worse in the third. This indicates that the software is improving the models and its use should be encouraged.

B. Scripts

a. Adduct-cleaner.py

```
#Ferran Casbas
#30/05/2017
#This program should read list of masses and retentention times and get rid of adducts.

def RTpicker(RT,lRT): #from a list of RT gets the two RT closer to the RT given
    lRT.sort()
    for i in range(len(lRT)):
        if lRT[i]==RT:
            if i==0:
                return(lRT[i],lRT[i+1])
            elif i>(len(lRT)-2):
                return(lRT[i],lRT[i-1])
            return(lRT[i-1],lRT[i],lRT[i+1])
    return False

def PPM(mass1,mass2,ppm):     #check if 2 masses are in a PPM range
    nppm=abs((mass1-mass2)/mass1*1000000)
    if nppm<ppm:
        return True
    else:
        return False

def Masspicker(Mass,lMasses): #Check if a mass
    for i in lMasses:
        if PPM(i,Mass,8):
            return i
    return False

def Adductchecker(Aductmasses,lMasses,iMass):   #Check if an adduct if found
    relatedMasses=[]
    for i in Aductmasses:
        #print(float(iMass),i,float(iMass)+i,lMasses)
        if Masspicker(float(iMass)+i,lMasses):

            relatedMasses.append(Masspicker(iMass+i,lMasses))
    return relatedMasses

def SUMlist(l1,l2): #adds the values of two list
    l3=[]
    for i in range(len(l1)):
        l3.append(l1[i]+l2[i])

    return l3


def listconvertSTRtoFLOAT(l): #converts de values on a list to floats
    l1=[]
    for i in l:
        l1.append(float(i))
    return(l1)
def listconvertFLOATtoSTR(l): #converts de values on a list to str
    l1=[]
    for i in l:
        l1.append(str(i))
    return(l1)
def dicpicker(Mass,l):  #pick the correct dictionary looks in the RT given
    true=[]
    for RT in l:
        try:
            true=DATA[(Mass,RT)]
            break
        except KeyError:
            pass
    return true



Lipids = open('lipid-masses.txt').readlines()   #The file should have the structure of Mode(POS/NEG) - m/z - RT
FULLDATA = open('FULL-DATA.txt').readlines()
AdductsPOS = open('Adducts-POS.txt').readlines()
AdductsNEG = open('Adducts-NEG.txt').readlines()
OUT = open('out.txt',"w")
POS={}  #This dictionaries will store the information of masses acording to their mode
NEG={}

RTl=[]
```

```
for line in Lipids:
    #print (line.split("\t"))
    Mode = line.split("\t")[0]
    Mass = float(line.split("\t")[1])
    RT = line.split("\t")[2].replace('\n','')
    RTl.append(float(RT))
    if Mode == "POS" or Mode == "Pos":
        #now we either add the new mass to its RT if it does not exist or add to another with same RT
        try:
            POS[float(RT)].append(Mass)
        except KeyError:
            POS[float(RT)]=[Mass]
    else:
        try:
            NEG[float(RT)].append(Mass)
        except KeyError:
            NEG[float(RT)]=[Mass]
RTl=list(set(RTl))
RTl=sorted(RTl)
FULLDATA.pop(0) #removes the first line
DATA={}

for line in FULLDATA:
    line=line.replace('\n','')
    Mode = line.split("\t")[0]
    Mass = float(line.split("\t")[1])
    RT = float(line.split("\t")[2])
    DATA[(Mass,RT)]=listconvertSTRtoFLOAT(line.split("\t")[3:])



#Now we have two dictionaries with the structure of:
#POS={RT1:[Mass1, Mass2, ...],RT2:[...],...}
adductPlist=[]
for adduct in AdductsPOS:
    adductp=adduct.split("\t")[-1].replace('\n','')
    adductPlist.append(float(adductp))
adductNlist=[]
for adduct in AdductsNEG:
    adductp=adduct.split("\t")[-1].replace('\n','')
    adductNlist.append(float(adductp))
massestobeRM=[]
finalelist=[]
c=0
for line in Lipids: #Now we read again the document to generate the new one.
    Mode = line.split("\t")[0]
    Mass = float(line.split("\t")[1])
    RT = float(line.split("\t")[2].replace('\n',''))
    #We will convert the mass to nominal
    if Mode=="POS" or Mode=="Pos":
        NominalMass = Mass - 1.007276
    elif Mode=="NEG"or Mode=="Neg":
        NominalMass = Mass + 1.007276
    RTimportant = RTpicker(RT,RTl)
    o=[]
    for i in RTimportant:
        lMasses=[]
        try:
            lMasses=POS[i]
            #we are in positive mode
            a=Adductchecker(adductPlist,lMasses,NominalMass)
            for j in a:         #This small section removes list formating
                o.append(j)
        except KeyError:
            lMasses=NEG[i]
            #we are in Neg mode
            a=Adductchecker(adductNlist,lMasses,NominalMass)
            for j in a: #becuare a can be more than one
                o.append(j)
```

```
    #add RT info to t
    for i in o:
        for rt in RTimportant:
            try:
                true=DATA[(i,rt)]
                break
            except KeyError:
                pass
        massestobeRM.append((i,rt))
    if len(o)==0:
        #no operations needed
        finalelist.append([Mode,str(Mass),str(RT)]+listconvertFLOATtoSTR(DATA[(Mass,RT)]))
        pass
    else:
        summ=DATA[(Mass,RT)]
        for a in o:
            summ=SUMlist(summ,dicpicker(a,RTimportant))

            #summ=DATA[(a,RT)]
            pass
        finalelist.append([Mode,str(Mass),str(RT)]+listconvertFLOATtoSTR(summ))
        #print(Mode+"\t"+str(Mass)+"\t"+str(RT)+"\t"+"\t".join(listconvertFLOATtoSTR(summ)))
    c+=1
    if c//100 == c/100:
        print(c,"/",len(Lipids))
    #print(DATA[(Mode,Mass,RT)],o)

print("Writing")
for line in finalelist:
    Mass = float(line[1])
    RT = float(line[2])
    passes=True
    for a in massestobeRM:
        if (Mass,RT) == a:
            passes=False
            break
    if passes:

        OUT.write("\t".join(line)+'\n')
    else:  #if lines want to be modified
    #    print("*"+"\t".join(line)+'\n')
        OUT.write("*"+"\t".join(line)+'\n')
```

b. List-mass-C13-check-ACT.py

The mzMatchOutput should start with Retention time in the same units and followed by the peak areas. The excel sheet can start with any column in most of our examples we had the metabolite name, followed by RT, C13 mass of the metabolite, n columns where n is the unlabelled data, followed by Average, Std Dev, Unlabelled CV, finally the n columns of labelled data if present.

```
#This program will try to use a series of predifined specifications to generate a better C13 model
#!/usr/bin/env python
#Last update 30/08/17
import xlrd
import numpy as np
import re
"""------------EQUATIONS-------------"""
def CV(L): #oefficient of variation for a list
    return(np.std(L)/np.average(L)*100)
def checkzeros(L): #checks if the list contains the value 0, when it appears will return True
    for a in L:
        if a==0:
            return True
    return False
def PPM(mass1,mass2,ppm):     #check if 2 masses are in a PPM range
    nppm=abs((mass1-mass2)/mass1*1000000)
    if nppm<ppm:
        return True
    else:
        return False
"""------------IMPORTING---------------"""

correctingDATA=open("mzMATCHoutput_Control1.txt",'r').readlines()     #data name
correctingDATA.pop(0)
book = xlrd.open_workbook("CV1.xlsx")                                 #excel name
print ("The number of worksheets is", book.nsheets)
sh = book.sheet_by_index(0)
print (sh.name, sh.nrows, sh.ncols)

colums=6                                                              #colum number
headers =[]
core=[]
D={}
c=0

out=open("sergey-analisis-important-peak-hight.txt",'w')                          #Output name

"""------------PROCESING--------------"""
for a in range(sh.nrows):
    if a==0:
        #header
        for i in range(sh.ncols-1):
            headers.append(sh.cell_value(rowx=a, colx=i))
    else:
        D[sh.cell_value(rowx=a, colx=0)]=[]
        for i in range(1,colums*2+6):
            D[sh.cell_value(rowx=a, colx=0)].append(float(sh.cell_value(rowx=a, colx=i)))
            #print ("Then we are in ", sh.cell_value(rowx=a, colx=i))

##Check if a metabolite needs repair
good=0
bad=0
improved=0
for key, value in D.items():

    #calculate the label/unlabel this set the start and finish line
    L=value[2:colums+3] #change to change the number of colums
    uL=value[colums+5:colums*2+6]   #change to change the number of colums
    LuL=[]
    #print(key,L,uL,checkzeros(uL))
    for i in range(len(L)-1):
        LuL.append(uL[i]/L[i])
        #print(key,CV(L),CV(LuL))
```

```
    if CV(LuL)<CV(L):
        print("good>>>>>>>", key)
        out.write(key+'\t')
        for k in value:
            out.write(str(k)+'\t')
        for k in LuL:
            out.write(str(k)+'\t')
        out.write(str(np.average(LuL))+'\t')
        out.write(str(np.std(LuL))+'\t')
        out.write(str(CV(LuL))+'\t')
        out.write(str(100*(1-CV(LuL)/CV(L)))+'\t')
        out.write('\n')
        good+=1
    else:
        bad+=1
        print("bad >>>>>>>", key)
        MASS=value[1]
        RT = value[0]
        Improved=False
        for l in correctingDATA:
            valuesl=[]
            for j in l.split('\t'):
                valuesl.append(float(j))
            MASSl=valuesl.pop(0)
            RTl=valuesl.pop(0)
            if PPM(MASSl,MASS,20) and checkzeros(valuesl)==False:           #MASS LIMIT
                if RTl> RT-9.001 and RTl<RT+9.001:                             #RT LIMIT
                    #print("possible match..")
                    uL=valuesl
                    LuL=[]
                    for i in range(len(L)-1):
                        LuL.append(uL[i]/L[i])
                    #print(value)
                    #print(key,CV(L),CV(LuL))
                    if CV(LuL)<CV(L):
                        print("improved", int(100-(CV(LuL)/CV(L))*100), "%")
                        out.write(key+'\t')
                        for k in value[:colums*2]:
                            out.write(str(k)+'\t')

                        for k in uL:
                            out.write(str(k)+'\t')
                        for k in LuL:
                            out.write(str(k)+'\t')
                        out.write(str(np.average(LuL))+'\t')
                        out.write(str(np.std(LuL))+'\t')
                        out.write(str(CV(LuL))+'\t')
                        out.write(str(100*(1-CV(LuL)/CV(L)))+'\t')
                        out.write('\n')
                        Improved=True
                        improved+=1
                        break
                    else:
                        pass
                        #print("not improved")

        if Improved==False:
            out.write(key+'\t')
            for k in value:
                out.write(str(k)+'\t')
            for k in LuL:
                out.write(str(k)+'\t')
            #out.write("BAD"+'\t')
            out.write(str(np.average(LuL))+'\t')
            out.write(str(np.std(LuL))+'\t')
            out.write(str(CV(LuL))+'\t')
            out.write(str(100*(1-CV(LuL)/CV(L)))+'\t')
            out.write('\n')
out.close()

print("\n\n\n\n\n\n>>>The final results show that:")
print("A total of ",good," metabolites were already correct")
print("A total of ",bad," metabolites were bad")
print("From the bad ones a total of ",improved," metabolites have been properly normalized")
print("Leaving a final ", str(bad-improved), " non improved metabolites and ", good+improved,"improved")
print("This means that ", str(int(float(good+improved)/float(bad + good)*100)) ,
    "% of the metabolites have been improved by the methodology")
```

c. Excelprocesing-2.0.py

```
#!/usr/bin/env python
import xlrd
import re
while True:
  try:
    FileName = input("Introduce the name of the excel sheet => ")
    book = xlrd.open_workbook(FileName)
    break
  except:
    print("Error the file was not found, check the name of the file and if it is in the same folder as this program.")
    pass
print ("The number of worksheets is", book.nsheets)
sh = book.sheet_by_index(0)
column=input("Introduce the column number you are interested in  => ")
compounds = []
once=True
l=[]
ll=[]
f = open('output.txt', 'w')
#f.write('\t')
#book.sheet_names() contains a list of all the names of each sheet 1carbox
for a in range(len(book.sheet_names())-1):
  sh = book.sheet_by_index(a)
  #print (sh.name)
  sheetName = sh.name
  try:
    f.write('\t'+sheetName)
  except  UnicodeEncodeError:
    print("Warning wierd characters detected in the sheet with name: " +sheetName)
    sheetName = re.sub('[^A-Za-z0-9-_*.]+', '', sheetName)
    f.write('\t'+sheetName)
    print("The special characters have been removed and the saved name is "+ sheetName)
  ll=[]
  #print(sh.nrows)
  for row in range(5,sh.nrows-6):
    ll.append(sh.cell_value(rowx=row, colx=int(column)-1)) #1 less than the one on excel sheet
  l.append(ll)
  if once:
    for row in range(5,sh.nrows-6):
      compounds.append(sh.cell_value(rowx=row, colx=0))
      #print(sh.cell_value(rowx=row, colx=0))
    once=False
f.write('\n')
for y in range(0,len(ll)):

  f.write(compounds[int(y)])
  for i in range(len(l)):
    #print(i,y)
    f.write('\t'+str(l[i][y]))#+' '+str(l[i][y+1])+' '+str(l[i][y+2]))
  f.write('\n')
f.close()
```

d. Mass-LipidMAPS-Checking.py

```
def RemoveLink(text):

    if len(text.split('(<a'))==2:
        newtext = text.split('(<a')[0]+text.split('</a>)')[-1]
        return(newtext)
    else:
        return text
import urllib2
import urllib
import subprocess

time=False
while True:
    try:
        File = str(raw_input('Write the name of the file containing the list of masses >>> '))  #Mass-list.txt
        FILE = open(File,'r').readlines()
        OUT = open('Metabolite.txt','w')
        if len(FILE[0].split('\t')) == 2 or len(FILE[0].split('\t')) == 3:
            if len(FILE[0].split('\t')) == 3:
                time=True
            break
        else:
            print('File format does not mach the requires')
    except IOError:
        print('\nFile name not found!')
        print('Make sure the name of the file is well writen and that is contained in the same folder that this program\n')
MaxPPM = int(input('Define the maximum ppm allowed >>> '))

PositiveMode = (['[M+H]',-1.0073],['[M+NH4]',-18.0338],['[M+Na]',-22.9892],['[M+H-H2O]',17.002])
NegativeMode = (['[M-H]',1.0073],['[M-H2O-H]',19.0183],['[M+Ac-H]',-59.0139],['[M-H+NaAc]',-82.0028],['[M-H+2NaAc]',-82.0028*2],['[M-H+3NaAc]',-82.0028*3])
c = 0.0
if time:
    OUT.write('Mass\tAdduct\tID\tName\tReal Mass\tFamily\tPPM\tRetention time\n')
else:
    OUT.write('Mass\tAdduct\tID\tName\tReal Mass\tFamily\tPPM\n')


for line in FILE:
    c+=1.0
    print str(int(c/len(FILE)*100))+'%'
    line = line.replace('\n','')
    if time:
        Time = line.split('\t')[2]
    Mass = line.split('\t')[1]
    Mode = line.split('\t')[0]
    if Mode == 'Pos' or Mode == 'POS' or Mode == 'pos' :
        playlist = PositiveMode
    elif Mode == 'Neg' or Mode == 'NEG' or Mode == 'neg':
        playlist = NegativeMode
    else:
        print ('Mode not understood in the Mass-list.txt. Please check the format')
        break
    for Adduct in playlist:
        AdductMass = Adduct[1]
        Adduct = Adduct[0]
        NewMass = str(AdductMass+float(Mass))
        req = urllib2.Request('http://www.lipidmaps.org/data/structure/LMSDSearch.php?&ExactMass='+NewMass+'&Mode=ProcessTextOntologySearch')
        response = urllib2.urlopen(req)
        the_page = response.read()
        the_pageL = the_page.split('<table class="datatable" border="0" width="100%">')

        if len(the_pageL) == 1:
            #print(Mass, "didn't get any result back")
            OUT.write(Mass+'\t'+Adduct+'\tNothing Found\n')
        else:#We found the table and we cut it to the important pieces for us
            table = the_pageL[1].split('</table>')[0]
            Compoundlist = table.split('<tr')
            Compoundlist.pop(0)
            Compoundlist.pop(0)
            NiceCompList=[]
            for i in Compoundlist:#Edit the lsist of compounds
                pieces = i.split('<td>')
                CompoundName = pieces[2].split('</td>')[0].replace('<br>','')
                CompoundName=RemoveLink(CompoundName)
                if CompoundName == '-':
                    CompoundName = pieces[3].split('</td>')[0].replace('<br>','')
                NiceCompList.append([pieces[1].split('</a>')[0].split('">')[1] \
                    ,CompoundName, \
                    pieces[5].split('</td>')[0],pieces[6].split('</td>')[0] \
                    ,(int(abs(float(pieces[5].split('</td>')[0])-float(NewMass))/float(NewMass)*1000000))])

            for i in NiceCompList:#Now we go through the table and ask for the lvl of accuracy
                #print(i[-1],MaxPPM)
                if abs(i[-1]) <= MaxPPM:
                    once = True
                    OUT.write(Mass+'\t'+Adduct+'\t')
                    for a in i:
                        OUT.write(str(a)+'\t')
                    if time:
                        OUT.write(Time+'\n')
                    else:
                        OUT.write('\n')
            if once == False:
                OUT.write(Mass+'\t'+Adduct+'\tNo results in this PPM range\n')
print('Processing done')
OUT.close()
```

e. Mass-LipidMAPS-Cleaner.py

```
def ListSpliter(l,Colum):       #The program gets as imput a list of list and give back a list of list
    oldparameter =  l[0][Colum]
    finallist = []
    pack= []
    for line in l:
        newparameter=line[Colum]
        #print(newparameter,oldparameter)
        if newparameter != oldparameter:
            oldparameter=newparameter
            finallist.append(pack)
            pack= []
        pack.append(line)
    finallist.append(pack)
    return finallist

def FamilyProcessing(l):
    if len(l) == 1:
        return(l)
    namelist=[]
    for i in l:
        namelist.append(i[3])
    if len(namelist[0].split('-')) ==2:
        namebase=namelist[0].split('-')[0]+'-'
    else:
        namebase=namelist[0].split('(')[0]
    #print(namebase)
    for name in namelist:   # in a family all should be equal
        if namebase != name.split('(')[0]:
            return(l)
    #first we check if s tri or not:
    name = namelist[0]
    name = name.replace(' / ',' - ')
    if len(name.split(r'/'))==2:
        #DG
        cutted = name.split(':')
        if len(cutted)==1:
            return([l[0]])
        else:
            P1=int(cutted[0].split('(')[-1].replace('/','').replace('d','').replace('t','').replace('m','').replace('O-',''))

            P2=int(cutted[1][:1])
            S1=int(cutted[1][-2:].replace('/',''))
            S2=int(cutted[2][:1])
            newname = namebase+'('+str(P1+S1)+':'+str(P2+S2)+')'
            l[0][3]=newname
            return([l[0]])
    elif len(name.split('/'))==3:
        #TG
        cutted = name.split(':')
        P1=int(cutted[0][-2:].replace('/',''))
        P2=int(cutted[1][:1])
        S1=int(cutted[1][-2:].replace('/',''))
        S2=int(cutted[2][:1])
        T1=int(cutted[2][-2:].replace('/',''))
        T2=int(cutted[3][:1])
        newname =namebase+'('+str(P1+S1+T1)+':'+str(P2+S2+T2)+')'
        l[0][3]=newname
        return([l[0]])
    else:
        #print(name+'Not compacted')
        return(l)
f = open('Metabolite.txt','r').readlines()
OUT = open('Clean-met.txt','w')
tittle=f[0]
f.pop(0)
data=[]
for line in f:
    data.append(line.split('\t'))
L1 = ListSpliter(data,0)

for i in L1:
    for x in i:
        if len(x) == 8:
            ContainsInfo=True
            break
        ContainsInfo=False
    if ContainsInfo==False:
        OUT.write(x[0]+'\t No results for this mass in any condition\n')
    L2 = ListSpliter(i,1)
    for t in L2:
        if len(t[0])!=3:
            L3 = ListSpliter(t,5)
            for k in L3:
                #print(k)
                k=FamilyProcessing(k)
                for j in k:

                    OUT.write('\t'.join(j).replace(r'&#946;','Beta').replace(r'&#945;','Alfa'))     #
    OUT.write('-------\n')
OUT.close()
```

f. Graph-RT.py

```
#Small program to do plots

def dicMOD(D):  #converts the data in the dic to %
    summ=0
    for key, value in AddDIC.items():
        summ+=value
    if summ==0:
        return AddDIC
    for key, value in AddDIC.items():
        AddDIC[key]=value/summ
    return AddDIC

data = open('Clean-met.txt').readlines()
o = open('out.txt',"w")
AddDIC={'[M-H2O-H]':0,'[M-H]':0,'[M+Na]':0,'[M+Ac-H]':0,'[M+H]':0,
'[M+H-H2O]':0,'[M+NH4]':0,'[M-H+NaAc]':0,'[M-H+2NaAc]':0,'[M-H+3NaAc]':0}


o.write("Mass\tRT\t")
for key, value in AddDIC.items():
    o.write(str(key)+"\t")
o.write("\n")
for l in data:
    l=l.replace("\n","")
    if l == '-------':
        #print(AddDIC)
        a=dicMOD(AddDIC)

        o.write(mass+"\t"+rt+"\t")
        for key, value in AddDIC.items():
            o.write(str(value)+"\t")
        o.write("\n")

        AddDIC={'[M-H2O-H]':0,'[M-H]':0,'[M+Na]':0,'[M+Ac-H]':0,'[M+H]':0,
        '[M+H-H2O]':0,'[M+NH4]':0,'[M-H+NaAc]':0,'[M-H+2NaAc]':0,'[M-H+3NaAc]':0}
    else:
        #print(l.split("\t")[2][0])
        mass = l.split("\t")[0]
        Add = l.split("\t")[1]
        rt = l.split("\t")[-1]#-1 rt -3 family
        if l.split("\t")[2][0]!="N":
            for key, value in AddDIC.items():
                if key == Add:
                    AddDIC[key]+=1
```

g. Graph-Family-2.py

```
#Gathers the family and adduct ifo together
def dicMOD(D):  #converts the data in the dic to %
    summ=0
    for key, value in AddDIC.items():
        summ+=value
    if summ==0:
        return AddDIC
    for key, value in AddDIC.items():
        AddDIC[key]=value/summ
    return AddDIC
data = open('Clean-met.txt').readlines()
intensity = open('intensity.txt').readlines()
o = open('out.txt',"w")
AddDIC={'[M-H2O-H]':0,'[M-H]':0,'[M+Na]':0,'[M+Ac-H]':0,'[M+H]':0,'[M+H-H2O]':0,
    '[M+NH4]':0,'[M-H+NaAc]':0,'[M-H+2NaAc]':0,'[M-H+3NaAc]':0}
FamDIC={'FA':0,'GL':0,'GP':0,'SP':0,'ST':0,'PR':0,'SL':0,'PK':0,'NI':0}

o.write("Mass\tIntensity\t")
for key, value in AddDIC.items():
    o.write(str(key)+"\t")
for key, value in FamDIC.items():
    o.write(str(key)+"\t")
o.write("\n")
c=0

for l in data:
    l=l.replace("\n","")
    if l == '-------':
        #print(AddDIC)
        a=dicMOD(AddDIC)
        imp=intensity[c].replace("\n","")
        o.write(mass+"\t"+imp+"\t")
        for key, value in AddDIC.items():
            o.write(str(value)+"\t")
        for key, value in FamDIC.items():
            o.write(str(value)+"\t")
        o.write("\n")
        c+=1
        AddDIC={'[M-H2O-H]':0,'[M-H]':0,'[M+Na]':0,'[M+Ac-H]':0,'[M+H]':0,'[M+H-H2O]':0
        ,'[M+NH4]':0,'[M-H+NaAc]':0,'[M-H+2NaAc]':0,'[M-H+3NaAc]':0}
        FamDIC={'FA':0,'GL':0,'GP':0,'SP':0,'ST':0,'PR':0,'SL':0,'PK':0,'NI':0}
        ForceID=False
    else:
        print(l.split("\t"))
        mass = l.split("\t")[0]
        Add = l.split("\t")[1]
        if l.split("\t")[2][0]!="N" and ForceID==False:
            Family= l.split("\t")[-3]
            try:
                Family=Family.split("[")[1]
                Family=Family[:-3]
            except IndexError:
                Family = 'NI'
            #print(Family)
            for key, value in AddDIC.items():
                if key == Add:
                    AddDIC[key]+=1
                    FamDIC[Family]+=1
                    if Add == '[M-H]' or Add== '[M+H]':
                        ForceID=True
```

h. Graph-FA-separator.py

```
def dicMOD(D):  #converts the data in the dic to %
    summ=0
    for key, value in D.items():
        summ+=value
    if summ==0:
        return D
    for key, value in D.items():
        D[key]=value/summ
    return D

data = open('Clean-met.txt').readlines()
out= open('Fa-out.txt','w')

AddDIC={'[M-H2O-H]':0,'[M-H]':0,'[M+Na]':0,'[M+Ac-H]':0,'[M+H]':0,'[M+H-H2O]':0,
    '[M+NH4]':0,'[M-H+NaAc]':0,'[M-H+2NaAc]':0,'[M-H+3NaAc]':0}
FAcat ={"01":0,"02":0,"03":0,"04":0,"05":0,"06":0,"07":0,"08":0,"09":0,"10":0,
    "11":0,"12":0,"13":0,"00":0}

out.write('Mass\t')
for key, value in FAcat.items():
    out.write(key+'\t')
for key, value in AddDIC.items():
    out.write(key+'\t')
out.write('\n')
c=0
FAPASS=False
for l in data:
    l=l.replace("\n","")
    if l[:3]=="---" and FAPASS:

        out.write(Mass+'\t')
        for key, value in FAcat.items():
            out.write(str(value)+'\t')
        for key, value in AddDIC.items():
            out.write(str(value)+'\t')
        out.write('\n')

        FAcat ={"01":0,"02":0,"03":0,"04":0,"05":0,"06":0,"07":0,"08":0,
        "09":0,"10":0,"11":0,"12":0,"13":0,"00":0}
        AddDIC={'[M-H2O-H]':0,'[M-H]':0,'[M+Na]':0,'[M+Ac-H]':0,'[M+H]':0,
        '[M+H-H2O]':0,'[M+NH4]':0,'[M-H+NaAc]':0,'[M-H+2NaAc]':0,'[M-H+3NaAc]':0}
        FAPASS=False

    try:
        if l.split("\t")[2][0]!="N":
            Mass = l.split("\t")[0]
            Add = l.split("\t")[1]
            Family= l.split("\t")[-3]
            Family=Family.split("[")[1]
            Family=Family[:-1]
            if Family[:2] == "FA":
                FAPASS = True
                FAcat[Family[2:]]+=1
                #print(Family[2:])
                AddDIC[Add]+=1
    except:
        pass

out.close
```

C. C13 Correction results

The effectiveness of this program has been tested in 6 experiments summarized in table 4.1. While all the tests have been done in metabolomics it can be applied in lipidomics perfectly, the analytical experiment method should change but all the data analysis process is exportable.

Table 4.14: Summary of the IDEOM experiments. Six experiments containing C13 normalization have used the program to improve on the number of metabolites that can get normalized.

Experiment name	Type of sample	Number of samples	N of metabolites before	N of metabolites after	% of improvement
Human Brain	Brain tissue	18	224	356	**61**
Mouse plasma	Plasma	12	320	440	**58**
Human Urine Tea	Urine	10	357	593	**62**
Human urine diabetes	Urine	36	332	640	**65**
Human Brain Glioma	Brain tissue	30	165	340	**67**
Mouse Liver Alzheimer	Liver tissue	24	268	353	**57**

In all studies, the number of metabolites that could be normalized has greatly increased (60% on average). So, the usage of this tool is greatly advised when working with any C13 normalization, but, of course, the control of the parameters set in order to achieve this normalization is absolutely crucia