Transomics: Integrating core ‘omics’ concepts

Joseph M Foster

European Bioinformatics Institute

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

15th of September 2012

We thought there was no more behind
But such a day tomorrow as today
And to be a boy eternal

The Winter's Tale. Shakespeare

I would like to dedicate this thesis to my mother and father; Alison and Michael Foster for their endless supply of encouragement and perseverence. To my partner Gillian I would like to thank her for putting up with me for the last 8 years, particularly throughout the more stresssful times of my studies and while I travelled to exotic locations in the name of science, leaving her behind with our two cats Poppy and Azrael.

Acknowledgements

I would like to begin by thanking my supervisors Dr. Rolf Apweiler, Prof. Lennart Martens and Dr. Matthieu Visser for giving me the opportunity to study under them, an opportunity I am still surprised I received today. I hope our time together has been mutually beneficial and that one day I can begin to repay the great service they have done me. I would like to thank them for the various opportunities to travel, present my work and meet new people. They have supplied the majority of my opportunities for new projects, always shrewdly judging the merits of taking on new work in favour of continuing old, regularly talking more sense in few sentences than I sometimes do days at a time. Their scientific wisdom and experience in research has been invaluable. I would like to thank them for the countless hours of discussion over small details and their patience while I found my feet.

The members of the PRIDE team have been like an extended family to me and I thank them for all the lunches, cups of tea and discouraging feedback on my half-brained schemes. Thanks to Richard Côté and Florian Reisinger for showing me the fundamentals of programming and the intricacies of the PRIDE database schema when I was working on proteomics quality control. Particular thanks go to Dr Juan Antonio Vizcaíno who while not being an official supervisor of mine, has surpassed that role, being both a great friend and a gifted listener. I thank him for his continued support in my lipidomics related work and the many opportunities I have had to present it as a result.

I would like to thank my fellow students without whose companionship and support none of us would have succeeded. Special mention goes to Pablo Moreno for his interest in my work and subsequent collaboration on the ‘LipidHome’ project. Similarly I would like to thank Antonio Fabregat for his time and unparalleled expertise in web application development. Without his input it would not have been possible to produce the ‘LipidHome’ database to nearly the same high standard. I would like to acknowledge his direct boss Henning Hermjakob for his support and allocating the project a portion of Antonio’s time.

To my Thesis Advisory Committee; Dr. Kathryn Lilley, Dr. Jeroen Krijgsveld, Dr. Alvis Brazma, Dr. Matthieu Visser and Dr. Rolf Apweiler, I would like to acknowledge that their hours spent reading my research summaries and listening to my presentations were very much appreciated. Their feedback was instrumental in consolidating my rapidly diverging work into a comprehensible story.

For proof-reading I would like to thank Prof. Lennart Martens, Dr. Juan Antonio Vizcaíno and Dr. Rolf Apweiler. Also, my mother Alison Foster as after 25 years I still show no sign of mastering my native language to even a fraction of the level she commands. Special thanks also to Tim Wiegels for his help compiling the thesis and help with LaTeX, with final proof reading by David Ovelleiro, Rui Wang and Johannes Griss.

I would like to thank Dr Qifeng Zhang for providing the colorectal cancer lipidomics data which has made a large positive impact on the content of my work. I am thankful to Prof. Mike Wakelam and Prof. Lennart Martens for organising this collaboration initially.

Many thanks go to my examiners Prof. Mike Wakelam and Prof. Juri Rappsilber for their insightful critique of my work, resulting in a thesis I am truly proud of.

Lastly, but by no means least, I would like to thank my partner Gillian for her continuous support throughout the PhD, particularly in the last few months before my submission.

Abstract

In recent years there has been an explosion in the number of biological fields grouped under the umbrella term ‘omics’. While seemingly disparate, they all share the same general approach: to perform the high-throughput identification, quantification and analysis of biological molecules. Most commonly, nucleic acids, proteins, small molecules or hybrid studies at the interface of these molecules. Technology has driven these fields from small scale pioneering efforts, analysing single samples and a small number of molecules, to incredibly fast, multi-components systems capable of chaining the analysis of many samples and multiple molecules. While the core concept of analysing biomolecules remains constant, the instrumentation required is as diverse as the aims of the ‘omics’ approaches as a whole. Instrumentation has developed at varying rates in different fields and the heterogeneity within fields has led to several bioinformatics challenges at different levels. However diverse, universal bioinformatics approaches can be applied to make simple, reusable software and analysis tools.

The work presented in this thesis aims to identify some of these universal bioinformatics challenges in dealing with high throughput ‘omics’ data. Once identified a field specific solution is generalised and applied to other ‘omics’ fields. The study focuses on three challenging areas of ‘omics’ research: Quality control, reference sequence databases and statistical data analysis. New applications and analysis approaches have been developed and implemented in the fields of proteomics and lipidomics.

First of all, with the aim of re-using data for new purposes, I performed a quality control (QC) analysis of publically available mass spectrometry (MS) derived data from the PRoteomics IDEntification (PRIDE)

database. The work highlighted several methods for thorough evaluation of this type of data. In addition, an open source R library was made available in order to make these methods accessible to the community.

Following on from QC, the concept of a reference space of all proteins (taken from UniProt) was translated to the field of lipidomics, culminating in the creation of a database of theoretical lipid species relevant to modern high throughput technologies, called ‘LipidHome’. Alongside its development and design, a web application was provided to easily propagate the information to bioinformaticians and wet lab scientists alike.

The final part of this work consisted of the statistical analysis of human colorectal cancer lipidomics data. Data transformations and analyses adapted from existing genomics and metabolomics approaches were applied to the analysis of quantitative lipid species measurements, in order to investigate the effect that colorectal cancer has on the lipidome. Several significant conclusions could be drawn from the analyses including the development of a robust machine learning classifier that predicts whether a sample is of tumour or normal origin based upon quantitative lipid data alone.

Contents

List of Figures

Nomenclature

General abbreviations

API Application Programming Interface.

ChIP-seq Combines Chromatin immunoprecipitation with high throughput DNA sequencing.

CID Collision induced dissociation.

Da . Dalton.

DNA Deoxyribonucleic acid.

ESI . Electrospray ionisation.

HPLC High Pressure Liquid Chromatography.

HUPO HUman Proteome Organisation.

IgY-12 Multi Protein Immunoaffinity subtraction.

iTRAQ Isobaric Tag for Relative and Absolute Quantitation.

LSA Latent Semantic Analysis.

m/z Mass to charge ratio.

MARS-6 Multiple Affinity Removal System.

MDS Multi Dimensional Scaling.

MS . Mass Spectrometry.

mzML A mass spectrometry data format.

mzXML A mass spectrometry data format.

PEG Polyethylene Glycol.

QC . Quality Control.

REST REpresentational State Transfer.

RNA Ribonucleic acid.

RNA-seq High throughput sequencing technology.

SCX Strong Cation Exchange.

SOAP Simple Object Access Protocol.

TMT Tandem Mass Tags.

XML eXtensible Markup Language.

Databases and bioinformatics resources

ArrayExpress A database of publically available experimental microarray and RNA-seq data. Available at http://www.ebi.ac.uk/arrayexpress/.

Bioconductor Open source software framework that develops R packages for the biosciences. Available at http://www.bioconductor.org/.

Biomart An opensource database system that integrates multiple data resources into one easy to query entity. Available at http://www.biomart.org/.

EBI European Bioinformatics Institute. A portal of bioinformatics tools, resources and training materials. Available at www.ebi.ac.uk.

ENA European Nucleotide Archive. A genetic sequence database. Available at http://www.ebi.ac.uk/ena/.

Ensembl Software and tools for the assembly of genomes, their visualisation and automatic annotation. Available at http://www.ensembl.org/index.html.

GenBank Genetic sequence database provided by the NIH. Available at http://www.ncbi.nlm.nih.gov/genbank/.

GEO Gene Expression Omnibus. A database of public funcitonal genomics data, including microarray and sequencing based experiments. Available at http://www.ncbi.nlm.nih.gov/geo/.

GPMDB Global Proteome Machine DataBase. A database of publically available proteomics data. Available at http://gpmdb.thegpm.org/.

IntAct Molecular interactions database. Available at http://www.ebi.ac.uk/intact/.

JFreeChart A java chart library. Available at http://www.jfree.org/jfreechart/.

LMSD Lipid Maps Structural Database. A database of lipid structures and metadata annotations. Available at http://www.lipidmaps.org/data/structure/index.html.

MASCOT Mass spectrometry search engine software, identifies peptides from spectra and infers proteins from sequence databases. Available at http://www.matrixscience.com/.

Metabolights A repository of publically avaialble metabolomics data. Currently in beta. Available at http://www.ebi.ac.uk/metabolights/.

MySQL A relational database management systems. Available at http://www.mysql.com/.

NCBI National Center for Biotechnology Information. A large resource containing multiple public databases and reference datasets. Available at www.ncbi.nlm.nih.gov/.

NIH National Institute of Health. An agency of the United States Department of Health. Available at http://www.nih.gov/.

PDB Protein Data Bank. A repository for experimentally derived structures of proteins, nucleic acids and complexes. Available at http://www.rcsb.org/pdb/home/home.do.

PeptideAtlas A database of publically avialable proteomics data. Available at http://www.peptideatlas.org/.

PRIDE PRoteomics IDEntifications database. A database of publically avialable proteomics data. Available at http://www.ebi.ac.uk/pride/.

PRIDE Inspector A tool for downloading and visualising both local mass spectrometry files in standard formats and data from the PRIDE database. Available at http://www.ebi.ac.uk/pride/webstart/pride-inspector.jnlp.

PSI . Proteomics Standards Initiative. A portal for proteomics data standard and reporting specifications. Available at http://www.psidev.info/.

R . . . A statistical programming language. Available at http://www.r-project.org/.

UniProt Protein sequence database. Available at http://www.uniprot.org/.

Xcalibur Mass spectrometry instument control and data analysis software. Available at http://www.thermoscientific.com.

Lipid abbreviations

FA . Fatty Acids.

Chapter 1

Introduction

1.1 Background

The "central dogma of molecular biology" was first described by Francis Crick, it stated that DNA is transcribed to RNA, which in turn is translated to proteins (Crick, 1970). From that era onwards modern biology has been the study of diverse molecular interactions spanning a wide range of biomolecules and the effects they have. Indeed, each level studied in isolation is no longer sufficient to provide profound insight into the highly interconnected inner workings of cells. Biomolecules can thus be put into context of their regulation and downstream effects, providing a better view of the entire cellular picture. Molecular biology is an umbrella term for the functional study of proteins, genes and the relationship between the two. Traditionally this involved the study of a single protein or gene target to fully elucidate its role within the organism. In the last decade instrumentation and bioinformatics advances have enabled the development of research approaches that aim to detect and quantify large numbers of biomolecules in parallel, creating a less focused but more comprehensive picture of a sample's temporal state. This high throughput, large scale analysis of biological samples is commonly referred to as 'omics' and can be split into a number of sub-disciplines depending on the biomolecule under study. The large scale study of biomolecular interactions and dynamics is also encompassed under the umbrella term 'omics', with fields such as interactomics (Alexeyenko *et al.*, 2012) and modificomics (Reinders & Sickmann, 2007) having a large temporal aspect to the characterisation of biomolecules. Simplified below are some of the most common 'omics' disciplines.

1.1.1 'Omics' fields

1.1.1.1 Genomics

Genomics is the study of whole genomes of organisms. It is the large scale identification and quantification of the total complement of genes in a sample and the interactions between them. From the first viral genomes sequenced in the 1970's by pioneers such as Fred Sanger (Sanger *et al.*, 1977), the sequencing of 2823 viral genomes has been completed and the sequences have been submitted to the NCBI GenBank(data collected 17/04/2012). The human genome project, initially completed in 2001 after 10 years, has evolved into a number of follow up projects. Inluding a number of efforts to sequence the cancer genomes, summarised by (Mardis & Wilson, 2009). The intial human genome project has since been superseded by the ten thousand human genomes project. Announced in 2010 to identify disease causing variants over the course of three years, this project shows the rate at which genome sequencing has advanced in the last decade in terms of speed and cost (Mardis, 2011). The field offers a variety of options in terms of instrumentation, including pyrosequencing (Ronaghi *et al.*, 1996), nanopore sequencing (Kasianowicz *et al.*, 1996) and ion semiconductor sequencing (Rothberg *et al.*, 2011). Modern instruments produce an astounding volume of data that must be analysed and interpreted in an automatic fashion. Recent developments in the field have therefore been achieved in the storage, analysis and dissemination of huge quantities of data (Stein, 2010). The field of genomics has a number of competing bioinformatics resources for the storage and presentation of whole genome sequence data, the most prominent of which is Ensembl (Flicek *et al.*, 2011). Sequence databases storing individual nucleotide sequences are also available from the European Nucleotide Archive (ENA) and Genbank. Aspects of bioinformatics in this field are amongst the most advanced in any field, this is largely a product of elevated funding and new technology uptake. As such it is an excellent template for the evolution of bioinformatics in other fields. Knowledge of a genome is clearly important for estimating the theoretical proteome and hence the molecular functionality available to a biological sample. However, due to its static nature the genome does not give information about the change in expression of the genome over time or its differential expression in different cell types.

1.1.1.2 Transcriptomics

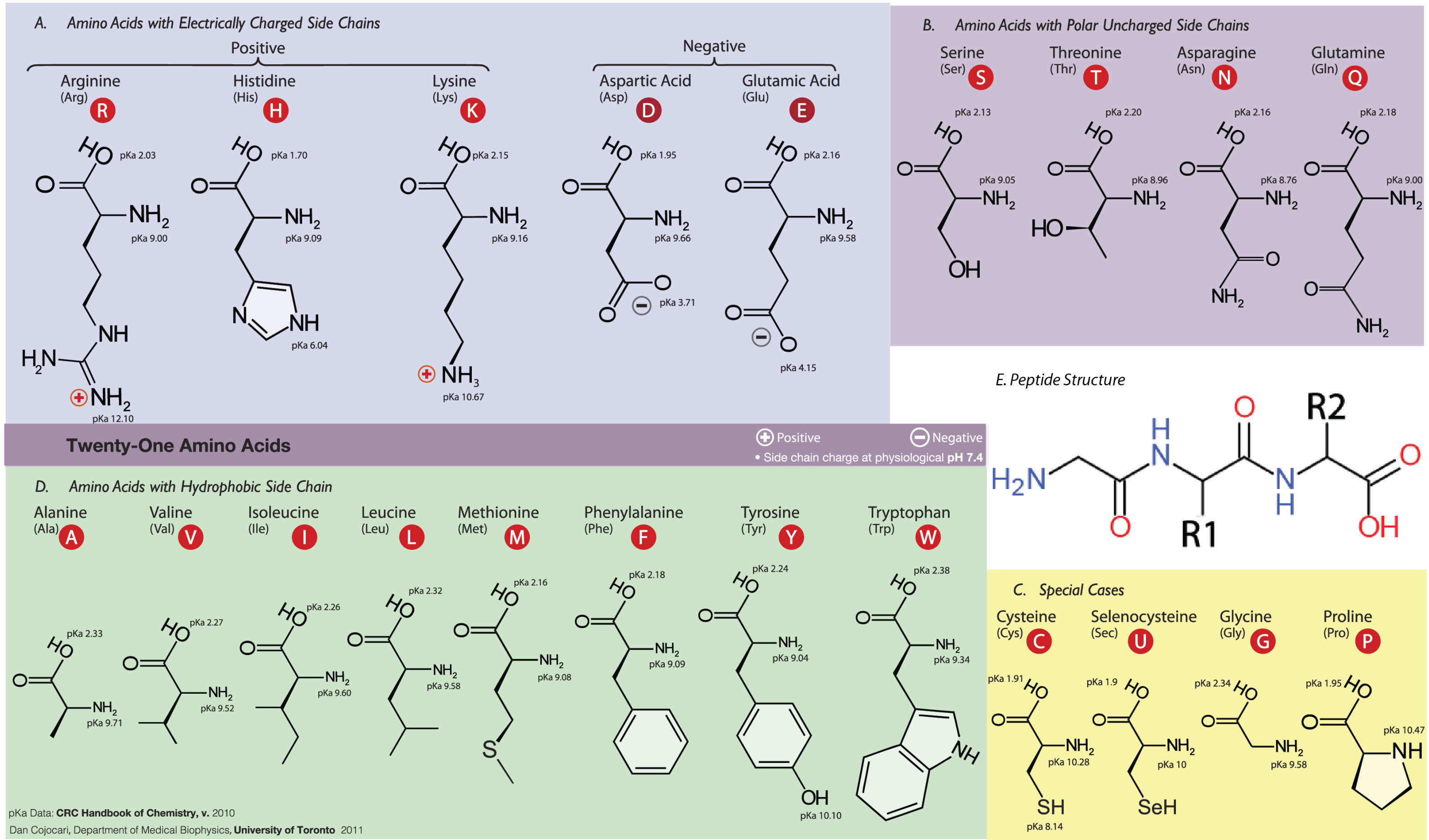
Transcriptomics is the study of all RNA molecules in a sample, whether a tissue or single cell. This includes messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA) and other non-coding RNA (e.g. piRNA). The transcriptome reflects not only the genome, which dictates what can be transcribed, but also the time and condition dependent expression of the genome. As such it is dynamic and can be studied in a time course or treatment-response manner. The expression level of mRNA which forms the template from which proteins are translated is measured using two main technologies.

The first of which to be developed was microarrays. Microarray technology evolved out of the Southern blot method (Southern, 1975) which separates DNA fragments by 2D gel electrophoresis, followed by transfer to a filter membrane of nitrocellulose and heating to fix the transferred DNA fragments. The membrane is exposed to a synthetic hybridisation DNA probe of known sequence, tagged by either radioactive isotopes or a fluorescent dye and the hyrbidisation visualised to confirm the presence of the complementary sequence to the probe. By contrast a microarray is a grid of known DNA probes fixed to a solid support such as glass or nylon, the probes are exposed to a prepared sample of mRNA. Hybridisation follows between probes and sample material and the level of hybridisation is detected to identify the mRNA. An estimate of the amount of mRNA as a surrogate for the amount of protein it translates to is also calculated. Individual DNA probes are mapped to genes and in this manner expression of genes can be investigated (Maskos & Southern, 1992). Early pioneers of the field were the California based company Affymetrix, who made commercially available the GeneChip in 1994 and have since generated competition from other manufactures of functionally similar platforms such as Illumina's bead based technology. However, the generalised approach and early dominance of the Affymetrix approach helped to consolidate data formats and analysis protocols and produce a much more mature field than proteomics. The second approach involves the more recent adaption of nucleotide sequencing technology used in genomics, this has enabled the quantitative sequencing of mRNA, known as RNA-seq (Wang *et al.*, 2009). However this is considerably more expensive and shares a problem with proteomics concerning high abundance proteins: 75% of all identifications are from only 5% of the total transcriptome (Łabaj *et al.*, 2011). mRNA as a surrogate for protein levels in the sample has been proven to be inaccurate in some instances and so proteins must also be identified and quantified directly (Rogers *et al.*, 2008). These transcriptome technologies are clearly dependent on accurate sequence database records for the organism under study. Repositories of experimental data are also available in this field, storing large numbers of publicly available microarrays, RNA-seq and ChIP-seq data; the main ones are ArrayExpress (Parkinson *et al.*, 2011) and Gene Expression Omnibus (GEO) (Barrett *et al.*, 2011).

1.1.1.3 Proteomics

Proteins consist of one or more polymer chains of amino acids folded in a specific conformation that imparts biological function, see figure 1.1 for a table of the amino acids and the generic structure of a peptide. The proteome is the entire complement of proteins in an organism or tissue. This includes post translational modifications e.g. phosphorylation and the effects they have on the system of study. Proteins interact with a large number of other molecules, in the case of enzymes, co factors assist approximately 45% of all reactions. The study of these molecular interactions in a high throughput manner also lies within the scope of modern proteomics. Additionally, new approaches to study protein structure and protein-protein interaction dynamics are becoming increasingly common to the field of proteomics (Rappsilber, 2011)

At the forefront of proteomics research is the mass spectrometer; a highly sensitive platform capable of identifying and quantifying thousands of proteins in a single experiment. The general principles of mass spectrometry (MS) are outlined in section 1.1.2 While proteomics measures biomolecules that are much closer to the biological phenomena witnessed under the microscope than for example mRNA, there are some trade-offs with its use. Notably the extremely complex downstream



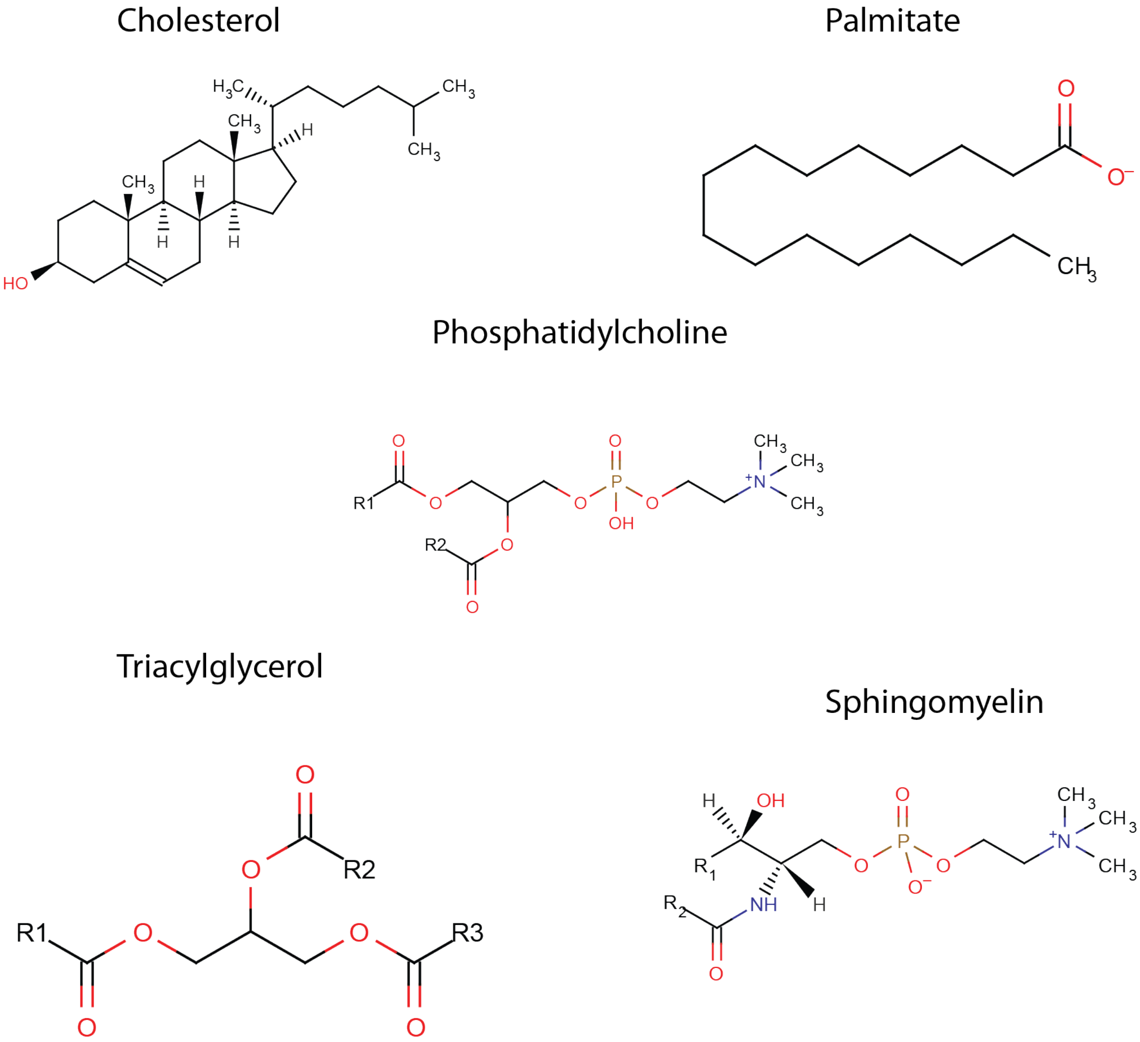
pKa Data: **CRC Handbook of Chemistry, v.** 2010
Dan Cojocari, Department of Medical Biophysics, **University of Toronto** 2011

Figure 1.1: A-D The twenty one amino acid side-chains grouped by their chenmical properties. E The structure of a generic peptide, R pseudo atoms indicate the position at which one of the twenty one side chains is attached.

data analysis: transforming raw spectra to peptide identifications and then to protein identifications. However proteomics remains a fast developing field building upon the bioinformatics and technological background of its sister fields. Proteomics trails behind genomics and transcriptomics fields largely due to the less restrictive nature of the instrument, for example, a microarray. A mass spectrometer is much more flexible in the number and types of biomolecules it can identify and the various compositions of the instrument e.g. FT-ICQ, triple quadrupole and Q-TOF vary considerably more than the various microarray platforms available. Access to organism specific protein sequence databases such as UniProt (UniProt-Consortium, 2012) is tightly integrated with the previously mentioned genomic sequence databases in section 1.1.1.1. Experimental data repositories also exist for mass spectrometry derived proteomics data. These share many of the core concepts for storage and public access to data with those mentioned in section 1.1.1.2. The main resources of this type are the PRoteomics IDEntifications database (PRIDE) (Martens *et al.*, 2005), PeptideAtlas (Desiere *et al.*, 2005) and the Global Proteome Machine DataBase (GPMDB) (Craig *et al.*, 2004). It is interesting to note that these three sister resources all originated from very similar time period due to neccessity in the field. Integrating existing ideas like experimental data repositories into other 'omics' fields, gives the opportunity to establish a concensus approach before there is an urgent need that results in multiple slightly differenet approaches. Protein-protein interactions data is generated in a high throughput manner by approaches such as the yeast two-hybrid system (Young, 1998). Knowledge of the protein-protein and protein-DNA interactions is crucial to a more detailed understanding of the proteome, above just the identifiaction of what proteins are present. Protein interaction data also has standardised data formats (Hermjakob *et al.*, 2004a) and resource to submit data and make it publically avaialble like IntAct (Hermjakob *et al.*, 2004b; Stark *et al.*, 2006).

1.1.1.4 Metabolomics

Metabolomics and proteomics share a lot of features in common due to their reliance on mass spectrometry as the high throughput platform of choice for detecting and quantifying biomolecules. Small metabolites are the focus of metabolomics and the detection of biologically relevant molecular fingerprints that result from some cellular processes. Mass spectrometry was pioneered in the small molecules field but has since fallen behind the curve in terms of technology development. This is largely the fault of a reduced focus on bioinformatics, particularly high throughput data analysis, storage and dissemination techniques (Wenk, 2005). As such it has a lot to learn from proteomics in particular but other 'omics' fields as a whole. A key area of metabolomics research is in the field of lipidomics: the portion of the metabolome dedicated to lipids, fatty acids and their derivatives. Lipids are a broad class of molecules with a variety of structures a selection of which are highlighted in Figure 1.2. They have wide ranging functions across the entire breadth of biological kingdoms, including roles in the structure of cell membranes, energy storage and cell signalling. The first lipidomics mass spectrometry publications emerged in 1994 (Han & Gross, 1994; Kim *et al.*, 1994) and have since gained considerable favour with 69 papers published in the first half of this year. This is partially due to the improved availability of reagents and synthetic standards necessary for quantitative lipidomics. However, integration with other fields and general appeal of the subject is still in a huge minority to genomics and proteomics, generating a comparative predicted 0.5% and 1% the number of citations respectively last year (Wenk, 2010). Metabolomic profiles offer detailed insight into the exact molecular state of a sample and have the ability to identify unique small molecules that define or contribute to a particular cellular phenomenon. Currently there is no published experimental data repository for metabolomics results like in other 'omics' fields, but if the other 'omics' field continue to be used as a template for metabolomics it is only a matter of time before one is released. However, there is a beta version of a metabolomics data repository avaliable from the EBI at `http://www.ebi.ac.uk/metabolights/`. Instrument vendors and wet lab scientists alike are not yet integrated in the standardised data approaches that make these resources easy to set up, slowing the fields progress. While small molecule reference databases do exist, the sparseness of them in comparison to this, the largest strata of biomolecules, is insufficient. In response small molecule databases have been developed for specific sub sets of small molecules that are interesting for a particular research focus. For the specific field of lipidomics the LIPID MAPS Structural Database (LMSD) (Sud *et al.*, 2007) is the current front-runner.



Cholesterol

Palmitate

Phosphatidylcholine

Triacylglycerol

Sphingomyelin

Figure 1.2: A small subset of lipid structures. Cholesterol is an important component of cell membranes, dictating fluidity and permeability of some ions. Palmitate is a common exit point of fatty acid synthesis by Fatty Acid Synthase (FAS) and constitutes a large proportion of the fatty acids incorporated into lipid structures. Phosphatidylcholine is another membrane lipid, often found on the exoplasmic leaflet and is responsible for membrane mediated cell signalling events. Triacylglycerol is a major dietary lipid, the degradation products of which are fatty acid species like palmitate, which undergo oxidation in the mitochondrial matrix to produce large amounts of energy. Sphingomyelin is the only phospholipid not derived from glycerol and is found in abundance in the myelin sheath that surrounds neuronal axons.

1.1.2 Mass spectrometry

For the purpose of completeness I will give a brief overview of the concept of MS because it features prominently in the subsequent chapters. However, for a more detailed description of 'omics' field specific applications of MS, chapters 2 and 4 describe a typical experiment in proteomics and lipidomics respectively. The fundamentals of mass spectrometry involve the measurement of mass-to-charge ratios of charged particles using controlled electrical fields. The process is broken down into the following steps:

Sample preparation The biological sample is mixed with a number of reagents to improve its behaviour when it enters the mass spectrometer. This involves everything prior to the sample actually entering the mass spectrometer.

Gas phase transfer Biomolecules embedded in solid media or solution enter the mass spectrometer and are transferred to the gas phase.

Ionisation Gaseous components of the sample are then ionised to produce charged particles. There are a number of ionisation techniques for producing both positively and negatively charged species. Choice of ionisation technique is largely influenced by the physical limits of the available mass spectrometer which is a decided upon based on what biomolecules are under investigation.

Separation Ions are separated in an electromagnetic field according to their mass-to-charge ratio. This part of the machine is often referred to as the analyzer, of which there are a number of specific implementations.

Detection Ions are detected and the count of each specific ion mass-to-charge ratio recorded.

Spectra generation Once the charged particles in a sample have been counted a spectrum is collated from the counts.

Interpretation Spectra must undergo some processing in order to translate the list of highly abundant mass-to-charge ratios into biomolecule identifications. This process is computationally complex and differs between the specific biomolecule types under study.

Prior to vaporisation a mass spectrometer coupled with reproducible front-end separation such as High Pressure Liquid Chromatography (HPLC) enables the simplification of complex samples by spreading the analytes over time. This enables the measurement of comparatively raw samples without the need for extreme sample preparation procedures that may impinge upon results. However, this does require specific ionisation techniques suitable for liquid samples such as electrospray ionisation (ESI). It is important to stress that mass spectrometry is an extremely diverse analytical platform with many different viable approaches, but the general principles outlined above remain the same; ionise, analyse and detect (Aebersold & Mann, 2003; Glish & Vachet, 2003).

1.2 Motivation

While instrumentation and data analysis have diversified between the various 'omics' approaches, they still share a lot of common ground. Each field has its own area of translatable expertise to be explored and cannot fairly be compared as neccessarily better than any other. However, it is clear that these fields have a lot to learn from one another, especially from the more mature fields such as genomics and transcriptomics. Rather than "reinventing the wheel" and developing successful aspects of a particular 'omics' field independently of similar work done in another field, ideas can be re-used. This applies most strongly to the portion of an 'omics' field attributed to bioinformatics, where reuse of well established concepts in one field and applying them to another field is not only time-saving but will create software with a similar user experience to that which already exists. As 'omics' fields become more integrated in their data processing workflows to unravel the wider effect of changes in one biomolecule to another, bioinformatics tools that integrate these fields must be established. It was with the motivation of finding examples of accepted solutions to bioinformatics problems, that can be translated to similar approaches in other 'omics' fields, that this PhD was undertaken.

1.3 Goal statement

The specific aims of this work involve three main areas of investigation. Firstly, the quality control (QC) of publicly available proteomics data. The aim of which is to assess the possibility of aggregating datasets and reusing them to answer general questions about the field of proteomics. QC is a much more prominent feature of genomic and transcriptomic data processing pipelines. These existing approaches can lead to adapted versions for proteomics, especially the style in which the approaches are represented e.g. similar graphical outputs and familiar summary statistics. Genomics has a large community of bioinformaticians built around the Bioconductor (Gentleman *et al.*, 2004) project that has yet to be fully embraced by the proteomics community but could be used to serve a similar purpose. Resources such as the Protein Data Bank (PDB) are well known for the extreme scrutiny new data undergoes when it is made publically available, with third party tools designed specifically for re-analysing the whole database (Joosten *et al.*, 2012); a rarity in the field of proteomics. While high throughput mass spectrometry based proteomics is a much newer and smaller field, age and size of the field are not the only explanatory variables for the relatively undeveloped state of proteomics quality control, the heterogeneity of the instrumentation and protocols plays a major role. Proteomics represents a considerable challenge in this respect, with so many instrument manufacturers producing proprietary data formats, followed by a further explosion of downstream data analysis software. Writing generic compatible QC pipelines is extremely challenging and is often a case of trying to serve as many people as possible, while appreciating there are many edge case that will remain unsupported. The problem is compounded in the context of public data, as explanatory metadata of how results were generated is often missing. The problem is rooted in the very early diversification of proteomics in comparison with transcriptomics, where competitive platforms do exist but the flexibility of the instrumentation (microarrays) does not allow for such varied experiments. Proteomics effectively diversified at a rate faster than the data standards and standardised analysis community could establish itself and accommodate, resulting in a disproportionately more difficult task. Complementary to the work on QC of proteomics data, insights into the diverse nature of the data will impact upon the curation of public data and highlight common problems in reusing these for new purposes.

Secondly, taking inspiration from the field of proteomics, in particular the sequence database UniProt (UniProt-Consortium, 2012), the concept of a database of theoretically generated biomolecules, will be ported to the field of lipidomics where a similar resource does not currently exist. The approach will involve a set of rules that define the chemical space in which lipids reside, e.g. maximum fatty acid chain lengths, number and position of double bonds and the initial categories of lipids to develop the service with. This differs from the approach in UniProt, which does not enumerate all theoretical proteins from the set of 21 amino acids between a feasible minimum peptide length and maximum peptide length. Instead, proteins are theoretically translated from the genome. The prospective proteins are clearly distinguished from experimentally validated proteins. It is this separation of the theoretical from the validated and the UniProt approach to supplementary metadata that will translate into a novel resource for the field of lipidomics.

The final aim of this work is the statistical analysis of human colorectal cancer lipidomics data. The dataset is unique in the level of detail measured of the tumour and normal lipidome and the large number of patient samples. However, analysis of this scale of lipidomic data is in its infancy with little existing literature available on the role of the lipidome in cancer or datasets of this size being publically available. Guidance will need to be taken from analogous endeavours in fields like genomics and proteomics, where there is some indication on the statistical rigour required to mine such data sets.

1.4 Thesis outline

The remainder of the thesis is split into three research chapters, a final conclusion about the significance of the overall results and a set of appendices that contain useful complementary information about the work carried out. Chapter 2 describes work on the QC of publicly available proteomics data extracted from the PRIDE database, specifically drawing on the more experienced field of QC in transcriptomics. Chapter 3 relates the integral nature of sequence databases to the field of genomics and proteomics and applies it to the field of lipidomics. The design and implementation of a web application and database of theoretical lipid species called ‘LipidHome’ is described. Chapter 4 continues on the work in the field of lipidomics, analysing a clinical lipidomics dataset of human colorectal cancer samples. Using standardised approaches to data storage and representation pioneered in Chapter 3 the data is mined for biological insight into the progression of cancer and its effects upon the lipidome. Chapter 5 concludes with some general remarks on the success of the work done and the prospect for its continued contribution to the field of proteomics and lipidomics in the future.

Chapter 2

Quality Control of Public Proteomics Data

The field of proteomics is expanding at an incredible rate (Csordas *et al.*, 2012), what was once the pioneering work of a few labs is now a huge industry of technology development (Anonymous, 2007; Armirotti & Damonte, 2010; McLafferty, 2011; Rabilloud *et al.*, 2010), basic biology research (Solit & Mellinghoff, 2010) and clinically relevant findings (Latterich & Schnitzer, 2011; Sigdel & Sarwal, 2011). The experimental diversity of the field has been no small challenge for the groups responsible for standardizing the data and storing it in an accessible and intuitive way (Craig *et al.*, 2004; Desiere *et al.*, 2005; Vizcaíno *et al.*, 2009). Multiple data formats have arisen from specific instrument vendors which has necessitated the creation of standard data formats such as mzXML (Pedrioli *et al.*, 2004) and mzML (Martens *et al.*, 2011) which can be read and processed by everyone without the need for proprietary software. Public proteomics repositories have played a key role in the cultivation of the field, allowing easy access to proteomics datasets, from the raw spectra all the way up to peptide identifications and most recently the quantitative estimates of protein abundance. However, actual submission to these services still only represents a tiny fraction of the actual data that is being generated and published in the wider proteomics community. With a greater focus on the clinical and biological relevance of studying the proteome, journals are beginning to mandate the public deposition of datasets to services such as PRoteomics IDEntifications database (PRIDE) prior to review. In order to realise the full potential of public proteomics data, simple storage of it must be supplemented with large scale analyses of general properties of proteomics data in parallel to the particular biological questions that individual studies were designed for. This chapter describes work designed to evaluate the inherent quality of public proteomics data with regard to its potential for re-use.

2.1 Introduction

Quality Control (QC) has wide ranging connotations within the biological sciences with regard to its scope, but QC is a continuous process that should be integral to a work flow from its conception through to its execution. Firstly, a system must be checked for basic technical functionality. When running calibrants on the various components of a whole system, experimentalists must ask the question "are the results reproducible and can the variance in the results be explained and possibly diminished?". Secondly, pilot experiments must be run to explore the variability of the system intended for measurement; "how many technical replicates are required to fairly sample the variation in the system?". Penultimately, "can the results from the pilot be reliably reproduced within the laboratory on the larger scale required for the study?". Finally, the results must be confirmed to be of biological origin either by some orthogonal technology or confirmation by an external laboratory. This sort of reproducibility has been achieved in proteomics studies such as The HUPO Test Sample Study (Bell *et al.*, 2009) where 27 laboratories took part in the attempted identification of an equimolar sample containing 20 proteins. Initiatives such as 'Fixing Proteomics' (`www.fixingproteomics.org/`) offer a portal of information for co-ordinating the design of experiments and making use of state of the art protocols that have been validated by multiple labs around the world to deliver reproducible and statistically reliable results. While proteomics is the focus of this chapter's research, the state of quality control in the field of genomics makes a valuable comparison from which lessons can be learnt and successful approaches applied to the fledgling field of proteomics QC. Scientists within the field of genomics have known for a long time that the quality of expression data must be carefully assessed over the course of an experiment (Ji & Davis, 2006) and have developed appropriate infrastructure for standard analysis tools and storage of data. Open source software frameworks like Bioconductor (Gentleman *et al.*, 2004) have been a breeding ground for standardised genomics data analysis protocols and QC checks of genomic data. Now in its 10th year of development Bioconductor has a significant market share of genomics data analysis and has built a community of interested wet-lab scientists and bioinformaticians to propagate the standardised analysis of not only genomics data, but increasingly also of other fields such as proteomics (Scheltema *et al.*, 2011) and metabalomics (Benton *et al.*, 2008; Smith *et al.*, 2006). In addition to analysis, the presentation and storage of public genomics data is a well established aspect of the field where specific repositories provide datasets of confirmed levels of quality suitable for reuse (Parkinson *et al.*, 2011). This type of service can be directly applied to the field of proteomics where a set of core experiments of known quality can be re-analysed to compare findings across datasets or large amounts of data aggregated with a view to find general properties of proteomics data.

2.1.1 A typical proteomics work flow

Before beginning an investigation into the inherent quality of proteomics data, it is imperative to understand the complex nature of a proteomics experiment and the potential sources of irreproducibility. Experimental diversity in proteomics is one of the many reasons it has so much to offer the scientific research community and while innovation in sample preparation and instrumentation are still being explored, the community has settled on mass spectrometry as the technology of choice for identifying proteins in a high throughput manner. The alternative is sequencing via Edman degradation, in which the N-terminal amino acid is detached from the peptide and identified by chromatography. Edman degradation is rarely used in modern proteomics labs and cannot be considered high throughput. Mass spectrometry based work flows break down into a number of core steps (* denotes an optional step) each of which is a potential checkpoint for quality control issues.

Sample acquisition When acquiring a sample of biological origin it is of critical importance to respect the environment from which it came. The proteome is dynamic and any extended period of sample handling or incorrect sample storage will detract from the biological phenomenon under study.

Sample preparation Raw biological samples rarely contain a concentrated representation of the proteins of interest, these are more typically mixed with other components of the cells such as phospholipids and nucleic acids. During sample preparation the protein component of the raw sample is boosted by removing contaminants and by solubilising the proteins. These preparation protocols and the chemicals involved vary a great deal depending on the raw sample being worked upon. Special care must be made to identify any preferential stabilisation of certain proteins or likely non-native chemical modifications forced upon the proteins during this process.

Protein pre-fractionation* High molecular weight proteins, highly abundant proteins and phosphorylated proteins can all be enriched at this stage depending on the aim of the study. It is at this point in a protocol that a single raw sample begins to multiply into many distinct protein fractions with particular properties.

Protein separation* Once the proteins of interest have been enriched they may then be separated further to reduce the complexity of the sample. This is typical of gel based approaches, where proteins are separated in one or two dimensions by their molecular weight and/or isoelectric point. For targeted experiments, individual protein spots or bands can be excised from the gel and independently analysed in the downstream stages. State of the art technology and less targeted ‘shotgun proteomics’ approaches now largely negate the need for the protein separation step due to much improved High Performance Liquid Chromatography (HPLC) at the peptide separation level.

Digestion Proteins then undergo a process of digestion by proteolytic enzymes. Trypsin is most commonly used due to its favourable cleavage properties, cutting proteins into peptides terminating in lysine or arginine residues which are readily ionised and detected by a mass spectrometer. Other proteolytic enzymes are sometimes used in niche situations but the stability, availability, cleavage properties and documentation available on trypsin make it by far the most popular. It is important to note that cleavage is not always 100% complete and downstream peptide identification should account for this. Heavy

Oxygen atoms may also be incorporated into peptides at this stage to differentially label a sample by performing the digestion in heavy water.

Peptide fractionation* Similar to protein fractionation, peptides may also be enriched for particular properties such as phosphorylation, glycosylation or containing a particular amino acid residue. With each fractionation step material is lost and incomplete separation must be accounted for. This stage also offers a last chance to modify the peptides with quantitative labels such as iTRAQ (Ross *et al.*, 2004) or TMT (Thompson *et al.*, 2003) reagents.

Peptide separation Prior to mass spectrometry the effective complexity of the peptide sample must be further reduced, a step usually carried out by HPLC. State of the art HPLC separates peptides based largely on their hydrophobicity. These HPLC columns are coupled directly to the mass spectrometer, the elution gradient and flow rate through them defining the rate at which peptides enter the mass spectrometer to be analysed. The low flow rate of nano-HPLC columns have revolutionised mass spectrometry allowing a direct interface to an Electrospray ionisation (ESI) ion source and the high through put analysis of samples. HPLC columns are typically re-used for multiple runs, so careful control should be kept on recording the performance of the column, recognising the first signs of degradation.

Mass spectrometry Mass spectrometers are complex multi-component systems, the general principles behind them being ionization (often ESI), separation (mass analysis), and detection. Additionally ions may undergo several rounds of separation and fragmentation prior to detection to gain more structural detail on a particular ion. Only ionized particles can be detected in a mass spectrometer and it is the mass-to-charge (m/z) that is recorded in a quantitative manner as intensity by the detector. A typical proteomics experiment can produce anywhere in the region of 100 to 100,00 spectra. With such a complex system the possible combination of instrument parameters is huge. Consequently, machines should regularly be tested for expected behaviour across a range of parameters.

Peptide identification Spectra undergo a complex routine of identification by a piece of software known as a database search engine e.g. MASCOT (Perkins *et al.*, 1999). The search engine correlates the experimental spectra with theoretical spectra of peptides to assign a best match for each spectrum. Quality metrics are provided with peptide identifications to remove false positives or spurious hits. Search engines take many parameters to optimise their performance such as defining known modifications and the accuracy of the instrument. Incorrect search engine parameters will result in improbable identifications with high scores or a severe lack of identifications from a dataset that would otherwise perform very well.

Protein inference In the shotgun approach peptide identifications are then aggregated and proteins are inferred from the peptide list. The process is complicated by the fact that a peptide sequence may originate from multiple distinct proteins, this must be accounted for in order to report the most likely proteins in the sample. In these circumstances there are multiple methods for reporting proteins. Either the entire set of possible proteins that contain the detected peptides, a single protein for each peptide or a flavour of the minimal set that accounts for the coverage/evidence of a protein when selecting the most likely protein to be represented by a peptide (Martens & Hermjakob, 2007). This is not a problem in top-down proteomics where whole proteins are ionised and then subject to several rounds of fragmentation and detection. However top-down approaches suffer unqiue challenges in the interpretation of spectra, particularly the estimation of charge state, which impacts heavily upon the identification process.

With so many distinct steps involved in a proteomics experiment combined with such sensitive samples and equipment, it is extremely important to appreciate the effect that any small alteration in protocol may have on the results.

2.1.2 Quality control of public data

In line with the deliverables of the PRIDE (Vizcaíno *et al.*, 2009) project at EBI and in response to the European Science Foundation Quality Control in Proteomics

Workshop, 2010 in Cambridge UK, this work was undertaken to investigate generic quality control markers of publicly deposited proteomics data to inform the future experiment filters that will define the experiments for inclusion into the proposed PRIDE Q database (Vizcaíno *et al.*, 2010). A paper was produced and published (Foster *et al.*, 2011) on these results from which I will regularly quote verbatim throughout the remainder of this chapter. PRIDE was chosen as the public proteomics data source due to its wide coverage of many different organism proteomes, most heterogeneous experiment protocol/instrumentation combinations. Most importantly, its superior support for metadata annotations of experiments that will be critical later on in the investigation to find the cause of any data inconsistencies.

2.2 Materials and Methods

A number of data retrieval, transformation and processing techniques were necessary to fulfill the very broad aims of this research to suggest tools to measure the inherent quality of proteomics datasets. Below the main algorithms are described for future reference in the results section.

2.2.1 Data retrieval from PRIDE

Prior to analysis, the appropriate data must be exported from the PRIDE database. To facilitate this process an R library was developed to retrieve comprehensive spectra objects that included a selection of metadata, spectral data and peptide identification data for MS2 spectra. This library is used repeatedly throughout the analyses, data is retrieved using the readPrideExperimentFunction() with a vector of PRIDE experiment accessions as the only parameter. The spectra object structure is as follows:

spectra$spectrumID$accession The experiment accession associated with the spectrum, used when retrieving multiple experiments data at once from the database.

spectra$spectrumID$preMZ Precursor ion m/z.

spectra$spectrumID$charge Precursor ion charge.

spectra$spectrumID$mzI The spectrum peak list, a two column matrix of m/z and intensity respectively.

In addition, a number of plotting functions are provided in this library to work with the experiment object and present an easy interface to visualise the data, including:

plotNPeaksVsTime() A spectrum-series plot of the peak count per spectrum. This assumes that the order MS2 spectra were deposited in the database is identical to the order they were acquired by the mass spectrometer.

plotPeakHistogram() Plots a histogram of binned total peak count per spectrum, the bins are as follows: 0-10, 11-50, 51-100, 101-200, 201-500, 501-1000, 1000+.

plotChargeBarPlot A bar plot of precursor ion charge frequency of all MS2 spectra in the spectra object.

plotMZHistogram() A histogram of 100 equal width m/z bins in the range 0-3000 of the combined MS2 spectra peaks.

plotAverageIntensity() A spectrum-series plot of log10 mean spectrum intensity for all MS2 spectra. This assumes that the order the MS2 spectra were deposited in the database is identical to the order they were acquired by the mass spectrometer.

plotAverageMZ() A spectrum-series plot of log10 mean spectrum m/z.This assumes that the order the MS2 spectra were deposited in the database is identical to the order they were acquired by the mass spectrometer.

plotPrecursorMZvsTime() A spectrum-series plot of the precursor ion m/z.This assumes that the order the MS2 spectra were deposited in the database is identical to the order they were acquired by the mass spectrometer.

plotPrecursorMZHistogram A histogram of 100 equal width m/z bins in the range 0-3000 of the combined precursor ion m/z of all MS2 spectra.

2.2.2 Latent semantic analysis

The protocol followed here for latent semantic analysis (LSA) follows the approach used by Klie et al (Klie *et al.*, 2008). Concretely identified peptide sequences are extracted from each 'PRIDE experiment', and used to generate a peptide *versus* PRIDE experiment occurrence matrix. In the case of the HUPO PPP2 dataset to which this is applied each PRIDE experiment represents a single SCX fraction, and so the peptide *versus* experiment occurrence matrix is actually equivalent to a peptide *versus* SCX fraction occurrence matrix. Term Frequency-Inverse Document Frequency (tf-idf) is subsequently applied to this matrix in order to attenuate the signal derived from high abundance peptides, thus allowing for greater sensitivity of effects produced by less abundant peptides. This method of weighting the frequency data is summarised by the set of formulas:

Term frequency Where $n_{i,j}$ is the number of occurrences of the peptide (t_i) in experiment d_j, and the denominator is the sum of the number of occurrences of all peptides in the experiment d_j (equation 1).

Inverse document frequency A measure of the general importance of a peptide calculated by taking the logarithm of the total number of experiments $|D|$ divided by the number of experiments containing the peptide (equation 2).

Term frequency inverse document frequency The product of the two previous expressions for each peptide in each experiment (equation 3)

1. Term frequency: $\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$

2. Inverse document frequency: $\mathrm{idf}_i = \log \frac{|D|}{|\{d : t_i \epsilon d\}|}$

3. Term frequency inverse document frequency: $\text{tf-idf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$

An iterative round of k reduction and LSA is carried out on the tf-idf weighted occurrence matrix, resulting in a signal boost by deriving a denoised and less sparse occurrence matrix. The selection of k is largely an arbitrary decision based loosely around defining the number of expected 'concepts' or clusters in the data, but the iterative approach allows visualisation of multiple values of k to estimate which produces the clearest polarisation of signal. This process effectively allows the inference of a peptide's frequency based on the frequency of other peptides present in the experiment and the co-occurrences of those peptides in other experiments. Based on the LSA transformed occurrence matrix, all possible experiment *versus* experiment distances are calculated using the cosine similarity function between each pair of experiment eigenvectors.

$$\text{similarity} = \cos(\Theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

The cosine similarity matrix can be plotted as a heat map to visualise the relationship of experiments to one another. The R implementation of LSA from the 'lsa' package provides a simple to use and accurate LSA algorithm. However, its accompanying package for plotting the heat map is no longer supported under the latest version of R. As an alternative the levelplot() function in the 'lattice' package is equivalent.

2.2.3 Tryptic missed cleavage rate

Proteins are cleaved into peptides by a set of enzymes called proteases, each has a specific amino acid recognition sequence that is cleaved. Occasionally a cleavage site is not cleaved, the resulting peptides contain a missed cleavage. Trypsin is the most commonly used protease, a tryptic missed cleavage is defined as a non-C-terminal arginine or lysine residue not directly preceded by a proline residue. The missed cleavage rate is the number of missed cleavages divided by the number of successful cleavages (each peptide is considered a successful cleavage). This does not account for C-terminal peptides and so the missed cleavage rate can be considered slightly conservative.

2.2.4 Empirical background distributions

The concept of an empirical background distribution against which a single experiment property, e.g. missed cleavage rate is compared (section 2.2.3), is applied throughout these analyses. An empirical background distribution is created for a metric by selecting a set of similar experiments of which the experiment of interest is a member. For each of these reference experiments the distribution of the metric is calculated and normalised, typically between zero and one using min-max normalisation. The normalised metrics of the reference experiments can be plotted as a box plot to show the mean value of the metric and the inter quartile range. The experiment of interest can then be compared to this global estimate of the metric and conclusions be drawn. As an example the MS2 m/z distribution for a set of experiments can be approximated by individually calculating the frequency of binned m/z values for each experiment. After normalisation of each experiment by dividing bin frequency by the sum of the bin frequencies. Each bin has a number of normalised values equal to the number of experiments, a box plot can be constructed from these values. A box plot for each bin is created and plotted in bin order to reveal the empirical background distribution of MS2 m/z frequency.

2.2.5 MS2 m/z delta

For a spectrum the top 10% most intense m/z fragment ions are selected. An m/z distance matrix is calculated, the Δ of each m/z ion with each other is flattened into a vector. This process is repeated for all the spectra in a PRIDE experiment. The vectors are concatenated and the Δ m/z values separated into one Dalton wide bins. The frequency of ions in each one Dalton bin is calculated and then plotted as a bar plot. Assuming the experiment has detected peptides, it is the expectation of the experimentalist that the most intense peaks belong to peptide sequence ions. By extension the Δ m/z between sequence ions ought to be the m/z of a single amino acid or short sequence of amino acids. Plotting the histogram of Δ in the region of 40-200 m/z spans the entire range that the +1 charged amino acid m/z lie in. It is expected that the bins corresponding to an amino acid +1 charge m/z are much more frequent than the other masses which are considered to be noise, or higher mass and charge Δ m/z that occur at a much lower rate.

2.2.6 Multi dimensional scaling

Multi dimensional scaling (MDS) is a powerful technique for reducing high dimensional data into fewer dimensions, often two or three to allow visualisation as a graph. For a set of vectors a dissimilarity matrix is constructed between them (in this case euclidian distance). This dissimilarity matrix is input for MDS (in this case metric MDS) which finds an optimization for faithfully translating the distances between elements while collapsing the data into two dimensions. This functionality is provided in the base package of R as the cmdscale() function, which takes the distance matrix and number of required dimensions as input parameters.

2.2.7 PRIDE Inspector

Alongside R, the main piece of software used to complete this work was the PRIDE Inspector tool (Wang *et al.*, 2012). A graphical user interface to the PRIDE public MySQL instance, it offers direct access to any public experiments. Experiments can be browsed by high level metadata e.g tissue type or specific experiments be selected by their PRIDE experiment accession. The tool provides three main views of the data: protein view, peptide view and spectra view. Depending on the interest of the user, any particular aspect of the data can be visualised in the context of its most relevant supporting data. A general overview of the experiment is summarised into three parts: “Experiment General”, “Sample and Protocol” and “Instrument and Processing”. Most relevant to this work are the “Experiment Summary Charts”, developed as a consequence to this work but also in tandem. These charts offer insight into the suitability of the metadata annotation and any potential problems between data acquisition, processing, conversion and submission to the database. The tool is available to download at `http://www.ebi.ac.uk/pride/webstart/pride-inspector.jnlp`.

2.3 Results

2.3.1 Depletion and Separation Analysis

Various types of liquid chromatography (LC) form key steps in most proteomics experiments. Reproducible separations are essential for several quantitative or targeted approaches, where alignment of chromatographic profiles or correct scheduling of peptide elution times are paramount to success (Bondarenko *et al.*, 2002; Chelius & Bondarenko, 2002; Schiess *et al.*, 2009). In order to assess chromatographic performance and reproducibility across many different elution runs, I analysed the Human

Proteome Organisation's Plasma Proteome Project 2 (HUPO PPP2) datasets (Liu *et al.*, 2006; Qian *et al.*, 2008) deposited in PRIDE (Vizcaíno *et al.*, 2009), since these contain several similar chromatographic runs, performed on different peptide subsets obtained from human plasma. The experiments and peptide identifications from this submission were performed by the Richard Smith Lab at Pacific Northwest National Laboratory, submitted to PRIDE under accession numbers 8172 to 8544 (inclusive) and retrieved from the public PRIDE MySQL instance. These 373 experiments represent the analysis of twelve human plasma samples, each subjected to some combination of MARS-6 or IgY-12 depletion and cysteine or N-glycosylated peptide selection prior to offline strong cation exchange (SCX) chromatography followed by LC-MS (for full protocol details please see (Liu *et al.*, 2006; Qian *et al.*, 2008)). Each PRIDE experiment accession number represents the analysis of a single SCX fraction. Based on the identified peptides in each PRIDE experiment, a peptide *versus* PRIDE experiment occurrence matrix was then constructed. In order to accommodate the differences in experimental design that preceeded the chromatography runs and that influenced the selection of proteins and peptides, latent semantic analysis (LSA, see section 2.2.2) was employed as a powerful signal booster and noise filter prior to experiment cosine similarity calculation. The experiments, each corresponding to an SCX fraction, were grouped by depletion technology and peptide fractionation method, and then ordered by their elution order. The resulting peptide-based experiment *versus* experiment similarity matrix is provided graphically as a heat map in Figure 2.1.

i A region of high similarity between the later SCX fractions in the first 'MARS Cysteine' analysis. It appears that peptides have begun to bleed across the fractions from a certain time point, indicating possible problems with the gradient or the column.

ii A highly similar smearing effect is seen for an independent sample. Furthermore, throughout the plot, various levels of such smearing (black to red blurring) can be seen along the diagonal, possibly indicating progressive evidence of SCX column degradation.

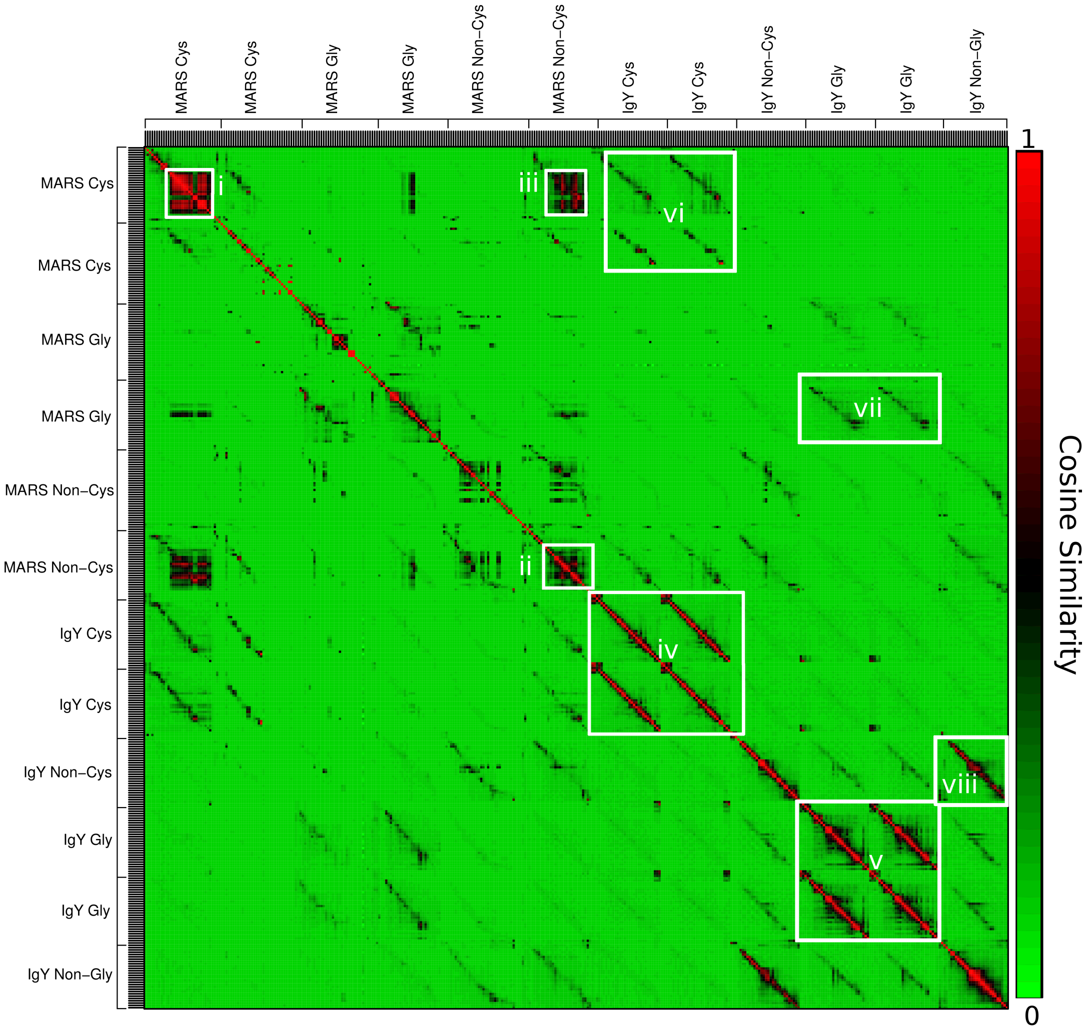


Figure 2.1: From the HUPO PPP2 data submission by the Richard Smith Lab at Pacific Northwest National Laboratory all 373 experiments (accession numbers 8172 to 8544 inclusive) were extracted from PRIDE. Each PRIDE experiment represents a single LC run of an SCX fraction, performed upon a sample that had undergone a combination of either cysteine or N-glycosylated peptide selection and either MARS-6 or IgY-12 depletion. An experiment-peptide occurrence matrix was subjected to the denoising algorithm Latent Semantic Analysis and subsequently transformed into an experiment correlation matrix.

iii Worryingly, this shows that there is a clear similarity in detected peptides between the first 'MARS Cysteine' sample and the second MARS Non-Cysteine sample. This contradictory similarity indicates a problem with the peptide selection protocol, as it should ideally result in two fully distinct peptide subsets. On closer inspection of the offending overlapping peptides we found that the majority are attributed to the various isotypes of immunoglobulin, thus indicating that the selection procedure specifically fails for highly abundant proteins. This demonstrates that slight deviations from perfect peptide selection efficiency can quickly result in substantial carry-over of undesired peptides for abundant proteins. Of course, high remaining levels of immunoglobins might in turn indicate possible problems with the MARS depletion (which should include affinity binders for the depletion of albumin, transferrin, haptoglobin, IgG, IgA, and alpha-1 antitrypsin) for these samples. This may be explained by overloading of the SCX column, in which too much peptide material exposed to the solid phase and binding site saturated. This saturation resulted in improper separation at the SCX level. Alternatively this could be a problem at the HPLC stage where previous SCX fraction were not washed from the column and eluted over the course of several subsequent runs of adjacent SCX fractions.

iv & v The excellent reproducibility of well-executed separations for identical depletion protocols is shown in these highlights, with equivalent SCX fractions yielding largely the same peptide identifications (hence the strongly red off-diagonal lines).

vi & vii Reproducibility across different depletion methods can also be found, albeit at less intensity, in these highlights where similar peptides were consistently identified regardless of the protein depletion strategy.

viii Unexpectedly, the peptides identified in 'IgY Non- Cysteine' and 'IgY Non N-Glycosylated' samples also show high similarity in this highlight. It is unclear why this similarity is so pronounced as it is unlikely that non-cysteinyl peptides are preferentially N-glycosylated.

2.3.2 Proteolytic Digestion and Precursor Mass Analysis

Between the steps for protein depletion and peptide selection, proteins are enzymatically cleaved into peptides, typically using the endoproteinase trypsin. However, trypsin digestion is not always completely efficient (Yang *et al.*, 2010), resulting in missed cleavages and therefore slightly longer than average tryptic peptide populations in most datasets. Two methods are suggested here for quality controlling the digest efficiency; analysis of the distribution of missed cleavages in the resulting peptides (see section 2.2.3) and comparison of the distribution of precursor ion masses to those empirically derived from a large set of representative experiments in PRIDE (see section 2.2.4). Note that the former introduces an additional dependence on the search engine used for identification, along with its parameters, while the latter is independent of the search engine. In order to generate a reference for the efficiency of a tryptic digest, all PRIDE experiments annotated to have a single digestion step involving trypsin were extracted from PRIDE along with their peptide identifications. This resulted in 4582 experiments, 1695 of which were discarded due to no missed cleavages being detected in those experiments, indicating that the search engine was configured to not tolerate any missed cleavages. The remaining 2887 experiments were used to estimate the overall background missed cleavage frequencies, and precursor ion mass distributions. In order to derive an empirical background for the frequency of occurrence of missed cleavages, the number of missed cleavages were counted for each peptide identification in an experiment, and the rate of missed cleavage for an experiment was then calculated as the number of missed cleavages over the total number of observed cleavages (each peptide terminus is considered a correct tryptic cleavage). The resulting distribution of missed cleavage rate, shown in Figure 2.2, then provides a useful empirical background to measure individual experiments against.

In Figure 2.2 we see two experiments retrieved from PRIDE and their missed cleavage rates, PRIDE accession 12914 shows approximately a 100% efficient digest while, experiment 12152 shows a missed cleavage rate of near 1 and hence 50% of all potential cleavages in the identified peptides were missed. Interestingly, closer inspection of the protocol employed for experiment 12152 indicates that free amines had been acetylated, thus excluding lysines from tryptic cleavage. As a result,

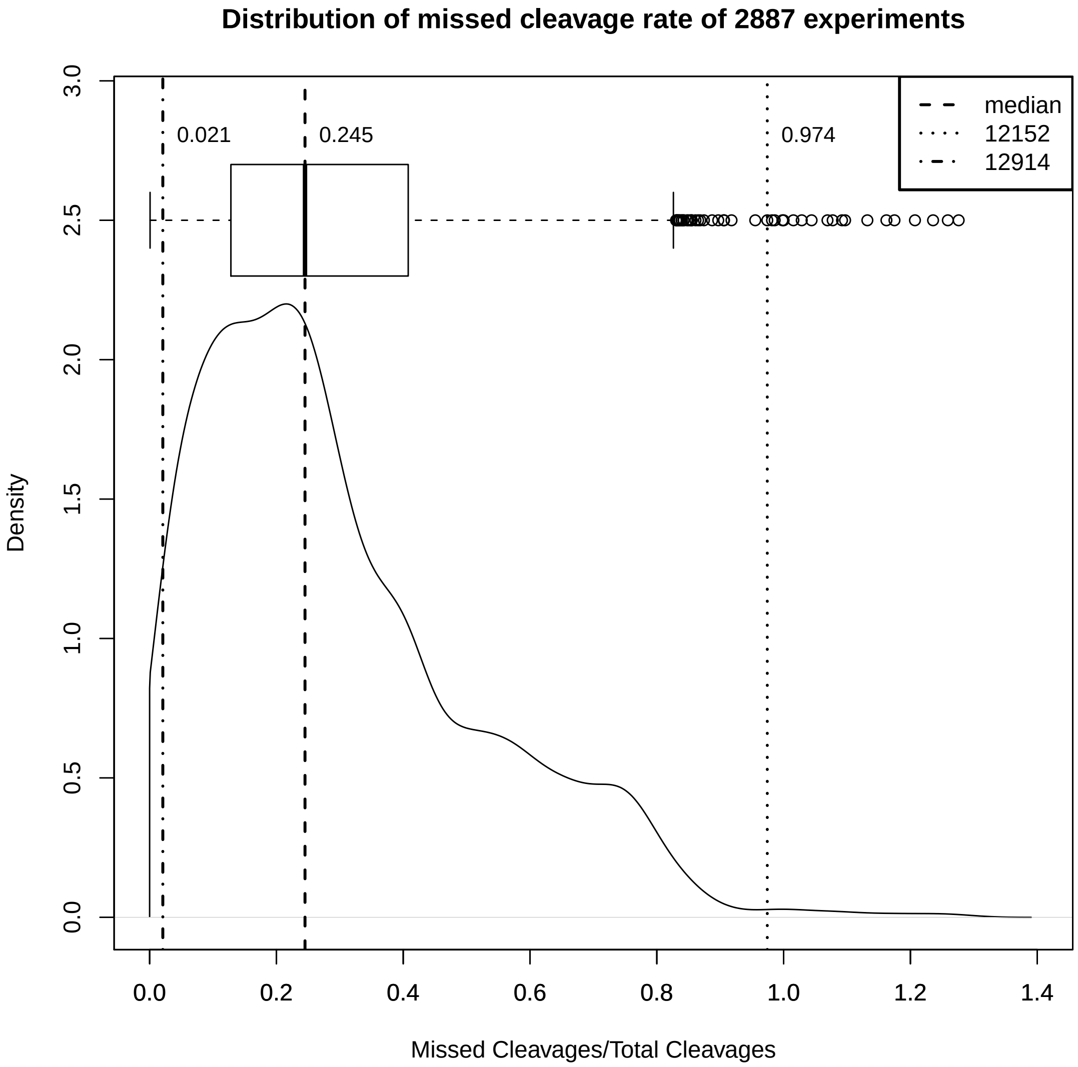


Figure 2.2: From PRIDE 2887 experiments are retrieved with a custom developed R library. For each experiment the peptide identifications are extracted and the missed cleavage rate calculated as the total number of missed cleavages (occurrence of an intra-peptide lysine/arginine residue not proceeded by proline) divided by the correct number of cleavages (each peptide is considered the result of a correct cleavage, there is no correction factor for terminal tryptic peptides.)

the high missed cleavage rate is in fact an expected effect of the protocol, and the experiment's status as an outlier in this case indicates an interesting case for data annotation as opposed to a QC issue. A background distribution can also be constructed from the precursor ion masses obtained from the spectra in each experiment. Such a background distribution can also be used in quality control as deviations in precursor masses are often indicative of issues with the data. In order to obtain the background distribution, the precursor ion masses per experiment were grouped into 60, 100 Da wide bins and the contents of each bin were then normalised by the total number of peptides. Since the bins were kept constant between the different experiments, a box plot could be created reflecting the distribution of normalised frequencies across all experiments for each bin. To demonstrate the usefulness of this test, the mass distributions from individual experiments of the HUPO Test Sample Study (Bell *et al.*, 2009) (PRIDE accession numbers 8130-8158) were compared to the local empirical background of this entire study of 27 samples. These experiments were chosen because the HUPO Test Sample Study was carried out on a single sample consisting of 20 equimolar proteins by a combination of 27 individual laboratories and instrument vendors worldwide. Furthermore, these 20 proteins were carefully chosen to yield peptides that reflect overall properties of the human proteome as best as possible (Bell *et al.*, 2009). An important detail is that individual peptide identifications are not listed for these experiments; only mass spectra and proteins are provided. As such, only an indirect measure such as the precursor ion mass used here, can be employed for such data.

Figure 2.3 correspondingly shows that the experimental precursor mass distribution from PRIDE experiment 8145 closely mimics the empirically derived precursor mass distribution for tryptic digests, with very little difference in mean precursor ion mass. Clearly seen in figure 2.4, the mass range 2000-3000 shows a larger number of identifications in PRIDE experiment 8146 than the empirical background distribution, suggesting the use of specific protocols or methods that would produce such deviating data. These could include a less efficient tryptic digest that resulted in many missed cleavages and hence longer, high mass peptides. Figure 2.5 on the other hand shows a precursor mass distribution from PRIDE experiment 8155 that lies to the left of the expected distribution, indicating that this analysis was particularly sensitive to smaller precursors. Yet with these deviations in hand, a mechanistic

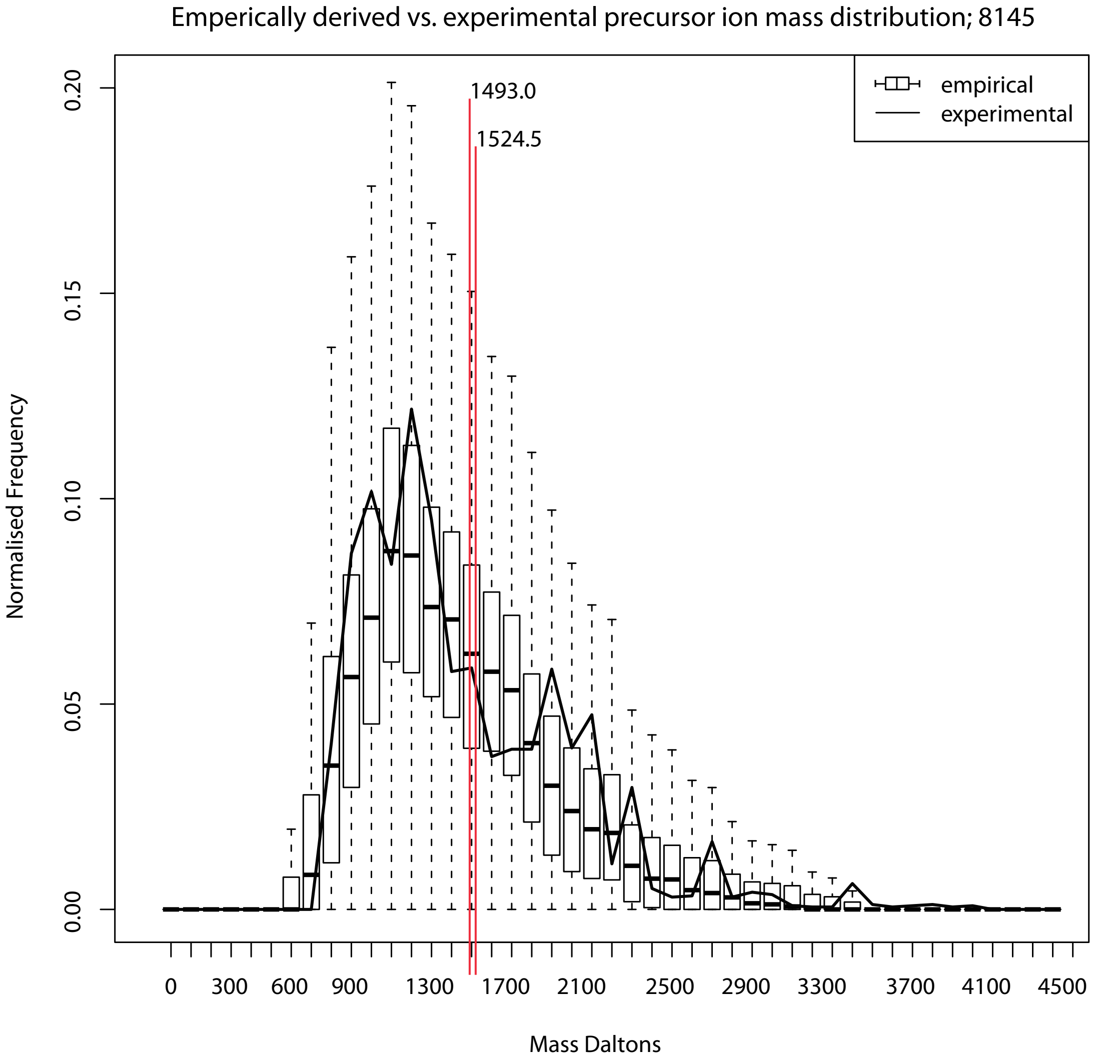


Figure 2.3: The precursor ion mass distribution of PRIDE experiment 8145 overlaid on the empirical precursor ion distribution of all samples within the HUPO Test Sample Study. The mean precursor ion mass (Daltons) is not significantly different between experiment 8145 and the empirically derived background. Hence, the detected peptide masses and by proxy, the tryptic digest efficiency are of expected quality.

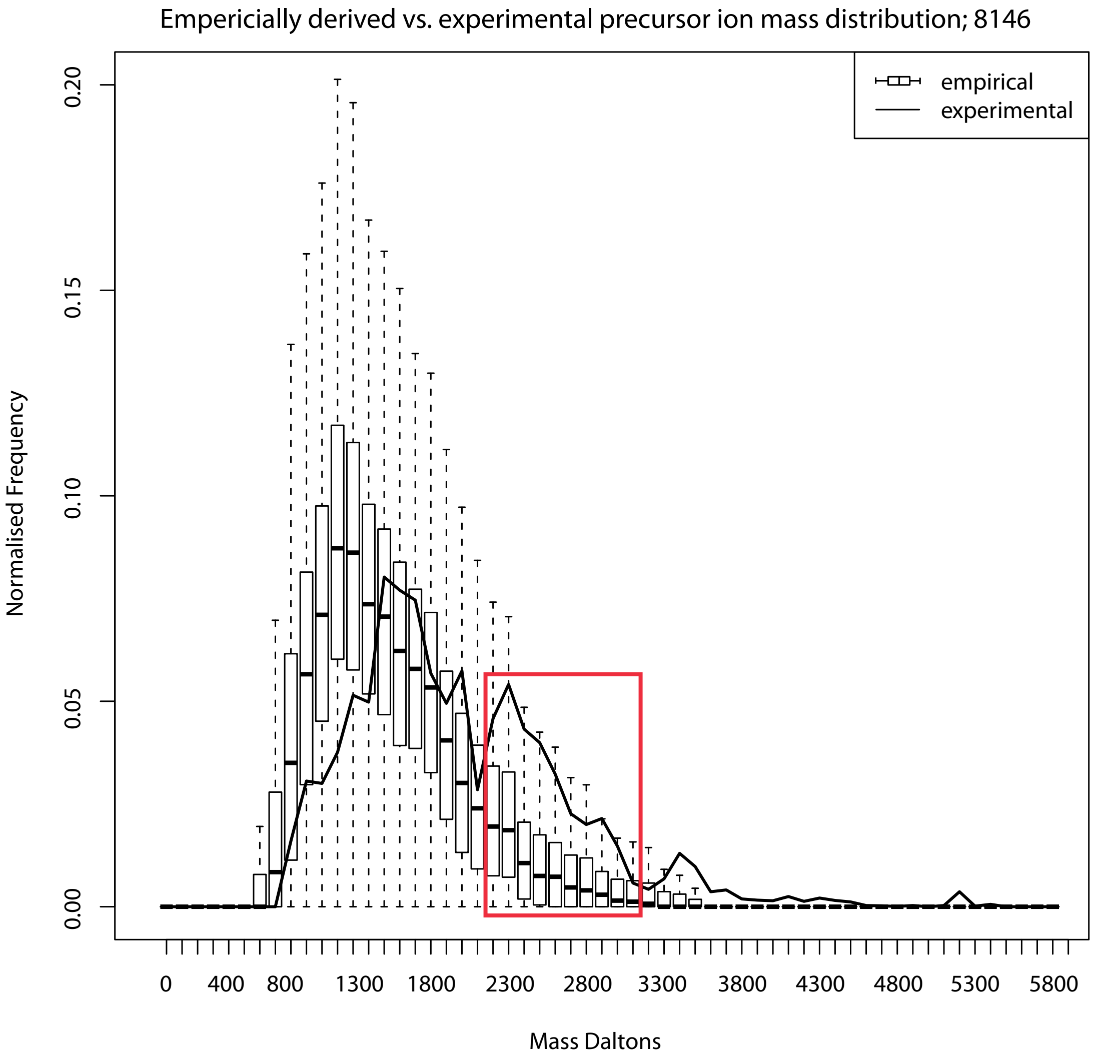


Figure 2.4: The precursor ion mass distribution of PRIDE experiment 8146 overlaid on the empirical precursor ion distribution of all samples within the HUPO Test Sample Study. The mean precursor ion mass (Daltons) is significantly different between experiment 8146 and the empirically derived background. For some reason the detected precursor ion masses are higher in the region 2000-3000 Da in experiment 8146. This could be explained by an inefficient tryptic digest that resulted in many missed cleavages and hence longer peptides with higher mass on average.

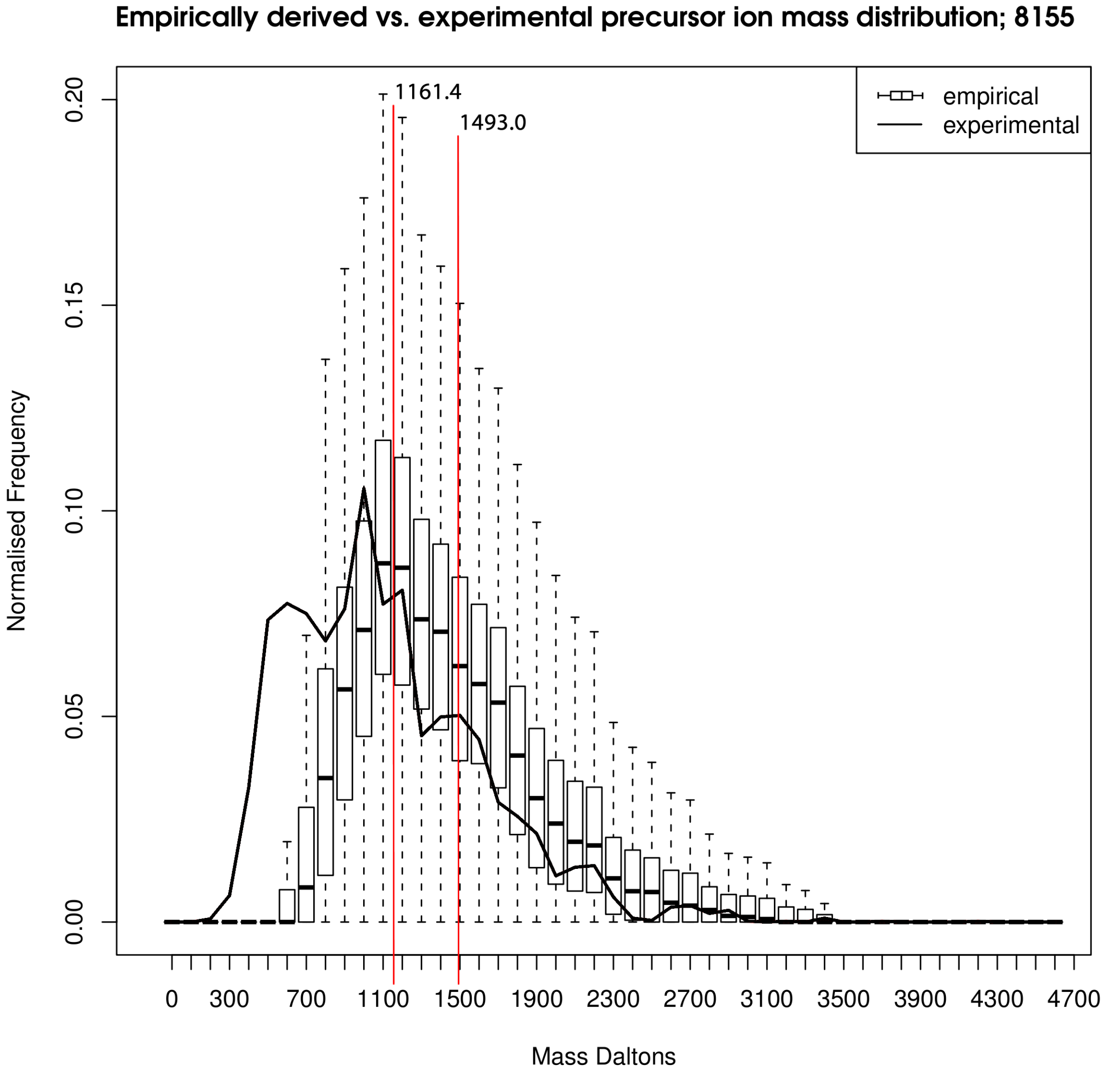


Figure 2.5: The precursor ion mass distribution of PRIDE experiment 8155 overlaid on the empirical precursor ion distribution of all samples within the HUPO Test Sample Study.The mean precursor ion mass (Daltons) is significantly different between experiment 8155 and the empirically derived background. For some reason the detected precursor ion masses are lower in experiment 8155. A mechanistic explanation requires a much more detailed analysis of the data but an explanation could be the use of an enzyme other than trypsin that has much more frequent cleavage sites throughout the human proteome or a combinatorial digestion protocol in which multiple proteases were used together.

explanation remains far off. If such a mechanistic explanation should be required, there is a need for the manual inspection of the dataset, and/or a thorough reading of the corresponding paper(s).

2.3.3 Mass Spectrometry Analysis

The tremendous sensitivity of mass spectrometry renders it highly susceptible to artefactual contaminants which enter into the sample during handling (Keller *et al.*, 2008), impairing the detection of *bona fide* peptides in the sample. Additionally, peptide fragmentation is not always efficient, sometimes yielding too few or too small fragments to be useful for identification. In order to measure the amount of non-informative MS2 spectra recorded during an analysis, I examined the distribution of the mass differences between peaks in each MS2 spectrum in an experiment (see section 2.2.5). To ensure analysis of the most significant signals, the MS2 spectra were first filtered to retain only the top 10% most intense peaks. M/Z difference (Δ m/z) distance matrices were computed from the filtered peak lists, these matrices were then combined resulting in a distribution of many Δ m/z and their frequencies. This was then plotted as a histogram, where a single bar represents a 1 Δ m/z difference between two peaks. The region 40 to 200 m/z was decided to be the most useful region of m/z difference distribution as it encompasses the m/z of all amino acid residues and several common contaminants at charge state 1+.

Figure 2.6 shows a typical high-quality example for such a distribution, with the Δ m/z corresponding to amino acid residue masses (grey) clearly rising well above the general noise level in the histogram. Figure 2.7 on the other hand, shows a distribution that provides a less favourable picture; the m/z differences corresponding to amino acid residue masses lie well within the noise and the extremely high peak at 44 Da corresponds to the mass of a polyethylene glycol (PEG) monomer building block, a common contaminant in mass spectrometry. The MS2 Δ m/z distribution is one of the best ways to quickly detect PEG contamination levels, since the actual polymers take a variety of precursor masses depending on the number of monomers they are composed of, but the MS2 spectra will always reveal the steady train of 44 Da differences. Interestingly, each experiment can now also be considered a multidimensional vector, with each dimension corresponding to a bar in the bar

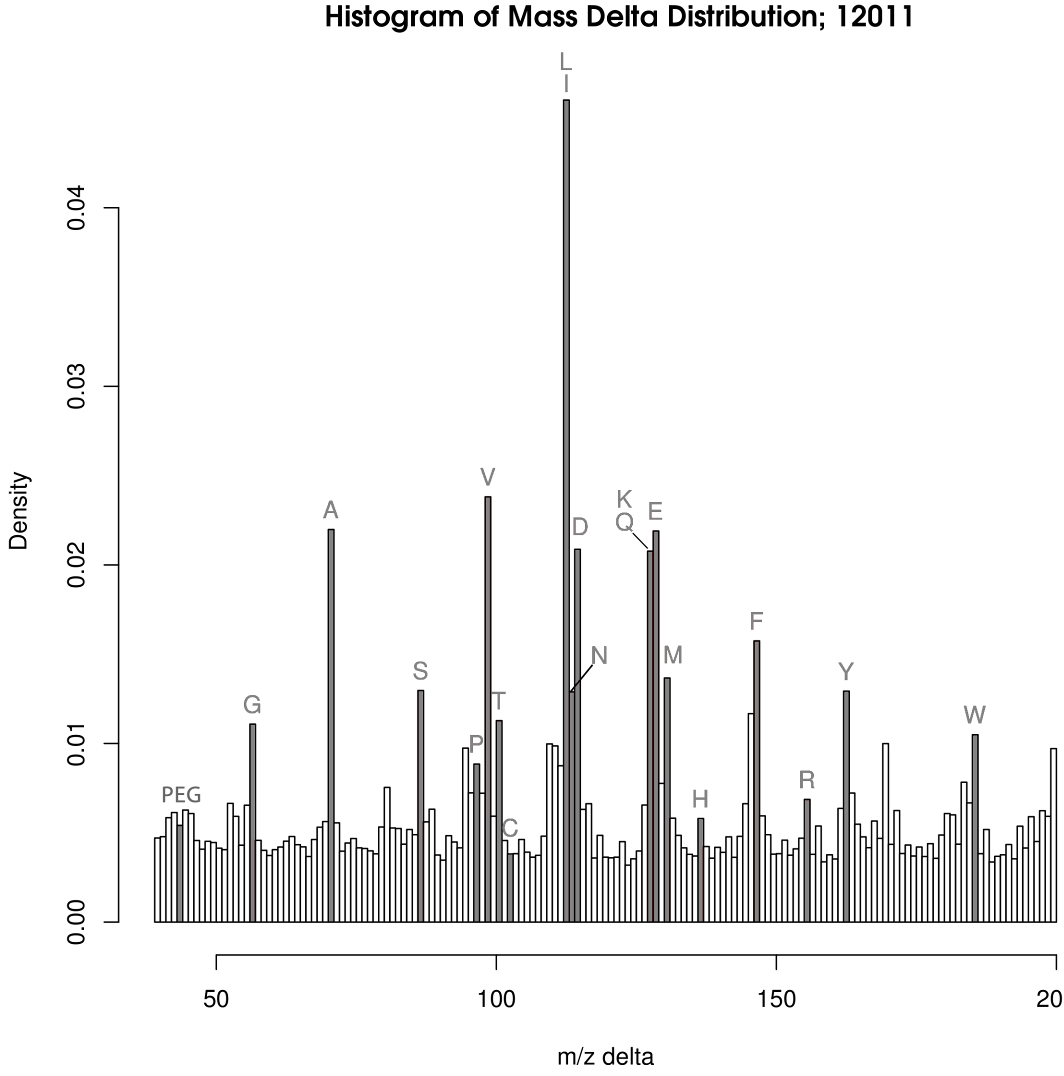


Figure 2.6: PRIDE experiment 12011 Δ m/z histogram. Constructed by creating an m/z distance map for the top 10% most intense peaks in each MS2 spectrum. Distance matrices are vectorized and concatenated to produce a vector of distances between peaks that is then transformed to a frequency table and plotted as a histogram. Specifically the Δ m/z region 40-200 is plotted as it contains the m/z of all +1 charged amino acid residue ions. Bars are labelled with one letter amino acid codes and with polyethylene glycol (PEG) at 44 m/z Δ. The amino acid bars lay clearly above the level of noise indicating that the top 10% most intense MS2 ions are sequence ions of peptides.

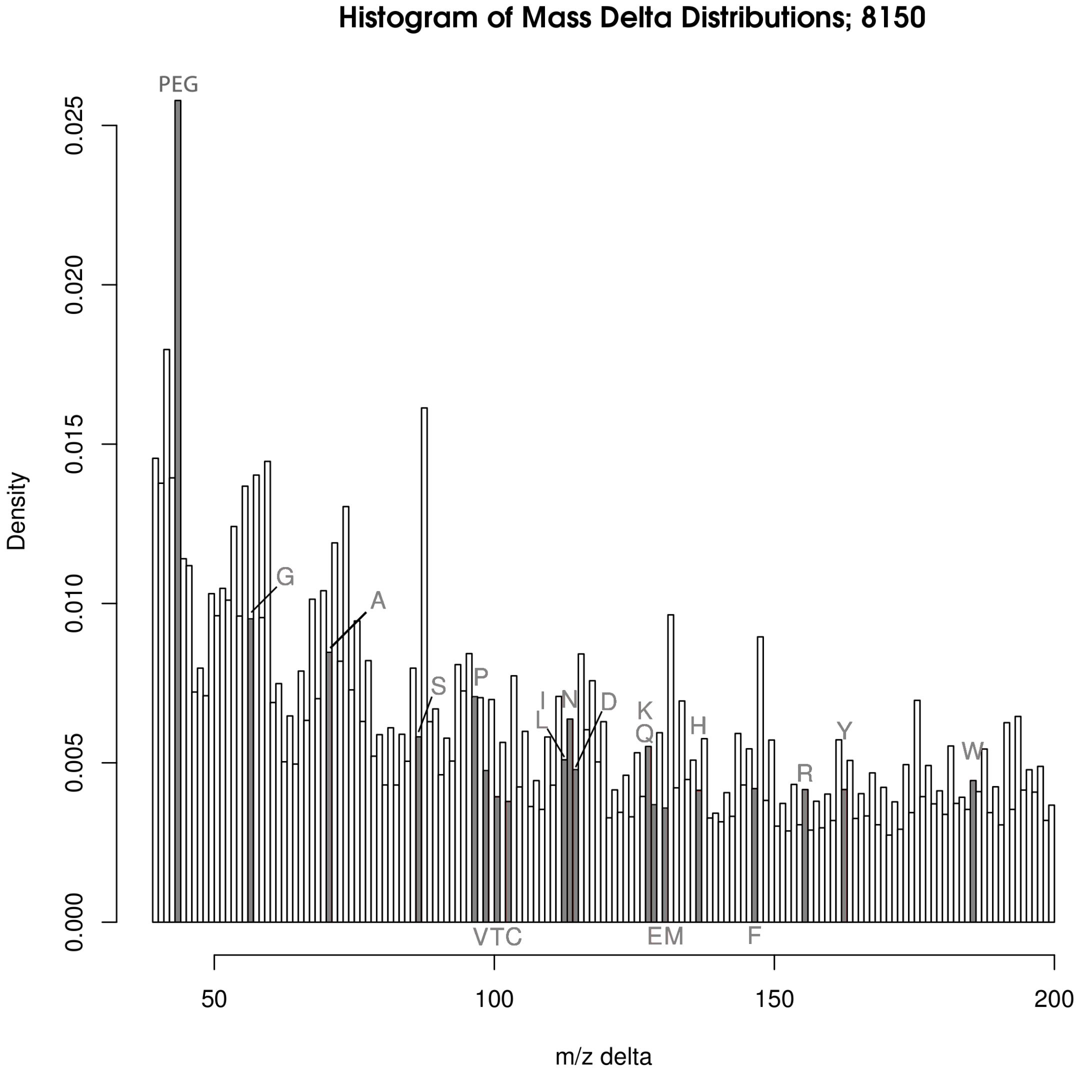


Figure 2.7: PRIDE experiment 8150 Δ m/z histogram. Constructed by creating an m/z distance map for the top 10% most intense peaks in each MS2 spectrum. Distance matrices are vectorized and concatenated, then transformed to a frequency table and plotted as a histogram. Specifically the Δ m/z region 40-200 is plotted as it contains the m/z of all +1 charged amino acid residue ions. Bars are labelled with one letter amino acid codes and with polyethylene glycol (PEG) at 44 m/z Δ. The amino acid bars lay amongst the general noise, the only clear signal is that of PEG a common contaminant of MS experiments.

plot. Since amino acid occurrence rates are species-specific, it should be possible to spot consistent patterns in experiments that are derived from the same species, and contrast these with the patterns obtained from experiments from other species.

Figure 2.8 depicts the two-dimensional projection of the experiment $\Delta m/z$ vectors for the twenty amino acid residue m/z through multi-dimensional scaling (see section 2.2.6). It is clear that specific species follow specific distributions, but that the separation in two dimensions lacks sufficient resolving power to clearly distinguish the species origin for any single experiment. Regardless, such approaches can be very useful in the *a posteriori* annotation of experiments and in the curatorial detection of possible misannotations. Also note that both *Saccharomyces cerevisiae* and *Drosophila melanogaster* populations seem to be split into two distinct categories.

The distribution of the m/z of the MS2 peaks by itself can also provide useful information, somewhat similar to analysing the precursor ion mass distribution as discussed previously. Typically, the distribution of MS2 m/z peaks is roughly normal. Performing this analysis across the portion of PRIDE experiments that contain MS2 spectra yielded unsurprising results in the majority of cases, but two experiments submitted sequentially by a single lab (PRIDE accession numbers 8927 and 8928) proved to be outliers.

Figure 2.9 shows the MS2 m/z distribution for experiment 8927, displaying an unusual truncated bimodal distribution. While the truncation is most likely the product of mass limits imposed by the analyzer, the bimodal character is unexpected. While there are multiple different charge states amongst the ions aggregated in this analysis, the sum of those charge state specific distributions should still be a near normal distribution. To check this was not a legitimate feature of the submitter's data, 150 other submissions from the same lab were investigated and a reference distribution of MS2 m/z created to compare against. Figure 2.10 clearly shows that these two experiments alone are cause for concern. After discussion with the data submitter the raw data was analysed and the instrument records cross referenced for potential causes. The cause of this phenomenon was ultimately attributed to a miscalibration of the mass spectrometer's analyzer due to a contamination of the transfer capillary, which resulted in overcharging effects.

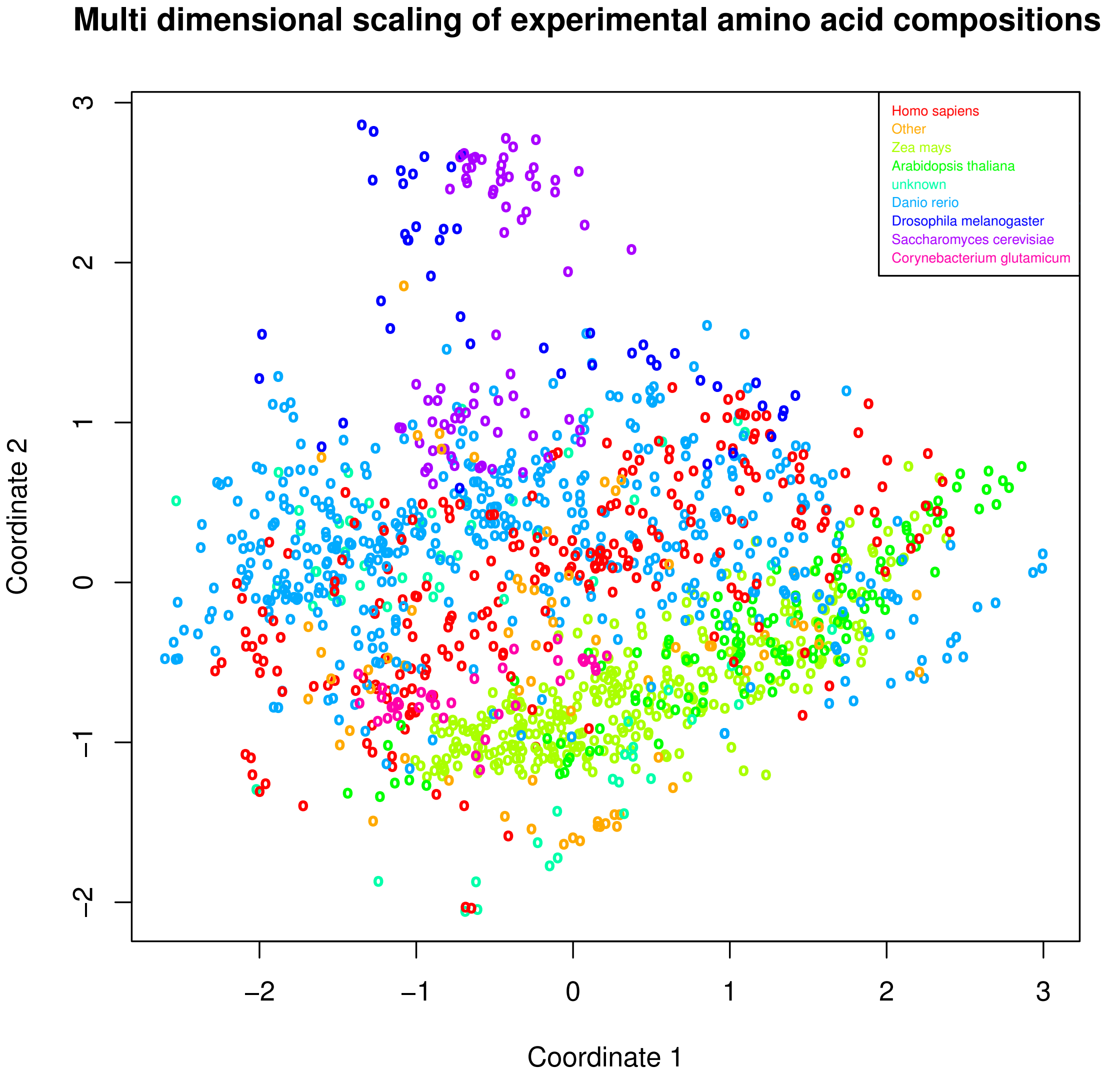


Figure 2.8: The Δ m/z frequency table for each experiment is a set of multidimensional vectors (a dimension for each Δ m/z window) that can be compared by multidimensional scaling in order that it can be projected into two dimensions and visualized. In this figure we see that experiments are coloured by the species under study and in only two dimensions there is an approximate separation of species, with the most dispersed group being a collection of multiple species called "other".

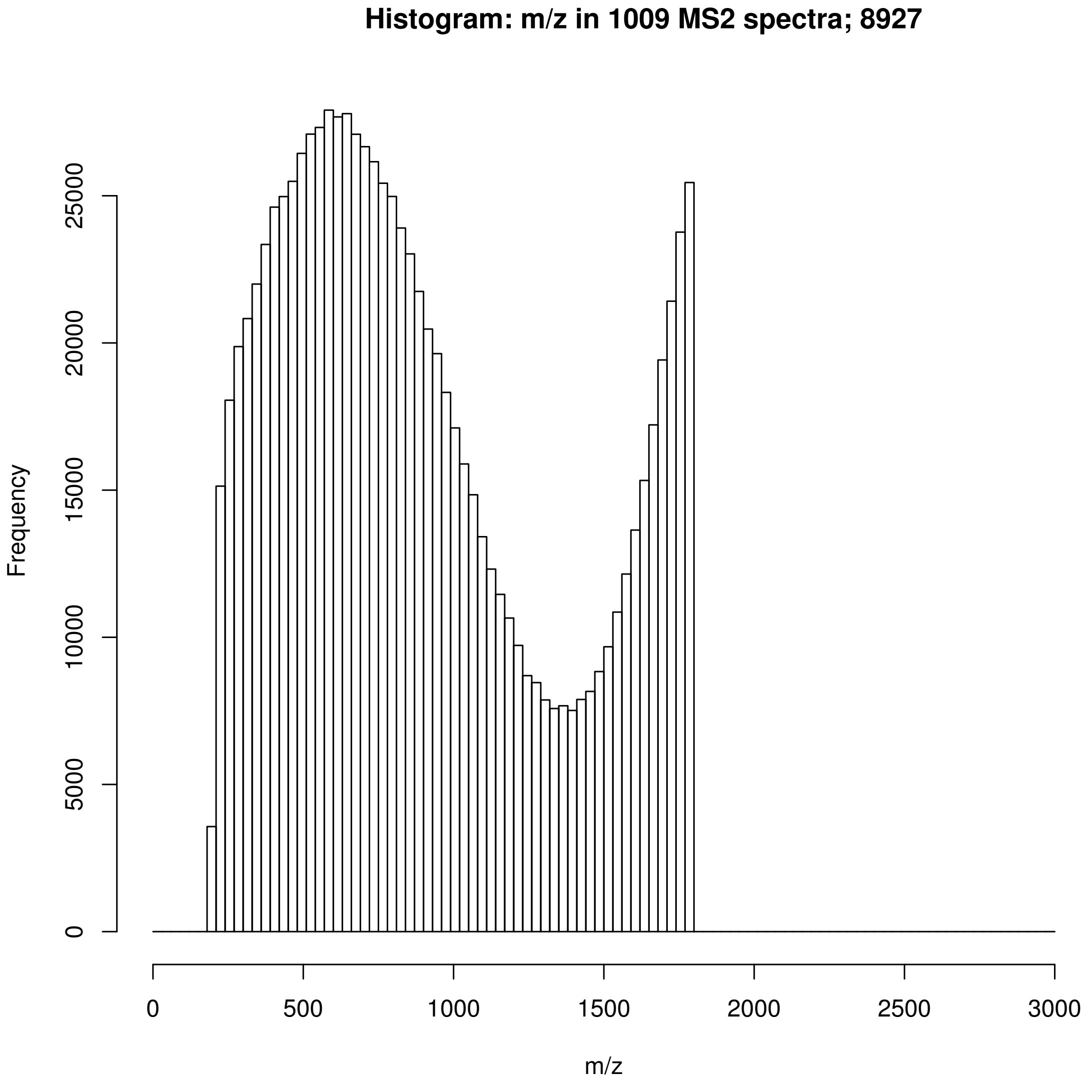


Figure 2.9: MS2 peak m/z distribution for PRIDE experiment 8927.

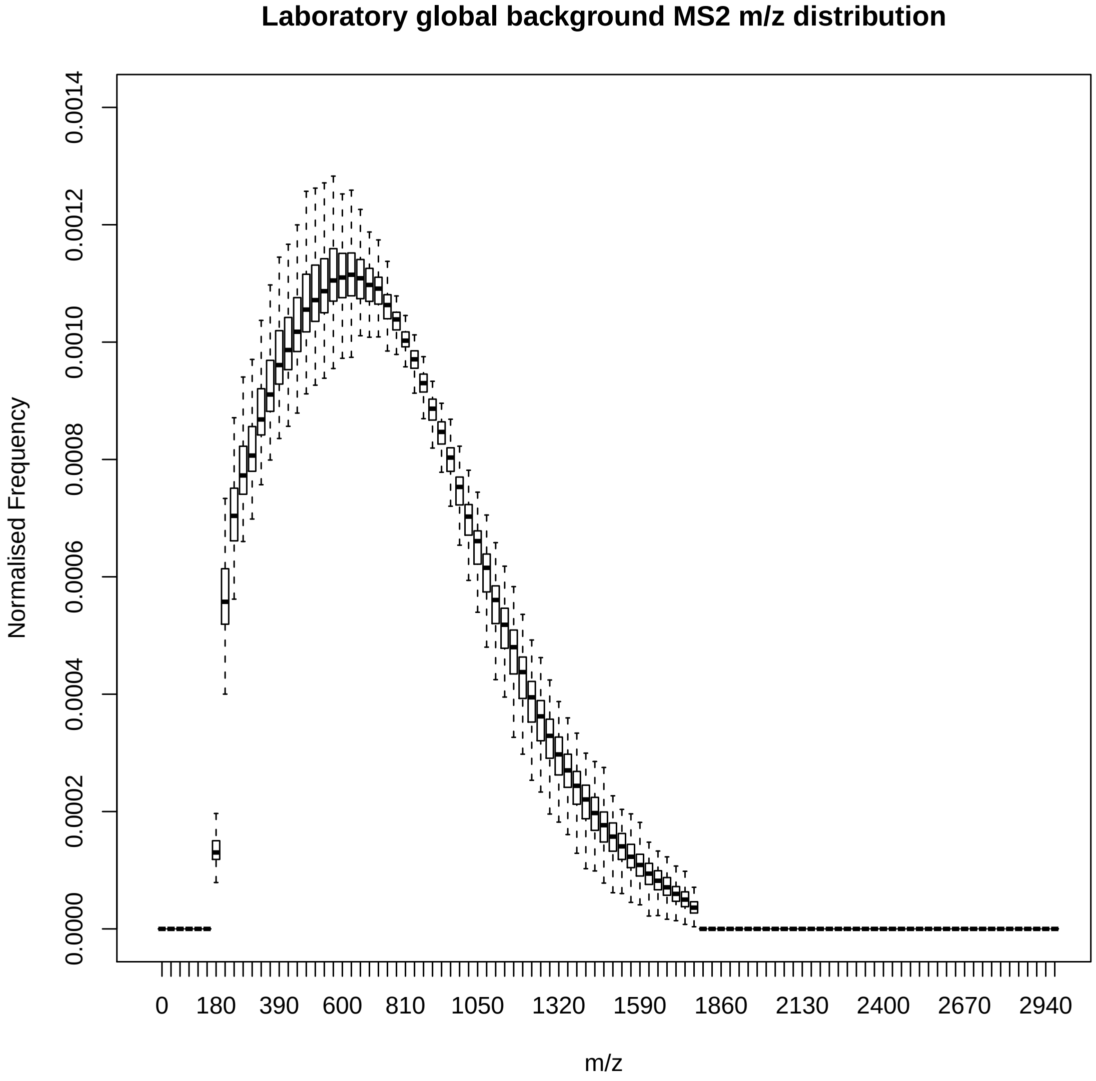


Figure 2.10: MS2 peak m/z distribution for 150 other PRIDE experiments attributed to the same lab as PRIDE experiment 8927

2.3.4 Quantitative Analysis

Quantification of proteins through peptides as surrogates has become increasingly popular over the last few years. In labelled proteomics tagging systems like iTRAQ (isobaric Tag for Relative and Absolute Quantification) (Ross *et al.*, 2004) or TMT tags (Tandem Mass Tags) (Thompson *et al.*, 2003) are routinely used to quantify peptides from their MS2 spectra. These tagging systems are multiplexed sets of isotope tags that are used to label all peptides generated from tryptic digestion. Since the tags are isobaric, differentially labeled versions of a peptide appear as a single precursor ion in MS mode. When labeled peptides are subjected to collision induced dissociation (CID), the tags release diagnostic, low-mass reporter ions that are used for quantification. See figure 2.11 for a schematic representation of the TMT methodology. With the increase in popularity of these and other quantitative methods, it is crucial to develop corresponding QC methodologies and metrics. Because the storage of quantitative proteomics data in publicly available proteomics data repositories currently trails behind their ability to store data that results from more traditional proteomics strategies, there is very little consistent data to be found in the public domain. Consequently, comparative QC on quantitative proteomics data remains a little further in the future, but independent QC checks can already be designed and applied with an aim to calculating them across multiple datasets to generate reference metrics when more data becomes available. For instance, by comparing the measured ratios with the expected ratios, the accuracy of the (relative) quantification can be determined. In an experiment performed in collaboration with my industrial partner Philips, a proteomic sample was diluted into 6 samples in the relative concentration ratio 1:1:3:3:9:9 and a unique TMT tag applied to each sample. After identification and quantification a number of analyses were preformed to assess the quality of the quantification. Assessing variables such as sample concentration, TMT tag specific quantification variance and TMT tag specific likelihood of detection.

In Fig. 2.12 a distribution of the ratios of the peak intensities at 126 to 130 Da with respect to the peak intensity at 131 Da is shown as box plots. On the left, data from identified spectra only was taken, whereas the box plots on the right also incorporate unidentified spectra. These are centered on y = -2, y = -1 and y = 0

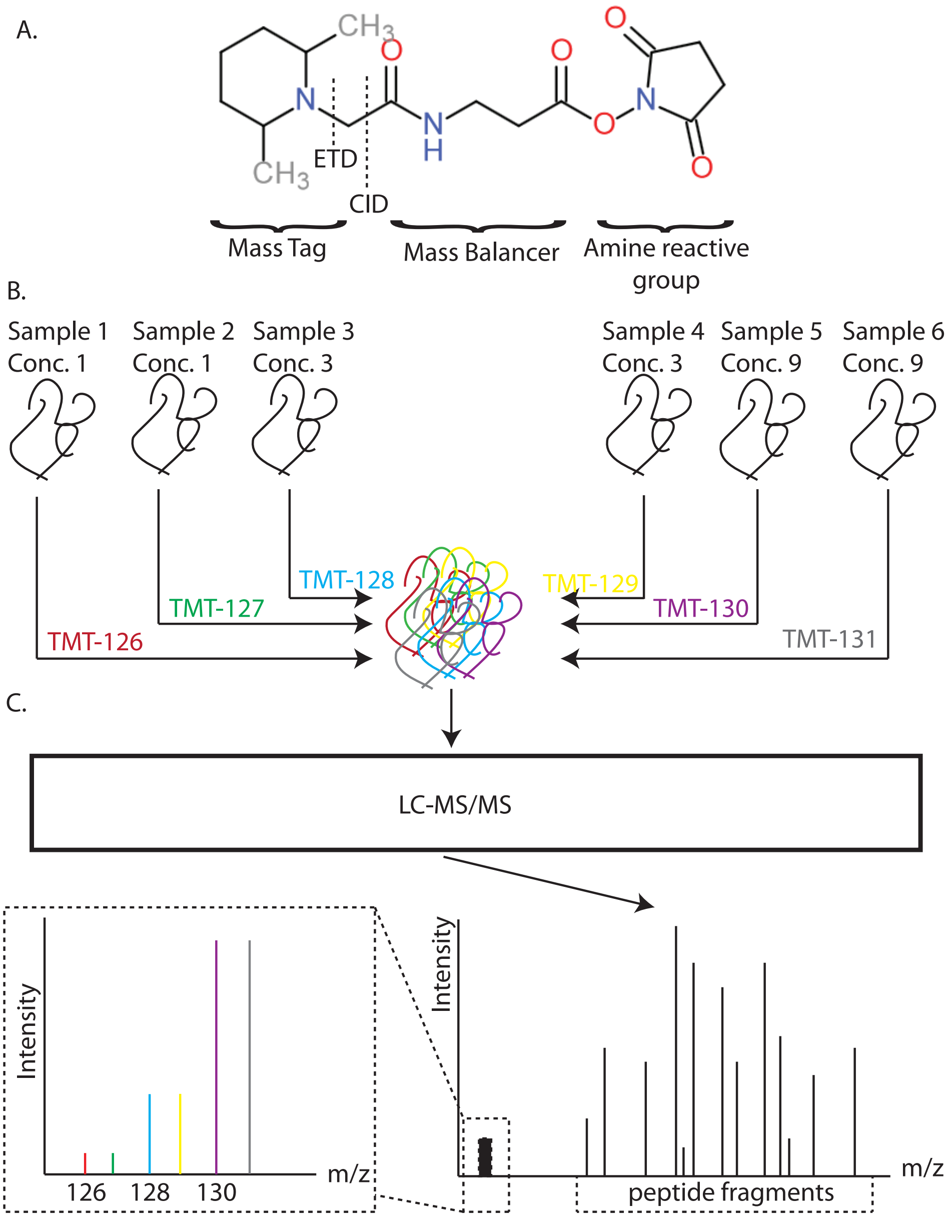


Figure 2.11: **A** The structure of amine reactive TMT reagents. All six tags tag have the same overall mass but it is distributed differently between the “mass tag” and “mass balancer” by the use of heavy carbon and nitrogen isotopes. **B** Up to six peptide samples are independently labelled with a distinct TMT reagent then pooled together. **C** Upon fragmentation the “mass tag” is released from each peptide and the peptide undergoes the normal fragmentation process. An MS2 spectrum is recorded and the sequence ions used for identification of the peptide. The m/z region 126-131 contains the fragmented “mass tag” ions, the intensity of which can be used to determine the relative concentration of that peptide (and by extension, protein) in the samples.

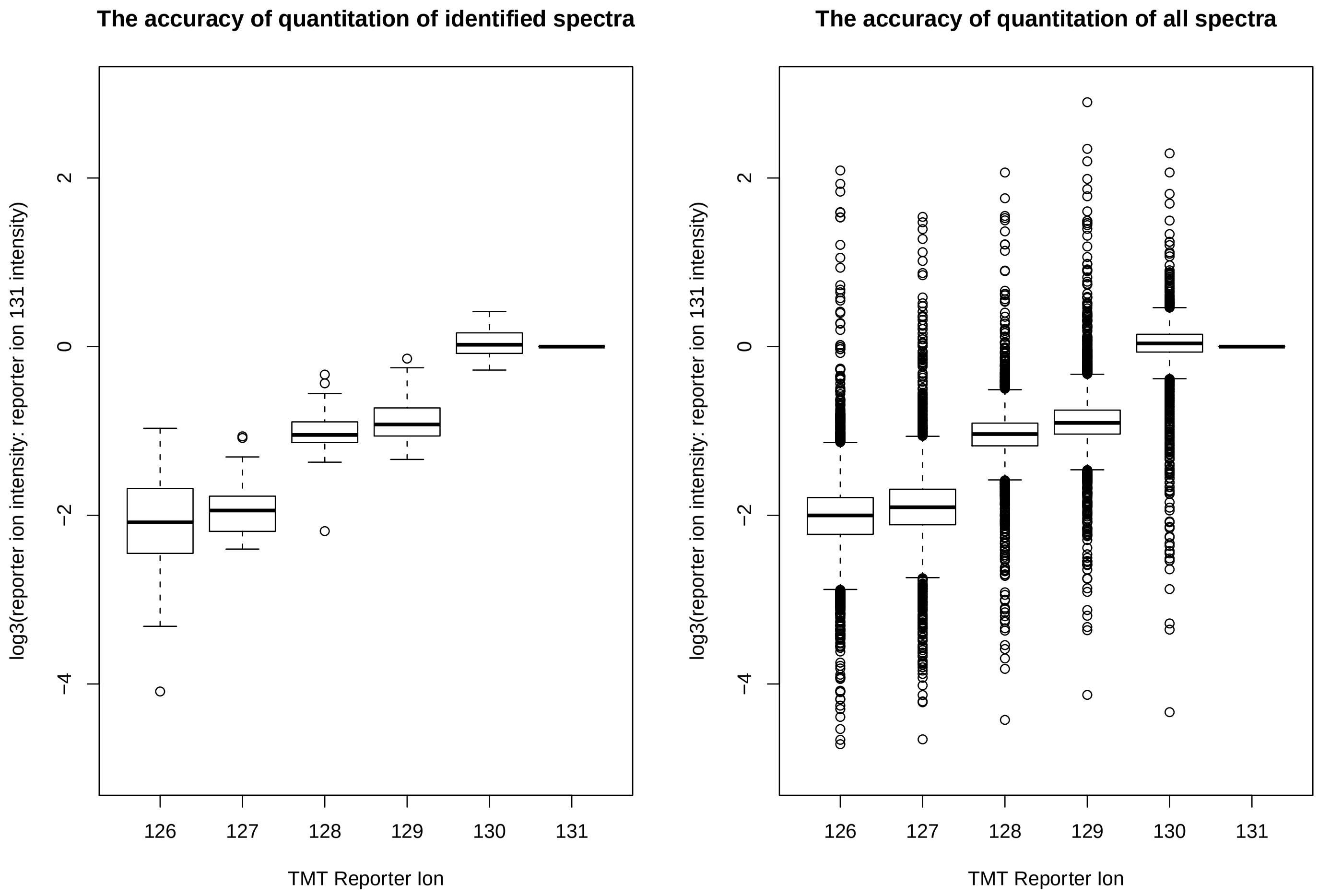


Figure 2.12: Each box plot is constructed by taking the ratio of a particular TMT tag peak to the TMT 131 peak. Ideally this would produce ratios in log3 scale of -2, -1, 0. This was done for 35 identified spectra only in the left panel and for all 4949 spectra in the right panel.

on a log3 scale as expected from the test sample's design, indicating good accuracy, even on the very heterogeneous data provided by summing across unidentified and identified spectra. The range of the box plots gives the precision in this test sample, corresponding to technical reproducibility. It appears that as the expected log3 ratio increased in magnitude as did the range of the results indicating that when doing this type of relative quantification, the standard against which something is quantified is best in similar concentration as the analyte. One must be careful to translate this ideal scenario to the more common situation where only a few components are regulated. In that more complex scenario, imperfect selection of the precursor peptide leads to overlapping TMT reporter peaks, incorrect quantification and hence calculation of the regulation. Particularly for low-abundance proteins, the effect will be an underestimation of the regulation. However, simultaneous selection of two or more precursor peptides is detectable from the MS2 analysis of the fragmentation spectra (Houel *et al.*, 2010), providing a means to identify such co-fragmentation events.

The histogram in Figure 2.13 shows the number of missing values at the m/z positions of the 6 TMT reporter ions. The vast majority of spectra contained all 6 reporter ion peaks, indicating overall efficient labelling. In only 0.8% of spectra, all 6 peaks were missing, either because the corresponding peptide was of low concentration, or because of incomplete labeling or fragmentation. This cannot be verified, as the analysis is done on all spectra, and generally we find that only a minority fraction (10-50%) of the MS2 spectra leads to an acceptable peptide identification. Limiting ourselves to spectra leading to accepted peptide identifications for this dataset however, lead to very similar results. Although the majority of the spectra was not identifiable in the first pass of a Mascot search, a large fraction of these could be mapped to peptides of proteins from the first search. This second search used Mascot's error-tolerant search, allowing different modifications, missed and non-tryptic cleavages (certainly likely due to the presence of proteases in serum (Yi *et al.*, 2007)). The more recent Protein Pilot software from Life Science (`http://www.absciex.com/products/software/proteinpilot-software`) typically allowed the identification of more than 80% of the spectra in the first pass.

Figure 2.14 is a histogram of the number of missing values for each of the six TMT reporter ions separately and reveals that missing peaks were much more common

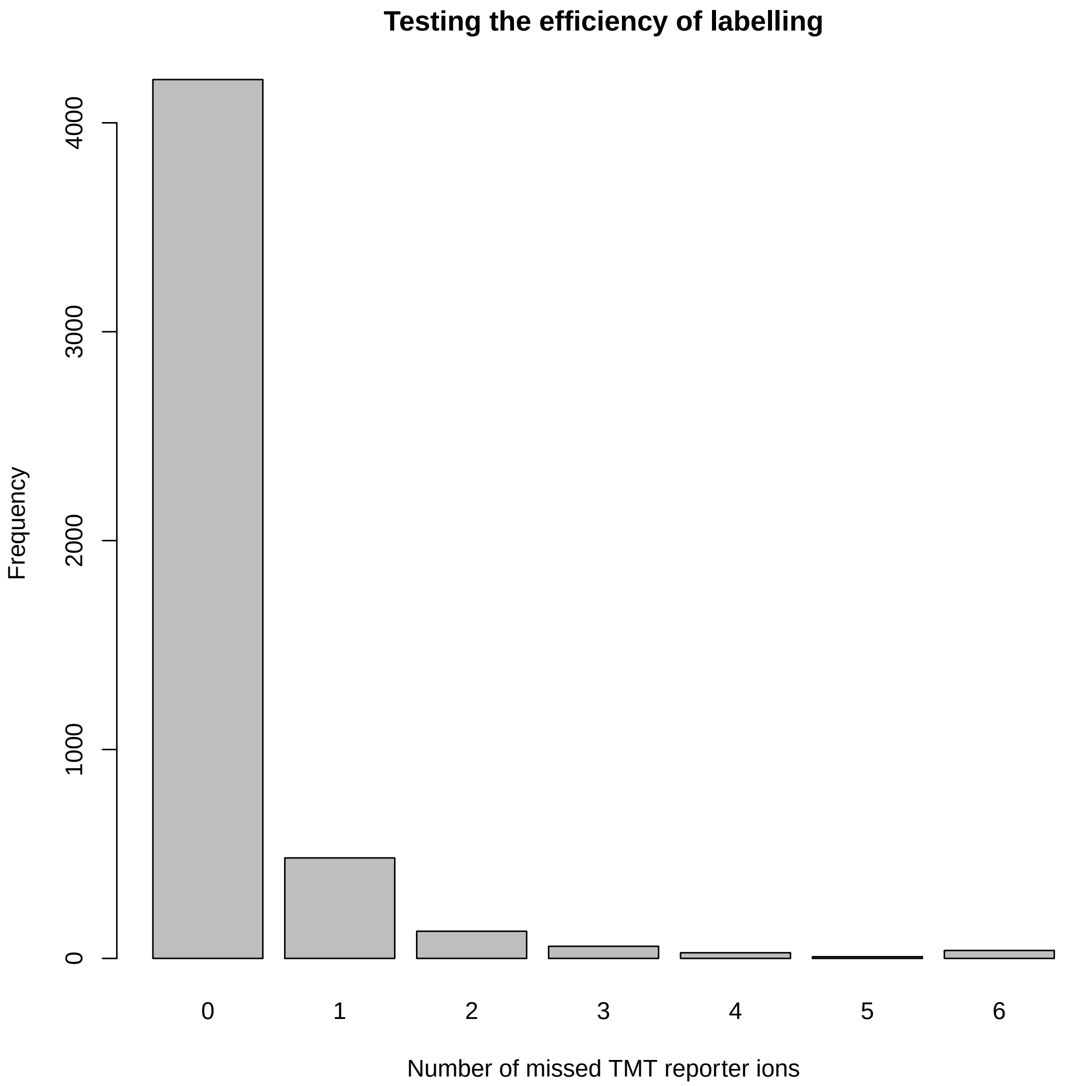


Figure 2.13: A histogram depicting the frequency of missing TMT reporter ions. Calculated by analyzing each spectrum for the presence of a peak in a small tolerance window around the expected m/z of a TMT reporter ion.

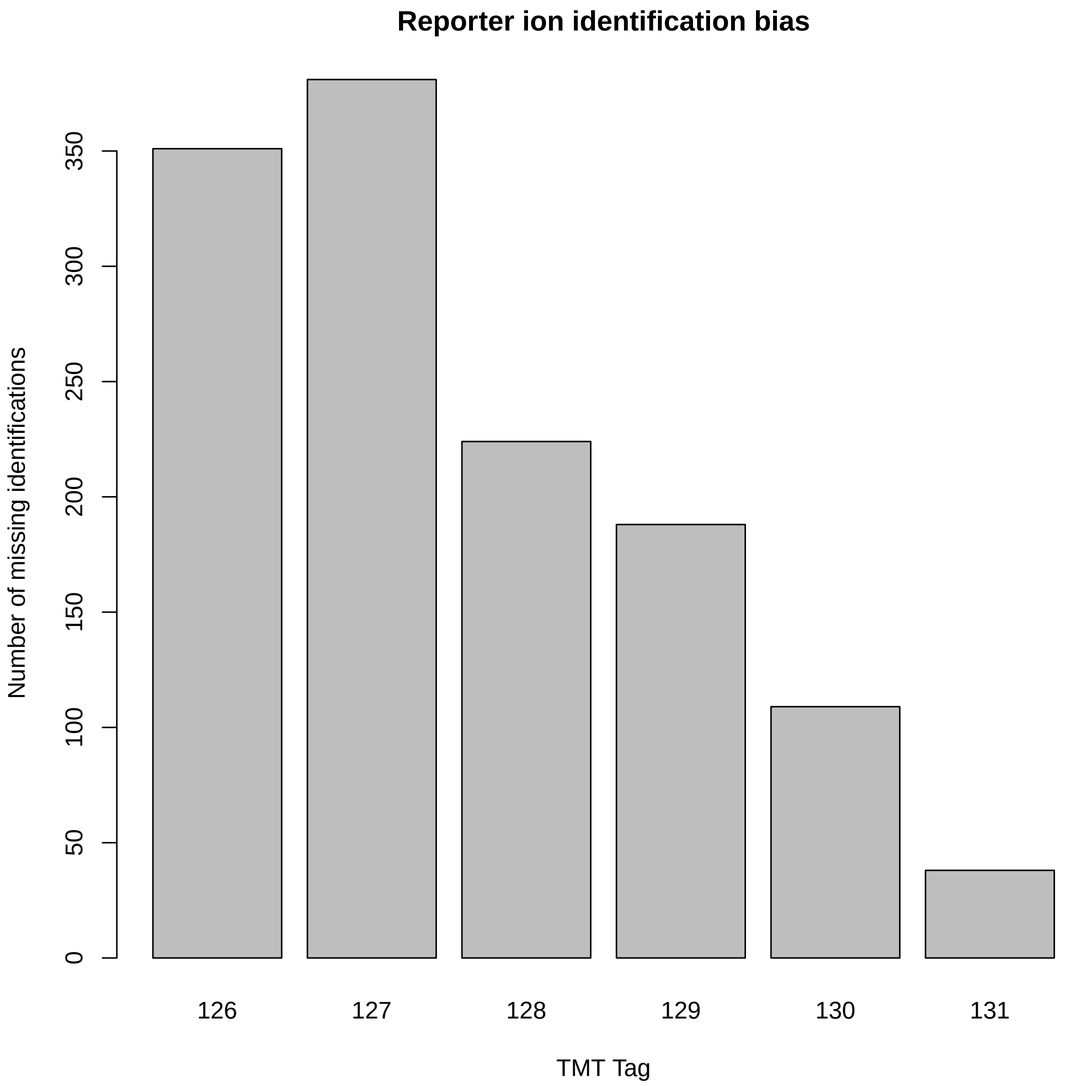


Figure 2.14: This histogram identifies the frequency at which individual reporter ions are missing.

amongst the masses 126 and 127. These were the samples with nine times lower concentration, showing that abundance is an important factor in the detectability of reporter ions.

The intensity of the different TMT reporter ion peaks against the average intensity of the top 10% most intense non-TMT peaks is plotted in Figure 2.15. A single dot on the charts represents a single MS2 spectrum, and axes have been drawn on log10 scale. This analysis shows whether a good balance between quantification and identification has been found. This appears to be the case here, as the reporter ions have an intensity that is clearly correlated to the metric. It is also apparent that the ratio between reporter ions to the top 10% most intense peaks is quite constant across a broad intensity range, revealing that reporters are good surrogates for peptide quantification. At the same time the 1:100 (Figure 2.15A and 2.15B) to 1:10 (Figure 2.15E and 2.15F) ratios indicate that peptide identification (typically based on the most intense peaks in the spectrum) has not been overshadowed by the reporter ion peaks at all, illustrating the absence of overly competitive ionisation. However, at the same time, the actual ratios of the different reporters versus the top 10% most intense peaks correlates very well with the sample composition; roughly 1/100 for 126 and 127, about 1/33 for 128 and 129, and a little under 1/10 for 130 and 131, corresponding to the 1:1:3:3:9:9 mixing ratio for these reporters respectively. The TMT peaks are thus well-suited to (relative) quantification purposes.

The R scripts are publicly available alongside the publication (Foster *et al.*, 2011) and can be downloaded from `www.ebi.ac.uk/~jfoster/MSQCLib.zip`.

2.3.5 Metadata Provision

For a repository, metadata is an important feature of the submitted spectra. Without the context provided by experimental and data processing metadata, re-usage of the data or even validation of the results is near impossible. In recent years attention has been foccused on the provision of metadata in publicly deposited proteomics experiments (Medina-Aunon *et al.*, 2011; Montecchi-Palazzi *et al.*, 2009; Taylor *et al.*, 2007). Similar to the methods section in a paper, the metadata should enable the interested party to understand the acquisition and processing of the data clearly enough to replicate its production. It is unheard of for research papers in

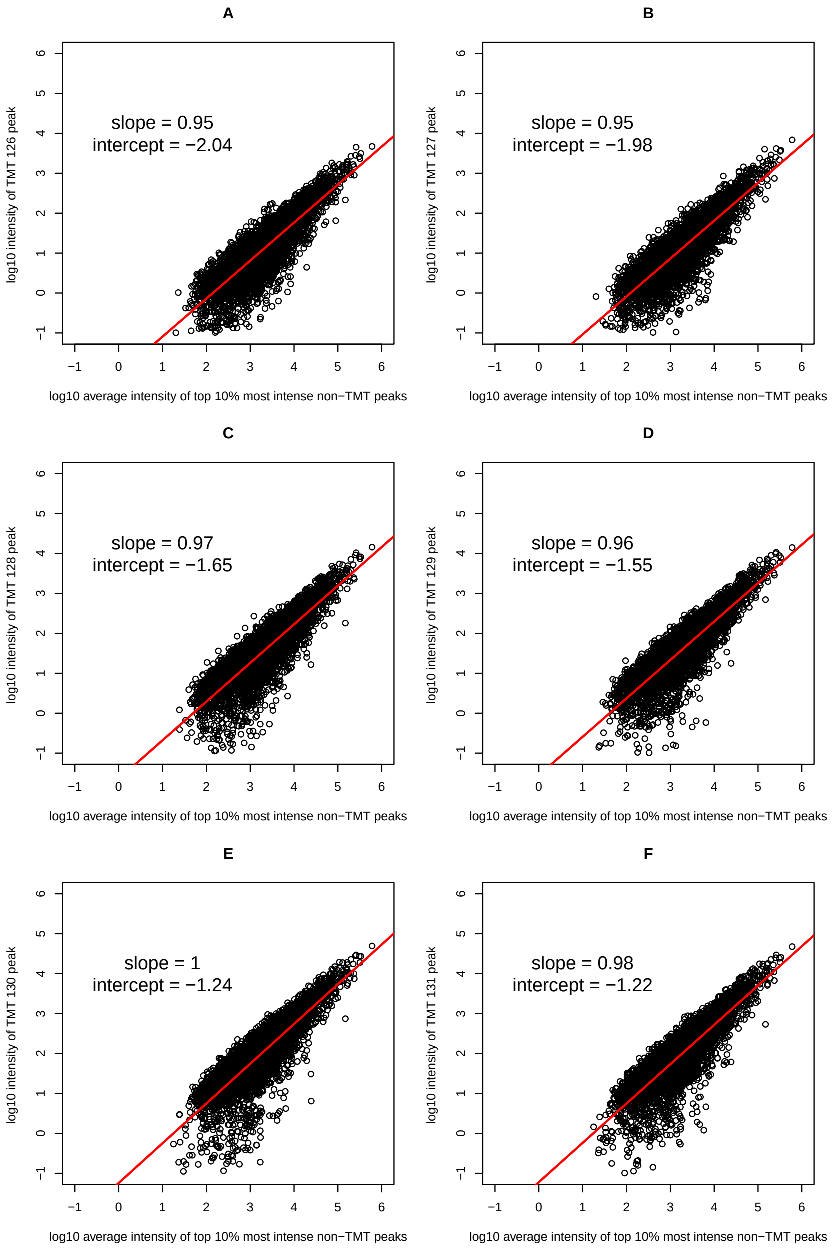


Figure 2.15: The intensity of the TMT reporter peaks *versus* the average peak intensity of the top 10% non-reporter ion peaks in each spectrum. TMT reporters shown: (A) 126 Da; (B) 127 Da; (C) 128 Da; (D) 129 Da; (E) 130 Da; (F) 131 Da. These plots relate quantification to identification, since the reporter ions have correlated intensities to the top 10% non-reporter ions that are most often used for identification purposes. The TMT reporters thus prove to be adequate estimators of quantity. Since the reporter ions are always at least an order of magnitude less intense than fragment ions, an overshadowing effect by reporter ions, reducing peptide identifications is unlikely.

which data has been collected to exclude a methods section, yet it is still commonplace in public data depositions in all ‘omics’ fields for metadata not to be present. References to papers are an excellent source of additional information for public proteomics data. However, in PRIDE only 4051 out of 12649 publicly accessible experiments are associated with a published manuscript. While its occurrence is declining, lacklustre metadata is still a significant burden upon curators and data re-users, who must first infer the metadata where possible in order to judge its suitability for either submission or analysis respectively. Currently 21/12649 public experiments in PRIDE do not have a species annotation to describe where the sample under study originated from. Of the 12628 PRIDE experiments with a species annotation 9570 do not have a tissue annotation; this is important since unlike the genome which remains effectively constant between cell types, the proteome can be extremely diverse between tissues. Other useful annotation for data re-users is the instrument type and identification software used to produce peptide identifications and the peptide to protein inference, 17% of all PRIDE experiments have an unspecified identification software annotation while 28% have an unknown instrument. Figures 2.16 and 2.17 show the seven day sliding mean percentage annotation of all PRIDE experiments in the last 6 years for the metadata parameters “instrument” and “software”. Instrument annotation in figure 2.17 shows a few periods when the local mean % of experiments not annotated rises extremely high, due to the submission of a large set of unannotated experiments. On the whole instrument annotation is very well reported by the proteomics community. In contrast, figure 2.16 demonstrates that software annotation (spectra identification software) is much less well reported. This may be the product of the common misconception that search engines are all very similar and produce identical identifications for the same spectra. While this is true in general terms, agreement between search engines is not 100% and as such is important metadata. By extension this highlights the fact that the actual search engine parameters used are almost certainly not reported in experiment where the software was not reported, without which the results cannot be reproduced independently.

The problem of good quality metadata does not end by simply providing all the relevant annotations, they must also be the most concise and conventional annotation. A problem that occurs frequently in bioinformatics and in particular in

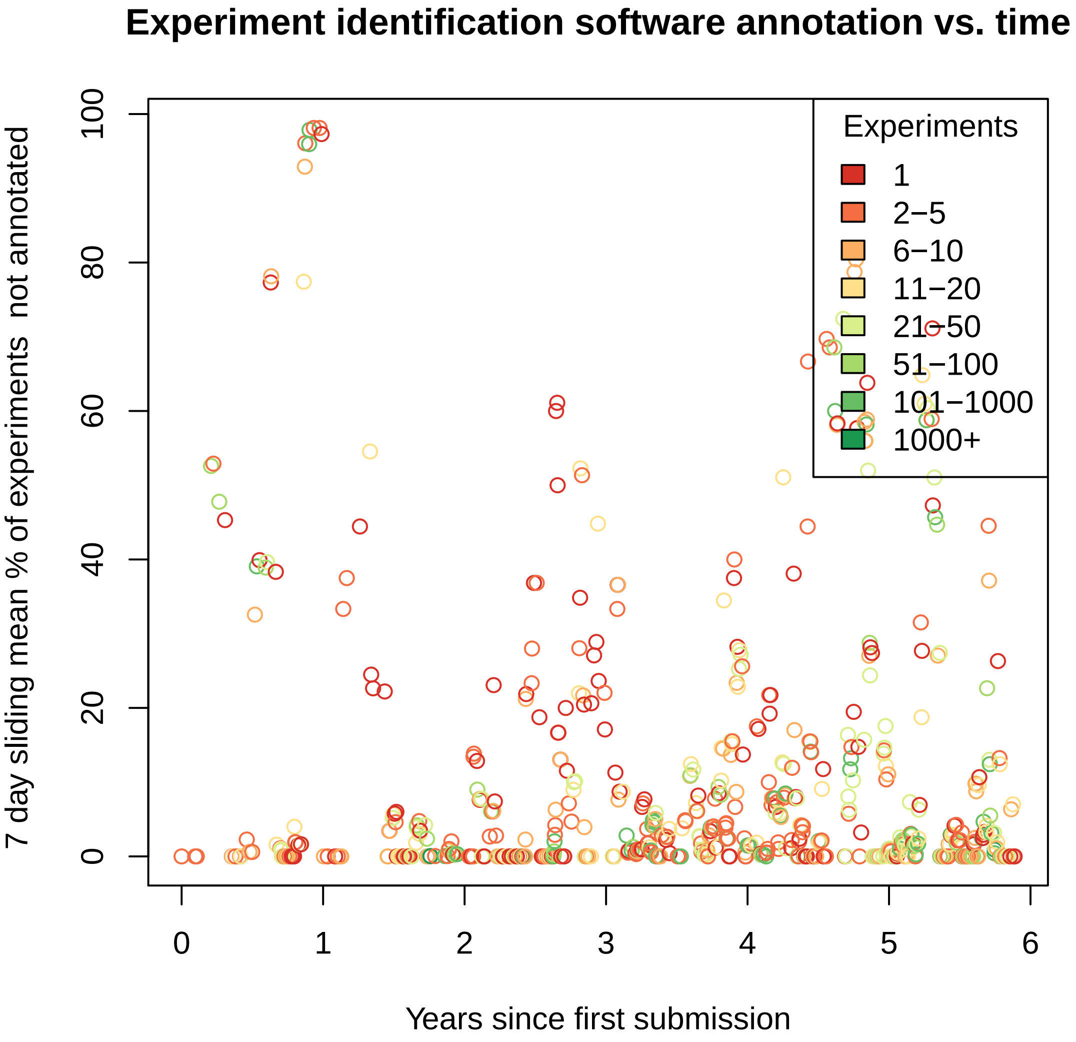


Figure 2.16: The seven day sliding mean percentage of experiments not annotated with spectral search engine software over time for the entire history of PRIDE submissions. Time zero is considered the date of the first experiment submission to PRIDE. Data for this graph is accurate as of the 2nd August 2012. The chaotic mean percentage of experiments not annotated with software metadata is a clear marker for the field's opinion on the provision of metadata. Results cannot be independently verified without this information and more detailed information on the search parameters.

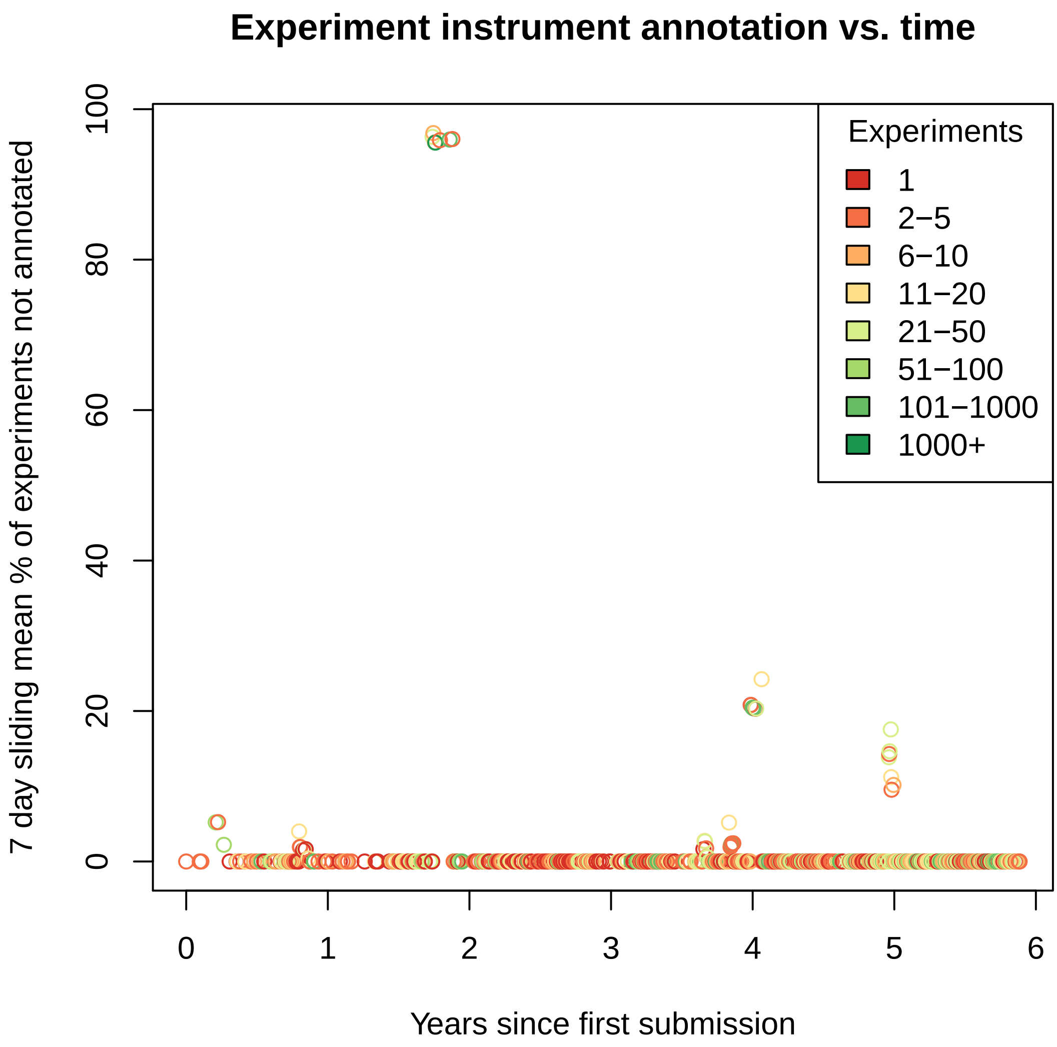


Figure 2.17: The seven day sliding mean percentage of experiments not annotated with instrument type over time for the entire history of PRIDE submissions. Time zero is considered the date of the first experiment submission to PRIDE. Data for this graph is accurate as of the 2nd August 2012. Instrument metadata is on the whole much better reported than search engine software, with the majority of recent experiments providing the instrument type. However there are still occasions where large submissions that do not contain this information are submitted to the database. Instrument annotation may be much more common in PRIDE experiment submissions because it may be perceived as more important by the submitters, but a number of other explanations may fit, including simpler reporting tools.

bioinformatics databases is that of nomenclature (this topic is also discussed in more detail in Chapter 3). While it is plainly obvious to an expert in the field that "Xcalbur","XCalibur" and "x calibur" represent the same identification software (the "Xcalibur" package from ThermoFisher Scientific), even the slightest difference in these terms will cause confusion for a computer. For example when searching a database like PRIDE for the instrument annotation "Waters QTof Premier", the user should in fact be searching for "QTof-Premier" and all of its 6 synonyms currently in the database to provide a complete view of the relevant data holdings. Synonymisation is no small task for a database such as PRIDE that has to accept submissions from the global proteomics community and account for discrepancies in nomenclature between different laboratories.

Pioneering efforts in the PRIDE team and the Proteomics Standards Initiative (PSI) (Orchard & Hermjakob, 2011) have therefore produced and encouraged the widespread use of community agreed upon ontologies to describe each individual metadata annotation (Hoogland *et al.*, 2010; Jones *et al.*, 2010). These are implemented in the PRIDE database through submission tools such as PRIDE Converter (Barsnes *et al.*, 2009, 2011), where ontology terms can be selected and searched for, reducing the amount of manual input necessary to describe the data. The result is a more homogeneous set of annotations in PRIDE that allow data re-users to aggregate larger amounts of data and perform more complex re-analyses of combined datasets to investigate the general state of proteomics as a whole.

However easy it is for data submitters to provide sufficient and accurate metadata there will always be some that fail to do so either by accident, or through a misunderstanding of how important it really is. There are several issues with encouraging authors to publish their results, currently there is very little re-use of the data in wider proteomics applications and repositories such as PRIDE are occasionally viewed as "data graveyards" which require tedious amounts of work to submit to with very little added value. These feelings in the community are somewhat substantiated, not all proteomics approaches are covered by the standard submission tools and data reuse does not come close to that in the fields of x-ray crystallography. Advances are being made to make these processes easier and the metadata annotation more clear in order to try and break the chicken and egg cycle, where data is not reused because there is not enough in the public domain with sufficient annotation and there is not enough submitted because there is currently little added value to doing so. In the case of accidental misannotation the errors can sometimes be identified and rectified, resulting in an accurate annotation. When metadata is simply not available, such as in the case of the HUPO PPP2 dataset and its SCX fraction order annotation (see figure 2.1), the individual PRIDE accession numbers had to be ordered to reconstitute the actual SCX fraction order of the various individual samples. This required a combination of looking at the submission order, the submission file name and the experiment name and was a largely manual job.

The PSI semantic validator (Montecchi-Palazzi *et al.*, 2009) provides a computationally quick framework to assess the metadata content of experimental data reported as XML. Not only does it check the syntax of XML but it allows rules to be created that enforce how controlled vocabularies are used within the document and that numerical values are in a reasonable range. Defining sub sets of viable annotations for specific metadata fields such as 'identification software' from controlled vocabularies ensures data is reported in the most standard way possible and helps eliminate the problems described previously.

2.3.6 Access to Public Data

Important lessons can also be learned from other fields regarding the accessibility of scientific data in public repositories in order to effectively propagate it to interested data consumers. Success in other fields has come from a variety of data accessibility options, including public MySQL access (Protein Data Bank Europe, `http://www.ebi.ac.uk/pdbe/` (Velankar *et al.*, 2011) and Ensembl, `http://www.ensembl.org/` (Flicek *et al.*, 2011)), Java APIs (UniProt, `http://www.ebi.ac.uk/uniprot/`, (Patient *et al.*, 2008)), REST & SOAP web services (Array Express `http://www.ebi.ac.uk/arrayexpress/` (Parkinson *et al.*, 2011)). Until recently data from PRIDE was only available to data consumers via FTP, the PRIDE web application and the PRIDE BioMart, none of which were readily accessible to the lay user. A more end-user friendly system was therefore developed for the PRIDE database in the form of PRIDE Inspector.

2.3.7 Retrieving data from PRIDE

Included in the R scripts previously described are a number of functions for accessing data directly from the PRIDE public MySQL instance ready for analysis. In response to end-user needs (including peer reviewers for journals that mandate data deposition in PRIDE for proteomics manuscripts), the curatorial requirements of the PRIDE team and the general shift in focus of the proteomics community towards more stringent QC, the stand-alone application PRIDE Inspector was developed (Wang *et al.*, 2012). Capable of downloading experimental data and displaying it in an intuitive and comprehensive manner, PRIDE Inspector was also designed to provide a QC overview of an experiment that allows data consumers to quickly assess the suitability of an experiment for re-use or comparison and also for curators to quickly identify faults in annotation, spectrum processing, peptide identification, protein inference and modification reporting prior to submission to the database. It is upon the work described previously that the QC module of PRIDE Inspector was founded, translating the original approaches in R to the JFreeChart framework (`http://www.jfree.org/jfreechart/`) and adding new quality control measures to suit the specific needs of the PRIDE curators.

2.3.8 Evaluating data using PRIDE Inspector

In what follows, a selection of faults found in PRIDE experiments through the use of the QC component of PRIDE Inspector are highlighted. They are a representative sample of issues that can occur when data is submitted to PRIDE. Through the QC undertaken by the curatorial team and the automatic checks performed on data prior to submission these errors can be captured and corrected.

2.3.8.1 Scrambled peptide to spectrum matches

From an early submission into the PRIDE database, a unique and interesting phenomenon can be seen by plotting the precursor ion 'Delta m/z' (the difference between the reported precursor ion m/z and the m/z of the identified peptide). Mass spectrometers are highly accurate instruments with a typical mass accuracy around ±4 Daltons for old ion trap instruments and ±2 Daltons for newer ion trap instruments. The expected distribution of the difference between precursor ion m/z and identified peptide m/z (calcualted *in silico* from the amino acid composition) should therefore center sharply around zero.

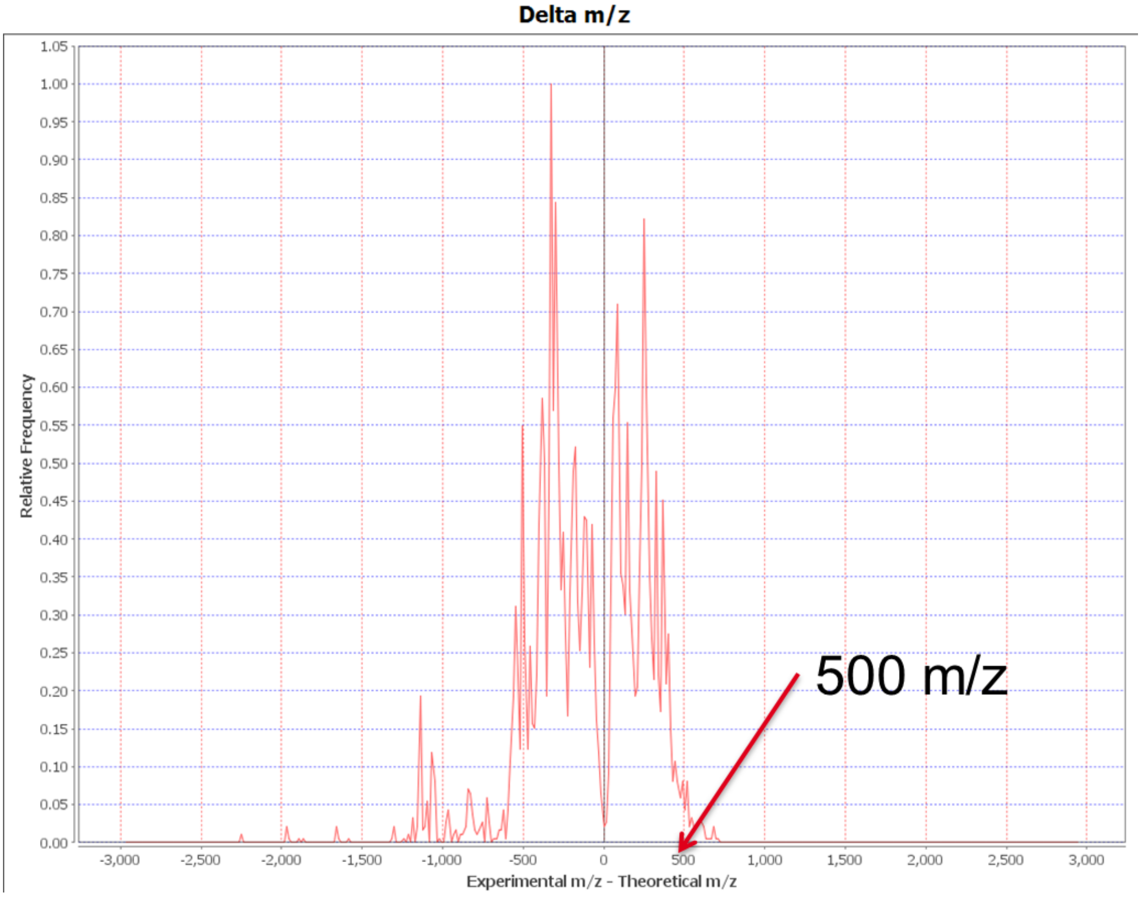
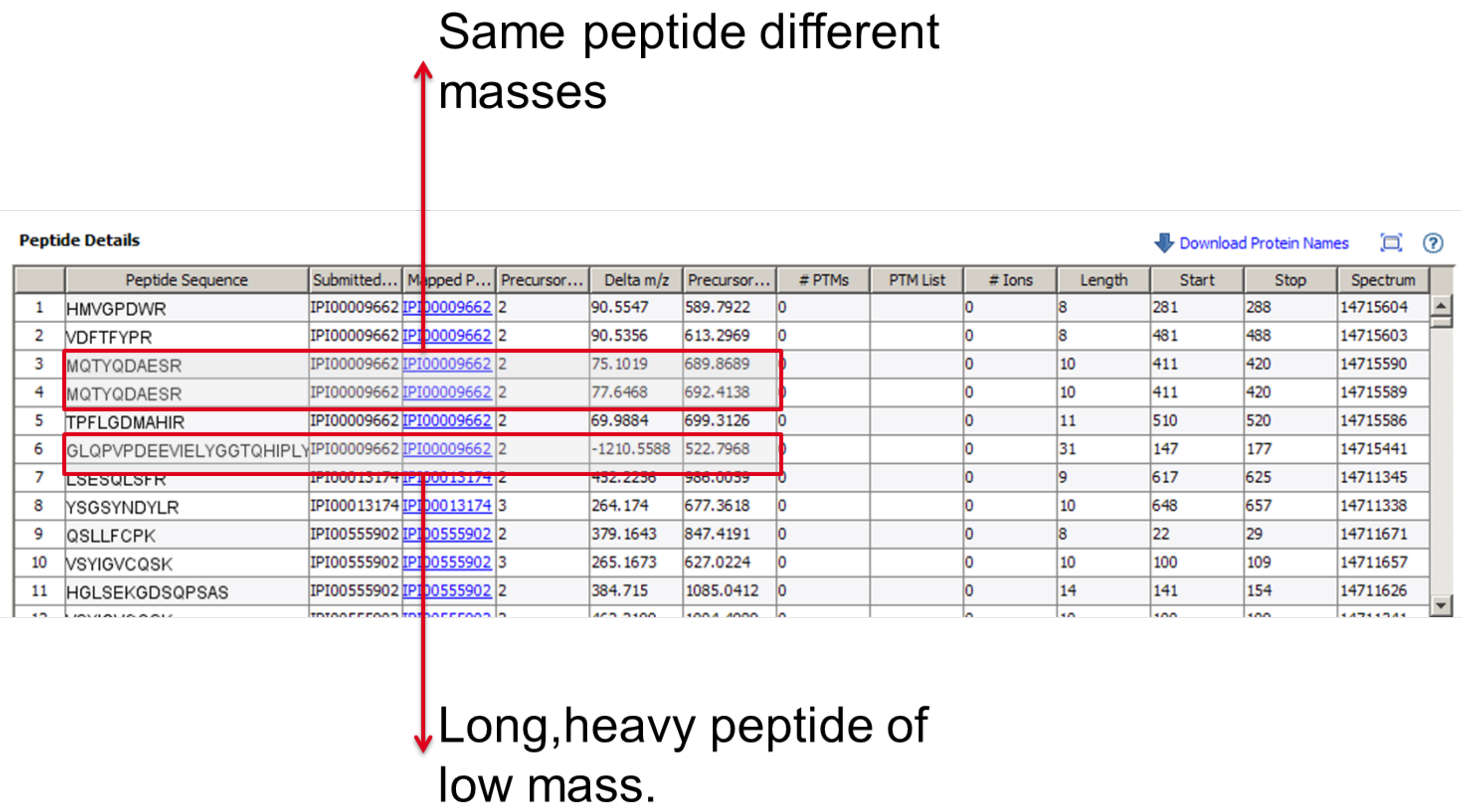


Figure 2.18: The distribution plot of the difference between precursor ion m/z and identified peptide m/z; part of a series of PRIDE Inspector QC charts. The distribution should center sharply around zero indicating very little difference between theoretical peptide mass and reported peptide mass.

Clearly the m/z error distribution from the experiment in Figure 2.18 is far from expected, with frequently occurring m/z differences in the hundreds of Daltons. This discrepancy is so large that it is unlikely to have been derived from an instrument fault or calibration error. It is much more likely that post acquisition data processing is at fault. To identify the source of the errors the “peptide view” of PRIDE inspector is employed.


Same peptide different masses
Long,heavy peptide of low mass.
Download Protein Names

Peptide Details

	Peptide Sequence	Submitted...	Mapped P...	Precursor...	Delta m/z	Precursor...	# PTMs	PTM List	# Ions	Length	Start	Stop	Spectrum
1	HMVGPDWR	IPI00009662	IPI00009662	2	90.5547	589.7922	0		0	8	281	288	14715604
2	VDFTFYPR	IPI00009662	IPI00009662	2	90.5356	613.2969	0		0	8	481	488	14715603
3	MQTYQDAESR	IPI00009662	IPI00009662	2	75.1019	689.8689	0		0	10	411	420	14715590
4	MQTYQDAESR	IPI00009662	IPI00009662	2	77.6468	692.4138	0		0	10	411	420	14715589
5	TPFLGDMAHIR	IPI00009662	IPI00009662	2	69.9884	699.3126	0		0	11	510	520	14715586
6	GLQPVPDEEVIELYGGTQHIPLY	IPI00009662	IPI00009662	2	-1210.5588	522.7968	0		0	31	147	177	14715441
7	LSESQLSFR	IPI00013174	IPI00013174	2	452.2256	986.0039	0		0	9	617	625	14711345
8	YSGSYNDYLR	IPI00013174	IPI00013174	3	264.174	677.3618	0		0	10	648	657	14711338
9	QSLLFCPK	IPI00555902	IPI00555902	2	379.1643	847.4191	0		0	8	22	29	14711671
10	VSYIGVCQSK	IPI00555902	IPI00555902	3	265.1673	627.0224	0		0	10	100	109	14711657
11	HGLSEKGDSQPSAS	IPI00555902	IPI00555902	2	384.715	1085.0412	0		0	14	141	154	14711626

Figure 2.19: The “peptide view” of PRIDE Inspector.

From the "peptide view" in figure 2.19 it is possible to see identical peptides showing different recorded m/z values, suggesting the error is not the fault of the database search engine. Similarly, a very large peptide of high theoretical m/z is shown to have a very low reported m/z. While this could be explained by a very high charge state, the precursor ion charge column clearly shows it to be a 2+ ion. Browsing the experiment metadata revealed that this submission was created by processing the search results with an early, unstable version of PRIDE converter (Barsnes *et al.*, 2009) (An application for transforming a multitude of accepted search engine result data formats into the internally managed PRIDE XML data format). Upon conversion the peptide to spectrum mapping was irrecoverably scrambled, and spectrum IDs assigned to peptides were mixed up, ultimately causing the extremely large discrepancies in delta m/z. To fix this problem the theoretical mass of each peptide can be calculated, for unique masses the peptide identification can be assigned to the correct spectrum by its reported precursor ion mass. This will solve a some of the peptide-spectrum mismatches in the dataset depeding on the precursor ion accuracy, which affects the number of unique peptides. For the peptides which are reported multiple times or isobaric peptides (with machine mass tolerence accounted for) a set of reasonable spectra can be assigned to each group. However, a concrete assignment of a single spectrum is impossible. The underlying problem has long since been rectified by the developers of PRIDE converter. But it is important to appreciate that repositories of public experimental data like PRIDE are pioneering efforts which sometimes involve the use of unfinished software, without which they could not be tested and fixed.

2.3.8.2 Residue double modification

Recent efforts in the PRIDE curation team have focused largely upon the correct annotation of peptides with their modifications. Post translational modifications represent an important facet of proteomics research, providing insight into biological phenomena such as cell signalling, protein localisation and degradation. However, they introduce an extra layer of complexity in proteomics data that must be accurately described when submitting data to a public repository, since erroneously submitted or missing modifications can completely alter or obviate biological conclusions derived from the data.

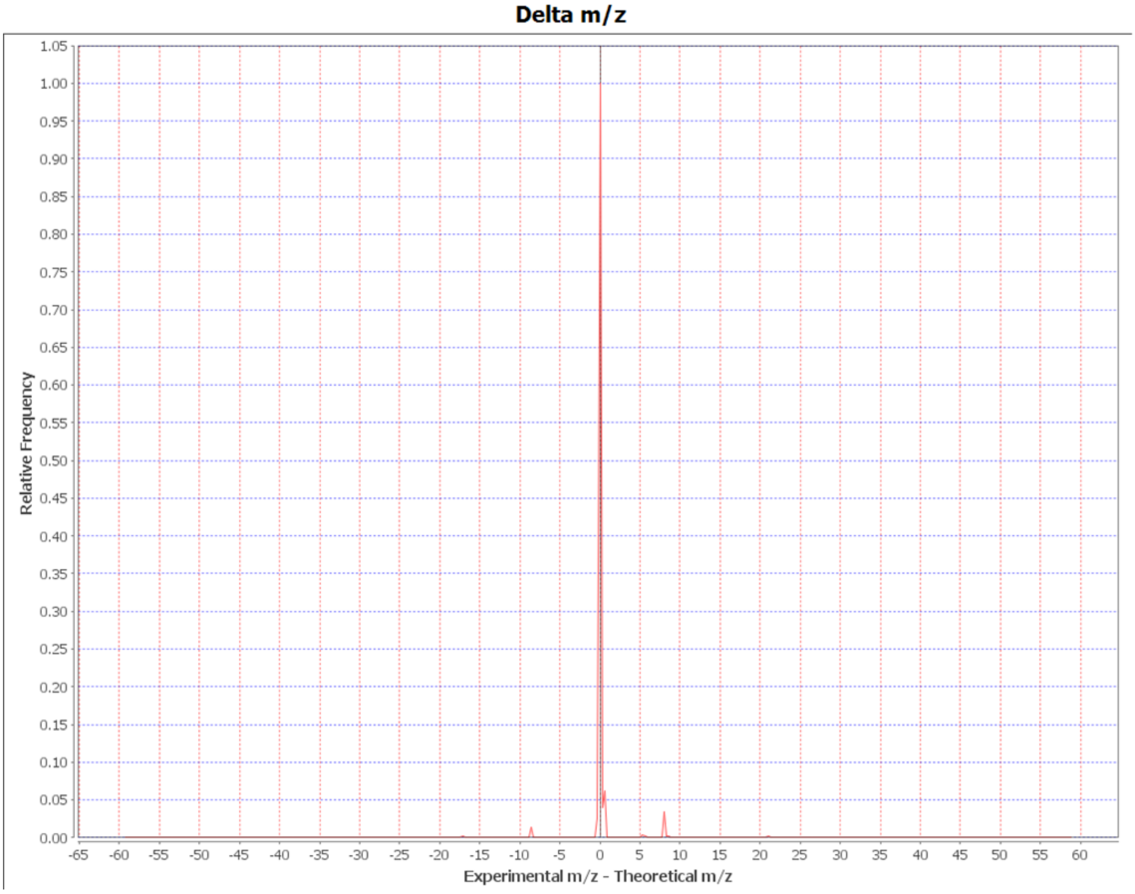


Figure 2.20: The distribution of the difference between precursor ion m/z and identified peptide m/z plot, part of a series of the PRIDE Inspector charts. The distribution is centered sharply around zero as expected, but two symmetrical peaks at 8 and −8 are cause for concern.

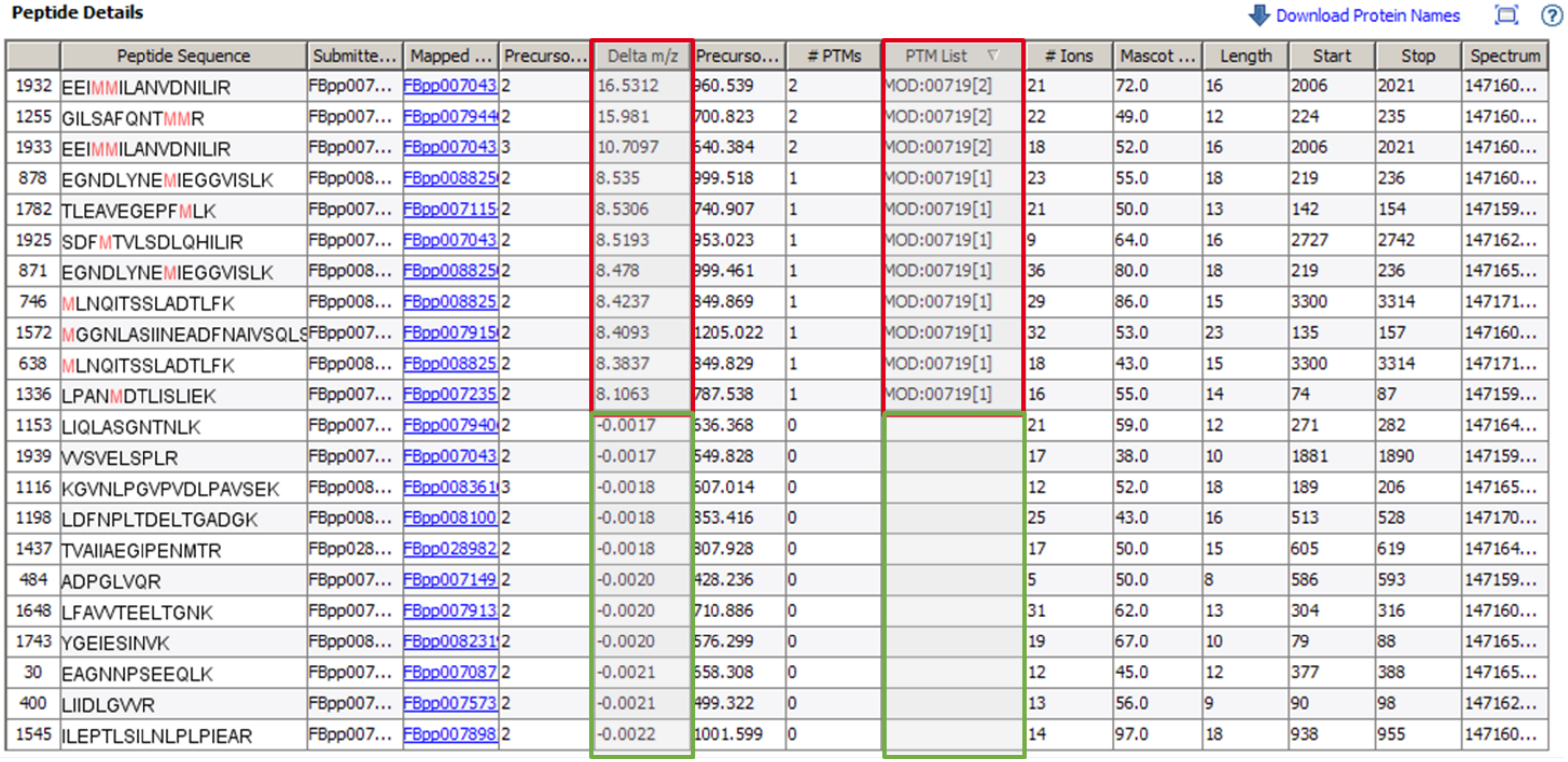
Figure 2.20 displays a delta mass distribution that appears as expected, centering sharply around zero. However, there are two symmetrical peaks around zero at 8 and -8. These typically represent an unreported methionine oxidation (mass delta of 16 Da) for peptide identifications with charge state 2+ (the most common charge state). Here too, the "peptide view" of PRIDE Inspector reveals the source of the error.

Figure 2.21 clarifies the existence of the anomaly and reveals its source. After sorting the columns by "Delta m/z", a selection of peptides highlighted in green at the bottom show acceptable "Delta m/z", while another set highlighted in red and located at the top display the erroneous delta m/z values. The peptides all have reported methionine oxidations (MOD:00719). However, even with the modification mass accounted for there is still a discrepancy of 8 Daltons for each charge 2+ identification with a single methionine residue and 16 Daltons for those with 2 methionine residues. Methionine oxidation has an approximate modification mass of 16 Daltons, this suggests that these methionine resiudes do not carry a single modification each, but infact carry two oxidations. This modification is known as L-methionine sulfone (MOD:00256) and in this case has been misreported. Utilising QC metrics like Delta m/z thus enables curators in the PRIDE team to quickly asses the quality of a submission and in cases like this, report any fault to the submitter who can quickly fix the submission. This problem would be easily rectified by including L-methionine sulfone as a variable modifcation in the search engine parameters.

2.3.8.3 No reported modifications

This final example highlights the importance of simple metadata annotation for submissions. Browsing the PRIDE Inspector QC charts in figure 2.22 revealed nothing untoward with the submission, good delta m/z distribution, excellent charge state distribution, etc.

Analysing the data more closely revealed a rather different story. Looking at the peptide view in figure 2.23 revealed that no modifications were reported in any of the peptides. In contrast with the metadata present in the "Protocol View" where an "N-ethylmaleimide derivitized cysteine" step was reported. This highly specific,



	Peptide Sequence	Submitte...	Mapped ...	Precurso...	Delta m/z	Precurso...	# PTMs	PTM List	# Ions	Mascot ...	Length	Start	Stop	Spectrum
1932	EEIMMILANVDNILIR	FBpp007...	FBpp007043	2	16.5312	960.539	2	MOD:00719[2]	21	72.0	16	2006	2021	147160...
1255	GILSAFQNTMMR	FBpp007...	FBpp007944	2	15.981	700.823	2	MOD:00719[2]	22	49.0	12	224	235	147160...
1933	EEIMMILANVDNILIR	FBpp007...	FBpp007043	3	10.7097	640.384	2	MOD:00719[2]	18	52.0	16	2006	2021	147160...
878	EGNDLYNEMIEGGVISLK	FBpp008...	FBpp008825	2	8.535	999.518	1	MOD:00719[1]	23	55.0	18	219	236	147160...
1782	TLEAVEGEPFMLK	FBpp007...	FBpp007115	2	8.5306	740.907	1	MOD:00719[1]	21	50.0	13	142	154	147159...
1925	SDFMTVLSDLQHILIR	FBpp007...	FBpp007043	2	8.5193	953.023	1	MOD:00719[1]	9	64.0	16	2727	2742	147162...
871	EGNDLYNEMIEGGVISLK	FBpp008...	FBpp008825	2	8.478	999.461	1	MOD:00719[1]	36	80.0	18	219	236	147165...
746	MLNQITSSLADTLFK	FBpp008...	FBpp008825	2	8.4237	849.869	1	MOD:00719[1]	29	86.0	15	3300	3314	147171...
1572	MGGNLASIINEADFNAIVSQLS	FBpp007...	FBpp007915	2	8.4093	1205.022	1	MOD:00719[1]	32	53.0	23	135	157	147160...
638	MLNQITSSLADTLFK	FBpp008...	FBpp008825	2	8.3837	849.829	1	MOD:00719[1]	18	43.0	15	3300	3314	147171...
1336	LPANMDTLISLIEK	FBpp007...	FBpp007235	2	8.1063	787.538	1	MOD:00719[1]	16	55.0	14	74	87	147159...
1153	LIQLASGNTNLK	FBpp007...	FBpp007940	2	-0.0017	636.368	0		21	59.0	12	271	282	147164...
1939	VVSVELSPLR	FBpp007...	FBpp007043	2	-0.0017	549.828	0		17	38.0	10	1881	1890	147159...
1116	KGVNLPGVPVDLPAVSEK	FBpp008...	FBpp008361	3	-0.0018	607.014	0		12	52.0	18	189	206	147165...
1198	LDFNPLTDELTGADGK	FBpp008...	FBpp008100	2	-0.0018	853.416	0		25	43.0	16	513	528	147170...
1437	TVAIIAEGIPENMTR	FBpp028...	FBpp028982	2	-0.0018	807.928	0		17	50.0	15	605	619	147164...
484	ADPGLVQR	FBpp007...	FBpp007149	2	-0.0020	428.236	0		5	50.0	8	586	593	147159...
1648	LFAVVTEELTGNK	FBpp007...	FBpp007913	2	-0.0020	710.886	0		31	62.0	13	304	316	147160...
1743	YGEIESINVK	FBpp008...	FBpp008231	2	-0.0020	576.299	0		19	67.0	10	79	88	147165...
30	EAGNNPSEEQLK	FBpp007...	FBpp007087	2	-0.0021	658.308	0		12	45.0	12	377	388	147165...
400	LIIDLGVVR	FBpp007...	FBpp007573	2	-0.0021	499.322	0		13	56.0	9	90	98	147162...
1545	ILEPTLSILNLPLPIEAR	FBpp007...	FBpp007898	2	-0.0022	1001.599	0		14	97.0	18	938	955	147160...

Figure 2.21: The peptide view of PRIDE Inspector, ordered by "Delta m/z". Identifications with the large Delta m/z are highlighted in a red box, this discrepancy is explained by the misreporting of methionine oxidation which should actually be the rarer double methionine oxidation known as L-methionine sulfone.

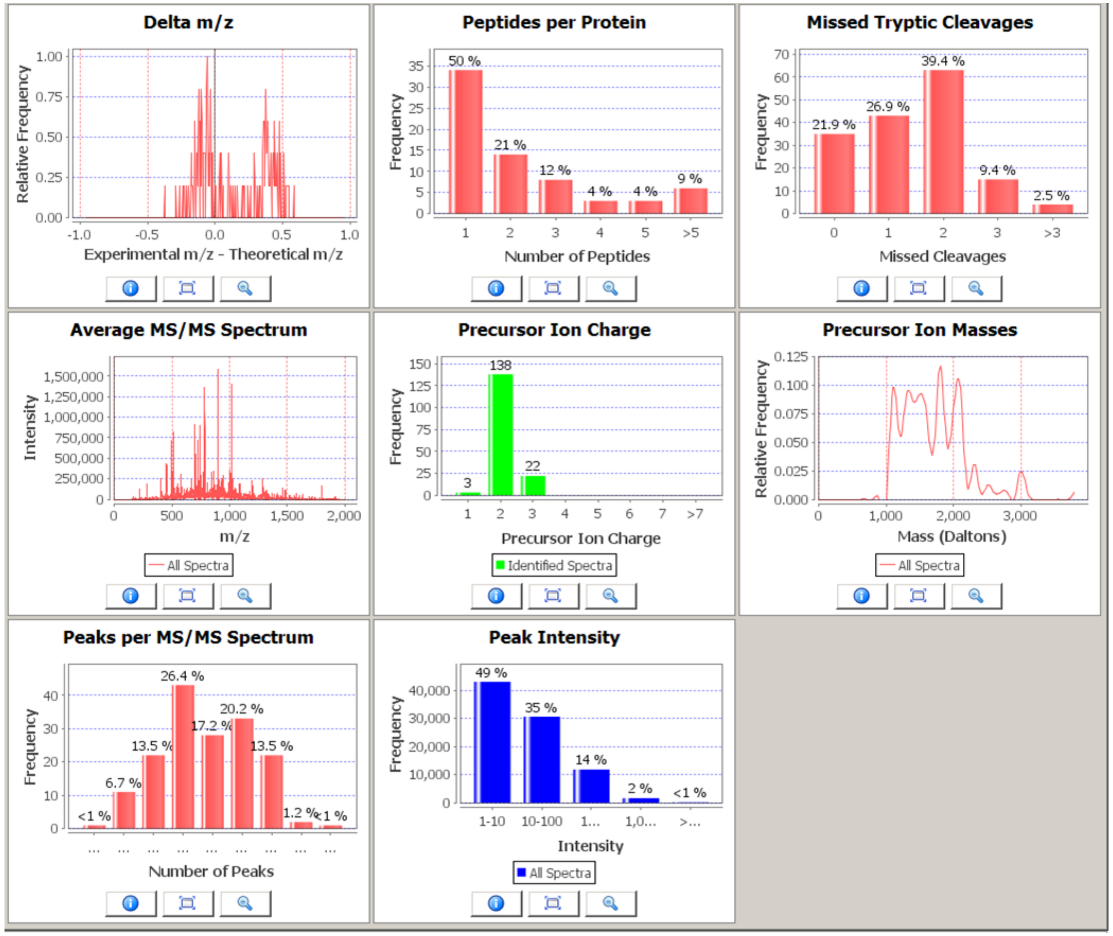
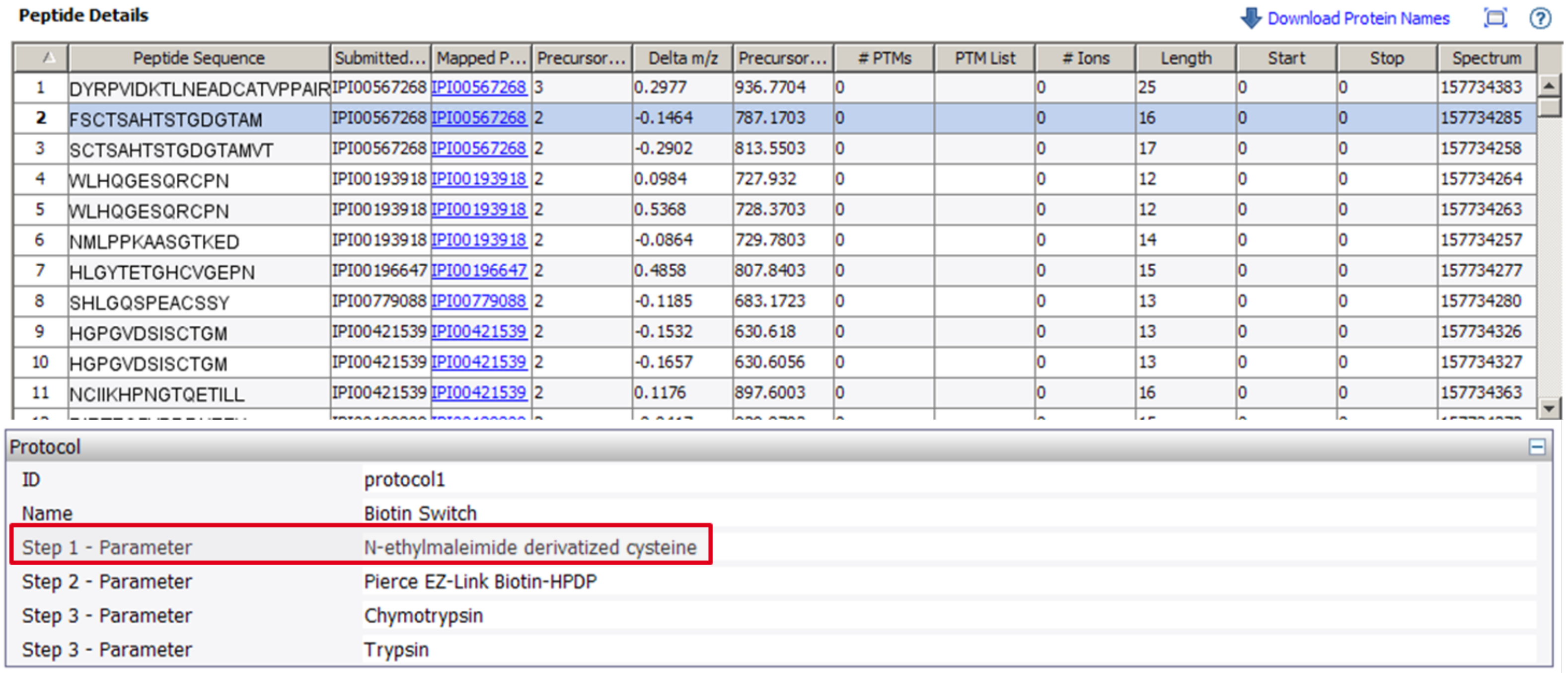


Figure 2.22: The PRIDE Inspector chart view, from top left to bottom right: delta m/z plot, peptides per protein frequency bar plot, missed tryptic cleavage percent bar plot, average ms/ms spectrum, precursor ion charge frequency bar plot, precursor ion mass distribution plot, percentage number of peaks per spectrum histogram and percentage peak intensity histogram.



Peptide Details

	Peptide Sequence	Submitted...	Mapped P...	Precursor...	Delta m/z	Precursor...	# PTMs	PTM List	# Ions	Length	Start	Stop	Spectrum
1	DYRPVIDKTLNEADCATVPPAIR	IPI00567268	IPI00567268	3	0.2977	936.7704	0		0	25	0	0	157734383
2	FSCTSAHTSTGDGTAM	IPI00567268	IPI00567268	2	-0.1464	787.1703	0		0	16	0	0	157734285
3	SCTSAHTSTGDGTAMVT	IPI00567268	IPI00567268	2	-0.2902	813.5503	0		0	17	0	0	157734258
4	WLHQGESQRCPN	IPI00193918	IPI00193918	2	0.0984	727.932	0		0	12	0	0	157734264
5	WLHQGESQRCPN	IPI00193918	IPI00193918	2	0.5368	728.3703	0		0	12	0	0	157734263
6	NMLPPKAASGTKED	IPI00193918	IPI00193918	2	-0.0864	729.7803	0		0	14	0	0	157734257
7	HLGYTETGHCVGEPN	IPI00196647	IPI00196647	2	0.4858	807.8403	0		0	15	0	0	157734277
8	SHLGQSPEACSSY	IPI00779088	IPI00779088	2	-0.1185	683.1723	0		0	13	0	0	157734280
9	HGPGVDSISCTGM	IPI00421539	IPI00421539	2	-0.1532	630.618	0		0	13	0	0	157734326
10	HGPGVDSISCTGM	IPI00421539	IPI00421539	2	-0.1657	630.6056	0		0	13	0	0	157734327
11	NCIIKHPNGTQETILL	IPI00421539	IPI00421539	2	0.1176	897.6003	0		0	16	0	0	157734363

Protocol

ID	protocol1
Name	Biotin Switch
Step 1 - Parameter	N-ethylmaleimide derivatized cysteine
Step 2 - Parameter	Pierce EZ-Link Biotin-HPDP
Step 3 - Parameter	Chymotrypsin
Step 3 - Parameter	Trypsin

Figure 2.23: Both the peptide view and protocol view of PRIDE inspector, detailing the identified peptides and the conditions under which the sample was handled.

reagent induced chemical modification affects cysteine residues with extremely high efficiency. Of the 163 peptides identified in the experiment, 155 contained cysteine yet had the mass of an unmodified peptide. Either these peptides represented the very few cysteine containing peptides for which the chemical modification did not occur, or they were false positives. Assuming they were false positives, it is likely that the spectra were not searched with N-ethylmaleimide derivitized cysteine as a modification parameter, since doing so would probably result in many hundreds more reliable identifications. This dataset would probably identify an order of magnitude more peptides were it re-searched with the appropriate modifications.

PRIDE inspector thus offers many ways to quickly assess the quality of a dataset and detect any potential problems that are not explained by the existing metadata. While correction of these metadata misannotations and missing annotations is currently a largely manual job, it is not difficult to imagine that future advances in understanding proteomics quality control will allow the automatic annotation of experiments. Hints at how this may be done are visible in figure 2.8 where the species of a sample can be estimated from its MS2 fragment ion Δ mass distribution. Likewise, the instrument used or search engines settings could be guessed from data analysis *a posteriori* as well. By aggregating datasets together that differ in very few experimental parameters, techniques like MDS, LSA and supervised machine learning approaches may be able to classify experiments with unknown parameters to give an estimate of the missing metadata.

2.4 Discussion

As public data repositories are getting increasingly populated with (published) proteomics datasets (Csordas *et al.*, 2012), large-scale data analysis becomes an ever more powerful tool for investigating and predicting the nuances implicit in proteomics methods and results. Yet the reliability of many downstream public data processing methods is crucially dependent on the validity of the data. So it is increasingly important to have properly matched, empirically derived reference metrics available for selecting and filtering the available datasets (described in section 2.2.4). This holds true for both computational users, as well as for database curators, since both have a vested interest in detecting, and possibly understanding, outlying datasets. The relative sparseness of the information currently deposited in public repositories compared to what is available in the lab during and immediately after sample processing and data acquisition (highlighted in section 2.1.1), requires the development of robust metrics that can be derived from the available data, preferentially as close to the acquisition point as possible. Correspondingly, I have illustrated here that publicly available data, spanning many individual experiments of diverse origin, can be *a posteriori* examined according to several easily and reliably obtainable metrics for a typical proteomics work flow. This includes analyses performed within the context of a larger study, as was shown for the protein depletion, peptide selection, and SCX separation procedures used in the HUPO PPP2 dataset (Liu *et al.*, 2006; Qian *et al.*, 2008) (see section 2.3.1), but also extends across individual studies, where a large body of only very loosely related experimental data (e.g., selected based solely on the enzyme used for proteolytic digestion) can be used to estimate empirical background distributions complete with tolerance ranges (shown in section 2.3.2). Furthermore, the obtained metrics can be used in surprising ways to compare experimental metadata annotations, as was illustrated for the m/z differences between MS2 peaks. The latter point is more than a gimmick; missing or incorrect annotation constitutes a serious downstream problem for data consumers and the ability to detect possible misannotation or to assign annotation where none is given will most likely be an important curatorial function for repositories in the years to come. The mass spectra can also be mined for additional data. Mass distributions of precursor and product ions can relate important information about biases or faults in the protocol (sections 2.3.2 and 2.3.3 respectively), while the isobaric labelling of peptides for MS2-based quantification approaches can be inspected easily as well. In the latter case, the possible trade-off between quantification and identification efficiency can be monitored by comparing the reporter ion intensities against the average intensity of the top 10% most intense non-reporter peaks (shown in section 2.3.4).

With proteomics coming ever more into the limelight of the life sciences as a powerful and sensitive analytical platform, the need for robust quality control practices is becoming ever more pressing. Such quality control must in most cases also extend beyond a single mass spectrometry analysis, necessarily encompassing several runs within or even across experiments or studies. As a result, metrics need to be obtained which can function at the level of the individual run, but also across many runs. The latter can directly benefit from already acquired data for the establishment of acceptance criteria. These criteria are of course open to interpretation and will depend on the downstream use case for the data. In time, clear guidelines will be established within the community to assess the quality of data prior to public deposition with tools like PRIDE Inspector (Wang *et al.*, 2012) and the approaches developed in the PRIDE team may inform such guidelines. Greater exposure of QC concepts to the wider proteomics community will hopefully transform it into a more expert and data complexity aware field ready to face upcoming challenges. The future of quality control in proteomics therefore is set to go hand in hand with that of data repositories and the standardized deposition of well annotated datasets.

In the next chapter I will expand upon the specifics of semantic validation and the use of defined nomenclature, in the context of another 'omics' field. I will concentrate on the field of lipidomics, where core bioinformatics and data sharing concepts are not yet well established and many lessons can be learned from existing efforts in the field of proteomics.

Chapter 3

LipidHome

3.1 Introduction

Central to all 'omics' disciplines is the identification of biological molecules followed by their accurate description and reporting. Key to this research is a centralized namespace against which to identify the molecules. Fields such as proteomics have a long history of providing databases of proteins, with standardised naming formats and intelligent mappings that relate protein names in one format to another. Perhaps the prime example of this is the UniProt Consortium's KnowledgeBase (UniProtKB) resource (Magrane & Consortium, 2011; UniProt-Consortium, 2011), self described as a comprehensive resource for protein sequence and annotation data'. It provides not only protein sequence information for a multitude of organisms but a considerable amount of metadata for those sequences, ranging from functional to domain/structural to disease related annotations and supporting evidence in the literature. UniProtKB is composed of 2 sections:

UniProtKB/Swiss-Prot Contains manually annotated protein records retrieved from the literature and expert reviewed computationally inferred annotations.

UniProtKB/TrEMBL A much larger set of unreviewed protein records typically theoretically translated from the genome that are automatically annotated entirely computationally with no human curation.

There is clearly some interplay between these two services where over time, likely candidates from UniProtKB/TrEMBL are manually validated by UniProtKB curators and upgraded into the UniProtKB/Swiss-Prot database. This separation of the plausible but theoretical from the validated is an important aspect of this type of catalogue, offering both a comprehensive view of the potential pool of identifiable molecules and a more conservative view of just the validated molecules. Each database has its specific use cases and neither can be argued as strictly better or worse than the other. What is important is the clear separation between them so that they can be used for the slightly different purposes intended. UniProt has formed a centralised hub, promoting the use of standard data formats, nomenclature, annotation protocols and auxiliary services around which proteomics has built itself and prospered. Genomics has similarly well developed resources in Ensembl (Flicek *et al.*, 2011; Kersey *et al.*, 2012) and its genome browser which provides automatic annotation of the genomes of several model organisms and organisms under significant levels of study. The emerging field of lipidomics is yet to have access to such a mature resource as UniProt. While products exist, they lack several important aspects of a resource like UniProt.

The LipidomicNet Consortium (`www.lipidomicnet.org`) is funded by an EU FP7 project with approximately 30 partners whose aims are to advance many aspects of the field of lipidomics, including novel experimental approaches, standardised data analysis, developing new experimental reagents and encouraging the use of standard data formats and nomenclature. As a result of the LipidomicNet Cambridge Bioinformatics workshop April 2010 and a presentation I gave there, it was agreed that a list of lipids detected within the consortium must be created so that the detectable lipidome could be estimated. At the same time I suggested that the theoretical lipidome be calculated based upon some agreed bounds to that chemical space. It is the aim of this chapter to describe work undertaken to create and present a solution to the storage of a catalogue of lipids. Throughout this chapter special consideration will be given to existing biological molecule reference databases for the creation of a new resource aimed specifically for the high throughput lipidomics community.

3.1.1 Mass spectrometry lipidomics

In section 2.1.1 of chapter 2 a typical proteomics mass spectrometry workflow is outlined. Following separation of peptides on a HPLC system a sample undergoes spectra acquisition in the mass spectrometer. This process typically involves tandem MS, where peptide ions are isolated and fragmented to record MS2 spectra, from which peptide identifications can be predicted. Identified peptides are assembled into partial proteins, and the most likely proteins reported as identifications from the sample. The resolution of detecting amino acids and often a selection of common modifications is the convention to almost all modern day proteomics experiments. While there are many parallels with mass spectrometry based identification of lipids the are some differences, particularly in the wide spread use of manual spectrum interpretation. The reported identifications typical of most lipidomics mass spectrometry experiments do not fully resolve the structure of the lipid. They regularly identify the main class (e.g glycerophosphocholine, largely resolved by the solid phase extraction step) and the total number of carbons and double bonds in the constituent fatty acids (e.g. 34:2 , 34 carbons and 2 double bonds spread in in unknown way among the two fatty acid attached to the glycerol back bone). More complex experiments may reveal the individual fatty acid components to a similar resolution (number of carbons and double bonds of fatty acids known, but the position of double bonds or *sn* positions of fatty acids upon the glycerol remain unknown) (Yoshinaga *et al.*, 2011). This process is described in figure 3.1. Further structural characterisation is achieved with highly specialist reagents and tools. One such method is the use of fragmentation in the presence of ozone to locate the double bond positions in fatty acid chains (Brown *et al.*, 2011), others include repeated ion selection and fragmentation (Hsu & Turk, 2008).

3.1.2 Nomenclature

As proposed in the LipidomicNet Cambridge Bioinformatics Workshop, four partners completed and returned a list of lipids identified in their laboratories (See Appendix 1 for the standardised documents). This document raised some clear points about the state of lipidomics as a whole and the disparity in reporting standards between different labs. The main points raised were:

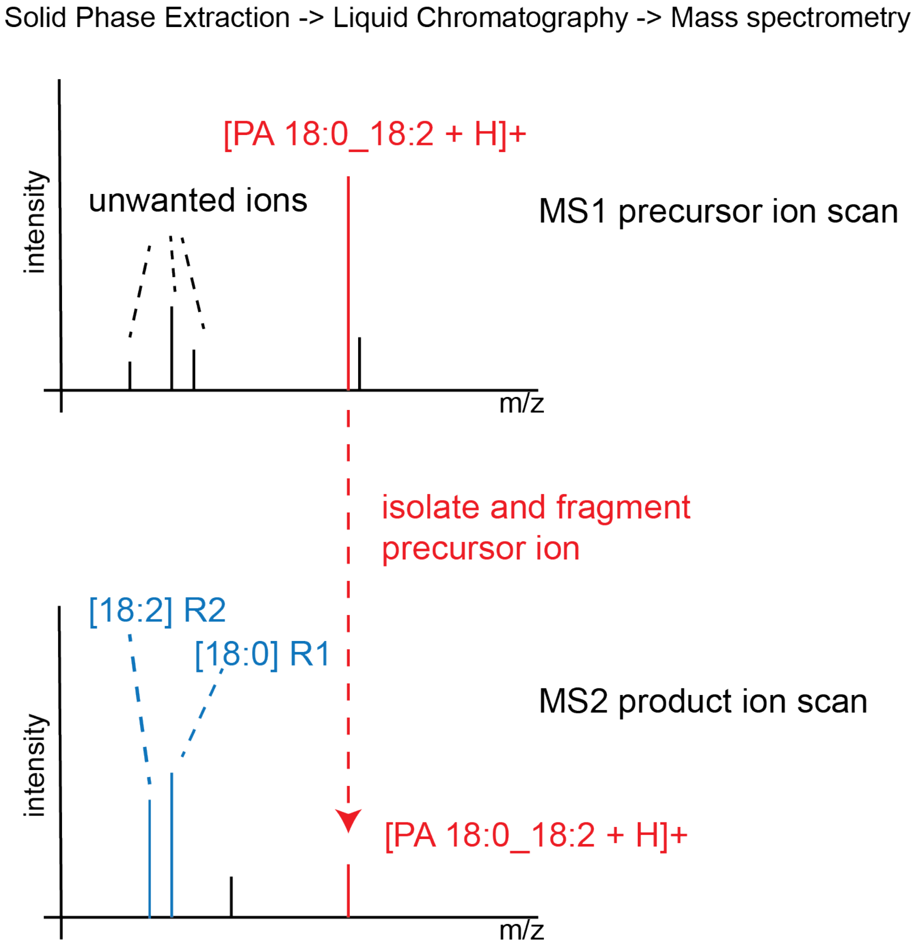


Figure 3.1: A diagram of MS1 and MS2 spectra aquisition and interpretation. Following solid phase extraction (SPE) in which crude fractions of a single lipid main class are isolated and High Performance Liquid Chromatography (HPLC) used to separate individual lipid species, lipids enter the mass spectrometer to undergo a process of selection, fragmentation and detection. In an initial scan, a precursor ion is isolated inside the mass spectrometer. The ion is then fragmented, primarily by CID. The resulting fragments are detected and the raw spectra recorded. Assuming unique isolation of a lipid after precursor ion scan, the fragment masses detected in the product ion scan can be assigned to constituents of the identified precursor ion. Glycerophospholipids typically yield intense fragment ions of their fatty acid moieties and hence product ions can be used to determine the fatty acid constituents of a lipid species such as PC 36:2 detection in the precursor ion scan. Position of the respective fatty acids has a tendency to be reflected in the product ion intensities, the R1 fatty acid being more intense than the R2 (Hsu & Turk, 2009)

.

Lipidome coverage The four laboratories made a combined total of 902 unique lipid identifications. The overlap in which a lipid was identified in all laboratories was only 65 (7.2%), while the experimental protocols differed and the the organism under study varied, the overlap was quite minimal.

Lipid name dialects Of the lipids recorded in the consortium by multiple different partners, a variety of naming dialects were in use. As an example from the document, glycerophosphocholine species with a single alkyl linkage and a single acyl linkage(as opposed to the more common diacyl linkages) totaling 36 carbons and 2 double bonds were described as: aPC 36:2, 36:2-aPC and PC a 36:2. Similarly the glycerophosphocholine main class was reported as PC, GPC and GlyPC, the case for Lysoglycerophospholipid was even more heterogeneous with LPC, LyPC, LysoPC and PC 0:0/18:1.

Lipid identification resolution Labs within the consortium are identifying lipids at two levels of structural resolution; precursor ion scans, where only the total number of carbons and double bonds in the lipid's fatty acids are known and product ion scans, where number of carbons and double bonds for each individual fatty acid are resolved. No labs reported lipid identifications at any higher resolution where the positions of double bonds could be detected.

Based upon work by the LIPID MAPS Consortium (Fahy *et al.*, 2005), an identification hierarchy was created that described all the information present in the various levels of lipid identification (see figure 3.2).

From all these different levels of structural detail, it became clear that reporting lipid identifications is not as simple as reporting a peptide sequence. A comparable level of complexity is found in protein post-translational modifications where additional detail may be recorded and must be displayed in a consistent and informative manner to reflect the complexity of the identification. With the levels of reported detail in mind, appropriate nomenclature must be developed to report this detail. While the reported lipid dialects were very similar and the structural resolution of identifications different, an expert would easily understand identifiers from different dialects and resolutions, however this process is much more difficult for a computer. To resolve this problem and standardise the reporting of lipid identifications across

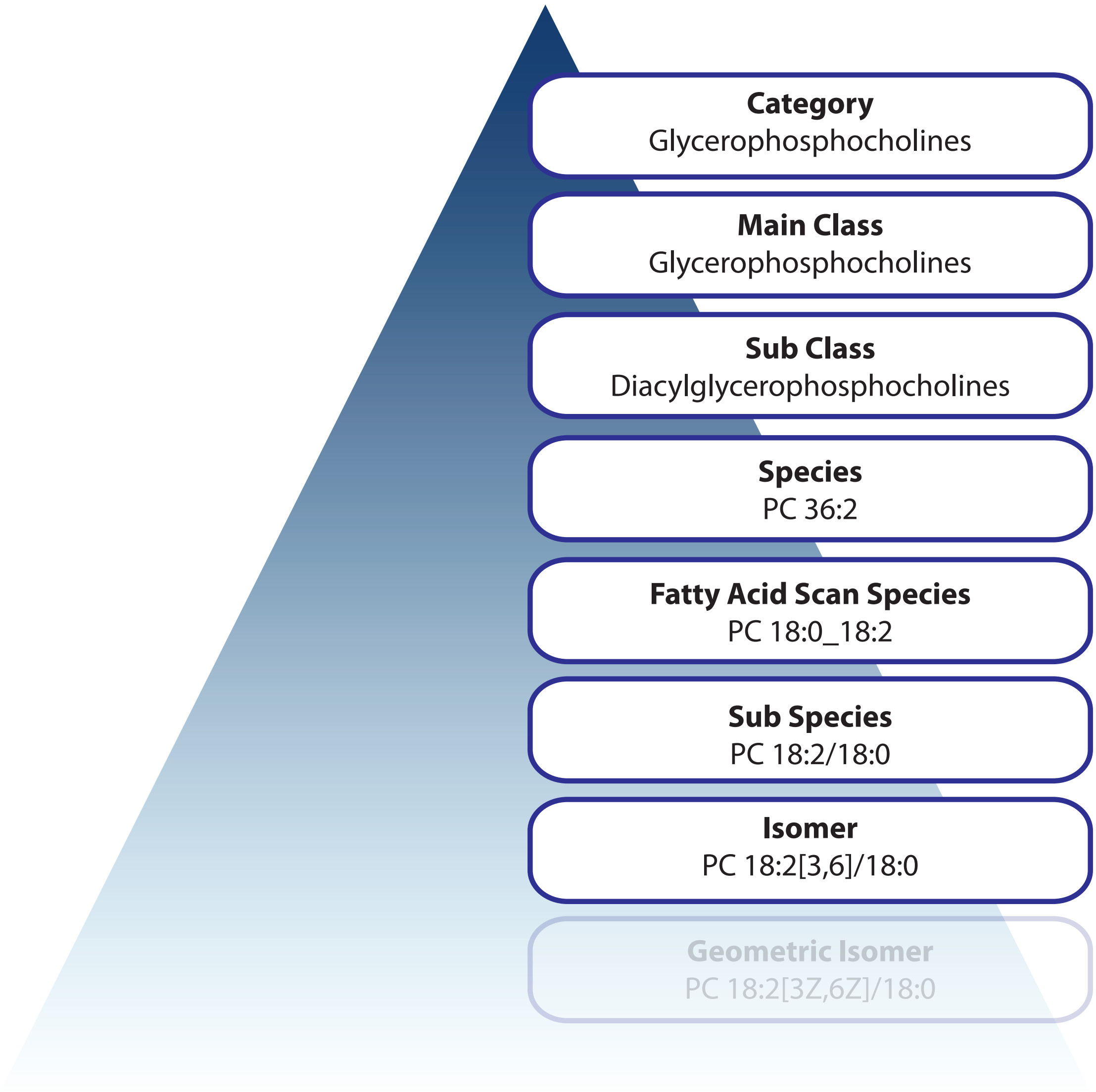


Figure 3.2: Lipids can be arranged into a hierarchy denoting the structural specificity of the identified lipid. Each level of the structural hierarchy has a name and an example. The structural resolution of the lipid increases as the hierarchy is descended. The transparent Geometric Isomer level is not supported by LipidHome because this level of strucutral detail is never achieved in high throughput lipidomics studies. However, it is supported in the LIPID MAPS Structural Database.

the consortium a nomenclature committee was established. Its aim was to create a nomenclature for the various levels of lipid identification. This work was successfully achieved but is under review currently. However, a brief overview is described here as it will form the basis of the database of lipid structures:

Species Species have the format: "*Headgroup carbons*:*double_bonds*", e.g. PC 36:2. Any known non-acyl fatty acid linkages are reported like so: "*Headgroup linkages carbons*:*double_bonds*", where linkages can be quantified by "*m,d,t* or *e*" (mono, di, tri, tetra respectively) and the non acyl linkages themselves can be any of "*a*" for alkyl or "*p*" for plasmalogen (alkenyl/acyl) e.g. PC da 36:2.

Fatty acid scan species Following the same rules as species these identifications differ by having known fatty acid constituents, but unknown *sn* positions of the fatty acids. Fatty acids with unknown positions are separated by "_" e.g PC 18:0_18:2. The same rules for defining linkages apply as for species.

Sub species With the position of the fatty acids known, sub species are written in *sn* position order separated by "/" e.g. PC 18:0/18:2. At this resolution non-acyl linkages are no longer written as before, the position of the linkages is reported as part of the fatty acid, where the linkages are defined by "*O*-" for alkyl and "*P*-" for plasmalogen type bonds. e.g. PC O-18:0/O-18:2.

Isomer Isomers are the highest resolution structure covered by the nomenclature, similar to sub species except that double bond positions are defined as comma separated integers within square brackets e.g. PC 18:0/18:2[3,6].

Proteomics faces a very similar problem, several protein databases exist, each with distinct identifiers for the same amino acid sequence, for example the 'Cellular tumour antigen p53' has the primary UniProt accession P04637 along with 26 secondary UniProt accessions, 16 International Protein Index (IPI) (www.ebi.ac.uk/IPI/) accessions and 18 NCBI RefSeq identifiers (Pruitt *et al.*, 2009). Clearly that is a lot of different names for the same thing. Lipidomics is not yet in that situation and is unlikely to face such a large challenge as lipid names convey the general structure of the molecule whereas proteins, due to the length of their amino acid sequence cannot be relayed in such a short name and cryptic accessions must be used instead. The field of proteomics has overcome the challenge with tools like the Protein Identification Cross Reference service (PICR) (Côté *et al.*, 2007). This web application accepts protein identifiers, amino acid sequences and some gene identifiers as input and returns a list of all known identifiers from the selected protein database resources. In the future it will be necessary for tools to be written to translate common and laboratory specific lipid nomenclatures into one standardised nomenclature to facilitate the adoption of the standardised lipid nomenclature and the universal interpretation of lipidomics data.

3.1.3 Extant lipid databases

With the requirements of the consortium, the structural hierarchy and nomenclature of lipid identifications clear, an overview of existing technologies in the field was undertaken. Three major resources are in common use, LIPID MAPS (Sud *et al.*, 2007) produced by the Nature Publishing Group and the LIPID MAPS Consortium (`www.lipidmaps.org`), Lipid Bank (Yasugi & Watanabe, 2002) a product of The Japanese Conference on the Biochemistry of Lipids `http://lipidbank.jp/` and a commercial catalogue of lipid products available from Avanti Polar Lipids Inc (`http://avantilipids.com/`).

3.1.3.1 LIPID MAPS

LIPID MAPS (Sud *et al.*, 2007)(`www.lipidmaps.org`) is perhaps the most comprehensive resource for lipidomics related information. It contains not only a database of lipid structures, but also a whole host of other features including references to recently published literature, lipid drawing tools, standard identification protocols and reference spectra, amongst others. It also provides a classification system for lipids that is recognised by the majority of the lipidomics community (Fahy *et al.*, 2005, 2009). Of most interest to this project is the LIPID MAPS Structure Database (LMSD): A database of lipid structures with associated metadata. It supports lipid identification at the geometric isomer structural hierarchy level, a level lower than isomer described in figure 3.2 and has a structural hierarchy equivalent to that shown in figure 3.3. Geometric isomer identifications are defined by known stereochemistry of double bonds e.g PC 18:0/18:2[3Z,6Z].

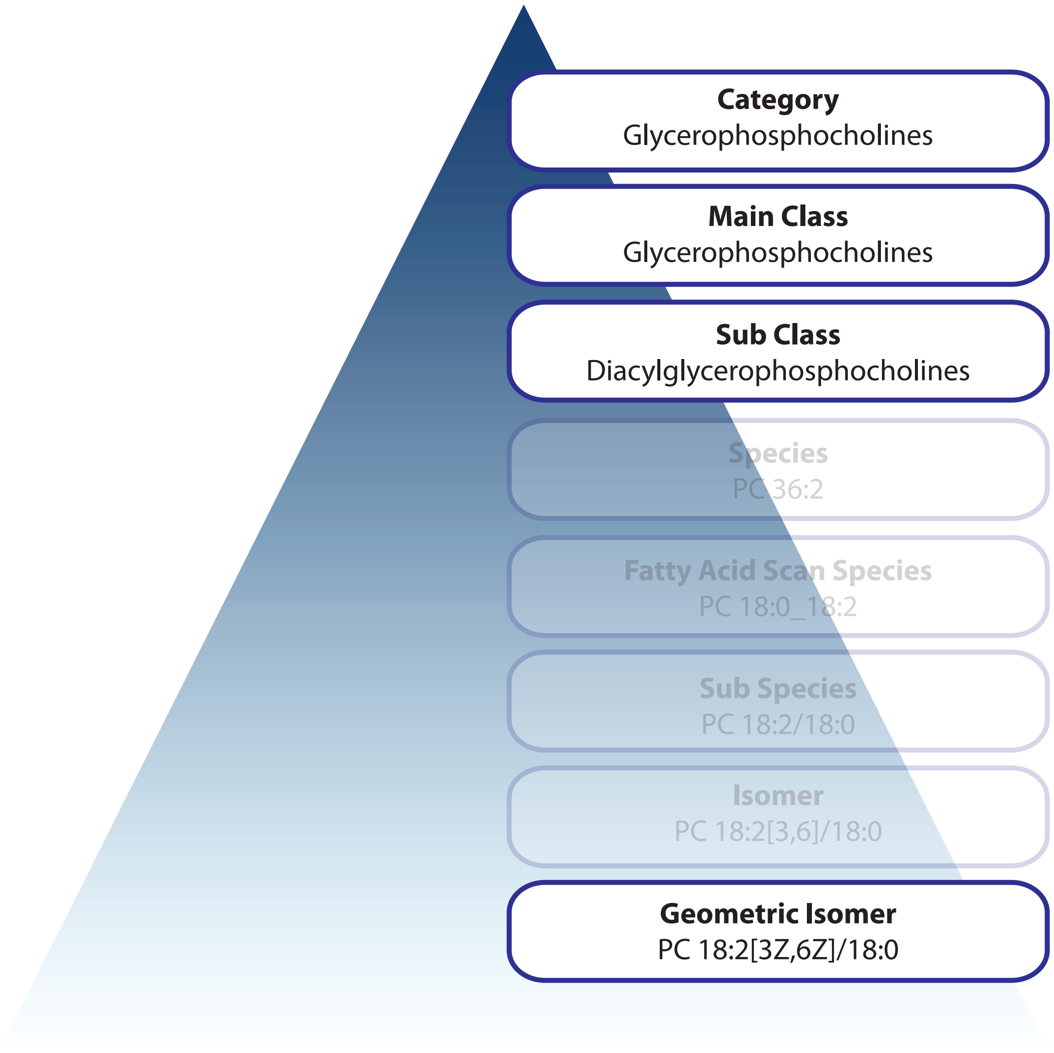


Figure 3.3: The lipid structural hierarchy of the LIPID MAPS Structural Database. Similar to the lipid structural hierarchy in figure 3.2, Category and Main Class represent the same identification resolution. Sub Class has the subtle difference that in LIPID MAPS the fatty acid linkage (e.g. acyl) is defined for each position. In the LipidHome structural hierarchy the position of linkages is not defined e.g. the LIPID MAPS sub classes 1-alkyl,2-acylglycerophosphocholines and 1-acyl,2-alkylglycerophosphocholines are combined into the composite sub class monoacyl,monoalkylglycerophosphocholines in LipidHome. LIPID MAPS also store lipid identification as the level of geometric isomers, a level below the lowest (isomer) proposed in 3.2. However, lipids are not systematically stored at the transparent levels species, fatty acid scan species or sub species.

It has the following main features:

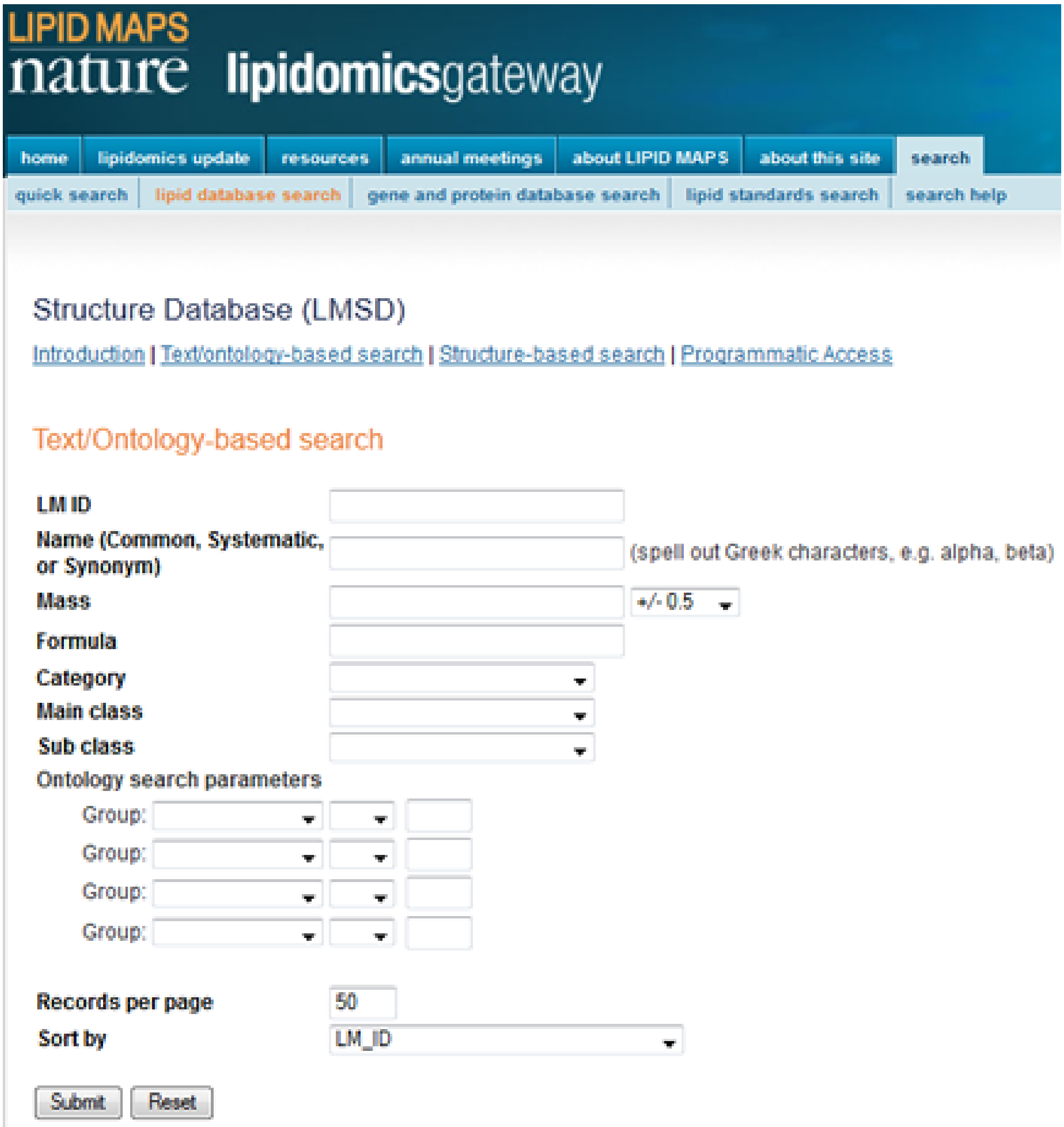
Searching for a lipid After selecting "Text/Ontology search" the user is presented with a selection of drop down boxes, which progressively narrow down the class of the lipid under search. Additional "Ontology search parameters" may be specified such as number of carbons/ double bonds, see figure 3.4.

Search results The search results are displayed in a table showing: ID, name, systematic name, formula, mass, main class and sub class. The ID of each row can be clicked to link to a more detailed page of a particular record. The results are at a structural heirarchy level even lower than isomer. In this level the bond positions are defined but also the stereochemistry of those double bonds (Z -cis, E -trans), these are known as geometric isomers. See figure 3.5.

Result details Clicking a LMID redirects the user to a more detailed view of the lipid record where a 2D image is displayed in addition to the IUPAC International Chemical Identifier (InChI) (urlhttp://www.iupac.org/inchi/); a system for describing the two dimensional structure of a molecule as an encoded string. Cross references from a variety of external databases are reported including; LIPIDAT (Caffrey & Hogan, 1992), PubChem (Wang *et al.*, 2009), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2012), The Human Metabolome Database (HMDB) (Wishart *et al.*, 2009) and Chemical Entities of Biological Interest (ChEBI) (de Matos *et al.*, 2010). In addition a selection of calculated chemical descriptors describe some feature of the molecule, all seen in figure 3.6.

Programmatic access Programmatic access is provided via two scripts that can return a single detailed record based on an LMSD ID or multiple simple records based upon search parameters shown in figure 3.4. The results can be retrieved as Comma Separated Value(CSV), Tab Separated Value (TSV) or Structure-Data File (SDF) (Dalby *et al.*, 1992).

Contrasting to the hierarchy structure in figure 3.2 highlights the key difference between LIPID MAPS and the identifications common in the lipidomics community.



Structure Database (LMSD)

Text/Ontology-based search

LM ID

Name (Common, Systematic, or Synonym) (spell out Greek characters, e.g. alpha, beta)

Mass +/- 0.5

Formula

Category

Main class

Sub class

Ontology search parameters

Group:

Group:

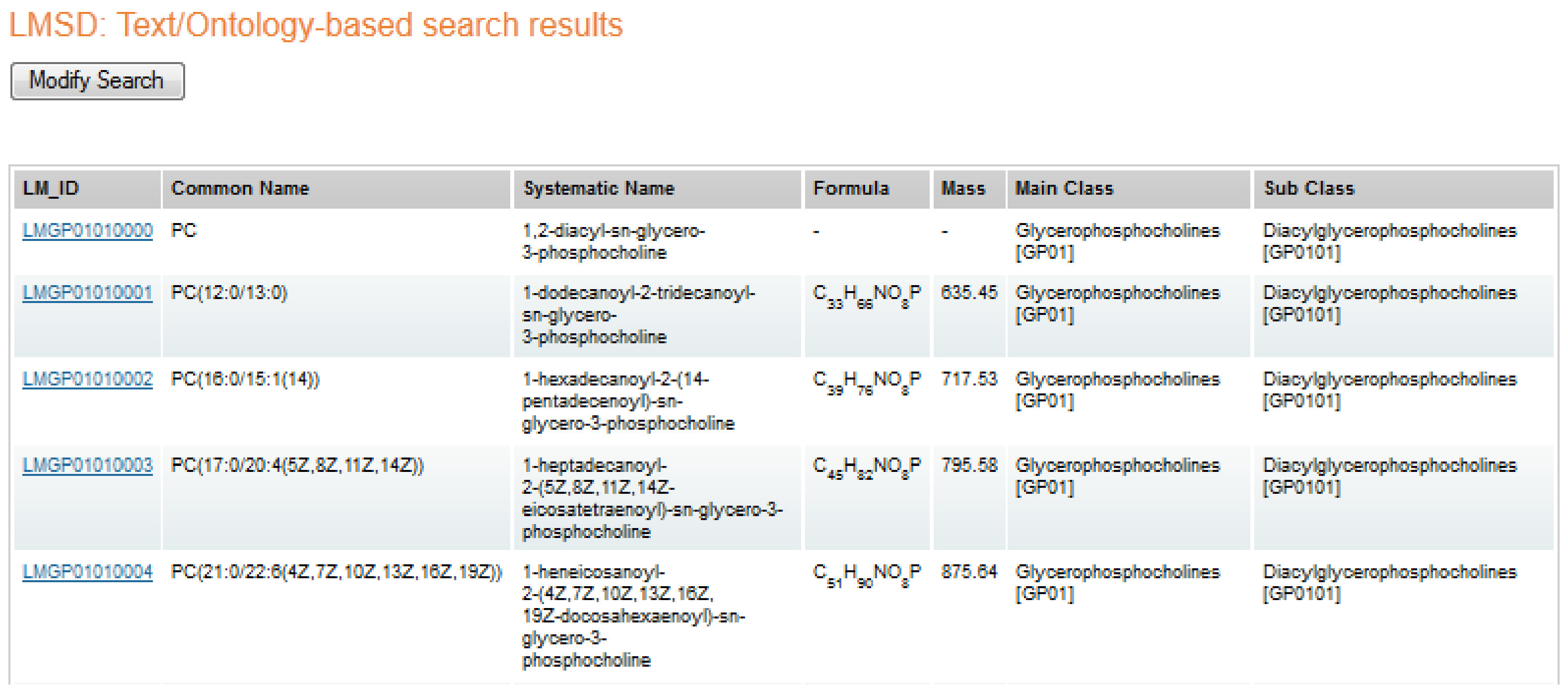
Group:

Group:

Records per page 50

Sort by LM_ID

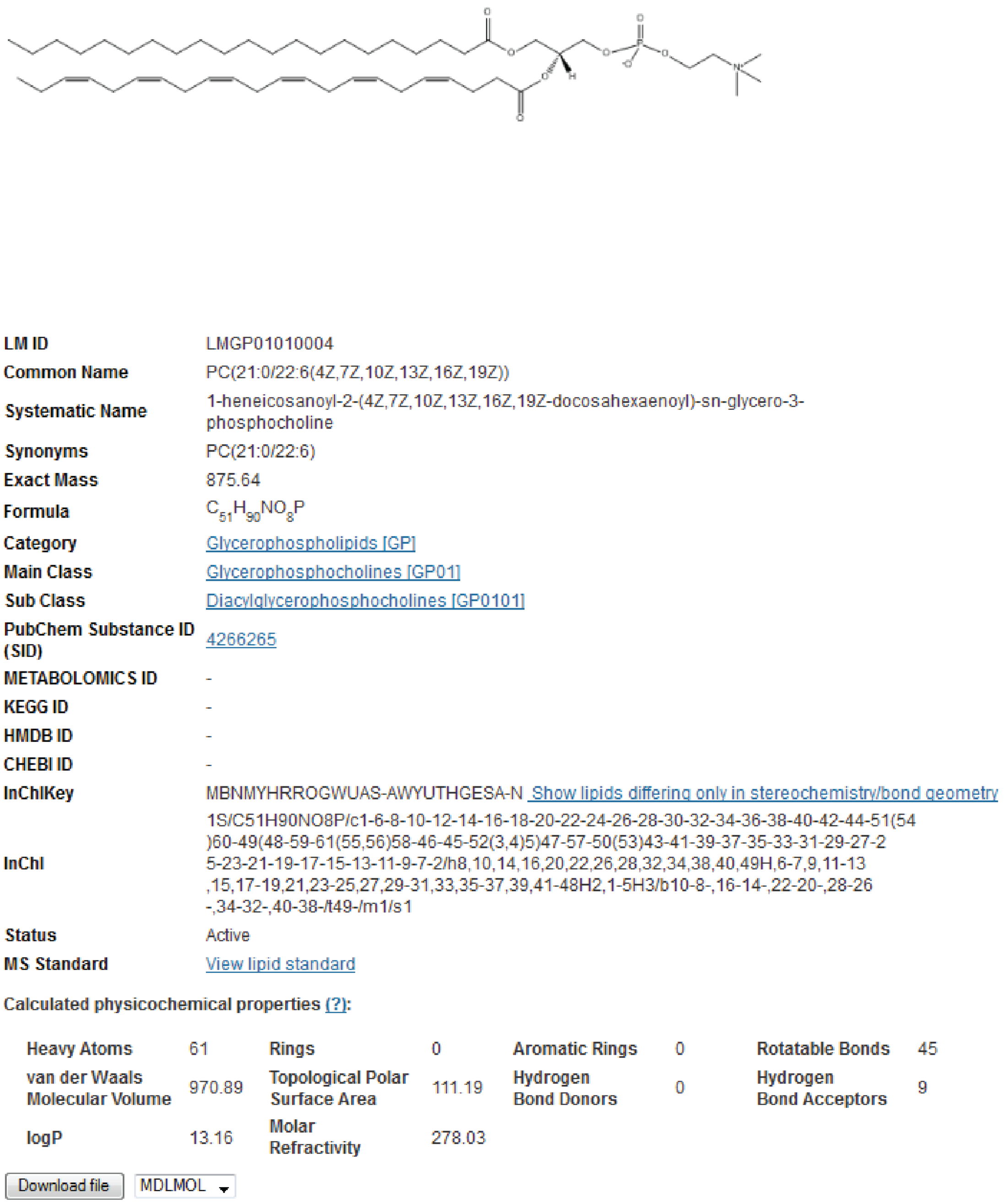
Figure 3.4: A screenshot of the LIPID MAPS Text/Ontology search. Users may select a lipid sub class and define chemical properties of lipids to refine search results from the LMSD.



LMSD: Text/Ontology-based search results

LM_ID	Common Name	Systematic Name	Formula	Mass	Main Class	Sub Class
LMGP01010000	PC	1,2-diacyl-sn-glycero-3-phosphocholine	-	-	Glycerophosphocholines [GP01]	Diacylglycerophosphocholines [GP0101]
LMGP01010001	PC(12:0/13:0)	1-dodecanoyl-2-tridecanoyl-sn-glycero-3-phosphocholine	$C_{33}H_{66}NO_8P$	635.45	Glycerophosphocholines [GP01]	Diacylglycerophosphocholines [GP0101]
LMGP01010002	PC(16:0/15:1(14))	1-hexadecanoyl-2-(14-pentadecenoyl)-sn-glycero-3-phosphocholine	$C_{39}H_{76}NO_8P$	717.53	Glycerophosphocholines [GP01]	Diacylglycerophosphocholines [GP0101]
LMGP01010003	PC(17:0/20:4(5Z,8Z,11Z,14Z))	1-heptadecanoyl-2-(5Z,8Z,11Z,14Z-eicosatetraenoyl)-sn-glycero-3-phosphocholine	$C_{45}H_{82}NO_8P$	795.58	Glycerophosphocholines [GP01]	Diacylglycerophosphocholines [GP0101]
LMGP01010004	PC(21:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z))	1-heneicosanoyl-2-(4Z,7Z,10Z,13Z,16Z,19Z-docosahexaenoyl)-sn-glycero-3-phosphocholine	$C_{51}H_{90}NO_8P$	875.64	Glycerophosphocholines [GP01]	Diacylglycerophosphocholines [GP0101]

Figure 3.5: A screen shot of the LIPID MAPS results page. LIPID MAPS ID (LMID), LIPID MAPS name, Systematic Name, Formula, Exact Mass ,Parent Main Class and parent Sub Class are provided.



LM ID	LMGP01010004
Common Name	PC(21:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z))
Systematic Name	1-heneicosanoyl-2-(4Z,7Z,10Z,13Z,16Z,19Z-docosahexaenoyl)-sn-glycero-3-phosphocholine
Synonyms	PC(21:0/22:6)
Exact Mass	875.64
Formula	$C_{51}H_{90}NO_8P$
Category	Glycerophospholipids [GP]
Main Class	Glycerophosphocholines [GP01]
Sub Class	Diacylglycerophosphocholines [GP0101]
PubChem Substance ID (SID)	4266265
METABOLOMICS ID	-
KEGG ID	-
HMDB ID	-
CHEBI ID	-
InChIKey	MBNMYHRROGWUAS-AWYUTHGESA-N Show lipids differing only in stereochemistry/bond geometry
InChI	1S/C51H90NO8P/c1-6-8-10-12-14-16-18-20-22-24-26-28-30-32-34-36-38-40-42-44-51(54)60-49(48-59-61(55,56)58-46-45-52(3,4)5)47-57-50(53)43-41-39-37-35-33-31-29-27-25-23-21-19-17-15-13-11-9-7-2/h8,10,14,16,20,22,26,28,32,34,38,40,49H,6-7,9,11-13,15,17-19,21,23-25,27,29-31,33,35-37,39,41-48H2,1-5H3/b10-8-,16-14-,22-20-,28-26-,34-32-,40-38-/t49-/m1/s1
Status	Active
MS Standard	View lipid standard

Calculated physicochemical properties (?):

Heavy Atoms	61	**Rings**	0	**Aromatic Rings**	0	**Rotatable Bonds**	45
van der Waals Molecular Volume	970.89	**Topological Polar Surface Area**	111.19	**Hydrogen Bond Donors**	0	**Hydrogen Bond Acceptors**	9
logP	13.16	**Molar Refractivity**	278.03				

Figure 3.6: A screen shot of the LIPID MAPS single record page. A 2D image, synonyms and chemical descriptors are displayed amongst other useful information.

LIPID MAPS supports identifications at the geometric isomer level, where double bond positions and stereochemistry are resolved. It does not contain identifications at the species, fatty acid scan species or sub species level and so cannot be referenced by the majority of mass spectrometry lipidomics experiments. Also considering LIPID MAPS contains lipids at the geometric isomer level (double bond position and geometry known) it is extremely sparse. Only 7755 gylcerophospholipids are recorded, a small proportion of the millions of physically feasible lipids. Of the lipids it does contain there is little or no evidence of whether they are theoretical, identified, bacterial or human etc and so its use, as a comprehensive reference of lipid identification, is limited. In the field of proteomics, UniProtKB has shown the necessity for clearly separating the theoretical from the experimentally validated, added stratification into organisms has also proven useful and could be applied to lipid databases. After using the resource for a while it is clear that the application is well designed around its structural hierarchy and that the functionality available through the web application is obvious. However, the appearance is not especially modern and the user experience could be vastly improved by using more current web technologies. Additional programmatic access could be simplified considerably while also making available much more complex queries upon the data using techniques such as RESTful web services.

3.1.3.2 Lipid Bank

Lipid Bank (Yasugi & Watanabe, 2002) (http://lipidbank.jp/) contains approximately 6000 lipid identifications at the geometric isomer level. Unfortunately this resource is no longer actively developed or new data made available and has subsequently been described as discontinued since June 2007. However it does contain useful information and embodies a slightly different strategy for cataloguing lipids than LIPID MAPS. All entries in Lipid Bank are manually added and curated with supplementary annotation, including physical/chemical properties and literature references. While discontinued for some time, it still shows the value of human curated data especially when it is distinct from any automatically generated/annotated data.

3.1.3.3 Avanti Polar Lipids

The Avanti Polar Lipids (http://avantilipids.com/) catalogue is not designed as a comprehensive list of all lipid species, only a list of synthetic lipid species available for purchase. The majority of its lipids are described at the isomer level with appropriate metadata such as structure, formula and mass.

3.1.3.4 Resource summary

In my opinion, of the available resources, no single resource is the most effective platform for storing and displaying a catalogue of lipid structures relevant to modern mass spectrometry based lipidomics and the identifications it is producing. It is clear that elements can be drawn from them and efforts in other fields to design a system which supports the lipidomics community and will drive high throughput approaches and bioinformatics within this field forward.

3.1.4 Data access solutions

Having a comprehensive database of lipid species is not enough to warrant its creation, data is only truly useful if it is publicly available and easily accessible. There are several solutions to exposing this type of data to the lipidomics community each with benefits and drawbacks:

Direct access Direct access to the database via a public read-only user account, the benefits include:

- Easy to implement
- Full access to all of the data
- Access to stored procedures that perform slightly more complex queries

The drawbacks include;

- User requires knowledge of SQL
- User must first understand the database schema before being able to query it

- Only simple processing of the data can be performed via SQL, if the user requires a complex view of the data it will have to be done in multiple rounds of retrieving the necessary data then processing it locally
- If the database schema is under regular revision, queries will not be stable and will need to be rewritten with each new version. This would temporarily break any software built on top of the database.
- Direct database access may cause some concern for security and integrity of the data

Graphical User Interface Providing data via some form of graphical user interface (GUI) whether a stand alone desktop application or a web application can provide an extremely easy environment in which to browse data. Its benefits include:

- Users less familiar with programming can access the data
- Data can be provided in multiple different views not just the unprocessed data directly from the database
- Large potential user base if a GUI is available
- For most tasks a well designed GUI radically decreases the effort required to achieve the same results programmatically

The drawbacks include:

- Developing an intuitive GUI requires a considerable investment of time
- Not all use cases can be supported. In a GUI there will always be a niche group of users who require complex functionality that is not cost effective to develop
- Because of the large time investment required it is important that either the underlying data to be displayed is stable (at least in its structure and format) to avoid large re-development periods as the goal posts for displaying data shift

Web services In this case web services could provide a layer between the database and the user by which requests for data can be made. The requests are not limited to simple data access methods but can also allow the remote execution of complex code to process data prior to providing it to the user. Web services have a strict set of methods which are accessible and return data often using eXtensible Markup Language (XML) or JavaScript Object Notation (JSON) (`http://json.org/`) as the data interchange format. An example is mentioned the previous chapter, section 2.3.6 regarding access to Array Express `http://www.ebi.ac.uk/arrayexpress/` (Parkinson *et al.*, 2011). Web services have a number of benefits:

- Web services can be accessed from anywhere with an internet connection
- The web services themselves are programming language independent, access is made using the common data interchange formats
- The core business logic of web services is typically very compact, each method handling a specific but small task. These types of methods are highly reusable for creating new applications or chaining existing simple methods together to provide a complex result. For common processing required on the raw data it is good practice to provide this functionality so the users do not have to develop it independently

Web services, while being a good intermediate between restricting direct access to the data source and while providing highly flexible access to the raw data do have some drawbacks:

- Web services typically provide data in plain text formats, this data can be considerably more verbose than requests encoded in a binary protocol
- Handling client sessions over a web service requires extra work because the HTTP and HTTPS protocols were designed to be stateless and not identify clients over long periods of time, only over the short period of the request
- In order to directly access a web service some programming expertise is required. However, this can be made remarkably easy

Providing a graphical user interface to the database is key to making its usage and future development a success. Of all the previously discussed option the best possible choice is a web application front end and a web service back end. Providing a browser and platform independent web application makes data access possible from anywhere in a consistent manner. It negates the need to maintain versions for different or out-dated operating systems. It also removes the need for a complex installation or setup that a desktop application may require which can act as a usage barrier to less technical users. Additionally it gives access to the functionality from any device with internet access. Developing a web application is not a short or simple task, but it can be broken down into two main sections:

Client A client application must be developed that presents the information to the user and acts as the user's point of interaction with the database.

Server A server must not only retrieve information from the database but also process it and provide it to the client so that it can be visualised. The same is true in reverse; data submitted by the client may also be processed by the server side code and persisted to the database.

By designing the client to receive data in a standard data interchange format and developing the server as a web service, the client can call the web service to retrieve data. In the process a free stand alone web service is created that can be programmatically accessed to retrieve the data in some data interchange format and new applications be easily built on top of it. The model around which the web application will be designed can be seen in figure 3.7.

3.1.5 Objectives

In order to support the lipidomics community and provide easy access to a comprehensive catalogue of lipids a system is needed to:

Provide stable identifiers for all common lipid structures. Consistent naming is a critical feature of a database, but consistent internal IDs of records is equally important to facilitate long lasting third party code written to use information in the database.

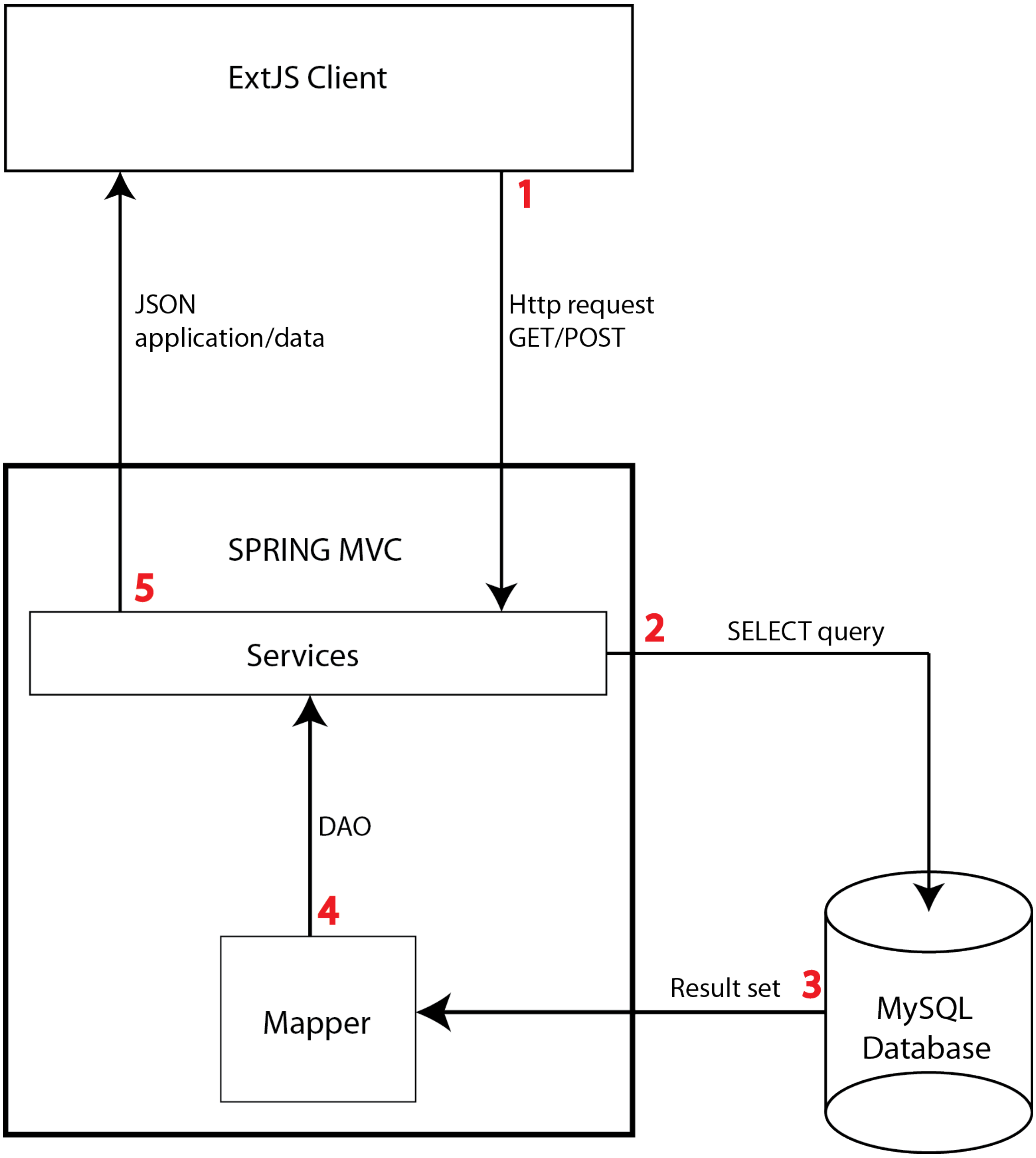


Figure 3.7: The schematic representation of client, server and datasource interaction. SPRING is an application development framework that handles the majority of data transformation and request/response generation automatically and EXTJS is a JavaScript library for building rich web applications. Following a request from the client or any other source to a REST service (1), the database is queried (2). Results are mapped to Data access objects (DAOs) (3) and transformed into JSON application data (4) before being returned as the response to the client (5) which then processes and renders the result.

Provide all theoretical lipid structures within agreed upon chemical bounds, while maintaining a clear separation between them and experimentally validated structures. Expressed by the LipidomicNet Consortium and seen in Lipid Bank, experimentally validated structures are of primary interest and the theoretical chemical space is not appropriate for some analyses.

Provide an evidence based system where lipids at all levels of the structural hierarchy can be annotated with cross-references into other resources, literature references, chemical descriptors and spectra. The main use of the database will not come from its dictionary functionality but its supporting metadata of lipids.

Provide a highly normalised database that allows complex queries of lipid components e.g. mass, fatty acids, modifications. Different users will approach the database with entirely different aims, some from a literature search point of view, others with a mass spectrometry peak list of masses to identify.

Provide programmatic access so that bioinformatic tools such as lipid identification software or Laboratory Information Management Systems (LIMS) can utilise it and promote its growth. A simple to use web service is key to accessing data from a complex relational database.

Provide a platform and browser independent web application to represent the information in the database in an intuitive graphical user interface.

With these aims in mind, the research gathered on the existing lipid databases and the requirements of the LipidomicNet Consortium, I set about designing a database of theoretical lipids and supporting web application here onward referred to as LipidHome.

3.2 Materials and Methods

3.2.1 Theoretical lipid generation

Commonly identified lipids within the LipidomicNet Consortium are mainly members of the glycerophospholipid category, these lipids all share a similar structure. It is this category and the sub classes described below that form the proof of concept for generating theoretical lipids,storing them in the database and accessing them via the flexible web application. The approach and methods involved were designed to be readily transferable to other lipid categories with only minor modification to the code. The following lipid main classes represent the base set of main classes from which theoretical lipids are generated:

Glycerophosphoethanolamines

Glycerophosphocholines

Glycerophosphoserines

Glycerophosphoinositols

Glycerophosphates

Glycerophosphoglycerols

Of these main classes the following sub class types are generated:

diacyl

dialkyl

monacyl,monoalkyl

This category of lipids is composed of a few parts; head group, phosphate, glycerol, linkages and fatty acids, see figure 3.8. By doing a Cartesian cross of all main classes and linkages, 18 lipid sub classes that are the inputs for lipid enumeration were generated. Within a lipid sub class it is the fatty acid composition of lipids that makes them different. Each of these subclasses has two fatty acid R groups, each of which can be anything from a pool of theoretical fatty acids. In order to estimate the chemical bounds for fatty acid generation a survey was conducted amongst LipidomicNet partners and additional lipid scientists at the Babraham Institute. A provisional set of parameters for fatty acid generation was agreed upon as:

Minimum number of carbons 2

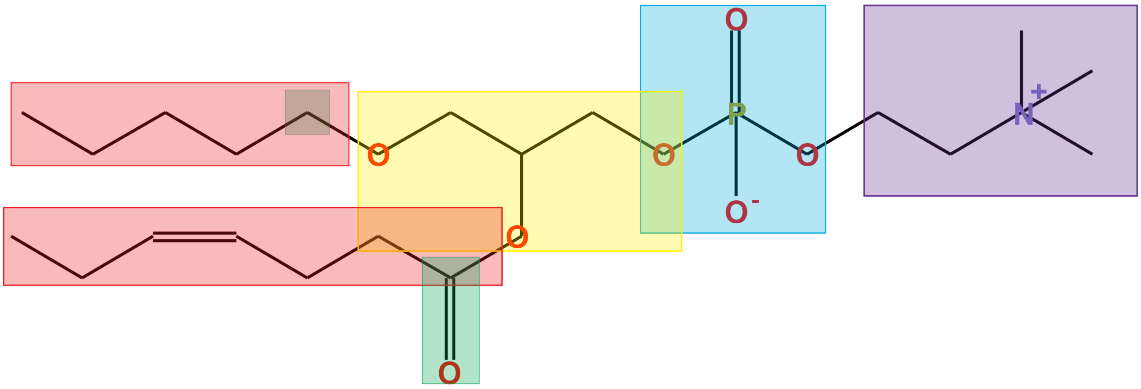


Figure 3.8: A Glycerophospholipid is theoretically constructed from a set of overlapping parts. Red: Fatty acid with linkage. Green: linkage. Yellow: Glycerol. Blue: phosphate. Purple: Head group..

Maximum number of carbons 30

Odd number of carbon fatty acids Fatty acid chains may have an odd number of carbons

Minimum number of double bonds 0

Maximum number of double bonds 10

Double bond spacing Must be spaced at least 3 carbons apart (i.e. conjugated double bonds not allowed, although the option to allow them is available.).

While these parameters do not reflect the enormous variety of fatty acids synthesised in nature e.g. branch-chain fatty acids, they do reflect a simple test case to prove the methodology upon and will be used to construct a large proportion of the lipids of interest within the LipidomicNet Consortium. From this pool of fatty acids a Cartesian cross was performed between fatty acids and sub classes to produce all viable lipids within the defined chemical space. (See figure 3.9.)

Originally this whole process was done using the Java programming language, simply generating the names but no structural information. This was done purely by string building, with code similar to the pseudo code:

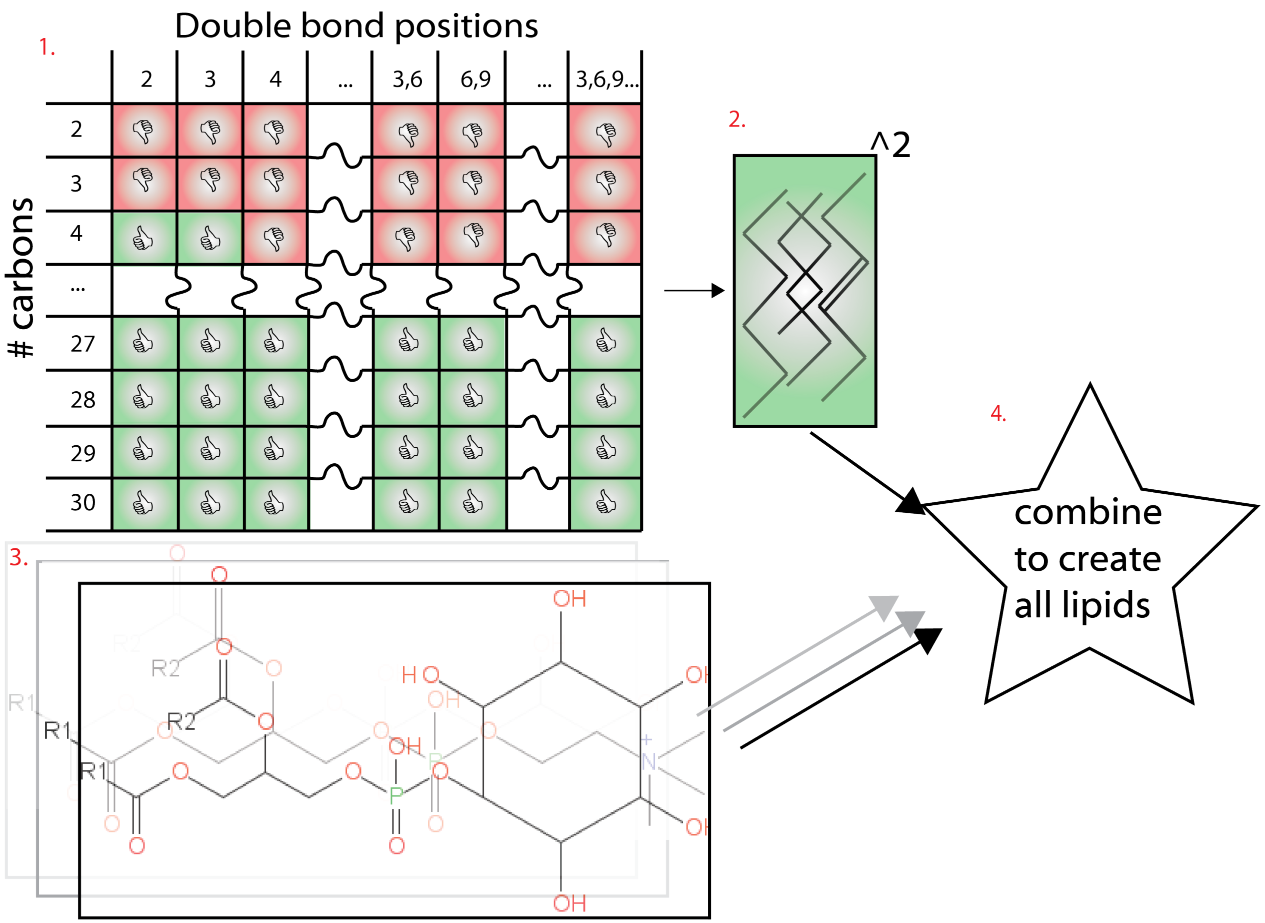


Figure 3.9: A diagram of the *in silico* construction of theoretical diradyl lipid sub species. 1. All viable potential fatty acids are generated from a set of starting parameters. 2. They are combined all against all. 3. The head groups with α-carbons and linkages are generated. 4. The head groups are crossed with the fatty acid pairs to produce all viable lipid structures within the predefined chemical space.

```
List allSubSpecies
fattyAcids = generateFattyAcids()
for(fattyAcidOne: fattyAcids){
  for(fattAcidTwo: fattyAcids)
    for(headGroup: headGroups){
      subSpecies = headGroup + fattyAcidOne + '/' + fattyAcidTwo
      allSubSpecies.add(subSpecies)
    }
  }
}
```

As an alternative to simply generating lipid names, lipid isomers were created directly as chemical objects using the Chemical Development Kit (CDK) (Steinbeck *et al.*, 2003). This Java library allows the efficient creation and manipulation of *in silico* molecules, with convenience functions to extract properties such as mass and formula. In addition, encoded structural descriptors were generated including Simplified Molecular-Iinput Line-Entry System (SMILES) codes and INternational CHemical Identifiers (InChIs). These encoded strings are community accepted, compact ways of describing a molecule that can be interpreted by the majority of existing software relating to molecule visualisation and chemical descriptor generation. After some early discussion, it was agreed that another PhD student and I work in collaboration on this project. My Collaborator created a stand alone Java library with all the basic functionality to generate lipid isomer objects given the previously described fatty acid generation parameters and the sub class head group + phosphate + glycerol structures as MOL format files. While the details of that library are out of the scope of this thesis, it is important to note that the *in silico* molecules it produces uses a similar logic to that shown in my original approach and the previous code snippet. Once the library was completed my code was re-written to provide it with the relevant parameters for which to generate lipids, the generated viable lipids are then stored in a MySQL database.

3.2.2 Database design and technical implementation

In order to store theoretical lipids generated by the process described in 3.2.1 a database schema was designed. Using the MySQL Workbench software suite version 5.2 (Oracle Corporation), a schema was designed with the graphical data modeling component. Core database design principles such as normalisation and indirection tables are employed to produce an optimised database that should be simple to query and fast. The most important feature of the data being generated is the lipid structural hierarchy described in figure 3.2. Modelling the core structural tables of the database on the structural hierarchy allowed users to search for lipids relevant to the structural hierarchy level they are interested in. The database schema was optimised by flagging some fields as unique that could not have duplicate values, this speeds up searches that involve these fields. The table sizes in a database are a key feature of performance, typically the larger the slower. For this reason lipid isomers which numbers are predicted to be in the millions are not stored in the database. Instead they are generated on the fly by server side code when requested. During application development the database was run on a MySQL 5.2 server instance, for reasons regarding the available computing resources the MySQL server exposed to the public via web services was version 5.1.

3.2.3 Metadata annotation

The strength of a resource like LipidHome lies in the quality and volume of annotation associated with its entries. The database contains two main types of annotation based upon research from the previous resources in section 3.1.3.

3.2.3.1 Cross references

The more integrated a resource is with other existing resources the better visibility it has to new users. Other resources effectively offer free information that can be linked to, at a cost much lower than regenerating the information they contain from scratch. As LIPID MAPS represents the most comprehensive lipid resource to date it is the primary cross reference for records in LipidHome. As discussed previously LIPID MAPS records lipids at the geometric isomer level, so the only cross references available from it are records for the isomer table, in a many to one relationship with their sub species. From LIPID MAPS it was not only possible to build references to it, but also harvest the cross references already provided by LIPID MAPS and store those as well. In order to harvest cross references from LIPID MAPS a set of scripts were written to utilise the LIPID MAPS programmatic access. The base set of subclasses for inclusion into LipidHome are found in LIPID MAPS and using the LMSDSearch.php script with the following parameters each time passing a sub class ID at "?":

```
http://www.lipidmaps.org/data/structure/LMSDSearch.php?Mode=Process
TextSearch&OutputMode=File&OutputType=CSV&OutputDelimiter=
semicolon&OutputQuote=no&SubClass=?
```

This retrieved a table of all lipids that are a member of the subclass in the format: LM_ID, COMMON_NAME, SYSTEMATIC_NAME, FORMULA, MASS, CATEGORY, MAIN_CLASS, SUB_CLASS

From each of the LIPID MAPS isomer records the LipidHome parent lipids were calculated, i.e. the sub species, fatty acid scan species and species it belonged to. A LipidHome isomer was then created and persisted to the database with the appropriate links to its parent lipids. At the same time a cross reference was created and persisted for this LipidHome isomer pointing to its LIPID MAPS record ID (LMID) and URL. While the LMID is known, a LIPID MAPS php script was executed to retrieve the details of the record and its secondary cross references, LMID is passed into "?":

```
http://www.lipidmaps.org/data/LMSDRecord.php?Mode=File&LMID=?
&OutputType=CSV&OutputDelimiter=semicolon&OutputQuote=no";
```

This retrieved a single semicolon delimited list of LM_ID, COMMON_NAME, SYSTEMATIC_NAME, FORMULA, MASS, CATEGORY, MAIN_CLASS, SUB_CLASS, SYNONYMS, KEGG_ID, HMDBID, CHEBI_ID, PUBCHEM_SUBSTANCE_ID, LIPIDBANK_ID, LIPIDAT_ID, STATUS, METABOLOMICS_ID.

Not all LIPID MAPS records contain cross references to every resource. Many of them point to external records at a completely different and inaccurate level of the structural hierarchy, e.g. PC 18:0/18:1[3Z] referencing the PC main class in KEGG. However, they were all harvested as it seems accepted in the field. The cross references were then persisted into the database. The pipeline is scheduled to run once a month identifying any new additions to LIPID MAPS or the secondary cross references it contains.

3.2.3.2 Literature

Similar to cross references, links to relevant literature gives users the ability to explore the existing information for a particular lipid, such as experimental protocols, associations with disease and interactions with other molecules. When searching for literature it is particularly difficult to make a search that is capable of returning all the relevant papers which use a slightly different nomenclature than the user is familiar with. To do the best possible literature search, search terms must first be synonymised so papers that contain all possible synonyms of the term searched for can be retrieved. Secondly, simply reading the keywords or title of a paper is not sufficient to return all the relevant literature. The abstract must be read to understand the context of the terms contained in the paper. Using the EBI Whatizit service (Rebholz-Schuhmann *et al.*, 2008), specifically the 'QBmarsdf' corpus processing pipeline, the entirety of MEDLINE abstracts was searched for each record in the database and all its synonyms. After extracting the name and ID of each record from each of the core structural tables of the database (specie, fatty acid scan species and sub specie), the process for each record was as follows:

Synonymise the name to create a list of names that conform to a number of commonly used nomenclatures. For example the lipid species PC 36:2 was synonymised to 36:2-PC, GPC 36:2, GlyPC 36:2 and PCho 36:2 etc.

MEDLINE abstracts have been indexed with lucene (an information retrieval software library http://lucene.apache.org/) to allow rapid searches of large text documetns. To search the indexed abstracts a Lucene search string must be created, but in order to reduce false positive search hits, the name of the main class of the record was synonymised and appended to the record name to give the search some context. For example, the main class glycerophosphocholine was synonymised to phosphatidylcholine and appended to give the two

Lucene search queries “glycerophosphocholine PC 36:2” and “phosphatidylcholine PC 36:2”.

A Lucene search was performed for each Lucene search string using the EBI whatizit? service programatically via the Simple Object Access Protocol (SOAP) web service. The Qbmarsdf pipeline was selected which is described as a “MEDLINE Abstract Retrieval Engine based on the Text Mining Index”. It retrieved an XML document that contained a list of search results.

The XML document was parsed to extract individual papers and their data, including PMID, title, journal, date of publication, abstract, authors and MeSH terms. This information was then persisted into the appropriate tables and fields in the database referencing the relevant LipidHome record.

This operation took a considerable amount of time to parse all the abstracts in MEDLINE for all the synonymised records in the LipidHome database. After synonymisation the process represented millions of Whatizit searches and approximately 1.5 days of computation time. Due to the considerable time to perform this entire operation and the likelihood that this will increase over time as the database stores additional lipid categories, main classes and sub classes, it was scheduled to execute on the first of every month to scan for new paper annotations in the MEDLINE abstracts.

3.2.3.3 Spectra

While mass spectrometers are very diverse instruments, with a considerable number of operational parameters that effect the detection of molecules, it is possible to build a simplified spectrum if its mass, adduct ions and fragmentation pattern are known. Fragmentation depends heavily upon the instrument settings and collision method but as a prototype for a more comprehensive theoretical spectral library in the future, fragmentation of glycerophospholipids was based on work by Hsu et al (Hsu & Turk, 2009). Wherein the fragmentation pathway of gylcerophospholipids was discovered for low-energy collisionally activated dissociation (CAD) fragmentation in both tandem quadrupole and ion-trap instruments for both positive and negative ion mode. Discussing this approach with members of the LipidomicNet Consortium it was apparent that detection of gylcerophospholipids is mostly done in positive ion mode. Figure 3.10 is a simplified fragmentation pathway that forms the theoretical basis for the generation of a theoretical spectral library.

Using the CDK, a single isomer for each sub species is created and the fragmentation bonds outlined in figure 3.10 were located. The bonds are then removed and the appropriate fragments created. Each fragment mass is recorded and assigned an arbitrary ion intensity of one. From this data a spectrum can be plotted. Further annotation of the peaks and processing of the ion intensity can be applied at a later stage to create spectra that are comparable with experimentally derived replicates. An example of post processing the ion intensities comes from the observation that the R1 fragment is always more intense than that of the R2 fragment (Hsu & Turk, 2009), due to the more favourable selection of the top fragmentation pathway shown in figure 3.10. The actual implementation of this approach is not yet complete, but I have included the methodology as I believe it to be unique and valuable to the future development of the database. Additionally the database design has been oriented around its inclusion as has the web application.

3.2.4 MS1 search Algorithm

A simple mass spectrometry experiment designed to identify lipid species typically results in a list of precursor ion m/z, where each m/z value represents the ion of a different lipid species. An MS1 search algorithm is a piece of software that translates recorded precursor ion m/z values into identified biomolecules with the same m/z. From the following input parameters; list of precursor ion masses, mass tolerance (Daltons) and applicable adduct ions an MS1 search can be performed to identify a set of experimental masses. A set of neutral [M] masses is generated by subtracting the applicable adduct ion masses from each precursor ion mass. All precursor ion mass are assumed to be of charge magnitude 1. For example the precursor ion mass 785.6 Da with the applicable adduct ions: [M+H],[M+K] and [M+Na], produces three neutral masses of the unknown precursor ion mass: 784.6 Da , 746.65 Da and 762.62 Da respectively. Each neutral mass is searched against the 'composition' table of the database with the tolerance value to modify the range of the neutral ion masses. Species names, identification status, exact mass, sub class and Δ mass are

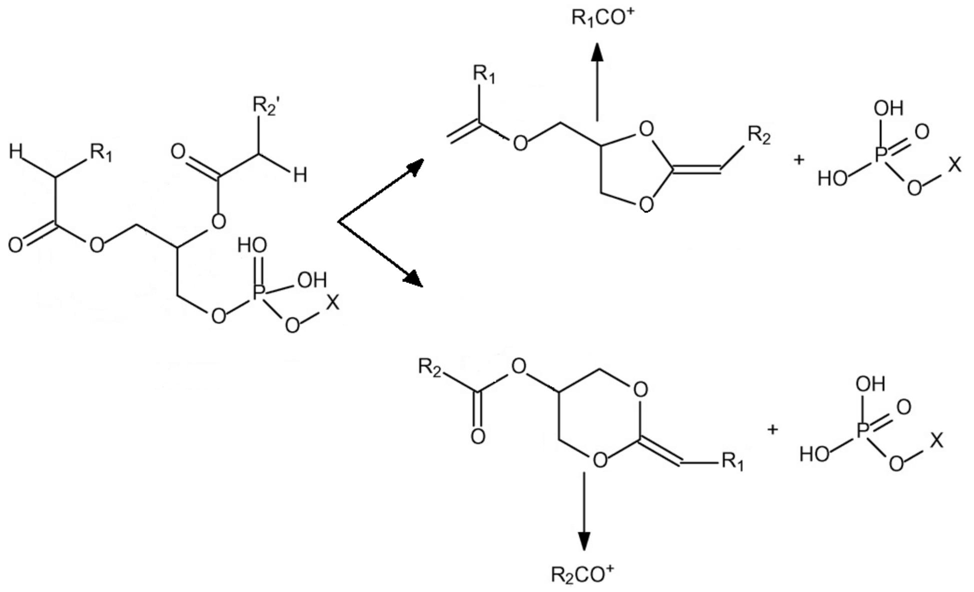


Figure 3.10: A simplification of the general fragmentation pathway for the majority of glycerophospholipids. Typically a Phosphate-X ion is created where X is any one of the head groups previously mentioned in 3.2.1, e.g. ethanolamine. The remainder of the molecule containing the bulk of the glycerol and fatty acids undergoes a further fragmentation to yield a positively charged RCO fragment that can be used to identify the fatty acid.

returned. It is worthy of note at this point that all the returned species are viable identifications by mass for the original precursor ion scan. It is the responsibility of the experimentalist to report the identifications in the most appropriate manner. For example, identified lipids (those with cross references or literature references) may be selected over all other viable hits for a mass, or the lipid with the smallest mass difference between the search mass and hit mass. Alternatively, all viable identifications may be reported. However, this is not the usual case in proteomics where a single peptide is reported per spectrum.

3.2.5 Client sketches

Before creating a web application it is standard practice to design the GUI layout as a reference during the implementation process and to gather general feedback before investing time into a potentially sub optimal user experience. The client design sketches were created using the Firefox add-on 'Pencil' version 1.0.6. The tool allowed the quick design of applications with prebuilt graphics such as buttons, scroll bars, frames and tabs. Design sketches were exported as Portable Network Graphics and printed for reference during client application development.

3.2.6 Molecule rendering

The graphical representation of the molecule presented a particularly difficult design decision, with so many existing libraries on offer it was unclear which to use. Originally I created a solution utilising the CDK Java library used to generate the lipid structures. Images were rendered on the server side after retrieving them as MOL formatted strings then returned to the client. This approach suffered two main problems; transporting images across the network is not optimal if the raw data used to render them can be transported and the actual computational power used to render them farmed out to the client. These 2D molecules were rendered in a pseudo random orientation and required a computationally intensive stepwise rotation and optimisation to render the molecule in the LIPID MAPS Consortium recommended orientation. To reduce server load and network traffic a client-side solution was initiated using the open source JavaScript library ChemDoodleWeb (http://web.chemdoodle.com/). Unfortunately, the library did not fully support our needs and after contact with the developer, it was clear that the basic functionality requested would not be developed by them. An open source fork of the original library was created and made available at http://code.google.com/p/lipidhome-molrender/. The library was improved to support R chains and other pseudo-atoms and the the parsing of a MOL file's description block, where charge states of atoms amongst other things are stored.

3.2.7 Client technical implementation

After a short prototyping period utilising the Google Web Toolkit (https://developers.google.com/web-toolkit/) software development environment,

specifically the ExtGWT Library (http://www.sencha.com/products/gxt), enough positive feedback from the lipidomics community was received to warrant the inclusion of a developer from the proteomics services team at EBI to aid in LipidHome's development. Analysis of the prototype code quickly revealed an unscalable application that was not designed to service many users simultaneously or be easy to develop and maintain in the future. With the help of Antonio Fabregat the previous code was overhauled entirely and client development ported to the JavaScript framework Ext JS 4 (http://www.sencha.com/products/extjs/). This framework provides a multitude of rich components (grids, forms, trees etc) for constructing complex web applications and communicating both data and events between them. The plethora of browser software currently available poses a significant problem to web developers. Websites may render differently between browsers and JavaScript in particular may be executed at very different speeds. Ext JS is a platform independent framework that takes care of browser cross compatibility of code so the the web designer is free to concentrate on the application itself. Development of the client code was written in IntelliJ IDEA versions 10.5 and 11; a professional Java oriented Integrated Development Environment (IDE) offering both an 'Ultimate' edition and free 'Community' edition, available at http://www.jetbrains.com/idea/. Primary testing of the application was done under Firefox version 10.0.2 with Firebug 1.9.1 and

Google Chrome version 17.0.963 with its native web developer tools. The application was also tested under Internet Explorer 8 and 9, Opera 11.6.1 and Safari version 5.1.2.

3.2.8 Server technical implementation

While the clients sole function is to present data to the user it is the servers responsibility to connect the data with the client. As far as application development frameworks go, Spring (http://www.springsource.org/) is an outstanding example which was recommended by members of the IntAct team at EBI. After further collaboration with the members of IntAct (Hermjakob *et al.*, 2004b) and reviewing their server code architecture, their usage of Spring seemed transferable to our demands (http://code.google.com/p/intact/). Building a RESTful web service presents a way of providing dynamic data to the client on a custom request-by-request basis, but also an avenue of computational access to exposed server side data processing methods (e.g. MS1 search engine) and data access code. The data exchange format provided by the web services is JSON, a compact well structured format in place of the more verbose XML. The server code was written using the Java programming language in IntelliJ IDEA version 10.5 and 11. During development, testing of the server and the web application, code was deployed to a local Jetty HTTP server/servlet container. This does not require lengthy Web application ARchive file (WAR) deployment times and is ideally situated for server code undergoing rapid development that needs to be constantly recompiled and tested. Once a stable release was created the WAR file was deployed to an Apache Tomcat web server where it was made publicly available over the web.

Data access web service methods were organised into the lipid structural hierarchy level for which they provide data. For example, requesting "/services/species/summary?=1" retrieves all the relevant information for populating the information panel (see section 3.3.1.2) for the species with database id 1 such as; papers, cross references and structure MOL string as JSON. Similarly requesting the URI "/services/species/fascanspecies?=1" returns the list of fatty acid scan species that are children of the species with database id 1, this is used to populate the ‘child list panel’(see section 3.3.1.2). This service approach exists for all structural hierarchy levels of the database. Other services were organised into Tools and Utilities, the former providing access to the MS1 search engine. Utility services does a number of structural hierarchy level independent actions, such as returning a tree structure of all parents of a requested database item and exposing the live search functionality of the server. A list of current services and their usage is provided in Appendix 1.

3.3 Results

Initially this work was undertaken alone, but after a prototype it became clear that my expertise alone was not enough to complete such a large project in a time frame reasonable for the PhD. I would like to make clear that the initial research, concept and database are my work and that the work I will describe subsequently is much more collaborative, wherein I took both an organisational role and an active part in architecting and implementing the current system. It would be amiss not to mention in detail the excellent work of Pablo Moreno in the generation of theoretical lipid isomers (section 3.2.1) and Antonio Fabregat in the web application development. These results represent a preliminary dataset and a robust platform on which to store and present them. Future iterations of the work have been considered when architecting this solution. Robust and dynamic code has been favoured over a multitude of superficial features that will not translate easily to changing requirements.

The structural hierarchy shapes our data and forms the core structure of the database. Experimentalists can perform an MS1 experiment and identify species. More complex experiments involving fragmentation to identify the fatty acids of a lipid to infer the sub species. Or even an extremely high resolution mass spectrometry or NMR experiment in which the structural details of the lipid are resolved to isomer level. To accomodate this a table is devoted to each level of the structural hierarchy, this allows fast seaching and flexible queries independent of the users demands. Some levels of the structural hierarchy have shared structural features that are stored as separate tables in the database, for example, fatty acid scan species and sub species share particular fatty acid species components associated with them.

Storing this information in separate tables will allow for queries such as "Which sub species contain the fatty acid 18:2?". Planning for future complexity, the database also contains a modification table for fatty acids in which branched fatty acid chains and glycosylated fatty acids will be supported. Similarly, many lipid species may share a mass and chemical formula. These 'composition' elements are abstracted into separate tables for better normalisation and the possibility to search for lipids by mass in a much more optmised manner. This will be a key feature of the final web application. Lipid isomers are not stored in the database as they represent an extremely large amount of data that does not have a particularly high potential to be accessed. If necessary, theoretical isomers for a sub species can be generated on the fly extremely quickly, using the custom theoretical lipid generation library described in 3.2.1. The isomer table is specifically reserved for isomer records that have evidence in the literature that they exist or for which other databases store them and thus have metadata annotations. The final database provides a highly normalised intuitive schema that, with reference to the structural hierarchy, should be quick and easy to query. In order to accommodate metadata annotation of the core structural tables of the database a schema revision was finalised that sufficiently represented the structure of the data intended to be stored and the metadata to be associated with it in figure 3.11.

Based upon the methodology and pre-defined parameters outlined in section 3.2.1 the lipid enumeration library generated a full complement of 10890 species, 645504 fatty acid scan species and 645504 sub species for 18 glycerophospholipid sub classes. The database population time was approximately 2 days on a Dual Core @ 2.60Ghz with 4GiB of memory. Typical generation time of the set of viable isomers for a lipid with many possibilities was under 5 seconds. The cross references annotation pipeline described in section 3.2.3.1 persisted 6670 isomers into the database and 13656 cross references. After the initial execution of the literature annotation pipeline described in 3.2.3.2, 204 unique papers were inserted into the database, annotating a total of 368 species, sub species, fatty acid scan species and fatty acid species. This breaks down into 0.62% of species and 0.01% of sub species annotated with a reference to the literature. The quality of the annotations were inspected by hand to check for accuracy and to manually reduce false positives. Unfortunately there were a collection of papers that when referring to a fatty acid scan

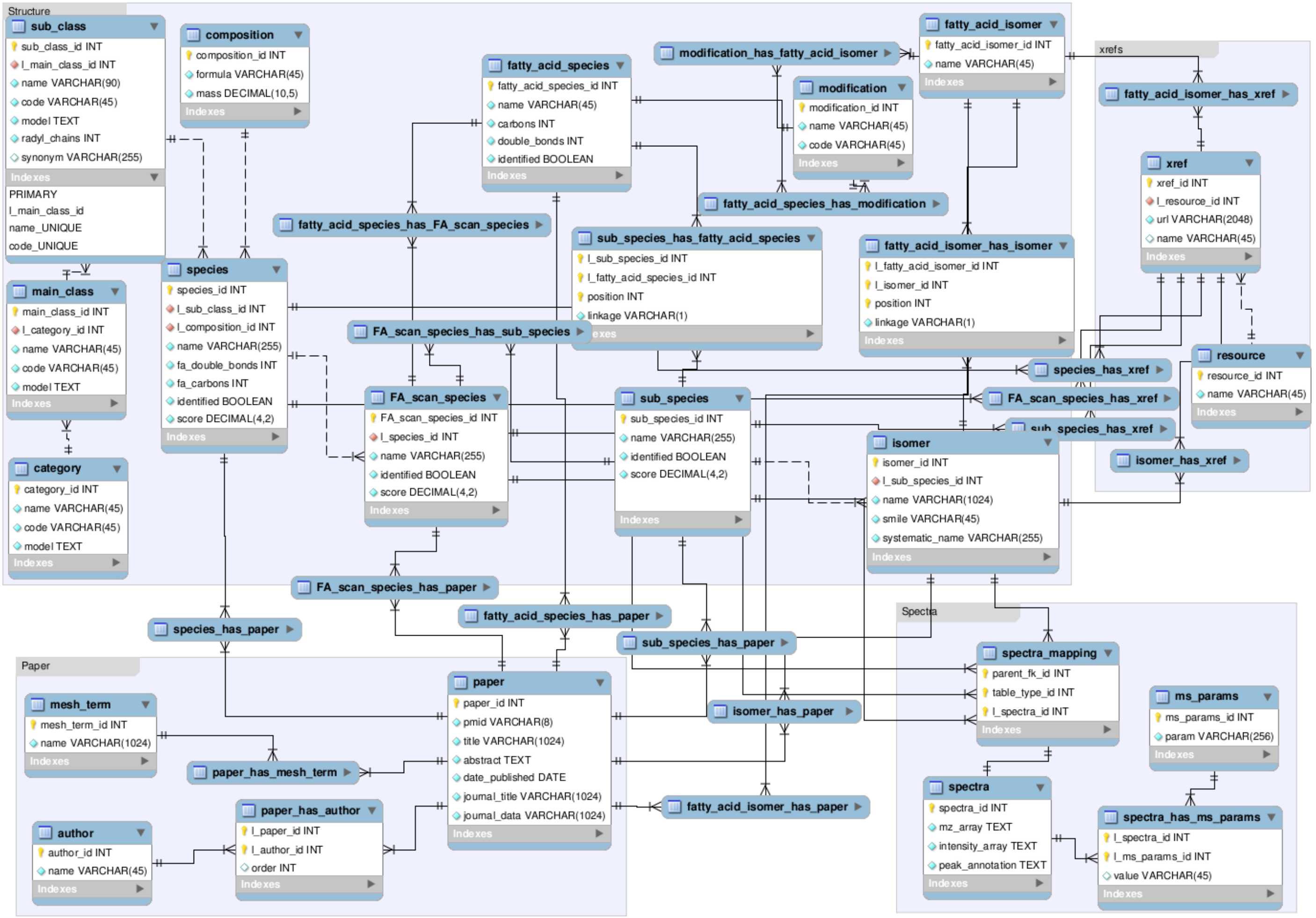


Figure 3.11: The LipidHome database schema. Tables are separated into the core structural data, relating levels of the lipid hierarchy to one another and the associated metadata.

species or sub species with identical fatty acids e.g. PC 18:1/18:1 it is shortened to PC 18:1. This causes a misannotation of the species PC 18:1 to the paper that is incorrectly referring to the sub species PC 18:1/18:1. These annotations were manually altered and marked for automatic corrective annotation in future execution of the literature pipeline.

3.3.1 Client design

Prior to developing the client using the ExtJS framework, each major view of the data was sketched using the 'Pencil' add-on for Firefox as described in section 3.2.5. The designs and a description of the various views of the data are described in the following sections.

3.3.1.1 Initial View

The application contains three main aspects; the browser, a selection of tools and help documentation on the data and methods used. These are organised in a tabbed panel with initial focus on the browser panel as it is the core of the application and the underlying database of lipid structures, see figure 3.12.

3.3.1.2 Browser

The browser is the core of the web application presenting the complex relationship of the lipids within the hierarchy of structural resolution. In this panel users can explore the data in the database and the associated metadata imported from other resources. The panel is comprised of five main sections each interconnected to reflect content (state) changes of one, in all the others (see figure 3.13).

Hierarchy Panel The hierarchical nature of the data lends itself perfectly to a tree structure, with one category having main classes, each having several sub classes, each with many species etc, as depicted in figure 3.2. Within the hierarchy panel a tree is constructed that upon initialisation contains only the highest level of information in the database; the category. These tree nodes can be selected and expanded to retrieve all its children in the next highest level of structural detail, in this case main class. Upon clicking a tree node,

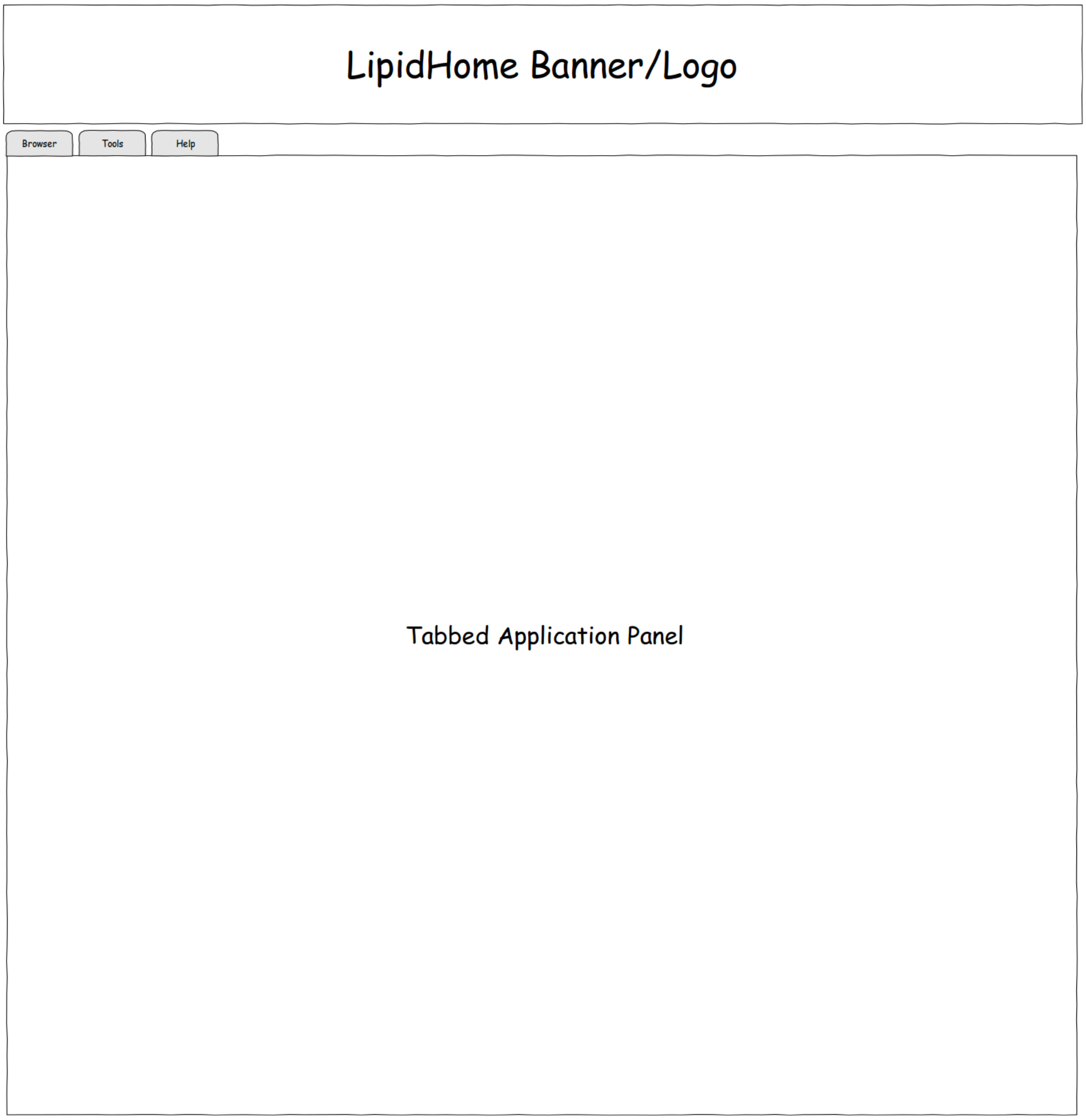
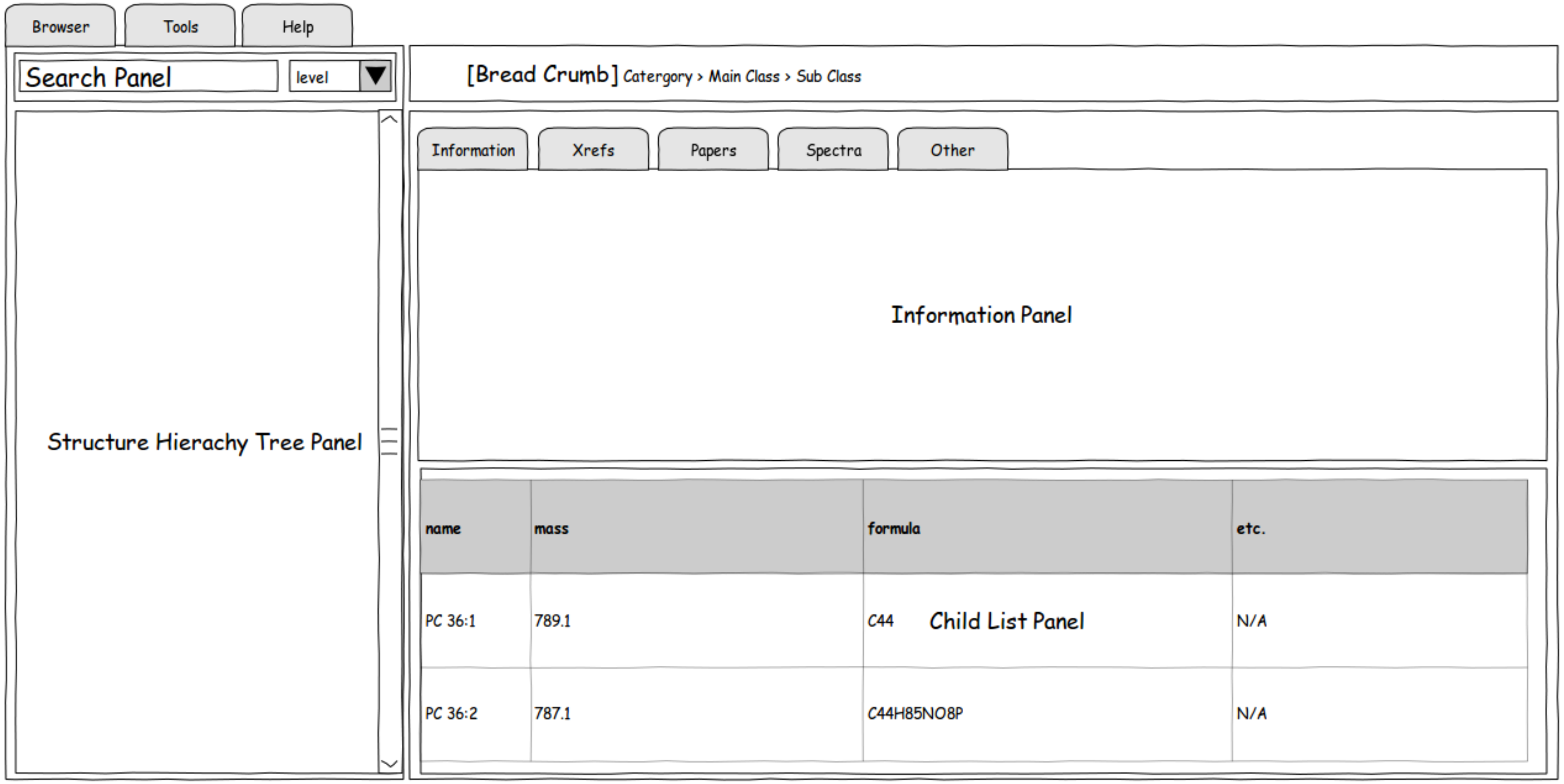


Figure 3.12: The initial view of the application follows a typical web page layout with a banner as header above the main application panel below. The three main components of the application are split into different tabs.



name	mass	formula	etc.
PC 36:1	789.1	C44	N/A
PC 36:2	787.1	C44H85NO8P	N/A

Figure 3.13: The Browser panel is the most important tab of the application panel and is split into five regions; Hierarchy Panel, Information Panel, Child List Panel, Path Panel and Search Panel.

information pertaining to that item in the database is displayed in the content panel. After a while of using the tree with many nodes and their children visible, it is important to keep a consistent view of the tree so that previously selected items can be easily found again. This is achieved by re-ordering the tree upon every insertion of a node alphabetically by the items name.

Information Panel Displays information about the item of interest in a selection of tabs depending on relevance to the structural hierarchy level of the item of interest. These tabs include general information such as number of children and a graphical representation of the lipid, along with more specialist tabs such as cross references to other resources and papers which mention it with a high probability. Molecules are rendered in the client side as described in section 3.2.6. This client-side solution renders dynamic molecule graphics that can be re-sized on the fly as the user re-sizes the information panel. To complement existing EBI products that render molecules, identical graphical parameters such as bond length, bond width, atom label font size and atom label colour were conserved where possible.

Child list Panel A table of all the children of the selected item and some simple information about them. In the case of selecting the sub class 'diacylglycerophosphocholine', a table of all of its species in the database is provided along with information such as their mass, formula and names. The items in this table are selectable and enable the user to traverse down the structural hierarchy to the level they require while updating the hierarchy panel to show the newly loaded children. Especially important to this application is the ability for a user to customise their view of the data due to the largely theoretical content of the database. To facilitate this, the child list's columns are filterable, sortable, removable and exportable to a number of data formats.

Path Panel The trail of items clicked to arrive at the current item e.g.
'Glycerophospholipids→Glycerophosphocholine→
Diacylglycerophosphocholine→PC 36:2→PC 18:0_18:2.'
Each element of the path panel is clickable to return to that parent item.

Search Panel While traversing the database via a combination of the hierarchy, bread crumb and child list for a known lipid it is also possible to search directly for that lipid in the search panel. Either all structural levels of the database can be searched or a specific one, e.g the species level with the query “36:2” will produce all species with the total number of carbons 36 and the total number of double bonds 2. The live search begins once a minimum of four characters have been typed into the search bar and a delay of 1 second since typing the last character has elapsed. Results are ordered primarily by structural hierarchy level, followed by identification status and finally alphabetically by name.

Communication between these disconnected components is clearly important. Items clicked and data requested in one panel must be propagated to all the others to maintain the state of the application and keep a consistent view of the data. This is managed by the ‘content manager’, a top level panel that contains all those previously mentioned in the browser. Events in one panel fire and ‘bubble up’ to the content manager which then passes them on to the relevant components who in turn listen for these remote events and execute the necessary response. For example clicking on a lipid in the child list will execute the loading of a new child list, loading of new information in the information panel, adding the requested child to the appropriate place in the hierarchy tree panel and appending the child to the bread crumb.

While the hierarchy panel and the search panel are singleton components, the others (path panel, information panel and child panel), collectively known as the content panel are numerous, see figure 3.14. Communication with the server and complex database queries are the slowest aspect of a web application and one of their major downfalls in comparison to a standalone desktop application which can rapidly access local resources. To avoid unnecessary communication with the server each view of an item (the combination of its bread crumb, information and children: the content panel) is stored in a panel. Upon requesting a new item this content panel is hidden and the new lipid requested and rendered. Upon attempting to view previously loaded data again in the same session no communication with the server is necessary, the current item is simply hidden and the selected existing data shown.

This works particularly well if some modification of the default components has been performed by the user (e.g. children ordered by mass and filtered for identified only), this modification will remain when the data is re-viewed.

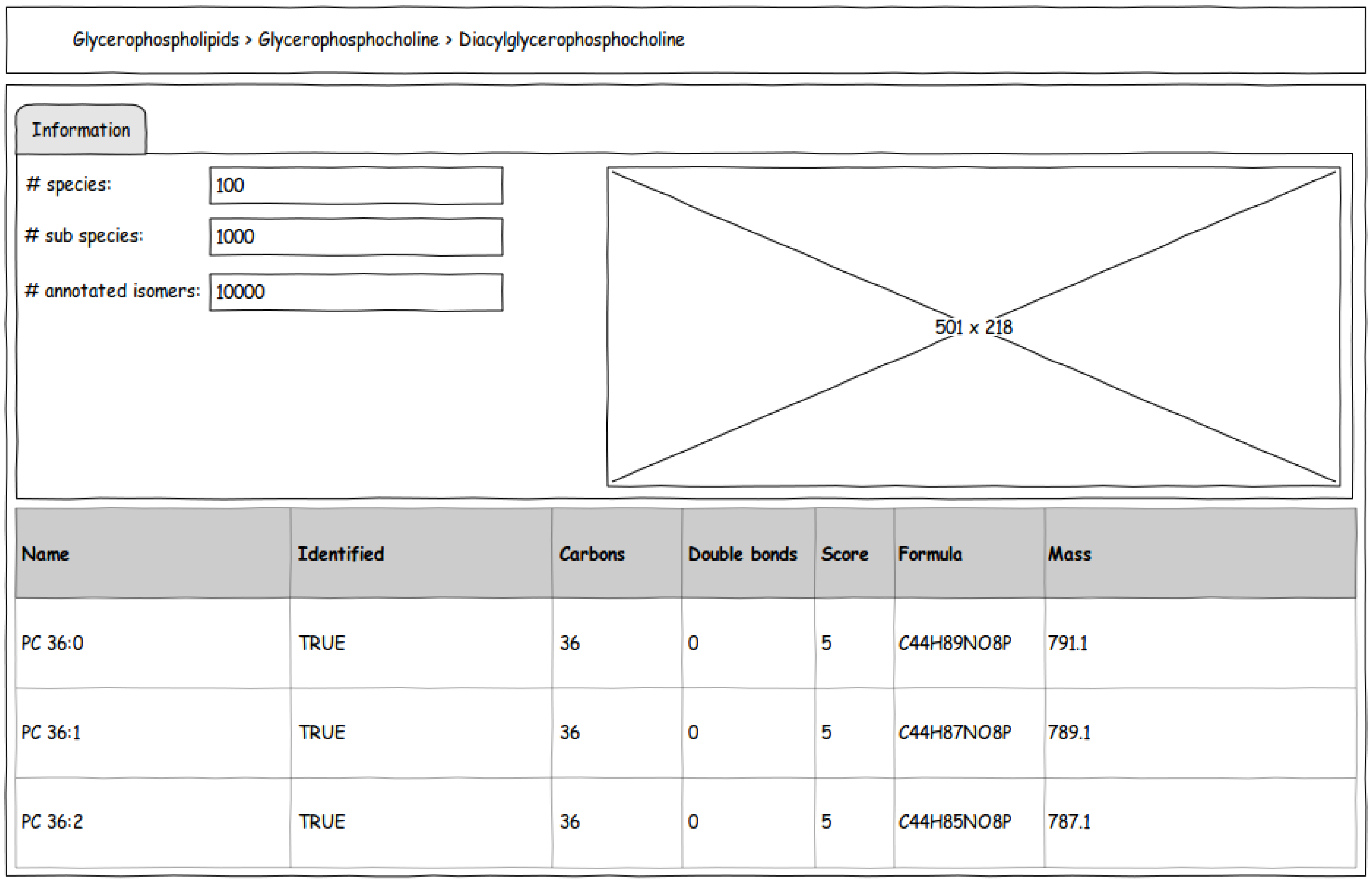
3.3.1.3 Tools

The tools panel is the location of a variety of lipidomics related tools for data analysis, data submission and visualisation. In order to improve usability, wherever possible, tools follow a standardised panel layout, so that while the analysis performed and the result produced by the tool may vary considerably the location of the core elements; input, output and description remain constant.

As a proof of concept a simple MS1 search is currently available allowing users to identify unknown precursor ion masses against lipids in the database and export the results.

MS1 Search Engine Unlike proteomics, automatic and high throughput identification of lipids is in its infancy, as is the resolution of lipid structure achieved by mass spectrometry. Amino acids represent a far less diverse set of building blocks and hence structural bounds than lipids where often such small differences as the position of double bonds in a fatty acid chain can have a significant biological effect, for example the ratio of omega-6:omega-3 essential fatty acids is attributed to a number of inflammatory/immune diease and cancer (Simopoulos, 2002). Described previously in section 3.1, MS1 precursor ion scan data can be collected where no fragmentation is performed and the structural resolution of the identified lipids is at the species level (total number of carbons and double bonds in fatty acids known). It is from this type of experiment that a list of precursor ion masses can be searched against the database to identify the likely species present in the sample. The MS1 search engine algorithm is outlined in section 3.2.4

The input panel comprises a text box in which a new line separated list of precursor ion masses for identification is pasted, a search parameters panel and an ion selection panel (see figure 3.15). The search parameters define the level of output, currently set to species only, as any identification at a more structurally defined level with MS1 data is not scientifically supported by the



Glycerophospholipids > Glycerophosphocholine > Diacylglycerophosphocholine

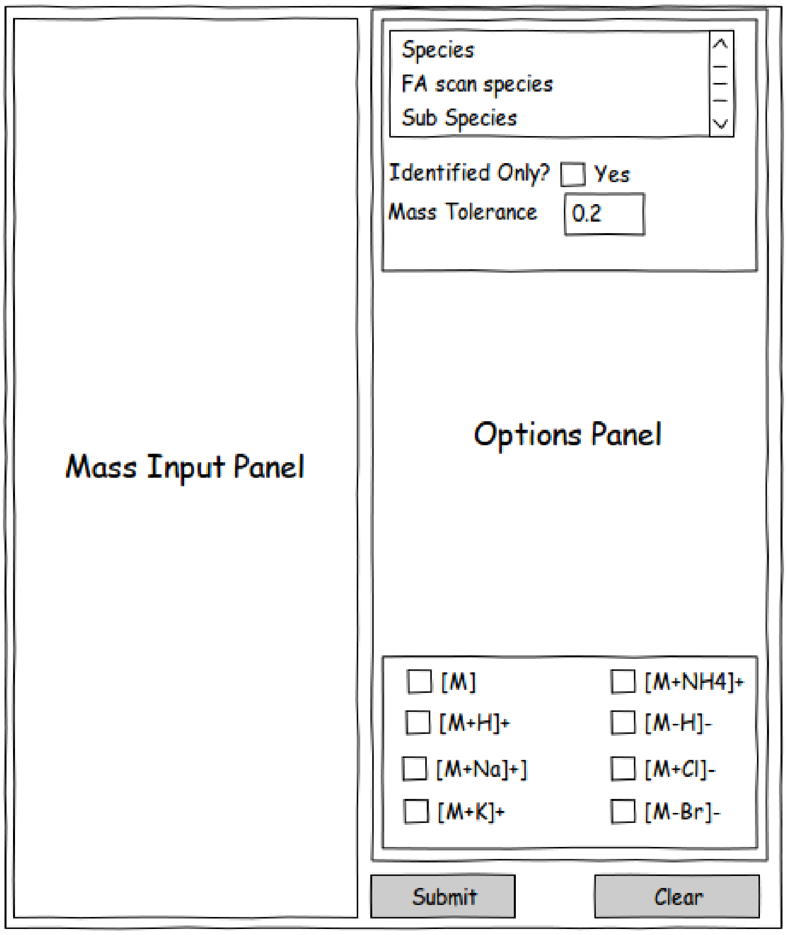
species: 100

sub species: 1000

annotated isomers: 10000

Name	Identified	Carbons	Double bonds	Score	Formula	Mass
PC 36:0	TRUE	36	0	5	C44H89NO8P	791.1
PC 36:1	TRUE	36	0	5	C44H87NO8P	789.1
PC 36:2	TRUE	36	0	5	C44H85NO8P	787.1

Figure 3.14: The Information Panel, Child List Panel and Path Panel are collectively known as the Content Panel. Information regarding the record requested is displayed along with an image in the top information panel and a list of all its children are displayed in a grid in the child list panel.



Mass Tolerance 0.2

Figure 3.15: The GUI design document for the MS1 search engine input panel. Masses are pasted into the scrollable Mass Input Panel, then the search parameters are configured in the options panel.

data. Also defined by the search panel is the mass tolerance selection spinner, where the machine dependent parameter mass tolerance can be selected in order to provide a fuzzy search of the composition table's mass records. Identified items can also be selected if the user requires a very strict search of the database and wants to reduce the number of valid but previously unreported lipids as hits in the result set. A technical detail that is extremely important to mass spectrometry is knowledge of the adduct ions in the sample under detection, these adducts alter the mass of the natural molecule while also providing charge to enable its detection inside the mass spectrometer. Any combination of adduct ions may be selected for inclusion in the search but each incurs an additional search for the parent search mass. For example, searching the database with the mass 785.59 Da and the adduct ions [M+H]+, [M+Na]+ and [M+K]+ requires 3 individual mass searches of the database; 784.59 Da, 746.63 Da and 761.61 Da respectively. See figure 3.15 for a design of the search input GUI.

The output panel displays the results of a successful MS1 search , grouping hits by the parent search mass in a series of tables, see figure 3.16. Like the child list of the browser, the columns of these tables are filterable, sortable and removable. Each hit gives information on the lipid's name, sub class, identification status, hit mass search mass - hit mass (Δ mass). Similarly to proteomics, it is often the case that a search mass may hit multiple different lipids. It is the responsibility of the experimentalist to report the final results and justify any arbitrary selection of the most likely hit. Lipid mass spectrometry is usually preceeded by a separation protocol wherein lipids from a complex sample can be separated into their main classes, often by solid phase extraction. Information from this sensitive and robust procedure can be used to filter out unlikely identifications. The output panel facilitates this with a check box tree containing the first three levels of the lipid structural hierarchy; category, main class and sub class. The user can choose an entire category or multiple main classes to select / deselect as viable hits, doing so to any parent, e.g. the main class 'glycerophosphoethanolamines' will automatically do this to all its children; 'diacyglycerophosphoethanolamines', 'dialkylglycerophosphoethanolamines' and 'monoacyl,monalkylglycerophosphoethanolamines'. Fil-

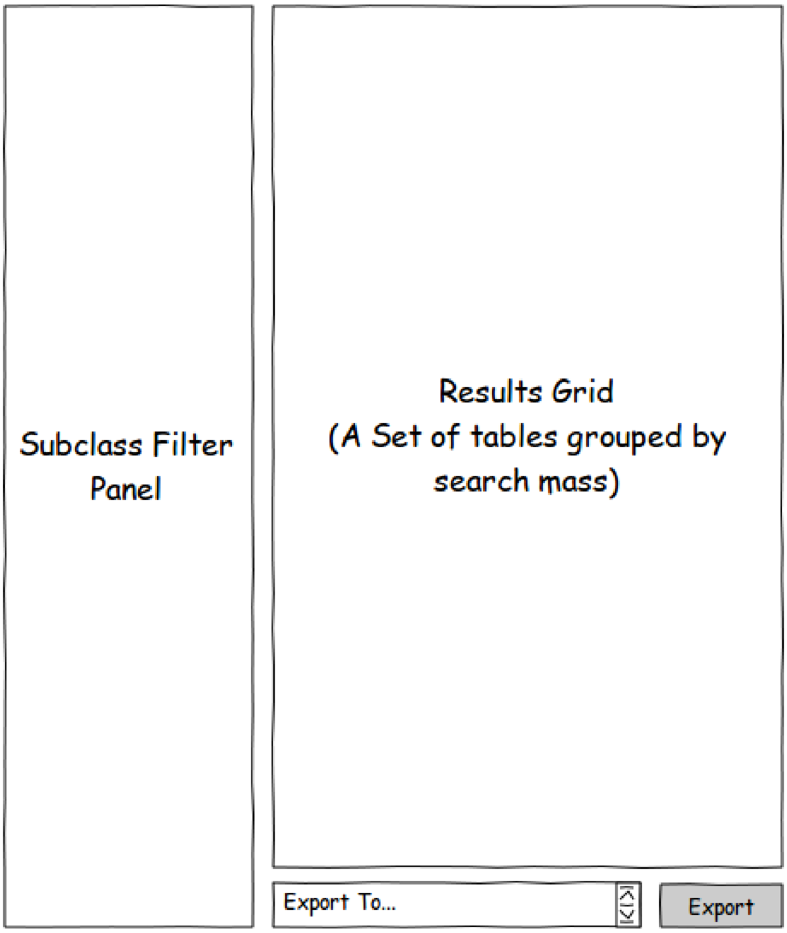


Figure 3.16: The GUI design document for the MS1 search engine output panel. Results are provided in a grid that shows the search mass, the species identified, the adduct ion type of the identification, identification status and the Δ mass.

tering via the tree is relayed to the results grid in a similar manner to that utilised by the browser for communication between its different elements with hits added and removed accordingly. Once the user has filtered, sorted and organised the results as they see fit, the data can be exported to multiple data formats for offline inspection/publication, preliminarily; CSV, TSV, XML, JSON and Microsoft Excel format (XLS).

As the project develops more tools will be added but for time constraints and sufficiency as a proof of concept, the MS1 search engine currently completes the tools section.

3.3.1.4 Documentation

Previously it was mentioned that this project received considerable interest from the lipidomics community, mainly via its regular presentation and discussion at LipidomicNet workshops. In order to produce a production piece of software and distinguish it from a prototype PhD concept, not only must the code base be well architected and implemented but documentation also encompasses a large part of the development. The quality of documentation of a new piece of software in the biological sciences can be the difference between a largely ignored nice idea to a popular brand around which a community of users can build itself. The documentation comes in several forms from simple tool tips that pop up on mouse over of important components to provide a short snippet of information, to a comprehensive entry describing the data, exactly how it was produced and any assumptions that were made in its production. The documentation contains articles on many topics that have already been explained in this chapter:

- LipidHome philosophy
- LipidHome content statistics
- Theoretical lipid generation
- Database schema
- A page for the major browser components, e.g. search bar and lipid hierarchy

- MS1 search algorithm
- MS1 search engine usage and data export
- Web services usage

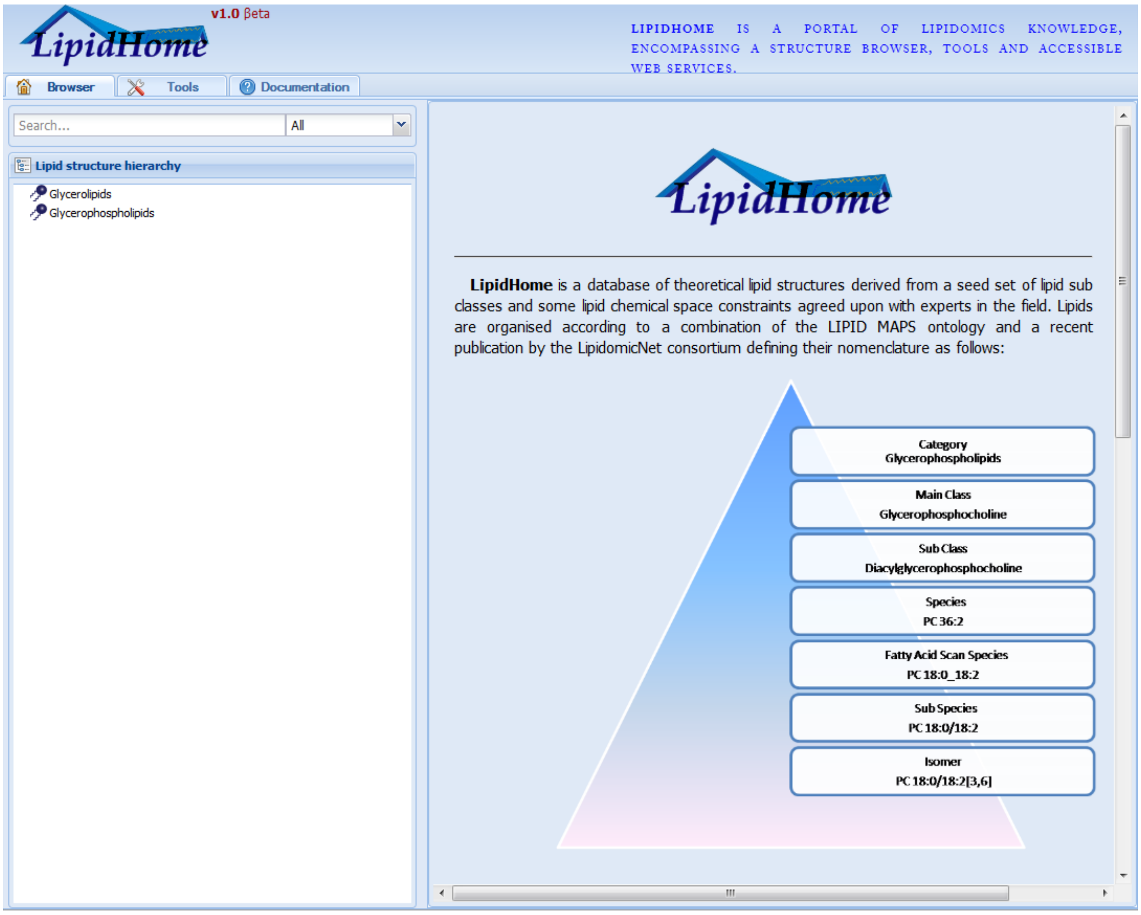
Along with scientific documentation and usage guidelines, technical documentation is also an important factor in creating a professional application, especially open source software, which has been designed to be re-implemented in other applications. Technical documentation is provided throughout the code itself, with additional installation/operational requirements documentation available as separate documents bundled with the code.

3.3.2 Client screenshots

Additional thought has been spent on the look and feel of the application, re-using essential design principles, to make using the application simple. Through the use of custom icons, standard component behaviour, colour and information repetition, this has been successfully achieved. At the time of writing the web application is currently accessible under the URL www.ebi.ac.uk/apweiler-srv/lipidhome and hosted by the EBI. Source code for the web application and web services code is available under an Apache 2.0 license hosted by Google at http://code.google.com/p/lipidhome/. Currently in external beta, the application has so far received excellent feedback from partners within the LipidomicNet Consortium, with multiple suggestions for future expansion of the application and its data. Present hereafter are a selection of screen shots of the application, but I would like to remind the reader that the flexibility, usability and potential of the application are best judged by using it.

3.4 Discussion

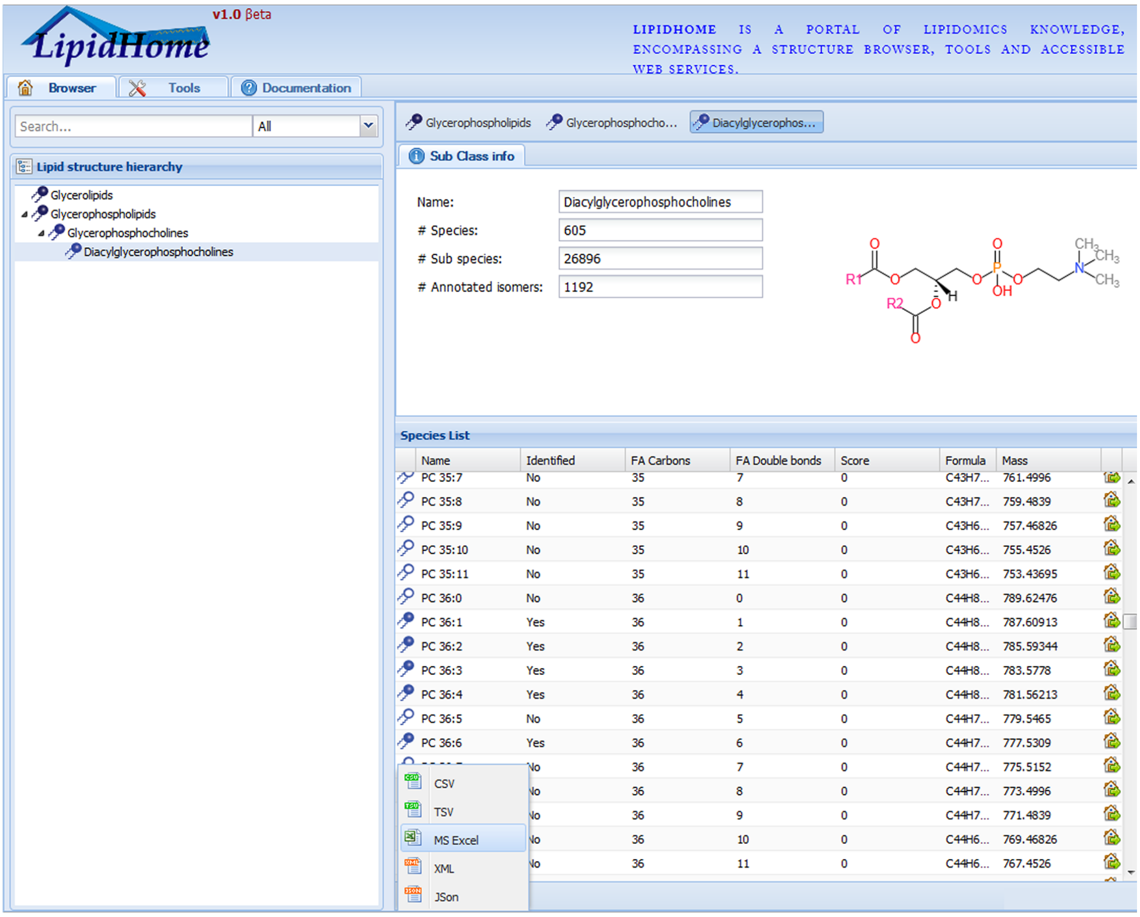
Comparison between LipidHome and the primary lipids identification database in the field, the LIPID MAPS LMSD gives a valuable summary of the new cutting edge of the field and insight into future work to improve LipidHome. The first



LIPIDHOME IS A PORTAL OF LIPIDOMICS KNOWLEDGE, ENCOMPASSING A STRUCTURE BROWSER, TOOLS AND ACCESSIBLE WEB SERVICES.

LipidHome is a database of theoretical lipid structures derived from a seed set of lipid sub classes and some lipid chemical space constraints agreed upon with experts in the field. Lipids are organised according to a combination of the LIPID MAPS ontology and a recent publication by the LipidomicNet consortium defining their nomenclature as follows:

Figure 3.17: The homepage of the LipidHome website, designed on the browser layout sketch in figure 3.13. Structural hierarchy panel on the left with search panel above it and the content panel to the right showing the homepage and a brief description of the structural hierarchy that underlies the data.



Name: Diacylglycerophosphocholines

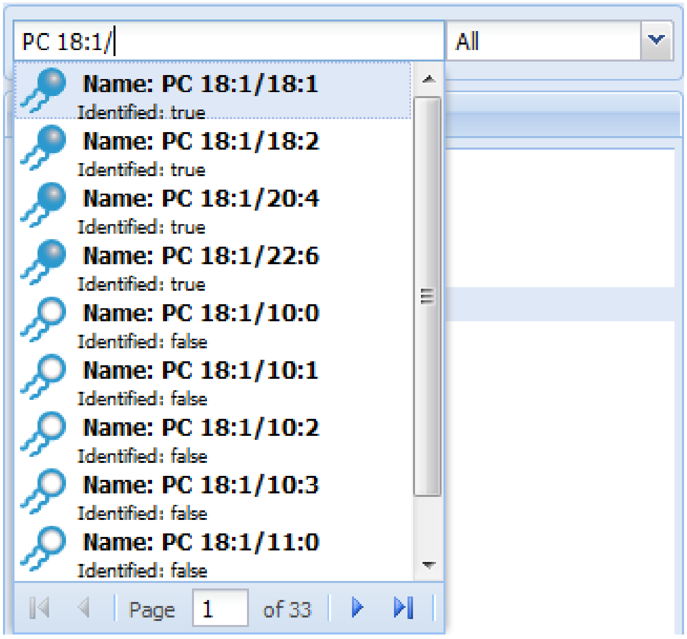
Species: 605

Sub species: 26896

Annotated isomers: 1192

Name	Identified	FA Carbons	FA Double bonds	Score	Formula	Mass
PC 35:7	No	35	7	0	C43H7...	761.4996
PC 35:8	No	35	8	0	C43H7...	759.4839
PC 35:9	No	35	9	0	C43H6...	757.46826
PC 35:10	No	35	10	0	C43H6...	755.4526
PC 35:11	No	35	11	0	C43H6...	753.43695
PC 36:0	No	36	0	0	C44H8...	789.62476
PC 36:1	Yes	36	1	0	C44H8...	787.60913
PC 36:2	Yes	36	2	0	C44H8...	785.59344
PC 36:3	Yes	36	3	0	C44H8...	783.5778
PC 36:4	Yes	36	4	0	C44H8...	781.56213
PC 36:5	No	36	5	0	C44H7...	779.5465
PC 36:6	Yes	36	6	0	C44H7...	777.5309
[illegible]	No	36	7	0	C44H7...	775.5152
[illegible]	No	36	8	0	C44H7...	773.4996
[illegible]	No	36	9	0	C44H7...	771.4839
[illegible]	No	36	10	0	C44H6...	769.46826
[illegible]	No	36	11	0	C44H6...	767.4526

Figure 3.18: A view of the sub class 'Diacylglycerophosphocholines'. Its parent lipids are visible in the structural hierarchy panel next to the general information about the sub class and all its children in the right hand top and bottom panels respectively. In addition, the export children list menu is visible highlighting the export capabilities of lipid information to CSV, TSV, MS EXCEL, XML and JSON formats.



PC 18:1/

All

Name: PC 18:1/18:1
Identified: true

Name: PC 18:1/18:2
Identified: true

Name: PC 18:1/20:4
Identified: true

Name: PC 18:1/22:6
Identified: true

Name: PC 18:1/10:0
Identified: false

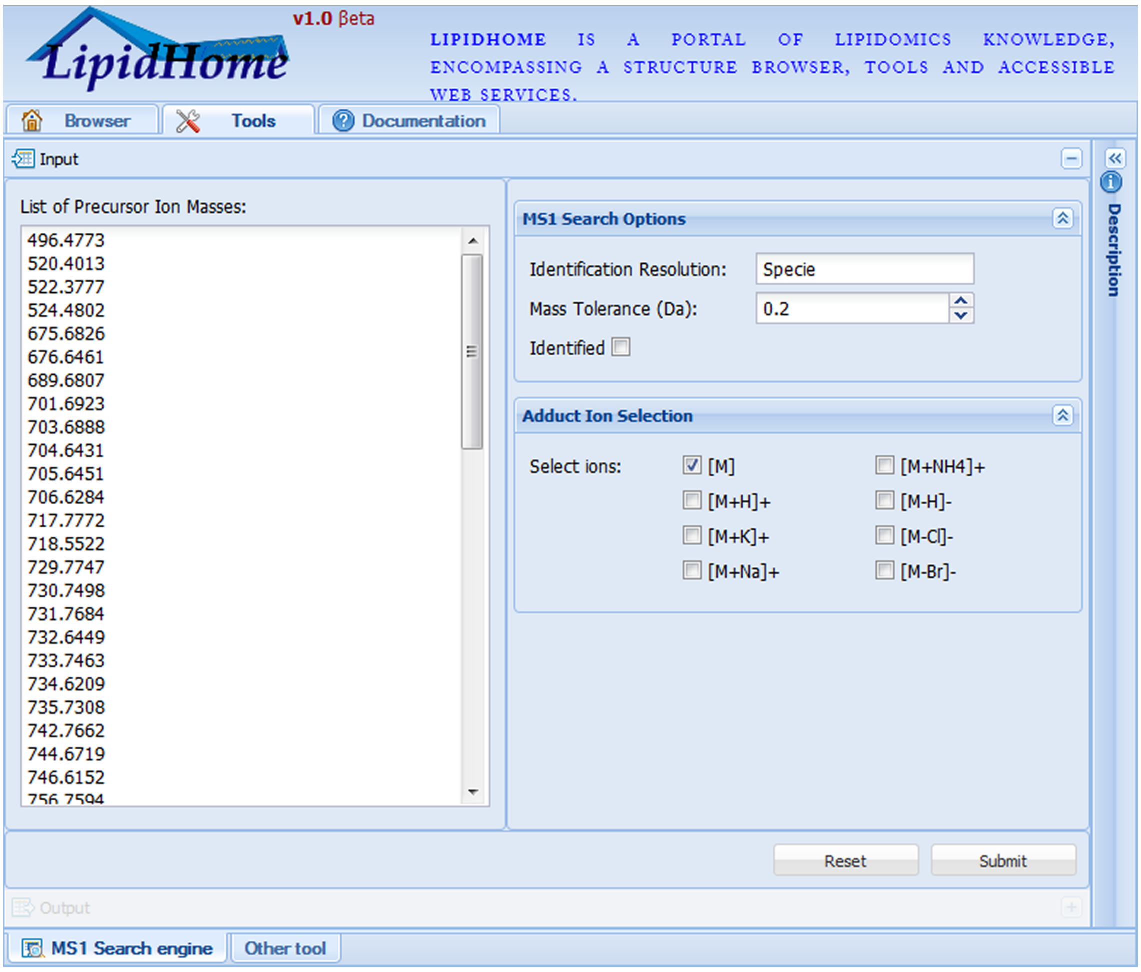
Name: PC 18:1/10:1
Identified: false

Name: PC 18:1/10:2
Identified: false

Name: PC 18:1/10:3
Identified: false

Name: PC 18:1/11:0
Identified: false

Figure 3.19: A screen shot of the LipidHome search bar and its results. The live search bar updates search results as the user types them in, results can be filtered to a particular structural hierarchy level using the combo box to the right of the search bar which defaults to searching the entire database. Results are ordered by structural hierarchy level followed by identification status (filled/unfilled lipid icon, identified/unidentified respectively) followed by alphabetical order. When more than ten results are returned they can be paged through. Clicking on a result will load the appropriate lipid into the hierarchy panel (recursively loading parent elements if necessary) and display the record and its children in the content panel.



LIPIDHOME IS A PORTAL OF LIPIDOMICS KNOWLEDGE, ENCOMPASSING A STRUCTURE BROWSER, TOOLS AND ACCESSIBLE WEB SERVICES.

List of Precursor Ion Masses:

```
496.4773
520.4013
522.3777
524.4802
675.6826
676.6461
689.6807
701.6923
703.6888
704.6431
705.6451
706.6284
717.7772
718.5522
729.7747
730.7498
731.7684
732.6449
733.7463
734.6209
735.7308
742.7662
744.6719
746.6152
756.7594
```

MS1 Search Options

Identification Resolution: Specie

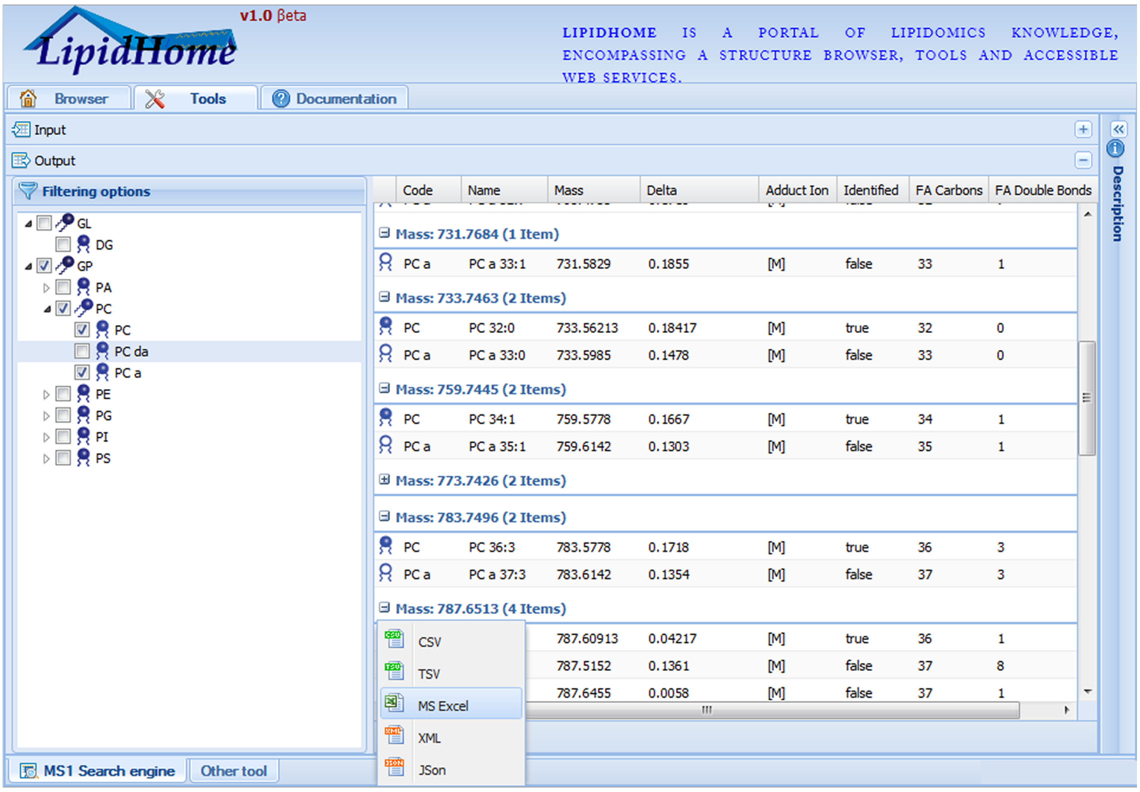
Mass Tolerance (Da): 0.2

Identified □

Adduct Ion Selection

Select ions: ☑ [M] □ [M+NH4]+ □ [M+H]+ □ [M-H]- □ [M+K]+ □ [M-Cl]- □ [M+Na]+ □ [M-Br]-

Figure 3.20: A screenshot of the input panel of the MS1 search engine, designed in figure 3.15. It is pre-loaded with the same test data as the equivalent service from LIPID MAPS so that a direct comparison can be made. Due to the more detailed coverage of the glycerophospholipids category, this tool identifies significantly more lipids. Adduct ions can be selected for and the mass tolerance appropriate to the instrument under which the precursor ions were detected is set using the spinner.



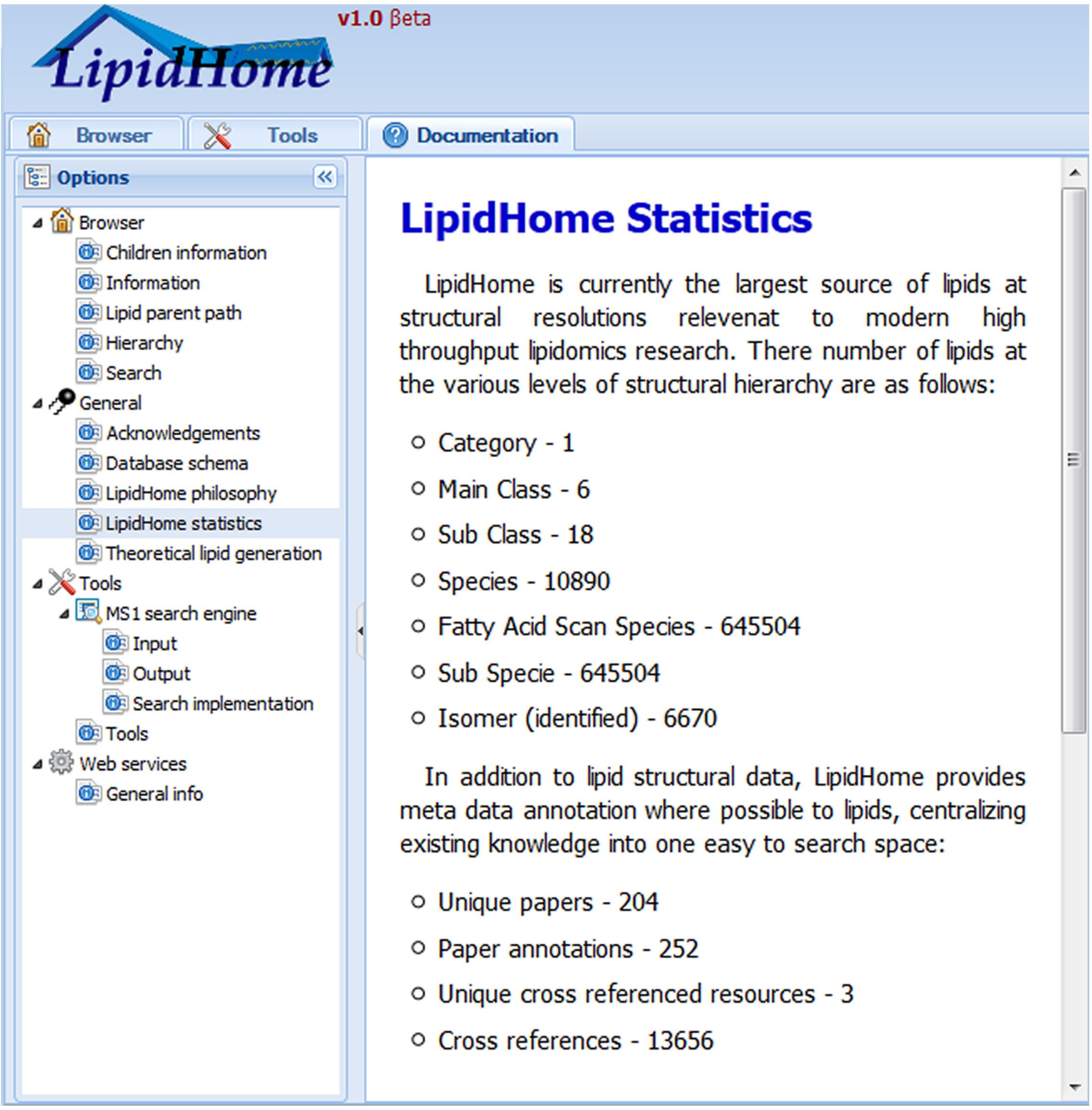
LIPIDHOME IS A PORTAL OF LIPIDOMICS KNOWLEDGE, ENCOMPASSING A STRUCTURE BROWSER, TOOLS AND ACCESSIBLE WEB SERVICES.

Filtering options

- ◢ □ GL
  - □ DG
- ◢ ☑ GP
  - ▷ □ PA
  - ◢ ☑ PC
    - ☑ PC
    - □ PC da
    - ☑ PC a
  - ▷ □ PE
  - ▷ □ PG
  - ▷ □ PI
  - ▷ □ PS

Code	Name	Mass	Delta	Adduct Ion	Identified	FA Carbons	FA Double Bonds
⊟ Mass: 731.7684 (1 Item)							
PC a	PC a 33:1	731.5829	0.1855	[M]	false	33	1
⊟ Mass: 733.7463 (2 Items)							
PC	PC 32:0	733.56213	0.18417	[M]	true	32	0
PC a	PC a 33:0	733.5985	0.1478	[M]	false	33	0
⊟ Mass: 759.7445 (2 Items)							
PC	PC 34:1	759.5778	0.1667	[M]	true	34	1
PC a	PC a 35:1	759.6142	0.1303	[M]	false	35	1
⊞ Mass: 773.7426 (2 Items)							
⊟ Mass: 783.7496 (2 Items)							
PC	PC 36:3	783.5778	0.1718	[M]	true	36	3
PC a	PC a 37:3	783.6142	0.1354	[M]	false	37	3
⊟ Mass: 787.6513 (4 Items)							
		787.60913	0.04217	[M]	true	36	1
		787.5152	0.1361	[M]	false	37	8
		787.6455	0.0058	[M]	false	37	1

Figure 3.21: A screen shot of the output panel of the MS1 search engine tool. After submitting a search the MS1 search engine output panel is expanded to reveal the result of the search. The left hand check box hierarchy panel is used to filter results to only view hits within specific sub classes. Results are grouped by the parent search mass, with each set of adduct ion hits being displayed in its own collapsible table. Results are primarily ordered by the identification state of the lipid hit, followed by the delta mass of the search mass and the adduct ion. The results in the right hand panel can be sorted and filtered by the various columns to reduce unlikely or uninteresting results. After the results have been prepared by the user they may be exported into a number of file formats for publication or further offline analysis.



LipidHome Statistics

LipidHome is currently the largest source of lipids at structural resolutions relevenat to modern high throughput lipidomics research. There number of lipids at the various levels of structural hierarchy are as follows:

- Category - 1
- Main Class - 6
- Sub Class - 18
- Species - 10890
- Fatty Acid Scan Species - 645504
- Sub Specie - 645504
- Isomer (identified) - 6670

In addition to lipid structural data, LipidHome provides meta data annotation where possible to lipids, centralizing existing knowledge into one easy to search space:

- Unique papers - 204
- Paper annotations - 252
- Unique cross referenced resources - 3
- Cross references - 13656

Figure 3.22: A screent shot of the documentation panel. The documentation section follows similar design principles to the rest of the application, navigation panel on the left and content on the right. Documentation is ordered into different categories to easily navigate to the topic of choice.

comparison is that of the data content. LipidHome contains the complete coverage of a small part of the chemical space that defines lipids as opposed to the LMSD which covers a much larger portion of the chemical space much more sparsely. This is the key difference in philosophy between the two resources, LipidHome provides the theoretical chemical space with no judgment on likelihood of existence or organism of origin. The LMSD contains a strict set of manually curated lipids of unknown and mixed origins with some spectral evidence but no systematic evidence system to annotate the others or justify their inclusion. The cross references available in LMSD and LipidHome for the intersecting lipids present in both resources are identical due to the way they were harvested primarily from the LMSD. Similar to the LipidBank, LipidHome contains paper annotations for each lipid that was mentioned in a paper's abstract. The LMSD does not contain paper information which can be useful in quickly researching a lipid of interest or obtaining protocol information for identifying a lipid. When viewing a single lipid's result page in the LMSD chemical descriptors are provided such as logP and topological polar surface area. LipidHome does not calculate chemical descriptors of isomers, but this would be a great addition. The computational accessibility of LipidHome is well supported through its flexible web services, comparable web services in the LMSD are less flexible and the output formats more restrictive. As it stands LipidHome is the only reference space for lipid identifications higher than the geometric isomer structural hierarchy level.

A number of use cases are supported by the web application and this is set to grow in the near future, they can be grouped as:

Literature search Users looking for literature on a particular lipid are served by typing the name or part of it into the search box and selecting the lipid in the search result. The lipid will be loaded into the browser, the user then selects the papers tab of the information panel to view the papers in which the lipid is mentioned. The list of papers can be expanded to read their abstracts and other information about authors, journal and year of publication.

Lipid identification reference When reporting lipid identifications a laboratory may refer directly to its entry in the LipidHome database. A lipid can also be seen if it has been identified before.

Nomenclature reference LipidHome acts as a reference point for the most up to date nomenclature information about lipids, and helps promote the use of standardised names to members of the community who have not yet adopted the naming system or are unfamiliar with its details.

Precursor ion scan Experimentalists with precursor ion scan lists can have them identified against the growing number of lipid in the database and export the results in standard file formats and data exchange formats.

Lipid research LipidHome is a centralised knowledge base of lipidomics information. From its lipid records the user can link out to other resources that may have additional information.

Lipid images Drawing lipid images, especially for use in publications, can be difficult without specialist software. LipidHome provides images for its lipids in standard formats with appropriate bond lengths and widths, and atom colours. They can be saved as portable network graphics (PNG), suitable for use in presentations and publications.

Bioinformatics LipidHome has been designed from its outset to be a platform for bioinformaticians to developed lipidomics software on top of e.g. the MS1 precursor ion search engine. A LIMS system could also be created to store lipidomic datasets that would reference LipidHome by its lipid identifications.

Originating as a quick solution to an isolated problem within the LipidomicNet Consortium, the project has evolved from a simple list of theoretically generated lipid names, to a production standard web application giving access to a wide array of lipid records and interesting metadata. It provides lipid identifiers relevant to the modern lipidomics community and encourages the use of standard data formats, standard lipid representation and improved lipid nomenclature. Through standardised nomenclature lipidomics information can be easily aggregated and shared between laboratories, this will become the first step in any data analysis protocol. In addition to the original demands of the database, it provides a number of tools but most importantly, a platform on which to build future applications for the benefit of the lipidomics community. The success of lipidomics in the biological sciences can be clearly extrapolated against the rise of proteomics and its foundation on equivalent services such as UniProt. It is through imitating the tools of other ‘omics’ fields, as has been done here, that lipidomics can be advanced to the next level and become truly high throughput.

3.5 Future work

The current version of the database and web application are a proof of concept for the effective storage of lipid identifications and their associated metadata. The database and its web services provide an effective platform for the continuation of this work, expanding the data and providing additional services. Future work could include:

3.5.1 Increased lipid coverage

A wider proportion of the lipidomics community could be supported by including new lipid classes and categories, most prominently sphingolipids, cardiolipins, glycerolipids and synthetic lipids, to supplement the current glycerophospholipids. The process of generating theoretical lipids described in section 3.2.1 will need to be extended. The current set of lipid sub classes all share the same structural features described in figure 3.8. The necessary changes will include:

Branched fatty acids Generation of fatty acids with branches. This can explode the chemical space in which viable lipids reside, so it will be very important to understand the rules imposed by nature, that have been witnessed by the community, to solve this.

Non-diradyl sub classes Cardiolipin, for example, has four fatty acids, triacylglycerol three fatty acids and lysoglycerophospholipids a single fatty acid. The code must be able to combine up to four fatty acids before crossing them with the core + head group. Conceptually, this problem is not much more complex than the diradyl case, but it may present problems in the time frame necessary to generate all theoretical possibilities with realistic computational power and time.

3.5.2 Improved metadata

More sources of lipid related data could be integrated into the database, each resource having a custom parser so that annotation could be kept up to date and of the highest possible detail. This would allow a wider array of lipid centric information to be displayed in one place to lipidomic scientists, exposing them to new resources they had previously not discovered. As an example, the Reactome database (Croft *et al.*, 2011) (`www.reactome.org`) is a database of human pathways and the reactions, translocations and complex formations that constitute them. It provides details of pathways involving lipid molecules at the sub class level. The information of which lipids form which parts of the human pathways would be an extremely valuable addition to LipidHome, and the data easily presented in a new tab of the information panel like papers are currently. Aside from simply more cross references, an individual parser for each external resource allows the independent inclusion of new information rather than depending upon LIPID MAPS to include it first. A very small precentage of records in the database are associated with a paper; 0.62% of lipid species. The text mining approach outlined in section 3.2.3.2 must be improved to provide evidence for a large portion of the theoretical lipidome.

3.5.3 MS2 search

Lipidomic research labs are increasingly performing MS2 experiments to characterise the fatty acid species on lipid species to result in fatty acid scan species and sub species identifications. The spectrum identification process is considerably more complex for these experiments, an open access and methodologically transparent MS2 search engine would be an excellent addition to the tool suite.

3.5.4 Community sourced annotation

Automatic annotation of lipids is quick and easy but the quality of results does not compare favourably to manually annotated data. Manual annotation of data often requires whole teams of expert curators to read the literature and dissect out lipid terms and store them in the database. This is not feasible for such an early stage project with no dedicated funding. As an alternative the web application could support community sourced annotation by providing a control panel wherein visitors could suggest cross references and papers to annotate lipids. Community sourced annotation quality could be quantified by 'thumbs up/down' rating system for each annotation (similar to the system popularised by YouTube), allowing users to make sure the most relevant annotation can be highlighted, and less informative annotation be flagged for deletion.

3.5.5 Usage tracking

A rather distant but necessary development marker will be the introduction of application usage logging. Using a library like Google Analytics (www.google.com/analytics/) it is possible to log which features of the application and web services are being most used, which lipids are being most heavily accessed and which lipids are being identified via the MS1 search engine. This information will help shape the future development of the application and give some much needed, albeit passive user feedback, on how the application is most used.

3.5.6 Chemical descriptors

For lipids in the isomer level of the structural hierarchy it is possible to calculate a number of chemical descriptors such as logP, molar refractivity and van der Waals molecular volume. These descriptors could be generated on the fly as isomers are selected and displayed in the client via a new web service. There are a number of libraries for generating chemical descriptors in various different programming languages, but the industry standard is JChem by ChemAxon (www.chemaxon.com/jchem/intro/index). Thankfully, the library is written in Java and as such would require no major changes to the server to integrate. A "Free web" license for academic purposes has already been negotiated with ChemAxon.

3.5.7 Spectral Library

Particularly important to lipid identifications is the avilability of experimental spectra which can be used to aid the identification process. A pertinent example from the field of proteomics can be found in its equivalent sequence database. UniProt contains cross references to PRIDE (Martens *et al.*, 2005) and other resources like Peptide Atlas (Desiere *et al.*, 2006) for protein sequences which have evidence from mass spectrometry experiments. The cross references link to the experiments in which the protein of interest is detected and then on to the peptide spectrum match and the spectrum itself. In summary there is no analagous resource like PRIDE for the storage and access of public lipidomics data. Although this is likely to change in the future as good practices from one 'omics' field are integrated into others. Without this type of resource available to the lipidomics community it is difficult to find publicly available spectra to cross reference LipidHome records with. LIPID MAPS contains a total of 244 glycerophospholipid standard spectra under measured and controlled conditions. However, this data is provided only as an image, which is not suitable for computational use in building a lipid identification search engine. Lacking the resources to create a spectral library the alternative is to generate a theoretical spectral library, as outlined in section 3.2.3.3. Should the resources or collaboration become available to build an experimentally derived spectral library for a set of commercially available lipids and other highly purified/synthetic lipids, LipidHome is well positioned to quickly integrate this data into the database and visualise it in the web application.

3.6 Acknowledgements

This work would not have been possible without the collaboration of my fellow PhD student Pablo Moreno in the revised implementation of theoretical isomer generation via CDK, and Antonio Fabregat without whom, the web application would be a considerably more amateur and prototype affair. Additionally, I would like to thank the Professor Michael Wakelam and Dr Juan Antonio Vizcaíno for their foresight into the potential of this resource and the standardised approaches it champions.

Chapter 4

Colorectal Cancer Lipidomics

4.1 Introduction

From a biological perspective the lipidome represents an interesting sub set of the metabolome, namely the hydrophobic and organic acid soluble fraction. Similar to the proteome it is much closer to the physiological control of an organism than the genome, dictating changes in signalling (Piomelli, 1993), energy metabolism (Hoch, 1998) and cellular trafficking (Indiveri *et al.*, 1991) amongst others. With this much finer and more rapid control comes greater complexity. It is estimated that 23,000 protein coding genes translate to 58,345 proteins (UniProt Human Reference Proteome) but quite likely considerably more (117,937 human proteins in UniProt KB; release 2012 04). The genome being only composed of four bases and the proteome only composed of twenty one amino acids has much more stringent limits to its potential theoretical complexity. While the sequences are typically very long, actual identification of sequences are usually quite short and later assembled into longer sequences. However, theoretical complexity of the proteome and genome is still astronomical, but the restriction in building blocks makes identification easier. In comparison, the lipidome is also immeasurably complex. From expert agreed criteria, 645,504 feasible glycerophospholipid sub species were generated (3.2.1) without any of the other lipid categories; glycerolipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids and polyketides. Hence, lipidomics could be described as the identification and quantification of those lipid species in a biological system that constitute the lipidome (Wenk, 2010).

4.1.1 The State of the Field

Originally cementing its success in the small molecules field, mass spectrometry has been the pioneering high throughput technology for most non-nucleotide ‘omics’ disciplines. Advances in mass spectrometry have propelled the measurement of lipids into the high throughput era, coupled with front end separation many hundreds of lipids can be isolated, identified and quantified in a single short experiment. Comparable to the instrumentation in the field of proteomics, lipid mass spectrometry is performed on almost identical machines with the main difference being in the separation technology that simplifies a complex sample in time prior to it entering the mass spectrometer for detection. Additionally, lipid mass spectrometry more regularly utilises the negative ion mode for detection of many negatively charged species which is primarily used for phosphorylation studies in proteomics (Old *et al.*, 2009). Like proteomics, lipid mass spectrometry has a few higher level flavours than simply MS1 or tandem mass spectrometry (MS2). Experiments can be described as untargeted, focused and targeted (Navas-Iglesias *et al.*, 2009).

Untargeted The comprehensive profiling of an entire biological sample. Useful for detecting a large proportion of the most abundant lipids in a sample. Very little sample preparation is required making this method favourable due to its simplicity.

Focused More detailed identification of a sub set of lipids, typically defined by the separation and isolation of a single category or main class of lipids. A combination of precursor ion scanning, product ion scanning and neutral loss scanning may be used to identify the lipids.

Targeted Analogous to the term used in proteomics, targeted analysis refers to the precise measurement of several distinct lipid species from a list of interesting candidates. This technique often employs Selected Reaction Monitoring (SRM) where a known precursor ion mass is selected for, fragmented and specific product ions detected to confirm the presence or absence of specific lipids one at a time.

Following acquisition of spectra, biomolecules must be identified using specialist software. Proteomics provides a plethora of complex algorithms for converting raw spectra into peptide identifications (Nesvizhskii, 2010). Lipidomics suffers greatly from a lack of robust informatics to identify mass spectra and resolve lipids to structural resolution. Typically, the filtered mass list is cross referenced with a laboratory specific list of known lipids which each mass is likely to correspond to, often by hand and at structural hierarchy level not supported by the data. For example; a single precursor ion mass of 786.6007 reporting an identification of the [PC 36:2 + H]+ ion is commonly misreported as one of its most abundant sub species PC 18:0/18:2, a level of detail which a single precursor ion mass clearly does not provide enough structural resolution to support. This is neither systematic nor accurate and careful scrutiny must be made of published results prior to re-using the data or further developing conclusions. In addition, lipidomics lacks the bioinformatics resources available in proteomics and genomics that have advanced the pace of these other fields of research, such as equivalent sequence databases (see chapter 3 for a prototype of such a resource) and public experimental data repositories. With time, lipidomics will reach the current standards of research found in other 'omics' disciplines, but for now it represents a great opportunity for early adopters to implement existing technology into this field and translate data analysis patterns to a much less explored area of modern high throughput data.

4.1.2 Quantitative Colorectal Cancer Lipidomics

This chapter will discuss the analysis of a dataset provided by the Babraham Institute, wherein lipids were identified and quantified in a large patient cohort suffering from some stage of colorectal cancer. Approaches from fields such as proteomics and genomics are recycled into effective analyses and data transformations of this lipidomics data. Previously described best data practices such as using defined nomenclature (see section 3.1.2) and quality control of the data prior to analyses (see chapter 2) are also adopted.

4.1.3 Colorectal cancer

Colorectal cancer presents a significant problem in the developed world with an estimated 1.24 million new cases diagnosed in 2008 of which almost 50% proved fatal (Ferlay *et al.*, 2010). While colorectal cancer predominantly affects the older proportion of society with almost 85% of diagnoses occurring in over 65-year-olds, it also shows high prevalence in younger people with prior inflammatory bowel conditions such as Crohn's disease and ulcerative colitis (Jawad *et al.*, 2011; Triantafillidis *et al.*, 2009). Hereditary factors constitute approximately 20% of all cases and provide a two to three fold greater risk in developing the disease. Aside from hereditary factors and predisposition due to inflammatory bowel diseases, diet represents a significant correlative factor in development of the disease, especially an excess of alcohol or red meat (Watson & Collins, 2011). The disease is stratified by a modified Duke's classification system, from the least severe to the most:

Adenoma A benign tumour that does not cause serious health problems

Stage A Confined to the intestinal mucosa

Stage B Growth into the muscle layer

Stage C1 Growth penetrating the muscle layer without involvement of the lymph nodes

Stage C2 Growth penetrating the muscle layer with involvement of the lymph nodes

Stage D Widespread metastases

Early diagnosis is critical to patient outcome with 5 year survival rate rapidly decreasing as stage severity progresses. Any abnormal swelling or adenomatous polyps are typically detected via rectal examination with a follow up colonoscopy. During a colonoscopy any tumour tissue is resected and staged, to determine any further action that may need to be taken e.g. chemotherapy in the case of stage C2 tumours involving the lymph nodes. Five year survival rate of different stages of colorectal cancer highlights the importance of early diagnosis:

Stage A 5 year survival rate: 100%

Stage B 5 year survival rate: 90%

Stage C1 5 year survival rate: 70%

Stage C2 5 year survival rate: 40%

Stage D 5 year survival rate: 5%

The Duke's classification system is not the only systematic way to describe a tumour. For example, the TNM system describes the tumour size (T0-4), regional lymph node involvement (N0-3) and distant metastasis (M0-1). The system is widely used in clinical databases and for communicating patient diagnosis in a standardised manner between the numerous clinicians involved in a cancer patient's treatment. TNM staging can be accomplished by a number of methods and combinations thereof, including; physical examinations, medical imaging such as computed tomography and magnetic resonance imaging, laboratory tests of blood, urine and other fluids and reports from surgery.

While much is known about the genetic factors associated with colorectal cancer (Parsons *et al.*, 2005), the role of the lipidome in the progression, prevention and rehabilitation of the disease is largely unstudied and as such is an extremely interesting area of data analysis to be involved in.

4.1.4 Existing literature

To the knowledge of my collaborators and I, this is the first investigation of its kind into the large scale analysis of a relatively large cohort of human colorectal cancer data. Significant research has been undertaken into human colorectal cancer in the field of proteomics, summarised concisely by (Jimenez *et al.*, 2010; Tjalsma, 2010). The traditional fecal occult blood test used for diagnosis of colorectal cancer amongst other diseases is not as specific as some of the most recent panels of proteomic biomarkers (Bosch *et al.*, 2011). There are many genomics efforts to categorise the difference between normal and colorectal cancer samples, each reporting a small and overlapping set of genes. However, a recent paper took a genomics approach to identify genes and then proteins relevant to colorectal cancer, which were then overlaid onto protein-protein interaction data (Li *et al.*, 2012). Twenty six patients, each with a tumour and a histologically matched adjacent normal sample were subject to microarray analysis. A list of 6 genes was identified as significantly different between the two sample types. The 6 genes were mapped onto a protein-protein interaction network and the shortest path that connected them all computed. The new proteins that were part of the interaction sub-network were included as interesting hits, bringing the total number to 41. These 41 gene targets showed an improved combined level of cancer related annotation from sequence database, thus validating the approach. A literature search using PubMed for "colorectal cancer lipidomics" returns a single paper which documents the increased abundance of long chain fatty acids and its derivatives in colorectal cancers (Gassler *et al.*, 2010). The molecular mechanisms behind this phenomenon are reported as not well understood. The equivalent search for proteomics returns 61 articles. There is clearly more than a single paper's worth of research into colorectal cancer lipidomics, but this does highlight the disparity between the effort spent on lipidomics and on other 'omics' fields. Outside of colorectal cancer there have been several recent human clinical lipidomics studies varying in patient size and number of lipids quantified, below is a summary:

Patients	**Lipid**	**Sample**	**PMID**
20	23	Plasma	22795978 (Stübiger *et al.*, 2012)
10	18	Macrophage	22728312 (Sewell *et al.*, 2012)
182	59	Plasma	22713803 (Draisma *et al.*, 2012)
9	122	Retina Optic Nerve	22496896 (Acar *et al.*, 2012)
53	251	Serum	22257447 (Orešič *et al.*, 2012)
83	81	Plasma	22009255 (Hu *et al.*, 2011)
38	250	Liver	21857953 (Gorden *et al.*, 2011)
47	33	Plasma	21779331 (Han *et al.*, 2011)

Table 4.1: A table of recent human clinical lipidomics studies showing the number of patients involved and the nubmer of lipids quantified in various samples.

Lipidomics has a lot of biological insight to offer a whole range of complex human diseases and biochemical phenomenons, including; Alzheimer's disease (Han *et al.*, 2011), liver disease (Gorden *et al.*, 2011), hypertension (Hu *et al.*, 2011) and schizophrenia (Orešič *et al.*, 2012) amongst others. Whilst inspiration can be taken from existing literature, large scale colorectal cancer lipidomics analysis is in its infancy and the work documented throughout the rest of this chapter is a largely exploratory attempt into the development of new avenues of more focused research.

4.2 Methods

4.2.1 Introduction to the Data

4.2.1.1 Sample Collection

From a cohort of 71 patients suffering a previously mentioned stage of colorectal cancer tissues samples were collected from resected tumours. Immediately after removal tissues were frozen in liquid nitrogen to preserve the composition of the cells and reduce the change in its volataile components. In addition to a tumour sample, a piece of histologically matched normal tissue was taken from around the tumour sample and similarly frozen. Samples were stored at -80°C until they were required for quantitification of the lipids. This sample size is the largest of any resected tissue in the last year of human clinical lipidomics publications. While some publications show large patient numbers they are typically of much easier to obtain samples, such as plasma or serum. Additionally, the studies in table 4.1 do not have the benefit of histologically matched, same patient controls, only age/gender matched controls.

4.2.1.2 Sample Preparation

Sample preparation was performed by Dr. Qifeng Zhang at The Babraham Institute. For each tissue sample, whole cells were pelleted and spiked with the following lipid sub class specific internal standards obtained from Avanti Polar lipids (http://www.avantilipids.com/).

TG 12:0/12:0/12:0 Spiked standard for Triacylglycerolipids, 100ng.

DG 12:0/12:0 Spiked standard for Diacylglycerolipids, 200ng.

MG 12:0 Spiked standard for Monoacylgylcerolipids, 100ng.

FA 17:0 Spiked standard for free Fatty Acids, 200ng.

Cer C17 Spiked standard for Ceramides, 100ng.

SG C17 Spike standard for Sphinganines, 50ng.

CL 14:0/14:0/14:0/14:0 Spiked standard for Cardiolipins, 200ng.

PG 12:0/12:0 Spiked standard for both Diacylglycerophosphoglycerols and monoalkyl, monoacylglycerophosphoglycerols, 100ng.

PE 12:0/12:0 Spiked standard for both Diacylglycerophosphoethanolamines and monoalkyl, monoacylglycerophosphoethanolamines, 200ng.

PS 12:0/12:0 Spiked standard for both Diacylglycerophosphoserines and monoalkyl, monoacylglycerophosphoserines, 200ng.

PI 17:0/20:4 Spiked standard for both Diacylglycerophosphoinositols and monoalkyl, monoacylglycerophosphoinositols, 400ng.

PA 12:0/12:0 Spiked standard for both Diacylglycerophosphates and monoalkyl, monoacylglycerophosphates, 100ng.

PC 12:0/12:0 Spiked standard for both Diacylglycerophosphocholines and monoalkyl, monoacylglycerophosphocholines, 400ng.

LPA 17:0 Spiked standard for Lysoglycerophosphates, 100ng.

LPC 17:0 Spiked standard for Lysoglycerophosphocholines, 100ng.

Cer1P 12:0 Spiked standard for Ceramide 1-phosphates, 100ng.

S1P C17 Spiked standard for Sphingoid base 1-phosphates 100ng.

SM C17 Spiked standard for Sphingomyelins 200ng.

SPC C17 Spiked standard for Sphingosylphosphorylcholines 50ng.

CE 17:0 Spiked standard for Cholesterol esters 100ng.

A modified Folch method was used to extract the lipids. This involved: first extraction with 4 mL chloroform, 2 mL methanol and 2 mL 0.88% NaCl for each sample, followed by extraction of the upper phase with 3 mL of synthetic lower phase of chloroform/methanol/0.88% NaCl 2:1:1. The combined lower phase of the lipid extract was dried with a Thermo SpeedVac at room temperature under vacuum and re-dissolved in 50 uL chloroform/methanol 1:1. Prior to LC, lipid sub classes were separated on a normal phase silica gel column (2.1x150 mm, 4 micro, MicoSolv Technology) with a hexane/ dichloromethane/ chloroform/ methanol/ acetanitrile/ water/ ethylamine solvent gradient based on the polarity of head group.

4.2.1.3 Sample Measurement

For both LC/MS and LC/MS/MS analysis, 7 uL of the extract was injected onto a Prominence HPLC (Shimadzu) column with the following instrumentation: Shimadzu IT-TOF LC/MS/MS system hyphenated with a five-channel online degasser, four-pump, column oven, and autosampler with cooler. Accurate mass (mass accuracy approximately 5 ppm) and tandem MS were used for molecular species identification and quantification. The identities of lipids were further confirmed by reference to appropriate lipids standards. The IT-TOF mass spectrometer was operated under the following conditions: ESI interface voltage +4.5 kV for positive ESI and -4 kV for negative ESI, heat block temperature 230 C, nebulising gas flow 1.4 L/min, CDL temperature 210 C, with drying gas at a pressure of 100 kPa. All the solvents used for lipid extraction and LC/MS/MS analysis were LC/MS grade and sourced from Fisher Scientific.

4.2.1.4 Lipid Quantitation

For each sample a total of 566 lipids were identified at the species level and quantified (see 3.2 for details of the structural hierarchy of these identifications). The quantitation procedure is repeated for each lipid class independently after they have been isolated by the normal phase silica gel column described in section 4.2.1.1. Over a 45 minute retention time window, spectra are acquired in the 200 m/z to 1200 m/z range and individual lipid species identified by their mass. For a small m/z window the intensity of all peaks across the entire retention time window are aggregated to produce a unique chromatogram for each individual lipid species. The peak of each chromatogram is isolated and the peak area integrated as a surrogate for direct quantitation. The integrated peak areas of each lipid species are compared to that of the internal standard and a measure of relative quantitation achieved. Because the spiked in sample is of known quantity, as described in section 4.2.1.2, the absolute quantitation of a lipid species is achieved by multiplying its ratio with the internal standard by the mass of the spiked in internal standard. This entire procedure is performed by LCMSsolution Version 3 software from Shimadzu and is outlined in figure 4.1.

The quantified data was exported to a custom excel file format and provided as such for statistical analysis. The measurement of the 566 lipids in each sample is disproportionately larger than data from the other publications in table 4.1 and gives a much more accurate view of the patient's lipidome and the interplay between lipids species.

4.2.2 Data standardisation

4.2.2.1 Data Storage

Exploratory data analysis requires a dynamic way to access and manipulate data in order to get a view appropriate for a particular analysis. For example, investigating the data from a Duke's stage perspective is different to that of an analysis focused solely on a particular patient. For this reason the static spread sheets provided are not optimal and so a MySQL database was used instead to house the results with a schema appropriate to the entities common to all the data: Stage, patient, lipid sub class and lipid species. The schema was designed using the MySQL workbench software suite version 5.1.19 and is shown in figure 4.2.

A Java application using the Apache POI-XSSF library was developed to parse the spreadsheet format file by file (each workbook representing a Duke's stage, each worksheet representing a lipid sub class) and populate the database with the quantitative measurement records associated with each patient and lipid species pair. Once completed the database provided a flexible platform to access the data and generate complex views of the data, mixing and matching the required entities as necessary. Additional to the quantitative data, a metadata file was supplied that

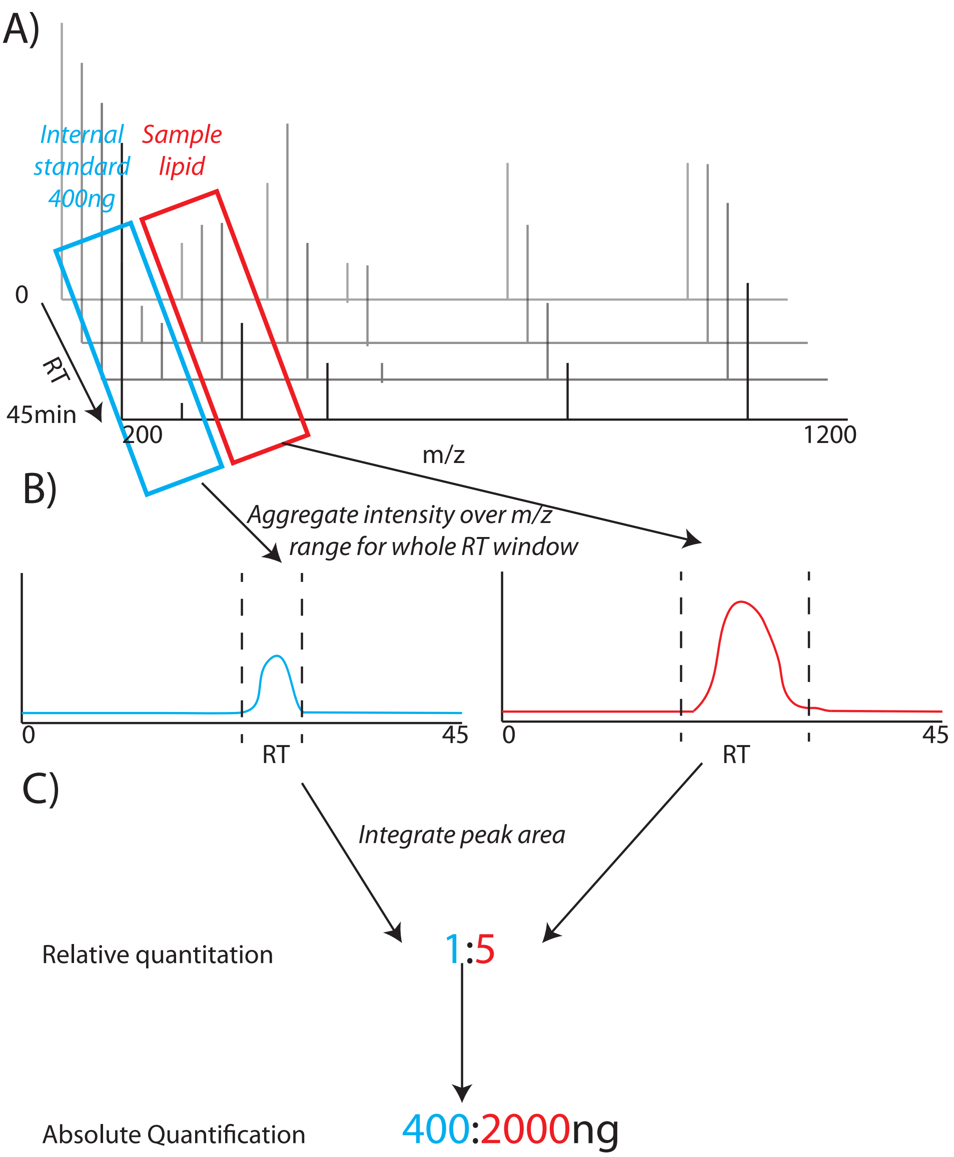


Figure 4.1: **A)** Over the 45 minute retention time window, spectra are acquired in the 200 m/z to 1200 m/z range and individual lipid species identified by their mass. **B)** For a small m/z window the intensity of all peaks across the entire retention time window are aggregated to produce a unique chromatogram for each individual lipid species. **C)** The peak of each chromatogram is isolated and the peak area integrated as a surrogate for direct quantitation. The integrated peak areas of each lipid species are compared to that of the internal standard and a measure of relative quantitation achieved. Because the spiked in sample is of known quantitythe absolute quantitation of a lipid species is acheived by multiplying its ratio with the internal standard by the mass of the spiked in internal standard.

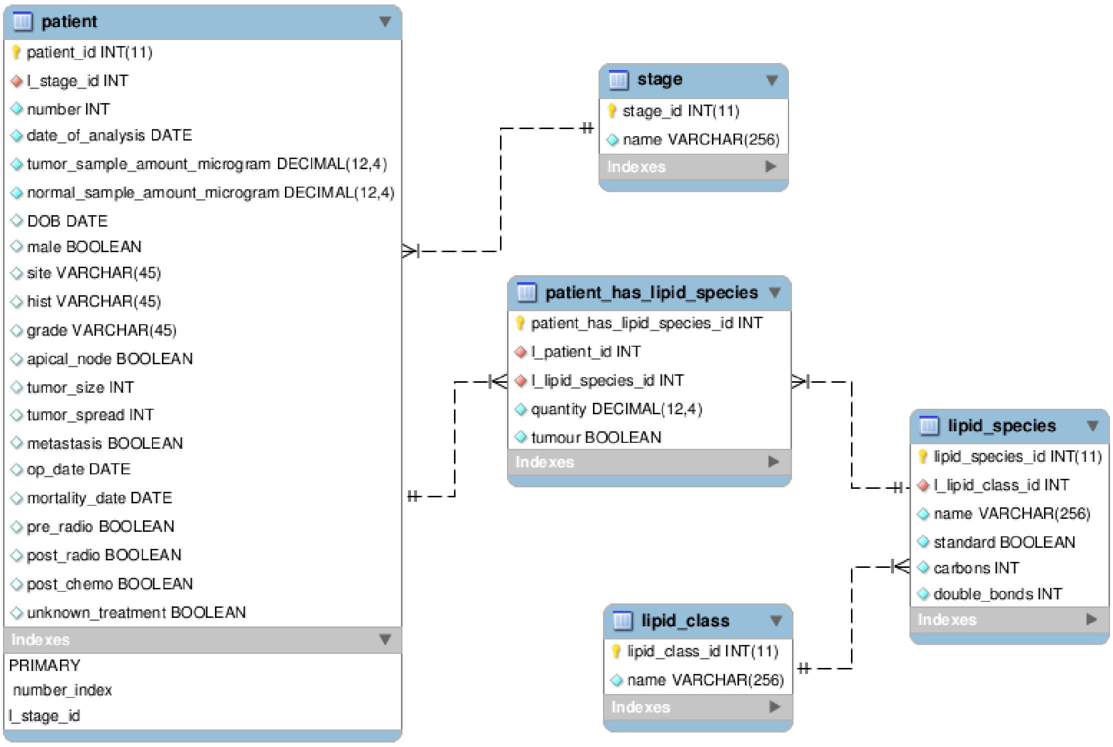


Figure 4.2: A graphical representation of the colorectal cancer lipidomics database schema.

detailed a number of patient specific features such as age, gender, mortality date, tumour size and tumour spread. This data was also incorporated into the patient table of the database and a selection of it is available in Appendix 2.

4.2.2.2 Nomenclature standardisation

Highlighted in the previous chapter, data from different laboratories is often described by a slightly different nomenclature. This dataset was no exception and before data analysis could be performed it was necessary to transform the non-standard lipid names into the community supported LipidomicNet nomenclature. This was necessary for publication and using existing tools to perform analysis. Integration with LipidHome was also carried out in order to use its literature search functionality for gathering information on interesting lipids identified during statistical analysis. An R script was developed to update the lipid names in the database with generic functionality for transforming Babraham Institute lipid nomenclature names into the more standard LipidomicNet nomenclature. The code is available for download from http://code.google.com/p/jmf-thesis/source/browse/trunk/SupplementaryFiles/LipidomicsLib.zip. While it is not yet common practice to utilise data standards in the field of lipidomics, a conscious effort was made here to do so to help promote work done in chapter 3.

4.2.2.3 R library

Much of the analysis requires the reuse of similar data structures. For example; a patient by species matrix of quantities for all normal samples, or a vector of patient metadata such as their stage, tumour size and gender. To reduce analysis development time, a reusable R library with dataset specific processing and access methods was developed. Future sections will regularly refer to usage of the library and its data processing methods. A commented version of the library is available for download from http://code.google.com/p/jmf-thesis/source/browse/trunk/SupplementaryFiles/LipidomicsLib.zip.

4.2.2.4 Quality control

The quantified data provided has already undergone several rounds of quality control prior to storage in the database during the initial experimental setup. As described in the sample collection and storage section, careful control has been kept over sample handling conditions and reagents used in the extraction and separation of lipids prior to mass spectrometry. For the purpose of quantification, spiked-in synthetic lipids of known concentration act as markers against which the natural lipids can be measured independently for each sample. With such a large proportion of the workflow completed prior to receiving the data there was little opportunity to advise upon the experimental design. For example, no technical replicates were supplied to estimate the reproducibility and sensitivity of the instruments. While biological replicates are provided as samples from the same Duke's stage, the inherent variability between such biological replicates, especially in such a heterogeneous source material as cancer in which a multitude of cellular perturbations can be responsible for the detected lipidome, make them less useful. Quality control on this dataset was integrated within individual analyses as outliers were detected and experimental design issues identified. Prior to analysis, quality control was performed upon the metadata to ensure that the sample annotations were consistent and sufficient for the explanation of trends and patterns within the data. From a first glance at the breakdown of tumour sample stages there was an uneven distribution with particularly few in the border stages (B/C and C/D) and adenoma, see figure 4.3. Uneven and particularly sample-sparse Duke's stages can complicate data analysis, especially machine learning classification approaches that would otherwise be well suited to these data.

The adenoma patient group offers a particularly interesting dataset because these samples are not strictly cancerous (not appearing on the Duke's staging classification system) and unlikely to have systematic effects on the body or surrounding tissue. They represent the closest thing to a technical replicate in the dataset, but are only four in number. Discussion with the data provider resulted in a further sample being measured to boost the statistical power of the stage in later analyses. An additional Duke's stage A sample was analysed and included in the dataset, bringing the total patient number to 71. In addition to Duke's stage, additional metadata was

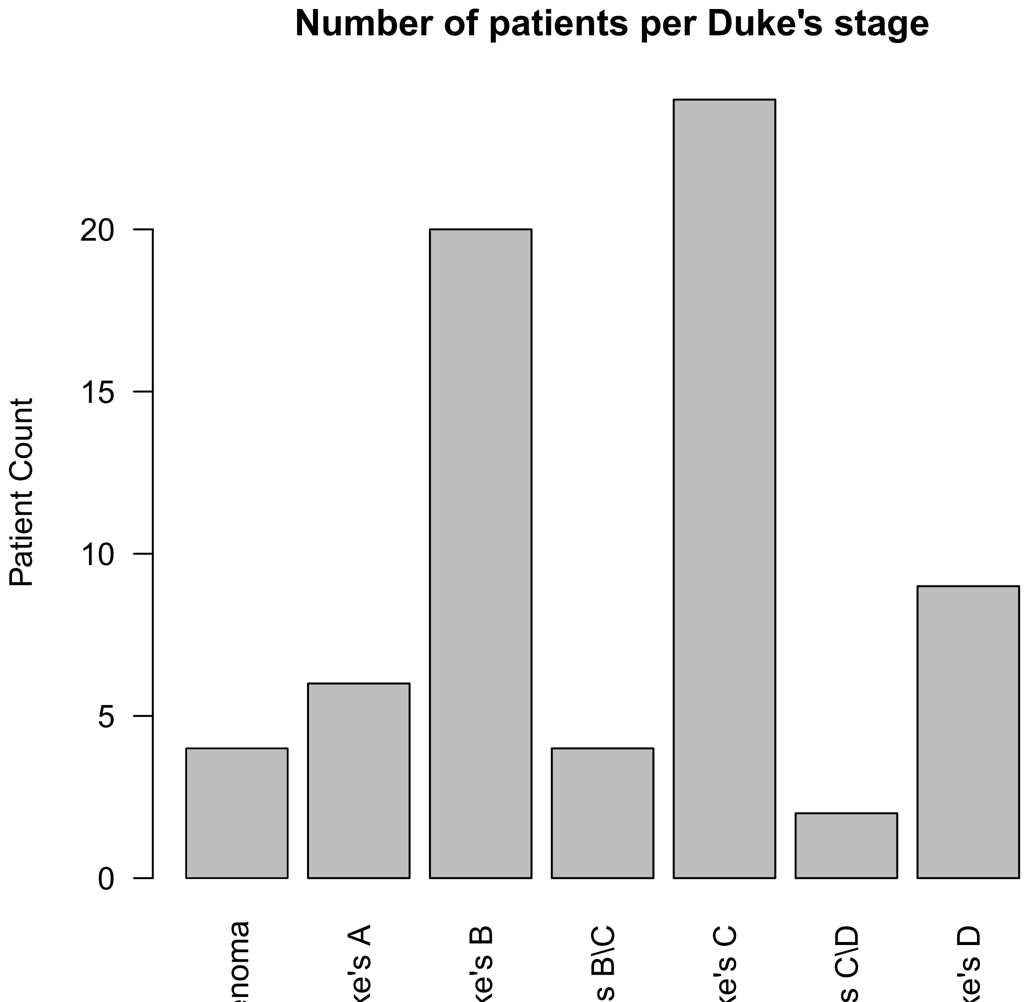


Figure 4.3: Distribution of Duke's stages in tumour samples of the 69 patients.

provided upon request by the Babraham Institute after onerous work to gather it directly from the archived paper records of each patient. In parallel to obtaining the patient records, border stage samples were re-examined under the microscope and their Duke's classification stage confirmed into a single category. The final Dukes stage sample distribution can be seen in figure 4.4.

4.2.2.5 Pathway development

Lipid species can be investigated in isolation, comparing the difference between treatment conditions. However, extra insight into their relationship with each other and the biology that governs their inter-conversion can be studied by grouping lipids together by their chemistry. A pathway was constructed from a number of known reactions that convert one lipid sub class to another. The majority of reactions are common knowledge within the field and as such are available in undergraduate texts such as Biochemistry 5th Edition Stryer *et al.* (2002). The network was also supplemented with other reactions from the literature; 'PC+Cer→SM+DAG' (Ooira *et al.*, 1974) and 'PC→LPC' (Adimoolam *et al.*, 1998). Each lipid sub class that was measured in the dataset with well documented reactions was added to the sketched network diagram as a node. Nodes were connected by directional reactions. The diagram was recreated using CellDesigner version 4.2 (Systems Biology Institute (SBI), Tokyo, Japan). Once the diagram was finished the software was used to reduce the network to a list of all the individual reactions. Additionally, start points and end points were defined in the network. The possible reactions that created a non-redundant path from a starting point to an end point were computed and these sets of reactions recorded. Patient data could then be overlaid onto the network and comparisons be made between them. A concept used throughout the reaction based analysis was the reaction ratio; after data normalisation the ratio of products to reactants (P:R) can be calculated. Reaction ratios can either be in favour of products (> 1) or in favour of reactants (< 1). Due to the design of the study; each patient having a normal and a tumour sample, P:R could be compared between tumour and normal samples to detect any significant shift in equilibrium. The magnitude of the shift towards products or the shift towards reactants does not necessarily change which is favoured but just reduces the ratio. For example

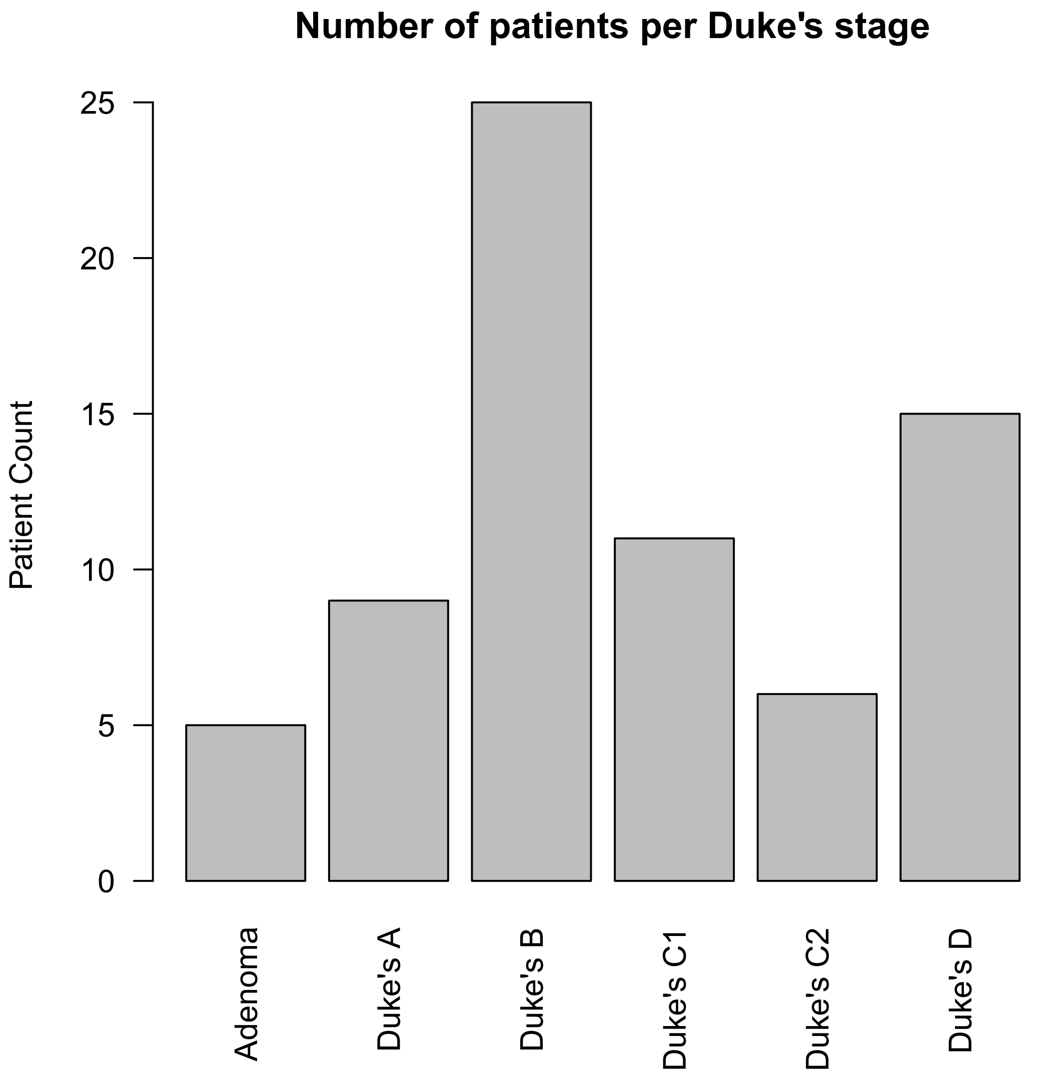


Figure 4.4: Updated distribution of Duke's stages in tumour samples of the 71 patients after re-classification of border stage samples and addition of adenoma samples.

the reaction 'PC→PE' may be in a 3:1 P:R ratio in a normal sample and 2:1 in a tumour sample, the product is favoured in both cases. However in the tumour sample there has been a shift towards reactants and a less pronounced favouritism for products. These shifts in equilibrium can be used to infer nodes of the network that are seeing considerable build up or depletion of mass. These can further be used to infer potential enzymatic reactions that may be the cause and hence relevant protein/gene targets.

4.2.3 Statistics implemented

Throughout the analysis of this dataset a number of common statistics and data transformation techniques were utilised, below each method is described followed by a brief explanation upon interpretation of the result.

4.2.3.1 Shapiro-Wilk test

Tests the null hypothesis that a sample originates from a normally distributed population. It has the formula:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Performed in R using the shapiro.test() function of the base package, it provides a p-value in its return object's $p.value attribute. The p-value can be interpreted as significant at the 5% level if below 0.05, hence the null hypothesis rejected and the sample concluded to be from a non-normally distributed population. This test is typically used to determine the suitability of non-parametric tests such as the Wilcoxon signed-rank test for subsequent use on the data.

4.2.3.2 Paired Student's t-test

A sample of matched pairs have their means compared to determine if they come from the same population distribution, the null hypothesis being that they do. The difference of the pairs D must be calculated for use in the formula:

$t = \frac{\bar{X}_D - \mu_0}{s_D/\sqrt{n}}$

Performed in R using the t.test() function of the base package, with the parameter "paired = TRUE", it returns a p-value in its return object's $p.value attribute. The p-value can be interpreted as significant at the 5% level if below 0.05, hence the null hypothesis rejected and the two samples assumed to originate from different population distributions. This test dictates that the data must be normally distributed and the two ditributions under test share equal variance.

4.2.3.3 Wilcoxon signed-rank test

Similar in application to the paired Student's t-test, the Wilcoxon signed-rank test is its non parametric equivalent for use when either one or both of the paired samples is proven to be sampled from a non normal population distribution. After calculating the paired differences, the zero differences are removed, followed by ordering and ranking of the absolute remaining non-zero differences. These differences are then partitioned into positive and negative and the ranks summed. The minimum of these two sums is compared to the critical value and the null hypothesis rejected if it is less than or equal to the critical value. Conveniently provided by R as the wilcox.test() function with the parameter "paired = TRUE", p-values are available in the return object's $p.value attribute.

4.2.3.4 Mann-Whitney U test

This test is the unpaired equivalent of the Wilcoxon signed-rank test described in section 4.2.3.3. Similarly, it tests whether two samples have differences in their distributions. Its calculation requires the ranking of all values in both samples. The sum of each samples rank is denoted as R_1 and R_2, the total number of observations in each sample is denoted as n_1 and n_2 respectively. U is calculated for each sample as follows:

$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$

$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$

The smaller value of U is compared to the critical values provided by a significance table from which p-values can also be calculated. The test is provided by R in the base package as the wilcox.test() function, with the a parameter for each sample and “PAIRED = FALSE”.

4.2.3.5 Bonferroni correction

When performing the same statistical test multiple times and assessing the results simultaneously the multiple comparisons problem occurs. This problem results in the null hypothesis being rejected at a higher rate than appropriate because the total number of tests has not been accounted for. The liberal rejection of the null hypothesis can be counteracted by applying some method of multiple testing correction. Bonferroni correction is one such method and is well known for its reduction of false positives at the expense of increased false negatives. P-values are compared to α (the cut-off for significance e.g. 0.05) divided by n (the number of tests). If the p value remains smaller than $\frac{\alpha}{n}$ then the null hypothesis is rejected.

4.2.3.6 Normalisation

In data analysis it is typical to divide data by a common variable or set of variables in order to remove the variable’s effect upon the data, allowing multiple datasets to be compared. In the case of the colorectal cancer data, each lipid measurement was divided by the sum of its patient’s lipid measurements. This effectively normalised the large differences in total lipid content between patients and gave a natural lipid level independent view of the data. Normalisation against sum was the method of choice here as it scaled all values between zero and one. On this particular dataset the normalised values represented the percentage contribution a lipid species made to the entire lipidome or its lipid sub class. For example, a patient sample made up of the following sub class totals 5ng/g PC, 10 ng/g PA, and 15ng/g TG is normalised to 16.67% PC, 33.33% PA and 50% TG.

4.2.3.7 Log ratio

When calculating the up or down regulation of two lipid measurements, dividing the two produces poorly scaled and asymmetric data that is difficult to interpret and plot on a graph. To remove this problem and relate negative changes or large fold changes, the log ratio was taken. This scales both positive and negative changes equally and given the large differences seen in the data, the log10 ratio was most often employed.

4.2.3.8 Pearson product-moment correlation coefficient

Used to measure the linear dependence upon two variables, Pearson correlation was regularly used throughout this analysis to relate two lipids, patients or samples to each other and determine if they were similar. It has the formula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Performed in R using the cor() function with the parameter “method = ’pearson’ ”, it returns the correlation coefficient r. Where $-1 \leq R \leq 1$, 1 indicating a strong positive correlation, -1 indicating a strong negative correlation and 0 indicating no correlation.

4.2.3.9 Data resampling

When performing a large number of correlations (see section 4.2.3.8) it is important to estimate what level of correlation constitutes a significant result, i.e. two patients are more correlated to each other than they are with all other patients on average. When estimating the significance of correlation coefficients it is simply not sufficient to use an arbitrary cut-off such as ‘0.8’. To estimate a reasonable significance cut-off for correlation, data can be sampled many thousands of times to produce chimeric samples. These theoretical samples are made up of parts of many real samples. When resampling data it is important to appreciate that the samples created must fairly sample the data and maintain any important structure of the data. The large number of theoretical samples can be correlated against each other to form a correlation matrix. The mean of the matrix is then the expected correlation of two randomly selected samples. Flattening the matrix into a correlation distribution and inspecting the tails (typically the 2.5 and 97.5 percentiles) can give a conservative estimate of the correlation coefficients that constitute a significant correlation at the 5% level. The critical values of correlation are applied to the original data and significant positive and negative correlations then remain to be further investigated.

In the case of resampling patients to produce theoretical patients, only a single measurement for each lipid species was sampled, rather than performing the sampling entirely by chance which would lead to the most abundant lipid being sampled more than once, creating a non-representative patient. Once an arbitrarily large number of theoretical patients had been fairly sampled, all pair-wise correlations were calculated, producing a large number of correlations that followed a certain distribution. This distribution, specifically its 2.5 and 97.5 percentiles, was used to estimate the cut-off for significant correlation at the 5% level.

4.2.3.10 Fisher's exact test

Used to analyse contingency tables, Fisher's exact test is used for testing the significance of association between two pieces of categorical data. For example, lipids are partitioned into significant and non-significant categories on the one hand, and of the sub class TG and not of the sub class TG on the other hand. This creates a 2x2 contingency table upon which Fisher's exact test can be calculated, in order to judge if the TG category is significantly overrepresented in its number of significant lipids in comparison to significant lipids of other lipid sub classes. It has the following formula for calculating the probability of obtaining such a result where a, b, c and d are found in the quadrants of the 2x2 contingency table and n their sum:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

Fisher's exact test can be performed using the R function fisher.test() in the base package. The tests has been criticised as being overly conservative, not actually testing significane at the desired 5% level, but it some cases testing it at the largest value smaller that 5% that is possible with the data. For larger data sets this is less problematic and so the test has been used in this analysis in favour of a more complicated but robust test such as Barnard's exact test.

4.2.3.11 Near zero variance feature removal

When using a large number of features for machine learning, particularly classification, it is common practice to begin with a round of feature selection. The purpose of this is to remove features which provide little information, or redundant information. A simple method for feature selection is near zero variance feature removal, wherein features that do not change very much and thus will contribute very little signal to classifying samples are removed. The exact criteria for removal are feature dependent and should not be pre-estimated. The function nearZeroVar() is provided by the 'caret' package in R. It requires the parameter 'freqCut'; a value equal to the ratio of the frequency of the most common value to the second most common.

4.2.3.12 Highly correlated feature removal

Highly correlated features provide very little unique signal when used for training models to classify data by machine learning techniques. To reduce the computation time of the classifier and simplify the results, one of a pair of a pair of highly correlated features is removed from the dataset. Estimating the cut-off that defines high correlation can be achieved through a resampling method described in section 4.2.3.9. The findCorrelation() function is provided by the 'caret' package in R. It requires the parameter 'cut-off' equal to an appropriate correlation threshold above which features will be considered similar enough for one to be randomly selected and removed from the feature set. Choosing the value for 'cut-off' can be achieved by plotting the correlation distribution between features and identifying differences from an expected normal distribution.

4.2.3.13 10-fold cross validation

In 10-fold cross validation the data is divided into ten subsets of approximately equal size. The sets are then used to train a classifier, each time using nine out of ten sets for training and a single set for validation. This process is repeated ten times, so that each dataset is used nine times for training but only once for validation. The results of the classification are compared and any discrepancy in the models is accounted for to alter its predictive accuracy. Using the 'caret' package in R the trainControl() function assigns the validation protocol. In this case the parameters ‘method’ and ‘number’ were “cv” and “10” respectively.

4.2.3.14 Random Forest

Based upon averaging the votes of multiple decision trees, Random Forest is an ensemble classifier capable of predicting categorical labels to numerical data. It is recognised as one of the most accurate machine learning techniques and is favourable over many others due to its less intensive computational requirements. The algorithm works by building a large number of decision trees (500 in this case). Each decision tree is built by taking a small number of random features and for each feature finding the boundaries that best separate the training data into the correct categorical classes. These decision boundaries are constructed into a tree with as many levels as the number of features selected. When hundreds of trees have been constructed from different feature sets and training samples, test data can be applied to each tree to predict its category. Each tree ‘votes’ which category a sample is in. The average of the votes is taken to be the classifiers final classification with some estimate of certainty. Wrapped by the ‘caret’ libraries train() function in R with the ‘method’ parameter set to “rf”, a call to the ‘RandomForest’ library’s Random Forest implementation is made.

4.2.3.15 Cohen’s Kappa statistic

Cohen’s Kappa statistic quantifies the agreement between a number of classifiers, categorising data into any number of mutually exclusive groups. It estimates the agreement between classifiers, taking into account the random chance of agreement and so is more conservative than simply calculating the percentage agreement of classifiers. It has the formula:

$\kappa = \frac{Pr(a)-Pr(e)}{1-Pr(e)}$

Where $Pr(a)$ is the observed probability of agreement between classifiers and $Pr(e)$ is the probability of agreement by chance. Available in R under the ‘concord’ package using the function cohen.kappa().

4.2.3.16 Brown-Forsythe test

The Brown-Forsythe test is a robust statistic to test the null hypothesis that the variance of two samples is the same. It is the alternative to the F-test when normality of data cannot be confirmed; it is also related to Levene's test substituting the use of mean for median. It has the following formula where N is the number of observations, n_j is the number of observations in group j, p is the number of groups and $z_{ij} = |y_{ij} - \tilde{y}_j|$ where $\tilde{y}_j$ is the median of group j:

$$F = \frac{(N-p)}{(p-1)} \frac{\sum_{j=1}^{p} n_j (z_{.j} - z_{..})^2}{\sum_{j=1}^{p} \sum_{i=1}^{n_j} (z_{ij} - z_{.j})^2}$$

The function is available in R as levene.test() from the 'lawstat' package with the parameter "location=median", the result object provides a p-value in the $p.value attribute. If the p-value is below 0.05, then the null hypothesis can be rejected and the variance between the two samples be considered significantly different at the 5% level.

4.2.3.17 Fisher transformation

In order to compare Pearson correlation coefficients, values must first be transformed using the Fisher transformation. The resulting transformed variable is normally distributed and the use of location tests much more appropriate to this distribution. Given that r is the correlation coefficient the transformation formula is as follows:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

Once all values in a correlation distribution are transformed they can be fairly compared by the appropriate location test e.g. Student's t-test (see section 4.2.3.2).

4.2.3.18 Hierarchical clustering

A correlation or distance matrix is clustered to group together similar vectors of data. As an example from this chapter, similar patients were clustered together based upon their lipidome composition. R provides agglomerative hierarchical clustering via the hclust() method, where clusters are built bottom-up, from each patient being an individual cluster, to all patients sharing the same cluster. There are a few technical parameters to the approach: distance metric and linkage criterion. After a round of trial and error, Pearson correlation (section 4.2.3.8) in combination with complete linkage criteria provided acceptable clustering results. Clusters are then identified by hand and the members of those clusters are interrogated for explanatory metadata or similar lipidome composition.

4.3 Results

4.3.1 Sample analysis

The broadest level discriminant between two samples is their sample type: Either normal or tumour. From the experimental design the reader is reminded that for each patient, both a normal and a tumour sample in close proximity were resected from the colon and that lipids were identified and quantified in both.

4.3.1.1 Total lipid change

The total lipid concentration for each patient was calculated for each sample, by summing each individual lipid concentration. The data was split into two sets, one set of the sum of normal sample lipid concentrations, the other the sum of tumour sample lipid concentrations. A Shapiro-Wilk test for normality was performed to estimate the appropriateness of a t-test to measure the difference between the two sets (see section 4.2.3.1). The Shapiro-Wilk test concluded both normal and tumour samples were non-normally distributed with p-values of 7.954e-07 and 1.308e-08, hence a paired Wilcoxon test was used instead of the t-test (see section 4.2.3.3). The paired Wilcoxon test concluded that there was a significant increase in the total concentration of lipids in the tumour sample with a p-value of 0.002534, see figure 4.5.

4.3.1.2 Lipid sub class change

The lipid species were organised into lipid sub classes analogous to those previously described in the LipidHome structural hierarchy (see figure 3.2). A table of lipid sub classes, their codes and full names is available in appendix 2. Using the same methodology as described previously, the data was stratified into the 25 sub classes,

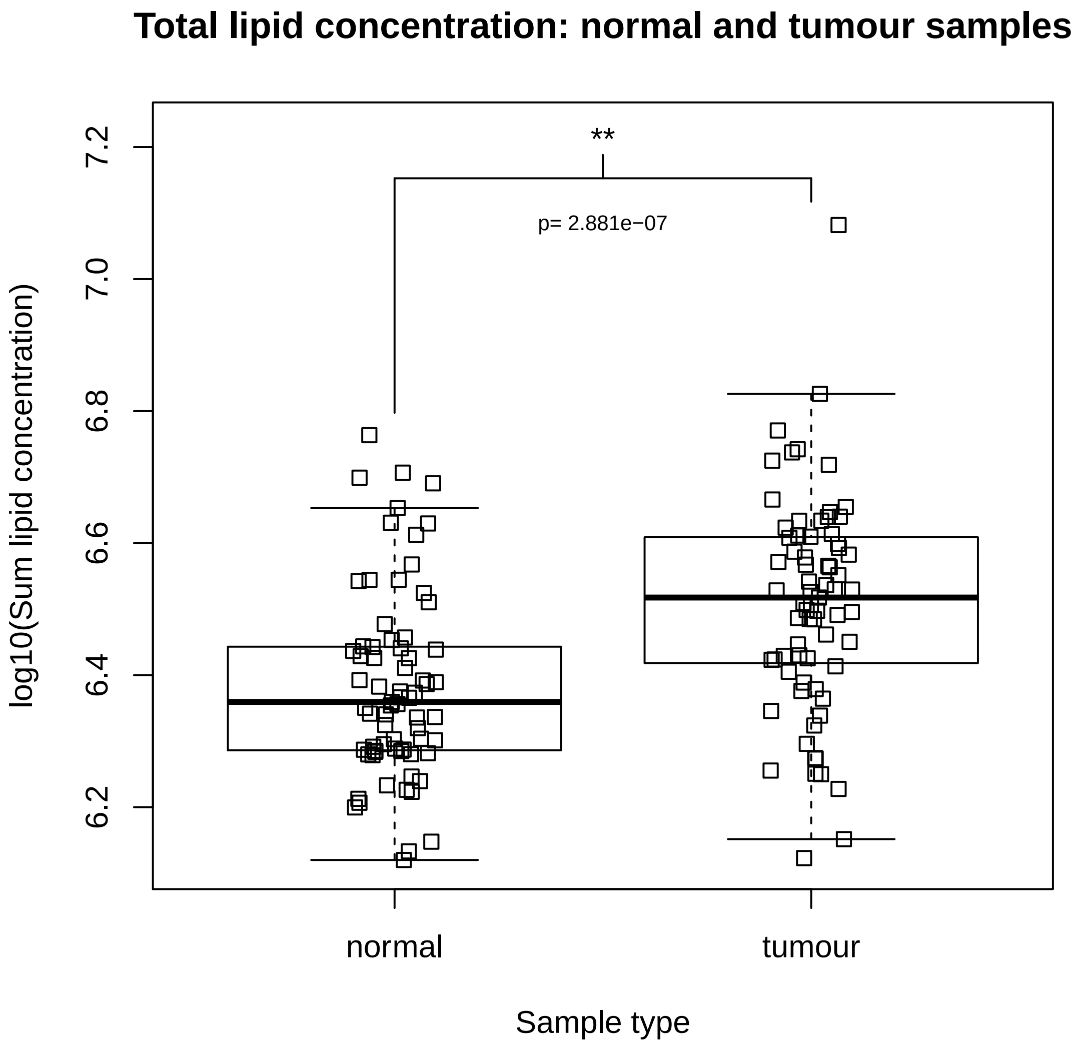


Figure 4.5: Based upon a paired Wilcoxon test the total lipid concentration is seen to be significantly higher in tumour samples than normal samples. Significance denoted by "*","**" and "***" at the 5%, 1%, and 0.1% levels respectively.

each analysed individually. Analysing the change in lipid concentration at the sub class level revealed individual sub classes which differed in concentration between normal and tumour samples. Of the 25 sub classes, 21 showed a statistically significant difference at the 5% level, only the sub classes LPC, PA a , PA and SM showed no changes. In all but three of the significant cases, the sum concentration of a sub class was higher in tumour than normal. The three exceptions to this trend were the LPA, MG and TG sub classes, see figure 4.6.

While global lipid concentration rose between normal and tumour and the majority of sub classes followed suit, the data were normalised and re-analysed to estimate the specific perturbation of the lipidome each sub class incurred as a percentage of the total lipidome. Normalising the sub class data by total sum (as described in section 4.2.3.6), gave the percentage each lipid sub class contributes to the total lipidome. Analysing this percentage difference using the previously described methodology allowed for the detection of sub classes which increased in concentration from normal to tumour but decreased in their percentage concentration in the lipidome. Statistical analyses of the normalised data revealed only 8 lipid sub classes for which there was a difference at the 5% level in the percentage of the lipidome that a sub class represented between normal and tumour samples. The data is summarised in figure 4.7, it shows a significant decrease in the percentage contribution to the total lipidome of the sub classes PA, LPA, LPC, TG and SM represented in tumour samples compared to normal. Conversely, there was a significant increase of the sub classes PE, LPC a and Cer above the general increase of lipids as a whole in tumour compared to normal samples.

4.3.1.3 Lipid species change

Reusing the previous methodology for finding statistically significant differences between sub classes of paired tumour and normal samples, the same was done for individual lipid species absolute data. This involved the comparison of all 547 lipid species (the 19 synthetic quantification standards were of course removed). The list of quantified lipids identified in this study is available in Appendix 2. Due to the large number of comparisons, a multiple testing correction strategy was applied to the resulting p-values to reduce the number of false positives and focus on

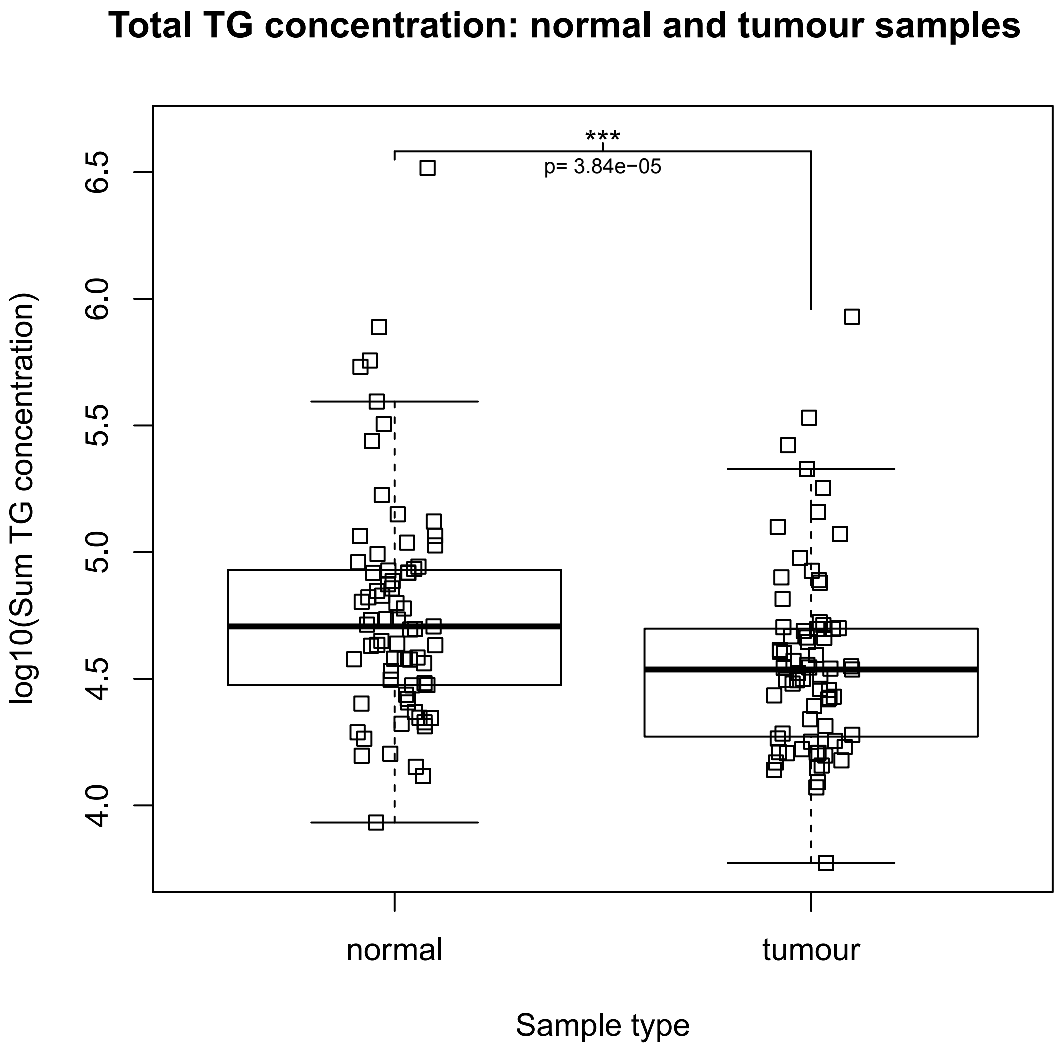
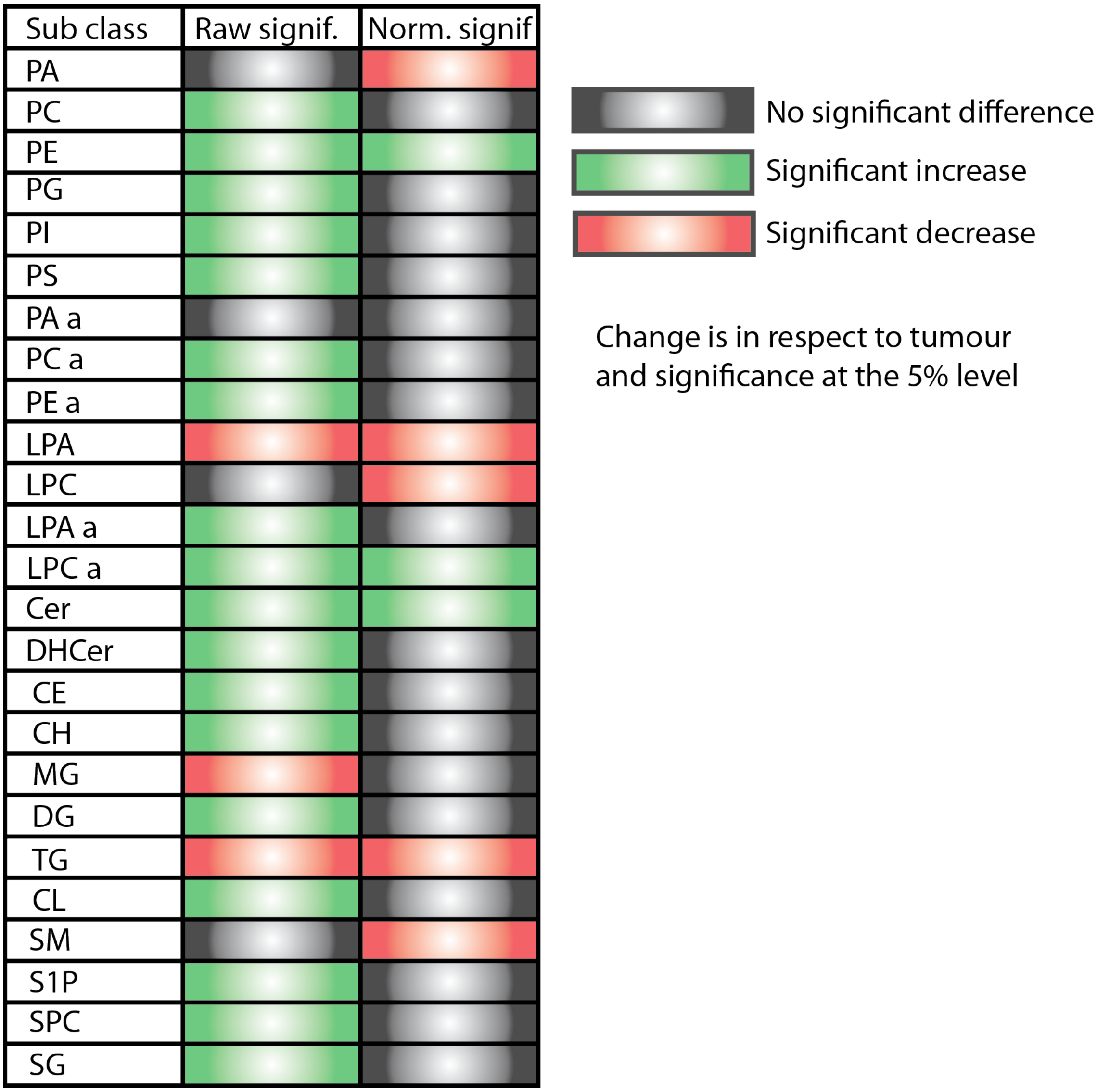


Figure 4.6: Based upon a paired Wilcoxon test the triacylglycerolipid concentration is seen to be significantly lower in tumour samples than normal samples. This contradicts the general trend of the other 18 lipid sub classes with a significant difference in normal and tumour that mimics the total lipid concentration increase from normal to tumour seen in figure 4.5. Significance denoted by "*","**" and "***" at the 5%, 1%, and 0.1% levels respectively.



Sub class	Raw signif.	Norm. signif
PA	⬛	🟥
PC	🟩	⬛
PE	🟩	🟩
PG	🟩	⬛
PI	🟩	⬛
PS	🟩	⬛
PA a	⬛	⬛
PC a	🟩	⬛
PE a	🟩	⬛
LPA	🟥	🟥
LPC	⬛	🟥
LPA a	🟩	⬛
LPC a	🟩	🟩
Cer	🟩	🟩
DHCer	🟩	⬛
CE	🟩	⬛
CH	🟩	⬛
MG	🟥	⬛
DG	🟩	⬛
TG	🟥	🟥
CL	🟩	⬛
SM	⬛	🟥
S1P	🟩	⬛
SPC	🟩	⬛
SG	🟩	⬛

⬛ No significant difference

🟩 Significant increase

🟥 Significant decrease

Change is in respect to tumour
and significance at the 5% level

Figure 4.7: A summary of the changes in sub class absolute lipid concentration and normalised (% of lipidome) differences between 71 tumour and normal samples. Green represents a significant increase in tumour samples at the 5% level; red represents a significant decrease in tumour samples at the 5% level.

the most interesting differences. A number of multiple testing correction methods were tested in a pilot phase, including: Bonferroni, Holm, Hommel and Benjamini-Hochberg FDR with Bonferroni producing the most conservative results. Prior to multiple testing correction, 423 lipids were shown to have differences between normal and tumour samples at the 5% level using a Wilcoxon signed-rank test. Due to the large amount of significance Bonferroni correction was accepted as the correction method of choice (see section 4.2.3.5). After multiple testing correction, only 273 remained significant at the 5% level. In comparison to microarray experiments upon which the analysis and data transformation I have described is based, a percentage of significant hits around 50% is extremely high. Microarray analysis shares the same problem of identifying many hundreds, sometimes thousands, of biomolecules (mRNA). However, microarrays typically report a significant hit percentage in the 1% region depending on the specificity of the array. Of the 273 lipids that were significantly different between tumour and normal samples, 97 were decreased in tumour and the remaining 176 increased.

Following on, the same was applied to the normalised dataset to detect lipids that significantly changed in their percentage lipid composition between tumour and normal samples. Prior to multiple testing correction 361 lipids were found to be significantly different between normal and tumour samples at the 5% level, a reduction from the absolute data due to the normalisation by the overall increase in lipids seen in figure 4.5. After Bonferroni correction, 161 lipid species were found to be significantly different between tumour and normal samples. The intersection of the significant lipids from the absolute data and the normalised data was 150/160. Of particular interest to the data collaborators was the PS (diacylglycerophosphoserine) sub class of lipids. The lipid species in this sub class were analysed in isolation to identify the cause of the significant increase from normal to tumour of total PS, highlighted in section 4.3.1.2. Using the same Wilcoxon signed rank test, followed by Bonferroni correction method detailed previously, PS species were ordered by mean log10 tumour:normal ratio and plotted as individual boxplots, visible in figure 4.8. There was a clear trend of extremely desaturated lipid species showing upregulation in tumour and a slightly weaker but significant tendency for the more saturated PS lipid species to be downregulated in tumour samples.

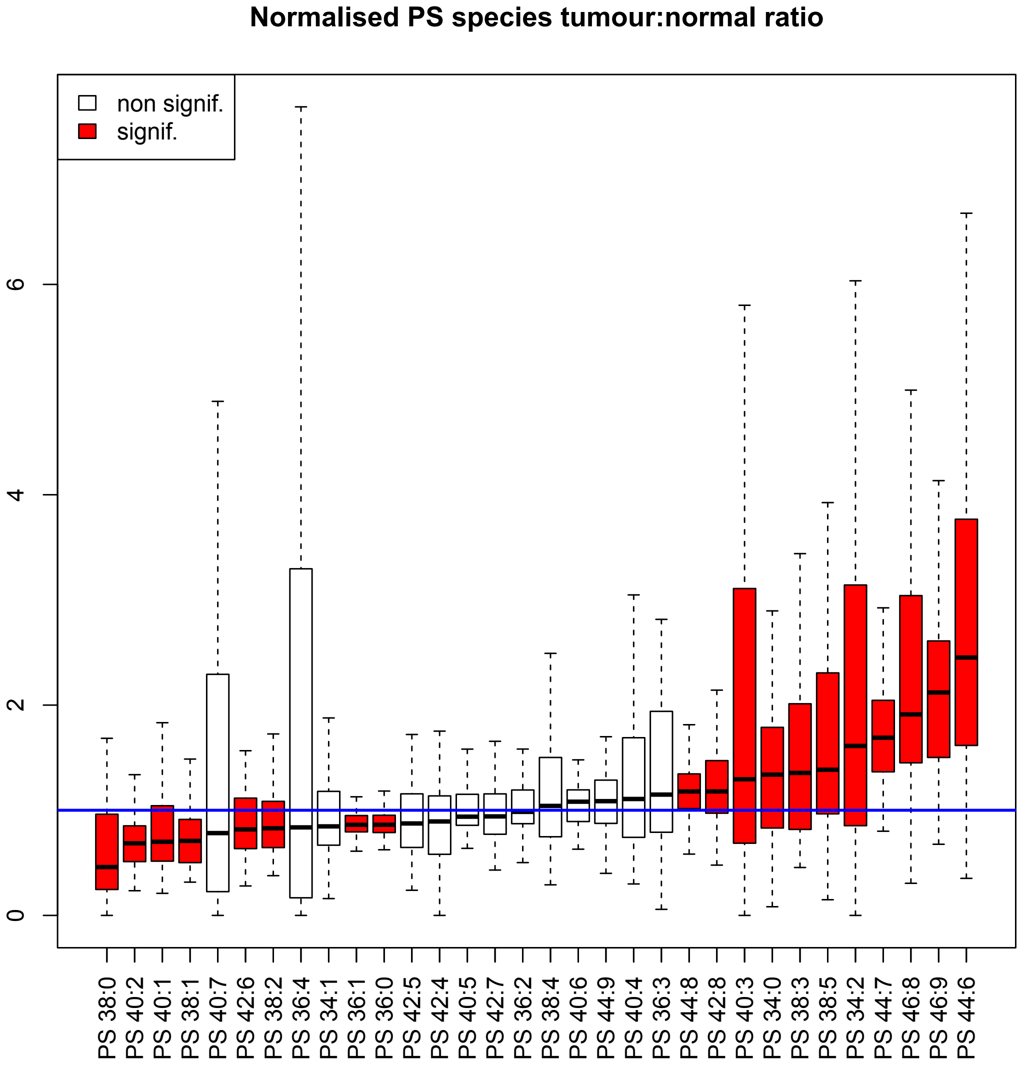


Figure 4.8: The log10(tumour:normal) ratio of all PS lipid species are ordered by the mean log10 ratio (outliers included). This shows a clear trend of highly desaturated PS species being upregulated and highly saturated PS species being downregulated in tumour samples.

4.3.1.4 Lipid feature overrepresentation

To identify any relationship between the significant lipid species, the statistics that underlie the Gene Ontology (GO) enrichment analysis, typically applied to genomics and proteomics datasets, were transferred to this case. A Fisher exact test (see section 4.2.3.10) was employed to determine if the number of lipids with a particular feature (e.g. fatty acid structure 36:0) are overrepresented in the list of significant lipids in comparison to the total number of lipid species that share that feature. For example ten lipid species share the 36:0 fatty acid structural feature, eight of which were found to be significantly different between normal and tumour samples. Of the remaining 200 lipids which did not have the 36:0 fatty acid structure feature, 50 were found to be significantly different between tumour and normal samples. The Fisher exact test is used to test if the proportion of significant lipids sharing the fatty acid feature 36:0 is different from the significant proportion of the 200 other lipids that do not share the 36:0 feature. If the proportion is significantly larger, the 36:0 feature is determined overrepresented amongst the lipids that are significantly different between normal and tumour samples. If the proportion is significantly lower, the 36:0 feature is determined as underrepresented amongst the lipids that are significantly different between normal and tumour samples. Similarly to the previous example, lipid sub classes can be tested for disproportionate representation in the set of significant lipids. For example, is the ratio of significant lipids of the PC subclass different to the ratio of significant lipids for all other sub classes combined.

Using the absolute data and testing each sub class if it has a disproportionate number of significant lipids, the sub classes PC, PE and TG were shown to be significantly overrepresented amongst the significant lipids, with odds ratios of 2.31, 7.50 and 2.69 respectively. This is interpretted as the sub classes PC, PE and TG are 2.31, 7.50 and 2.69 times more likely to contain lipid species significantly different between normal and tumour samples. Taking the same approach for the fatty acid structure gave no significant results at the 5% level, although the fatty acid structure '38:3', '40:6' and '42:8' did come very close to significance. As discussed in section 4.2.3.10 the Fisher exact test can be rather conservative on determining significance with small sample sizes and so it was deemed reasonable to report this result. An analysis was devised to look in slightly more detail, at the significantly different lipids between normal and tumour samples. The significant lipids were segregated into 'increased in tumour' and 'decreased in tumour'. Performing the Fisher exact tests again on these segregated sets identified if particular lipid sub classes have a disproportionate number of significantly upregulated lipids, or a disproportionate number of significantly downregulated lipids. The upregulated set revealed the over-representation of PE species that significantly increased in tumour samples, whereas TG was identified as overrepresented in the significantly downregulated lipids. That left PC that was neither disproportionate in the upregulated or downregulated sets, but was disproportionate (overrepresented) in the significant lipids over all. This hinted at a polarising effect of the tumour where some PC species were upregulated and some down.

The identical analysis was performed upon the normalised data to find sub classes and lipid structural features that show a disproportionate number of lipids that differ significantly between tumour and normal samples. The sub classes TG and PI were found to be overrepresented amongst the lipids that significantly changed between tumour and normal samples. PI is a particularly interesting sub class due to its involvement with cell signalling. The fatty acid structures '38:0', '44:6', '44:7', '46:7' and '46:8' were overrepresented in the lipids that significantly changed between tumour and normal samples. This was an interesting result because the majority shared a relatively high number of carbons and level of unsaturation. Lipids were further segregated into those that increase and those that decrease on average in percentage composition of the lipidome in tumour samples. Analysing only the lipids that increase, the Fisher exact test was applied as before. The lipid sub classes TG and SM were significantly overrepresented, hinting that while TG was previously seen to decrease overall from normal to tumour (see figure 4.6), of those that increase, a large number of them were significant. In the case of the downregulated lipids PE, PI and TG were found to be overrepresented amongst significant lipids. It is not impossible for TG to be overrepresented in both upregulated lipid species and downregulated lipid species. This result highlighted the very interesting possibility that different TG species were being regulated independently or the effects in one group were somehow altering the levels of another. To better understand the special case of TG, the lipid species significantly different between tumour and normal are displayed in figure 4.9 as a series of boxplots ordered by mean log10(tumour:normal) ratio.

Interpretation of figure 4.9 revealed that the Fisher's exact test results were correct. There were two groups of TG species that significantly differed between normal and tumour; those that increased and those that decreased. The blue vertical line indicates the boundary between those that were significantly downregulated in tumour (left of blue line) and those that were significantly upregulated in tumour (right of blue line). Looking at the lipid species at each side of the boundary clearly revealed that the lipid species that decreased were primarily the shorter chain TGs while those that increased were predominantly the longer chain TGs.

4.3.1.5 Sample classification

Analysis in the previous sections highlighted some biologically interesting perturbations to the lipidome of tumour colorectal cancer cells. As an extension of this work a machine learning approach was used to build a binary classifier to test if an accurate model could be built to predict the origin of unknown colorectal cancer samples. Using all 547 normalised lipid species as initial features for training, features underwent a process of selection. Firstly, removal of near zero variance features (see section 4.2.3.11), where the cut-off for the ratio of the most common value to the second value was 95/5. This was followed by a round of removing any feature that was not significantly different between normal and tumour at the 5% level after Bonferroni correction of Student's t-test results. Finally, highly correlated features were removed by first plotting the pair-wise correlation distribution of all features (see section 4.2.3.12). In this case, no clear correlation cut-off could be estimated from the distribution and so a conservative 0.95 was used. The three feature selection steps left 186 features useful for classification. Patients were split into training and test data in a 3:1 ratio. Using a 10-fold cross validation (see section 4.2.3.13), a set of Random Forest classifiers were trained (see section 4.2.3.14). To estimate the optimal starting parameter '.mtry' (the number of features randomly sampled to build the various levels of the decision tree) a grid expansion approach was used, testing multiple starting values over a specified domain. Models were evaluated based upon their accuracy and Cohen's Kappa statistic (see section 4.2.3.15) with secondary

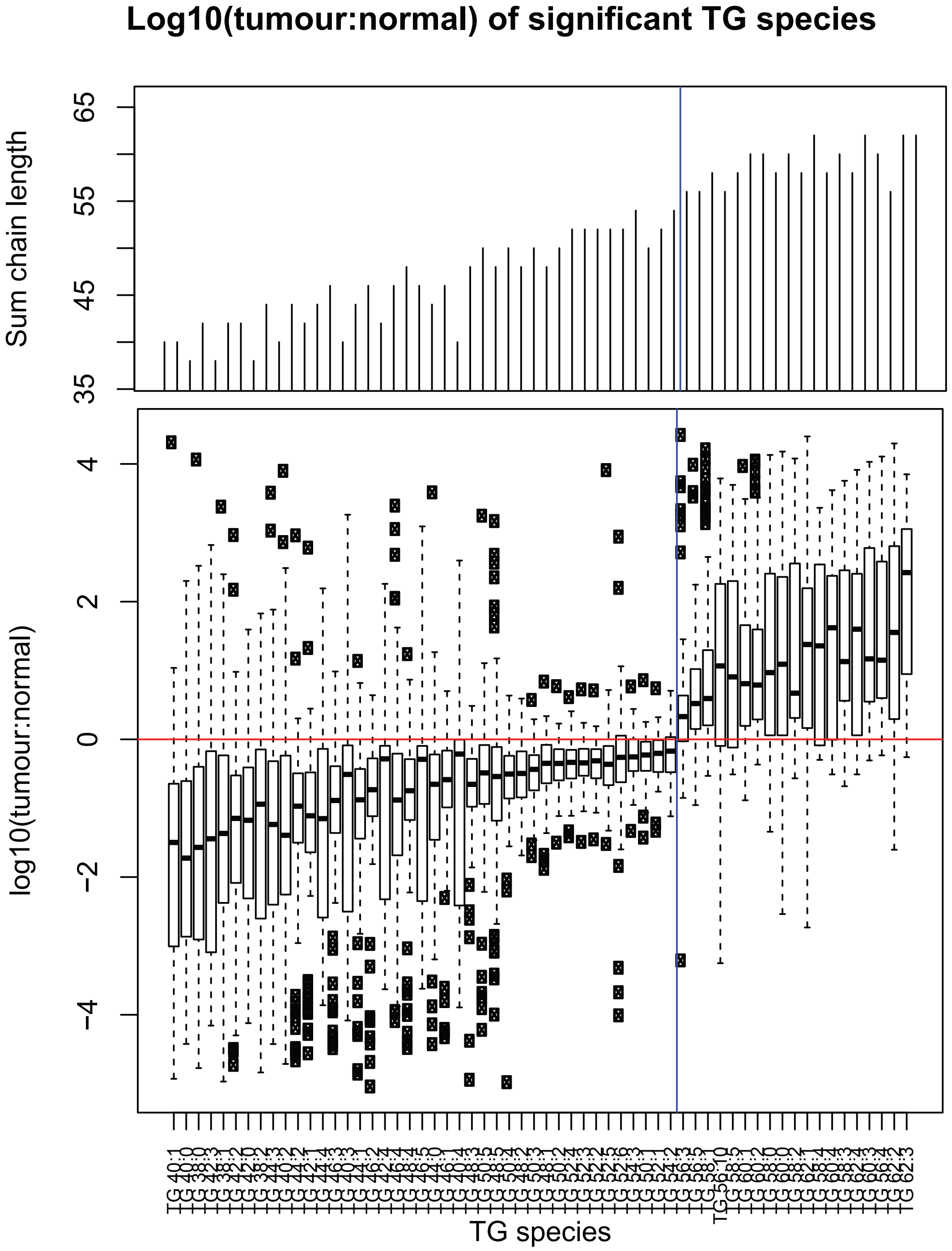


Figure 4.9: The log10(tumour:normal) ratio of the TG species that are significantly different between tumour and normal samples at the 5% level after Bonferroni correction. Lipid species are ordered by the mean log10 ratio (outliers included), and a clear trend is visible of short chain TGs being downregulated and high chain length TGs being upregulated in tumour samples.

consideration for minimising '.mtry' in order to have the most optimal model that required the least computation. In the case of these data '.mtry' was found to be optimal at a value of 10 and the resulting classifier was accurate 99.1% of the time in correctly distinguishing between a normal and a tumour sample. Classification accuracy in this region is extremely rare and was a strong signal that colorectal cancer is represented by changes to the composition of the lipidome. Looking deeper at the approximate 1% misclassification rate revealed the systematic misclassification of the tumour sample of patient '1383'. Upon consultation of the metadata of this patient's tumour, it was shown to be staged as adenoma. Mentioned previously in section 4.1.3, adenoma is a benign tumour that does not pose a serious health risk unless it becomes too large and an obstruction to blood flow or transport of faeces through the bowel. By comparison, carcinoma is a much more serious condition where the genome is considerably altered and the tumour develops malignant properties that if left untreated would be fatal. Indeed, adenoma should provide a quite distant cell state from true colorectal cancer and so misclassification was somewhat unsurprising. However, the more regular occurrence of correct classification of adenoma samples as tumour raised many questions into when exactly changes in the lipidome occur between normal and tumour cells and could some adenoma samples be sub classified into those likely to develop further into Duke's stage A and those that would not?

In order to extend this work and provide robust statistics on the model, 100 Random Forest models were trained using the same methodology as described above, the only difference being the random partitioning of samples into training and test set. Once all 100 models were trained, the Out-Of-Bag accuracy (the mean accuracy of the 10 "leave one" out samples in the 10 fold cross validation) and the test data accuracy were plotted as boxplots in figure 4.10. The original model was not overtrained, as a similar accuracy was reported for the 100 other models. In addition, the test sets that had no part in training the models were accurately predicted on average 100% of the time. To see the internal performance of the models, a single model that performed the highest in terms of OOB accuracy was selected and the individual votes for each training sample counted. Figure 4.11 shows how the percentage of the 500 trees in the model voted for each sample. Clearly some samples are much easier to classify as normal or tumour than others but the ensemble of votes from all the trees constitutes an incredibly accurate binary classification model. A visualisation of the raw chromatograms of a selection of patient tumour and normal samples is avialble in Appendix 2 figures 6, 7, 8 and 9. There is a clear upregulation of the PC 28:0 species in the tumour sample of these patients in comparison to the normal samples.

A Random Forest approach is not only useful in building a predictive model, but the relevance of training features can be extracted to give a semi-quantitative impression of which lipid species contribute to distinguishing tumour and normal samples. Figure 4.12 displays the normalised importance of the top 50 most important features from the 100 Random Forest models. A large number of the most important features were from the TG sub class. Noteworthy also, were the short chain PC species PC 28:0,PC 30:0 and PC 30:1, along with the long chain LPC species 24:0, 24:1, 26:0 and 26:1.

4.3.2 Patient centric analysis

In parallel to work aimed at detecting differences between tumour and normal samples in the lipidome, a patient centric slant on the analysis can also be taken to detect similarities between patients based on the underlying lipidome. This type of analyses may help to corroborate known metadata or identify groups of patients for which remarkable similarity (or dissimilarity) is shown within lipid samples.

4.3.2.1 Patient similarity

To get a feel for the data, lipid species from the same sub class and patients within the same Duke's stage were plotted. Lipid species were ordered by ascending number of carbons (sub ordered by ascending number of double bonds). Figure 4.13 shows the absolute lipid quantities for the PC sub class for normal samples of adenoma patients. Whilst absolute amount of lipids were different between patients, the distribution of lipids within the sub class was very similar. The same plot for the TG species of the Duke's B stage patients normal samples in figure 4.14 showed a surprising outlier in patient 90 (light blue solid line). This patient had orders of magnitude higher TG levels. This result gave strong support for using the normalisation by sum method previously described in section 4.2.3.6 in order to make

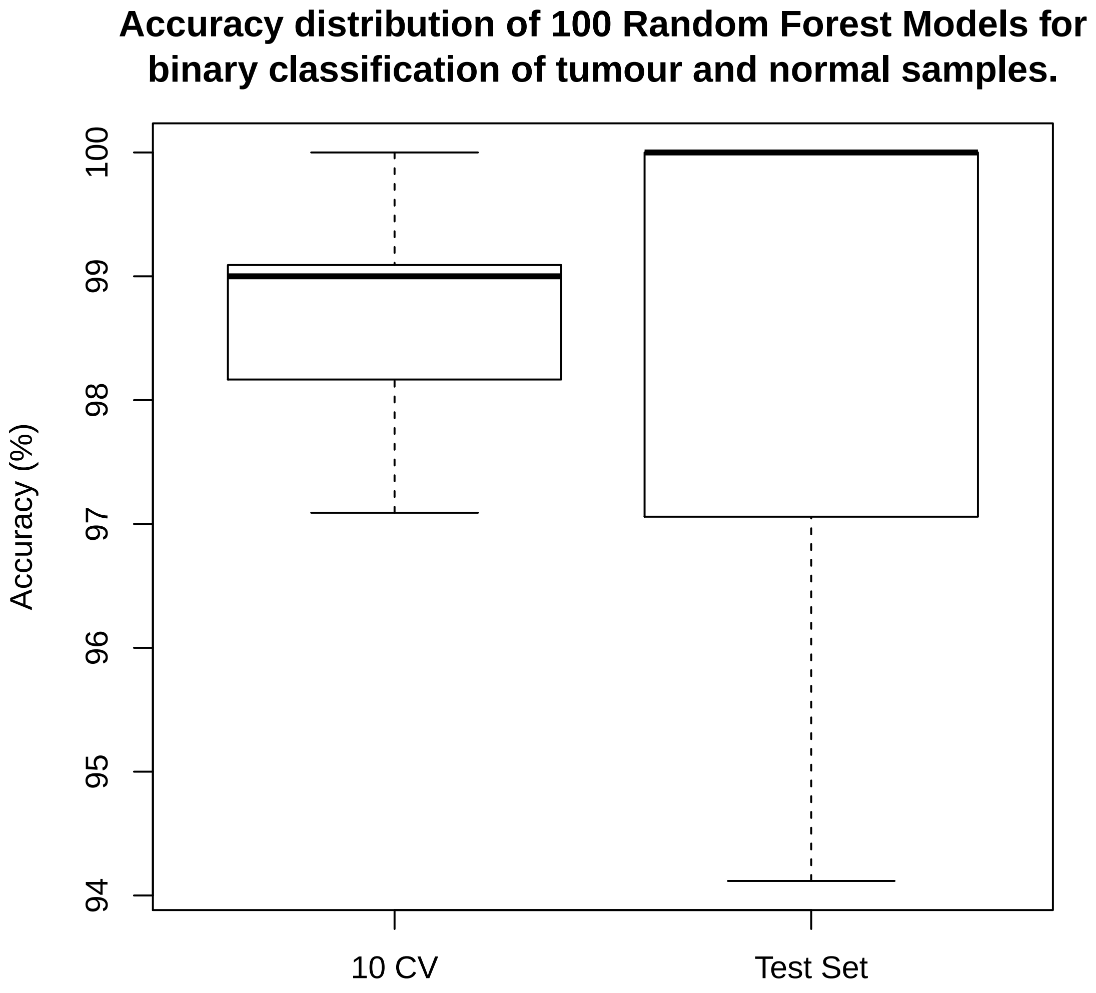


Figure 4.10: The mean accuracy of 100 Random Forest models classifying the difference between tumour and normal samples from a filtered set of 547 normalised lipid species. Samples were partitioned into 3:1 training:test sets and a 10 fold Cross Validation (10 CV) performed. The process was repeated 100 times with different partitions. For each model the Out-Of-Bag (OOB) accuracy is calculated and displayed as a box plot (left). Additionally, the test set accuracy for each of the 100 models is plotted as box plot (right). The mean accuracy is 99.1% and 100% for the OOB set and the test set respectively.

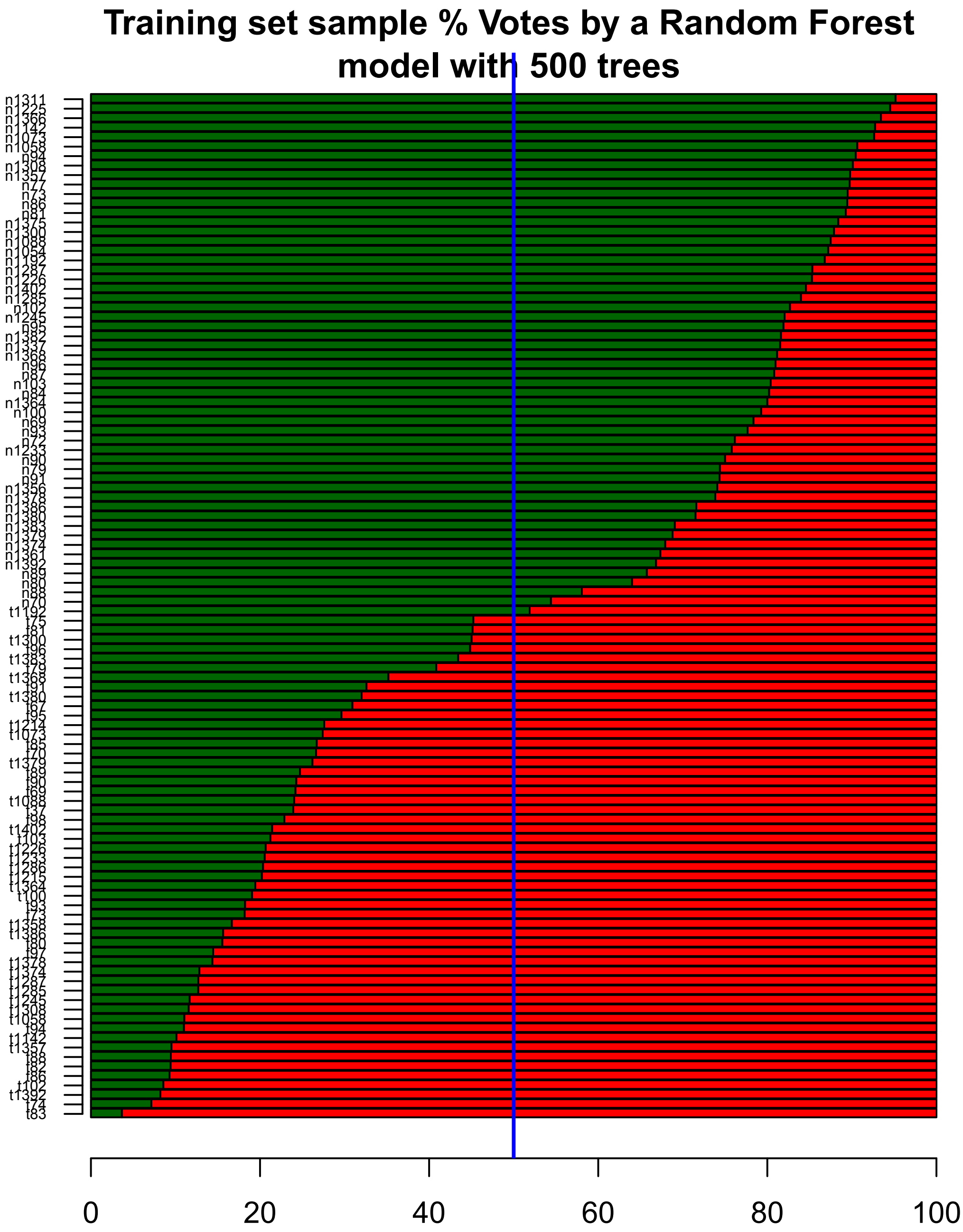


Figure 4.11: For the single Random Forest model with the highest OOB accuracy, the percentage of classification votes for each training sample from the 500 trees is plotted. Green indicates percentage of votes for normal classification, whilst red represents tumour classification. It is clear that not all samples are as easy to classify as others, but the result is still 100% classification accuracy.

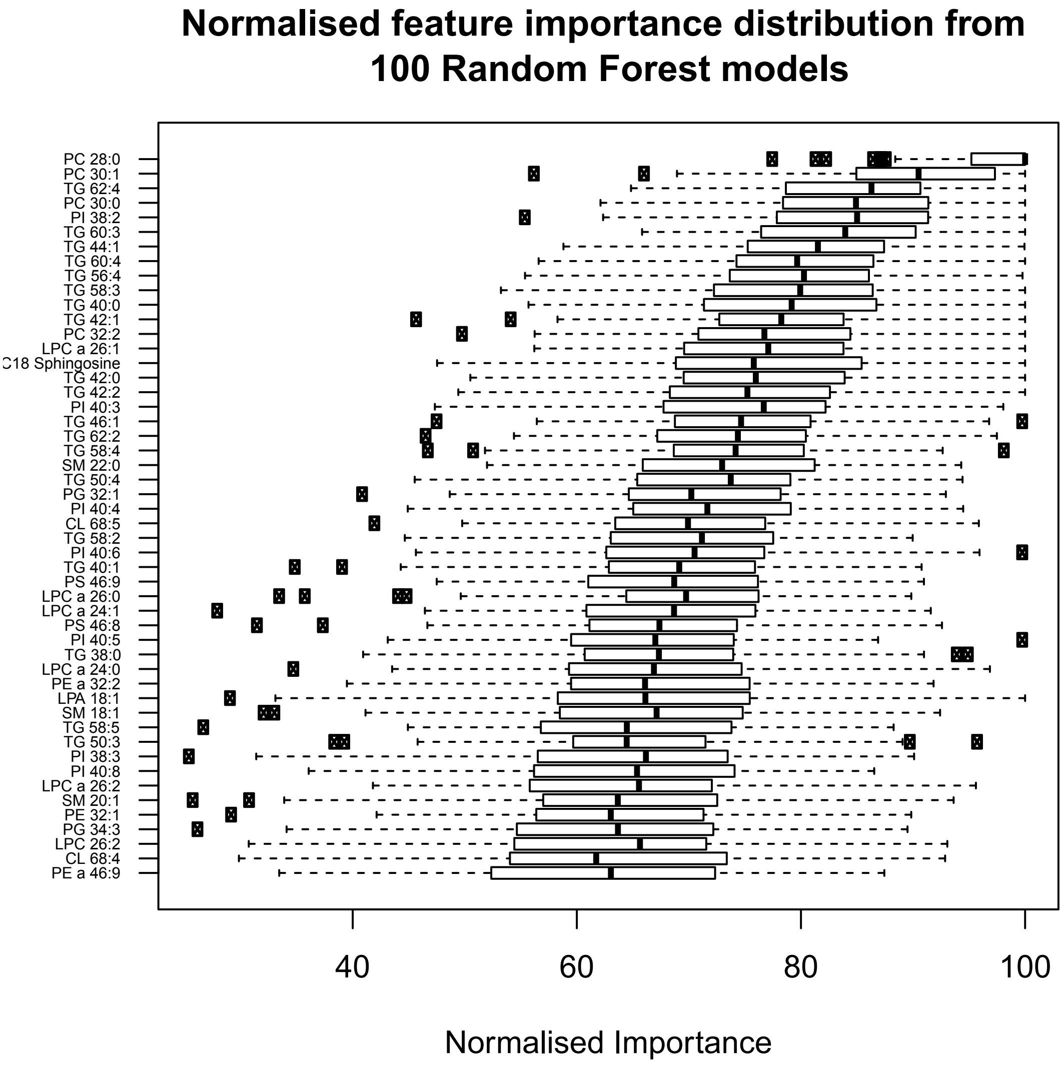


Figure 4.12: The normalised importance distribution of the top 50 most important features in differentiating between tumour and normal samples from 100 Random Forest Models trained on different data partitions. The short chain PC species and the TG species are particularly important at distinguishing the difference between tumour and normal samples.

patient samples comparable against one another. The right hand panel of figure 4.14 clearly shows that post normalisation the patients were much more comparable in their lipid species profiles and that the normalisation procedure was suitable to enable later analyses.

The chromatograms for these five patients are available in Appendix 2 and show the data prior to quantitation. The first track is that of the internal standard. The PC a species and LysoPC species are also included as they are separated with the PC species and analysed together by LC/MS.

4.3.2.2 Patient clustering

Extending the approach in section 4.3.2.1, patient tumour samples can be compared in a statistical manner using Pearson product-moment correlation coefficients (see section 4.2.3.8) in order to identify which patients are similar to each other. Lipid species were split into their sub classes and 25 patient lipid species matrices were constructed (one for each lipid sub class). The patient vectors were then correlated all against all to produce 25 correlation matrices. These matrices and the correlation values in them represented how similar two patients were in the percentage contribution of each lipid species to its lipid sub class. Before conclusions could be drawn about which patients were correlated with each other, a cut-off was estimated to define the level of correlation deemed significant. This was done fairly by resampling the data to generate theoretical hybrid patient vectors, which contained a single measurement for each species in a lipid sub class from a random patient. One thousand hybrid theoretical patient vectors were generated and correlated against each other, and the resulting 499,000 correlations (half the matrix minus the self-correlations) formed a correlation distribution that was cut at the 2.5 and 97.5 percentiles to give significance cut-offs at the 5% level (see section 4.2.3.9). Applying the two cut-offs to the real patient correlation matrix (97.5 for significant positive similarity, 2.5 for significant dissimilarity) a matrix that held the discrete values 1,-1 and 0 was constructed where 1 was significant similarity, -1 significant dissimilarity, and 0 an insignificant correlation. These matrices were clustered using Ward's hierarchical method (section 4.2.3.18) and plotted for each lipid sub class.

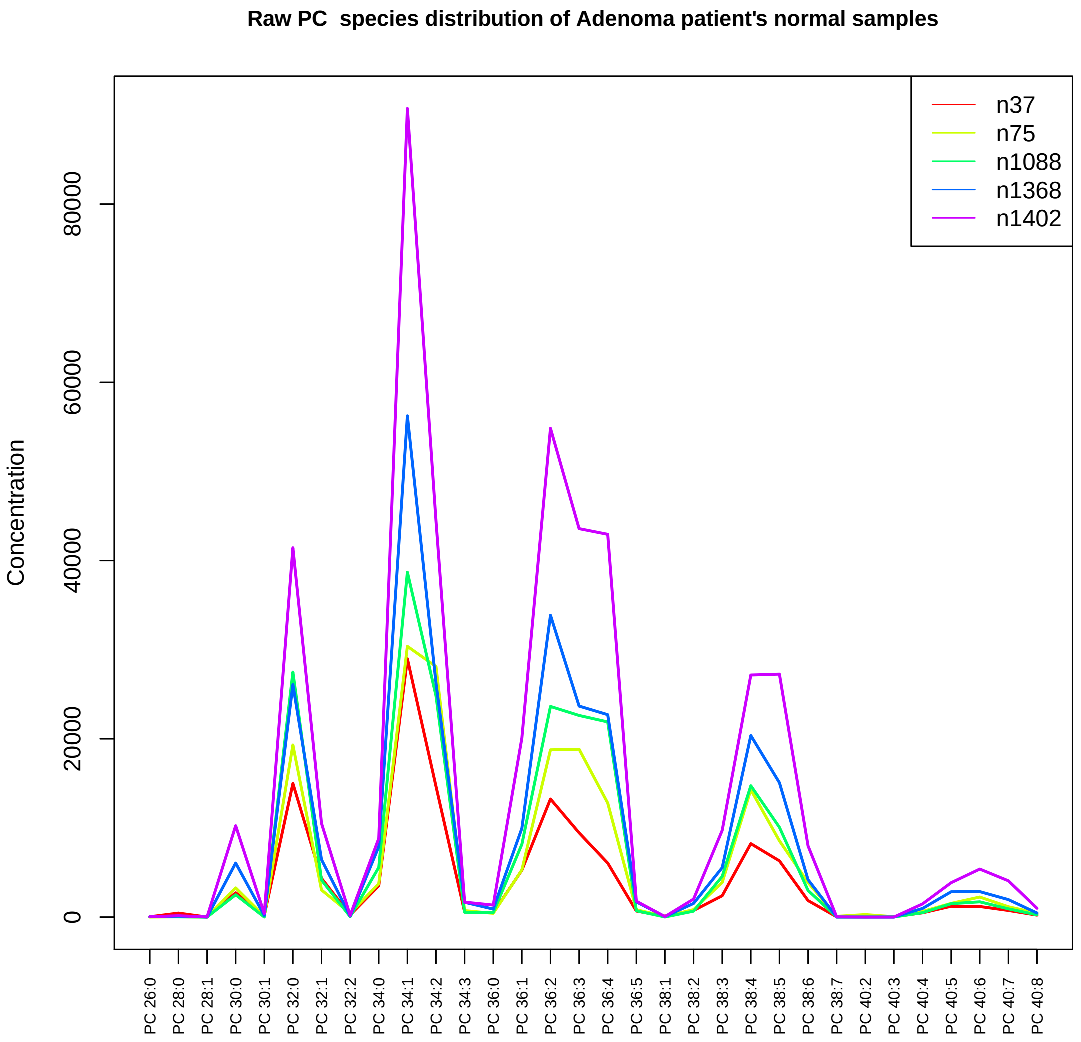


Figure 4.13: Adenoma staged patients's normal sample absolute PC species distributions shows the general trend of almost all lipid sub classes and Duke's stages that patients share a similar general distribution of lipid species. While the absolute concentration of lipid species differs, the overall percentage contribution to the sub class specific lipidome is very similar.

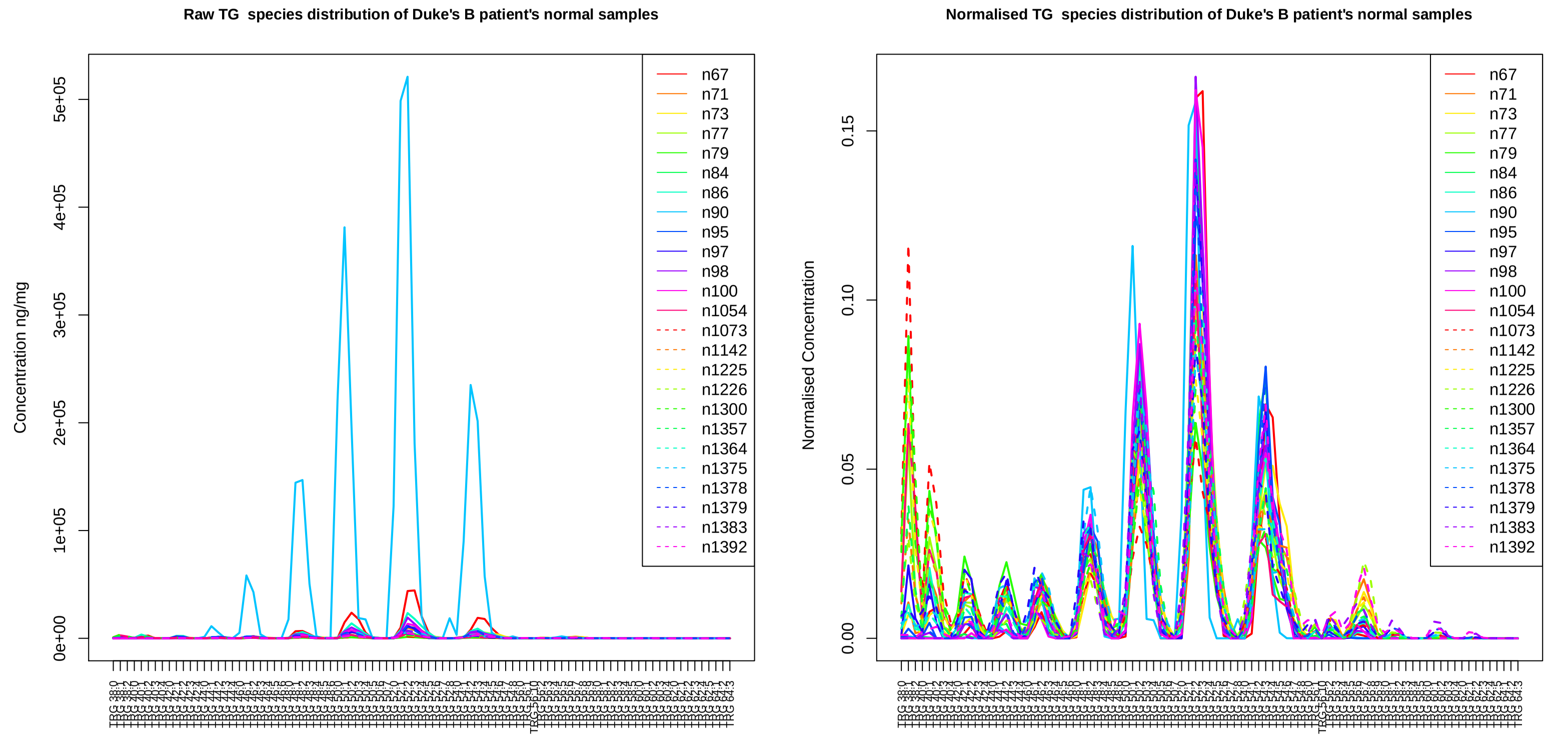


Figure 4.14: Similar to figure 4.13 this shows the absolute (left) and normalised (right) concentration of lipid species in the TG lipid sub class of the normal samples of patients categorised as Duke's B. Patient 90 (light blue solid line) has a much higher absolute TG concentration than the others, but after normalisation shows a much more consistent pattern of lipid expression. This is a key example of the importance and suitability of the normalisation procedure outlined in section 4.2.3.6.

Clusters were located by eye and the patients that they contained analysed for overrepresentation of metadata terms using Fisher's exact test (see section 4.2.3.10). For each metadata type, e.g. gender, tumour size and tumour spread, each individual level of that type is tested for over and underrepresentation. Clusters found to be over or underrepresented by an explanatory metadata term were investigated in more detail to find the exact composition of the sub class that seemed to define the cluster. To do this, the mean sub class lipidome was calculated and plotted for the samples within the identified cluster. For reference, the mean sub class composition of the background (all other patients) or the opposing of two dissimilar clusters was calculated and plotted on the same axis.

Figure 4.15 shows the heatmap of tumour sample similarity for the ceramide sub class. Only significant similarity is shown, where green represents significant similarity and blue represents significant dissimilarity. The cluster highlighted in the red box shows a significant overrepresentation of the 'tumour size: 1' metadata type at the 5% level. The levels of the tumour size metadata type are 1, 2, 3, 4 and NA. The ceramide lipidome of this cluster of similar patients (numbers 1042, 71, 1374, 72, 1366, 1380 and 1245, see Appendix 2 table 6) was explained by small tumour size. This cluster implicated ceramide species in the small size of some tumours. Looking in detail at the ceramide sub class composition in figure 4.16 revealed that samples in this cluster had a significantly higher proportion of Cer 18:0 and a significantly lower proportion of Cer 16:0 in comparison to the background of all other tumour samples at the 5% level based upon a Student's t-test.

PE showed an interesting cluster of tumour samples which are significantly dissimilar from a large proportion of the other tumours, as seen in figure 4.17. The cluster highlighted in red had a significant overrepresentation of the metadata annotation "pre-radio=TRUE" and contained the patient sample numbers: 1402, 81, 96, 1383, 1364, 1300, 1234, 94, 85 and 1361 (see Appendix 2 table 6). The samples seem to be clustered by whether the patient's underwent radiotherapy prior to tissue collection and identified the sub class PE potentially being effected by radiotherapy. The specific difference in this cluster from the other samples and the inferred effect of radio therapy is visualised in figure 4.18. Of the mean lipid composition of the clustered patients, there was a large shift in composition towards longer chain length species such as PE 38:- and PE 40:- and the associated decrease in the shorter chain

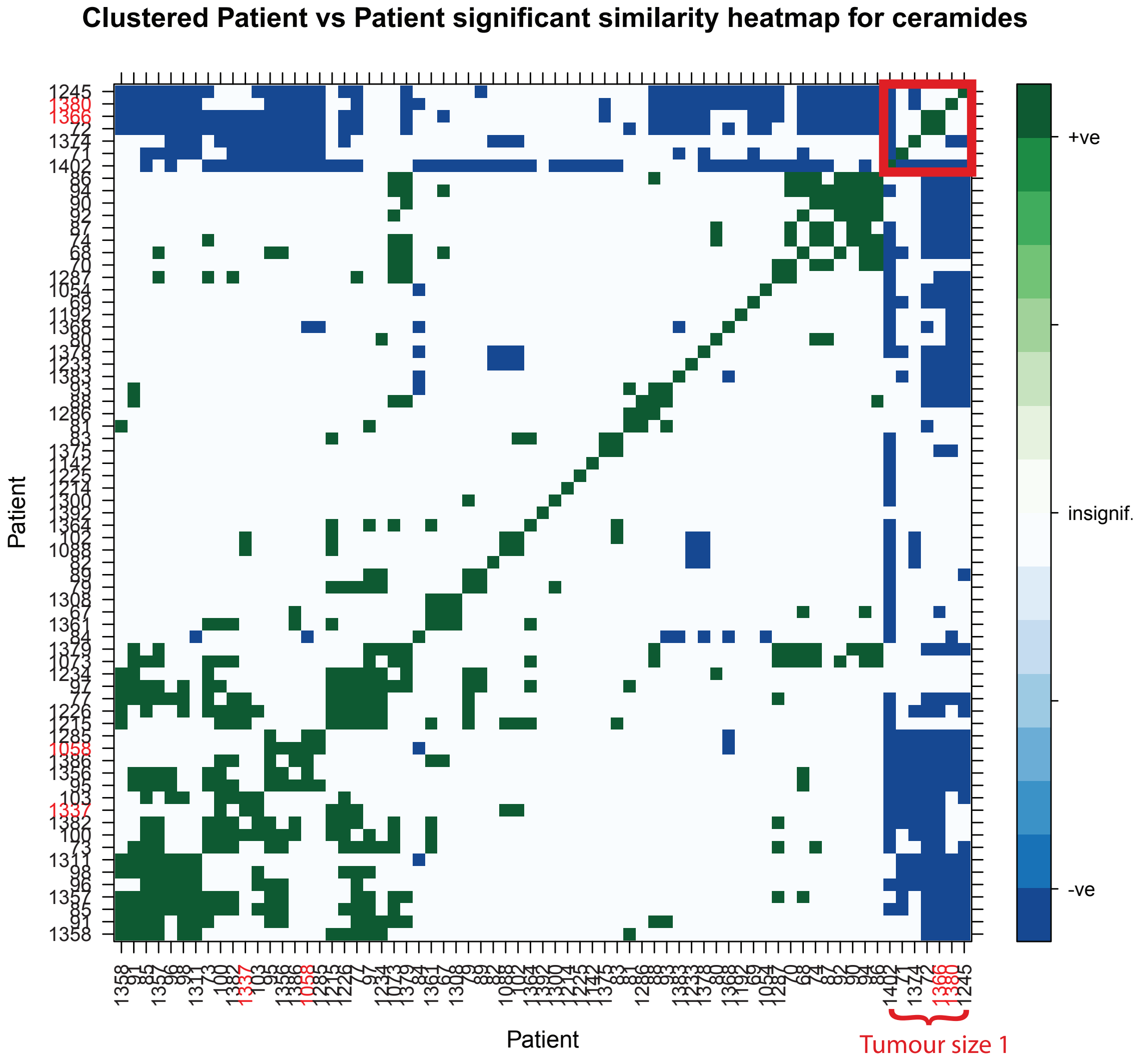


Figure 4.15: The significant correlation heatmap of all tumour samples against one another for the ceramide sub class. Green represents a significant similarity between two samples, blue represents a significant dissimilarity between two samples, while white represents an insignificant correlation. The red highlighted group of patients show a significant overrepresentation of the metadata 'tumour size: 1'.

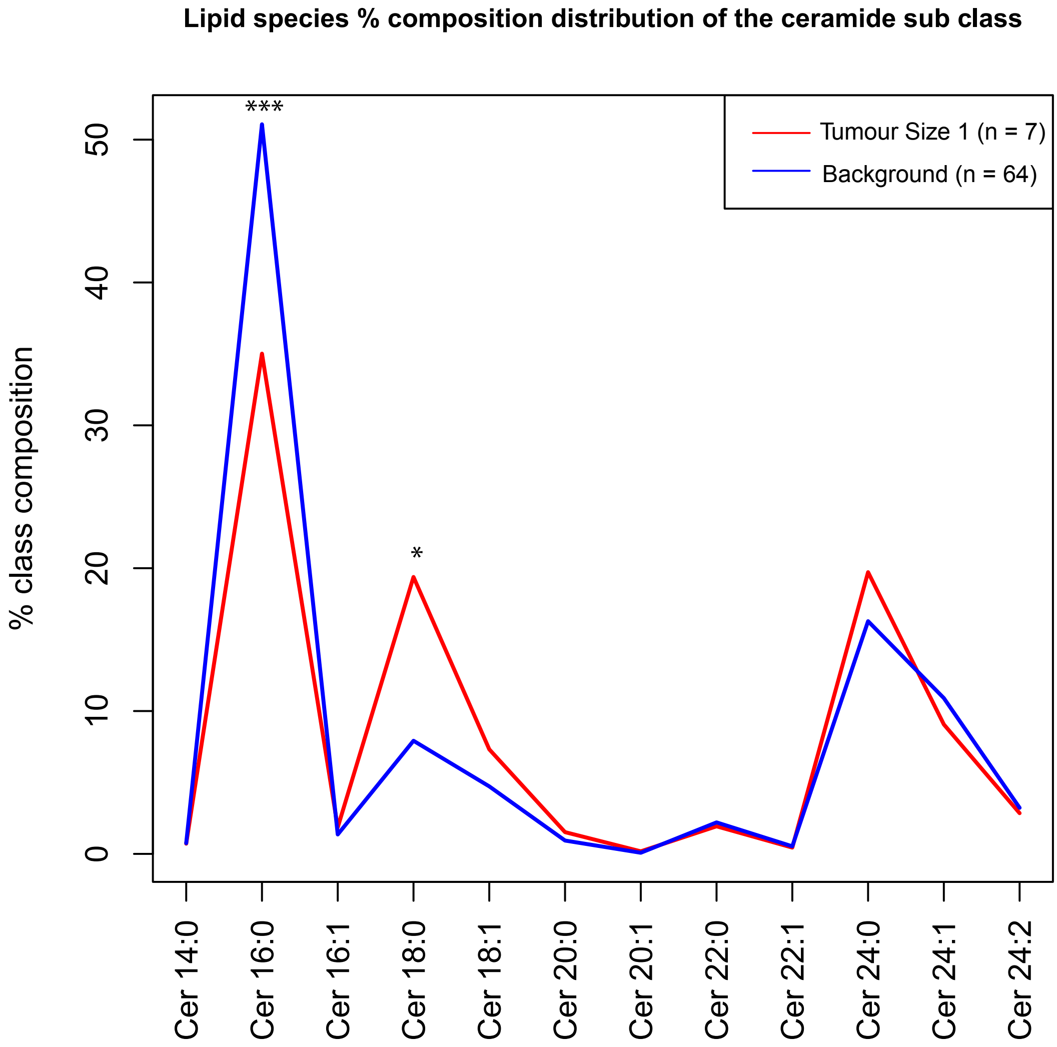


Figure 4.16: The mean species composition of the ceramide sub class for tumour samples in the cluster (red) and all other samples (blue). There is an increased composition of Cer 18:0 and a decreased composition of Cer 16:0 in the 'tumour size 1' cluster. Both are significantly different to all other tumour samples, denoted by '*','**' and '***' representing 5%, 1%, and 0.1% levels respectively.

length species PE 34:- and PE 36:-, all significant at a minimum of the 5% level based upon Student's t-tests.

Tumour samples in the PI sub class were arranged into an interesting pair of clusters which show significant differences between each other, see figure 4.19. The cluster highlighted in red has an overrepresentation of the 'tumour size 3' metadata annotation and contained the patient sample numbers: 1287, 71, 1382, 1215, 1386, 1356, 91, 70, 79, 1379, 1192, 1285, 77, 72, 102, 1226 and 1214 see Appendix 2 table 6. Conversely, the cluster in purple has an overrepresentation in the 'tumour size 2' metadata annotation and contained the patient sample numbers: 1392, 98, 92, 1378, 1364, 88, 69, 82, 1088, 1234, 1311, 1368, 94, 1058, 1383, 68, 1402 and 81 see Appendix 2 table 6. It is interesting to see the significant dissimilarity these two clusters of patients show between one another. This highlighted the potential role of the sub class PI in the size of a tumour. More detailed analysis of the sub class composition for both clusters in figure 4.20 revealed a significant increase in PI 34:- and 36:- species from tumours size 2 to tumour size 3 and a decrease in PI 38:- with no real change in PI 40:-. All were significant to at least the 5% level.

The final lipid sub class with interesting tumour sample clusters was PS, seen in figure 4.21. Highlighted in red is a cluster of tumour samples that showed significant dissimilarity to the sample cluster highlighted in purple. Underlying these two clusters is the significant overrepresented metadata annotations 'tumour size: 2' and 'tumour size: 3' respectively. The 'tumour size: 2' cluster highlighted in red contained the patient sample numbers: 89, 93, 1358, 94, 68, 1058, 1392, 84, 83, 88, 90, 1380, 92, 4102, 1311, 82 and 81 see Appendix 2 table 6. The 'tumour size: 3' cluster highlighted in purple contained the patient sample numbers: 1214, 1192, 80, 1287, 74, 77, 79, 1215, 73, 67, 1226 and 1234 see Appendix 2 table 6. This suggests that PS has a significant role in differentiating smaller tumour samples from larger ones. Isolating the specific differences in lipid sub class composition at the species level, seen in figure 4.22, revealed a less clear trend as PI but overall there is a comparable shift towards shorter chain fatty acid species between tumour size 2 to 3.

It is clear previously, that chain length has been an important feature in differentiating tumour and normal samples, with the general trend of longer chain species being more abundant in tumour samples. At the level of tumour size, the discrete

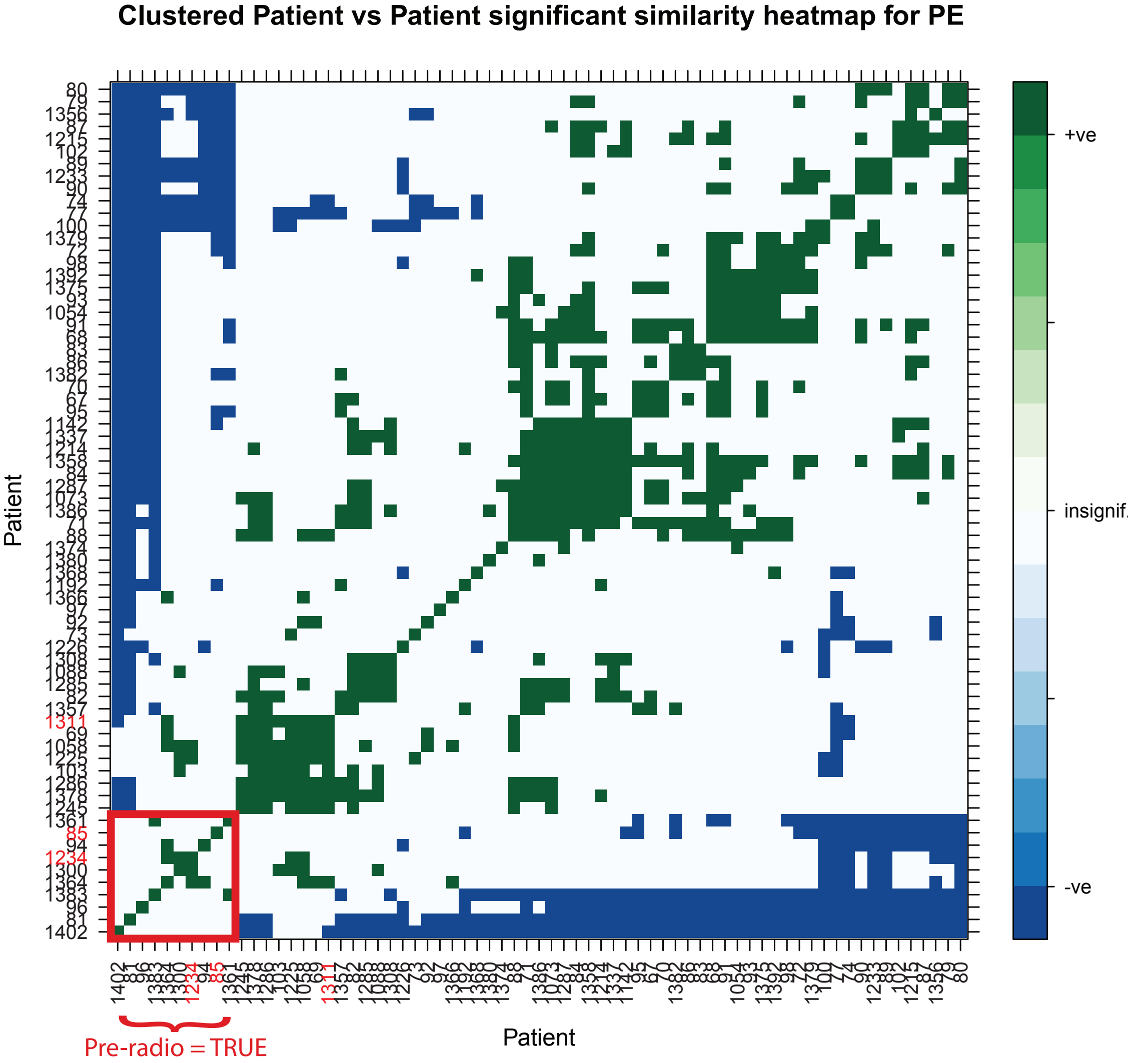


Figure 4.17: The significant correlation heatmap of all tumour samples against one another for the diacylglycerophosphoethanolamine sub class. Green represents a significant similarity between two samples, blue represents a significant dissimilarity between two samples, while white represents an insignificant correlation. The red highlighted group of patients show a significant overrepresentation of the metadata feature "pre-radio = TRUE", which translates to the patient undergoing a course of radiotherapy prior to sample collection.

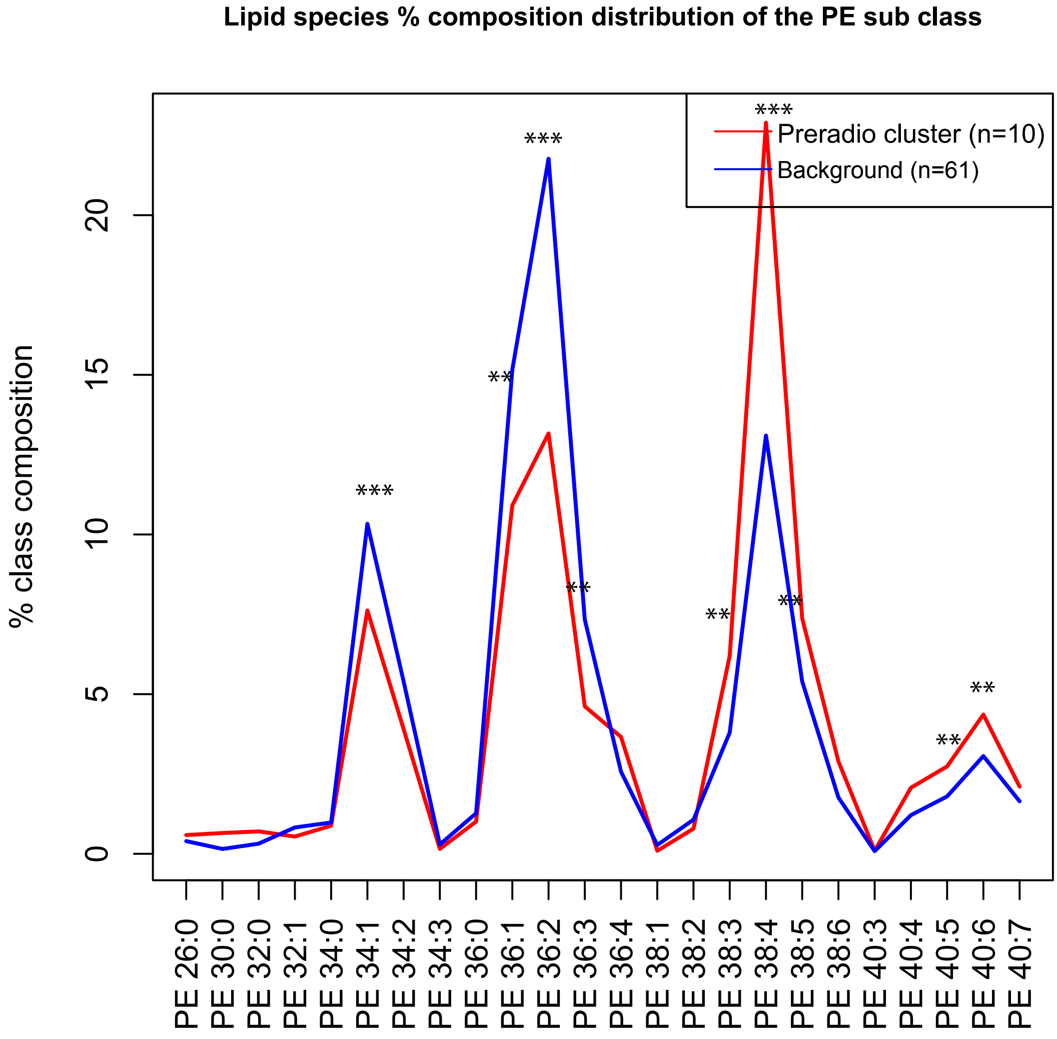


Figure 4.18: The mean species composition of the diacylglycerophosphoethanolamine sub class for the tumour samples in the cluster (red) and all other samples (blue). There is a clear shift in composition with PC 34:- and PC 36:- species decreasing and longer chain PC 38:- and PC 40:- species increasing in comparison to all other tumour samples with significance denoted by “*”,“**” and “***” representing 5%, 1%, and 0.1% levels respectively.

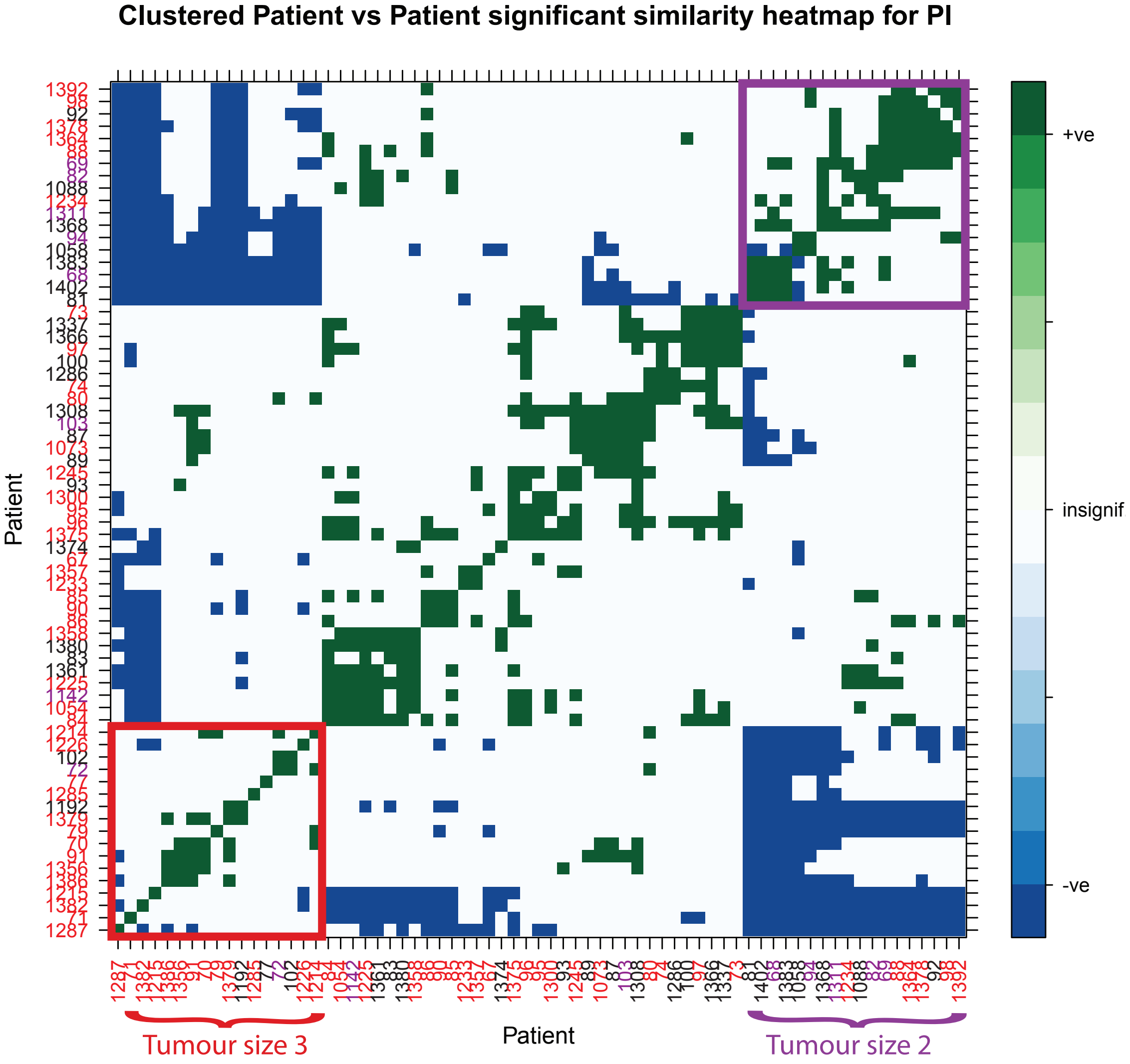


Figure 4.19: The significant correlation heatmap of all tumour samples against one another for the diacylglycerophosphoinositol sub class. Green represents a significant similarity between two samples, blue represents a significant dissimilarity between two samples, while white represents an insignificant correlation. The red highlighted group of patients show a significant overrepresentation of the metadata feature 'tumour size: 3'. The samples highlighted in the purple cluster show and overrepresentation of the 'tumour size: 2'.

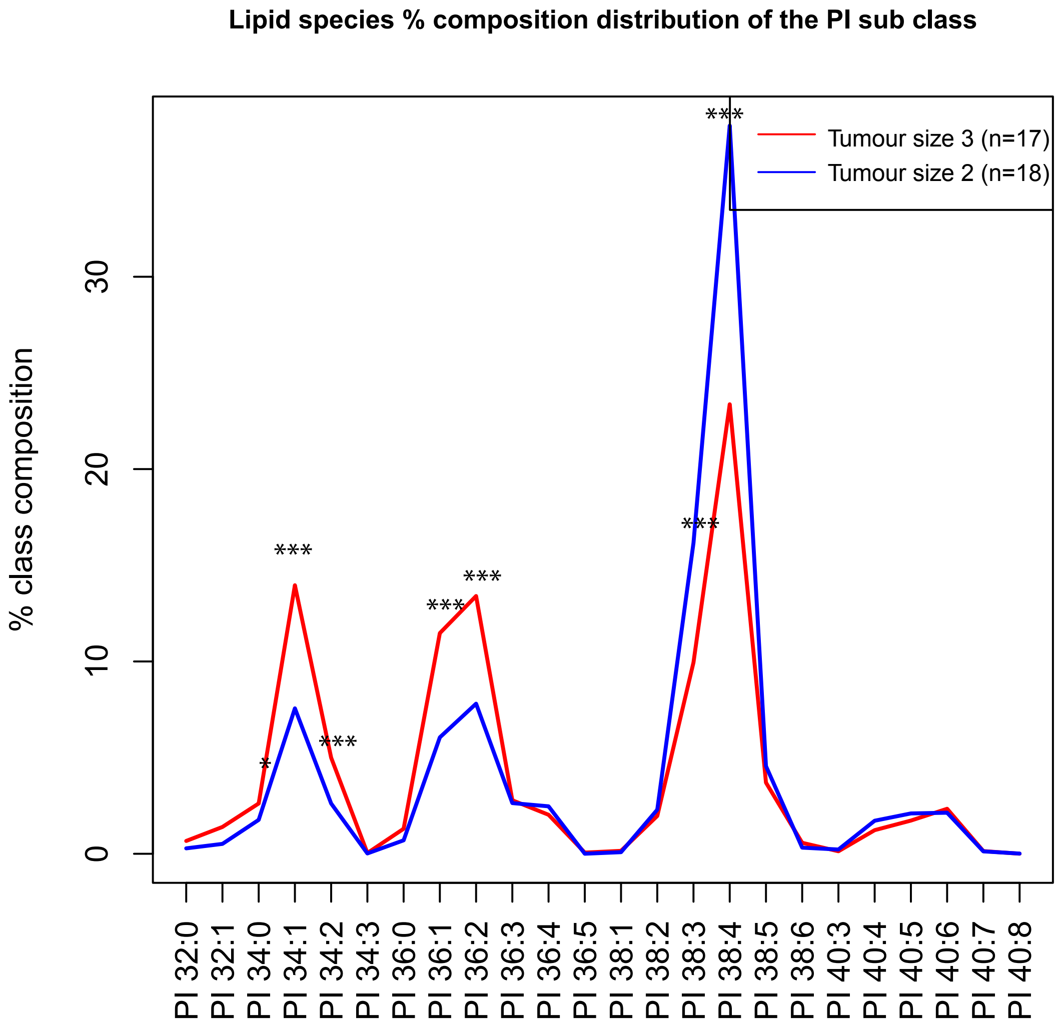


Figure 4.20: The mean species composition of the diacylglycerophosphoinositol sub class for the tumour samples in the 'tumour size: 3' red cluster (red line) and the 'tumour size: 2' purple cluster (blue line). There is a clear shift in composition, namely PI 38:4 is significantly decreased in the cluster predominated by tumour size 3, which shows a bias towards the small chain length PI species. Significance at the 5%, 1% and 0.1% levels are denoted by "*", "**" and "***" respectively.

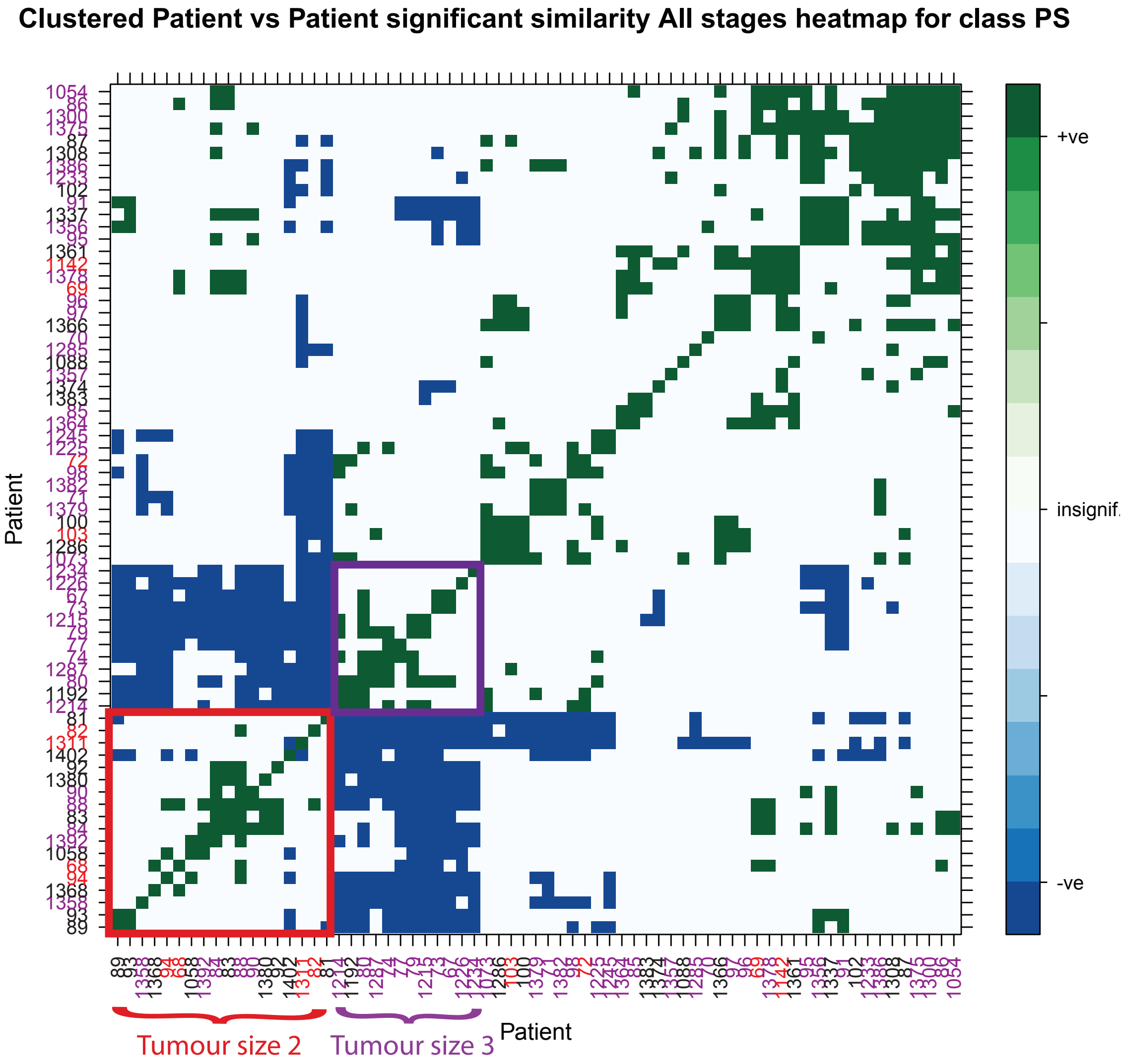


Figure 4.21: The significant correlation heatmap of all tumour samples against one another for the diacylglycerophosphoserine sub class. Green represents a significant similarity between two samples, blue represents a significant dissimilarity between two samples, while white represents an insignificant correlation. The cluster of patients highlighted in red are overrepresented in the metadata annotation 'tumour size 2', while the cluster highlighted in purple overrepresent the metadata annotation 'tumour size 3'.

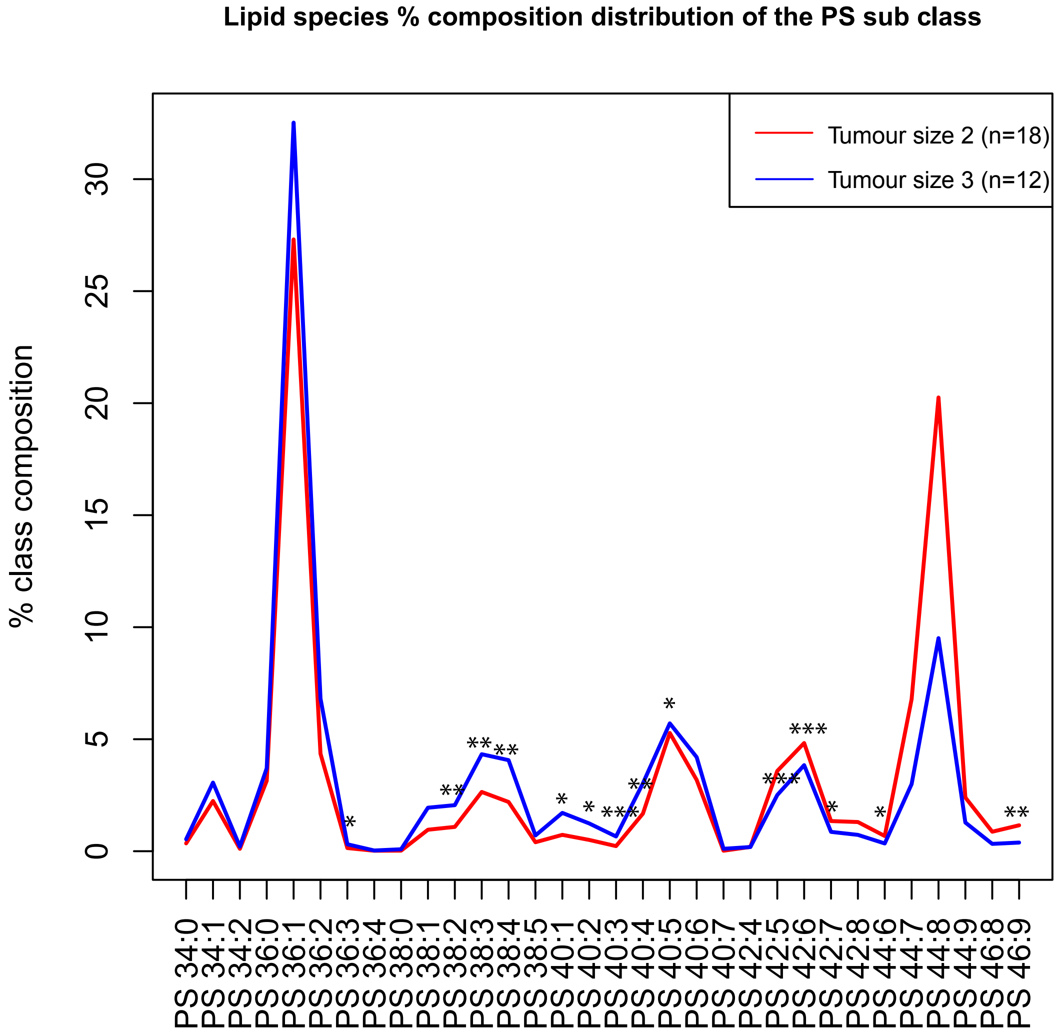


Figure 4.22: The mean species composition of the diacylglycerophosphoserine sub class for the tumour samples in the red highlighted cluster (red line) and the purple highlighted cluster (blue line). PS species with 42 or greater carbons appear to be reduced in the 'tumour size 3' cluster, while PS species with 38 to 40 carbons show an increase in the 'tumour size 3' cluster. Significance at the 5%, 1% and 0.1% levels are denoted by "*", "**" and "***" respectively.

transitions between size shows sub class specific regulation of chain length preference. For the subclass PI in figure 4.20, the transition from tumour size 2 to 3 is defined by an increase in average chain length of PI species. The opposite is true of the PS subclass which shows a decrease in mean chain length in the transition from tumour size 2 to 3.

4.3.3 Heterogeneity of cancer

Cancer in general is accepted to be an extremely heterogeneous disease at the cellular level. To test the chaotic effects of cancer on the lipidome, I proposed two methods; an analysis of variance and an analysis of mean patient correlation between normal and tumour samples. Based on the assumption that cancer is very diverse and what appears on the surface to be a single disease, can actually be very many. These analyses tested the hypothesis that normal samples are statistically less diverse in the composition of their lipidome.

4.3.3.1 Analysis of variance

Due to the data having exhibited a non-normal distribution in some parts, the classical F-test of equality of variance did not suffice, as it is highly sensitive to non-normal data. As a substitute, the Brown-Forsythe test (see section 4.2.3.16) was used as a robust test for the equality of variance. As a first comparison, data was normalised as described in section 4.2.3.6. The normalised lipid sub class total was calculated by summing together all the normalised lipid species concentrations in each sub class for each sample.This resulted in each sample having a set of values that described the percentage each lipid sub class contributed to the entire lipidome. Aggregating this data by sample type allowed the comparison of variance of a total lipid sub class' contribution to the lipidome between normal and tumour samples. A Brown-Forsythe test was calculated on normal and tumour samples between each sub class (25 sub classes; 25 tests). This resulted in six lipid sub classes showing significant difference in their variance between normal and tumour samples: LPC a, PC a, CL, PE, PG and LPA. All of these showed an increase in variance in tumour samples except the lipid sub class LPA, confirming tumour samples seem to have a more chaotic lipidome. The same analysis can be performed on lipid species, to test which individual lipid species have a significant difference in variance and can be considered more or less regulated in tumour samples. Due to the large number of comparisons, Bonferroni correction was employed again as a conservative corrector of p-values to reduce the false positive rate. After correction, 87 lipid species were found to have significant difference in variance at the 5% level. Of the significant lipids, 70 out of 87 show increased variance in tumour samples and agree with the earlier hypothesis. Of the interesting set of 17 lipids that have reduced variance in tumour samples, 14 of them are TGs, specifically the short chain TGs in the fatty acid structure range TG 40:0 to TG 44:2. The reduction in variance could be partially accounted for by lipids not being detected in the normal sample, but this does not diminish the fact that these lipids are systematically being regulated with tighter control than other lipids in normal samples. The polarisation of the TG sub class in general originally seen in figure 4.9 was endorsed in this result with the long chain fatty acid TGs showing significant increase in variance in tumour samples. In addition to TG, long chain gylcerophospholipids were also significantly increasing in variance in tumour samples.

4.3.3.2 Analysis of correlation

In a second attempt to justify the perceived chaotic cell state cancer induces, a patient correlation based analysis was devised. This analysis was done in two stages; first on total lipid sub classes and then on individual lipid species. The mean correlation of data within normal samples and within tumours was measured to detect any difference between groups. Absolute data was first normalised as described in section 4.2.3.6. The lipid sub class totals were then calculated by summing together the normalised lipid species. This equated to each sample having a vector of normalised sub classes, where each value represented the percentage a sub class made of the total lipidome. For all normal samples these vectors were correlated using Pearson product-moment correlation coefficient (see section 4.2.3.8), the process was repeated for tumour samples. The two correlation matrices (one of normal samples, the other tumour samples) were flattened into two correlation distributions. Prior to testing their difference they were transformed using the Fisher transformation (see section 4.2.3.17), transforming the correlation r distribution into a normal distributed z distribution. These Fisher z score distributions were tested for differences in location by a paired Student's t test (see section 4.2.3.2). Figure 4.23 shows the Fisher z score distribution of the transformed intra-sample Pearson correlations. There was a highly significant difference between the mean normal sample correlation and the mean tumour sample correlation. Specifically, tumour samples were significantly less correlated, providing further evidence of the concept of cancer having a heterogeneous effect on the sub class composition of the lipidome.

Applying the same methodology to the normalised lipid species data, tests if the same difference can be seen at the higher resolution of lipid species and whether the difference there is perhaps more pronounced. Figure 4.24 shows the same trend of significantly higher correlation between normal samples than tumour samples. The correlation has decreased overall in comparison to the lipid sub class data in figure 4.23, showing that the lipid species level is more diverse than the lipid sub class level as would be expected from a more heterogeneous feature set.

4.3.4 Network centric analysis

The concentration of a lipid is a function of both its forward and reverse reactions, the partners of which are also subject to equilibria defined by their associated reactions. This creates a large number of distantly connected effects on a single lipids concentration, some of which may be unmeasured and hence difficult to locate the origin of a perturbation. Like most small molecules, lipids are in a constant state of shifting equilibrium based upon the changing cellular conditions as a result of gene expression, environmental change and a whole host of other factors. The lipids detected in this investigation share some structural features and their synthesis reflects this common building plan by repeated use of similar enzymes to construct them. The conversion of one lipid to another is key to the dynamic adaptation of the cell to its environment and the reactions that define this process can be broken down into three simplified categories:

Fatty acid elongation/reduction A lipid may become another lipid by elongation of its fatty acid chains e.g. PC 18:0/18:2 can be elongated to PC 20:0/18:2. The reverse is also possible where fatty acid chains are reduced in length.

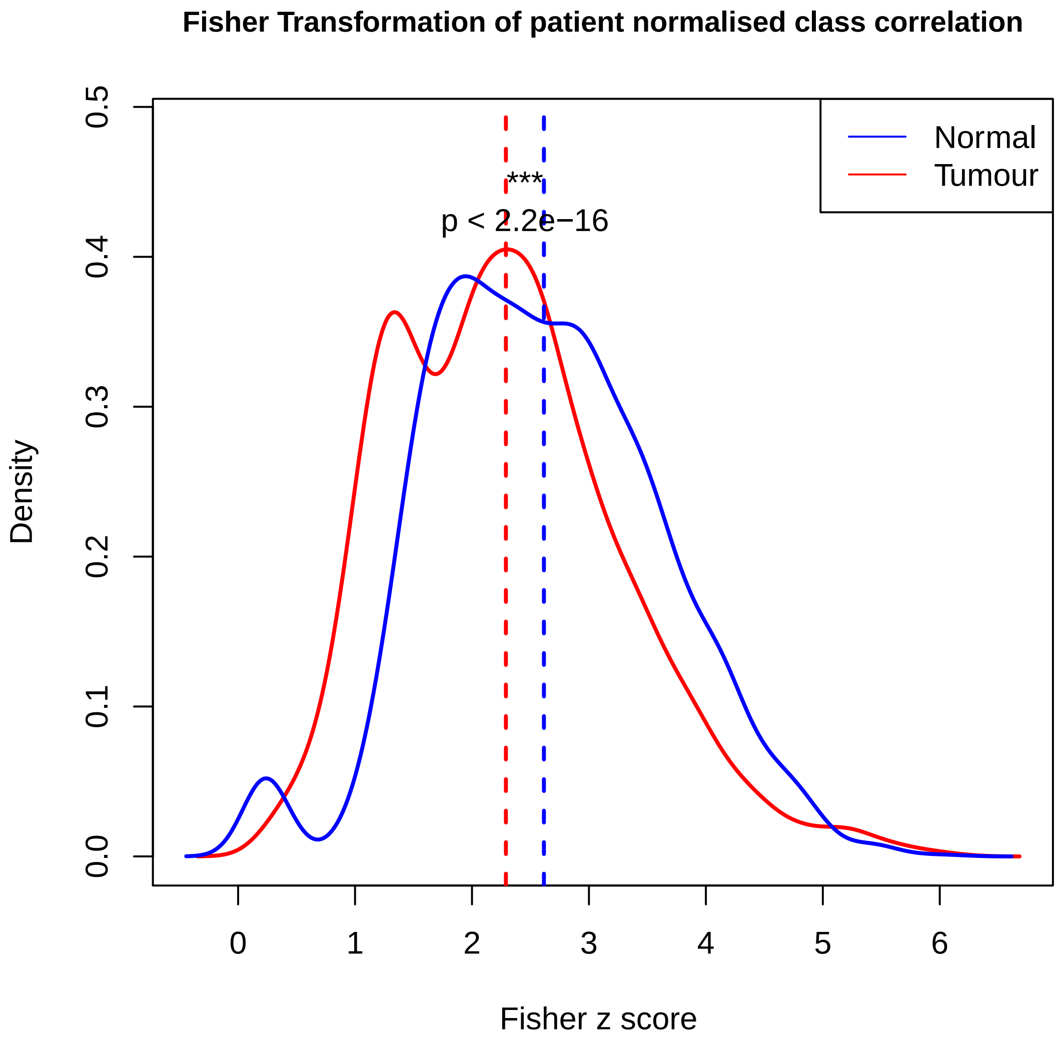


Figure 4.23: The lipid sub class Pearson correlation distribution of normal and tumour samples are transformed via the Fisher transformation to a comparable z score distribution. Plotting the density of the z score distribution shows a significant difference in their means. The sub class compositions of tumour samples are significantly less correlated to one another than those of normal samples are to one another. The p-value calculated is the smallest number possible in R, so its exact value should be disregarded. Significance denoted by “*”,“**” and “***” at the 5%, 1%, and 0.1% levels respectively.

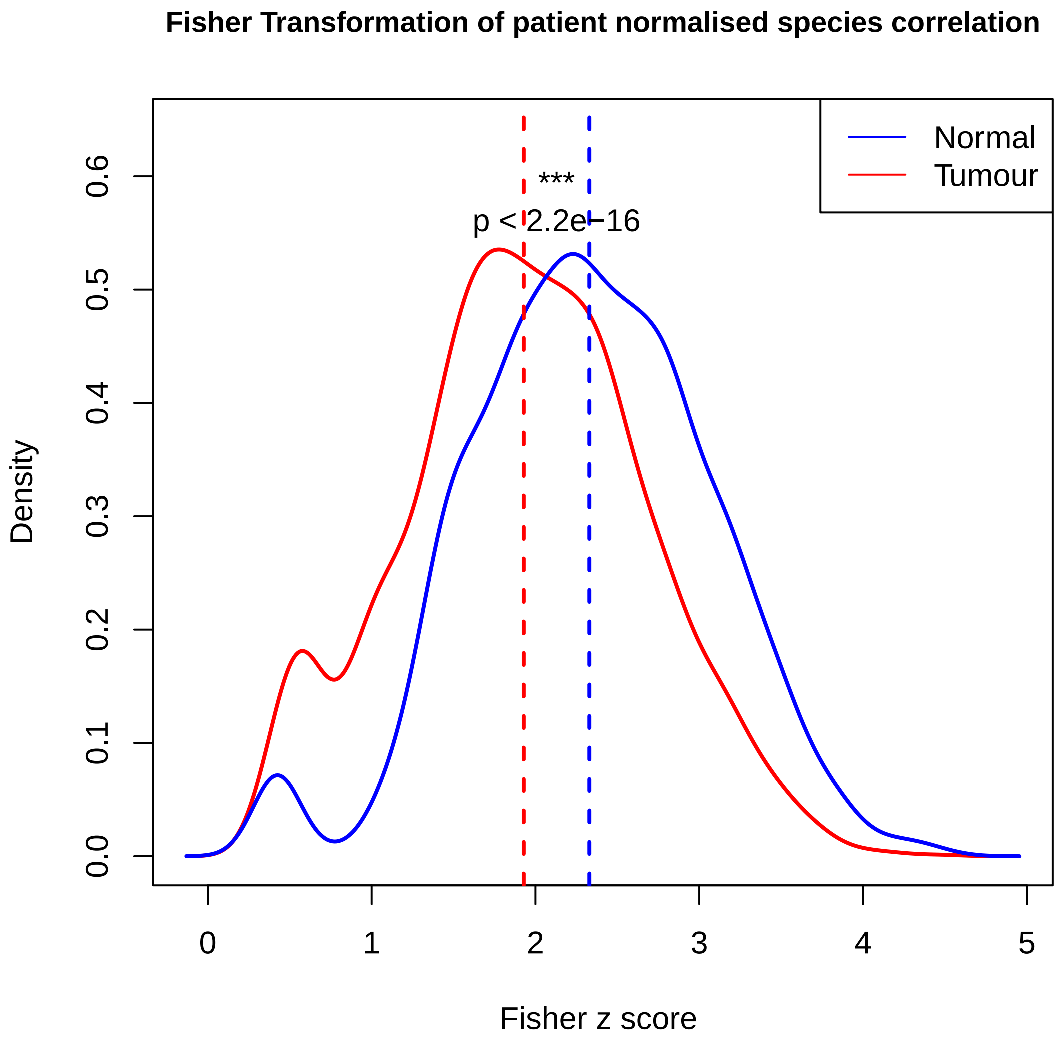


Figure 4.24: The lipid species Pearson correlation distribution of normal and tumour samples are transformed via the Fisher transformation to a comparable z score distribution. Plotting the density of the z score distribution shows a significant difference in their means. The lipid species composition of tumour samples are significantly less correlated to one another than normal samples are to one another. The p-value calculated is the smallest number possible in R, so its exact value should be disregarded. Significance denoted by "*", "**" and "***" at the 5%, 1%, and 0.1% levels respectively.

Fatty acid saturation/desaturation Fatty acid chains may have their carbon backbone saturated or desaturated to produce a variety of distinct lipids, for example PC 18:0/18:1 may be desaturated to PC 18:0/18:2.

Headgroup substitution The critical head group that distinguishes glycerophospholipid main classes from one another may be replaced to create a new lipid of a different main class, e.g. PC 36:2 to PS 36:2 by substitution of the choline headgroup for serine.

Accounting for these three types of reaction, the list of 566 identified lipids could be modelled as an extremely complex highly connected network, where all lipids can be converted into each other through a short series of reactions. For instance PA 34:1 to PC 36:2 could follow any number of distinct reaction paths, of which this is one example:

PA 34:1 to PS 34:1 Headgroup substitution

PS 34:1 to PE 34:1 Headgroup substitution

PE 34:1 to PC 34:1 Headgroup substitution

PC 34:1 to PC 36:1 Fatty acid elongation

PC 36:1 to PC 36:2 Fatty acid desaturation

This highly connected network of lipids is far too complex to attempt an investigation into the conversion of lipid species to one another. A simplified network that shows just the headgroup substitution reactions was constructed using CellDesigner version 4.2 (Systems Biology Institute (SBI), Tokyo, Japan) as described in section 4.2.2.5 and is visible in figure 4.25. The network was distilled from the combined knowledge of Prof. Michael Wakelam and Dr. Qifeng Zhang.

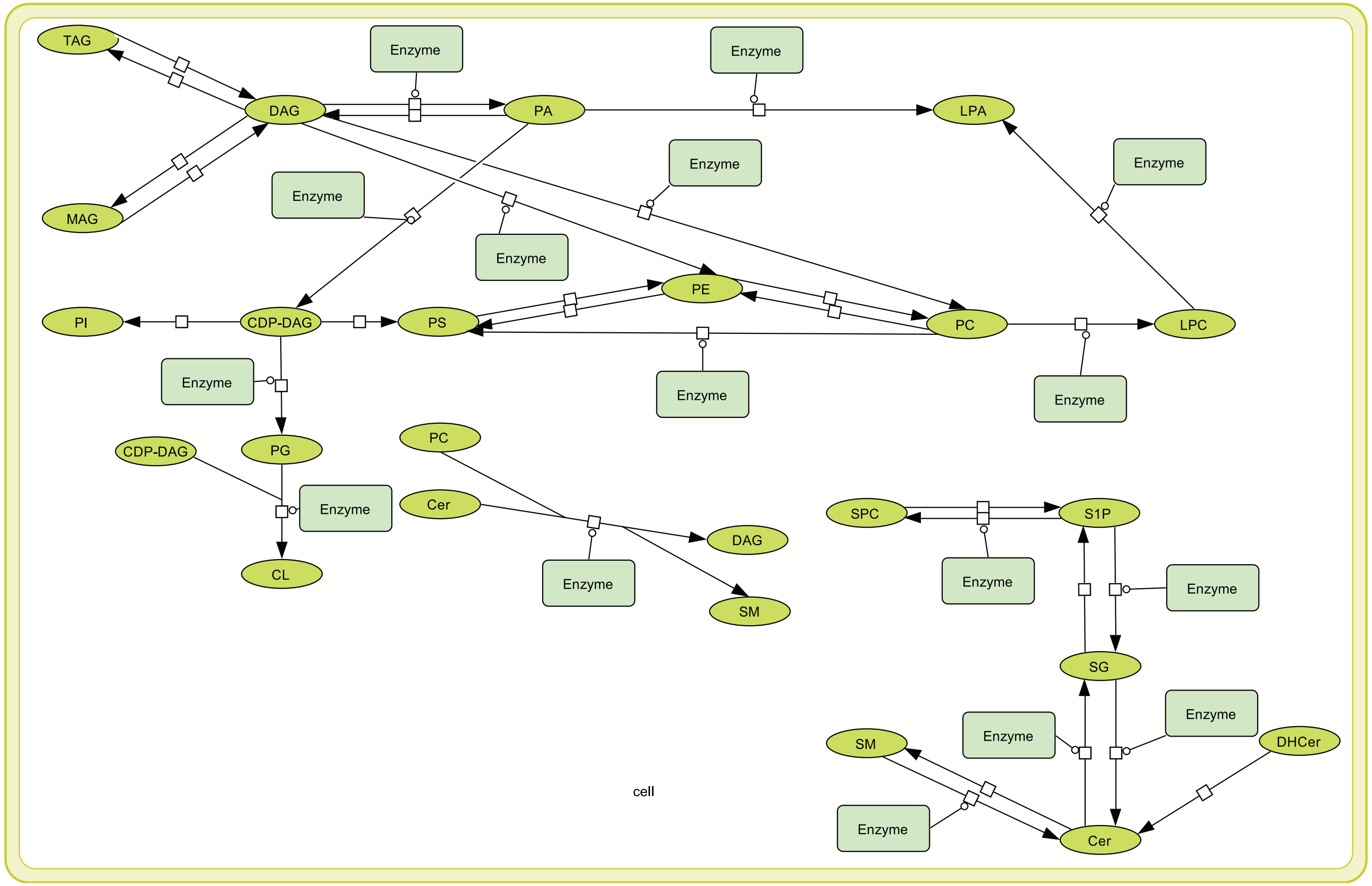


Figure 4.25: A network of sub class interactions that define the conversion of one lipid sub class to another, constructed from the combined knowledge of reactions reported in the literature. In order to simplify the highly connected network at the lipid level, reactions altering the fatty acid structure of a lipid species are disregarded. This effectively creates a set of layers of the network, one for each fatty acid structure e.g. 36:2 with no reactions between layers.

4.3.4.1 Lipid class reactions

In order to assess the change in equilibrium of lipid sub classes, the reaction ratio of products to reactants can be compared between paired normal and tumour samples. Assuming the samples reflect the cell state at equilibrium, any significant difference in reaction ratio indicates a potential enzyme, or set of enzymes that govern the reaction as protein/gene targets playing a role in colorectal cancer. The data was divided into tumour and normal samples as before and the percentage each lipid sub class constitutes of a sample's lipidome calculated. The network diagram was reduced to a set of binary reactions with a single product and single reactant. For each reaction separately, the log10 ratio of products to reactants was calcualted for each sample. The paired reaction ratios of a tumour and normal sample were compared using a paired Student's t-test (see section 4.2.3.2), p-values were subsequently corrected for multiple testing by Bonferroni correction. Of the twelve reactions that only involve diradyl lipid sub classes (requiring a single headgroup substitution for the reaction to take place), three reactions had a significantly different reaction ratio between normal and tumour samples. They are: 'PA→PG', 'PS→PE' and 'PC→PE', with 15%, 10% and 10% shifts towards products respectively. All show a shift towards increased products in tumour samples at the 5% level, in this case higher accumulation of PG and PE.

4.3.4.2 Lipids species reactions

Disregarding the reactions which alter fatty acid structure, the network can be conceptualised into repeated layers of the same network in figure 4.25 stacked on top of each other, where a single 'layer' reflects a single fatty acid structure e.g. 36:2. These 'layers' are independent of each other, and the lipids within them (that share the same fatty acid structure) can be investigated for fatty acid structure specific differences in lipid species reaction ratios. Individual reactions between lipid species sharing the same fatty acid structure were investigated in the same manner as in the previous section 4.3.4.1. Not only were the head group substitution reactions important, but also the layer in which significant differences of reaction equilibria between tumour and normal samples occured. Different enzymes have specificity for lipids with particular fatty acid structures, so any significant overrepresentation of a particular fatty acid structure layer of the network may indicate a specific set of enzymes that may be involved in colorectal cancer. For this analysis it was only possible to investigate binary reactions with a single reactant and product, as well as reactions where the fatty acid structure did not change i.e. conversion of one diradyl species to another 'PC 36:2→PE 36:2'. Twelve sub class reactions translates to 261 lipid species reactions, based upon the identified lipid species in the dataset.

Using Student's t-tests followed by Bonferroni correction to test for differences between the product to reactant ratios of reactions of tumour and normal samples resulted in 40 out of 261 being significant at the 5% level, and these are summarised in table 4.2. The general trends that were extracted from these 40 significant reactions were the large shift of 'DG→PE' and 'DG→PC', the tumour samples seemed to be funnelling resources away from DG and into the rest of the pathway. Highly desaturated PA species were being converted to PG and PI in tumour samples. Perhaps the most consistent result was the build-up of PE; with the reactions 'PS→PE' and 'PC→PE' show a bias towards the build-up of PE in tumour samples.

A Fisher Exact test on the significant generic reaction (e.g. 'PS→PE') and an independent one on the fatty acid structures of those reactions (e.g. 36:2) confirmed the significant overrepresentation of 'PC→PE' and the fatty acid structure 40:8.

A number of sub class and species specific reactions have been shown to be perturbed in tumour samples. A follow up proteomics experiment could be devised to confirm the source of these alterations in the lipidome. A targeted Selected Reaction Monitoring (SRM) approach where transition sets are designed for the group of known lipid metabolism enzymes only. However, this approach has the flaw that any replacement polymorphisms in the transition sets, will return no identification and miss potentially, very valuable genetic markers. A phospho-proteomics approach may reveal changes in phosphorylation that has specific effects to lipid metabolic enzymes or upstream proteins responsible for lipid uptake and transport.

Reactant	Product	Shift	% Shift
DG 40:4	PE 40:4	+	35.4
DG 40:5	PE 40:5	+	32.9
DG 40:6	PE 40:6	+	28.2
DG 32:0	PC 32:0	-	-66.6
DG 32:2	PC 32:2	+	67.9
DG 38:7	PC 38:7	+	47.6
DG 40:4	PC 40:4	+	20.9
DG 40:8	PC 40:8	+	13.0
PA 36:3	DG 36:3	+	31.4
PA 36:3	PG 36:3	+	20.7
PA 40:7	PG 40:7	+	29.2
PA 40:8	PG 40:8	+	20.9
PA 40:7	PI 40:7	+	54.5
PA 40:8	PI 40:8	+	102.2
PS 36:0	PE 36:0	+	20.9
PS 36:1	PE 36:1	+	18.8
PS 38:1	PE 38:1	+	53.7
PS 38:2	PE 38:2	+	37.7
PS 40:4	PE 40:4	+	31.6
PS 40:5	PE 40:5	+	33.7
PS 40:6	PE 40:6	+	34.2
PE 38:2	PS 38:2	-	-37.7
PE 38:3	PC 38:3	-	-19.6
PE 40:5	PC 40:5	-	-27.4
PC 32:0	PE 32:0	+	32.3
PC 32:1	PE 32:1	+	30.0
PC 34:2	PE 34:2	+	13.9
PC 38:3	PE 38:3	+	19.6
PC 38:4	PE 38:4	+	17.4
PC 38:5	PE 38:5	+	14.8
PC 40:4	PE 40:4	+	15.3
PC 40:5	PE 40:5	+	27.5
PC 40:6	PE 40:6	+	22.4
PC 40:7	PE 40:7	+	21.6
PC 34:2	PS 34:2	+	23.4
PC 36:0	PS 36:0	-	-18.4
PC 36:1	PS 36:1	-	-12.5
PC 38:2	PS 38:2	-	-26.0
PC 38:5	PS 38:5	+	17.0

PC 40:3	PS 40:3	-	-71.5

Table 4.2: A table of reactions which show significant differences in reactant:product ratio between normal and tumour sample. The “Shift” column represents whether the reaction has shifted in favour of products in tumour samples, ‘+’ indicating it has shifted towards favouring products, and ‘−’ indicating it has instead shifted to favouring reactants more. This should not be confused with the actual reaction ratio; this is only the shift in the reaction ratio. A reaction ratio may still remain well in favour of reactants but be ‘+’ because the reaction has less strong emphasis on favouring reactants in tumour samples.

4.3.4.3 Reaction routes

Reactions make up the components of a network. In figure 4.25 from one lipid sub class to another there are a few viable reaction routes. Constructing these reaction routes for each sample and correlating them against each other can cluster patient samples based upon the composition of their pathways. Similar to section 4.3.2.2, where whole lipid sub class compositions were correlated against one another to reveal patient clusters that supported the metadata, this approach correlated samples on a much smaller highly related set of measurements. For the reaction route ‘TG→DG→PA→PG→CL’ a vector of the % contribution that the sub class made to the total lipidome for each tumour sample was constructed. These vectors were correlated against each other using Pearson correlation (see section 4.2.3.8) to produce the correlation matrix seen in figure 4.26. A cluster of three patients was found to be very dissimilar to the other patients; upon scanning for overrepresented metadata the annotation ‘Tumour size: 4’ was found to be significant at the 5% level. In order to find the nature of this dissimilarity the median sub class % composition of the lipidome was plotted for both groups in figure 4.27 (members of the cluster against a background of all other samples). Differences between the normal and tumour distributions were tested by a Mann-Whitney U test (described in section 4.2.3.4) and show the sub classes TG, PA and CL to be significant at the 5% level.

The clustered patients all with tumour size 4 have an abundance of TG, PA and CL in comparison to the other samples.

4.3.5 Stage centric analysis

Patients were divided into discrete groups based upon the Duke's staging classification system outlined in section 4.4. These stages represent a discrete non-linear time course, and transitions between states are of biological and therapeutic interest. The lipid species and sub classes that differ significantly in their percentage contribution to the total lipidome between tumour samples from a group of patients of one stage to that of an adjacent stage provided insight into the progression of cancer and potential cellular events that increase the severity of the disease.

4.3.5.1 Adjacent stage comparison

For tumour samples, the percentage each sub class represents of the whole lipidome was calculated. Patients were stratified into their stages, and Student's t-tests calculated to compare the difference in mean % of the lipidome each sub class represented between adjacent stages. The transition from Adenoma to Duke's A showed a significant decrease in the amount of PS at the 5% level. The transition from Duke's C1 to Duke's C2 showed a significant increase in the level of PE at the 5% level, conversely the subsequent transition from Duke's C2 to Duke's D shows a significant decrease in PE.

The same analysis performed at the lipid species level revealed a set of lipids that significantly differ in their percentage of the total lipidome between Duke's stages. In figure 4.28, the significant decrease in PS seen in the previous analysis between adenoma and Duke's stage A tumour samples was attributed to PS 36:1, PS 38:0 and PS 38:1. In addition CL 68:4, PC 28:0, PC 30:1 and PG 46:7 showed a significant increase from adenoma to Duke's stage A. The transition between Duke's stage A and B was not distinguished previously at the lipid sub class level. However the lipid species level showed CH, PC a 38:1, PE a 38:1, TG 40:2, TG 40:3 and TG 42:4 have a significant increase in Duke's stage B tumour samples as shown in figure 4.29. Figure 4.30 shows the lipid species CE 18:3 and PC a 42:3 underwent a significant increase between Duke's B and Dukes C1 stages. The transition between

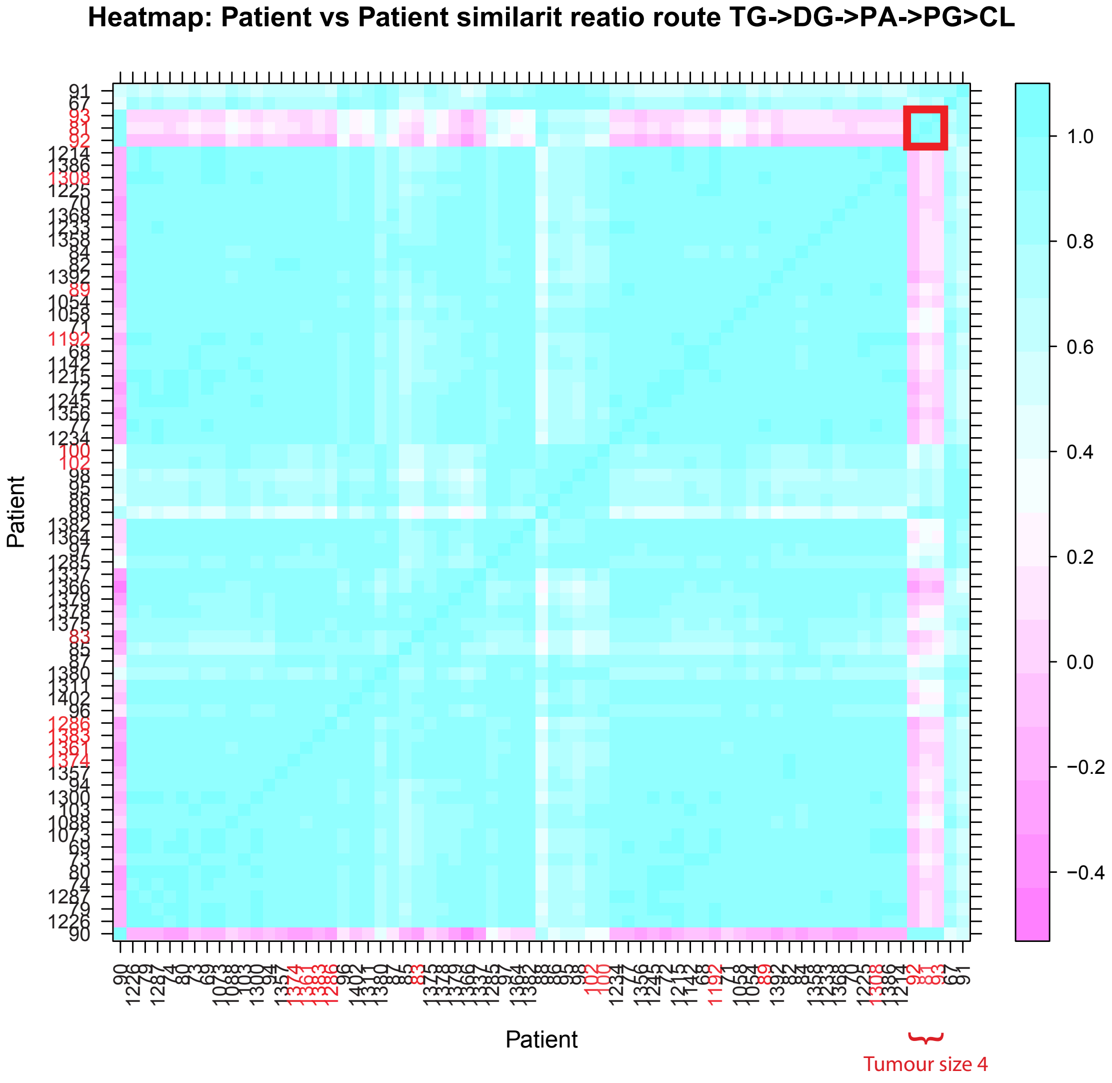


Figure 4.26: The Pearson correlation matrix of tumour samples for the reaction route ‘TG→DG→PA→PG→CL’. The cluster highlighted in red shows significant overrepresentation of the metadata annotation tumour size 4 at the 5% level.

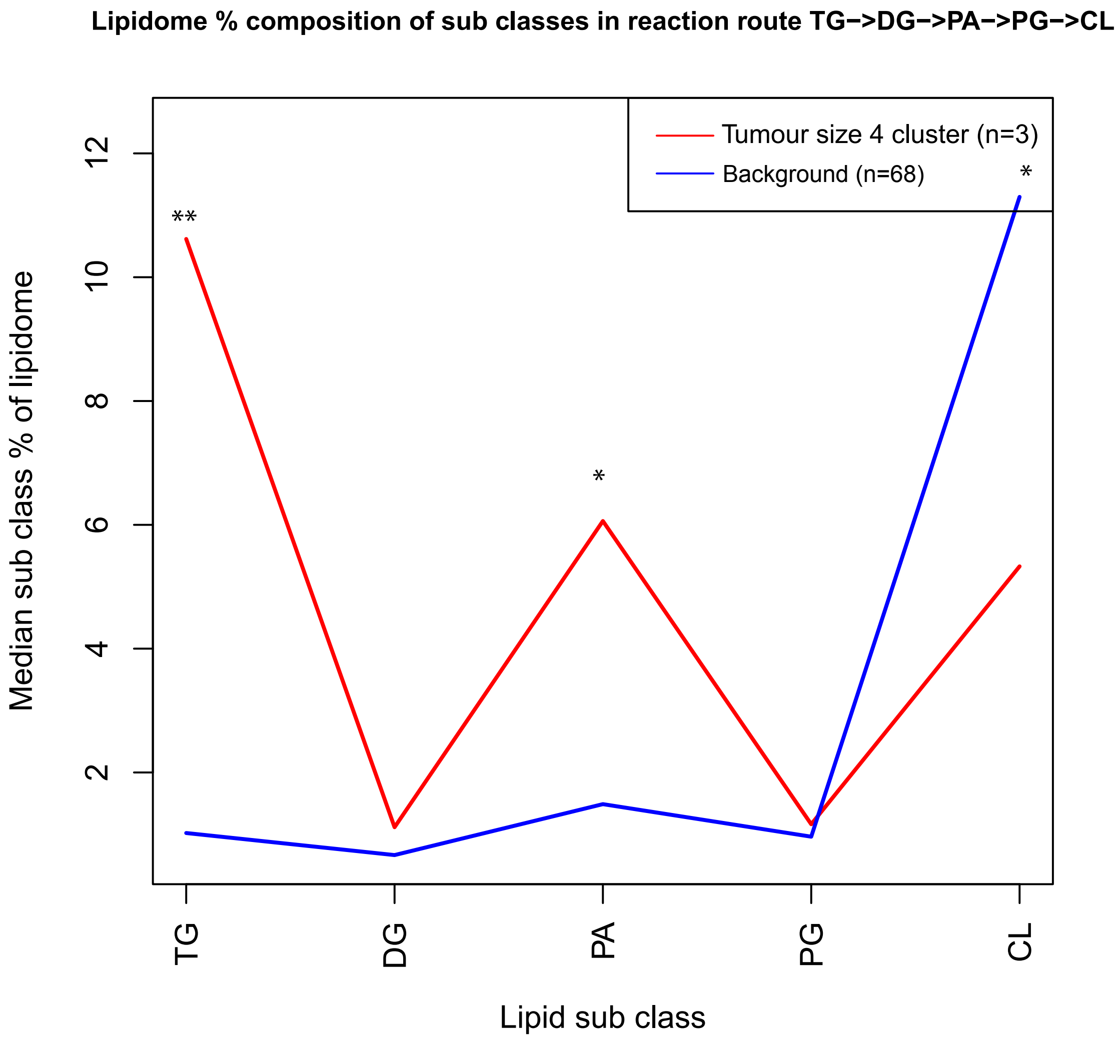


Figure 4.27: The median lipidome % sub class composition for the cluster of tumour samples highlighted in red on figure 4.26 (red line) and the remaining samples as the background (blue line). Significant differences are shown between PA and CL at the 5% level (indicated by "*") and TG at the 1% level (indicated by "**").

Duke's stages C1 and C2 was defined by an increase in PE at the sub class level. At the species level this pattern was no longer supported, this may be the product of insignificant increases in all PE species that accumalted into a significant change at the sub class level. Significant decreases were shown in the cholesterol esters CE 16:0, CE 18:2 and CE 18:3 and the ceramides Cer 20:0 and DHCer 18:0. Also, there were multiple significantly increased long chain triacylglycerols, namely TG 60:0, TG 60:4, TG 62:1, TG 62:3 and TG 62:4 visible in figure 4.31. The final stage comparison was between Dukes stages C2 and D seen in figure 4.32, the decrease in PE was significant in the species PE 32:1, PE 34:0, PE 34:2 and PE 38:3. Other lipid species significantly increased between these stages, including DHCer 18:0, PI 38:5, PI 40:8 and TG 40:2.

4.3.5.2 Species desaturation

Different lipid species shared varying levels of desaturation in their fatty acids. Saturation level of fatty acids impacts the fluidity of membranes, some fatty acids also have specific downstream signalling roles. This investigation aimed to determine if the level of desaturation within a lipid sub class was statistically different between Duke's stages of tumour samples. For example: Are the PE lipid species that all share three double bonds a much larger proportion in adenoma than Duke's stage A samples? For each lipid sub class, lipid species were grouped by their number of double bonds e.g. PC 34:3, PC 36:3, PC 38:3 and PC 40:3 are all grouped because they all contain 3 double bonds. The percentage that a group of lipids represents of the total quantity of lipids in that sub class was calculated for each sample. Samples were separated into the Duke's stages they had been assigned to and adjacent stages compared to each other. Each individual group of species sharing the same number of double bonds within a lipid sub class underwent a Student's t-test between the samples in the two adjacent Duke's stages. Lipid sub classes which showed a large number of double bond groups with significant differences of their percentage contribution to the entire sub class between the two stages, represent biologically interesting cases, where the level of desaturation is potentially linked to the progression of cancer. Figure 4.33 shows the comparison of cardiolipins between adenoma and Duke's A samples. There was a significant increase in tumour in the

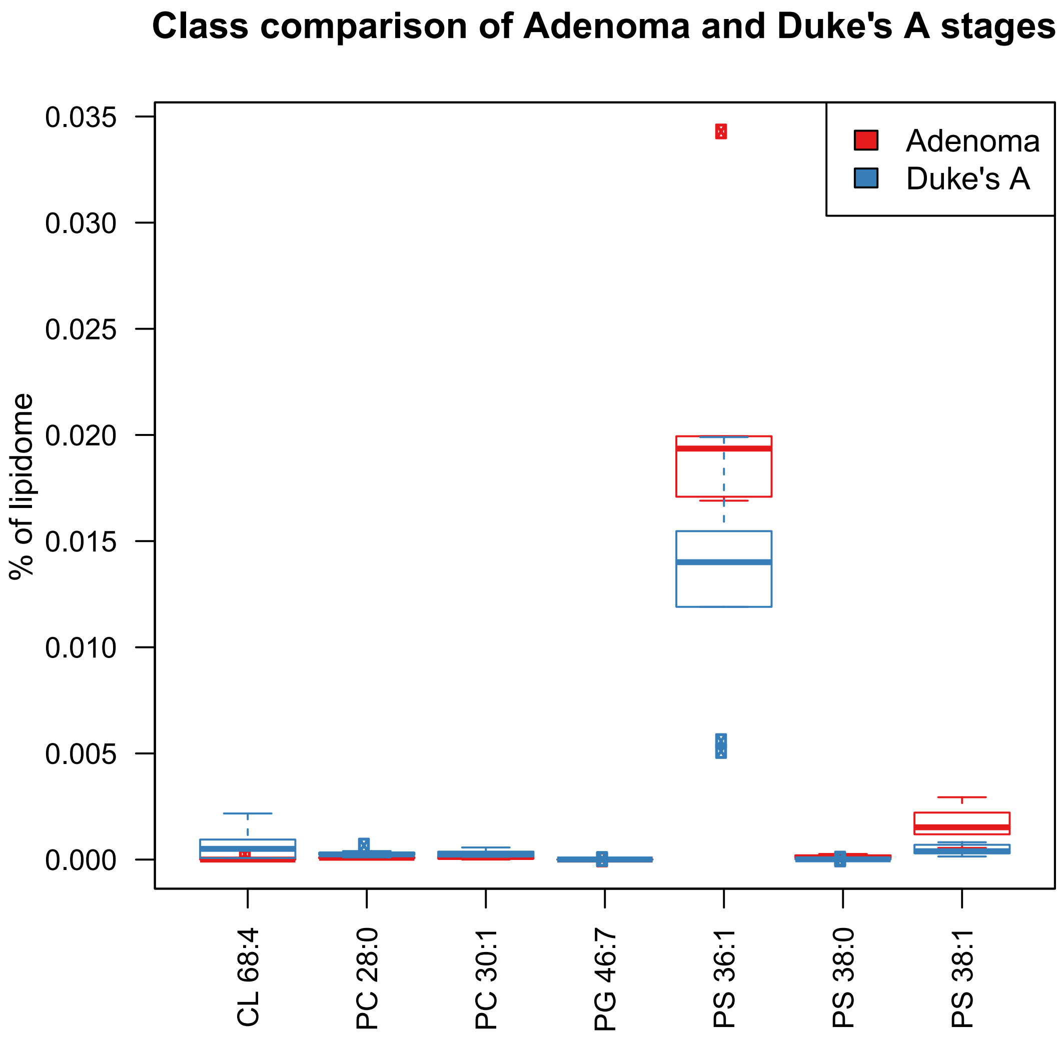


Figure 4.28: A set of box plots showing lipid species with significant differences at the 5% level (based upon Student's t-test) between Adenoma and Duke's stage A tumour samples.

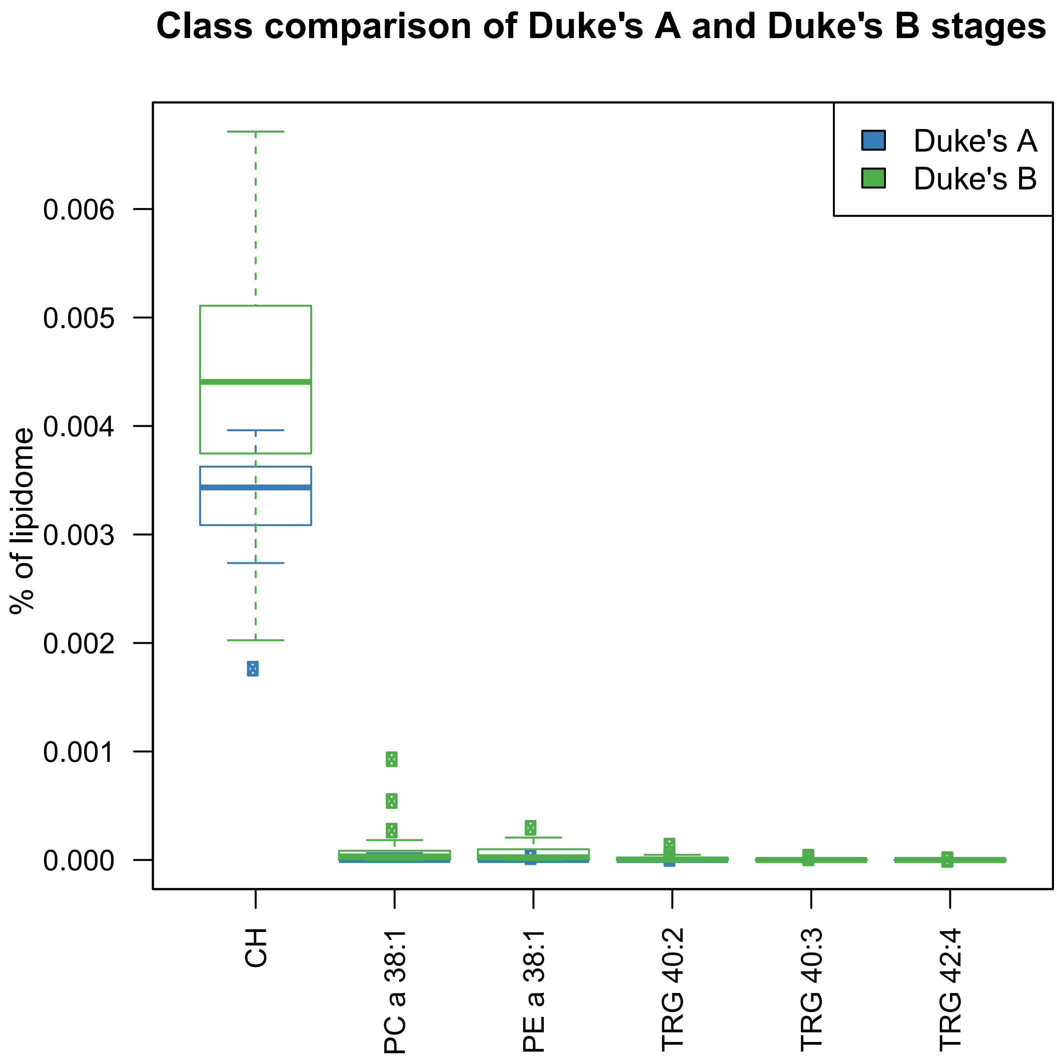


Figure 4.29: A set of box plots showing lipid species with significant differences at the 5% level (based upon Student's t-test) between Duke's stage A and Duke's stage B tumour samples.

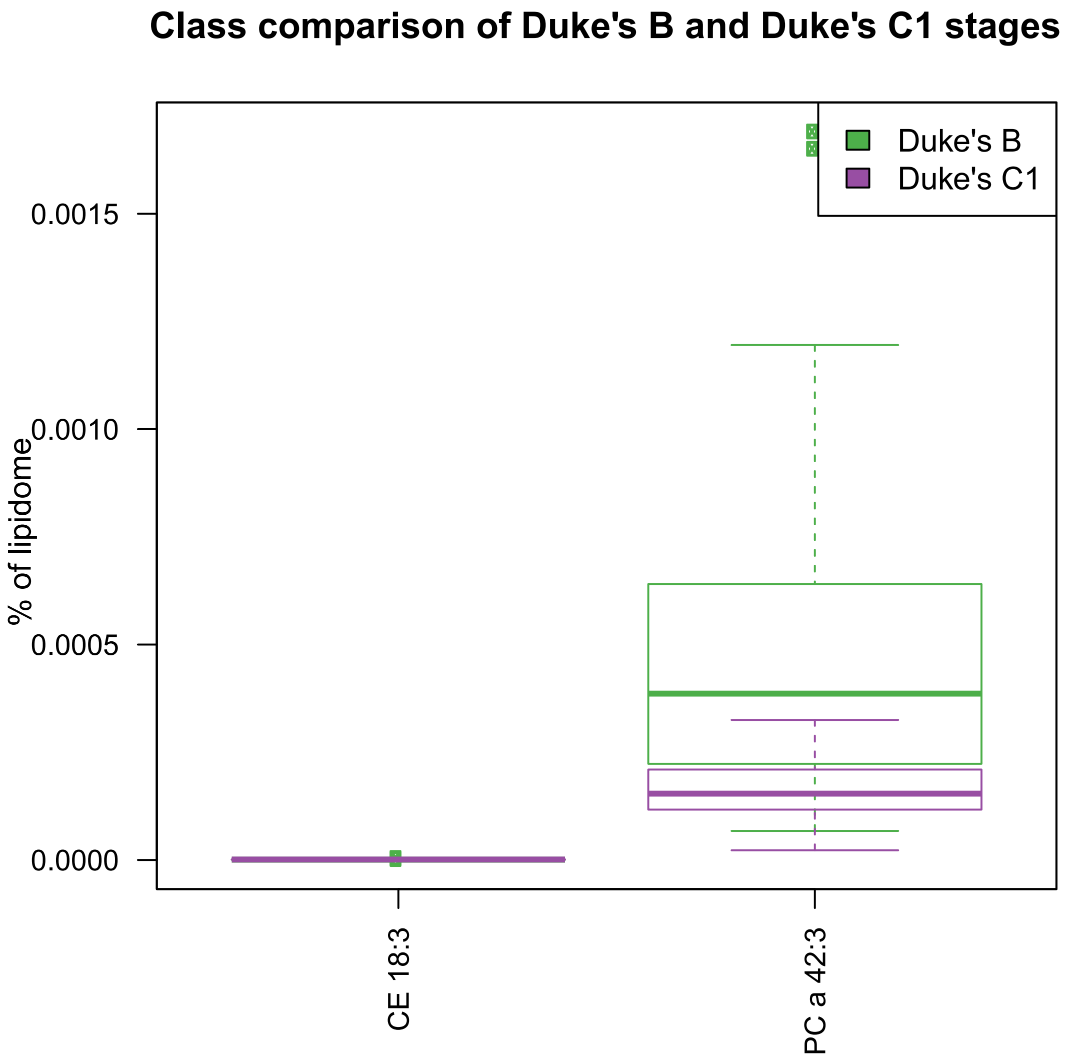


Figure 4.30: A set of box plots showing lipid species with significant differences at the 5% level (based upon Student's t-test) between Duke's stage B and Duke's stage C1 tumour samples.

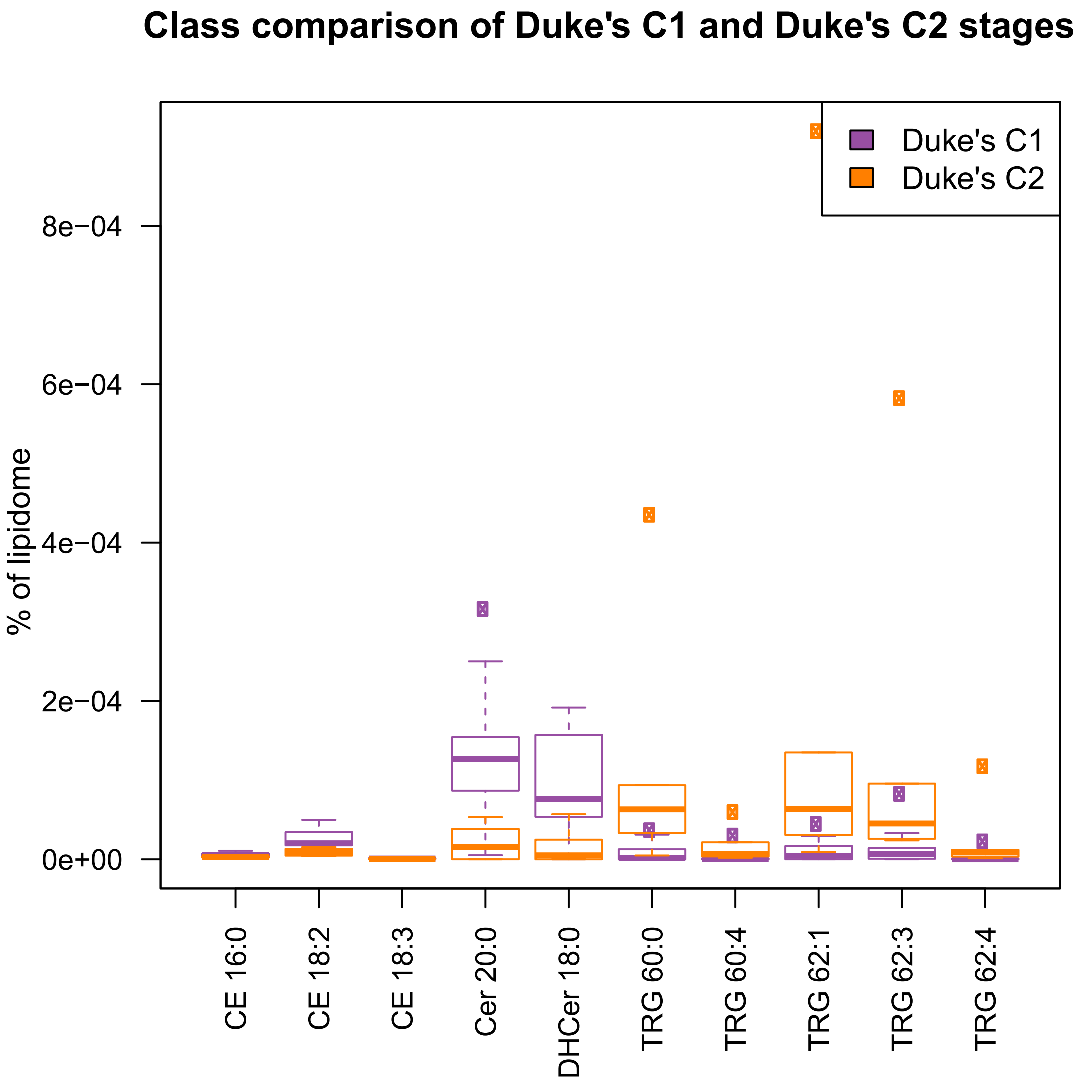


Figure 4.31: A set of box plots showing lipid species with significant differences at the 5% level (based upon Student's t-test) between Duke's stage C1 and Duke's stage C2 tumour samples.

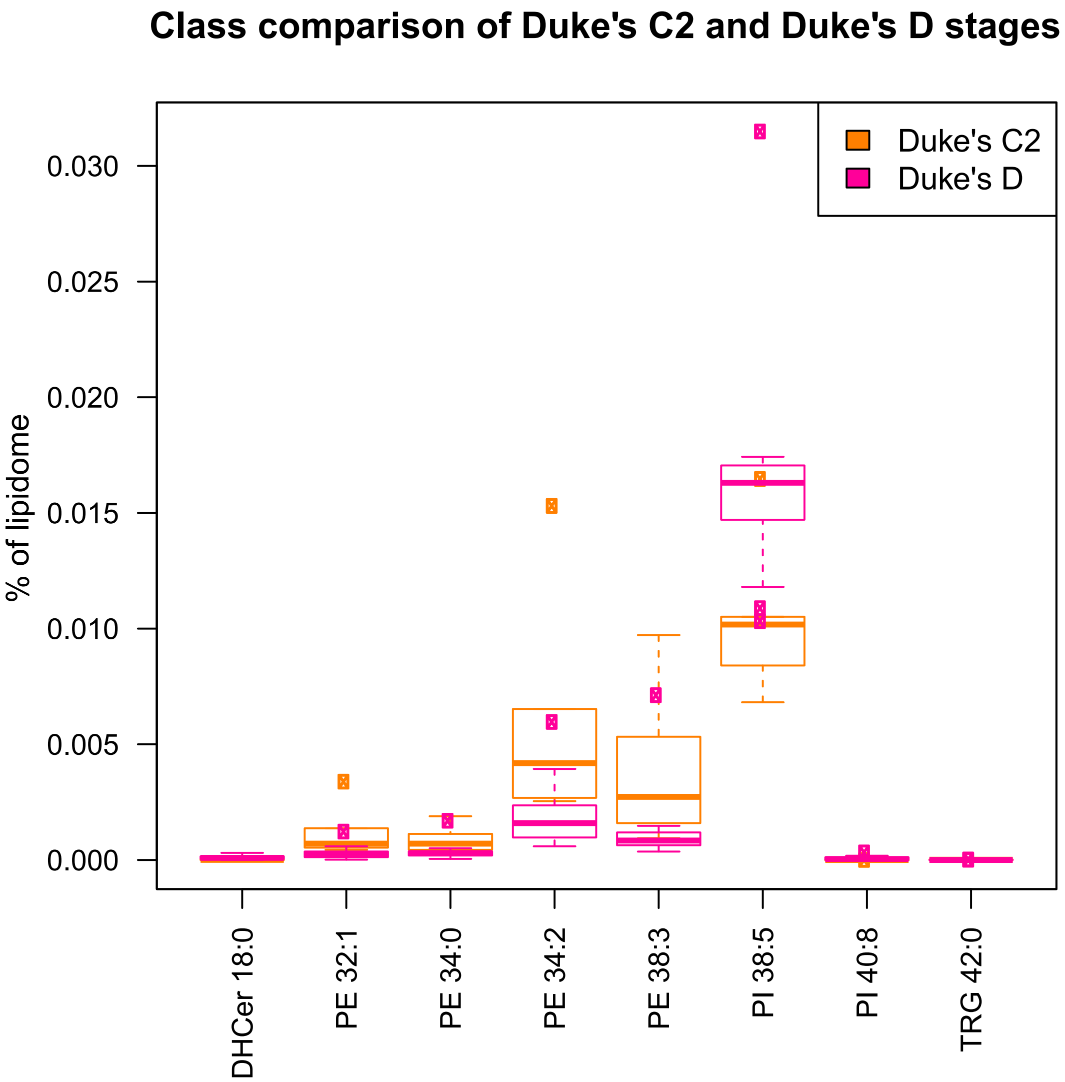


Figure 4.32: A set of box plots showing lipid species with significant differences at the 5% level (based upon Student's t-test) between Duke's stage C2 and Duke's stage D tumour samples.

quantity of species with four and five double bonds and a significant decrease in the quantity of species with seven and eight double bonds. Comparison of Duke's A to Duke's B tumour samples in figure 4.34 reveals a significant increase in 'PE a' lipid species with zero and one double bonds and the reverse for lipid species with 2 and 3 double bonds. Figure 4.35 shows that these stages also had a significant increase in the quantity of zero, two and four double bond species for the PS sub class and a significant decrease in species sharing seven and eight double bonds. The transition from Duke's C2 to Dukes D showed a significant decrease in PI lipid species sharing zero and one double bonds, and a significant increase in species sharing five, seven and eight double bonds, visible in figure 4.36.

4.3.6 Lipid generation analysis

Following on from the concept of theoretical lipid generation from a pool of free fatty acids and glycerophospho-headgroup units described in section 3.2.1, an analysis to measure how different lipid sub classes were related to the free fatty acid pool was performed. The majority of lipid species in the dataset were from diradyl sub classes (i.e. lipid species belonging to the sub class having two attached fatty acids). Cholesterol esters have a single fatty acid chain and so can be used to model the quantitative distribution of the free fatty acid pool in samples. This assumed that the synthesis of cholesterol esters is a process dependent entirely upon the concentration of the various fatty acids and is in no way preferentially selective for particular fatty acids. A typical normalised CE distribution can be seen in figure 4.37 for a normal sample (patient number 80). Performing a Cartesian cross of the fatty acids present in these CE lipids provided a theoretical distribution of the expected diradyl lipid species, given the free fatty acid pool modelled on CE species. An example of the theoretical diradyl distribution generated from figure 4.37 is available in figure 4.38. The theoretical diradyl distribution represented a vector of predicted percentages each diradyl species should contribute to the lipidome. This vector was compared against the same distribution of each measured diradyl lipid sub class using Pearson correlation (see section 4.2.3.8). Repeated for all patients, each sub class had a correlation distribution comparing it to the theoretical expected lipid species distribution. Figure 4.39 shows each sub class correlation as a boxplot, higher

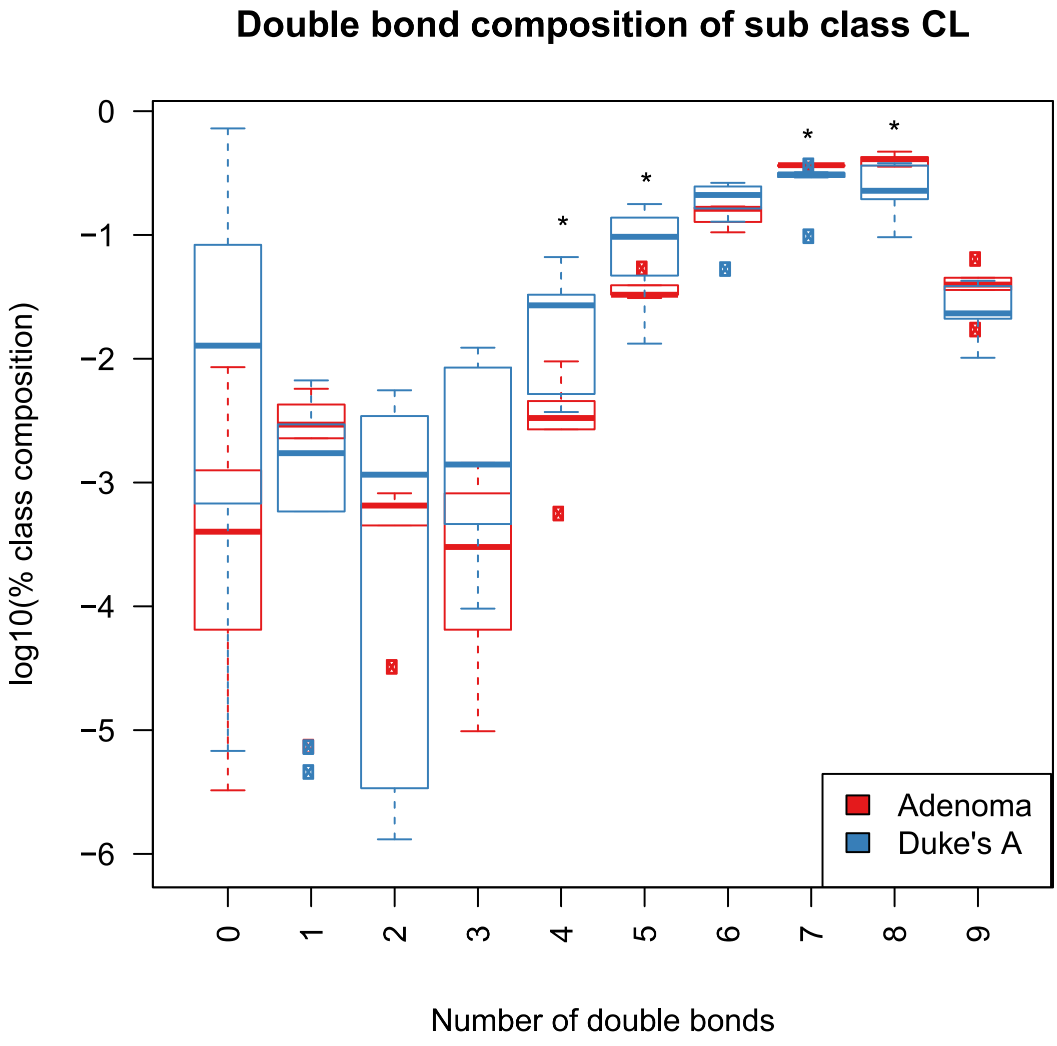


Figure 4.33: A comparison between the distribution of grouped species that share the same number of double bonds between adenoma and Duke's stage A tumour samples for the lipid sub class cardiolipin (CL). There was a significant increase at the 5% level of species sharing four and five double bonds (denoted by *). There was also a significant decrease of species sharing seven and eight double bonds in Duke's stage A tumour samples in comparison to adenoma.

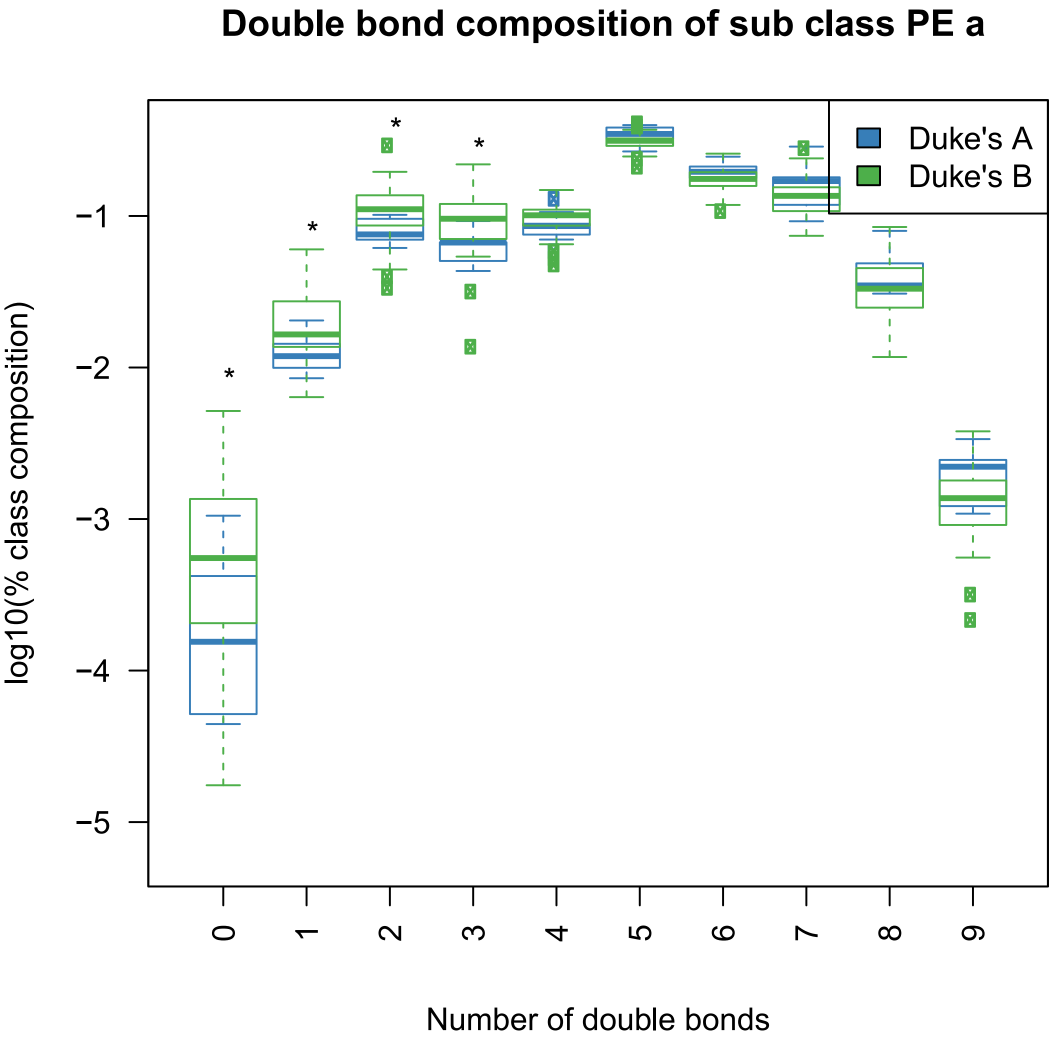


Figure 4.34: A comparison between the distribution of grouped species that share the same number of double bonds between Duke's stage A and Duke's stage B tumour samples for the lipid sub class 'PE a'. Lipid species sharing one, two, three and four double bonds show a significantly increased proportion of the sub class in Duke's B tumour samples.

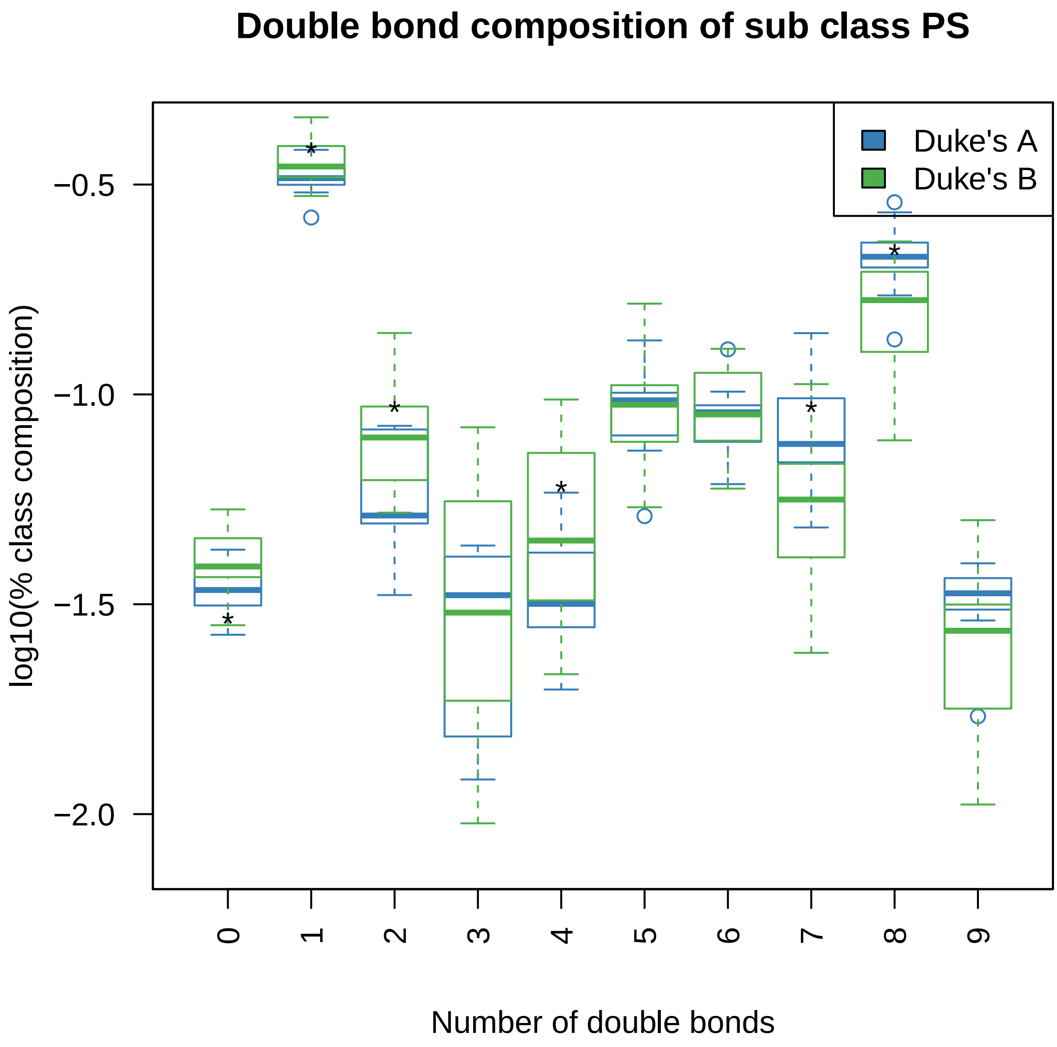


Figure 4.35: A comparison between the distribution of grouped species that share the same number of double bonds between Duke's stage A and Duke's stage B tumour samples for the lipid sub class 'PS'. Lipid species sharing one, three and four double bonds show a significantly increased proportion of the sub class in Duke's B tumour samples. The opposite is true of lipid species sharing seven and eight double bonds which show a decreased proportion of the total amount of PS in Duke's B tumour samples.

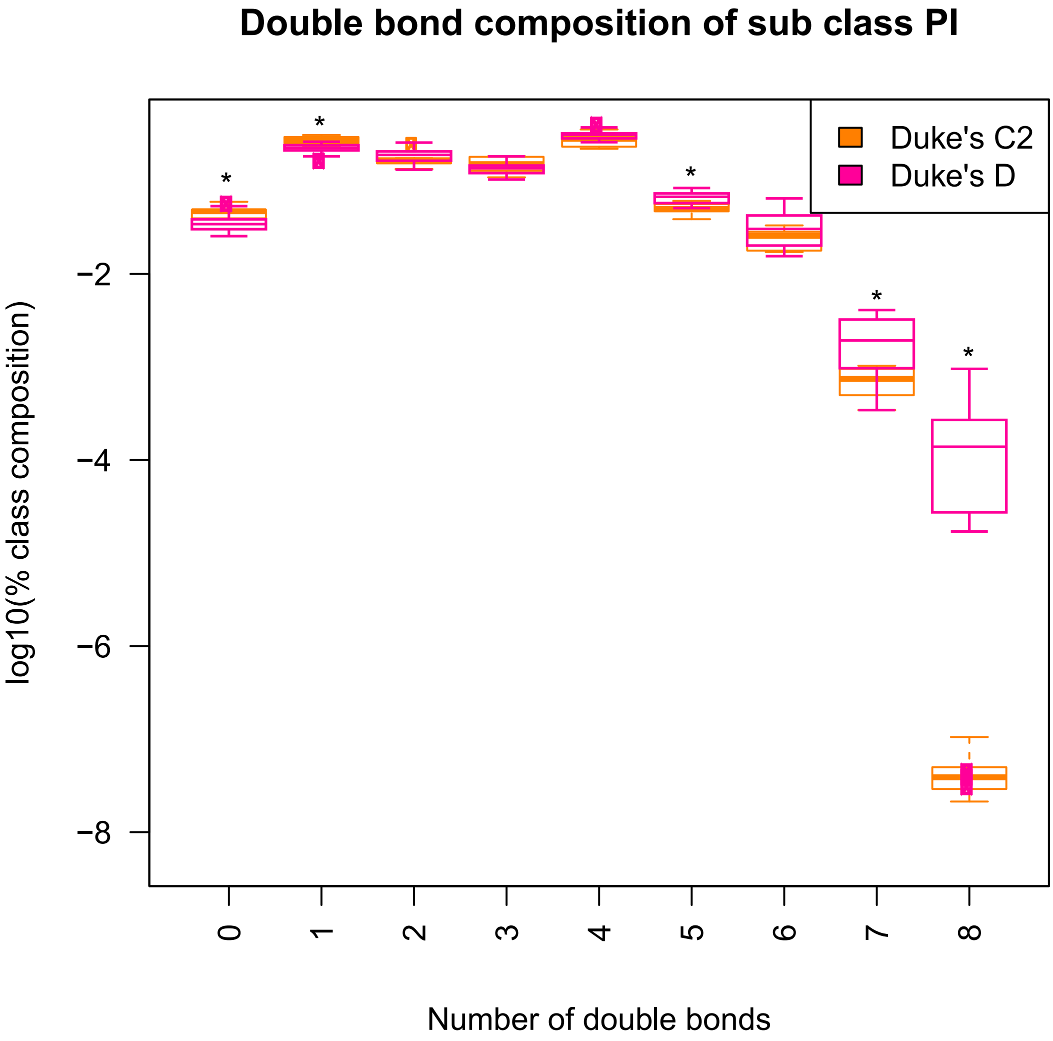


Figure 4.36: A comparison between the distribution of grouped species that share the same number of double bonds between Duke's stage C2 and Duke's stage D tumour samples for the lipid sub class 'PI'. Lipid species sharing one and two double bonds show a significantly increased proportion of the sub class in Duke's D tumour samples. The opposite is true of lipid species sharing five, seven and eight double bonds which show a decreased proportion of the total amount of PI in Duke's D tumour samples.

correlation represented more similarity to the predicted lipid species distribution, lower correlation represented less similarity and perhaps some biological processes that preferentially combined particular fatty acids to produce specific diradyl lipid species in a non-random way.

The results showed the sub class PS has much lower correlation to the theoretical diradyl species distribution. This suggested that the construction of PS species may not be a concentration dependent event, but a much more actively regulated event where particular fatty acids were targeted for PS construction or particular other glycerophospholipids are targeted for head group substitution to create PS. The same can possibly be said of PA but to a lesser degree.

4.4 Discussion

Throughout this chapter robust statistics have provided an insight into aspects of the extremely complex biology of colorectal cancer. Identifying specific components of the lipidome whose proportions alter, to distinguish the differences between normal and tumour samples. Also the transition between adjacent Duke's stages and the characteristic features of sample metadata annotations such as tumour size. From basic principles, a simplified reaction network was constructed of the lipid sub classes measured in this dataset. This network was deconstructed into reactions; each of which was probed for differences in reactant to product ratios between normal and tumour sample to reveal potentially interesting reactions and the machinery that regulates them.

The field of lipidomics is only recently coming in to the limelight of high-throughput biological sciences, generating huge quantities of quantitative lipid data at the species structural level. Many of the previous experiments in the field have targeted a few species specifically or a larger group of lipids at the sub class only level. In the future more datasets will appear in the literature where the analysis of large numbers of lipid species over many samples is done in tandem. This makes for an extremely exciting topic of research, unfortunately complicated by a number added difficulties. These difficulties include a lack of surrounding literature with which to corroborate the findings of statistical tests and transform them into biologically comprehensible events in the cells that can be targeted for further research.

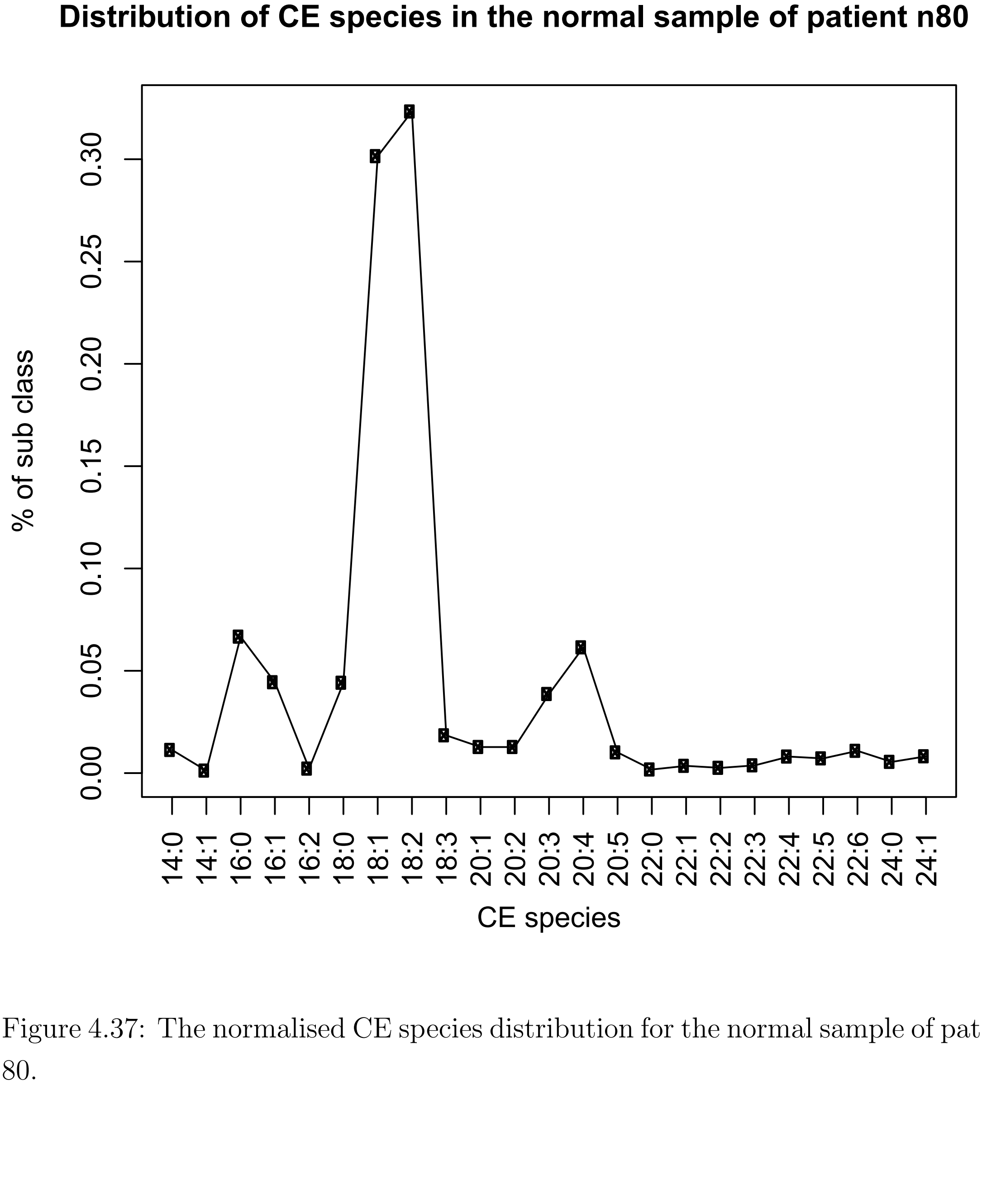


Figure 4.37: The normalised CE species distribution for the normal sample of patient 80.

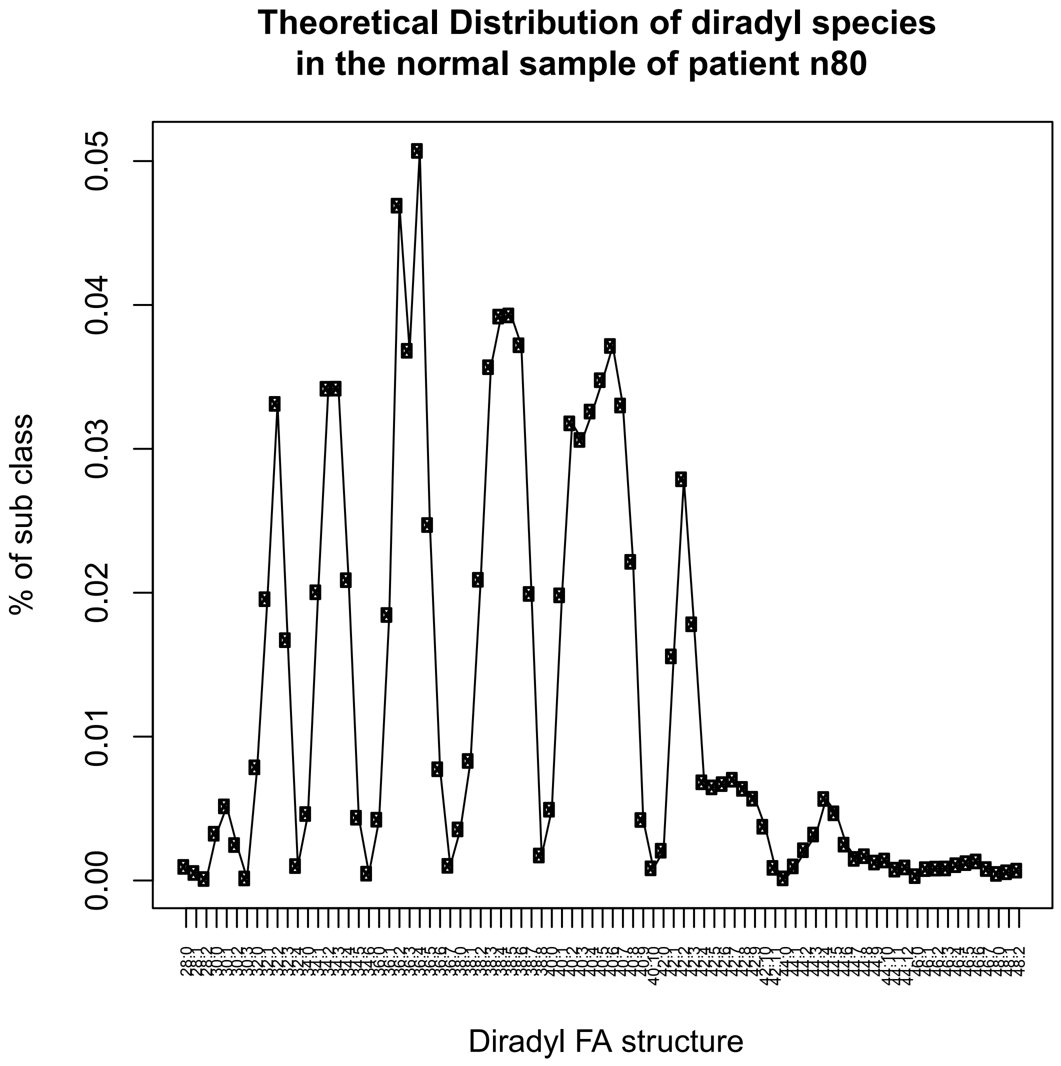


Figure 4.38: The theoretical diradyl species distribution for the normal sample of patient 80, calculated by a Cartesian cross of the CE species distribution shown in figure 4.37.

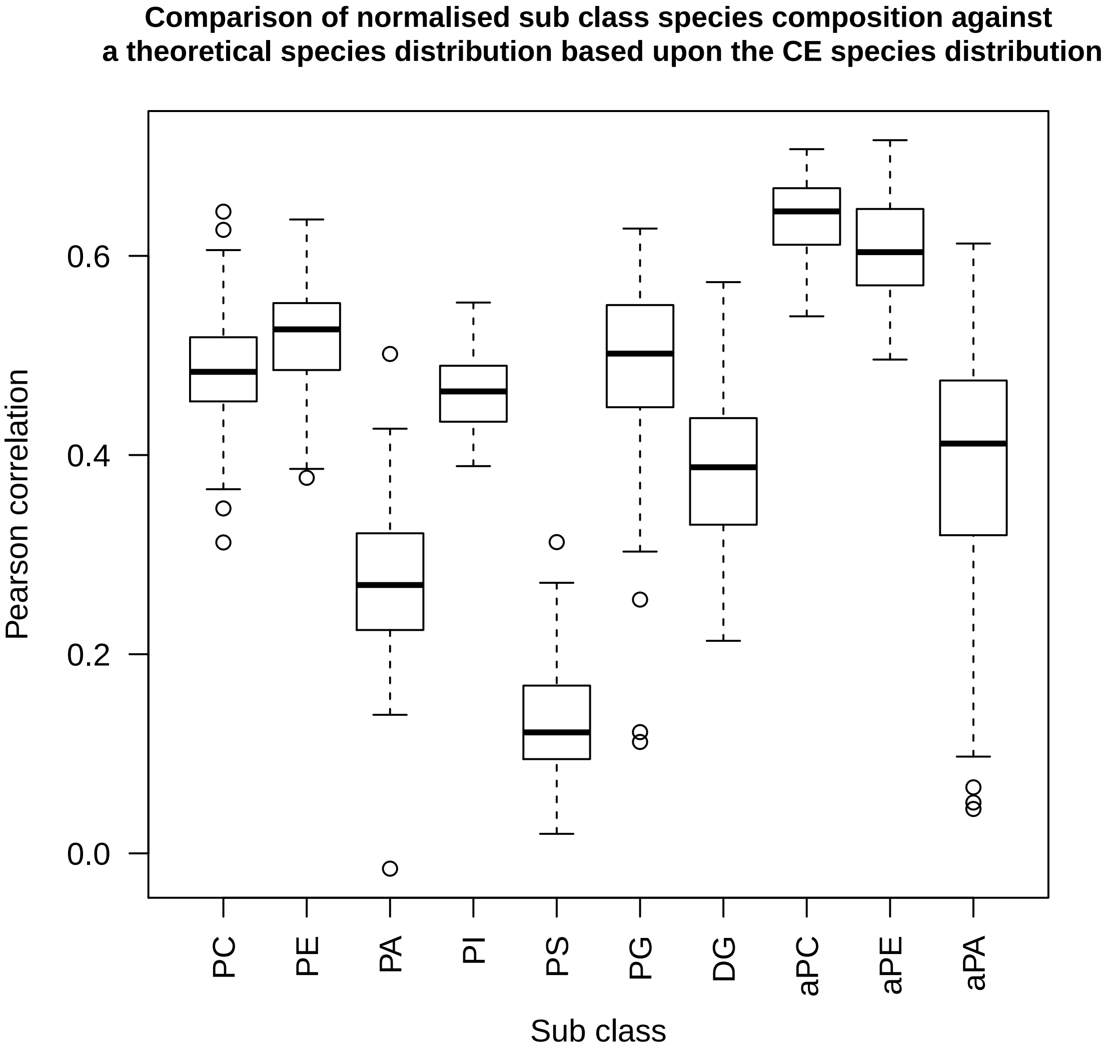


Figure 4.39: For each patient, the correlation between their specific theoretical diradyl distribution and the empirical lipid species distribution of a diradyl sub class is calculated. This process is repeated for each lipid sub class to produce a set of correlation distributions, each one displayed as a box plot. Lipid sub classes with lower correlation deviate more from the expected lipid species distribution based upon the fatty acids found in CE lipid species.

It is with caution that statistical significance must be associated with biological significance. Regardless of these difficulties, the following is an attempt to add some context to a selection of the results of the statistical analysis.

4.4.1 Glycerophosphoserine

Throughout the investigation glycerophosphoserine lipid species have shown a number of significant differences in tumour samples. In figure 4.8 short chain PS saturated species decrease in tumour samples while long chain desaturated species increase. PS plays a role in monocyte recognition of tumour cells, with 3-7 fold higher expression (Utsugi *et al.*, 1991). Hypoxia and inflammatory cytokines have also been shown to increase PS exposure on endothelial cell walls (Ran & Thorpe, 2002). This may be evidence that longer chain PS species are the ones exposed on the cell surface. This makes some conformational sense in that longer chain fatty acids will make a thicker bilayer and cause the membrane to fold. If enough of these longer chain species were localised together the longer chain PS species surrounded by shorter chain structural lipids would protrude very slightly from the membrane. This proposed protrusion would give slightly better access of the serine head group to biological entities that bind it, such as monocytes. This suggests that the overexpression of longer chain PS species is a defence mechanism by the cell to make itself an easier target for recognition by monocytes and downstream immune response. An independent discovery from figure 4.22 shows that PS species undergo a significant shift in their contribution to the total lipidome in the transition from tumour size 2 to tumour size 3. This transition is defined by the reduction of long chain PS species and the increase of short chain species. Expanding on the previous model of monocyte recognition, perhaps for a tumour to reach size 3 it must evade monocytes by expressing less long chain PS species on the cell surface.

4.4.2 Increased Total Lipid content

Figure 4.5 describes the significant increase of total lipid content of tumour samples in comparison to normal. Cancer cells in a growing tumour are forced to survive in a state of hypoxia as a response to the deprivation of oxygen owing to a lack of blood supply. To survive this challenge lipids and their generic components are stock piled. This is coordinated by hypoxia-inducible factor HIF-1 and epigenetic changes (Brahimi-Horn *et al.*, 2011). Tumour cells can divide extremely rapidly. A condition of this is the availability of structural components with which to build new cells, and since lipids form the majority of all membranes, this may also explain the increase in tumour samples.

For a comparative abundance of lipids to occur there must be upstream changes that facilitate it. Fatty acid synthase (FAS) is the enzyme responsible for the formation of long chain fatty acids from acetyl-CoA, malonyl-CoA and NADPH. Without fatty acids, glycerolipids and glycerophospholipids cannot be created and so this enzyme is likely to be one of those upregulated upstream components. It is known from the literature that when FAS is overexpressed in human cell lines, myristic acid (14:0) is a common exit point for the elongation process. The results of the Random Forest classifier highlighted PC 28:0 as being one of the most important lipid species in differentiating tumour and normal samples, as seen in figure 4.12. Referencing LipidHome (see chapter 3), the species PC 28:0 has been identified by the lipid sub species PC 14:0/14:0 and PC 10:0/18:0 automatically by the text mining process described in section 3.2.3.2. It is interesting to note that PC 14:0/14:0 is one of the few identified PC 28:0 sub species and this suggests that in tumour samples there is a significant over-expression/activity of FAS. In addition PC 28:0 was found to be one of the lipid species that shows a significant increase in its contribution to the total lipidome in the transition from adenoma to Duke's stage A. Increased activity of FAS may thus occur as early as the adenoma stage and might there differentiate adenomatous polyps that possess the ability to progress into a more severe carcinoma.

4.4.3 Glycerophosphoethanolamine

Phosphatidylethanolamine has shown interesting trends at both the sub class level and at the level of individual species. Figure 4.7 shows that PE exhibits a significant increase at the sub class level in the absolute amount of lipid in tumour compared to normal tissues. This pattern is also represented in the percentage that PE makes up of the total lipidome. Build-up of PE may well be a defence response by cells. Exogenous PE is known to induce cell apoptosis in human hepatoma Hep G2 cells via the bcl-2/bax pathway (Yao *et al.*, 2009). This is specifically shown in the stage analyses in the transition C1 to C2 to D where there is a significant increase in PE at the sub class level C1 to C2, followed by a decrease from C2 to D. This is shown in detail at the species level by figure 4.32, with a notable significant decrease in specific PE species from Duke's stage C2 samples to Duke's D stage samples. Increased PE concentration is a known feature of neoplastic tissues (Smith *et al.*, 2008). Cancer is a general example of neoplasticity but the transition from C1 to C2 which shows the marked increase of PE represents the invasion of the lymph system which requires a large amount of very specific cell proliferation. The results of the analysis preformed in section 4.3.4 gives clues to where the excess of PE may originate. The reactions in which PE is a product showed significant shifts in the product:reactant ratio to favour the build up of PE. These reactions were DG to PE, PS to PE and PC to PE. This suggests the additional PE is the result of funnelling mass through the pathway towards PE. The decline of DG has been previously reported in colonic adenomas (Sauter *et al.*, 1990) and this new evidence may suggest that it is diverted to PE.

4.4.4 Lysoglycerophospholipids

Alongside PC 28:0, the Lysoglycerophosphocholines (LPC) were quite prominent in the top 50 most important lipids in distinguishing between normal and tumour samples, visible in figure 4.12. Lysoglyerophospholipids have shown multiple effects in atherosclerosis and diabetes with particular effects on G-Protein Coupled Receptors (GPCRs). LPC generated by phospholipase A2 promotes inflammatory effects such as expression of growth factors and activation of monocytes (Kita *et al.*, 1999). Inflammation is a key side effect or even a cause of cancer so it is not outside the realm of possibility to apply a similar role of LPC in colorectal cancer.

4.4.5 Arachidonic acid

Arachidonic acid (AA) is a fatty acid with twenty carbons and four double bonds and has a well know role in cell signalling. Figure 4.20 shows the significant decrease in PI 38:4 in a group of tumour samples that represent the transition of tumour size 2 to 3. This species could be the sub species PI 18:0/20:4, a decrease in which could indicate a lack of cleavable species that produce AA which is known to regulate genes with roles in cell survival and apoptotic response (Monjazeb *et al.*, 2006). Alternatively, the reduced PI 38:4 may indicate the cleavage response has already occurred in the transition from tumour size 2 to 3 and the anti-cancer response has not been effective. PI is a precursor to the Phosphatidyinositolphosphate (PIP) series of second messengers, specifically PIP2, the major species of which are 38:4 in human T cells (82% and 70% respectively) (Haag *et al.*, 2012).

During the analysis of reactions between species in section 4.3.4.2, the concept of distinct layers within lipid pathways was introduced. The layer that accounts for all lipid species with the fatty acid structure 40:8 is significantly overrepresented among the list of reactions that show a significant difference in the reaction ratios of products and reactants between normal and tumour samples. These species may well be 20:4/20:4 and the perturbation in the reactions they take part in could thus be the result of phospholipase activity yielding AA for cell signalling purposes.

4.4.6 Future work

This work represents the foundation upon which a number of follow up experiments must be performed to elucidate the upstream events that define the perturbations to the lipidome described here. The dataset is of enormous scientific value and its potential has not yet been fully realised. With further work I aim to implement robust machine learning algorithms to better determine the exact changes in the lipidome that dictate the transition between Duke's stages. The adenoma group of samples represents an especially interesting set since adults over the age of 40 are expected to begin to develop harmless adenomatous polyps, some of which may advance to Duke's A stage of the cancer. A follow-up study could be designed to detect sub groups of adenomatous tissue in an unsupervised manner, in an attempt to develop a clinical diagnostic that can predict whether an adenoma is likely to advance to a more serious stage of cancer. However adenomas are the most difficult samples to acquire as they are often discovered during an exploratory colonoscopy but are not removed at the time due to their negligible health risk if not obstructively large.

Chapter 5

Conclusion

5.1 Contribution to the field

The work presented in this thesis has advanced the state of three fields, encompassing two 'omics' disciplines. Firstly chapter 2 demonstrated the standardised retrieval of publically available proteomics data from the PRIDE database via a newly developed R library. The remainder of the R library was dedicated to the development of statistical analyses for the detection of features within an experiment that indicate inherent data quality. One of the features included was provision of metadata, which led to an analysis of the provision of instrument type and software annotations over time in the PRIDE database. Spectral features such as MS2 m/z distribution were shown to find rare faults in the instrument setup. Analysing the m/z difference between the most abundant MS2 peaks revealed a novel method for the detection of common contaminants to mass spectrometry experiments such as PEG. This approach can also be transferred to the detection of other polymeric contaminants such as iron sulphur cluster.

The work on quality control was specifically aimed at analysing public data with the aim of reusing it for new purposes often with the aim to group large numbers of distinct experiments together to perform analyses that probe the field of proteomics as a whole. Chapter 2 also outlined the process of retrieving data from the PRIDE database using the PRIDE Inspector data visualisation tool. In this tool many of the quality control graphs were based upon earlier work laid out in this chapter and the accompanying paper. In addition, use of the QC tools was outlined for the detection of anomalies in experiment metadata and peptide/protein identifications, highlighting the plight of data curators and dangers of blindly using public data without first checking its consistency.

Following work on QC of data in proteomics databases, an overview of the state of bioinformatics resources in the field of lipidomics was summarised in Chapter 3. During this investigation it became apparent that there was not a 'sequence database' of theoretical lipids to act as a comprehensive dictionary of the physically feasible lipid molecules. In addition, existing resources were critiqued for the high resolution of structures stored, arguing that modern high throughput mass spectrometry based lipidomics is not identifying high resolution lipid molecules, only generalisations of the fine structural details provided by these resources. As such a resource focussed on serving the rapidly expanding mass spectrometry lipidomics community was undertaken. This involved the adaption of new nomenclature guidelines and the creation of a novel structural hierarchy with which to define the various levels of lipid identifications. Built around the structural hierarchy a database was designed to store lipids and associated metadata annotations. A novel approach to the generation of theoretical lipid structures was devised and implemented in the Java programming language. As a proof of concept a large set of glycerophospholipids were generated and stored in the database. To supplement the lipids a literature mining pipeline was implemented to automatically annotated lipids in the database with references to the literature. This paper annotation pipeline was designed to account for the discrepancies of lipid nomenclature and search whole abstracts for mention of lipids. Lipid records in existing resources were also cross-referenced in the database to mimic resource in other 'omics' fields with similar beneficial funcitonality. Significant work was attributed to the development of a web application to serve the information in the database graphically to interested data consumers. Intuitive web services were also implemented to enable simple computational access to the database for bioinformaticians. A new MS1 search engine was also developed to showcase the potential functionality of this resource and its strong ties to mass spectrometry. The database, web application and theoretical generation of lipids are new concepts to the field of lipidomics that have shown promise for future development.

The final contribution of the work described in this thesis is attributed to lipidomics again. Combining lessons learnt from the statistical analyses presented in the proteomics QC chapter 2 and data storage and standardisation principles from chapter 3, work was undertaken to explore a dataset of human colorectal cancer lipidomics measurements. Literature search revealed very few leads on standard approaches to mining the set of 71 patients each with over 500 distinct quantified lipids for both a tumour and adjacent, histologically matched normal samples. Analysis was largely exploratory, asking many questions of the data simultaneously. Focussing on patients, samples, stages, lipid species, lipid sub classes and lipid structural features the analysis was highly diverse. A novel approach to perform lipid network analysis was described providing insight into the flux of molecules through the pathway of known lipid inter-conversion reactions. Interesting differences between paired tumour and normal samples highlighted the key differences in the lipidome associated with colorectal cancer. The progression of tumours from harmless adenomas to critical Duke's stage D was investigated and stage-specific differences in their lipidome composition were found. Machine learning approaches were successfully applied to the problem of classifying normal and tumour samples and the key lipid features that underpin their distinction were extracted. The work represented a number of novel analysis methods applied to high throughput large scale lipidomics data. Not only new approaches but also new results that identify previously unknown perturbations to the colorectal cancer lipidome. In addition some light was shed on their significance and wider biological role.

In total the thesis provides a case study of transferal of knowledge and ideas from one 'omics' field to another and the development of those ideas into novel approaches to solve similar problems.

5.2 Opportunities for future work

Once completed, the work on proteomics quality control formed the basis of the quality control element of PRIDE Inspector. As new proteomics experiments and curation challenges arise within the PRIDE team there is potential to extend this work to help identify a larger portion of experiments with data integrity issues. For example the soon to be released PRIDE Q database contains peptide identifications of a predetermined high quality level. Work in the team at the moment is focussed on quantifying peptide spectrum match quality. Given the opportunity expertise gained during my work would definitely aid that process.

The main body of future work lies in the development of the 'LipidHome' web application, which is currently under discussion for a dedicated grant application to fund it. A more detailed description of future work for this project is available in 3.5 of which only a summary is described here. Increased coverage of lipids, specifically the inclusion of Glycerolipids, Sphingolipids and branched chain fatty acids will increase the potential user base. Improved automatic annotation of metadata, including a more robust paper annotation pipeline that accounts for more lipid synonyms and is more sensitive to false positive hits will make 'LipidHome' a first base for researching specific lipids. Cross referencing records in 'LipidHome' with other lipid databases and bioinformatics resources will be improved by writing individual cross referencing pipelines for each external resource rather than relying on harvesting all cross references from the LIPID MAPS Structural Database. The lipidomics community is the best source of lipid related information; if the process of submitting annotations can be made simple and rewarding, community sourced annotation could be a really pragmatic solution pioneered by the 'LipidHome' web application. 'LipidHome' was designed with flexibility and future development in mind, as such it is a solid platform on which to build new tools and integrate more data. Chemical descriptors provide useful estimates of predicted charge state, Van der Waals volume and the behaviour of biomolecules on chromatography columns, among many others. The most important feature of 'LipidHome' is its flexible and open source code base. Future work could be contributed by any number of interested developers, alternatively whole side projects can be stemmed from its data model and theoretical lipid generation code. An example lies in a library of lipid spectra to annotate lipid records with. These spectra could be experimentally derived or theoretically generated based upon lipid fragmentation and detection behaviour in a mass spectrometer. This resource could also act as the base of a high throughput spectral search algorithm for the quick and systematic identification of lipids for MS2 data, a well-established approach yet to be adopted in a widespread manner across the field of lipidomics. This work will soon be written into a set of papers describing, theoretical lipid generation and the ‘LipidHome’ web application, its use and content.

Finally the work on colorectal cancer lipidomics is in its infancy. New samples are regularly being added to boost the statistical power of patient groups and whole new classes of lipids are being analysed to shed new light on unexplored areas of the lipidome. However, with the current dataset the majority of future work will come in the from detailed orthogonal studies originating from non-mass spectrometry platforms to explain the list of interesting observations encountered so far. This may involve microarrays to detect the up or downregulation of key lipid metabolising enzymes, to identify a potential source for the build-up of specific lipids or groups of lipids. Quantitative proteomics could also be employed to analyse the patient samples to help correlate changes in the lipidome with the proteins that govern their inter-conversion and metabolism. Aside from extending the sample analysis to orthogonal platforms a comparison between other tumour data would be extremely valuable. Clearly the histology of different tissues complicates their direct comparison e.g. colorectal cancer and breast cancer. If a method of normalisation could be determined the ‘core’ lipids that define cancer’s perturbation to the lipidome could be detected and translated to generic biomarkers or drug targets for cancer therapy. The dataset described in chapter 4 is a “goldmine” of previously unknown and unexplored human colorectal cancer leads. As I have improved as an analyst it has yielded improved results and I expect it will continue to do so. As such I will continue to work with the dataset, applying new analyses on new research angles as I learn or invent them. This work will soon be written into a paper describing the unique dataset, the novel analyses performed and the conclusions that could be drawn.

5.3 Final word

Starting this PhD directly from a Molecular Biology Bachelors (Hons) degree from the University of Sheffield with little statistical or programming experience I have come a long way in a short while. I am extremely pleased with the scope, outcome and results of my work during the PhD and hope that this document is testament to my hard work. It seems the future inevitably involves the integration of ‘omics’ fields and studies encompassing multiple ‘omics’ fields becoming the norm. I hope that this work demonstrates that the hard work that underpins some ‘omics’ fields can readily be translated to others, for their mutual benefit. Additionally I hope the specific work demonstrated here and the discussion it has provoked will continue to benefit the fields of proteomics and lipidomics by providing cutting edge insight into their standardisation.

Appendix 1

Species	Identificaton 1	Identificaton 2	Identification 3
DG 24:0	Babraham		
DG 28:0	Espoo		
DG 28:1	Espoo		
DG 30:0	Espoo		
DG 30:1	Espoo		
DG 30:2	Espoo		
DG 32:0	Espoo	Babraham	Graz
DG 32:1	Espoo	Babraham	Graz
DG 32:2	Espoo	Babraham	Graz
DG 32:6	Babraham		
DG 34:0	Babraham	Graz	
DG 34:1	Espoo	Babraham	Graz
DG 34:2	Babraham	Espoo	Graz
DG 34:3	Babraham		
DG 34:4	Babraham	Graz	
DG 36:0	Babraham	Graz	
DG 36:1	Espoo	Graz	
DG 36:2	Espoo	Graz	
DG 36:3	Babraham	Graz	
DG 36:4	Babraham	Graz	
DG 36:5	Babraham		
DG 38:0	Graz		
DG 38:1	Graz		
DG 38:2	Babraham	Graz	
DG 38:3	Babraham	Graz	
DG 38:4	Babraham	Graz	
DG 38:5	Babraham	Graz	
DG 38:6	Babraham	Graz	
DG 38:7	Babraham		
DG 40:4	Babraham		
DG 40:5	Babraham		
DG 40:6	Babraham	Graz	
DG 40:7	Babraham	Graz	
DG 40:8	Babraham	Graz	
DG 42:1	Babraham		
DG 42:5	Babraham		
LPA 10:0	Espoo		
LPA 12:0	Espoo		
LPA 14:0	Espoo		

LPA 14:1	Espoo		
LPA 16:0	Espoo	Babraham	
LPA 16:1	Espoo		
LPA 16:2	Espoo		
LPA 17:0	Babraham		
LPA 18:0	Espoo	Babraham	
LPA 18:1	Espoo	Babraham	
LPA 18:2	Espoo		
LPA 22:0	Espoo		
LPA 24:0	Espoo		
LPA 26:0	Espoo		
LPA a 18:0	Babraham		
LPC 10:0	Espoo		
LPC 12:0	Espoo		
LPC 14:0	Espoo		
LPC 14:1	Espoo		
LPC 16:0	Espoo	Babraham	Graz
LPC 16:1	Espoo	Graz	
LPC 18:0	Espoo	Graz	
LPC 18:1	Espoo	Graz	
LPC 18:2	Graz		
LPC 20:0	Graz		
LPC 20:1	Graz		
LPC 20:3	Graz		
LPC 20:4	Graz		
LPC 22:5	Graz		
LPC 22:6	Graz		
LPC a 14:0	Babraham		
LPC a 14:1	Babraham		
LPC a 16:0	Babraham		
LPC a 16:1	Babraham		
LPC a 16:2	Babraham		
LPC a 18:0	Babraham		
LPC a 18:1	Babraham		
LPC a 18:2	Babraham		
LPC a 18:3	Babraham		
LPC a 20:0	Babraham		
LPC a 20:1	Babraham		
LPC a 20:2	Babraham		
LPC a 22:0	Babraham		
LPC a 22:1	Babraham		

LPC a 22:2	Babraham		
LPC a 24:0	Babraham		
LPC a 24:1	Babraham		
LPC a 24:2	Babraham		
LPC a 26:0	Babraham		
LPC a 26:1	Babraham		
LPC a 26:2	Babraham		
LPC a 26:3	Babraham		
LPE 10:0	Espoo		
LPE 12:0	Espoo		
LPE 14:0	Espoo		
LPE 14:1	Espoo		
LPE 16:0	Espoo		
LPE 16:1	Espoo		
LPE 18:0	Espoo		
LPE 18:1	Espoo		
LPE 22:0	Espoo		
LPE 24:0	Espoo		
LPE 26:0	Espoo		
LPI 12:0	Espoo		
LPI 14:0	Espoo		
LPI 14:1	Espoo		
LPI 16:0	Espoo		
LPI 16:1	Espoo		
LPI 18:0	Espoo		
LPI 18:1	Espoo		
LPI 22:0	Espoo		
LPI 24:0	Espoo		
LPI 26:0	Espoo		
LPS 10:0	Espoo		
LPS 12:0	Espoo		
LPS 14:0	Espoo		
LPS 14:1	Espoo		
LPS 16:0	Espoo		
LPS 16:1	Espoo		
LPS 18:0	Espoo		
LPS 18:1	Espoo		
PA 28:0	Espoo		
PA 28:1	Espoo		
PA 30:0	Babraham		
PA 30:1	Espoo		

PA 32:0	Espoo	Babraham	
PA 32:1	Espoo	Babraham	
PA 32:2	Espoo	Babraham	
PA 34:0	Babraham		
PA 34:1	Espoo	Babraham	
PA 34:2	Espoo	Babraham	
PA 34:3	Babraham		
PA 36:0	Babraham		
PA 36:1	Espoo	Babraham	
PA 36:2	Espoo	Babraham	
PA 36:3	Babraham		
PA 36:4	Babraham		
PA 36:5	Babraham		
PA 38:0	Babraham		
PA 38:1	Babraham		
PA 38:2	Babraham		
PA 38:3	Babraham		
PA 38:4	Babraham		
PA 38:5	Babraham		
PA 38:6	Babraham		
PA 40:3	Babraham		
PA 40:4	Babraham		
PA 40:5	Babraham		
PA 40:6	Babraham		
PA 40:7	Babraham		
PA 40:8	Babraham		
PA 42:6	Babraham		
PA a 34:0	Babraham		
PA a 34:1	Babraham		
PA a 34:2	Babraham		
PA a 34:3	Babraham		
PA a 36:0	Babraham		
PA a 36:1	Babraham		
PA a 36:2	Babraham		
PA a 36:3	Babraham		
PA a 36:4	Babraham		
PA a 38:4	Babraham		
PA a 38:5	Babraham		
PA a 38:6	Babraham		
PC 26:0	Babraham	Espoo	
PC 26:1	Espoo		

PC 28:0	Regensburg	Espoo	
PC 28:1	Espoo		
PC 30:0	Regensburg		
PC 30:1	Regensburg		
PC 32:0	Regensburg	Espoo	Graz
PC 32:1	Regensburg	Espoo	Graz
PC 32:2	Regensburg	Espoo	Graz
PC 32:3	Regensburg		
PC 34:0	Regensburg	Espoo	Graz
PC 34:1	Regensburg	Espoo	Graz
PC 34:2	Regensburg	Espoo	Graz
PC 34:3	Regensburg	Graz	
PC 34:4	Regensburg	Graz	
PC 36:0	Regensburg	Espoo	Graz
PC 36:1	Regensburg	Espoo	Graz
PC 36:2	Regensburg	Espoo	Graz
PC 36:3	Regensburg	Graz	
PC 36:4	Regensburg	Graz	
PC 36:5	Regensburg	Graz	
PC 36:6	Regensburg	Graz	
PC 38:0	Regensburg		
PC 38:1	Regensburg	Graz	
PC 38:2	Regensburg	Graz	
PC 38:3	Regensburg	Graz	
PC 38:4	Regensburg	Graz	
PC 38:5	Regensburg	Graz	
PC 38:6	Regensburg	Graz	
PC 38:7	Regensburg	Graz	
PC 40:0	Regensburg		
PC 40:1	Regensburg		
PC 40:2	Regensburg	Graz	
PC 40:3	Regensburg	Graz	
PC 40:4	Regensburg	Graz	
PC 40:5	Regensburg	Graz	
PC 40:6	Regensburg	Graz	
PC 40:7	Regensburg	Graz	
PC 42:0	Regensburg		
PC 42:4	Regensburg		
PC 42:5	Regensburg		
PC 44:0	Regensburg		
PC 44:1	Regensburg		

PC a 32:1	Babraham		
PC a 34:0	Babraham		
PC a 34:1	Babraham		
PC a 34:2	Babraham		
PC a 34:3	Babraham		
PC a 36:1	Babraham		
PC a 36:2	Babraham		
PC a 36:3	Babraham		
PC a 36:4	Babraham		
PC a 36:5	Babraham		
PC a 38:1	Babraham		
PC a 38:2	Babraham		
PC a 38:3	Babraham		
PC a 38:4	Babraham		
PC a 38:5	Babraham		
PC a 38:6	Babraham		
PC a 38:7	Babraham		
PC a 40:1	Babraham		
PC a 40:2	Babraham		
PC a 40:3	Babraham		
PC a 40:4	Babraham		
PC a 40:5	Babraham		
PC a 40:6	Babraham		
PC a 40:7	Babraham		
PC a 40:8	Babraham		
PC a 42:0	Babraham		
PC a 42:1	Babraham		
PC a 42:2	Babraham		
PC a 42:3	Babraham		
PC a 42:4	Babraham		
PC a 42:5	Babraham		
PC a 42:6	Babraham		
PC a 42:7	Babraham		
PC a 42:8	Babraham		
PC a 44:3	Babraham		
PC a 44:4	Babraham		
PC a 44:5	Babraham		
PC a 44:6	Babraham		
PC a 44:7	Babraham		
PC a 44:8	Babraham		
PC a 44:9	Babraham		

PC a 46:4	Babraham		
PC a 46:5	Babraham		
PC a 46:6	Babraham		
PC a 46:7	Babraham		
PC a 46:8	Babraham		
PE 24:0	Babraham		
PE 26:0	Babraham		
PE 28:0	Espoo		
PE 28:1	Espoo		
PE 30:0	Babraham		
PE 30:1	Espoo		
PE 32:0	Espoo	Babraham	Graz
PE 32:1	Espoo	Babraham	Graz
PE 32:2	Espoo	Graz	
PE 34:0	Babraham		
PE 34:1	Espoo	Babraham	Graz
PE 34:2	Espoo	Babraham	Graz
PE 34:3	Babraham		Graz
PE 34:4	Graz		
PE 36:0	Babraham	Graz	
PE 36:1	Espoo	Babraham	Graz
PE 36:2	Espoo	Babraham	Graz
PE 36:3	Babraham	Graz	
PE 36:4	Babraham	Graz	
PE 36:5	Graz		
PE 38:1	Babraham	Graz	
PE 38:2	Babraham	Graz	
PE 38:3	Babraham	Graz	
PE 38:4	Babraham	Graz	
PE 38:5	Babraham	Graz	
PE 38:6	Babraham	Graz	
PE 38:7	Graz		
PE 40:3	Babraham		
PE 40:4	Babraham	Graz	
PE 40:5	Babraham	Graz	
PE 40:6	Babraham	Graz	
PE 40:7	Babraham	Graz	
PE 40:8	Graz		
PE a 32:0	Babraham		
PE a 32:1	Babraham		
PE a 32:2	Babraham		

PE a 34:0	Babraham		
PE a 34:1	Babraham		
PE a 34:2	Babraham		
PE a 34:3	Babraham		
PE a 36:1	Babraham		
PE a 36:2	Babraham		
PE a 36:3	Babraham		
PE a 36:4	Babraham		
PE a 36:5	Babraham		
PE a 36:6	Babraham		
PE a 38:1	Babraham		
PE a 38:2	Babraham		
PE a 38:3	Babraham		
PE a 38:4	Babraham		
PE a 38:5	Babraham		
PE a 38:6	Babraham		
PE a 38:7	Babraham		
PE a 40:2	Babraham		
PE a 40:3	Babraham		
PE a 40:4	Babraham		
PE a 40:5	Babraham		
PE a 40:6	Babraham		
PE a 40:7	Babraham		
PE a 40:8	Babraham		
PE a 40:9	Babraham		
PE a 42:2	Babraham		
PE a 42:3	Babraham		
PE a 42:4	Babraham		
PE a 42:5	Babraham		
PE a 42:6	Babraham		
PE a 42:7	Babraham		
PE a 42:8	Babraham		
PE a 42:9	Babraham		
PE a 44:5	Babraham		
PE a 44:6	Babraham		
PE a 44:7	Babraham		
PE a 44:8	Babraham		
PE a 44:9	Babraham		
PE a 46:5	Babraham		
PE a 46:6	Babraham		
PE a 46:7	Babraham		

PE a 46:8	Babraham		
PE a 46:9	Babraham		
PG 24:0	Babraham		
PG 32:0	Babraham		
PG 32:1	Espoo	Babraham	
PG 32:2	Espoo		
PG 34:0	Babraham		
PG 34:1	Espoo	Babraham	
PG 34:2	Espoo	Babraham	
PG 34:3	Babraham		
PG 36:0	Babraham		
PG 36:1	Espoo	Babraham	
PG 36:2	Espoo	Babraham	
PG 36:3	Babraham		
PG 36:4	Babraham		
PG 36:5	Babraham		
PG 38:1	Babraham		
PG 38:2	Babraham		
PG 38:3	Babraham		
PG 38:4	Babraham		
PG 38:5	Babraham		
PG 38:6	Babraham		
PG 38:7	Babraham		
PG 40:4	Babraham		
PG 40:5	Babraham		
PG 40:6	Babraham		
PG 40:7	Babraham		
PG 40:8	Babraham		
PG 40:9	Babraham		
PG 42:6	Babraham		
PG 42:7	Babraham		
PG 42:8	Babraham		
PG 42:9	Babraham		
PG 44:4	Babraham		
PG 44:5	Babraham		
PG 46:3	Babraham		
PG 46:4	Babraham		
PG 46:5	Babraham		
PG 46:6	Babraham		
PG 46:7	Babraham		
PI 26:0	Espoo		

PI 26:1	Espoo		
PI 28:0	Espoo		
PI 28:1	Espoo		
PI 30:0	Espoo		
PI 30:1	Espoo		
PI 30:2	Espoo		
PI 32:0	Espoo	Babraham	Graz
PI 32:1	Espoo	Babraham	Graz
PI 32:2	Espoo	Graz	
PI 34:0	Babraham	Graz	
PI 34:1	Espoo	Babraham	Graz
PI 34:2	Espoo	Babraham	Graz
PI 34:3	Babraham	Graz	
PI 36:1	Espoo	Babraham	Graz
PI 36:2	Espoo	Babraham	Graz
PI 36:3	Babraham	Graz	
PI 36:4	Babraham	Graz	
PI 36:5	Babraham	Graz	
PI 38:1	Babraham		
PI 38:2	Babraham	Graz	
PI 38:3	Babraham	Graz	
PI 38:4	Babraham	Graz	
PI 38:5	Babraham	Graz	
PI 38:6	Babraham	Graz	
PI 38:7	Graz		
PI 40:3	Babraham	Graz	
PI 40:4	Babraham	Graz	
PI 40:5	Babraham	Graz	
PI 40:6	Babraham	Graz	
PI 40:7	Babraham	Graz	
PI 40:8	Babraham	Graz	
PS 24:0	Babraham		
PS 28:0	Espoo		
PS 28:1	Espoo		
PS 30:1	Espoo		
PS 32:0	Espoo		
PS 32:1	Espoo		
PS 32:2	Espoo		
PS 34:0	Babraham		
PS 34:1	Espoo	Babraham	
PS 34:2	Espoo	Babraham	

PS 36:0	Babraham		
PS 36:1	Espoo	Babraham	Graz
PS 36:2	Espoo	Babraham	Graz
PS 36:3	Babraham		
PS 36:4	Babraham	Graz	
PS 38:0	Babraham		
PS 38:1	Babraham		
PS 38:2	Babraham		
PS 38:3	Babraham		
PS 38:4	Babraham	Graz	
PS 38:5	Babraham		
PS 38:6	Graz		
PS 40:1	Babraham		
PS 40:2	Babraham		
PS 40:3	Babraham		
PS 40:4	Babraham		
PS 40:5	Babraham	Graz	
PS 40:6	Babraham	Graz	
PS 40:7	Babraham		
PS 42:4	Babraham		
PS 42:5	Babraham		
PS 42:6	Babraham		
PS 42:7	Babraham		
PS 42:8	Babraham		
PS 44:6	Babraham		
PS 44:7	Babraham		
PS 44:8	Babraham		
PS 44:9	Babraham		
PS 46:8	Babraham		
PS 46:9	Babraham		
TG 36:0	Babraham		
TG 38:0	Babraham		
TG 38:1	Babraham		
TG 38:2	Babraham		
TG 40:0	Babraham		
TG 40:1	Babraham		
TG 40:2	Babraham		
TG 40:3	Babraham		
TG 40:4	Babraham		
TG 42:0	Espoo	Babraham	
TG 42:1	Espoo	Babraham	

TG 42:2	Espoo	Babraham	
TG 42:3	Babraham		
TG 42:4	Babraham		
TG 44:0	Espoo	Babraham	
TG 44:1	Espoo	Babraham	
TG 44:2	Espoo	Babraham	
TG 44:3	Espoo	Babraham	
TG 44:4	Babraham		
TG 46:0	Espoo	Babraham	Graz
TG 46:1	Espoo	Babraham	Graz
TG 46:2	Espoo	Babraham	Graz
TG 46:3	Espoo	Babraham	Graz
TG 46:4	Babraham		
TG 46:5	Babraham		
TG 46:6	Babraham		
TG 48:0	Espoo	Babraham	Graz
TG 48:1	Espoo	Babraham	Graz
TG 48:2	Espoo	Babraham	Graz
TG 48:3	Espoo	Babraham	Graz
TG 48:4	Babraham	Graz	
TG 48:5	Babraham		
TG 48:6	Babraham		
TG 50:0	Espoo	Babraham	Graz
TG 50:1	Espoo	Babraham	Graz
TG 50:2	Espoo	Babraham	Graz
TG 50:3	Espoo	Babraham	Graz
TG 50:4	Babraham		Graz
TG 50:5	Babraham		Graz
TG 50:6	Babraham		
TG 50:7	Babraham		
TG 52:0	Espoo	Babraham	Graz
TG 52:1	Espoo	Babraham	Graz
TG 52:2	Espoo	Babraham	Graz
TG 52:3	Espoo	Babraham	Graz
TG 52:4	Babraham	Graz	
TG 52:5	Babraham	Graz	
TG 52:6	Babraham	Graz	
TG 52:7	Babraham	Graz	
TG 52:8	Babraham		
TG 54:0	Espoo	Babraham	
TG 54:1	Espoo	Babraham	Graz

TG 54:2	Espoo	Babraham	Graz
TG 54:3	Espoo	Babraham	Graz
TG 54:4	Babraham	Graz	
TG 54:5	Babraham	Graz	
TG 54:6	Babraham	Graz	
TG 54:7	Babraham	Graz	
TG 54:8	Babraham	Graz	
TG 56:0	Espoo	Babraham	
TG 56:1	Espoo	Babraham	Graz
TG 56:10	Babraham		
TG 56:2	Espoo	Babraham	Graz
TG 56:3	Babraham	Graz	
TG 56:4	Babraham	Graz	
TG 56:5	Babraham	Graz	
TG 56:6	Babraham	Graz	
TG 56:7	Babraham	Graz	
TG 56:8	Babraham	Graz	
TG 56:9	Babraham	Graz	
TG 58:0	Espoo	Babraham	
TG 58:1	Espoo	Babraham	Graz
TG 58:2	Espoo	Babraham	Graz
TG 58:3	Babraham	Graz	
TG 58:4	Babraham	Graz	
TG 58:5	Babraham	Graz	
TG 58:6	Graz		
TG 58:7	Graz		
TG 58:8	Graz		
TG 58:9	Graz		
TG 58:10	Graz		
TG 58:11	Graz		
TG 60:1	Espoo	Babraham	
TG 60:2	Espoo	Babraham	Graz
TG 60:3	Babraham	Graz	
TG 60:4	Babraham	Graz	
TG 60:5	Graz		
TG 60:6	Graz		
TG 60:7	Graz		
TG 60:8	Graz		
TG 60:9	Graz		
TG 60:10	Graz		
TG 60:11	Graz		

TG 60:12	Graz		
TG 62:0	Espoo	Babraham	
TG 62:1	Espoo	Babraham	
TG 62:2	Espoo	Babraham	
TG 62:3	Babraham		
TG 62:4	Babraham		
TG 62:5	Babraham		
TG 64:1	Babraham		
TG 64:2	Babraham		
TG 64:3	Babraham		

Table 1: A table lipid species identified by the LipidomicNet Consortium by mass spectrometry, and the individual institutes that identified them.

Fatty acid scan species	Identificaton 1	Identificaton 2
DG 12:0_16:0	Espoo	
DG 12:0_16:1	Espoo	
DG 14:0_16:0	Espoo	
DG 14:0_16:1	Espoo	
DG 14:1_16:1	Espoo	
DG 16:0_16:0	Espoo	
DG 16:0_16:1	Espoo	
DG 16:0_18:1	Espoo	
DG 16:1_16:1	Espoo	
DG 16:1_18:1	Espoo	
DG 18:0_18:1	Espoo	
DG 18:1_18:1	Espoo	
PA 12:0_16:0	Espoo	
PA 12:0_16:1	Espoo	
PA 14:0_16:1	Espoo	
PA 14:1_16:0	Espoo	
PA 16:0_16:0	Espoo	
PA 16:0_16:1	Espoo	
PA 16:0_18:1	Espoo	
PA 16:1_16:1	Espoo	
PA 16:1_18:1	Espoo	
PA 18:0_16:1	Espoo	
PA 18:0_18:1	Espoo	
PA 18:1_18:1	Espoo	
PC 10:0_16:0	Espoo	
PC 10:0_16:1	Espoo	
PC 12:0_16:0	Espoo	
PC 12:0_16:1	Espoo	
PC 14:0_16:0	Regensburg	
PC 14:0_16:0	Espoo	
PC 14:0_16:1	Regensburg	
PC 14:0_18:0	Regensburg	
PC 14:0_18:1	Regensburg	Espoo
PC 14:0_18:2	Regensburg	Espoo
PC 14:0_20:2	Regensburg	
PC 14:0_20:3	Regensburg	
PC 14:0_20:4	Regensburg	
PC 14:0_24:0	Regensburg	
PC 14:1_16:1	Espoo	

PC 16:0_20:4	Regensburg		
PC 16:0_14:0	Regensburg		
PC 16:0_16:0	Regensburg		
PC 16:0_16:1	Regensburg		
PC 16:0_18:0	Regensburg		
PC 16:0_18:1	Regensburg		
PC 16:0_18:2	Regensburg		
PC 16:0_18:3	Regensburg		
PC 16:0_18:4	Regensburg		
PC 16:0_20:0	Regensburg		
PC 16:0_20:1	Regensburg		
PC 16:0_20:2	Regensburg		
PC 16:0_20:3	Regensburg	Espoo	
PC 16:0_20:4	Regensburg	Espoo	
PC 16:0_20:5	Regensburg		
PC 16:0_22:0	Regensburg		
PC 16:0_22:1	Regensburg		
PC 16:0_22:4	Regensburg		
PC 16:0_22:5	Regensburg		
PC 16:0_22:6	Regensburg		
PC 16:0_24:1	Regensburg		
PC 16:0_24:2	Regensburg		
PC 16:0_24:4	Regensburg	Espoo	
PC 16:0_24:5	Regensburg		
PC 16:0_24:6	Regensburg		
PC 16:1_14:0	Regensburg		
PC 16:1_16:0	Regensburg		
PC 16:1_16:1	Espoo		
PC 16:1_18:0	Regensburg		
PC 16:1_18:1	Regensburg		
PC 16:1_18:2	Regensburg		
PC 16:1_18:3	Regensburg		
PC 16:1_20:0	Regensburg		
PC 16:1_20:1	Regensburg		
PC 16:1_20:2	Regensburg		
PC 16:1_20:3	Regensburg		
PC 16:1_20:4	Regensburg		
PC 16:1_22:0	Regensburg		
PC 16:1_22:1	Regensburg		
PC 16:1_22:4	Regensburg		
PC 16:1_22:5	Regensburg		

PC 16:1_22:6	Regensburg		
PC 16:1_24:0	Regensburg		
PC 16:1_24:1	Regensburg	Espoo	
PC 16:1_24:4	Regensburg	Espoo	
PC 16:1_24:5	Regensburg		
PC 16:1_24:6	Regensburg		
PC 16:3_20:3	Regensburg		
PC 18:0_14:0	Regensburg		
PC 18:0_16:0	Regensburg		
PC 18:0_16:1	Regensburg		
PC 18:0_18:0	Regensburg		
PC 18:0_18:1	Regensburg		
PC 18:0_18:2	Regensburg		
PC 18:0_18:3	Regensburg		
PC 18:0_18:4	Regensburg		
PC 18:0_18:5	Regensburg		
PC 18:0_20:0	Regensburg		
PC 18:0_20:1	Regensburg		
PC 18:0_20:2	Regensburg		
PC 18:0_20:3	Regensburg		
PC 18:0_20:4	Regensburg		
PC 18:0_20:5	Regensburg		
PC 18:0_20:6	Regensburg		
PC 18:0_22:1	Regensburg		
PC 18:0_22:2	Regensburg		
PC 18:0_22:3	Regensburg		
PC 18:0_22:4	Regensburg	Espoo	
PC 18:0_22:5	Regensburg		
PC 18:0_22:6	Regensburg		
PC 18:0_22:7	Regensburg		
PC 18:1_14:0	Regensburg		
PC 18:1_16:0	Regensburg		
PC 18:1_16:1	Regensburg		
PC 18:1_18:0	Regensburg		
PC 18:1_18:1	Regensburg		
PC 18:1_18:2	Regensburg		
PC 18:1_18:3	Regensburg		
PC 18:1_18:4	Regensburg		
PC 18:1_20:0	Regensburg		
PC 18:1_20:1	Regensburg		
PC 18:1_20:2	Regensburg		

PC 18:1_20:3	Regensburg	
PC 18:1_20:4	Regensburg	
PC 18:1_20:5	Regensburg	
PC 18:1_22:0	Regensburg	
PC 18:1_22:1	Regensburg	
PC 18:1_22:2	Regensburg	
PC 18:1_22:4	Regensburg	
PC 18:1_22:5	Regensburg	
PC 18:1_22:6	Regensburg	
PC 18:2_14:0	Regensburg	
PC 18:2_16:0	Regensburg	
PC 18:2_16:1	Regensburg	
PC 18:2_18:0	Regensburg	
PC 18:2_18:1	Regensburg	
PC 18:2_18:2	Regensburg	
PC 18:2_18:3	Regensburg	
PC 18:2_18:4	Regensburg	
PC 18:2_20:0	Regensburg	
PC 18:2_20:1	Regensburg	
PC 18:2_20:2	Regensburg	
PC 18:2_20:3	Regensburg	
PC 18:2_20:4	Regensburg	
PC 18:2_20:5	Regensburg	
PC 18:2_22:0	Regensburg	
PC 18:2_22:1	Regensburg	
PC 18:2_22:2	Regensburg	
PC 18:2_22:3	Regensburg	
PC 18:2_22:4	Regensburg	
PC 18:2_22:5	Regensburg	
PC 18:3_22:0	Regensburg	
PC 18:3_22:4	Regensburg	
PC 20:0_16:1	Regensburg	
PC 20:0_18:0	Regensburg	
PC 20:0_18:1	Regensburg	
PC 20:0_18:2	Regensburg	
PC 22:0_16:0	Regensburg	
PC 22:0_16:1	Regensburg	
PC 22:0_18:2	Regensburg	
PC 22:0_18:3	Regensburg	
PC 22:4_16:0	Regensburg	
PC 22:4_18:0	Regensburg	

PC 22:4_18:1	Regensburg	
PC 22:4_18:2	Regensburg	
PC 22:4_18:3	Regensburg	
PE 12:0_16:0	Espoo	
PE 12:0_16:1	Espoo	
PE 14:0_16:1	Espoo	
PE 14:1_16:0	Espoo	
PE 16:0_16:0	Espoo	
PE 16:0_16:1	Espoo	
PE 16:0_18:1	Espoo	
PE 16:1_16:1	Espoo	
PE 16:1_18:1	Espoo	
PE 18:0_16:1	Espoo	
PE 18:0_18:1	Espoo	
PE 18:1_18:1	Espoo	
PG 16:0_16:1	Espoo	
PG 16:0_18:1	Espoo	
PG 16:1_16:1	Espoo	
PG 16:1_18:1	Espoo	
PG 18:0_18:1	Espoo	
PG 18:1_18:1	Espoo	
PI 10:0_16:0	Espoo	
PI 10:0_16:1	Espoo	
PI 10:0_18:0	Espoo	
PI 12:0_14:0	Espoo	
PI 12:0_16:0	Espoo	
PI 12:0_16:1	Espoo	
PI 12:0_18:0	Espoo	
PI 14:0_14:0	Espoo	
PI 14:0_14:1	Espoo	
PI 14:0_16:0	Espoo	
PI 14:0_16:1	Espoo	
PI 14:0_18:1	Espoo	
PI 14:1_16:0	Espoo	
PI 14:1_16:1	Espoo	
PI 14:1_18:0	Espoo	
PI 16:0_16:0	Espoo	
PI 16:0_16:1	Espoo	
PI 16:0_18:1	Espoo	
PI 16:1_16:1	Espoo	
PI 16:1_18:1	Espoo	

PI 18:0_16:1	Espoo	
PI 18:0_18:1	Espoo	
PI 18:1_18:1	Espoo	
PS 12:0_16:0	Espoo	
PS 12:0_16:1	Espoo	
PS 14:0_16:1	Espoo	
PS 14:1_16:0	Espoo	
PS 16:0_16:0	Espoo	
PS 16:0_16:1	Espoo	
PS 16:0_18:1	Espoo	
PS 16:1_16:1	Espoo	
PS 16:1_18:1	Espoo	
PS 18:0_16:1	Espoo	
PS 18:0_18:1	Espoo	
PS 18:1_18:1	Espoo	
TG 12:0_12:0_18:1	Espoo	
TG 12:0_14:0_16:0	Espoo	
TG 12:0_14:0_16:1	Espoo	
TG 12:0_14:1_16:1	Espoo	
TG 12:0_14:1_22:0	Espoo	
TG 12:0_16:0_16:0	Espoo	
TG 12:0_16:0_16:1	Espoo	
TG 12:0_16:0_18:1	Espoo	
TG 12:0_16:1_16:1	Espoo	
TG 12:0_16:1_18:1	Espoo	
TG 12:0_16:1_22:0	Espoo	
TG 12:1_16:1_16:1	Espoo	
TG 14:0_16:0_16:0	Espoo	
TG 14:0_16:0_16:1	Espoo	
TG 14:0_16:0_18:1	Espoo	
TG 14:0_16:0_26:0	Espoo	
TG 14:0_16:1_16:1	Espoo	
TG 14:0_16:1_18:1	Espoo	
TG 14:0_16:1_22:0	Espoo	
TG 14:0_16:1_26:0	Espoo	
TG 14:0_18:0_18:1	Espoo	
TG 14:0_18:1_20:0	Espoo	
TG 14:0_18:1_22:0	Espoo	
TG 14:1_16:1_16:1	Espoo	
TG 14:1_16:1_18:1	Espoo	
TG 14:1_16:1_22:0	Espoo	

TG 16:0_16:0_16:0	Espoo	
TG 16:0_16:0_16:1	Espoo	
TG 16:0_16:0_18:0	Espoo	
TG 16:0_16:0_18:0	Espoo	
TG 16:0_16:0_26:0	Espoo	
TG 16:0_16:1_16:1	Espoo	
TG 16:0_16:1_18:0	Espoo	
TG 16:0_16:1_18:1	Espoo	
TG 16:0_16:1_22:0	Espoo	
TG 16:0_16:1_24:0	Espoo	
TG 16:0_16:1_26:0	Espoo	
TG 16:0_18:0_18:1	Espoo	
TG 16:0_18:0_22:0	Espoo	
TG 16:0_18:0_22:0	Espoo	
TG 16:0_18:0_26:0	Espoo	
TG 16:0_18:1_18:1	Espoo	
TG 16:0_18:1_20:0	Espoo	
TG 16:0_18:1_22:0	Espoo	
TG 16:0_18:1_24:0	Espoo	
TG 16:0_18:1_26:0	Espoo	
TG 16:1_16:1_16:1	Espoo	
TG 16:1_16:1_18:0	Espoo	
TG 16:1_16:1_18:1	Espoo	
TG 16:1_16:1_20:0	Espoo	
TG 16:1_16:1_22:0	Espoo	
TG 16:1_16:1_24:0	Espoo	
TG 16:1_16:1_26:0	Espoo	
TG 16:1_18:0_18:0	Espoo	
TG 16:1_18:0_18:1	Espoo	
TG 16:1_18:1_18:1	Espoo	
TG 16:1_18:1_20:0	Espoo	
TG 16:1_18:1_22:0	Espoo	
TG 16:1_18:1_24:0	Espoo	
TG 16:1_18:1_26:0	Espoo	
TG 18:0_18:0_18:0	Espoo	
TG 18:0_18:0_18:1	Espoo	
TG 18:0_18:0_26:0	Espoo	
TG 18:0_18:1_18:1	Espoo	
TG 18:0_18:1_22:0	Espoo	
TG 18:0_18:1_26:0	Espoo	
TG 18:1_18:1_18:1	Espoo	

TG 18:1_18:1_22:0	Espoo	
TG 18:1_18:1_26:0	Espoo	

Table 2: A table lipid fatty acid scan species identified by the LipidomicNet Consortium by mass spectrometry, and the individual institutes that identified them.

Fatty acid	Identificaton 1	Identificaton 2
10:0	Espoo	
12:0	Espoo	
12:1	Espoo	
14:0	Regensburg	Espoo
14:1	Espoo	
16:0	Regensburg	Espoo
16:1	Regensburg	Espoo
16:2	Espoo	
16:3	Regensburg	
18:0	Regensburg	Espoo
18:1	Regensburg	Espoo
18:2	Regensburg	Espoo
18:3	Regensburg	
18:4	Regensburg	
18:5	Regensburg	
20:0	Regensburg	Espoo
20:1	Regensburg	
20:2	Regensburg	
20:3	Regensburg	
20:4	Regensburg	
20:5	Regensburg	
20:6	Regensburg	
22:0	Regensburg	Espoo
22:1	Regensburg	
22:2	Regensburg	Espoo
22:3	Regensburg	
22:4	Regensburg	
22:5	Regensburg	
22:6	Regensburg	
22:7	Regensburg	
24:0	Regensburg	Espoo
24:1	Regensburg	
24:2	Regensburg	
24:4	Regensburg	
24:5	Regensburg	
24:6	Regensburg	
26:0	Espoo	

Table 3: A table lipid fatty acid species identified by the LipidomicNet Consortium by mass spectrometry, and the individual institutes that identified them.

.1 Webservices of LipidHome

The Webserice of the ‘LipidHome’ web application are a simple way to computationally retrieve data from the underlying database in a standardised manner. There are several different request URLs with different parameters and return values. However, they all share in common the data interchange format JSON as the return type. It is worth noting that the web services will become more advanced and evolve over time, as such it is worth keeping an eye on the help documentation in the main application hosted at http://www.ebi.ac.uk/apweiler-srv/lipidhome/. All web services are available under the root path http://www.ebi.ac.uk/apweiler-srv/lipidhome/service and the following section’s location should be appended to this root to access them.

.1.1 Category services

Category services all relate to the retrieval of information about category level lipids and their direct children; main classes. Available under /category there are three methods accessible;

.1.1.1 /summary

Takes a Long named ‘id’ as parameter, this is the database id of the category of interest. The method returns information about the specific category such as the number of main classes, sub classes and species that are its members. Example:http://www.ebi.ac.uk/apweiler-srv/lipidhome/service/category/summary?id=1.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.1.2 /list

This method takes no parameters and returns list of categories; names and ids. The method returns a list of all main classes that are members of the selected category.

.1.1.3 /mainclasses

Takes a Long named ‘id’ as parameter, this is the database id of the category of interest.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.2 Main class services

Main class services all relate to the retrieval of information about main class level lipids and their direct children; sub classes. Available under /mainclass there are two methods accessible;

.1.2.1 /summary

Takes a Long named ‘id’ as parameter, this is the database id of the main class of interest. The method returns information about the specific main class such as the number of sub classes, species and sub species that are its members. Example: http://www.ebi.ac.uk/apweiler-srv/lipidhome/service/mainclass/summary?id=1.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.2.2 /subclasses

Takes a Long named ‘id’ as parameter, this is the database id of the main class of interest. The method returns a list of all sub classes that are members of the selected category.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.3 Sub class services

Sub class services all relate to the retrieval of information about sub class level lipids and their direct children; species. Available under /subclass there are two methods accessible;

.1.3.1 /summary

Takes a Long named ‘id’ as parameter, this is the database id of the sub class of interest. The method returns information about the specific sub class such as the number of species, sub species and annotated isomers that are its members. Example: http://www.ebi.ac.uk/apweiler-srv/lipidhome/service/subclass/summary?id=1.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.3.2 /species

Takes a Long named 'id' as parameter, this is the database id of the sub class of interest. The method returns a list of all species that are members of the selected sub class.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.4 Species services

Species services all relate to the retrieval of information about species level lipids and their direct children; fatty acid scan species. Available under /specie there are two methods accessible;

.1.4.1 /summary

Takes a Long named 'id' as parameter, this is the database id of the species of interest. The method returns information about the specific species such as the number of sub species, annotated isomers that are its members and papers that mention it. Example: `http://www.ebi.ac.uk/apweiler-srv/lipidhome/service/specie/summary?id=1`.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.4.2 /fascanspecies

Takes a Long named 'id' as parameter, this is the database id of the species of interest. The method returns a list of all fatty acid scan species that are members of the selected species.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.5 Fatty acid scan species services

Fatty acid scan species services all relate to the retrieval of information about fatty acid scan species level lipids and their direct children; sub species. Available under /fasscanspecie there are two methods accessible;

.1.5.1 /summary

Takes a Long named ‘id’ as parameter, this is the database id of the fatty acid scan species of interest. The method returns information about the specific fatty acid scan species such as the number of sub species, annotated isomers that are its members and papers that mention it. Example: http://www.ebi.ac.uk/apweiler-srv/lipidhome/service/fascanspecie/summary?id=1.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.5.2 /subspecies

Takes a Long named ‘id’ as parameter, this is the database id of the fatty acid scan species of interest. The method returns a list of all sub species that are members of the selected fatty acid scan species.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.6 Sub species services

Sub species services all relate to the retrieval of information about sub species level lipids and their direct children; isomers. Available under /subspecie there are two methods accessible;

.1.6.1 /summary

Takes a Long named 'id' as parameter, this is the database id of the sub species of interest. The method returns information about the specific sub species such as the number of annotated isomers that are its members and papers that mention it. Example: http://www.ebi.ac.uk/apweiler-srv/lipidhome/service/subspecie/summary?id=1.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.6.2 /isomers

Takes a Long named 'id' as parameter, this is the database id of the sub species of interest. The method returns a list of all isomers that are members of the selected sub species. These isomers are both retrieved from the database (the identified ones harvested from other resources) and theoretically generated on the fly using the Chemical Development Kit.

Parameter	Values	Purpose
id	Any integer	Database record ID

.1.7 Tools services

Tools services relate to the tools panel of the web application where currently the only application hosted is an MS1 search engine. There are two methods available under /tools:

.1.7.1 /ms1search

Takes several parameters and the result is a simple JSON object, each element of which is a hit against the database. This service is a POST request and must be requested as such.

Parameter	Values
masses	"New line" separated list of masses
level	Any of "specie","faScanSpecie" or "subSpecie'
tolerance	Any float
identified	Boolean
adductions	String of "," separated integers. List available via the /adductions method.

.1.7.2 /isomers

Takes a Long named 'id' as parameter, this is the database id of the sub species of interest. The result is a downloadable file containing /ms1search results in the specified data format.

Parameter	Values
data	Json repsonse of /ms1search
format	any one of "CSV", "TSV", "Excel" and "XML"

.1.8 Utility services

Utility services is a catch all group of services necessary for varying aspect of the web application to run, most of them are application specific and uninteresting to interested data consumers. However the search functionality of the 'LipidHome' database is provided as a web service under /utils. The result is a list of items in the database that match the search text at the level specified and some basic information about them such as whether they are identified or not.

.1.8.1 /search

Takes several parameters;

Parameter	Values
query	Any String
type	Any of "CATEGORY", "MAIN_CLASS", "SUB_CLASS", "SPECIE", "FA_SCAN_SPECIE", "SUB_SPECIE", "ISOMER" AND "ALL"
start	Any positive integer
page	Any positive integer

Appendix 2

Name	Code
Cholesterol ester	CE
Ceramides	Cer
Cardiolipin	CL
Diacylglycerolipid	DG
Dihidroxyceramide	DHCer
Monoacylglycerophosphate	LPA
Monoalkylglycerophophate	LPA a
Monoalkylglycerophosphocholine	LPC
Monoacylglycerophosphocholine	LPC a
Monoacylglycerolipid	MG
Diacylglycerophosphate	PA
Monoacyl,monoalkylglycerophosphate	PA a
Diacylglycerophosphocholine	PC
Monoacyl,monoalkylglycerophosphocholine	PC a
Diacylglycerophosphoethanolamine	PE
Monoacyl,monoalkylglycerophosphoethanolamine	PE a
Diacylglycerophosphoglycerol	PG
Diacylglycerophosphoinositol	PI
Diacylglycerophosphoserine	PS
Sphingoid base 1-phosphate	S1P
Sphinganine	SG
Sphingomyelin	SM
Sphingosylphosphorylcholine	SPC
Triacylglycerolipid	TG

Table 4: A table of lipid sub classes identified by the Babraham colorectal cancer data and their abbreviations

Name	Name	Name	Name
C18 Sphingosine	LPC a 18:0	PC a 46:7	PS 36:2
CE 14:0	LPC a 18:1	PC a 46:8	PS 36:3
CE 14:1	LPC a 18:2	PE 26:0	PS 36:4
CE 16:0	LPC a 18:3	PE 30:0	PS 38:0
CE 16:1	LPC a 20:0	PE 32:0	PS 38:1
CE 16:2	LPC a 20:1	PE 32:1	PS 38:2
CE 18:0	LPC a 20:2	PE 34:0	PS 38:3
CE 18:1	LPC a 22:0	PE 34:1	PS 38:4
CE 18:2	LPC a 22:1	PE 34:2	PS 38:5
CE 18:3	LPC a 22:2	PE 34:3	PS 40:1
CE 20:1	LPC a 24:0	PE 36:0	PS 40:2
CE 20:2	LPC a 24:1	PE 36:1	PS 40:3
CE 20:3	LPC a 24:2	PE 36:2	PS 40:4
CE 20:4	LPC a 26:0	PE 36:3	PS 40:5
CE 20:5	LPC a 26:1	PE 36:4	PS 40:6
CE 22:0	LPC a 26:2	PE 38:1	PS 40:7
CE 22:1	LPC a 26:3	PE 38:2	PS 42:4
CE 22:2	MG 16:0	PE 38:3	PS 42:5
CE 22:3	MG 18:0	PE 38:4	PS 42:6
CE 22:4	MG 18:1	PE 38:5	PS 42:7
CE 22:5	MG 20:4	PE 38:6	PS 42:8
CE 22:6	PA 30:0	PE 40:3	PS 44:6
CE 24:0	PA 32:0	PE 40:4	PS 44:7
CE 24:1	PA 32:1	PE 40:5	PS 44:8
Cer 14:0	PA 32:2	PE 40:6	PS 44:9
Cer 16:0	PA 34:0	PE 40:7	PS 46:8
Cer 16:1	PA 34:1	PE a 32:0	PS 46:9
Cer 18:0	PA 34:2	PE a 32:1	S1P C18
Cer 18:1	PA 34:3	PE a 32:2	SM 12:0
Cer 20:0	PA 36:0	PE a 34:0	SM 14:0
Cer 20:1	PA 36:1	PE a 34:1	SM 16:0
Cer 22:0	PA 36:2	PE a 34:2	SM 16:1
Cer 22:1	PA 36:3	PE a 34:3	SM 18:0
Cer 24:0	PA 36:4	PE a 36:1	SM 18:1
Cer 24:1	PA 36:5	PE a 36:2	SM 20:0
Cer 24:2	PA 38:0	PE a 36:3	SM 20:1
CL 62:0	PA 38:1	PE a 36:4	SM 22:0
CL 64:0	PA 38:2	PE a 36:5	SM 22:1
CL 66:0	PA 38:3	PE a 36:6	SM 22:2

CL 66:3
CL 66:4
CL 66:5
CL 68:1
CL 68:2
CL 68:3
CL 68:4
CL 68:5
CL 68:6
CL 70:0
CL 70:1
CL 70:2
CL 70:3
CL 70:4
CL 70:5
CL 70:6
CL 70:7
CL 72:3
CL 72:4
CL 72:5
CL 72:6
CL 72:7
CL 72:8
CL 72:9
CL 74:10
CL 74:5
CL 74:6
CL 74:7
CL 74:8
CL 74:9
DG 32:0
DG 32:1
DG 32:2
DG 32:6
DG 34:0
DG 34:1
DG 34:2
DG 34:3
DG 34:4
DG 36:0
DG 36:1
PA 38:4
PA 38:5
PA 38:6
PA 40:3
PA 40:4
PA 40:5
PA 40:6
PA 40:7
PA 40:8
PA 42:6
PA a 34:0
PA a 34:1
PA a 34:2
PA a 34:3
PA a 36:0
PA a 36:1
PA a 36:2
PA a 36:3
PA a 36:4
PA a 38:4
PA a 38:5
PA a 38:6
PC 26:0
PC 28:0
PC 28:1
PC 30:0
PC 30:1
PC 32:0
PC 32:1
PC 32:2
PC 34:0
PC 34:1
PC 34:2
PC 34:3
PC 36:0
PC 36:1
PC 36:2
PC 36:3
PC 36:4
PC 36:5
PC 38:1
PE a 38:1
PE a 38:2
PE a 38:3
PE a 38:4
PE a 38:5
PE a 38:6
PE a 38:7
PE a 40:2
PE a 40:3
PE a 40:4
PE a 40:5
PE a 40:6
PE a 40:7
PE a 40:8
PE a 40:9
PE a 42:2
PE a 42:3
PE a 42:4
PE a 42:5
PE a 42:6
PE a 42:7
PE a 42:8
PE a 42:9
PE a 44:5
PE a 44:6
PE a 44:7
PE a 44:8
PE a 44:9
PE a 46:5
PE a 46:6
PE a 46:7
PE a 46:8
PE a 46:9
PG 32:0
PG 32:1
PG 34:0
PG 34:1
PG 34:2
PG 34:3
PG 36:0
PG 36:1
SM 24:0
SM 24:1
SM 24:2
SM 26:0
SM 26:1
SM 26:2
SPC C18
TG 38:0
TG 38:1
TG 38:2
TG 40:0
TG 40:1
TG 40:2
TG 40:3
TG 40:4
TG 42:0
TG 42:1
TG 42:2
TG 42:3
TG 42:4
TG 44:0
TG 44:1
TG 44:2
TG 44:3
TG 44:4
TG 46:0
TG 46:1
TG 46:2
TG 46:3
TG 46:4
TG 46:5
TG 46:6
TG 48:0
TG 48:1
TG 48:2
TG 48:3
TG 48:4
TG 48:5
TG 48:6
TG 50:0
TG 50:1

DG 36:2
DG 36:3
DG 36:4
DG 36:5
DG 38:2
DG 38:3
DG 38:4
DG 38:5
DG 38:6
DG 38:7
DG 40:4
DG 40:5
DG 40:6
DG 40:7
DG 40:8
DG 42:1
DG 42:5
DHCer 14:0
DHCer 16:0
DHCer 18:0
DHCer 22:0
DHCer 24:0
LPA 16:0
LPA 18:0
LPA 18:1
LPA 18:2
LPA a 18:0
LPC 14:0
LPC 16:0
LPC 16:1
LPC 18:0
LPC 18:1
LPC 18:2
LPC 18:3
LPC 20:0
LPC 20:1
LPC 20:2
LPC 20:3
LPC 20:4
LPC 20:5
LPC 22:1
PC 38:2
PC 38:3
PC 38:4
PC 38:5
PC 38:6
PC 38:7
PC 40:2
PC 40:3
PC 40:4
PC 40:5
PC 40:6
PC 40:7
PC 40:8
PC a 32:1
PC a 34:0
PC a 34:1
PC a 34:2
PC a 34:3
PC a 36:1
PC a 36:2
PC a 36:3
PC a 36:4
PC a 36:5
PC a 38:1
PC a 38:2
PC a 38:3
PC a 38:4
PC a 38:5
PC a 38:6
PC a 38:7
PC a 40:1
PC a 40:2
PC a 40:3
PC a 40:4
PC a 40:5
PC a 40:6
PC a 40:7
PC a 40:8
PC a 42:0
PC a 42:1
PC a 42:2
PG 36:2
PG 36:3
PG 36:4
PG 36:5
PG 38:1
PG 38:2
PG 38:3
PG 38:4
PG 38:5
PG 38:6
PG 38:7
PG 40:4
PG 40:5
PG 40:6
PG 40:7
PG 40:8
PG 40:9
PG 42:6
PG 42:7
PG 42:8
PG 42:9
PG 44:4
PG 44:5
PG 46:3
PG 46:4
PG 46:5
PG 46:6
PG 46:7
PI 32:0
PI 32:1
PI 34:0
PI 34:1
PI 34:2
PI 34:3
PI 36:0
PI 36:1
PI 36:2
PI 36:3
PI 36:4
PI 36:5
PI 38:1
TG 50:2
TG 50:3
TG 50:4
TG 50:5
TG 50:6
TG 50:7
TG 52:0
TG 52:1
TG 52:2
TG 52:3
TG 52:4
TG 52:5
TG 52:6
TG 52:7
TG 52:8
TG 54:0
TG 54:1
TG 54:2
TG 54:3
TG 54:4
TG 54:5
TG 54:6
TG 54:7
TG 54:8
TG 56:0
TG 56:1
TG 56:10
TG 56:2
TG 56:3
TG 56:4
TG 56:5
TG 56:6
TG 56:7
TG 56:8
TG 56:9
TG 58:0
TG 58:1
TG 58:2
TG 58:3
TG 58:4
TG 58:5

LPC 22:2
LPC 22:3
LPC 22:4
LPC 22:5
LPC 22:6
LPC 24:0
LPC 24:1
LPC 26:1
LPC 26:2
LPC 28:0
LPC 28:1
LPC a 14:0
LPC a 14:1
LPC a 16:0
LPC a 16:1
LPC a 16:2
PC a 42:3
PC a 42:4
PC a 42:5
PC a 42:6
PC a 42:7
PC a 42:8
PC a 44:3
PC a 44:4
PC a 44:5
PC a 44:6
PC a 44:7
PC a 44:8
PC a 44:9
PC a 46:4
PC a 46:5
PC a 46:6
PI 38:2
PI 38:3
PI 38:4
PI 38:5
PI 38:6
PI 40:3
PI 40:4
PI 40:5
PI 40:6
PI 40:7
PI 40:8
PS 34:0
PS 34:1
PS 34:2
PS 36:0
PS 36:1
TG 60:0
TG 60:1
TG 60:2
TG 60:3
TG 60:4
TG 62:0
TG 62:1
TG 62:2
TG 62:3
TG 62:4
TG 62:5
TG 64:1
TG 64:2
TG 64:3

Table 5: A table of lipid speciers identified by the Babraham colorectal cancer data set in the proposed LipidomicNet nomenclature.

Number	Stage	Operation Date	Gender	Size	Spread	Metastasis	Radiotherapy	Site
1368	Adenoma	09/02/2000	M	NA	NA	NA	0	Adenoma
1088	Adenoma	21/10/1998	M	NA	NA	NA	0	Adenoma
1402	Adenoma	Unknown	NA	NA	NA	NA	NA	NA
37	Adenoma	Unknown	NA	NA	NA	NA	NA	NA
75	Adenoma	Unknown	NA	NA	NA	NA	NA	NA
1380	Duke's A	20/06/2000	F	1	0	0	0	Sigmoid
68	Duke's A	22/04/1997	F	2	0	0	0	Rectal
69	Duke's A	29/04/1997	M	2	0	0	0	Rectal
1311	Duke's A	01/03/2000	M	2	0	0	0	Sigmoid
82	Duke's A	28/07/1997	M	2	0	0	0	Sigmoid
94	Duke's A	14/08/1997	M	2	0	0	0	Rectal
1337	Duke's A	02/12/1999	M	1	0	0	0	Adenoma
103	Duke's A	09/09/1997	F	2	0	0	0	Rectal
1058	Duke's A	17/03/1999	M	1	0	0	0	Rectal
1375	Duke's B	03/05/2000	F	3	0	0	0	Sigmoid
1364	Duke's B	02/02/2000	F	3	0	0	0	Rectal
1225	Duke's B	08/09/1999	F	3	0	0	0	Sigmoid
1226	Duke's B	08/09/1999	M	3	0	0	0	Rectal
1357	Duke's B	13/01/2000	M	3	0	0	0	Rectal
1392	Duke's B	07/11/2000	M	3	0	0	0	Caecal
1142	Duke's B	07/07/1999	M	2	0	0	0	Sigmoid
1383	Duke's B	19/07/2000	F	4	0	0	0	Caecal
1378	Duke's B	03/05/2000	F	3	0	0	0	Rectal
1379	Duke's B	22/06/2000	M	3	0	0	0	Rectal
71	Duke's B	22/05/1997	F	3	0	0	0	Caecal
67	Duke's B	22/04/1997	F	3	0	0	0	Sigmoid
77	Duke's B	01/07/1997	M	3	0	0	0	Caecal

79	Duke’s B	04/07/1997	M	3	0	0	0	Rectal
73	Duke’s B	10/06/1997	M	3	0	0	0	Rectal
1300	Duke’s B	03/11/1999	F	3	0	0	0	Rectal
84	Duke’s B	31/07/1997	M	3	0	0	0	Rectal
86	Duke’s B	05/08/1997	M	3	0	0	0	Rectal
1054	Duke’s B	09/03/1999	M	3	0	0	0	Caecal
95	Duke’s B	19/08/1997	F	3	0	0	0	Sigmoid
90	Duke’s B	09/08/1997	F	3	0	0	0	Rectal
100	Duke’s B	04/09/1997	F	4	0	0	0	Caecal
98	Duke’s B	02/09/1997	F	3	0	0	0	Rectal
97	Duke’s B	21/08/1997	F	3	0	0	0	Caecal
1073	Duke’s B	02/03/1999	M	3	0	0	0	Rectal
1366	Duke’s C1	09/02/2000	F	1	1	0	0	Colonic
1358	Duke’s C1	12/01/2000	M	3	1	0	0	Rectal
1234	Duke’s C1	15/09/1999	M	3	1	0	1	Rectal
72	Duke’s C1	06/06/1997	F	2	1	0	0	Rectal
85	Duke’s C1	05/08/1997	F	3	1	0	1	Rectal
81	Duke’s C1	15/07/1997	M	4	3	0	0	Rectal
92	Duke’s C1	12/08/1997	F	4	1	0	0	Colonic
89	Duke’s C1	09/08/1997	M	4	1	0	0	Rectal
96	Duke’s C1	21/08/1997	M	3	2	0	0	Rectal
1214	Duke’s C1	25/08/1999	M	3	1	0	0	Rectal
1215	Duke’s C1	25/08/1999	F	3	1	0	0	Caecal
1374	Duke’s C2	12/04/2000	M	4	2	0	0	Sigmoid
1382	Duke’s C2	29/06/2000	M	3	2	0	0	Sigmoid
74	Duke’s C2	12/06/1997	F	3	3	0	0	Rectal
1286	Duke’s C2	20/10/1999	F	4	2	0	0	Rectal
1287	Duke’s C2	21/10/1999	F	3	3	0	0	Rectal
1285	Duke’s C2	19/10/1999	F	3	1	0	0	Sigmoid

1361	Duke’s D	26/01/2000	F	4	1		1	Rectal
1356	Duke’s D	12/01/2000	M	3	1	1	0	Sigmoid
1233	Duke’s D	15/09/1999	M	3	2	1	0	Rectal
1245	Duke’s D	12/01/2000	M	3	0	1	0	Rectal
1386	Duke’s D	20/09/2000	M	3	2	1	0	Rectal
70	Duke’s D	01/05/1997	F	3	2	1	0	Rectal
1308	Duke’s D	08/12/1999	F	4	2	1	0	Rectal
87	Duke’s D	Unknown	NA	NA	NA	NA	NA	NA
80	Duke’s D	12/07/1997	F	3	1	1	0	Rectal
83	Duke’s D	30/07/1997	M	4	2	1	0	Caecal
93	Duke’s D	14/08/1997	F	4	1	1	0	Sigmoid
88	Duke’s D	07/08/1997	M	3	2	1	0	Sigmoid
91	Duke’s D	12/08/1997	M	3	1	1	0	Rectal
102	Duke’s D	09/09/1997	M	4	2	1	0	Caecal
1192	Duke’s D	30/06/1999	M	4	2	1	0	Caecal

Table 6: A table of patient metadata for the Babraham colorectal cancer data set.

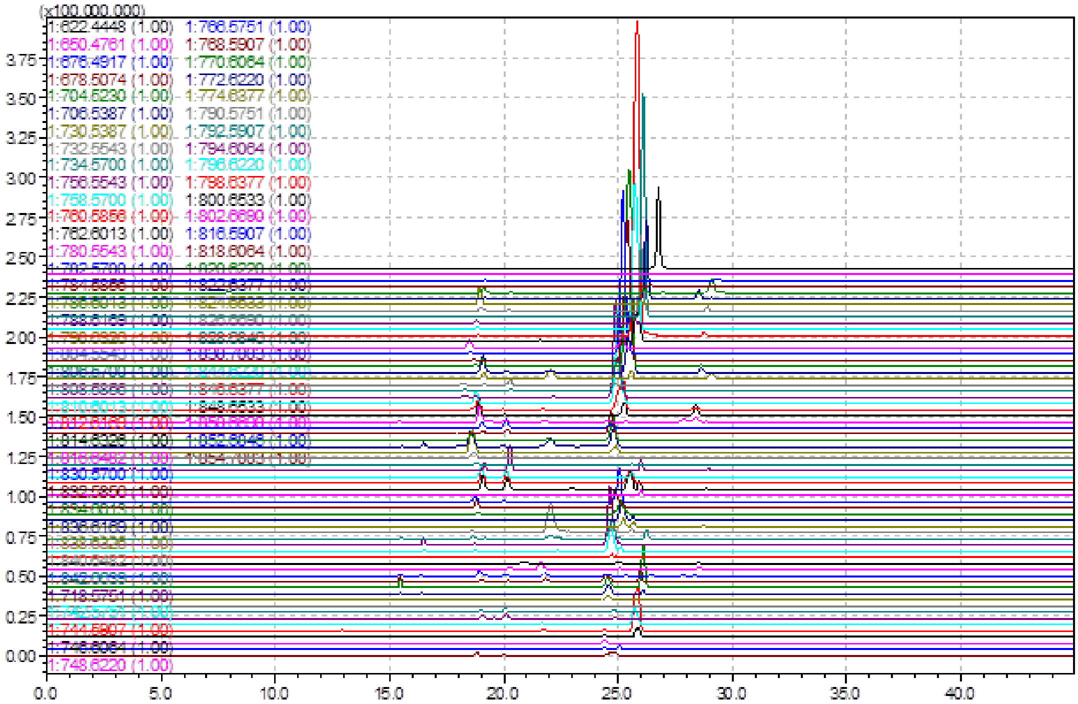


Figure 1: Chromatogram of PC and PC a lipid species for the normal sample of patient 37. This accompanies figure 4.13.

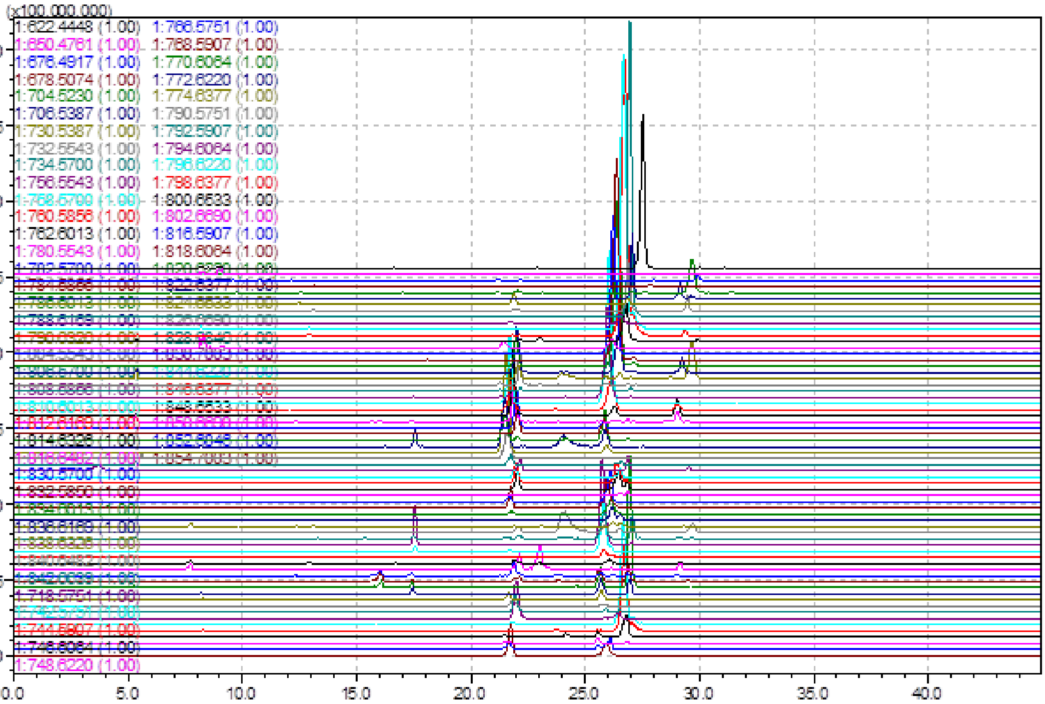


Figure 2: Chromatogram of PC and PC a lipid species for the normal sample of patient 75. This accompanies figure 4.13.

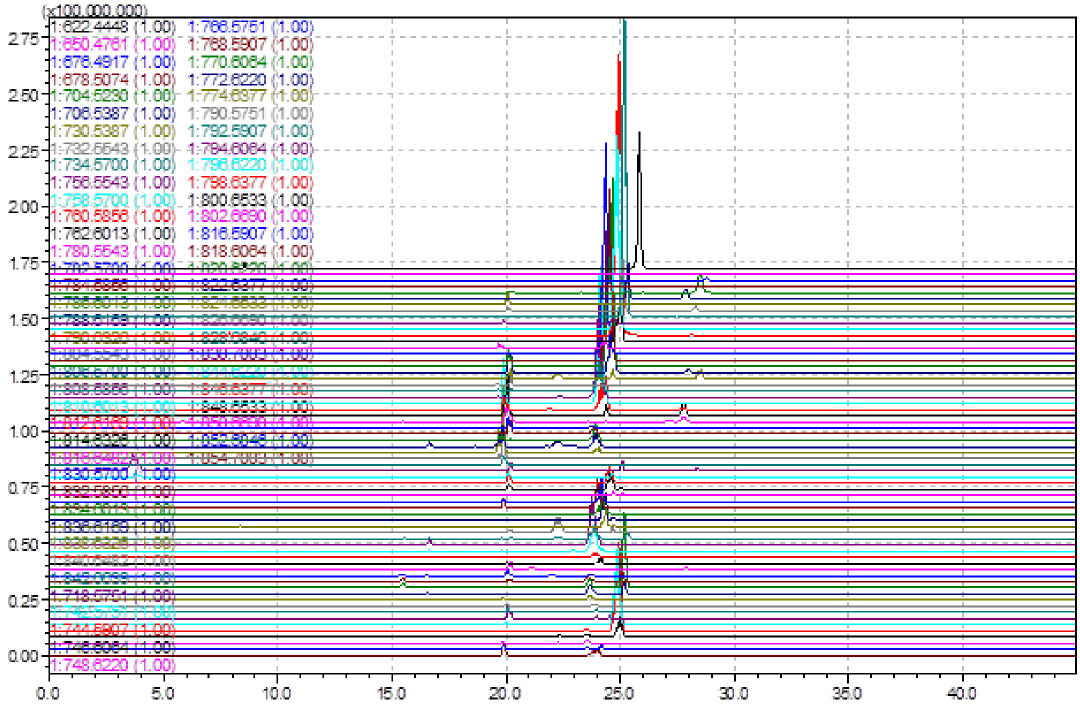


Figure 3: Chromatogram of PC and PC a lipid species for the normal sample of patient 1088. This accompanies figure 4.13.

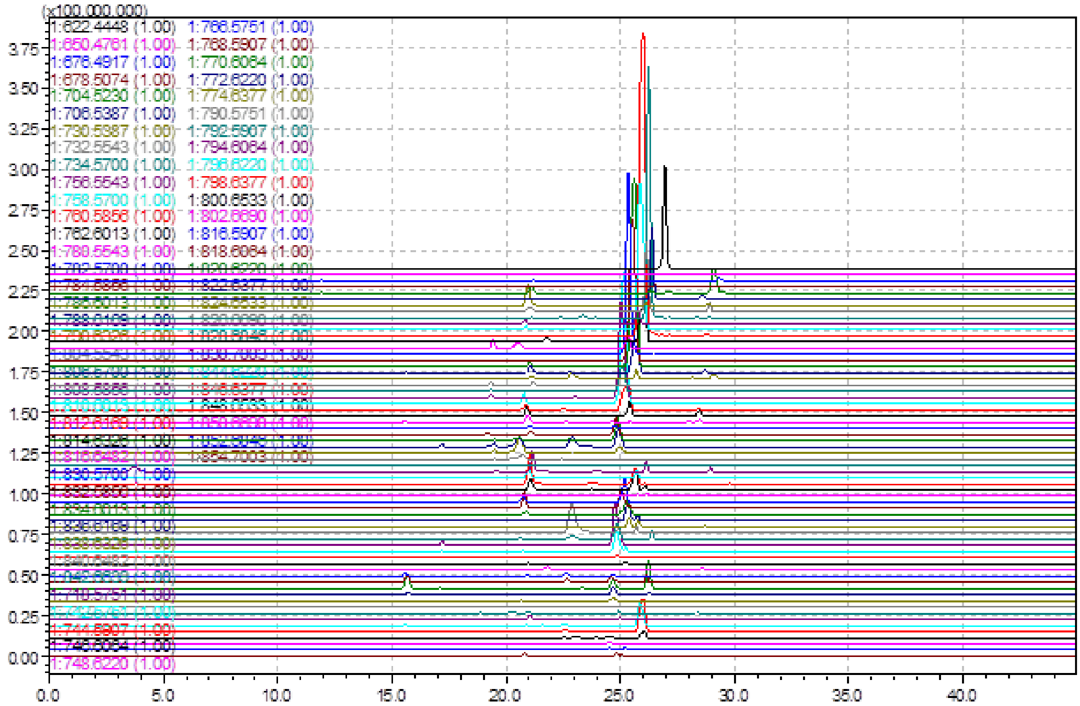


Figure 4: Chromatogram of PC and PC a lipid species for the normal sample of patient 1368. This accompanies figure 4.13.

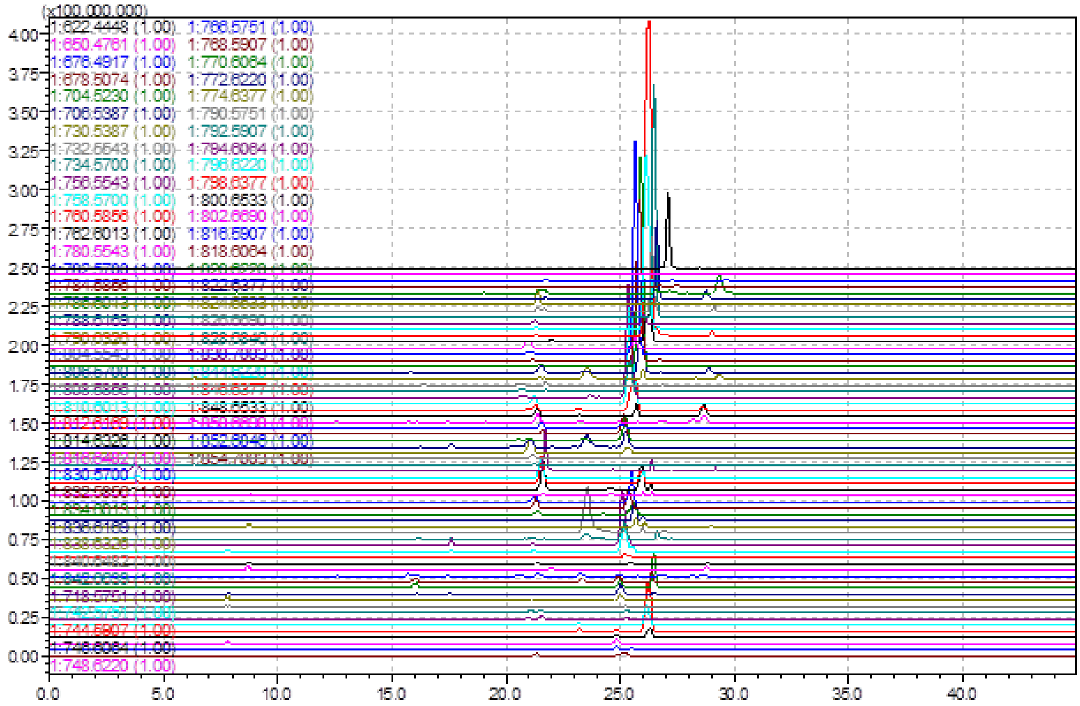


Figure 5: Chromatogram of PC and PC a lipid species for the normal sample of patient 1402. This accompanies figure 4.13.

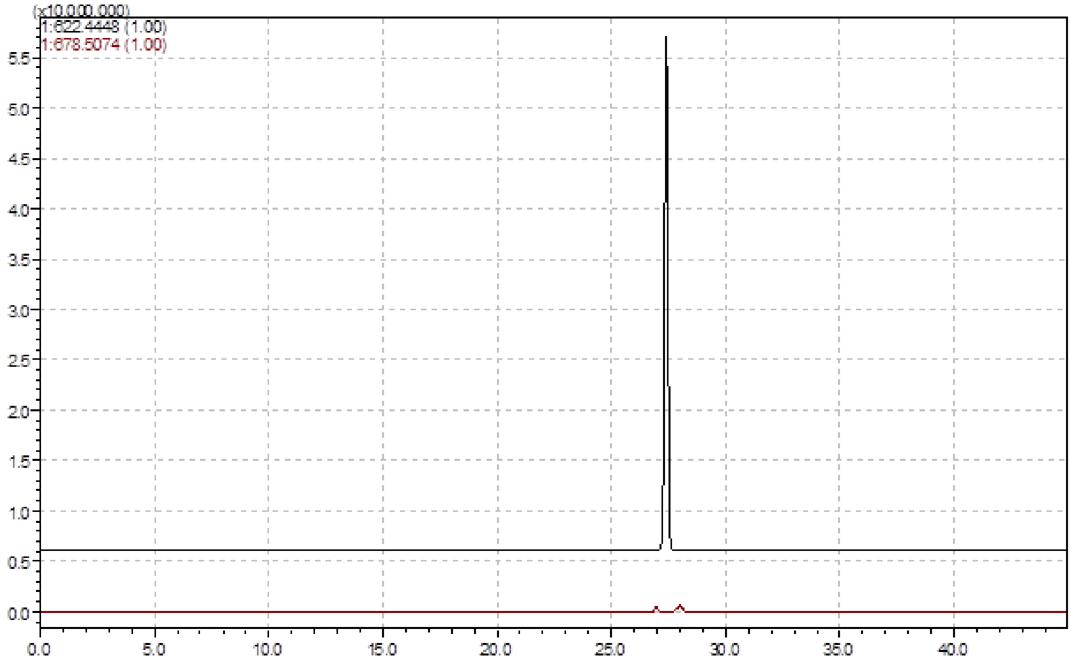


Figure 6: Chromatogram of the PC sub class internal standard PC 24:0 (black line) and the most important classification feature from the Random Forest analysis in section 4.3.1.5 PC 28:0 (brown line) for the normal sample of patient 1374.

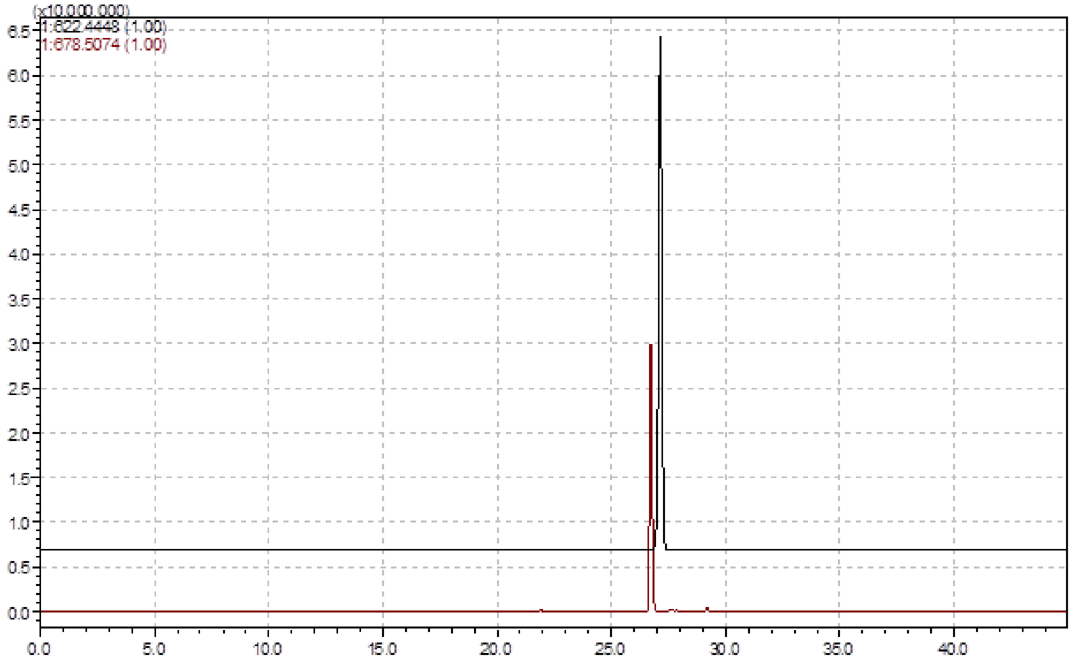


Figure 7: Chromatogram of the PC sub class internal standard PC 24:0 (black line) and the most important classification feature from the Random Forest analysis in section 4.3.1.5 PC 28:0 (brown line) for the tumour sample of patient 1374. The lipid is clearly upregulated in the tumour sample compared to the normal sample in figure 6

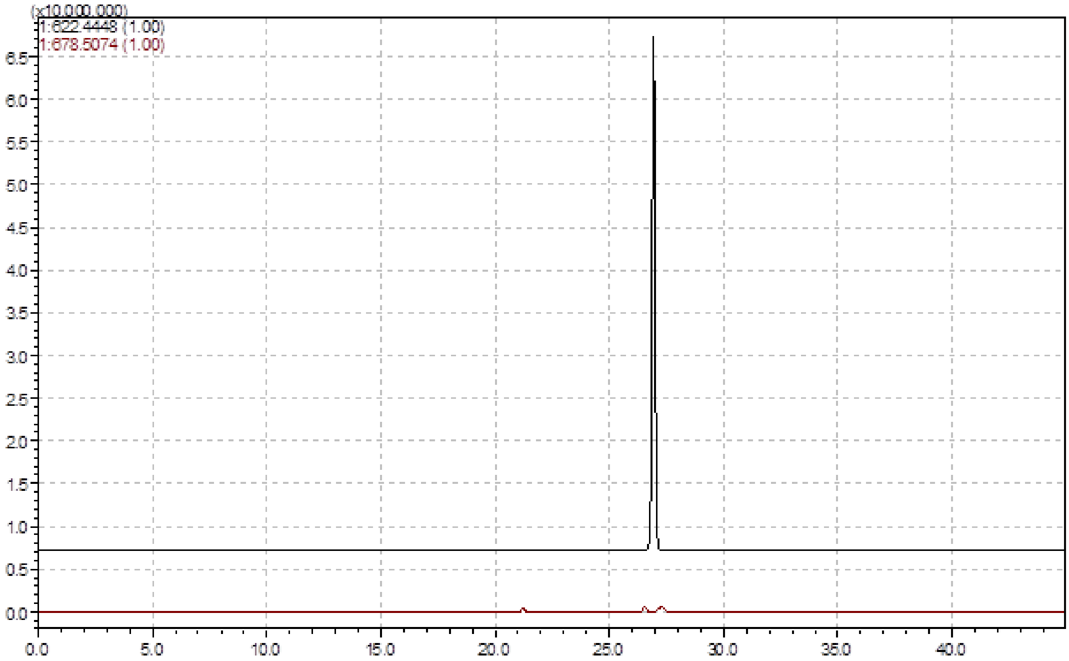


Figure 8: Chromatogram of the PC sub class internal standard PC 24:0 (black line) and the most important classification feature from the Random Forest analysis in section 4.3.1.5 PC 28:0 (brown line) for the normal sample of patient 1392.

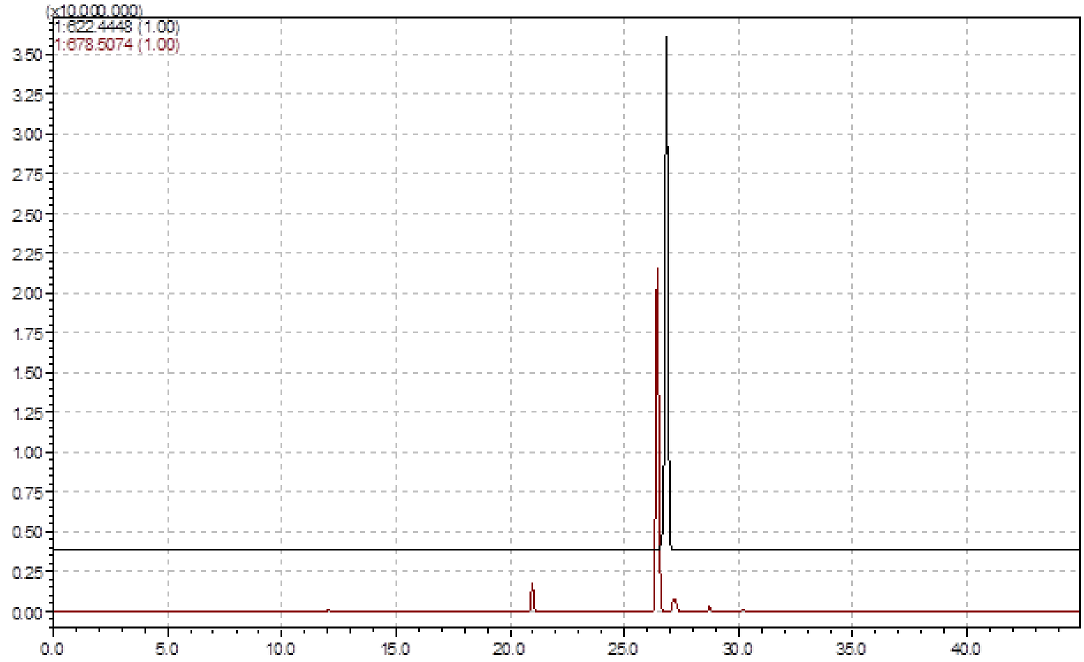


Figure 9: Chromatogram of the PC sub class internal standard PC 24:0 (black line) and the most important classification feature from the Random Forest analysis in section 4.3.1.5 PC 28:0 (brown line) for the normal sample of patient 1392. The lipid is clearly upregulated in the tumour sample compared to the normal sample in figure 8

References

Acar, N., Berdeaux, O., Grégoire, S., Cabaret, S., Martine, L., Gain, P., Thuret, G., Creuzot-Garcher, C.P., Bron, A.M. & Bretillon, L. (2012). Lipid composition of the human eye: are red blood cells a good mirror of retinal and optic nerve fatty acids? *PloS One*, **7**. 134

Adimoolam, S., Jin, L., Grabbe, E., Shieh, J.J. & Jonas, A. (1998). Structural and functional properties of two mutants of lecithin-cholesterol acyltransferase (T123I and N228K). *The Journal of biological chemistry*, **273**, 32561–7. 144

Aebersold, R. & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, **422**, 198–207. 10

Alexeyenko, A., Schmitt, T., Tjämberg, A., Guala, D., Frings, O. & Sonnhammer, E.L. (2012). Comparative interactomics with Funcoup 2.0. *Nucleic Acids Research*, **40**, 821–828. 1

Anonymous (2007). Recent patent applications in proteomics. *Nature biotechnology*, **25**, 425. 15

Armirotti, A. & Damonte, G. (2010). Achievements and perspectives of top-down proteomics. *Proteomics*, **10**, 3566–76. 15

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muertter, R.N., Holko, M., Ayanbule, O., Yefanov, A. &

SOBOLEVA, A. (2011). NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic acids research*, **39**, D1005–10. 4

BARSNES, H., VIZCAÍNO, J.A., EIDHAMMER, I. & MARTENS, L. (2009). PRIDE Converter: making proteomics data-sharing easy. *Nature biotechnology*, **27**, 598–9. 54, 59

BARSNES, H., VIZCAÍNO, J.A., REISINGER, F., EIDHAMMER, I. & MARTENS, L. (2011). Submitting proteomics data to PRIDE using PRIDE Converter. *Methods in molecular biology (Clifton, N.J.)*, **694**, 237–53. 54

BELL, A.W., DEUTSCH, E.W., AU, C.E., KEARNEY, R.E., BEAVIS, R., SECHI, S., NILSSON, T. & BERGERON, J.J.M. (2009). A HUPO test sample study reveals common problems in mass spectrometry-based *proteomics*. *Nature methods*, **6**, 423–30. 16, 32

BENTON, H.P., WONG, D.M., TRAUGER, S.A. & SIUZDAK, G. (2008). XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. *Analytical chemistry*, **80**, 6382–9. 17

BONDARENKO, P.V., CHELIUS, D. & SHALER, T.A. (2002). Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Analytical chemistry*, **74**, 4741–9. 26

BOSCH, L.J.W., CARVALHO, B., FIJNEMAN, R.J.A., JIMENEZ, C.R., PINEDO, H.M., VAN ENGELAND, M. & MEIJER, G.A. (2011). Molecular tests for colorectal cancer screening. *Clinical colorectal cancer*, **10**, 8–23. 133

BRAHIMI-HORN, M., BELLOT, G. & POUYSSÉGUR, J. (2011). Hypoxia and energetic tumour metabolism. *Current opinion in Genetics and Development*, **21**, 67–72. 213

BROWN, S.H.J., MITCHELL, T.W. & BLANKSBY, S.J. (2011). Analysis of unsaturated lipids by ozone-induced dissociation. *Biochimica et biophysica acta*, **1811**, 807–17. 71

CAFFREY, M. & HOGAN, J. (1992). LIPIDAT: a database of lipid phase transition temperatures and enthalpy changes. DMPC data subset analysis. *Chemistry and physics of lipids*, **61**, 1–109. 78

CHELIUS, D. & BONDARENKO, P.V. (2002). Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *Journal of proteome research*, **1**, 317–23. 26

CÔTÉ, R.G., JONES, P., MARTENS, L., KERRIEN, S., REISINGER, F., LIN, Q., LEINONEN, R., APWEILER, R. & HERMJAKOB, H. (2007). The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC bioinformatics*, **8**, 401. 76

CRAIG, R., CORTENS, J.P. & BEAVIS, R.C. (2004). Open source system for analyzing, validating, and storing protein identification data. *Journal of proteome research*, **3**, 1234–42. 6, 15

CRICK, F. (1970). Central dogma of molecular biology. *Nature*, **227**, 561–3. 1

CROFT, D., O'KELLY, G., WU, G., HAW, R., GILLESPIE, M., MATTHEWS, L., CAUDY, M., GARAPATI, P., GOPINATH, G., JASSAL, B., JUPE, S., KALATSKAYA, I., MAHAJAN, S., MAY, B., NDEGWA, N., SCHMIDT, E., SHAMOVSKY, V., YUNG, C., BIRNEY, E., HERMJAKOB, H., D'EUSTACHIO, P. & STEIN, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, **39**, D691–7. 126

CSORDAS, A., OVELLEIRO, D., WANG, R., FOSTER, J.M., RIOS, D., VIZCAÍNO, J.A. & HERMJAKOB, H. (2012). PRIDE: quality control in a proteomics data repository. *Database : the journal of biological databases and curation*, **2012**, bas004. 15, 65

DALBY, A., NOURSE, J.G., HOUNSHELL, W.D., GUSHURST, A.K.I., GRIER, D.L., LEALAND, B.A. & LAUFER, J. (1992). Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Science*, **3**, 244–55. 78

de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Spiteri, I., Turner, S. & Steinbeck, C. (2010). Chemical entities of biological interest: an update. *Nucleic acids research*, **38**, 249–254. 78

Desiere, F., Deutsch, E.W., Nesvizhskii, A.I., Mallick, P., King, N.L., Eng, J.K., Aderem, A., Boyle, R., Brunner, E., Donohoe, S., Fausto, N., Hafen, E., Hood, L., Katze, M.G., Kennedy, K.A., Kregenow, F., Lee, H., Lin, B., Martin, D., Ranish, J.A., Rawlings, D.J., Samelson, L.E., Shiio, Y., Watts, J.D., Wollscheid, B., Wright, M.E., Yan, W., Yang, L., Yi, E.C., Zhang, H. & Aebersold, R. (2005). Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome biology*, **6**, R9. 6, 15

Desiere, F., Deutsch, E.W., King, N.L., Nesvizhskii, A.I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S.N. & Aebersold, R. (2006). The PeptideAtlas project. *Nucleic acids research*, **34**, D655–8. 128

Draisma, H.H., Reijmers, T.H., Meulman, J.J., van der Greef, J., Hankmeier, T. & Boomsma, D.I. (2012). Hierarchical clustering analysis of blood plasma lipidomics profiles from mono- and dizygotic twin families. *European Journal of Human Genetics*. 134

Fahy, E., Subramaniam, S., Brown, H.A., Glass, C.K., Merrill, A.H., Murphy, R.C., Raetz, C.R.H., Russell, D.W., Seyama, Y., Shaw, W., Shimizu, T., Spener, F., van Meer, G., VanNieuwenhze, M.S., White, S.H., Witztum, J.L. & Dennis, E.A. (2005). A comprehensive classification system for lipids. *Journal of lipid research*, **46**, 839–61. 73, 76

Fahy, E., Subramaniam, S., Murphy, R.C., Nishijima, M., Raetz, C.R.H., Shimizu, T., Spener, F., van Meer, G., Wakelam, M.J.O. & Dennis, E.A. (2009). Update of the LIPID MAPS comprehensive classification system for lipids. *Journal of lipid research*, **50**, S9–14. 76

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C. & Parkin, D.M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN

2008. *International journal of cancer. Journal international du cancer*, **127**, 2893–917. 132

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Denise, C.S., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A.K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H.S., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y.A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X.M., Harrow, J., Herrero, J., Hubbard, T.J.P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A. & Searle, S.M.J. (2011). Ensembl 2012. *Nucleic acids research*, **40**, D84–90. 2, 55, 70

Foster, J.M., Degroeve, S., Gatto, L., Visser, M., Wang, R., Griss, J., Apweiler, R. & Martens, L. (2011). *A posteriori* quality control for the curation and reuse of public proteomics data. *Proteomics*, **11**, 2182–94. 21, 49

Gassler, N., Klaus, C., Kaemmerer, E. & Reinartz, A. (2010). Modifier-concept of colorectal carcinogenesis: lipidomics as a technical tool in pathway analysis. *World journal of gastroenterology : WJG*, **16**, 1820–7. 134

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H. & Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**, R80. 11, 17

Glish, G.L. & Vachet, R.W. (2003). The basics of mass spectrometry in the twenty-first century. *Nature reviews. Drug discovery*, **2**, 140–50. 10

Gorden, D.L., Ivanova, P.T., Myers, D.S., McIntyre, J.O., VanSaun, M.N., Wright, J.K., Matrisian, L.M. & Brown, H.A. (2011). Increased diacylglycerols characterize hepatic lipid changes in progression of human nonalcoholic fatty liver disease; comparison to a murine model. *Molecular bioSystems*, **7**, 3271–3279. 134

Haag, M., Schmidt, A., Sachsenheimer, T. & Brgger, B. (2012). Quantification of signaling lipids by nano-electrospray ionization tandem mass spectrometry (nano-esi ms/ms). *Metabolites*, **2**, 57–76. 215

Han, X. & Gross, R.W. (1994). Electrospray ionization mass spectroscopic analysis of human erythrocyte plasma membrane phospholipids. *PNAS*, **91**, 10635–10639. 7

Han, X., Rozen, S., Boyle, S.H., Hellegers, C., Cheng, H., Burke, J.R., Welsh-Bohmer, K.A., Doraiswamy, P.M. & Kaddurah-Daouk, R. (2011). Metabolomics in early alzheimer's disease: identification of altered plasma sphingolipidome using shotgun lipidomics. *PloS One*, **6**. 134

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S.G.N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C. & Apweiler, R. (2004a). The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nature biotechnology*, **22**, 177–83. 6

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. & Apweiler, R. (2004b). IntAct: an open source molecular interaction database. *Nucleic acids research*, **32**, D452–5. 6, 101

Hoch, F.L. (1998). Cardiolipins and mitochondrial proton-selective leakage. *Journal of bioenergetics and biomembranes*, **30**, 511–32. 129

Hoogland, C., O'Gorman, M., Bogard, P., Gibson, F., Berth, M., Cockell, S.J., Ekefjärd, A., Forsstrom-Olsson, O., Kapferer, A., Nilsson, M., Martínez-Bartolomé, S., Albar, J.P., Echevarría-Zomeño, S., Martínez-Gomariz, M., Joets, J., Binz, P.A., Taylor, C.F., Dowsey, A. & Jones, A.R. (2010). Guidelines for reporting the use of gel image informatics in proteomics. *Nature biotechnology*, **28**, 655–6. 54

Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K., Ahn, N.G. & Old, W.M. (2010). Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *Journal of proteome research*, **9**, 4152–60. 46

Hsu, F.F. & Turk, J. (2008). Elucidation of the double-bond position of long-chain unsaturated fatty acids by multiple-stage linear ion-trap mass spectrometry with electrospray ionization. *Journal of the American Society for Mass Spectrometry*, **19**, 1673–80. 71

Hsu, F.F. & Turk, J. (2009). Electrospray ionization with low-energy collisionally activated dissociation tandem mass spectrometry of glycerophospholipids: mechanisms of fragmentation and structural characterization. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, **877**, 2673–95. xix, 72, 96, 97

Hu, C., Kong, H., Qu, F., Li, Y., Yu, Z., Gao, P., Peng, S. & Xu, G. (2011). Application of plasma lipidomics in studying the response of patients with essential hypertension to antihypertensive drug therapy. *Molecular bioSystems*, **7**, 3271–3279. 134

Indiveri, C., Tonazzi, A. & Palmieri, F. (1991). Characterization of the unidirectional transport of carnitine catalyzed by the reconstituted carnitine carrier from rat liver mitochondria. *Biochimica et biophysica acta*, **1069**, 110–6. 129

Jawad, N., Direkze, N. & Leedham, S.J. (2011). Inflammatory bowel disease and colon cancer. *Recent results in cancer research. Fortschritte der Krebsforschung. Progrès dans les recherches sur le cancer*, **185**, 99–115. 132

Ji, H. & Davis, R.W. (2006). Data quality in genomics and microarrays. *Nature biotechnology*, **24**, 1112–3. 16

Jimenez, C.R., Knol, J.C., Meijer, G.A. & Fijneman, R.J.A. (2010). Proteomics of colorectal cancer: overview of discovery studies and identification of commonly identified cancer-associated proteins and candidate CRC serum markers. *Journal of proteomics*, **73**, 1873–95. 133

Jones, A.R., Carroll, K., Knight, D., Maclellan, K., Domann, P.J., Legido-Quigley, C., Huang, L., Smallshaw, L., Mirzaei, H., Shofstahl, J. & Paton, N.W. (2010). Guidelines for reporting the use of column chromatography in proteomics. *Nature biotechnology*, **28**, 654. 54

Joosten, R.P., Joosten, K., Murshudov, G.N. & Perrakis, A. (2012). PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallographica. Section D, Biological Crystallography*, **68**, 484–496. 11

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. (2012). Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, **40**, 109–114. 78

Kasianowicz, J.J., Brandin, E., Branton, D. & Deamer, D.W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 13770–3. 2

Keller, B.O., Sui, J., Young, A.B. & Whittal, R.M. (2008). Interferences and contaminants encountered in modern mass spectrometry. *Analytica chimica acta*, **627**, 71–81. 36

Kersey, P.J., Staines, D.M., Lawson, D., Kulesha, E., Derwent, P., Humphrey, J.C., Hughes, D.S.T., Keenan, S., Kerhornou, A., Koscielny, G., Langridge, N., McDowall, M.D., Megy, K., Maheswari, U., Nuhn, M., Paulini, M., Pedro, H., Toneva, I., Wilson, D., Yates, A. & Birney, E. (2012). Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic acids research*, **40**, D91–7. 70

Kim, H.Y., Wang, T.C. & Ma, Y.C. (1994). Liquid chromatography/mass spectrometry of phospholipids using electrospray ionization. *Analytical Chemistry*, **66**, 3977–3982. 7

Kita, T., Kume, N., Ishii, K., Horiuchi, H., Arai, H. & Yokode, M. (1999). Oxidized LDL and expression of monocyte adhesion molecules. *Diabetes research and clinical practice*, **45**, 123–6. 214

Klie, S., Martens, L., Vizcaíno, J.A., Côté, R., Jones, P., Hinneburg, A. & Hermjakob, H. (2008). Analyzing large-scale proteomics projects with latent semantic indexing. *Journal of proteome research*, **7**, 182–191. 23

Łabaj, P.P., Leparc, G.G., Linggi, B.E., Markillie, L.M., Wiley, H.S. & Kreil, D.P. (2011). Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics (Oxford, England)*, **27**, 383–91. 4

Latterich, M. & Schnitzer, J.E. (2011). Streamlining biomarker discovery. *Nature biotechnology*, **29**, 600–2. 15

Li, B.Q., Huang, T., Liu, L., Cai, Y.D. & Chou, K.C. (2012). Identification of Colorectal Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction Network. *PLoS ONE*, **7**, e33393. 134

Liu, T., Qian, W.J., Gritsenko, M.A., Xiao, W., Moldawer, L.L., Kaushal, A., Monroe, M.E., Varnum, S.M., Moore, R.J., Purvine, S.O., Maier, R.V., Davis, R.W., Tompkins, R.G., Camp, D.G. & Smith, R.D. (2006). High dynamic range characterization of the trauma patient plasma proteome. *Molecular & cellular proteomics : MCP*, **5**, 1899–913. 27, 66

Magrane, M. & Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database : the journal of biological databases and curation*, **2011**, bar009. 69

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, **470**, 198–203. 2

Mardis, E.R. & Wilson, R.K. (2009). Cancer genome sequencing: a review. *Human molecular genetics*, **18**, R163–8. 2

Martens, L. & Hermjakob, H. (2007). Proteomics data validation: why all must provide data. *Molecular bioSystems*, **3**, 518–22. 20

Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C.F., Gevaert, K., Vandekerckhove, J. & Apweiler, R. (2005). PRIDE: the proteomics identifications database. *Proteomics*, **5**, 4046. 6, 128

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Römpp, A., Neumann, S., Pizarro, A.D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.A. & Deutsch, E.W. (2011). mzML–a community standard for mass spectrometry data. *Molecular & cellular proteomics : MCP*, **10**, R110.000133. 15

Maskos, U. & Southern, E.M. (1992). Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised *in situ*. *Nucleic acids research*, **20**, 1679–84. 3

McLafferty, F.W. (2011). A century of progress in molecular mass spectrometry. *Annual review of analytical chemistry (Palo Alto, Calif.)*, **4**, 1–22. 15

Medina-Aunon, J.A., Martínez-Bartolomé, S., López-García, M.A., Salazar, E., Navajas, R., Jones, A.R., Paradela, A. & Albar, J.P. (2011). The ProteoRed MIAPE web toolkit: a user-friendly framework to connect and share proteomics standards. *Molecular & cellular proteomics : MCP*, **10**, M111.008334. 49

Monjazeb, A.M., High, K.P., Connoy, A., Hart, L.S., Koumenis, C. & Chilton, F.H. (2006). Arachidonic acid-induced gene expression in colon cancer cells. *Carcinogenesis*, **27**, 1950–60. 215

Montecchi-Palazzi, L., Kerrien, S., Reisinger, F., Aranda, B., Jones, A.R., Martens, L. & Hermjakob, H. (2009). The PSI semantic validator: a framework to check MIAPE compliance of proteomics data. *Proteomics*, **9**, 5112–9. 49, 55

Navas-Iglesias, N., Carrasco-Pancorbo, A. & Cuadros-Rodríguez, L. (2009). From lipids analysis towards lipidomics, a new challenge for the analytical chemistry of the 21st century. Part II: Analytical lipidomics. *Trends in Analytical Chemistry*, **28**, 393–403. 130

Nesvizhskii, A.I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics*, **73**. 131

Old, W.M., Shabb, J.B., Houel, S., Wang, H., Couts, K.L., Yen, C.Y., Litman, E.S., Croy, C.H., Meyer-Arendt, K., Miranda, J.G., Brown, R.A., Witze, E.S., Schweppe, R.E., Resing, K.A. & Ahn, N.G. (2009). Functional proteomics identifies targets of phosphorylation by B-Raf signaling in melanoma. *Molecular cell*, **34**, 115–31. 130

Ooira, A., Kimata, K., Suzuki, S., Takata, K. & Suzuki, I. (1974). A correlation between synthetic activities for matrix macromolecules and specific stages of cyto-differentiation in developing cartilage. *The Journal of biological chemistry*, **249**, 1637–45. 144

Orchard, S. & Hermjakob, H. (2011). Data standardization by the HUPO-PSI: how has the community benefitted? *Methods in molecular biology (Clifton, N.J.)*, **696**, 149–60. 54

Orešič, M., Seppänen-Laakso, T., Sun, D., Tang, J., Therman, S., Viehman, R., Mustonen, U., van Erp, T.G., Hyötyläinen, T., Thompson, P., Toga, A.W., Huttunen, M.O., Suvisaari, J., Kaprio, J., Lönnqvist, J. & Cannon, T.D. (2012). Phospholipids and insulin resistance in psychosis: a lipidomics study of twin pairs discordant for schizophrenia. *Genome Medicine*, **4**. 134, 135

Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., Kurbatova, N., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rustici, G., Sharma, A., Williams, E., Adamusiak, T., Brandizi, M., Sklyar, N. & Brazma, A. (2011). ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic acids research*, **39**, D1002–4. 4, 17, 55, 85

Parsons, D.W., Wang, T.L., Samuels, Y., Bardelli, A., Cummins, J.M., DeLong, L., Silliman, N., Ptak, J., Szabo, S., Willson, J.K.V., Markowitz, S., Kinzler, K.W., Vogelstein, B., Lengauer, C. & Velculescu, V.E. (2005). Colorectal cancer: mutations in a signalling pathway. *Nature*, **436**, 792. 133

Patient, S., Wieser, D., Kleen, M., Kretschmann, E., Jesus Martin, M. & Apweiler, R. (2008). UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics (Oxford, England)*, **24**, 1321–2. 55

Pedrioli, P.G.A., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R., Cheung, K., Costello, C.E., Hermjakob, H., Huang, S., Julian, R.K., Kapp, E., McComb, M.E., Oliver, S.G., Omenn, G., Paton, N.W., Simpson, R., Smith, R., Taylor, C.F., Zhu, W. & Aebersold, R. (2004). A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology*, **22**, 1459–66. 15

Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–67. 20

Piomelli, D. (1993). Arachidonic acid in cell signaling. *Current opinion in cell biology*, **5**, 274–80. 129

Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic acids research*, **37**, D32–6. 75

Qian, W.J., Kaleta, D.T., Petritis, B.O., Jiang, H., Liu, T., Zhang, X., Mottaz, H.M., Varnum, S.M., Camp, D.G., Huang, L., Fang, X., Zhang, W.W. & Smith, R.D. (2008). Enhanced detection of low abundance human plasma proteins using a tandem IgY12-SuperMix immunoaffinity separation strategy. *Molecular & cellular proteomics : MCP*, **7**, 1963–73. 27, 66

Rabilloud, T., Chevallet, M., Luche, S. & Lelong, C. (2010). Two-dimensional gel electrophoresis in proteomics: Past, present and future. *Journal of proteomics*, **73**, 2064–77. 15

Ran, S. & Thorpe, P.E. (2002). Phosphatidylserine is a marker of tumor vasculature and a potential target for cancer imaging and therapy. *International journal of radiation oncology, biology, physics*, **54**, 1479–84. 212

Rappsilber, J. (2011). The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *Journal of Structural Biology*, **173**, 530–540. 4

Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H. & Jimeno, A. (2008). Text processing through Web services: calling Whatizit. *Bioinformatics (Oxford, England)*, **24**, 296–8. 95

Reinders, J. & Sickmann, A. (2007). Modificomics: posttranslational modifications beyond protein phosphorylation and glycosylation. *Biomolecular engineering*, **24**, 168–177. 1

Rogers, S., Girolami, M., Kolch, W., Waters, K.M., Liu, T., Thrall, B. & Wiley, H.S. (2008). Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics (Oxford, England)*, **24**, 2894–900. 4

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry*, **242**, 84–9. 2

Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A. & Pappin, D.J. (2004). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics : MCP*, **3**, 1154–69. 19, 43

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., Simons, J.F., Marran, D., Myers, J.W., Davidson, J.F., Branting, A., Nobile, J.R., Puc, B.P., Light, D., Clark, T.A., Huber, M., Branciforte, J.T., Stoner, I.B., Cawley, S.E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K.J., Williams, A., Roth, G.T. & Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–52. 2

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. & Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, **265**, 687–95. 2

Sauter, G., Nerlich, A., Spengler, U., Kopp, R. & Pfeiffer, A. (1990). Low diacylglycerol values in colonic adenomas and colorectal cancer. *Gut*, **31**, 1041–5. 214

Scheltema, R.A., Jankevics, A., Jansen, R.C., Swertz, M.A. & Breitling, R. (2011). PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. *Analytical chemistry*, **83**, 2786–93. 17

Schiess, R., Wollscheid, B. & Aebersold, R. (2009). Targeted proteomic strategy for clinical biomarker discovery. *Molecular oncology*, **3**, 33–44. 26

Sewell, G.W., Hannun, Y.A., Han, X., Koster, G., Bielawski, J., Goss, V., Smith, P.J., Rahman, F.Z., Vega, R., Bloom, S.L., Walker, A.P., Postle, A.D. & Segal, A.W. (2012). Lipidomic profiling in Crohn's disease: Abnormalities in phosphatidylinositols, with preservation of ceramide, phosphatidylcholine and phosphatidylserine composition. *The International Journal of Biochemistry and Cellular Biology*. 134

Sigdel, T.K. & Sarwal, M.M. (2011). Recent advances in biomarker discovery in solid organ transplant by proteomics. *Expert review of proteomics*, **8**, 705–15. 15

Simopoulos, A.P. (2002). The importance of the ratio of omega-6/omega-3 essential fatty acids. *Biomedicine & pharmacotherapy = Biomédecine & pharmacothérapie*, **56**, 365–79. 110

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry*, **78**, 779–87. 17

Smith, R.E., Lespi, P., Di Luca, M., Bustos, C., Marra, F.A., de Alaniz, M.J.T. & Marra, C.A. (2008). A reliable biomarker derived from plasmalogens to evaluate malignancy and metastatic capacity of human cancers. *Lipids*, **43**, 79–89. 214

Solit, D.B. & Mellinghoff, I.K. (2010). Tracing cancer networks with phosphoproteomics. *Nature biotechnology*, **28**, 1028–9. 15

Southern, E.M. (1975). Detection of specific sequences among dna fragments separated by gel electrophoresis. *Journal of Molecular Biology*, **98**, 503–517. 3

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**, D535–9. 6

Stein, L.D. (2010). The case for cloud computing in genome informatics. *Genome biology*, **11**, 207. 2

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. & Willighagen, E. (2003). The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *Journal of chemical information and computer sciences*, **43**, 493–500. 92

Stryer, L., Berg, J.M. & Tymoczko, J.L. (2002). *Biochemistry*. W.H.Freeman & Co Ltd, 5th edn. 144

Stübiger, G., Aldover-Macasaet, E., Bicker, W., Sobal, G., Willfort-Ehringer, A., Pock, K., Bochkov, V., K, W. & O., B. (2012). Targeted profiling of atherogenic phospholipids in human plasma and lipoproteins of hyperlipidemic patients using MALDI-QIT-TOF-MS/MS. *Atherosclerosis*. 134

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E.A., Glass, C.K., Merrill, A.H., Murphy, R.C., Raetz, C.R.H., Russell, D.W. & Subramaniam, S. (2007). LMSD: LIPID MAPS structure database. *Nucleic acids research*, **35**, D527–32. 7, 76

Taylor, C.F., Paton, N.W., Lilley, K.S., Binz, P.A., Julian, R.K., Jones, A.R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E.W., Dunn, M.J., Heck, A.J.R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T.A., Patterson, S.D., Ping, P., Seymour, S.L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T.M., Whitelegge, J.P., Wilkins, M.R., Xenarios, I., Yates, J.R. & Hermjakob, H. (2007). The minimum information about a proteomics experiment (MIAPE). *Nature biotechnology*, **25**, 887–93. 49

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A.K.A. & Hamon, C. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical chemistry*, **75**, 1895–904. 19, 43

Tjalsma, H. (2010). Identification of biomarkers for colorectal cancer through proteomics-based approaches. *Expert review of proteomics*, **7**, 879–95. 133

Triantafillidis, J.K., Nasioulas, G. & Kosmidis, P.A. (2009). Colorectal cancer and inflammatory bowel disease: epidemiology, risk factors, mechanisms of carcinogenesis and prevention strategies. *Anticancer research*, **29**, 2727–37. 132

UniProt-Consortium (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic acids research*, **39**, D214–9. 69

UniProt-Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research*, **40**, D71–5. 6, 12

Utsugi, T., Schroit, A.J., Connor, J., Bucana, C.D. & Fidler, I.J. (1991). Elevated expression of phosphatidylserine in the outer membrane leaflet of human tumor cells and recognition by activated human blood monocytes. *Cancer research*, **51**, 3062–6. 212

Velankar, S., Alhroub, Y., Best, C., Caboche, S., Conroy, M.J., Dana, J.M., Fernandez Montecelo, M.A., van Ginkel, G., Golovin, A., Gore, S.P., Gutmanas, A., Haslam, P., Hendrickx, P.M.S., Heuson, E., Hirshberg, M., John, M., Lagerstedt, I., Mir, S., Newman, L.E., Oldfield, T.J., Patwardhan, A., Rinaldi, L., Sahni, G., Sanz-García, E., Sen, S., Slowley, R., Suarez-Uruena, A., Swaminathan, G.J., Symmons, M.F., Vranken, W.F., Wainwright, M. & Kleywegt, G.J. (2011). PDBe: Protein Data Bank in Europe. *Nucleic acids research*, **40**, D445–52. 55

Vizcaíno, J.A., Côté, R., Reisinger, F., Foster, J.M., Mueller, M., Rameseder, J., Hermjakob, H. & Martens, L. (2009). A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics*, **9**, 4276–83. 15, 20, 27

Vizcaíno, J.A., Côté, R., Reisinger, F., Barsnes, H., Foster, J.M., Rameseder, J., Hermjakob, H. & Martens, L. (2010). The Proteomics Identifications database: 2010 update. *Nucleic acids research*, **38**, D736–42. 21

Wang, R., Fabregat, A., Rios, D., Ovelleiro, D., Foster, J.M., Côté, R., Griss, J., Csordas, A., Perez-Riverol, Y., Reisinger, F., Hermjakob, H., Martens, L. & Vizcaíno, J.A. (2012). PRIDE inspector: a tool to visualize and validate ms proteomics data. *Nature biotechnology*, **30**, 135–7. 26, 56, 67

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, **10**, 57–63. 4, 78

Watson, A.J.M. & Collins, P.D. (2011). Colon cancer: a civilization disorder. *Digestive diseases (Basel, Switzerland)*, **29**, 222–8. 132

Wenk, M.R. (2005). The emerging field of lipidomics. *Nature reviews. Drug discovery*, **4**, 594–610. 7

Wenk, M.R. (2010). Lipidomics: new tools and applications. *Cell*, **143**, 888–95. 7, 129

Wishart, D., Knox, C., Guo, A., Eisner, R., Young, N., Gautam, B., Hau, D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J., Lim, E., Sobsey, C., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., Souza, A.D., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. & Forsythe, I. (2009). Hmdb: a knowledgebase for the human metabolome. *Nucleic acids research*, **37**, 603–610. 78

Yang, H.J., Hong, J., Lee, S., Shin, S., Kim, J. & Kim, J. (2010). Pressure-assisted tryptic digestion using a syringe. *Rapid communications in mass spectrometry : RCM*, **24**, 901–8. 30

Yao, Y., Huang, C., Li, Z.F., Wang, A.Y., Liu, L.Y., Zhao, X.G., Luo, Y., Ni, L., Zhang, W.G. & Song, T.S. (2009). Exogenous phosphatidylethanolamine induces apoptosis of human hepatoma HepG2 cells via the bcl-2/Bax pathway. *World journal of gastroenterology : WJG*, **15**, 1751–8. 214

Yasugi, E. & Watanabe, K. (2002). [LIPIDBANK for Web, the newly developed lipid database]. *Tanpakushitsu kakusan koso. Protein, nucleic acid, enzyme*, **47**, 837–41. 76, 82

Yi, J., Kim, C. & Gelfand, C.A. (2007). Inhibition of intrinsic proteolytic activities moderates preanalytical variability and instability of human plasma. *Journal of proteome research*, **6**, 1768–81. 46

Yoshinaga, M.Y., Kellermann, M.Y., Rossel, P.E., Schubotz, F., Lipp, J.S. & Hinrichs, K.U. (2011). Systematic fragmentation patterns of archaeal intact polar lipids by high-performance liquid chromatography/electrospray ionization ion-trap mass spectrometry. *Rapid communications in mass spectrometry : RCM*, **25**, 3563–74. 71

Young, K.H. (1998). Yeast two-hybrid: so many interactions, (in) so little time... *Biology of reproduction*, **58**, 302–11. 6