SUPPORTING ANALYSIS, VISUALISATION AND BIOLOGICAL INTERPRETATION OF METABOLOMICS DATASETS

YOANN GLOAGUEN

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

INSTITUTE OF INFECTION, IMMUNITY AND INFLAMMATION
COLLEGE OF MEDICAL, VETERINARY AND LIFE SCIENCES
UNIVERSITY OF GLASGOW

MARCH 2017

Abstract

Over the past decades, the emerging omics technologies have enabled scientists to take a step further in the investigation of biological systems. From food safety to stratified medicine, omics technologies are now an essential and powerful means to study biological processes. Omics technologies are however at different stages of maturity, and the most recent field of the omics family, metabolomics, is still in its infancy. Metabolomics attempts to catalogue, characterise and quantify all small molecules constitutive of a biological system. Liquid Chromatography - Mass Spectrometry (LCMS) is now the most commonly used technique to generate metabolomics data. The method allows the detection of hundreds of metabolites from a single sample and can provide a rapid assignment of formulae to detected masses using high accuracy mass spectrometers. While analytical methods are well developed, support for linking metabolites to detected features and interpreting the results of a data analysis in a biological context is still poorly developed. Significant challenges also arise from the additional steps required to export the data to third party environments to create a biological context. The study of integrated omics datasets as a single system has also shown to provide greater inferences than the study of each omics separately. Methods to integrate the different omics layers of biological systems are, however, at an early stage of development and no standard approach currently exists to provide a holistic view of organisms systems organisation.

The objective of this thesis is to formalise, standardise and unify the data analysis of the metabolomics field, by providing to biologists the tools to support them from planning to analysis to biological impact reporting. The work presented here focuses particularly on untargeted LC-MS metabolomics approaches and attempts to assist non-expert users in performing their own analysis of metabolomics datasets. The project also aims to enable systematic biological interpretation of metabolomics datasets. The first part of the thesis focuses on creating the foundation of a unified environment for LC-MS metabolomics data analysis. Subsequently, the created environment will be expanded to integrate and support the latest technological advances in the field and provide better support for both designing studies and

interpreting analysis results in a biological context. Finally, the last part of this thesis concentrates on integrating metabolomics data with other omics datasets in an attempt to provide a holistic view of a biological system.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Over the past four years, I have received support and encouragement from many individuals. Karl Burgess, my supervisor, has been incredibly helpful, regularly giving me valuable and insightful advice, and always being supportive and positive which made this experience a thoughtful and rewarding journey. I would like to thank Mike Barrett who led me during the first year and has been very supportive throughout these four years. I have also benefited from the support of many bioinformaticians and software developers within Glasgow University, Fraser Morton, David Wilson, Gavin Blackburn, Ronan Daly, Joe Wandy, Ross Gurden, Graham Hamilton, Simon Rogers, Manikhandan Mudaliar. I also like to thank Isabel Vincent and Justin van der Hooft amongst many people who tested over and over my software always giving meaningful feedback. Glasgow Polyomics provided me with a dynamic environment, and I would like to thank Glasgow Polyomics staff for the many insightful discussions we had over the past years. I also benefited from a four-year collaboration with Fabien Jourdan who is always welcoming and creative in his approaches to solving problems and surrounded by a great team of bioinformaticians. Finally, I would like to thank my family and Geraldine Camus who have been incredibly supportive for the past four years.

Author's declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Printed Name: YOANN GLOAGUEN

Signature: _____ 

List of Abbreviations

α7nAChR Oral keratinocyte alpha 7 nicotinic receptor

AJAX Asynchronous JavaScript and XML

CODA Component detection algorithm

COSMOS COordination Of Standards In MetabOlomicS

COW Correlation optimal warping

CPM Continuous profile mode

CRP C-reactive protein

DDA Data dependent acquisition

DIA Data-independent acquisition

DTW Dynamic time warping

EIC Extracted-ion chromatogram

FrAnK Fragmentation annotation kit

GCMS Gas chromatography - mass spectrometry

GP Glasgow Polyomics

HILIC Hydrophilic interaction liquid chromatography

HPLC High-performance liquid chromatography

JSON Javascript object notation

LCMS Liquid chromatography - mass spectrometry

LIMS Laboratory information management systems

MS Mass spectrometry

MSI Metabolomics Standard Initiative

MS/MS Tandem mass spectrometry

MVT Model view template

m/z Mass-to-charge ratio

NIST National Institute of Science and Technology

NMR Nuclear magnetic resonance

OBI-Warp Ordered bijective interpolated warping

ORM Object-relational mapping

PCA Principal component analysis

PiMP Polyomics Metabolomics integrated Pipeline

PTMS Post-translational modification

PTW Parametric time warping

TIC Total ion chromatogram

ToF Time of flight

Chapter 1

Introduction

This chapter presents the necessary background to understand the organisation of biological systems into omics layers with an emphasis on liquid-chromatography mass spectrometry metabolomics (LCMS). Measurement technologies, data formats and data processing of LCMS metabolomics are discussed, highlighting the challenges that the field is facing. The representation of omics data into biological networks is also introduced and reviewed in this section.

1.1 Omics technologies

Omics technologies attempt to characterise, quantify and help understanding relationships between all molecules constitutive of an organism. Over the past decades, the collection and interpretation of large-scale datasets have been powering new discoveries across all disciplines in biomedical sciences. The recent advances in high-throughput omics technologies such as genomics, transcriptomics, proteomics and metabolomics and improvement in bioinformatics have enabled the investigation of thousands of genes, proteins and metabolites simultaneously. Omics technologies have now an essential role in many fields of biological research: toxicology [1] and environmental health [2], biomarker discovery [3] and cancer diagnostics [4], food safety [5] and nutrition [6] are some examples of the disciplines that now make systematic use of omics technologies to drive their research.

1.1.1 Omics layers

Omics technologies are divided into four main disciplines, each of them allowing the investigation of four different parts of an organism or biological systems.

Genomics focuses on the large-scale study of the genes and how they interact with one another by sequencing DNA molecules to determine the order of nucleotides. This technology enabled the Human Genome Project [7] whose aim was to determine the sequence of nucleotides that make human DNA. Transcriptomics is the study of messenger RNA (mRNA). The identification and quantification of these molecules provide a way to understand the expression of genes better. The technology has evolved to allow the investigation of all species of transcripts such as small RNAs, non-coding RNAs as well as mRNAs [8]. Proteomics aspires to the large-scale identification and quantitation of the entire set of proteins present in an organism [9, 10]. Proteins are the reflection of gene expression through transcription and play a major role in the regulation of cell processes. Finally, the latest technology of the omics family, metabolomics, is the large-scale study of the metabolites - small molecules - present in an organism [11].

The complexity of omics technologies is however not linear. As illustrated in Figure 1.1, the complexity increases as the omics get closer to the phenotype. This is due to the growing number of arrangements the building blocks of each omics can take. Similarly, the number of molecules constitutive of each layer is not linear either. A single gene can indeed encode for hundreds of protein isoforms (due to post-translational modification (PTMS) and alternative splicing) [12] which makes the system incredibly intricate to study as a whole. Moreover, mRNA transcript levels do not always correlate with respective protein expression levels [13]. Omics technologies also attempt to explain the modification mechanisms that happen at different layers in the cells such as DNA methylation [14] which plays a significant role in gene expression, alternative splicing [15] and RNA editing [16], or post-translational modification of proteins [17].

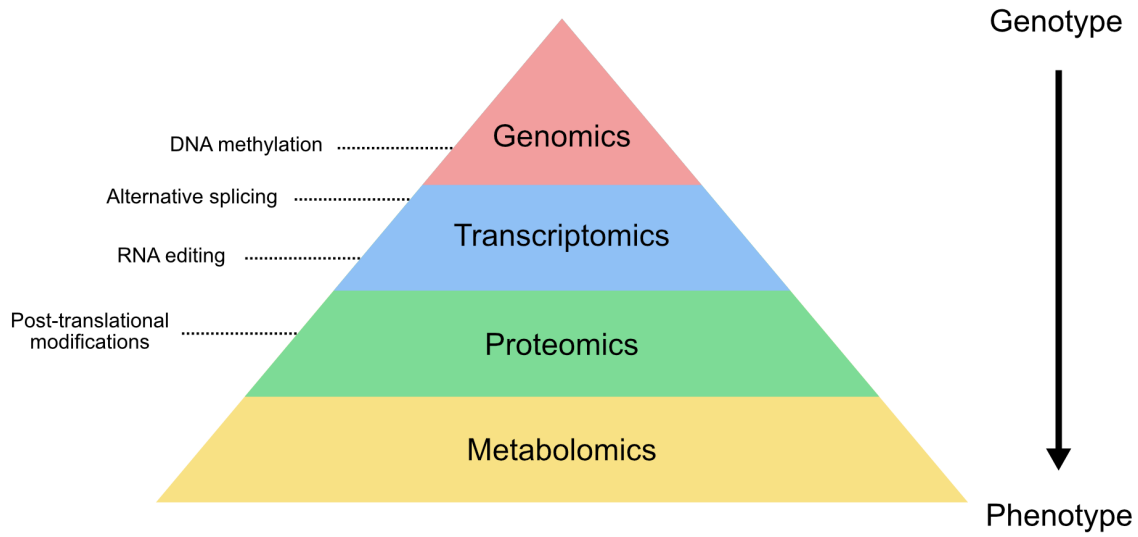

Figure 1.1: Omics layers organisation

1.1.2 Omics interactions

While individual omics datasets are informative, and combined analysis of genomics, transcriptomics, proteomics and metabolomics data has been found to be useful for a deeper understanding of fundamental biological processes, greater inferences can be obtained by integrating those datasets that are collected at different levels of biological organisation. Indeed, the first approach to multi-omics experiment has been to analyse them separately, in isolation of one another, and use the results as informative means to interpret another layer [18]. This combined analysis approach is an iterative process consisting of using information from one layer to focus the analysis of an another layer on a specific and narrow part. It, however, does not allow the study of omics layers as one system like integrative analyses. Indeed, integrative approcahes attempt to bring the datasets together in deifferent ways to interpret them as a whole. Various correlation-based approaches have been explored in an attempt to integrate multiple omics layers and extract meaningful information [19, 20]. Similarly, methods and software have been developed to assist and automate these approaches [21, 22, 23]. These methods have been successful in many cases, but all face the same challenge of connecting the different layers in a biological context, representing the interaction and process happening between and within layers. Although progress is made towards that goal [24], standard methods to integrate multiple omics datasets in a biological context is yet to be developed to unleash the full potential of omics technologies [25].

1.2 Metabolomics

Metabolomics aims to provide a snapshot, at a specific point in time, of all chemical activities occurring in a cell, tissue or organism, allowing the study of the biological processes in place in response to a stress. The metabolome, however, unlike the genome, is not static, it reflects the changes happening at every level of a biological system and can be influenced by environmental factors. Two cells of the same organism can indeed reveal an entirely different set of small molecules while sharing the same genome. Thus, metabolomics is often seen as the layer linking genotypes and phenotypes [26].

Two leading measurement technologies are currently used in metabolomics, Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS). These two types of analysis offer different views of the metabolome and are used for varying purposes. NMR, a non-destructive technology, is highly reproducible, provides structural information and absolute quantitation of the compounds observed. Mass spectrometry, in comparison, is a destructive technology, and requires isotopically labeled standards to provide anything other than relative abundance. However, MS offers higher sensitivity than NMR which allows the detection of many more metabolites. The choice of the technology usually depends on the design of the study and

the question addressed. NMR can be used for studies requiring the absolute quantitation of a definite set of metabolites while mass spectrometry is preferred for the exploration of the metabolome in a more untargeted approach.

1.2.1 Mass spectrometry metabolomics workflow

Over the past decade, many MS approaches based on different instrumentations have been implemented to study the metabolism. LCMS is often used for untargeted approaches due to the diverse range of separation available, its large sample capacity, and straightforward sample preparation methods. Gas chromatography - mass spectrometry (GCMS) is generally for targeted approaches as it offers absolute quantification for known compounds, a very high retention time reproducibility, but requires a derivatization step to make compounds volatile. Other analytical methods are available such as capillary electrophoresis - mass spectrometry (CEMS) which offers high separation power but poor retention time reproducibility, or direct infusion - mass spectrometry which offers rapid analysis but no separation. These different type of approaches can be coupled with tandem mass spectrometry to provide a better structural elucidation of the compounds analysed. Tandem MS can be performed following different protocol, Data dependent acquisition (DDA) allow the fragmentation and elucidation of the structure of a set predefined compounds, while data independent acquisition (DIA) proceed to the fragmentation of all ions present in the matrix. The measurement technologies, separation techniques and fragmentation procedures are introduced in more detailed in section 1.3 of this introduction.

Metabolomics laboratories and core facilities across the world use very similar overall workflows in term of data handling. The raw data acquired on the MS instrument is generally stored in in-house servers and archived, the vendor formatted files are then duplicated and converted to an open format for processing purposes. Two types of processing workflows have been adopted by different laboratories, one uses commercial software such as Compound Discoverer (Thermo Fisher Scientific) or Progenesis QI (Nonlinear Dynamics - Waters), the other makes use of freely available tools. Amongst the laboratories that use the free option, some use end to end data processing pipelines, other prefer to build their own pipeline using different tools for each step of the analysis. The different data analysis tools and platforms are discussed in section 1.3.4. Although there is a wide range of tools available, they all follow the same data processing steps. The analysis pipeline generally consist in 6 main steps: peak detection, peak alignment, data filtering, peak grouping, peak identification and statistical analysis. The order of these steps can slightly differ from one tool to another and several quality control steps can be introduced at a different stages of the pipeline. Those steps are detailed in section 1.3.3 of this introduction.

While the data processing steps are conserved across metabolomics facilities, data capture is

not yet standardised in LCMS. Laboratories use different ways to capture data and document studies which can go from electronic lab books to internal Laboratory Information Management Systems (LIMS). No tool or LIMS is however used across metabolomics community as they are often very specific to the need of laboratory which implemented it. This creates disparity in the way studies are documented. The Metabolomics Standard Initiative (MSI) is however attempting to standardise the reporting of studies by providing rules and best practices guidelines. Data repositories such as MetaboLights also now provides strict guidelines regarding the type of data that need to be captured to properly document a metabolomic study.

1.3 LCMS Metabolomics

1.3.1 Measurement and separation technologies

Mass spectrometry is an analytical technique that separates ionised chemical compounds by their mass-to-charge ratio (*m/z*) [27]. A mass spectrometer is constituted of 3 principal components with different purposes: the ion source imparts a charge to a molecule, the mass analyser separates ions, and the detector records ion signals. Several types of ionisation techniques are available, they are however not all appropriate for LCMS. Electron ionisation, for example, which produces a high degree of fragmentation is ordinarily coupled to gas chromatography as it cannot be used at atmospheric pressure and require the entire system to be under high vacuum [28]. Electrospray ionisation [29] is the most widely used ion source for LCMS metabolomics and produces soft ionisation (which reduces fragmentation). Alternatively, matrix-assisted laser desorption/ionisation (MALDI) [30] is used for imaging, to inform on the spatial distribution profiles of metabolites in tissues [31]. Many mass analysers exist with different characteristics; however, modern instruments used in LCMS share high mass resolving power. The mass resolving power is the ability of the mass spectrometer to separate ions with close *m/z* and evaluated using mass accuracy. Mass accuracy is measured in parts per million by calculating the ratio of the *m/z* measurement error to the real *m/z*. Three types of mass analysers are widely used for LCMS: time of flight (ToF), quadrupoles, and ion traps. ToF analysers create an electric field to accelerate the ions and measure the time ions take to reach the detector. Quadrupole analysers use oscillating electrical fields and a changing potential allowing only ions in a particular range of *m/z* to reach the detector at a given time scanning a wide mass range in a short period. Several types of ion traps exist; three-dimensional quadrupole ion traps, linear quadrupole ion traps and Orbitraps are examples.

Many of the instruments can also perform tandem MS (MS/MS). MS/MS is the succession

of at least two rounds of mass spectrometry separated by fragmentation. Fragmentation data can inform on the structure of the molecule analysed and is, therefore, a valuable resource in metabolomics to help with the identification of metabolites. Two types of fragmentation can be performed, fragmentation in time and fragmentation in space. Fragmentation in space can be done by using three quadrupoles (Triple Quadrupole) as seen in Figure 1.2; the first mass analyser isolates an ion, the second analyser acts as a collision cell to fragment the ion, the third analyser isolates a fragment ion. This means a signal will only occur if a characteristic molecular mass is detected, followed by a diagnostic fragment ion [32]. Fragmentation in time is done using one ion trap mass analyser over time such as quadrupole ion trap, and typically involves trapping the ions, selecting an ion of interest by manipulating the electrostatic field in the trap, then collisionally dissociating the analytes using a neutral gas.

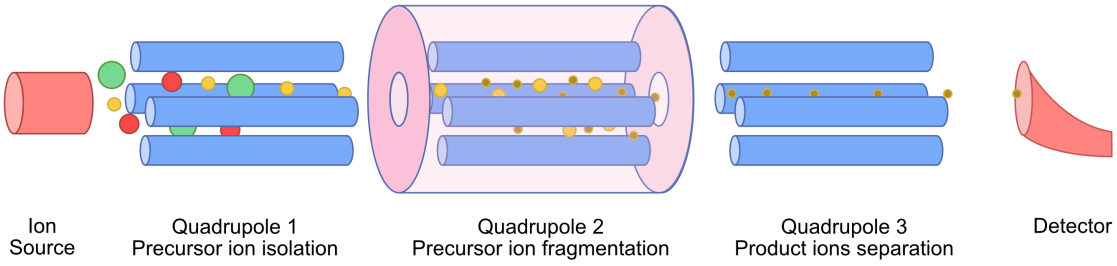

Figure 1.2: Representation of a triple quadrupole performing fragmentation in space. The precursor ion is isolated by the first quadrupole, then fragmented in a collision cell, and the fragments are separated by the third mass analyser. Simpler mass spectrometer only have one mass analyser.

Liquid chromatography adds another dimension to the compound separation (Figure 1.3). In LCMS, this separation is made using High-Performance Liquid Chromatography (HPLC). The sample to be analysed is injected into the stream of mobile phase and passes through the stationary phase, part of the chromatographic column. Analytes are infused into the mass spectrometer for mass separation as they elute from the column. Diverse columns with different stationary phase properties are used. Hydrophilic interaction liquid chromatography [33] (HILIC) columns are used in LCMS metabolomics and separate compounds by increasing polarity. Alternatively, Reversed-phase chromatography methods, which uses a hydrophobic stationary phase is also often used to separate non-polar compounds [34].

1.3.2 Data format

The data produced by mass spectrometers during an LCMS experiment can be very large. The different mass spectrometer manufacturers have developed their own proprietary data format to store and process the data. However, these data formats are not adequate for an

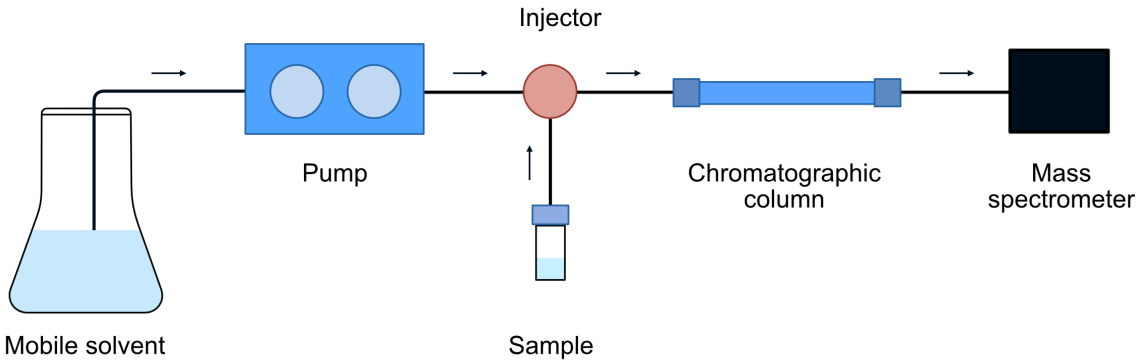

Figure 1.3: Liquid Chromatography - Mass Spectrometry system

academic research environment as they are binary, which make them difficult to read without dedicated software provided by their respective manufacturers. However, several open source data formats have been created over the years in an attempt to provide a unified standard format for MS. Over the past 15 years, two open formats were concurrently developed by the Proteomics Standard Initiative and Seattle Proteome Center, respectively called mzData and mzXML [35, 36]. A joint effort has however emerged since to create a unified open format, mzML [37], which integrates and extend mzData and mzXML data formats. Instrument manufacturers now all provide software libraries to access the data within the binary files and convert it to an open format. This task can be handled by the tool MSConvert [38, 39], part of ProteoWizard Software.

While some specifications such as metadata information change between the different open formats, the LCMS data itself is stored in a similar manner and can be described as a 3-dimensional chromatogram. Figure 1.4 illustrates this data by plotting *m/z* versus the retention time in y and x-axes respectively. The z-axis represents the intensity of the signal corresponding to ion counts, the highest signal being set at 100%. This data represents one polarity only. Two of these data structures are, therefore, present if the instrument is operated in polarity switching mode. Alternatively, positive and negative polarity data can be stored in two separate files.

This complex data structure allows to approaching the data in two different manners. Figure 1.5 illustrates a mass spectrum and the information it contains in the context of LCMS data structure. For each of the time points there is a corresponding mass spectrum containing MS peaks. Those peaks in a mass spectrum represent the molecules that eluted from the column at a specific retention time. Mass spectra become more complex as the number of compounds eluting from the column at the same retention time.

The data can also be approached in a transversal manner from mass spectra in order to look at a single ion (*m/z*) over time. Chromatographic peaks observed in extracted ion chromatograms as shown in Figure 1.6 show the elution of a single ion through the chromato-

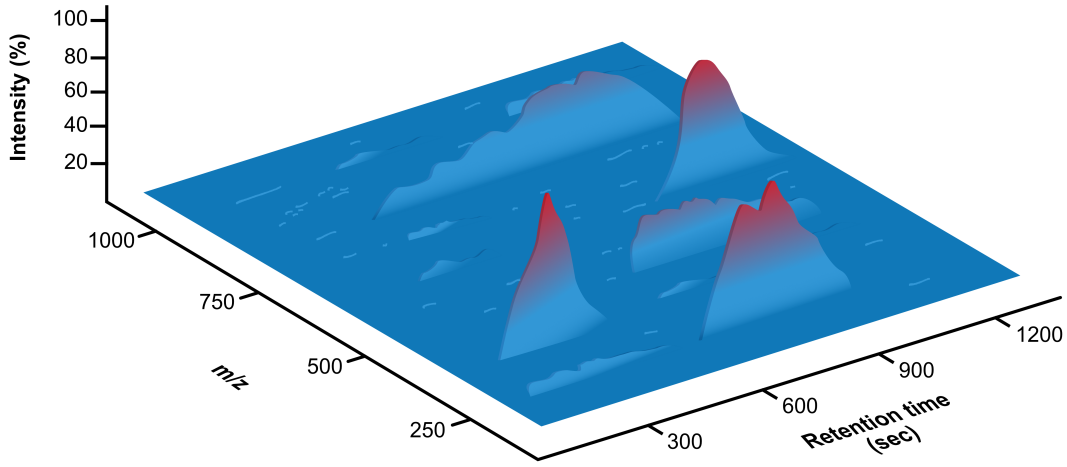

Figure 1.4: Representation of a 3-dimensional chromatogram produced by LCMS data acquisition. The x axis represent the elution time in the chromatographic column (retention time), the y axis represent the *m/z* measured by the mass spectrometer, and the z axis represent the intensity of the detected ions.

graphic column. It can, therefore, show the separation of two species of the same mass but different affinity with the column.

1.3.3 Data processing

Many tools have been developed to support LCMS data processing, while they do not always provide the same features, a common data processing pipeline is conserved across those different tools. The user interface can, however, vary from command lines to a dedicated graphical user interface for stand-alone tools. In the recent years, web-based data processing pipelines have also emerged, providing a graphical user interface through web browsers such as Galaxy based pipelines [40] and overcoming any installation requirements. This section below describes the different steps of LCMS data processing although some tools can provide some variations of this general pipeline.

Peak detection

The peak detection is applied to LCMS data as it reduces considerably the size of the data to handle. During this step, the data structure previously presented is converted to a list of peaks, each entry being characterised by its *m/z*, retention time and intensity. Peak detection is complex, and many tools tend to use the same algorithm to perform this task. CentWave [41], the most widely used peak detection algorithm is implemented as part of XCMS [42] and is based on centroid mode spectra. Other peak detection algorithms are

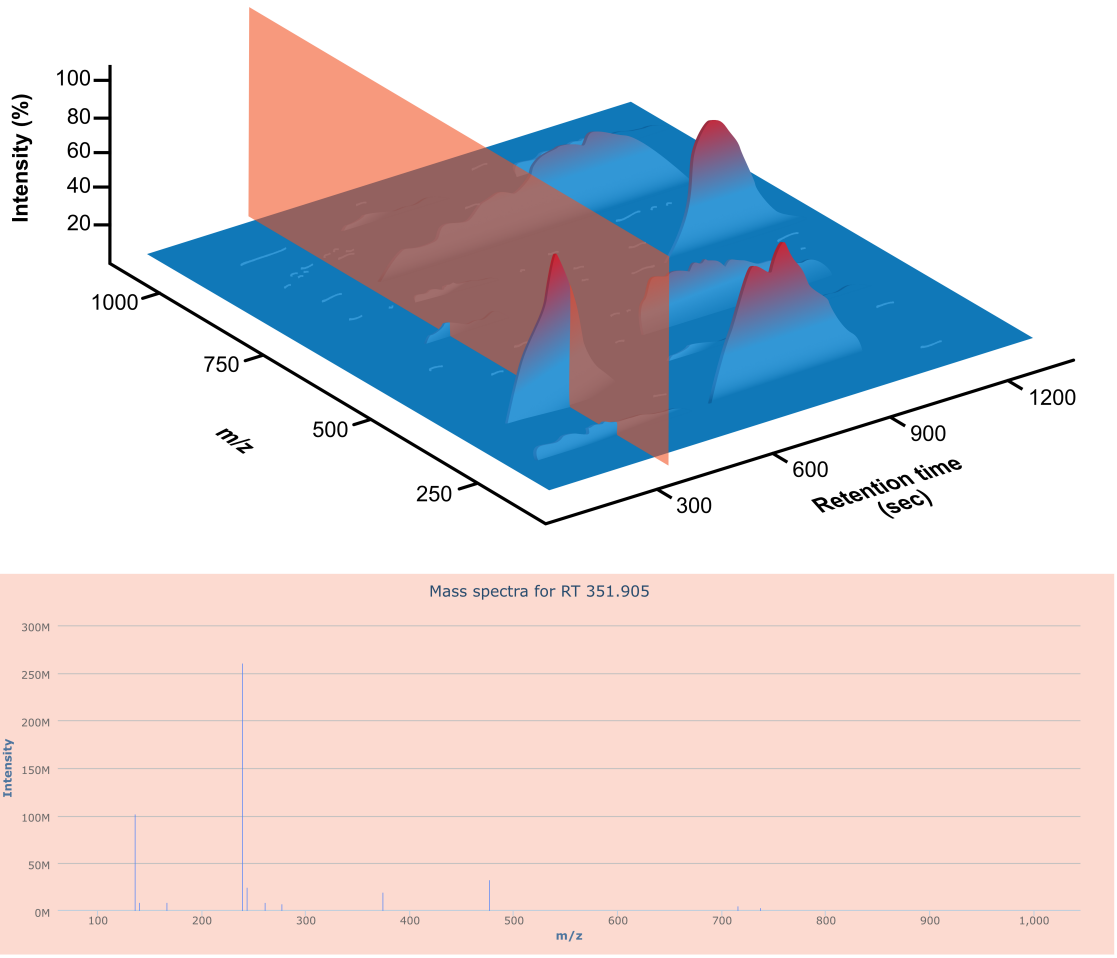

Figure 1.5: Single mass spectrum selected from an LCMS data file. This figure plots the intensity of ions against the *m/z*

available such as MetAlign [43] or CentroidPicker in MZmine [44], but cannot be used by third party tools as they are part of stand-alone software. These algorithms treat each sample spectrum independently and rely on manually defined parameters that have a significant impact on the quality of the peak detection. Some of these parameters are in direct relation to specifications of the instrument used for the data acquisition such as the mass deviation parameter in CentWave which requires knowing the mass accuracy of the instrument. These algorithms are therefore aimed to be used by experienced users for optimal results.

Peak alignment

The vast majority of metabolomics experiments are based on the assumption that differences will be observed between two different experimental group of samples, defined by either biological or technical replicates. This assumption implies that metabolite levels are comparable within and across experimental groups. As the peak detection is performed independently for each replicate sample (LCMS run), matching peaks across LCMS runs corresponding to

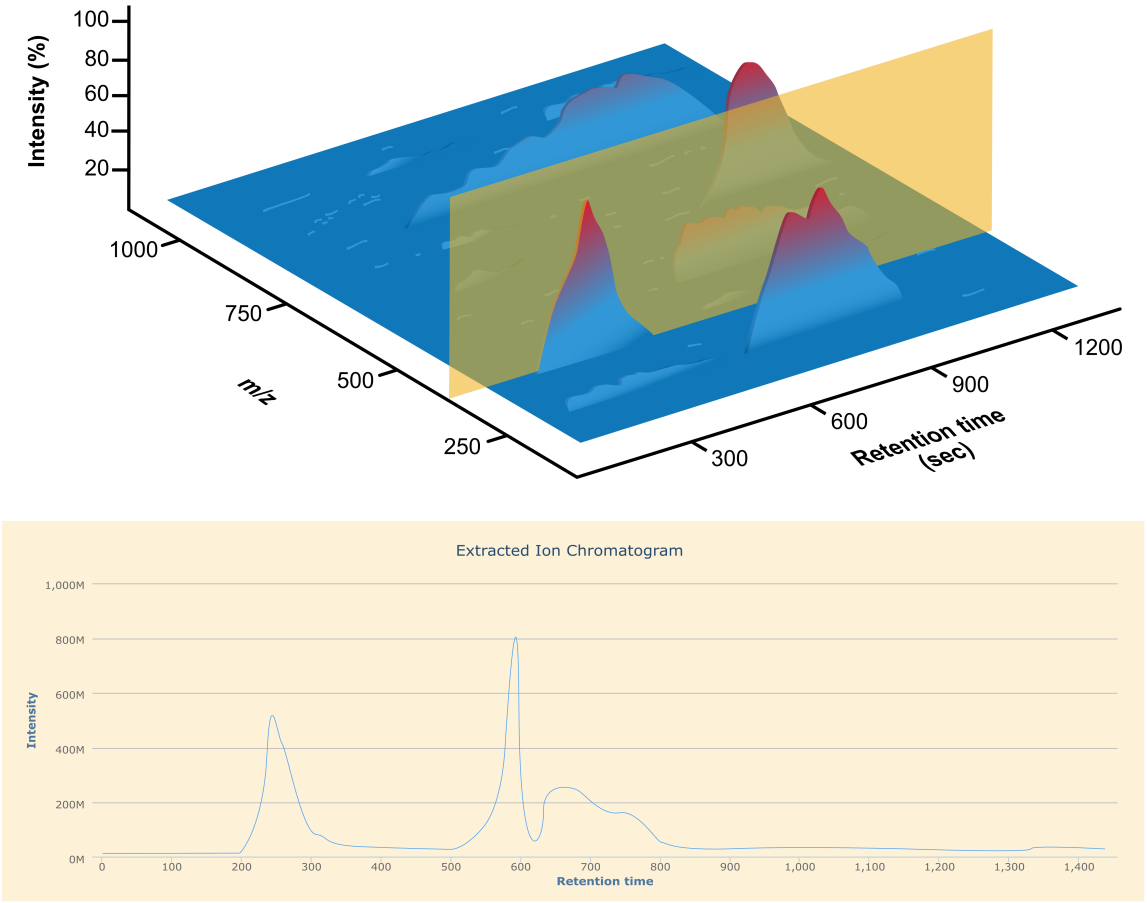

Figure 1.6: Extracted ion chromatogram of a LCMS file. This figure plots the intensity of one ion against the retention time.

the same molecular ion is essential for the downstream analysis. However, analytical platforms used in LCMS can produce data with large, non-linear retention time drift between LCMS runs. The peak alignment step addresses this issue and has been implemented using different methods. It produces as an output a list of *peaksets* containing the aligned peaks from each LCMS run.

Different warping-based alignment methods are widely used for LCMS data processing. These methods attempt to model the retention time drift between runs to correct it. Two main types of warping based alignment have been implemented and are based either on the total ion current (TIC) or the extracted peaks themselves. TIC-based algorithms such as Dynamic Time Warping (DTW) [45], Parametric Time Warping (PTW) [46, 47], Correlation Optimal Warping (COW) [48] and Continuous Profile Mode (CPM) [49] are not used by modern software as they take a reductive approach by using TICs only and ignoring the complex information of LCMS data. These methods were found to be inadequate for LCMS data as they often fail to align overlapping peaks (co-eluting compounds). An improvement of the

COW methods which combine it to a component detection algorithm (CODA) was however developed and showed a higher alignment quality [50]. The same type of approach was taken to improve PTW and DTW algorithms for LCMS data by combining it with CODA [51]. An extension of DTW termed Ordered Bijective Interpolated Warping (OBI-Warp) [52], available in XCMS, shows improved alignment results and is now commonly used.

Alternative alignment methods can also be used such as Direct Matching which compares peaks across LCMS runs based on similarities without warping. Many implementations of this method have been proposed using different similarities measures [53, 54, 55] and are available in various LCMS data analysis tools such as Join Aligner in MZmine [44].

Finally, a simpler labelled LCMS data alignment can also be used; it, however, requires the injection of internal standards in the experimental samples, which increases the complexity of sample preparation.

While several algorithms can suit the task of LCMS peak alignment, it is hard to assess what algorithm provides the best solution. This is in great part due to the lack of comparative evaluation at the time of publication as outlined by R. Smith, et al. [56]. The choice of an alignment algorithm can, however, be made by using recent comparative reviews [57, 58].

Data filtering

Many data filters can be used to remove undesired signal [59]. For example, Reproducibility Standard Deviation filter [60] available in mzMatch [61] helps to eliminate signal that is too variable between replicates. More common filters are available in the different data processing tools such as blank filter, noise filter or a minimum number of detection. The blank filter discards any peak that is higher in the blank samples (generally extraction solvent) than the experimental samples as they can be considered as contaminants. The minimum detection number allow discarding peaks that are present in a limited number of samples, which often correspond to noise signals.

Gap filling

In some cases, peaks can be missing from a peakset due to a misalignment or rejection during peak detection because of a poor shape or high background signal. The gap filling step aims to recover this missing signal directly from the raw files. This step gives better insurance on the true absence of a peak.

Peak grouping

Undesired in-source fragmentation often happens during the ionisation process which results in the production of multiple peaks per metabolite. Similarly, the sample preparation can cause the formation of adducts formed by the adduction of an ionic species such as different salts to a molecule. Beyond the ion suppression resulting from these formations [62], it also results on the production of multiple peaks for a single molecule. Finally, naturally occurring isotopes such as ^{13}C can produce several peaks that follow the isotopic distribution of the element. The signal generated by these products of the precursor ion caused by those different mechanisms are commonly called related peaks.

The peak grouping step attempts to identify these related peaks and group them together with the precursor ion. Different methods often based on known chemical relationships can be used to create these peak groups. For example, mzMatch uses a clustering method based on intensity and peak shapes while CAMERA [63] groups related peaks using multiple integrated methods reconstructing a similarity graph.

This grouping step can be applied to the data at different stages of the pipeline but results in a consistent reduction in the number of relevant peaks which facilitate the peak identification stage.

Peak identification

Peak identification is crucial in order give a biological meaning the data generated. This process attempts to match peaks from a given LC-MC dataset to molecular formulas and compound identities. It is however not a trivial task due to a high number of possible associations between a peak and metabolites. The Chemical Analysis Working Group as part of the Metabolomics Standard Initiative (MSI) created a 4 level scheme to help to report metabolite identification and annotation in a uniformed manner between studies [64]. The first level, considered as highest ranked identification, necessitate a match of a minimum of two independent and orthogonal data relative to an authentic compound analysed under identical experimental conditions such retention time and accurate mass. Authentic compounds data is acquired by running authentic standards mix on the instrument. The second level of identification is based upon spectral similarity with a public spectral library. The level 3 corresponds to putatively characterised compound classes, and the level 4 designate unknown compounds.

In standard untargeted approaches, a finite set of authentic standard compounds is run, which limits the number of peaks that can be annotated as level 1 identification. The majority of the other peaks are identified using accurate mass from public databases, KEGG [65], PubChem [66], HMDB [67] and LIPID MAPS [68] are some examples amongst many available.

However, mass accuracy is often not enough for unambiguous identification [69].

While *in silico* retention time prediction can help with the identification process [70], the most promising avenue for addressing this issue is the use of fragmentation data acquired by tandem MS (MS/MS). Fragmentation data can indeed offer structural information about compounds and therefore provide better support for peak identification. The same approach can be used for the identification process, matching fragmentation spectra against publicly available libraries. Many libraries are available with different degree of curation, matching options, and a varying number of spectra. MassBank [71] and ChemSpider [72] figure among the most widely used spectral databases.

Peak identification process is improving every day as spectral libraries cover an increasing number of compounds, it remains, however, one of the biggest challenges the metabolomics community has to overcome to lead to a better data interpretation in biological context.

Statistical analysis

A normalisation step is often required proceeding to the differential analysis of the dataset. The complexity of this task is highly dependent upon the size and the property of the dataset. Over a certain number of samples analysed, the data collection needs to be performed in separate batches before being merged into one large dataset. This procedure results in biased dataset values due to the variation of LCMS platforms over time. Solutions proposed are still in their infancy although the problem has been addressed many times over the past few years [73, 74]. Several methods for single batch data normalisation use different approaches that can be divided into two main approaches [75]. Methods-driven approaches use internal standard material references to base the normalisation upon. The standard used as reference rarely cover all metabolite classes present in the samples which limit the normalisation efficiency. This method is also not cost effective as the stable isotopes used as internal standards are expensive. Data-driven approaches are the most widely used normalisation methods and are based on the assumption that most metabolites produce a constant signal across samples, these methods have the benefit not to require to know the identity of the metabolites.

Once normalised, statistical analysis such as ANOVA, t-test, false discovery rate [76] and principal component analysis can be performed.

1.3.4 Data analysis platforms

Data generated by LCMS experiments is very complex, and its visualisation is essential at many steps of the analysis pipeline. Visualisation of the overall signal or raw files produced by the instruments is well supported by proprietary software provided by manufacturers.

This software also provides search and curation tools to explore the data and revealed to be technical. They are therefore aimed at trained users which are expert in the field. Similarly, some software also offers visualisation tools for open source raw data formats [77, 44]. Most software, however, provides visualisation tools corresponding to the specific analysis tasks they support. For instance, mzMatch supports the visualisation of extracted peaks with PeakML Viewer, and XCMS allow the visualisation of peaks before and after alignment.

There is, however, no standard when it comes to visualisation of data analysis results of metabolomics experiments. The typical representation of the data being a matrix where each row corresponds to a metabolite or an unannotated peak and each column a biological sample, and each matrix entry the intensity value of a metabolite in a sample. This very crude data representation is very limiting for the interpretation and omits major biological and statistical related information. Many attempts have been made to organise the results in a coherent manner to highlight the different type of information connected to the metabolites. These efforts are specifically made in end-to-end data analysis software which integrates all processing steps. IDEOM [78] proposes an organisation into tabs in an excel spreadsheet (Figure 1.7), with one of them summarising the metabolites found in the dataset along with t-test p-values and biological pathway information.

	A	B	C	D	E	F	H	I	J	K	L	M	N	O	P	Q	R
	Sort	Trend Sort	Import Peaks	Search	Tools	Graphs	Export										
1	Mass	RT	FORMU	Isome	Putative metabolite	confidence	Pathway	max intensity	C4	C9	C24	PHA4	PHA9	PHA24	ttest: C4	ttest: C9	ttest: C24
5	226.99	10.44	C5H9NO5S2	1	2-(sulfomethyl)thiazolidin	7	coenzyme M biosynthesis	144130	1.00	**8.35**	26.46	0.00	5.49	18.07	NA	0.010609	0.112343
6	147.04	6.474	C5H9NO2S	1	Thiomorpholine 3-carboxylate	7	0	1732878	1.00	6.60	0.40	4.45	1.55	0.43	1	0.217808	0.453235
7	161.07	20.37	C6H11NO4	10	N-Methyl-L-glutamate	8	Methane metabolism	268230	1.00	5.84	6.96	2.63	6.21	**9.49**	1	0.200112	0.22289
8	416.16	3.424	C26H24O5	1	Calophyllolide	7	0	68932	1.00	4.70	6.67	2.17	3.56	2.10	NA	0.068231	0.087758
9	211	9.664	C5H9NO4S2	1	3-butenyl-thiohydroximate-O-	7	0	82389	1.00	4.50	10.18	3.04	4.23	6.92	NA	0.052372	0.08266
10	126.01	9.562	C2H7O4P	2	2-Hydroxyethylphosphon	8	Aminophosphonate	1493763	1.00	4.02	9.45	1.69	1.93	7.07	1	0.099159	0.172278
11	146.06	9.409	C6H10O4	16	(R)-3-Hydroxy-3-methyl-2-oxopentanoate	8	Valine, leucine and isoleucine	352268	1.00	3.39	0.93	0.92	3.82	0.65	1	0.438898	0.732211
12	120.01	10.87	C3H4O5	1	2-Hydroxymalonate	7	0	18392	1.00	3.28	4.06	2.08	2.81	3.70	NA	0.214195	0.083259
13	179.08	21.49	C6H13NO5	10	1-Amino-1-deoxy-scyllo-inositol	7	Streptomycin biosynthesis	1067956	1.00	3.22	2.19	5.62	**5.73**	9.47	1	0.201368	0.222634
14	219.07	7.567	C8H13NO6	4	O-Succinyl-L-homoserine	6	Methionine metabolism	1941480	1.00	3.20	7.01	1.61	3.71	6.17	1	0.148464	0.121937
15	324.23	3.532	C19H32O4	3	Decylubiquinol	7	thiosulfate oxidation III	178447	1.00	2.99	1.41	1.38	4.26	1.42	1	NA	0.736301
16	132.09	14.61	C5H12N2O2	6	L-Ornithine	10	Arginine and proline	3516386	1.00	2.62	**3.34**	1.02	1.55	**2.74**	1	0.154477	0.002837

Figure 1.7: Comparison tab showing the results of LCMS data analysis in IDEOM. The columns A and B display the Mass and retention time of the peaks. Columns C to I show information about the identity of the compound corresponding to the peak such as the formula, the metabolite name or its pathway. Columns J to O show the fold changes between the different experimental conditions. The other columns give statistical information.

Other standalone applications propose similar approaches such as MAVEN [79] which give information about the biological compound under investigation in its pathway view.

More recently, web-based software has been developed which enable an easy access with no

installation requirements for the user. This very accessible software has attracted considerable interest from the biological research community and helps disseminate and systematise the use of metabolomics in biological science. Some of these programs support the entire data analysis such as XCMS Online [80] or Workflow4Metabolomics [81], others tend to focus particularly on a particular task. MetaboAnalyst [82] for example, offer extensive statistical tools for metabolomics data. XCMS Online is the first end-to-end data analysis software which attempts to allow non-experts to perform their own data analysis [83]. It was in part achieved by introducing simplified parametrisation and interactive visualisation (Figure 1.8). Other software such as OpenMS [84] which was first developed for proteomics provides now support for metabolomics data analysis.

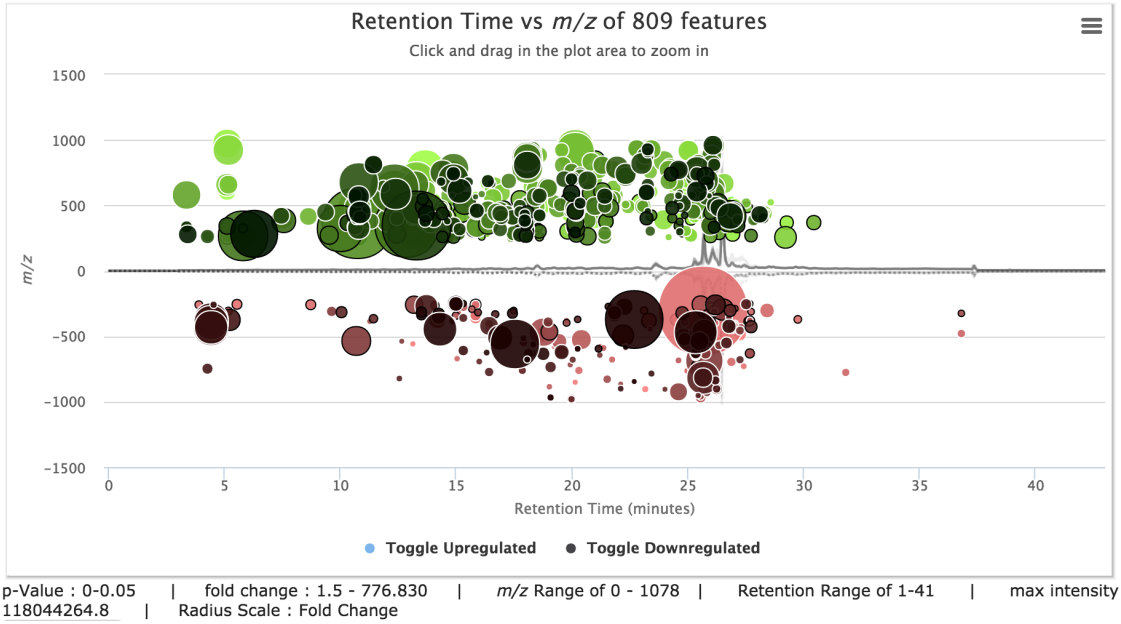

Figure 1.8: Innovative interactive visualisation tool available in XCMS Online. In this figure, *m/z* is plotted against retention time. Each bubble correspond to a feature, the colour is indicative of the directionality of the fold change between two experimental conditions, and the size is indicative to the extent of the fold change. The intensity of the color correspond to the statistical significance.

Shortcoming of current data analysis platforms

This section presents the shortcoming of the current software attempting to support end-to-end LCMS metabolomics data analysis, from raw data to biological interpretation.

The first common limitation of the tools presented in the previous section is the level of understanding required for the user to perform an analysis. Indeed, many settings needing an in-depth knowledge of LCMS technology has to be manually entered by the user. Mass and retention time window for feature detection, alignment parameters, are some of many

examples. This requirement currently limits the usage of these tools to mature audience forcing inexperienced users such as biologists or clinicians to outsource their metabolomics analyses to bioinformaticians.

The second limitation shared by most tools is the static and fragmented structure of the applications. While modular designs can be useful to expand the feature set that software has to offer responsively and can provide the user with many analysis options, the lack of connectivity between these modules results in a fragmented overall architecture. Two major problems arise from this approach. First, tools such as MetaboAnalyst present the data analysis pipeline in the form of functional modules that the user has to choose from, which implies some level of understanding from the user in order to run modules in a coherent, sequential manner. Other tools such as IDEOM present the workflow as an integrated pipeline. However, user intervention is still required at each step of the pipeline which limits the turnaround time for a complete analysis of the data considerably. Finally, Galaxy based software also necessitates basic understanding on how to organise a data analysis pipeline.

The same static approach is often taken with regards to data visualisation. While most software provide features to generate figures such as Principal component analysis (PCA) or volcano plots which can be interpreted on their own without surrounding information, the same method is often taken to present extracted-ion chromatograms (EICs) or mass spectra. This approach of generating static pictures to display specific information isolates the data from the general context of the analysis which makes it harder to interpret. XCMS Online started addressing these issues by organising the analysis pipelines into jobs and creating dynamic and interactive visualisation tools which can help users in better understanding their data.

Currently, available tools can be divided into two main groups, stand-alone software and web-based applications. Stand-alone applications necessitate the local installation of the software and its library dependencies on personal computers. This task is often difficult for users with no fundamental skills in computer science. Collaborative work within such environments can also become a challenge as it requires every party involved to have access to the same version of the same tool. Moreover, sharing large metabolomics datasets is not a trivial task due to the size of the raw and processed data. Web-based applications do not suffer from these limitations as they usually offer sharing features and direct access to the data through a web browser.

One of the key components to enable users to extract meaningful biological insight from metabolomics datasets is the biological context under which the results are investigated. While the fragmented structure previously discussed substantially limits this interpretation process, some tools are beginning to integrate pathway enrichment and analysis tools. However, many still require the use of third-party software to replace metabolomics data into

a larger biological context. Manual export, formatting, and import of the data is required whether these applications are web-based [85, 86, 87, 88], stand-alone [89] or Cytoscape plugins [90, 91], which creates yet another barrier for inexperienced users to interpret their data entirely.

1.4 Programming languages, libraries and frameworks

Many programming languages are available for developing bioinformatics tools. The choice of a programming language can depend on many parameters such as the performance required (i.e. computation time, hardware requirement), the development time and the aim of the tool developed. However, Perl and Python have been the two languages of choice for bioinformaticians as they need fewer lines of code than other languages such as C, C++ or Java. They, therefore, enable faster development. Those two languages benefit now from a wide range of biology-oriented libraries which provide many commonly used algorithms in the field. Python presents, however, advantages over Perl due to a syntax more straightforward and less permissive, which facilitate the development of reusable scripts and collaborative work. All languages are, however, used in bioinformatics, and the choice depends on the aims and requirements of the tool or script being developed [92].

Statistical languages are also commonly used in bioinformatics, MATLAB (matrix laboratory) [93] and R [94] are the two most popular languages. MATLAB presents the disadvantage of being a proprietary language which makes it expensive. For this reason, R, which is open source, can reach a wider audience than Matlab and is often preferred by bioinformaticians.

Nowadays, web technologies are also commonly in bioinformatics as they enable the straightforward development of user interface available through web browsers. Frameworks have been developed over the past decades to standardise and ease the development of web applications. Like the programming languages, the choice of a web framework is made according to the project, its aim and its target audience. For example, Shiny [95], an R web framework, allow the rapid development of a web interface around an R script. However, larger applications tend to use frameworks such as Ruby on Rails [96] or Django [97] which provide a more structured environment and enable scalability.

The choice of languages is always closely related to the objective of the tool being developed and the available libraries and frameworks. In-house scripts aimed to be used by a bioinformatics laboratory will not necessarily need a user interface; applications with an audience of biologists, however, require streamlining and interfacing to facilitate their use.

1.5 Biological networks

High throughput omics technologies allow the large-scale study of the systems organisation of organisms. Each omics technology attempts to describe the state of a particular layer of a system and the interaction between their constitutive components. Representing these interactions using networks helps to understand the different relationships between omics components. Thus, biological networks are often used to understand the process occurring in a system. Each omics layer can be represented by a different network to inform on various interactions. Gene-gene interaction networks, for example, are often used to attempt to understand the different relationship between genes [98]. In transcriptomics, gene co-expression networks are widely used to understand the processes regulating the expression of genes [99]. Other types of networks are commonly used to explore genomics and transcriptomics data such as co-localization or gene regulatory networks [100].

While these networks offer comprehensive support to study the interactions occurring in a biological system, they need to be processed to extract meaningful biological information. For example, as illustrated in Figure 1.9 a protein-protein interaction network can be formed of thousands of nodes highly connected with one another, which limits the amount of information that can be extracted from it. Reducing the size of the network can, for instance, be done by a semantic enrichment using Gene Ontology annotations [101]. The resulting network would highlight proteins sharing particular biological processes, molecular functions or found in the same cellular component; and interacting with one another, conveying a greater biological meaning than the initial network.

At the metabolome level, genome-scale reconstructions of metabolic networks [102, 103] can be used for studying the flux of metabolites within a system using flux balance analysis [104]. This type of network is also used to make different predictions on biological systems using *in silico* constraints-based approaches [105]. Figure 1.10 illustrates the genome-scale reconstruction of the human metabolic network [106].

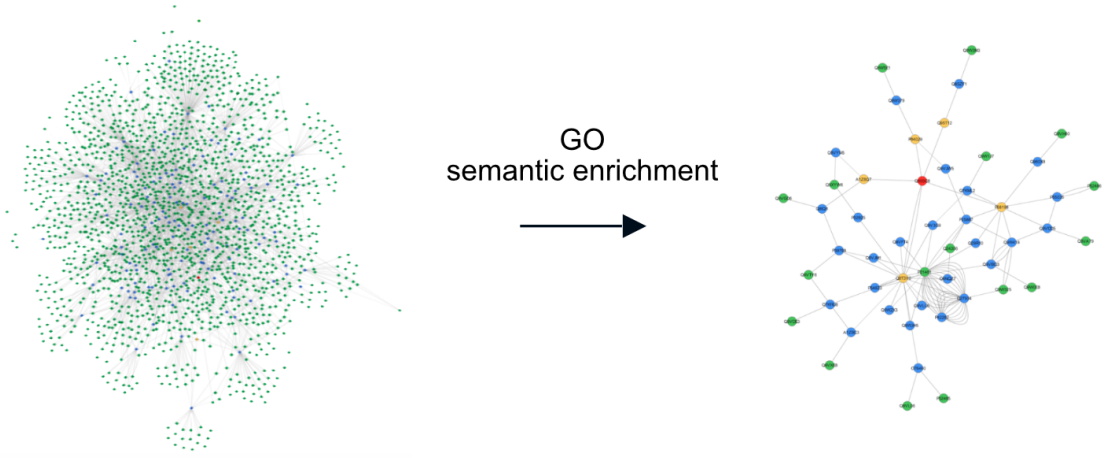

Figure 1.9: Example of a Gene Ontology semantic enrichment of protein-protein interaction network. On the left, a three degrees separation protein-protein interaction network was reconstructed from one seed protein (in red, in yellow are proteins with one degree separation, in blue two degrees, in green three degrees). On the right, the same network was reconstructed using a Gene ontology semantic enrichment to keep only proteins involved in the same molecular processes. The network was reconstructed for illustration purposes, the seed protein and ontologies were selected randomly. The network was reconstructed using a python script developed by the author and visualised using Cytoscape.

Figure 1.10: Genome-scale reconstruction of the human metabolic network using MetExplore. The network as such is not informative but can serve as a support to study flux data

1.6 Related work

The work presented in the following chapters aims to better support researchers in their LCMS metabolomics experiments. From data capture to result interpretation, the tools implemented and presented here provide platforms using state-of-the-art technology to facilitate the LCMS data capture, data analysis and interpretation. Every result chapter focuses on a different stage of the LCMS metabolomics workflow, the specific aims and objectives are defined in the related work section of each chapters.

Chapter 2

Materials and methods

The research discussed in this document mainly involve the development of new approaches to analyse, visualise and interpret metabolomics and to some extent omics data. These approaches are based on the development of new tools to support the different operations to perform on the data. The section below, therefore, outlines the programming languages, the programming libraries, the existing tools that were used for the development of the different part of the software and the data analysis presented in Chapter 3, Chapter 4 and Chapter 5.

2.1 Software engineering

The software presented in this document was developed following the agile software development method. Requirements were initially captured through extensive interaction with collaborators of Glasgow Polyomics (GP) metabolomics facility, and gathering feedback from Glasgow Polyomics data analysts. Agile development method was put in place once the first working prototype was developed. A pool of 10 test users with different background were given an early access to the tool to analyse their own LCMS data and to provide feedback on the features and report issues encountered.

Acceptance testing was performed in the form of a one day workshop, gathering test users as well as new users. Two acceptance tests were organised during the development of the software presented in this document, they allowed to refine different part of the project from the data structure to the user interface.

All software developed as part of this work were put under version control using Git and a private GitLab repository hosted on GP servers. A production environment and several development environments were created for each tools, every features newly developed was tested before being deployed on the publicly available production server.

Besides being under version control, the tools were encapsulated into docker containers to facilitate their deployment. This encapsulation was not carried out by the author and will, therefore, not be discussed in this document.

Unless specified in the text, the work presented below has been carried out by the author.

2.2 Data format

Several data formats were used for different purposes. The metabolomics data file format that the developed software uses is "mzXML" format. All instruments used for Liquid-Chromatography Mass-Spectrometry data acquisition produce result files in a different type of proprietary format. This files can be converted to the "mzXML" open file format by using the freely available tool ProteoWizard [38]. mzML format [99] was used for fragmentation data.

Web transactions involving data transfer use Javascript Object Notation [100] (JSON) which is a data-interchange format commonly used for web data exchange as it is language independent. This format is used for client-server asynchronous communication.

The data analysis pipeline used in Chapter 3 creates intermediary "PeakML" [61] files that contain pre-processed data of "mzXML" or "mzML" files.

One XML [107] based exchange format was also created as an intermediary data format between the software developed and the data analysis pipeline, and explained in Chapter 3. The exact purpose of the file format (pimpxml) is detailed in section 3.4.5.

2.3 Data analysis pipeline

The data analysis pipeline of the software described in chapter 3 is implemented in R [108, 94] and based around XCMS [42] for feature detection and mzMatch.R [61] for metabolomics data pre-processing tasks. mzMatch.R uses backend functions implemented in Java through the rJava library. The analysis pipeline also uses a collection of other R libraries; the full list is available in Appendix A.1. Extra pipeline functions are implemented in R.

The data analysis pipeline is run asynchronously by the implemented software using Celery [109]. Celery is an asynchronous task queue allowing both scheduling and concurrent tasks to run on several worker nodes. RabbitMQ [110] is used as the message broker for Celery. Reversed communication from the pipeline to the program is done through a dedicated "XML" format.

2.4 Web framework

The software and tools presented in Chapter 3 and 4 are developed in Python 2.7 [111] using Django (version 1.7) [97] web framework. Django is an open source web framework written in Python, which follows the Model View Template (MVT) architectural pattern. It is developed and maintained by the independent Django Project Foundation as a 501(c)(3) non-profit. Django consists of an object-relational mapper that mediates between data models and a relational database, a web templating system with a HTTP requests processor, and a regular-expression-based URL dispatcher. A MySQL [112] relational database is used to store data in production environments and SQLite on development environments. Nginx web server [113] is used in conjunction with Django in production environments, interfaced by Gunicorn [114], a Python Web Server Gateway Interface HTTP server written in Python. As mentioned in the previous section, long running asynchronous tasks and queueing systems are handled by Celery.

Mathematical operations are performed on the server side (in Django) using NumPy [115], SciPy [116] and Scikit-learn [117] Python libraries. NumPy extends Python support to large multidimensional arrays and matrices, and high-level mathematical functions. SciPy is built on NumPy array objects and expand the mathematical and scientific functions. Scikit-learn implements visualisation, preprocessing, cross-validation and machine learning algorithms.

Communication between Python and R is performed using rpy2 [118] to enable the use of XCMS functions in Python. Rpy2 is a python library interfacing Python with R using NumPy array objects.

The full list of Python libraries and Django plugins used for the implementation of the program and tools presented in Chapter 3 and 4 are available in Appendix A.2.

2.5 Data visualisation

The user interface of the tools presented in the different Chapter 3 and 4 are developed using common web standards such as HTML and CSS. The user interface represents the template layer of Django web framework. An extra layer developed using JavaScript programming language to create an interactive user interface. This layer is based on the jQuery JavaScript library and uses AJAX (Asynchronous JavaScript and XML) web development techniques to improve perceived response time and create a dynamic user interface. Charts and plots are designed using Highcharts and D3.js libraries to allow interactions such as zooming and download features. The tables are based on DataTables library for interaction purposes. The full list of JavaScript used is available in appendix A.3

Chapter 3

A semi-automated pipeline for untargeted metabolomics

3.1 Introduction

Metabolomics is a relatively new field which requires the combination of different scientific disciplines. From analytical chemistry to systems biology, metabolomics combines complex analytical applications to advanced bioinformatics and biochemistry expertise. This interdisciplinary breadth of metabolomics creates tremendous challenges in making it approachable to the scientific community as very few people are experts in all of those fields. Performing adequate and well-designed experiments to obtain good quality data that can be taken forward for analysis and interpretation becomes, therefore, an obstacle for non-experts. Indeed, the high complexity of Liquid Chromatography Mass Spectrometry data compared to other 'omics' data necessitates expert knowledge in the field to plan an experiment to be able to process the data post acquisition. This concept of post-acquisition data analysis to reduce noise or filter unprocessed raw data is also poorly understood by biologists who need to be guided through the design of their experiment as well as through the different data analysis steps. The biggest challenge, however, lies in the data analysis and interpretation of the results [119, 120]. Indeed, due to data complexity and currently available tools, the analysis of metabolomics data is usually performed by expert bioinformaticians or data analysts with a strong knowledge of the data structure, format and analysis process as well as advanced computer skills. These challenges can, however, be addressed by creating a tool to accompany and guide biologists from designing their experiments to interpreting their data. In order to overcome these issues, the tool needs to streamline the data analysis process into a semi automated tool reducing the necessary user interventions; presenting a simple 'step by step' pipeline to allow users them to proceed to their own analysis, but also assist them in the data interpretation. The second point can be achieved using data visualisation techniques to

provide the user with contextual information and create a self-learning environment.

Metabolomics is a rapidly evolving field, analytical tools and algorithm need therefore to adapt to meet new requirements continuously. This is usually achieved in two different manners, existing algorithms and tools can be modified to meet the new requirements and provide an alternative data analysis, or for entirely new approaches, new modules need to be created and incorporated into an existing pipeline. For this to happen, however, the tool needs to be developed in a structured and highly scalable manner. This can be achieved by creating modular software in which the modification of one module would not affect the rest of the software; such orthogonal design would allow an easy addition or removal of modules from the pipeline to adapt to new requirements in a responsive way.

3.2 Related work

At the commencement of this project several metabolomics data analysis pipelines necessitating different levels of understanding of the metabolomics and bioinformatics fields were available. XCMS [42] introduced in 2006 as an R package allows the analysis of untargeted metabolomics. While the tool provides all features necessary for full data processing from peak detection to statistical analysis, it requires prior knowledge in programming and is therefore limited to the use of bioinformaticians. mzMatch [61], a Java and R library provide a collection of small tools that enrich the features available in XCMS. The tool provides filters in order to improve data processing such as Relative Standard Deviation filter during peak grouping. mzMatch also provides a peak annotation tool which allows basic biological interpretation of the dataset. IDEOM [78] was the first tool introducing more extended biological interpretation feature. IDEOM is a wrapper around XCMS and mzMatch presented in the form of an Excel spreadsheet. A collection of macros allows data processing from peak detection to biological interpretation through calls to XCMS, mzMatch and independent algorithms. IDEOM displays the results of the data processing as tables within Excel; it provides pathway information and export functionalities to analyse the results further using external software. MZmine [44], an alternative to XCMS was first released in 2006 as a stand alone application for Mass Spectrometry data analysis. MZmine is written in Java and presented to the user in a dedicated user interface; it allows users to analyse MS data from peak detection to statistical analysis without the requirements of prior knowledge in programming once installed. Results can be exported for further analysis such as data interpretation using external tools.

The first and shared limitation of these tools is the installation requirement. The installation process of these tools requires advanced knowledge in informatics systems and can present significant challenges to biologist without skills in computer science. MZmine and

IDEOM give the advantage of providing a dedicated user interface to interact with the software once installed. However, analysing LCMS data using these programs still requires extensive knowledge of the underlying data to process it from end to end. All the tools also require user interventions at each and every step of the data processing to progress in the analysis pipeline; this limits the time efficiency of the analysis considerably. Other limitations related to stand-alone software such as specific hardware requirements for running intensive processing task also arise from these types of tools. Another major limitation for biologists to use the presented software is the basic capabilities they offer for assisting the user in their data interpretation and providing biological context. This critical step of transforming LCMS data into valuable biological insight currently necessitate third party software and the intervention of biochemistry experts to provide the biological context in which the data should be interpreted. Finally, as metabolomics data is extremely complex, the size of raw data files and analysis results present a barrier to collaborative projects. Although tools exist to transfer big data files, collaborating on a metabolomics data analysis project requires every party involved to have the same software installed on their personal computers to make collaborative approaches to metabolomics studies possible.

The work presented in this chapter will, therefore, try to answer these research questions:

- Can bioinformatics tools support non-expert users in the analysis and interpretation of metabolomics datasets?
- Can software solutions be scalable enough to support the rapid expansion of the metabolomics field and its ever growing requirements?
- Can software solutions overcome issues related to big data and enable world-wide collaboration in the field of metabolomics?

Five main project aims have been drawn to attempt answering these questions:

1. Support end users in their metabolomics data capture and analysis.
   Objective 1: Develop an installation free tool with user friendly UI to allow researchers with no computing skills to set up their own metabolomics data analysis.
2. Streamline the data analysis pipeline to limit or eliminate the need for user interaction after initial data capture.
   Objective 2: Create a wrapper and data exchange format to enable the encapsulation of data analysis pipeline within the developed tool.
3. Develop a modular tool to allow responsive feature integration.
   Objective 3: Decouple the various part of the tool using object oriented and model-view-controller design pattern to enable module integration.

4. Support end users in the interpretation of the results of the metabolomics data analysis.
   Objective 4: Develop a metabolomics data specific exploration environment to enable analysis results visualisation and interpretation within the developed tool.

5. Enable end users to collaborate by easily sharing metabolomics data from study design to analysis results.
   Objective 5: Develop a user session system with sharing capabilities.

The source code of the tool developed in the context of this work and described below is freely available on GitHub at https://github.com/RonanDaly/pimp and licence under GPL. An instance of the tool running on Glasgow Polyomics servers is available at http://polyomics.mvls.gla.ac.uk, access is freely available on request, 50 GB of space is allocated for data storage, with unrestricted number of samples and analyses per user. Over 60 active users are currently using the tool (September 2017), which has been used for LCMS data analysis in published work [121, 122]. The tool has been published as an application note in bioinformatics [123].

3.3 Integrated metabolomics workflow

3.3.1 Data analysis workflow

The aims of the project specify that the software solution needs to support end users for the entire metabolomics workflow, from experiment design to data interpretation. To achieve this aim, a clearly defined integrated metabolomics workflow was designed and is presented in Figure 3.1.

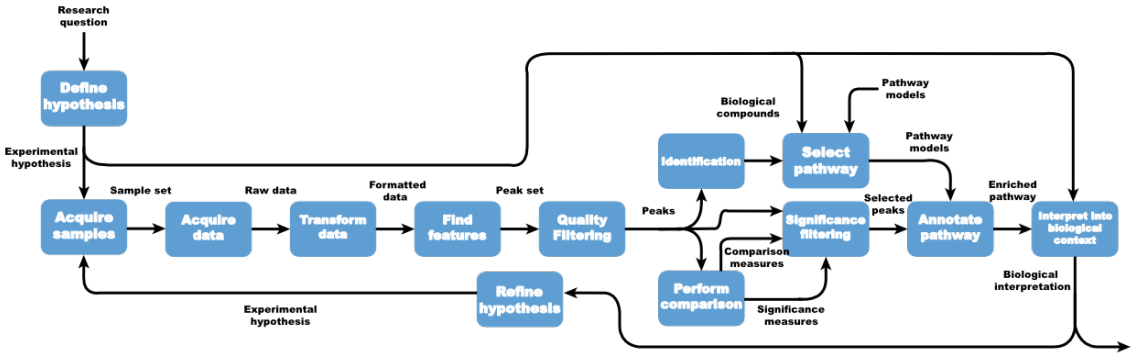

Figure 3.1: Model of an integrated metabolomics workflow from hypothesis generation to biological interpretation.

Most of the limitations are data dependent and appear to be downstream of the data acquisition step in the metabolomics workflow. As shown in Figure 3.1, the data processing requires

many steps from data acquisition to interpretation, and most of these steps require the user intervention in the existing software solutions. However, these steps can be automated and run sequentially: all parameters and information required from the user can be captured at once at the beginning of the processing pipeline. Another major limitation appears to be at the very end of the pipeline and concerns the data interpretation. Interpreting the data within their biological context currently, requires exporting the data to other tools completely separate from the data processing pipeline. Integrating this support within the pipeline and providing visualisation tools to help the user in the interpretation of the data could overcome this limitation. Finally, although the experiment design comes before the data acquisition, guiding the user on how best to design their experiment during the data capture step after the data acquisition could help biologists understand how to perform metabolomics experiment in an optimal manner. Figure 3.2 shows the different part of the workflow that can be improved by addressing these limitations.

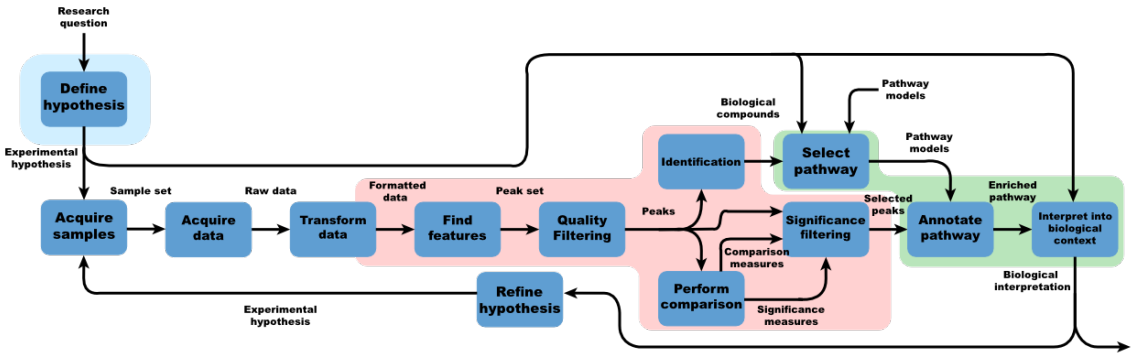

Figure 3.2: Area of limitation in a standard untargeted metabolomics workflow that need support. Highlighted in light blue is the hypothesis definition that need guidance support. Highlighted in red is the data processing which require centralised data capture and streamlining. In green is biological context and interpretation limitation that need integration within the analysis software, and visualisation tools required to support the user in the interpretation.

3.4 Untargeted metabolomics pipeline

3.4.1 Data structure

All the aims of the project rely on having all data, either captured or computed, accessible at all time. The data underlying the entire metabolomics workflow needs, therefore, to be centralised in a common structure. Sharing data between users (aim number 5) is highly dependent on the data structure and is addressed in this section. The modularity of the tool being another key objective of the system (aim number 3), the data structure needs to follow a modular design to support scalability and rapid development. The data structure proposed below is implemented as a relational database developed using MySQL.

The structure is organised in modules to separate the different types of data. Captured data and computed data have been identified as being the two main data types. Those two data types are then organised in sub-modules to form a coherent data structure supporting the capture of all the information and parameters required for an untargeted metabolomics experiment. Figure 3.3 shows the general organisation of the data structure implemented in PiMP. The following modules structure the data capture: projects, fileupload, groups and experiments; the computed data which form the results of the analysis pipeline is stored in the data and compound modules.

The first module called "projects" showed in Figure 3.4 allows the recording of a project's metadata such as its name, creation and edition dates as well as its owner. The module also captures the users that are granted access to the project through the UserProject table, recording also the level of permission a user may have to an individual project.

The second module represented in Figure 3.5 allows the storage and organisation of raw files generated by the instrument. The user may upload two types of samples, hence the separation of the module in two similar structures. The first type of sample supported and simply called "sample" corresponds to the biological samples of the experiment. Each sample when run on the instrument can contain either one or both positive and negative polarities, each polarity being contained in a different file. These files describing the same sample are stored using the "file" table and are organised using the "SampleFileGroup" joining table. The other sample type stored by the table "CalibrationSample" is used for quality control purposes; it is used to store and organise pooled, blank and external standard samples. The main difference with the biological samples is that the standard files can be stored in csv file containing both polarities. This difference is reflected by the "data" field in the "StandardFileGroup" table. This table and its equivalent for the biological samples, the "SampleFileGroup", also allow defining the format of the file that is stored. Finally, the "Curve" table is used to store the total ion chromatogram of each sample. This table was created for optimisation purposes as the TIC is also accessible from the file itself but requires more time and processing power than a simple query.

The "groups" module (Figure 3.6) captures the experimental information of a particular study. This adds an extra layer to the organisation of the samples. Two primary information is stored in this module which is the levels and factors respectively stored in the Attribute and Group tables. The factor represents a category of a biological sample and the level its condition within the category. For instance, if a factor is "gender", the level could be "male", "female" or "undefined". To be flexible, sample entries are attached to the attribute table through a joining table. This structure allows the storage of one level per factor for each sample. For example, sample A could be annotated with the level "male" under the factor "gender", and "time 0" under the factor "time". One sample can only be attached to one level under a specific factor; however, the number of factors is not limited to allow the definition

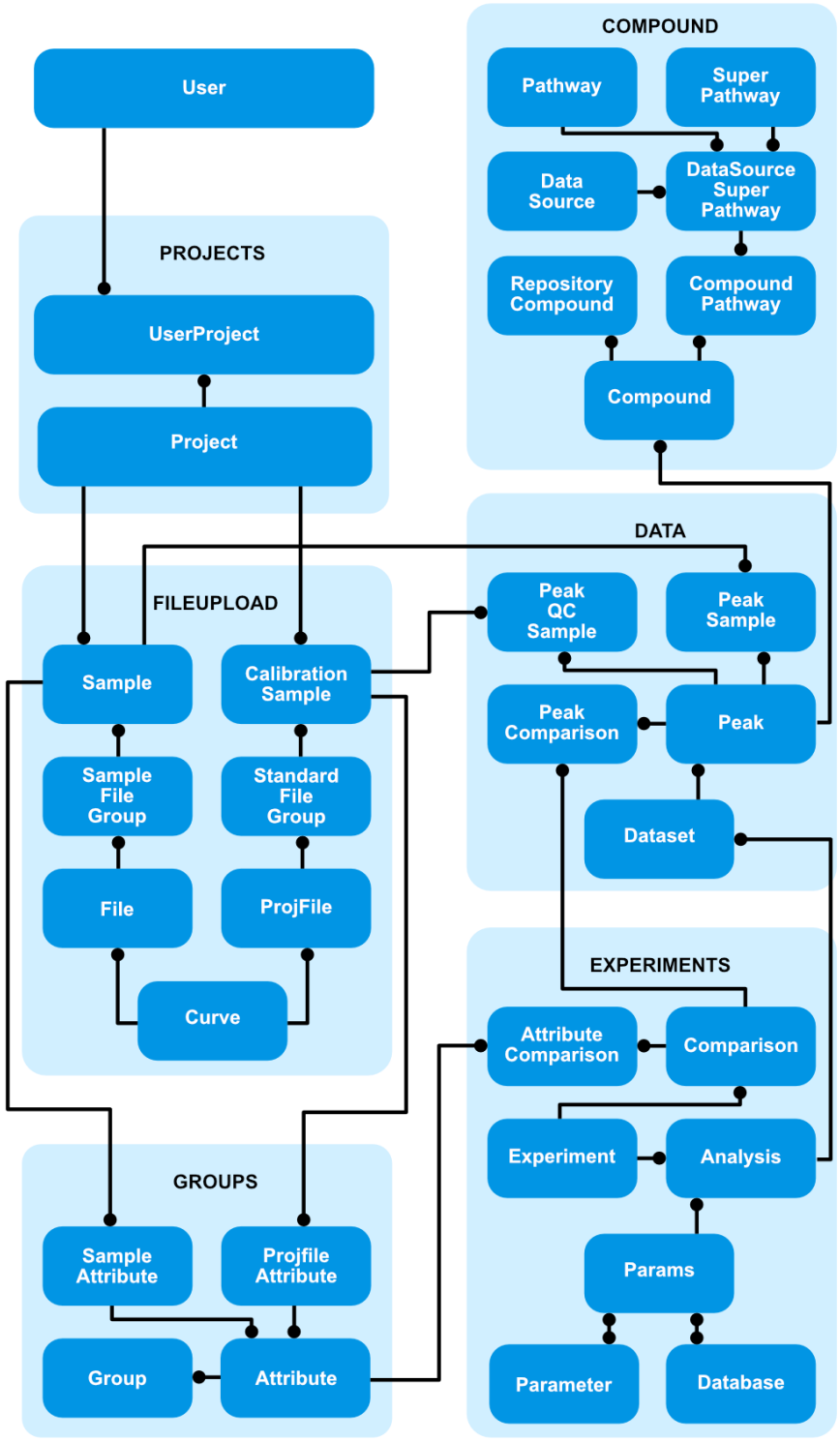

Figure 3.3: Database structure showing the general organisation of the data storage in modules. Four data modules (projects, fileupload, groups and experiment) are used to store the data captured from the user, the other two modules (data and compound) are used to store the processed and biological data generated by the the data processing pipeline.

of complex experiments that contain many factors and levels.

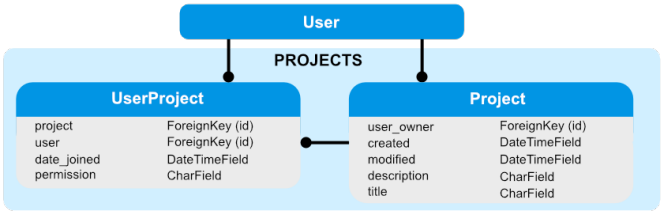

Figure 3.4: Detailed structure of the Projects module showing the organisation of meta data and user permission capabilities.

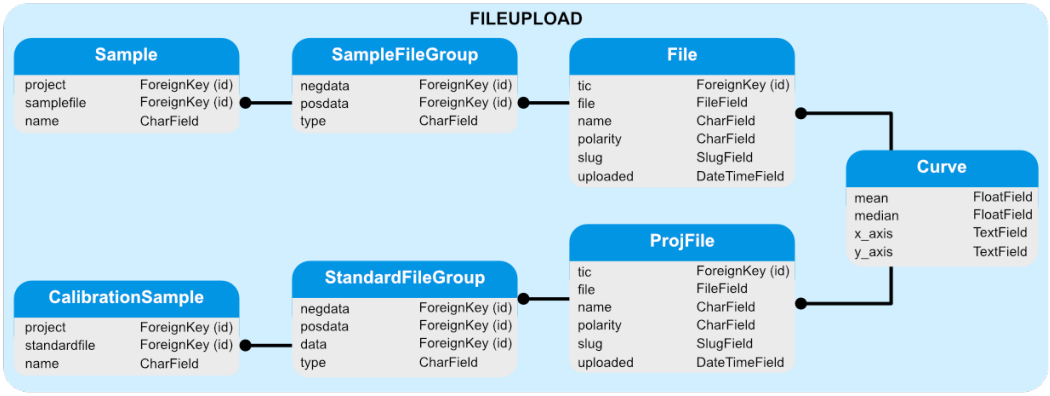

Figure 3.5: Detailed structure of the Fileupload module showing the organisation of the different files required for LCMS data analysis

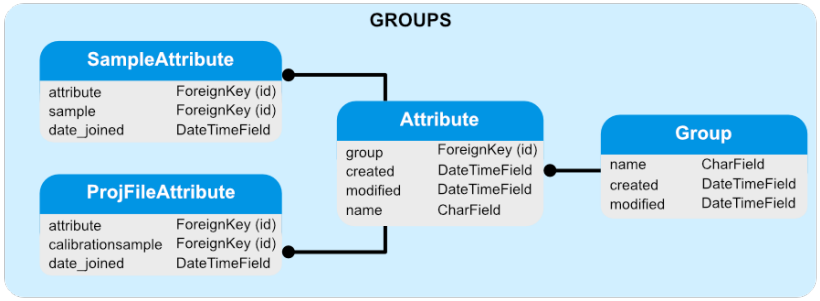

Figure 3.6: Detailed structure of the Groups module showing the organisation of the biological samples between factors (groups) and levels (attributes)

The next module represented in Figure 3.7 and named “experiments” captures two types of information, the analysis parameters and the different levels to compare. The “params” table store all the parameters and information required for the back-end pipeline to run, the ”experiment” table store the information about the comparisons to perform. The analysis table brings together those two sets of information along with extra fields such as the status of the analysis (i.e. “Running” or “Finished”) and time stamps.

The “data” module in Figure 3.8 corresponds to the extracted and computed data resulting from running the sample files through the back-end pipeline with the selected set of parameters and comparisons. The main information stored in this module is the peaks (in the “peak” table). The other tables are all joining tables that store extra information in relation to other

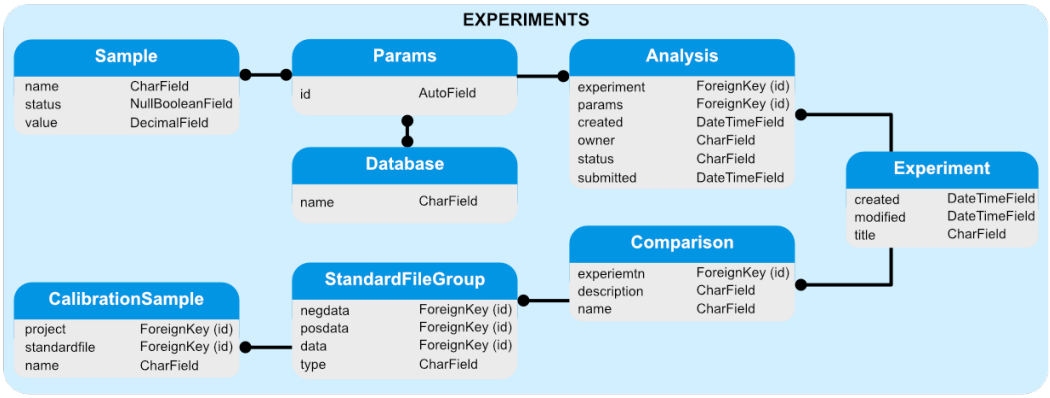

Figure 3.7: Detailed structure of the Experiments module showing the organisation of the levels to compare and the analysis parameters.

data entries. The dataset table represents the set of peaks that has been extracted from the sample files for a particular analysis. The presence or absence of a peak (and its intensity if present) in a specific sample is stored in the joining table called "peakDTsample". More information about the peak is also stored directly within the peak table such as the mass, the retention time or the polarity. The "peakQCsample" table stores the same information as the "peakDTsample" table but for the calibration samples (pooled and blank samples). The last table of this module ("PeakComparison") stores precomputed data resulting from the analysis pipeline such as the p-value and log fold change of two peaks in two different conditions. This table is a joining table between the peak and comparison table.

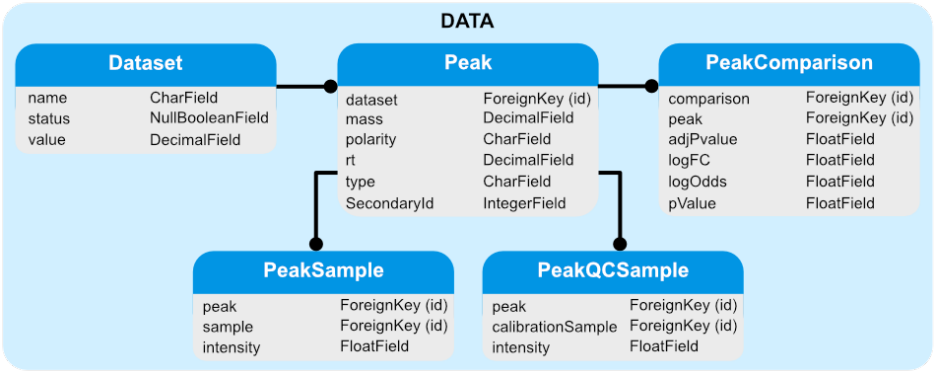

Figure 3.8: Detailed structure of the Data module showing the organisation of extracted and processed raw data into features (peaks) with attached values.

The last module is only attached to the rest of the data structure through the peak table. The "compound" module which structure is shown in Figure 3.9 stores data from external resources about biological compounds and pathways. Entries in the compound table must be unique, and the information about the external database can be found in the "RepositoryCompound" table. One compound can have many repository entries to be flexible and extendable with any external database. The four other tables in this module allow the storage and organ-

isation of pathway information. The “pathway”, “superPathway” and “DataSource” tables respectively store the name of pathways, super-pathways which are a set of pathways related with one another, and the external data source from which the information has been extracted. One big joining table brings all the information together by joining a pathway to a super pathway, a data source and many compounds.

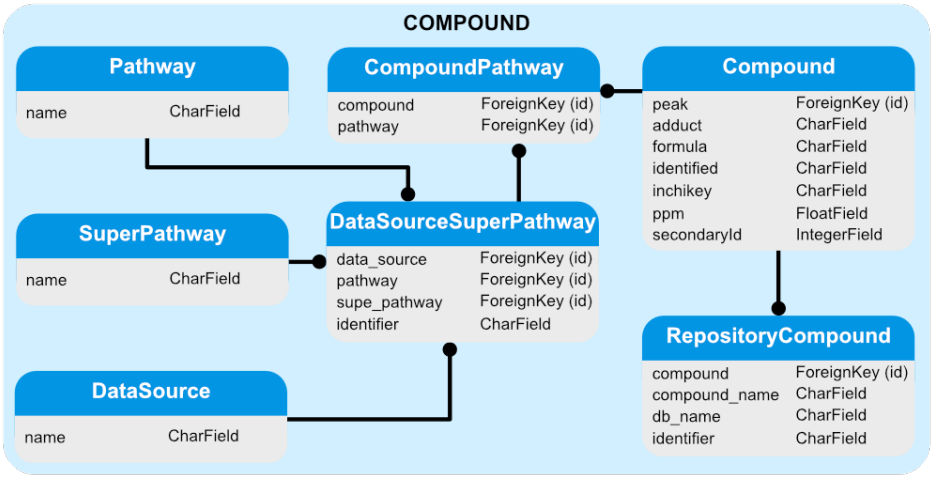

Figure 3.9: Detailed structure of the Compound module showing the organisation of the biological data used to enriched the processed data explained in the previous Data module

The modular design of the data structure has been implemented for flexibility and scalability purposes (Figure 3.3). Although the current structure supports any type of metabolomics data, the field evolves rapidly, and it is important that modules can be extended or replaced with ease to support long-term changes. For example, the file module supports the current files generated by mass spectrometers. However, it is possible that the file structure generated by these instruments changes in the future, it is, therefore, important to be able to adapt the module with minimum repercussion on the rest of the database.

3.4.2 Context-sensitive visualisation

To address the objectives 2, 3, and 4, data visualisation tools are essential as they can help users capturing the required data, performing quality control or any necessary tasks on the data before data analysis. Data visualisation tools can also provide support for data interpretation. Ultimately, the approach to data and task support presented below can help streamlining the analysis pipeline by limiting user interaction upstream of the data analysis pipeline.

Existing programs are focusing on specific activities within the workflow such as raw data visualisation (i.e. Xcalibur, Thermo Fisher), peak set visualisation and annotation [61], and pathway annotation [87, 86]. Each of these different activities involves context-sensitive forms of data visualisation. However, none of the existing solutions support all types of visualisation and more importantly, none of them connect them together within the same environment. The complexity of metabolomics data and the relationships between its different components makes the field difficult for novices to approach. End users often focus on interpreting the presented data and are not sufficiently critical its reliability itself. As other omics are more advanced, they provide a high confidence in their results; however, metabolomics is still at its early stage and the community needs to educated on the uncertainty of compound annotation [124, 125].

First, a series of quality control tasks must be performed before the data can be processed to assess the quality of the overall signal and the reproducibility between samples. Then, during data interpretation, manual validation is often required to evaluate the quality of specific features. This manual validation relies on many variables such as peak shape, mass and retention time error, and therefore need a strong understanding of the underlying data. Data visualisation can help performing these activities by providing different visualisation types according to the task to perform; we say that the visualisation is context-sensitive. The data visualisation tools described in this section aim to support the user in these validation tasks and basic interpretation tasks. These visualisation tools are developed as modules to be reusable anywhere within the software. The advanced data interpretation environment presented in section 3.4.6, in particular, make use of many of these visualisation tools. The parts of the metabolomics workflow supported by those tools are shown in Figure 3.10.

Raw data visualisation

The raw data contains the unprocessed signal measured by the instrument (mzXML, mzML or mzData). Raw data visualisation and curation tools are well supported by proprietary software such as Xcalibur (Thermo Fisher) that allows the user to drive the mass spectrometer itself. They are usually designed to support analytical chemists to run samples through the

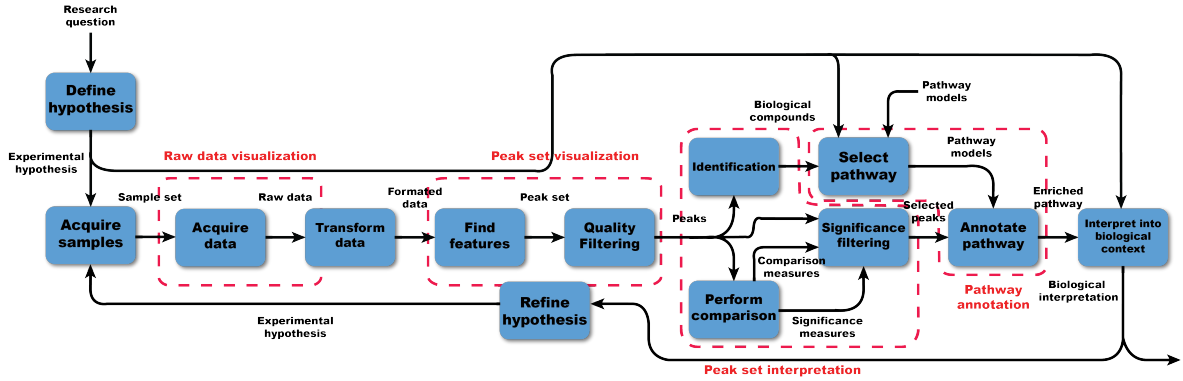

Figure 3.10: Metabolomics workflow activities that need support by context-sensitive data visualisation.

instrument and assess the quality of the data acquired. In the context of PiMP, raw data visualisation is essential at the pre-processing stage (Figure 3.11) to quality control the data and allow internal standard checks. To support these tasks, several visualisation tools are required.

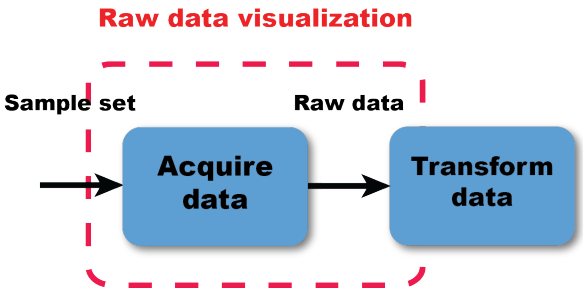

Figure 3.11: Raw data visualisation support within the metabolomics workflow.

The first module supports the visualisation of the total ion current (TIC) of individual samples to assess the quality of the signal (Figure 3.12).

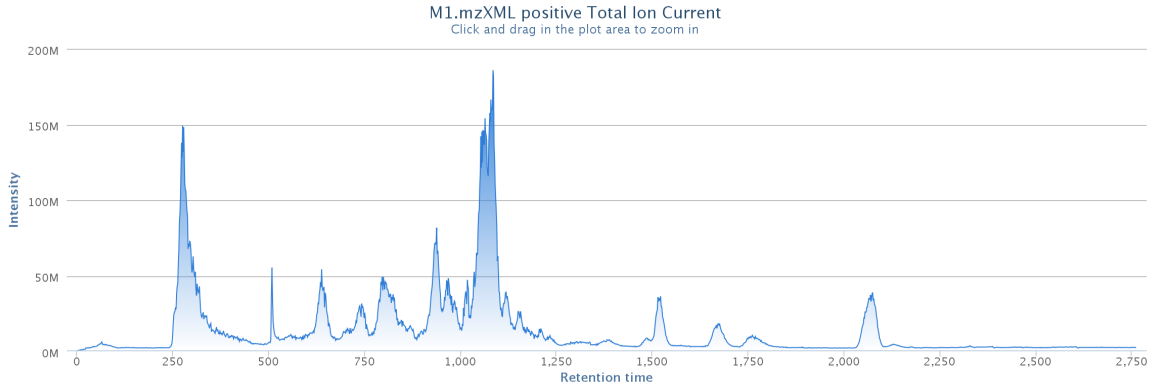

Figure 3.12: Total ion current of a single sample as viewed in PiMP.

This basic representation of the raw data is particularly useful to visualise the overall signal present in the file as well as the background noise. The evaluation of individual mass scans

(Figure 3.13) allows the identification of potential contaminants by locating specific signals throughout the run.

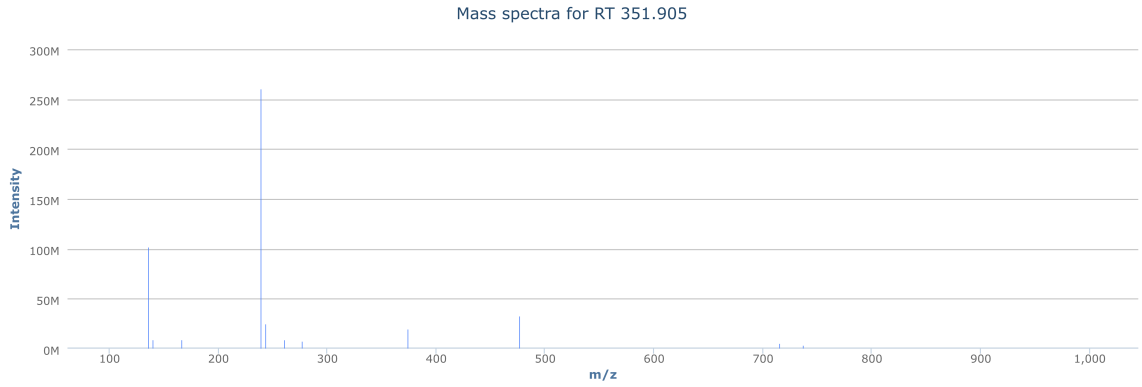

Figure 3.13: Mass spectra of the sample displayed in Figure 3.12 at retention time 2064.05 seconds .

As metabolomics studies require biological replicates to be run to be statistically relevant and interpretable, the reproducibility of these replicates is a critical part the quality control stage. Three types of charts are usually used for the user to achieve this task easily. A stacked line plot of the TIC of each replicates (Figure 3.14 a) allows the identification of potential signal exclusive to one replicate which can be the signature of a contaminant. This visualisation also allows the user to easily identify potential time drift that might have appeared throughout the run of the samples during data acquisition, in which case a realignment of the signal would be necessary during data processing (Figure 3.14 b).

Two alternative plots of the mean and median TICs are also presented to the user and allow the assessment of the reproducibility of the overall signal of the replicates (Figure 3.15).

The search for known peaks of internal standards or other compounds in the samples is also an important task that the user may want to perform before proceeding to a more global analysis of the dataset. The control of the presence of specific features in the samples can have several purposes such as assessing the presence of essential compounds for the study or the reproducibility of an internal standard. A simple tool presented in Figure 3.16 was created allowing the user to browse through the raw data by defining the following parameters: the mass, the retention time, the retention time and mass windows, the ionisation mode and the samples in which the search must be performed.

Once the search has finished processing, the signal found in every sample is presented within the same window as shown in Figure 3.17, allowing the user to assess the peak shape or identifying samples with missing signals. Users can refine their search by clicking the New screening button and changing the parameters. The tool also allows the download of each figure as a picture for presentation purposes.

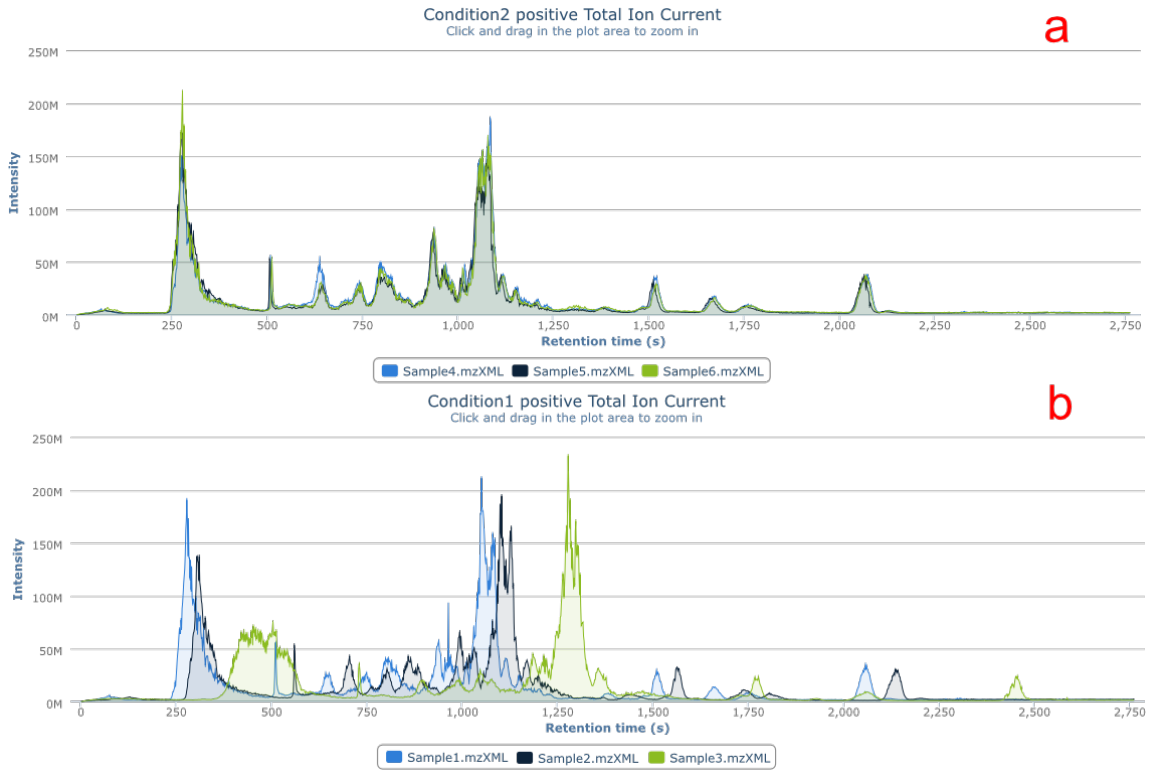

Figure 3.14: Total Ion Current of the positive ionisation of biological replicates as seen in PiMP. **a**. Replicate samples show high reproducibility, no time drift or contaminants can be identified. **b**. Replicate samples show good signal reproducibility but a time drift is clearly visible between the three replicates. Retention time correction will therefore be required during the analysis.

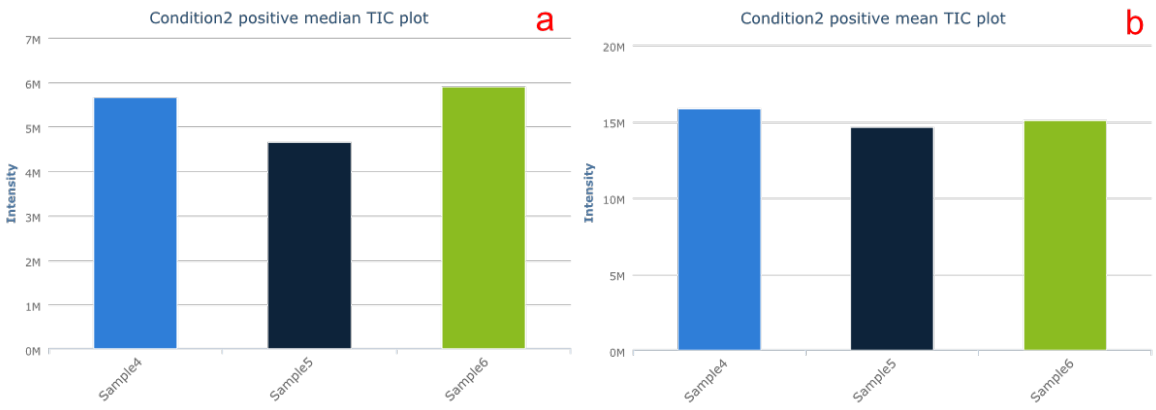

Figure 3.15: **a**. Mean Total Ion Current of the replicate samples of condition 2. **b**. Median Total Ion Current of the replicate samples of condition 2.

Peak set visualisation

During the data analysis pipeline, peaks are extracted from the raw files of each samples using a peak picking algorithm. At this step, the data describing each sample is a list of peaks, one peak being identified by two values which are the retention time and the mass of

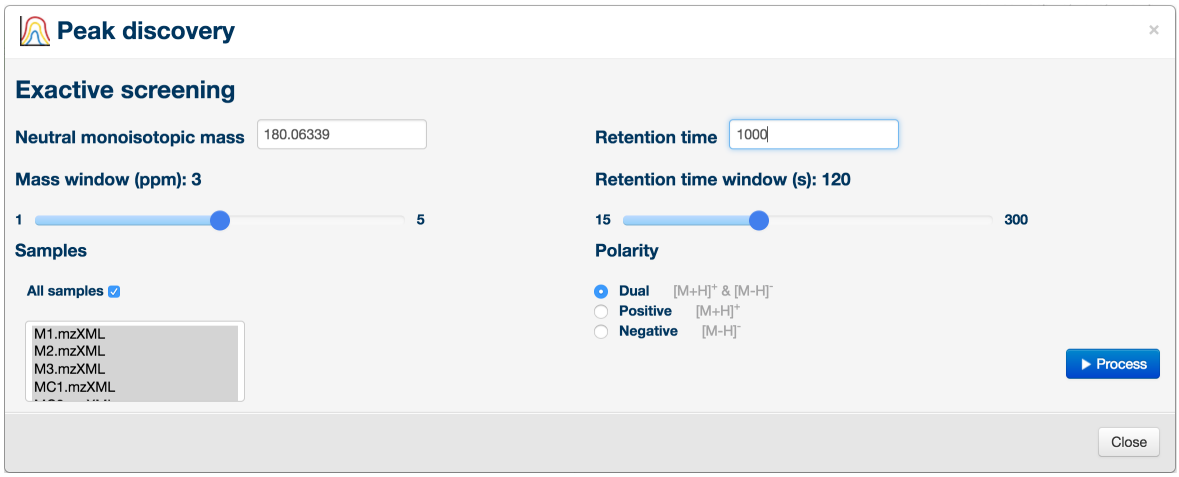

Figure 3.16: Peak discovery tool interface allowing the user to search for specific features. In this example, the user is performing a search for the signal corresponding to the glucose compound or its isomers.

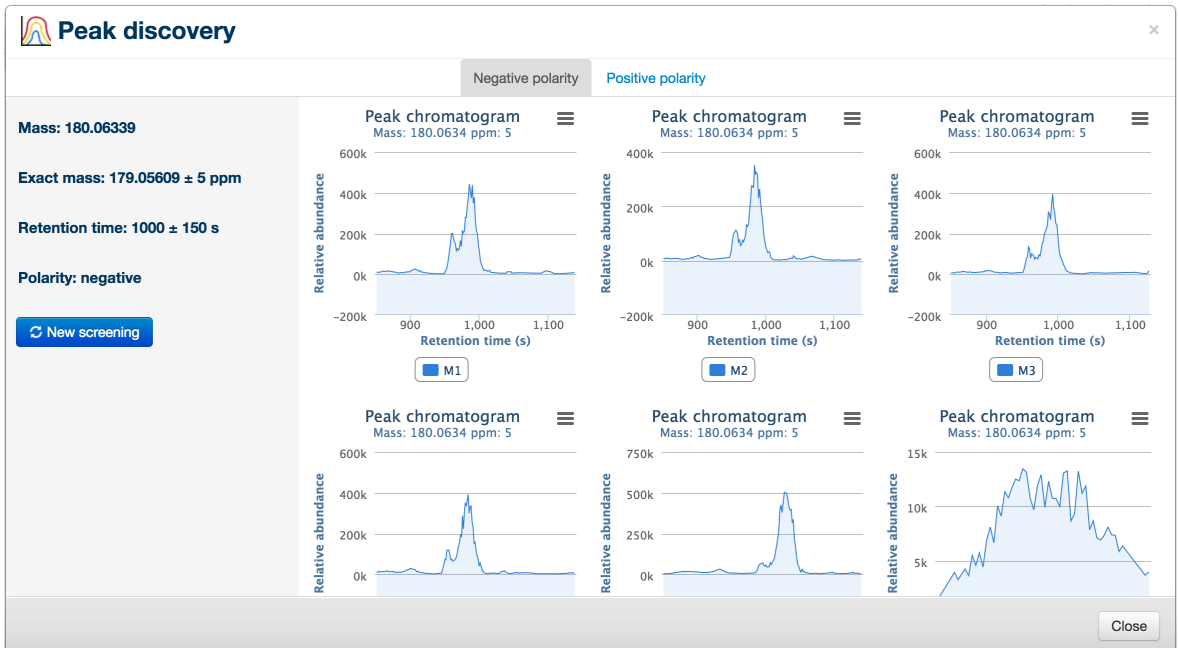

Figure 3.17: Peak discovery tool result interface displaying the extracted ion current for each individual sample.

the peak. Peaks need to be aligned across samples and replicates in order to be comparable for differential analysis at a later stage. Peaks are matched between replicates and samples using a retention time and mass window to account for time drift and mass shift across the samples, a set of peaks aligned together across samples is called peak set. The quality of the alignment is highly dependent on the parameters selected during the alignment step. For example, if a wide mass and retention time window is used, a higher number of peak set will be retrieved compare to applying narrow windows; however, the number of misassignment will also consistently increase. The visualisation of the peak sets resulting from the matching

algorithm can help assess the relevance and quality of the alignment. The task that the user may want to perform at this stage is to visualise the peak sets resulting from the grouping algorithm; this allows to assess the quality of peaks that have been grouped together, and identify potential time drift or peaks missing from specific samples. As the intensity of the peak can significantly vary between samples, it is also important that the visualisation tool supports a zooming feature as a low-intensity signal from certain samples may appear as a flat line when compare to high-intensity signal in other samples. The Figure 3.18 shows the interactive visualisation tool that was developed within PiMP to allow the user to perform this task.

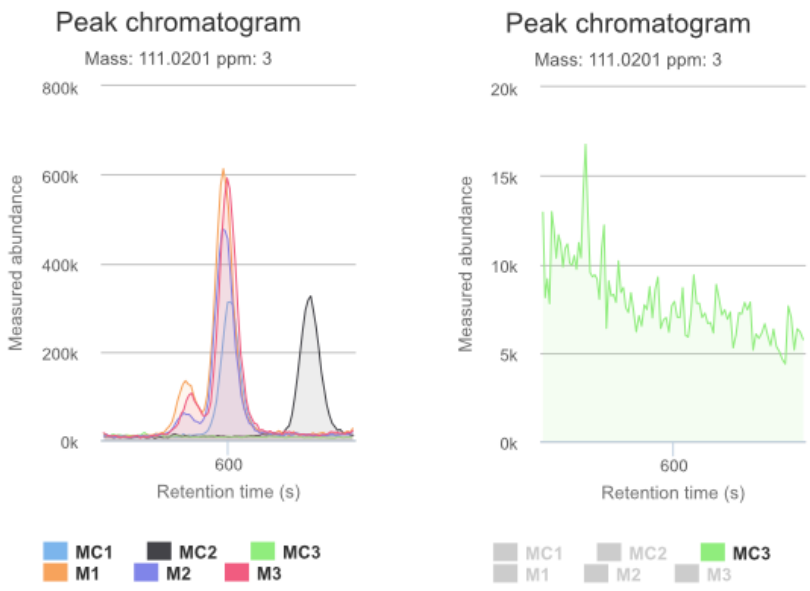

Figure 3.18: Peak set visualisation showing on the left a clear time drift in the sample MC2. The figure on the right shows the same tool with all samples hidden except the MC3, this last sample does not contain any peak and only noise signal is detected.

Peak set interpretation

The interpretation of peak sets consists of identifying what compounds could correspond to the observed signal of a feature using all the data available as a support. It is usually done by using the mass to derive the potential chemical formula, and the retention time to rank the potential compounds corresponding to this formula if many isomers exist. The second piece of information that needs to be reported is the difference of intensities between the different conditions and their significance. This is usually done by using p-values or adjusted p-values and log fold changes. The standard way of displaying this information and support the user in peak sets interpretation is a table as described in Table 3.1. This type of visualisation is however not optimal to support biologists as their primary interest lies in the compounds

	LogFC C1/C2	LogFC C1/ C3	p-value C1/C2	p-value C1/C3
Peak 1	1.34	2.31	0.0012	0.043
Peak 2	-0.52	-1.05	0.05001	0.021
Peak 3	0.19	0.15	0.237	0.1267

Table 3.1: Example of a typical table presenting changes three biological groups (C1, C2, C3). Here, C1 was compared to C2 and C3 and the log fold changes for the first three peaks are presented in the second and third columns. Only the p-value for the two comparison is shown in this table in column 4 and 5.

rather than the peaks themselves. A more appropriate interpretation module that supports the action of interpreting the compounds rather than the peak is detailed in section 3.4.6.

Pathway annotation

Pathway annotation is used to place putatively identified metabolites into biological context and enable biologists to extract meaningful biological insight. As discussed in Chapter 1.3.4, this type of analysis often necessitates the export of the data and the use of external tools. To avoid the problems that arise from exporting the data to external tools such as data formatting or controlled vocabulary, a pathway annotation tool was developed and embedded into PiMP. The second benefit of having this tool incorporated into the pipeline is the data loss-free nature of it. Indeed, the external software requires specific formats which do not support data such as peak shapes or other LCMS specific data. This makes the interpretation of LCMS data into biological context disconnected from the LCMS data itself as it only uses a list of compound names that were derived from the raw data. Having all the contextual data available and accessible in one place during the interpretation such as all the evidence supporting the annotation of a compound by a particular peak can give more power and confidence to the user. The pathway annotation in PiMP uses KEGG pathways, and the visualisation tool that supports it has been integrated in two different manners in the data exploration environment. The first one, dedicated to the pathway interpretation is described in section 3.4.6, the second tool implementing search and filter functions is part of the metabolite interpretation environment developed as a new approach for biologists to interpret their data. The second tool is also described in section 3.4.6.

3.4.3 Module based pipeline

Continuous development being a key objective of the project (aim number 3), the modularity of the tool proposed is essential. Modularity of the data structure and its implementation was discussed in section 3.4.3, however, this represent only the first step towards responsive feature integration capabilities. The rest of the software need therefore to follow a modular

design in order to address those problems. The back-end of the pipeline is written in Python (version 2.7) using Django web framework (version 1.7), it therefore follows a Model View Template (MVT) architecture [126]. This allow to separate the data structure (Model) from the logic applied to it (View) and its visualisation (Template) as shown in Figure 3.19.

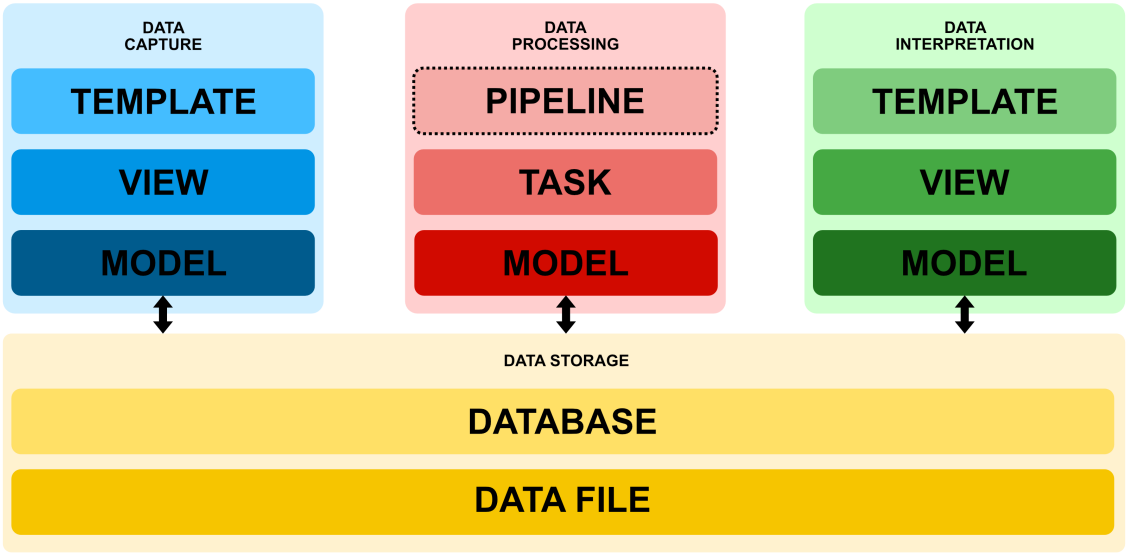

Figure 3.19: Modular organisation of the PiMP software separated in three major units corresponding to the main activities performed by the software, those activities are the data capture, data processing and data interpretation. The data storage is handled by a fourth module which includes the database and data files uploaded by the user.

The model layer fills the role of object-relational mapping (ORM) [127] which creates a virtual version of the database allowing the rest of the program to access the data. This type of architecture restrict the access to the database to the model layer only, which means no other part of the program can read, write or modify the data persistently in the database. This separation of tasks per modules or layers limits potential conflicts that can happen if different parts of the software try to perform a transaction on the same data at the same time. It is especially important in a modular software such as PiMP in which new modules can be added as conflicts can emerge rapidly if data transaction is not supervised. Limiting the access to a defined part of the software facilitate greatly the integration of new modules.

The view layer therefore only manipulates the virtual objects created by the model layer. This layer has multiple roles: (i) send commands to the model to update, modify, delete or create data, (ii) organise the data retrieved from the model and send it to the template for presentation, (iii) receive commands from the user through the template layer and perform the associated task such as performing actions on some data through the model layer, perform calculation and update the presentation of the data to the user through the template.

The template layer is the interface between the user and the program. Its role is to present the data and allow the user to interact with it. The template communicate to the rest of the

software through the view layer only, if an action is requested by the user on certain data, the template will send the request to the view, which in turn will send the command to the model layer to apply the changes. The view will then communicate to the template about the outcome of the request which will generate a new output that will reflect the changes to the user.

The high complexity of metabolomics data generated by LCMS analysis and its processing requires a robust data structure as explained in Chapter 3.4.1. Different approaches need to be supported for the user to visualise and interpret the data in an optimal manner. For example a data model might need several visualisation modules in order to present the data to the user in several different ways. The user may also require to visualise several data models at the same time within the same visualisation module, hence the need to separate the data itself from the logic applied to it and its visualisation. This flexible design also give the possibility to easily add or remove modules without affecting the rest of the software and allows high responsiveness to user needs.

As illustrated in Figure 3.19, PiMP is divided in three main units. Those units correspond to the three main tasks required to support the user in the analysis and interpretation of metabolomics data.

Data capture

The data capture unit supports the user in providing all the information necessary to perform the data analysis. The unit itself is separated in four ”apps”, each supporting a specific task.

As shown in Figure 3.20, each app follow the MVT design. The first app supports the user in the administration of his project such as giving access to collaborators and defining meta data about the design of its experiment, specifying the organism, tissue or disease studied. The second app assists the users in uploading the raw data files to the system supporting them with raw data visualisation as discussed in Chapter 3.4.2. The third app allows the user to organise the files into groups representing the different conditions of the experiment. The last app assists the user in defining the analysis to perform, specifying the conditions to compare and optionally defining custom parameters for the data processing.

Data processing

The data processing unit follows a different architecture from the two other units, this is due to the fact that the data processing needs to happen asynchronously without interaction with the user. The design of this unit is detailed in section 3.4.4.

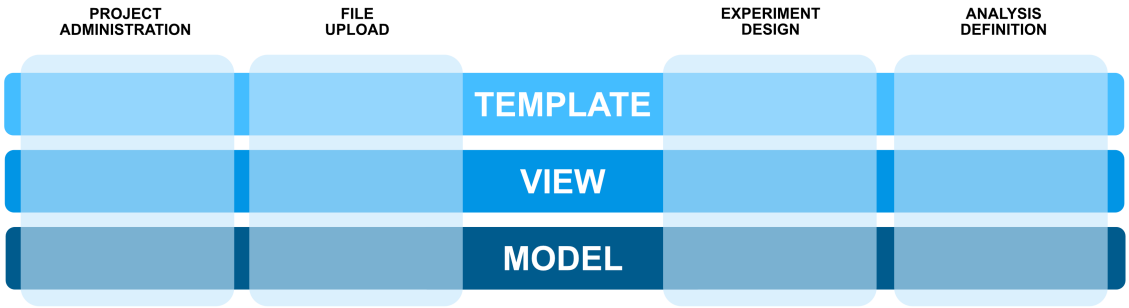

Figure 3.20: The data capture unit is fragmented in four apps, each one of them is organised using the MVT layers architecture. Each of the app also correspond to a data module in the database.

Data interpretation

The last unit also follows the MVT design and support the user in his data interpretation. This unit is made of one complex app that bring together both experiment design, raw data information and analysis results. The views present in this app relate therefore to most data models present in the program and organise them in order to best support the users in their interpretation. The complex data visualisation which form the template layer of this app is described in section 3.4.6.

3.4.4 Data analysis

Figure 3.19 shows that the data processing unit does not follow the MVT standard design. The first main difference is the view which is replaced by a ”Task” layer. This is in fact a distributed task queueing system that allow to distribute work across several threads. There are several benefits of using this system when a program needs to run intensive data processing such as PiMP. The first one is to limit the processing power given to one process in order to make sure that the rest of the program can run as intended. This queueing system also allows the program to run a predefined number of analyses at the same time, keeping the other ones in a queue waiting to be processed. Another advantage of such system for PiMP is its asynchronous nature, this means that this type of design is non-blocking for the user. Indeed, metabolomics data processing is heavy and can take several hours to complete, hence the need to make sure that the interface is still reactive to allow the user to perform other tasks during data processing. As PiMP is written in Python, running the pipeline within the Django framework would be limiting the number of tools that can be used for the processing of the data, the queueing system also give the possibility to export the data analysis pipeline outside Django, and therefore use pipelines built in other languages. An extra interesting feature of this system is that the data processing can be exported to another machine such as a computational cluster, this allows to separate the computationally intensive tasks from the

system in direct communication with the client side (the user). However, due to hardware limitations, PiMP currently does not take advantage of this feature.

Figure 3.21 shows in more details the architecture of the data processing unit. The data analysis pipeline runs from end to end and therefore does not require any user interaction, hence the absence of a template layer. As explained for the ”view” layer, the ”task” layer retrieve the necessary data and parameters through the model layer to launch the pipeline. The task layer supervise the pipeline and will update the model layer in case of interruption to inform the user. In order to limit conflicts, the pipeline does not access to the data directly as explained in section 3.4.3. There is therefore a need for the pipeline to communicate the results back to the task in order to be persistently saved in the database. This is made possible through a data exchange format specifically designed for PiMP. This format is discussed in section 3.4.5. As a result, the main requirement for the analysis pipeline is to be able to export the results to the PiMP specific file format. This approach follows the same modular design of the entire software to allow easy replacement or addition of new analysis pipeline to the software. Although only one analysis pipeline is currently available in PiMP, developers can easily integrate new pipelines as long as the output is formatted correctly.

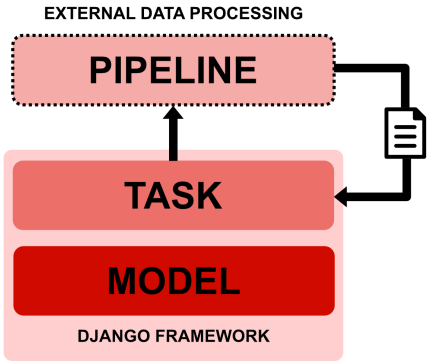

Figure 3.21: Data processing unit showing the external and independent nature of the pipeline to increase the modularity and flexibility of the software. The communication between the pipeline and the task is made using a standardised and PiMP specific data format.

The current data analysis pipeline present in PiMP is not detailed in this document as the author only developed the communication system between the framework and the pipeline. However, the overall architecture and tools used within the data analysis pipeline is given below for comprehension purposes. The backend pipeline in R was developed By Fraser Morton and Ronan Daly, the author developed the communication system to integrate it into PiMP.

The data analysis pipeline is written in R and make use of two main metabolomics packages. The feature detection is perform using XCMS. Peak alignment, filters and peak identification are then applied using mzmatch.R. Pathway annotation was developed specifically for PiMP using R and KEGG pathways. An R module to export the results from R to the PiMP

specific format was also developed in order to connect the analysis pipeline to the Django framework.

3.4.5 Data exchange and data sharing

Data exchange

Data exchange is the transformation process of data structured under a source schema into data structured under another schema called target schema. The target data must be an accurate representation of the source data and therefore require an exchange format that captures every piece of information that can be found in the source data structure. In the field of metabolomics, the Metabolomics Standards Initiative (MSI) support and coordinate the development of data formats for metabolomics through the coordination of standards in metabolomics (COSMOS) global effort [128]. COSMOS work aims to create standard formats for data exchange and is primarily focusing on raw data in MS, metabolite quantification and identification, and experimental metadata.

The PiMP data structure contains a large amount of information such as statistical values and biological pathways data that are not currently supported by any standard format. As PiMP is a modular software, the backend pipeline running the analysis can be replaced, or alternative analysis pipelines can be created to give a larger range of options and more flexibility to the user. Hence, to ease the integration of new analysis pipeline modules in the PiMP platform, a robust and PiMP specific data exchange format is required to transform the data structure of the analytical tool to the PiMP data structure. A new exchange format called pimpxml was therefore created allowing the transfer of all the information about the experimental design, data analysis parameters as well as the results of analysis such as the peaks information, statistical values, metabolite identification and biological context data. This data format is then used to populate the PiMP database and store the data permanently. In order to help the development of new pipelines, a python parser for the pimpxml format was created, and a database population function can be called within the Django framework to transform the data from the pimpxml file to the database. The pimpxml schema is an XML representation of the data structure discussed in the data structure section of this chapter. The schema is constantly evolving to support extra information created by external modules such as fragmentation data. However, as external modules are optional, the schema has a flexible design, and the definition of extra information such as fragmentation data is not mandatory. Therefore, although the schema is enriched to hold data from new modules, the minimal information required for a pimpxml file to be validated against the schema remain static and correspond to the standard pipeline output currently in place.

Data sharing

Collaborative research has become more and more common, and the size of the data in biotechnologies is a well-defined problem and an apparent obstacle to collaborative work. Raw data in metabolomics often exceed one gigabyte, and although the analysed data may be smaller, sharing the data results implies that all parties involved in the project have access to the same software to appropriately explore the data. As PiMP is a web-based application with a login system, it is by nature the type of application that can be used to share data between collaborators working on the same project. A sharing system has therefore been developed allowing users to share one or more projects with a some level of permission. Three permissions rights can be set when sharing a project with another user. The ”read” access permission is the most restrictive one; it only allows the invited user to view the project information, experiment design and access the results. The ”write” access is more permissive and allows the guest user to upload new samples and perform new analyses with chosen parameters. The ”admin” permission gives the guest user the same right as the project owner except the possibility to delete the project.

The sharing capability of PiMP is enabled by the user interface and its modular design, it is supported by the structure of the database and powered by a user session system. This capability enable multiple researchers without limitation to work at the same time on the same project and share any piece of data, allowing users to access the analysis design, analysis parameters, results or raw data of a single project. As the system does not require any installation, PiMP is not affected by versioning issues encountered by desktop applications. Furthermore, once the raw data is uploaded to the system, it becomes accessible to all collaborators of a project which overcome the problem of large data file transfer.

The architecture of PiMP also allows the possibility of making projects and analysis results publicly available, which would make them accessible to everyone without the necessity to create an account. Although this feature is currently disabled, as each analysis result page is given a unique url, it will allow easy data publication once activated.

3.4.6 Data interpretation

The interpretation of metabolomics data is a complex task that can require knowledge in different domains such as mass spectrometry, analytical chemistry, biochemistry or metabolism. It also often requires expertise in the specific question that the experiment is trying to address such as a broad knowledge of the organism, tissue or disease studied. However, guiding the user in the interpretation of the data resulting from a metabolomics experiment can help overcome the lack of knowledge in those fields. The work described is this section present the data exploration environment that was developed within the PiMP software to assist the user

in the interpretation of its results; it, therefore, addresses the aim number 4 of this project. The data exploration environment makes use of the modular model of PiMP and more specifically of the context-sensitive visualisation modules discussed in chapter 3.4.2 to present the data to the user in a coherent and intuitive manner.

The data exploration environment is presented as one unified page to the user and structured into tabs, each tab being laid out as a table with the exception of the summary page which provides an overview of the experiment in the format of a scientific paper, containing key findings of the experiment and associated metadata.

Summary page

The summary page is designed to attract the user attention to potential findings in his dataset as well as another quality control of the analysis performed. Three sections are presented to the user, each of which having a different purpose.

The first section contains metadata about the experiment provided by the user, a summary of the comparisons performed and a table containing the experiment design. This section also provides information about the method use for processing the data.

The second section is for quality control purposes. A principal component analysis (PCA) plot showed in Figure 3.22 is provided and allow the assessment of reproducibility between biological replicates and separation or clustering between biological conditions. TIC plots of the samples grouped by biological conditions as shown in Figure 3.14 are also provided to assess the reproducibility of the replicates in both positive and negative ionisation mode.

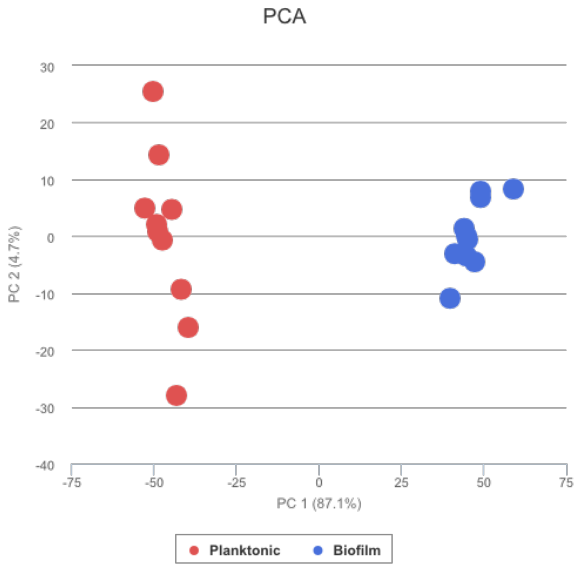

Figure 3.22: Principal Component analysis plot showing the clear separation of the two biological conditions being compared.

The third and last section of the summary page provides the user with the most significant quantitatively changing metabolites for each comparison. Those differences are highlighted using histograms and interactive volcano plots as shown in Figure 3.23. Zooming into the volcano plot and accessing to the annotated metabolite corresponding to a particular feature by clicking on dots on the figure allows quick exploration of the most changing compounds in the dataset.

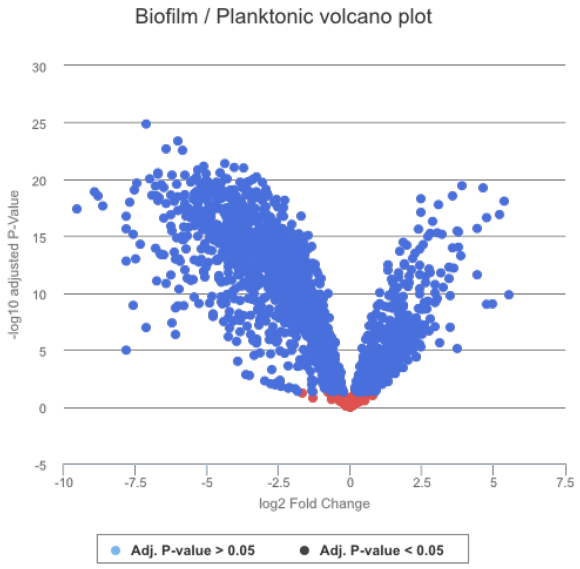

Figure 3.23: Interactive volcano plot allowing a rapid assessment of the number of peaks that significantly changing between the two conditions. Clicking on a dot allows access to the list of compound annotated by the feature. The visualisation tool also has zooming features.

Raw data

The list of extracted peaks from the raw data, often referred to as the raw data at this stage of analysis is available in the ”peaks” tab. The table contains the mass and retention time of the extracted features, the polarity in which it was detected, and the relative abundance value of all the replicates. This page contains no extra statistical or biological data. The purpose of this page is to allow the users to export it in a comma separated format to perform their own statistical analysis using external tools. The evidence side panel provides access to the peak chromatograms and quantitative information as interactive bar plots (Figure 3.24).

Statistical analysis

The results of the statistical analysis performed during the analysis pipeline are presented in two different manners to support two different tasks. The first visualisation module displays a separate table for each performed comparison containing all statistical values available attach

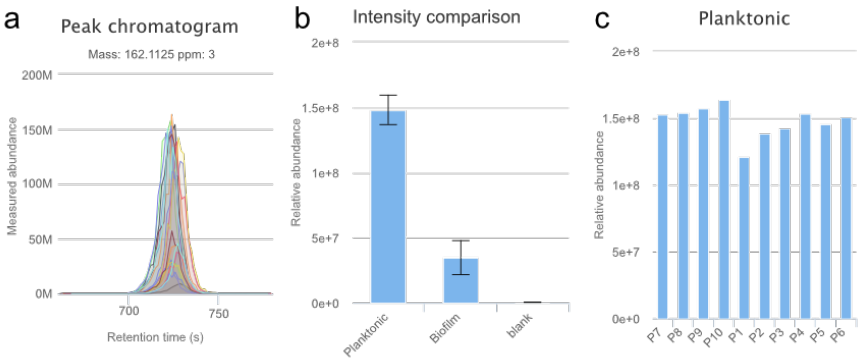

Figure 3.24: Main figures accessible in the evidence panel of the peak tab, showing a typical dataset derived from a metabolomics experiment comparing biofilm and planktonic *staphylococcus aureus* [129]. **a.** Extracted Ion Chromatogram of the peak using the peak set visualisation module. **b.** Bar plot showing the average intensity and standard deviation of the two conditions and the blank samples. **c.** Bar plot showing the intensity of the investigated feature in each sample of the planktonic condition.

to the peaks. The table contains the peak id, the log fold change, the p-value and adjusted p-value, and the log odds. While this table is not straightforward to interpret, it gives access to the user to all statistical values available that are essential for reporting purposes.

The second module is a unique table summarising the changes for all comparisons performed. For each peak entry represented as a row in the table, the log fold change values for every comparison is given in individual columns (Figure 3.25). A heat map type visualisation is overlayed on top of the comparison cells when the adjusted p-value is under the 0.05 significance limit. This module allows the user to identify peaks that are significantly changing between biological groups quickly. As no biological information is attached to the table, it allows potential discovery of unknown compounds that would not be reported when interpreted within biological context. However, if potential compounds have been matched to a particular peak, the information is reported in the right panel for the user's information.

Peak id	Stage_1 / Control (logfc)	Stage_2 / Control (logfc)	Stage_1 / Stage_2 (logfc)
520	-2.84	-2.33	-0.51
1111	-2.61	-2.87	0.26
1132	-1.32	-1.34	0.02
232	-1.29	-0.66	-0.63

Figure 3.25: First 4 entries of the summarised comparison table. The first column shows the peak id, the following three column show the log fold change values of the peak in every comparison. The number of column is dependant on the number of comparison performed.

Biological pathways

The Metabolic maps tab replaces the data in the context of biological pathways. Biological pathways used in this module originate from KEGG. The table displays the list of all KEGG pathways alongside with the number of compounds that form this pathway, the number of annotated and identified compounds found in the dataset that are part of the pathway, and the of coverage of the pathway by the dataset. A visual version of the coverage is available on the side panel in the form of a pie chart. The side panel also gives access to the pathway visualisation tool which displays the KEGG pathway with contextual data extracted from the dataset.

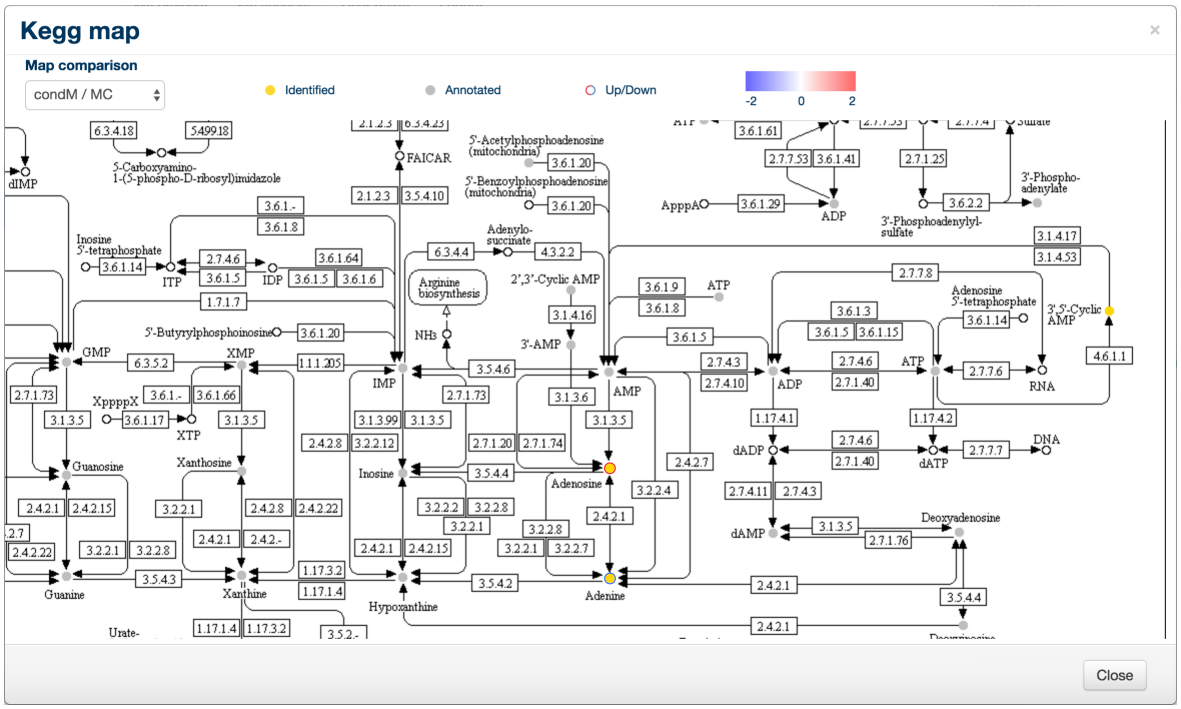

Figure 3.26: Pathway visualisation accessible from the evidence panel of the Metabolic map tab. This example shows the Kegg pathway of Purine metabolism with identified metabolites in yellow and annotated metabolites in grey. A specific comparison was selected (top left of the window) resulting on the display of log fold change indication on the representing the metabolite with a coloured border. The colour scale is indicated at the top of the window.

Figure 3.26 shows the visualisation of a KEGG pathway within the PiMP data environment, identified and annotated metabolite can be identified by gold and silver dot colour respectively. The user can select a specific comparison that was performed during the data analysis to overlay log fold changes data. The log fold changes value are represented by a red and blue border of the dot representing the compound. If several peaks annotate a compound with conflicting log fold changes (one negative and one positive log fold change), the border of the dot is then coloured in purple to inform the user.

Metabolites

The metabolite tab presented in Figure 3.27 brings together all the data such as the statistical values, biological pathways, peak information or compound structure in a unified environment to assist the user in the interpretation of his results. The interface follows the same structure as the other tabs with a main table and an interactive and contextual side panel displaying pieces of evidence for the compound being investigated.

The table contains the minimum information of interest to the user to guide its interpretation. A row in the table correspond to a compound, with the number of columns depending on the number of comparisons performed. The first two columns respectively contain the name of the compound detected and its formula. The following columns contain the log fold change values for all the comparisons, and the last column contains the level of identification of this compound. The identification level can take two values, ”identified” if the peak matched by mass and retention time to a standard compound, and ”annotated” if the peak matched by mass only to an external compound database. As several peaks can annotate one compound, there can be several possibilities available for the log fold change value. The value displayed is selected according to several criteria: (i) M+H and M-H, the first criterion filters the peaks corresponding mono-isotopic mass calculated from the molecular formula (M) of the compound with an added or subtracted proton (H) depending on the ionisation mode. (ii) When both ions (M+H and M-H) are available, the peak set with the highest intensity value is kept.

The evidence panel is structured into collapsible cards that give contextual information and evidence on the compound being investigated by the user. This information is accessible to the user by simple a click on a row of the metabolite table. Three main cards are displayed by default in the evidence panel, each of which relating to different data. The first card, called compound card, inform on the external or internal (if standards compounds were provided) database in which the compound was found. There can be several databases available with different names given to the same compound. For example, the lactic acid in HMDB is named lactate in KEGG. The structure of the compound is also available when the card is expanded, as seen in Figure 3.28.

The second card lists all the pathways in which the compound is found, in its collapsed state, the card only gives the number of pathways.

The last card informs on the peaks which annotate the compound under investigation; there are as many cards as there are peaks. This card is designed to give the user the essential information to quickly assess the quality of the peak and the relevance of the match. Each peak card contain the following information: (i) Retention time, (ii) Mass, (iii) Polarity or ionisation mode, (iiii) Type of peak which can be for example a base peak, a related peak or an adduct, (iiiii) The ion information (i.e. M+H, M+Na), (iiiiii) the mass error in ppm. The

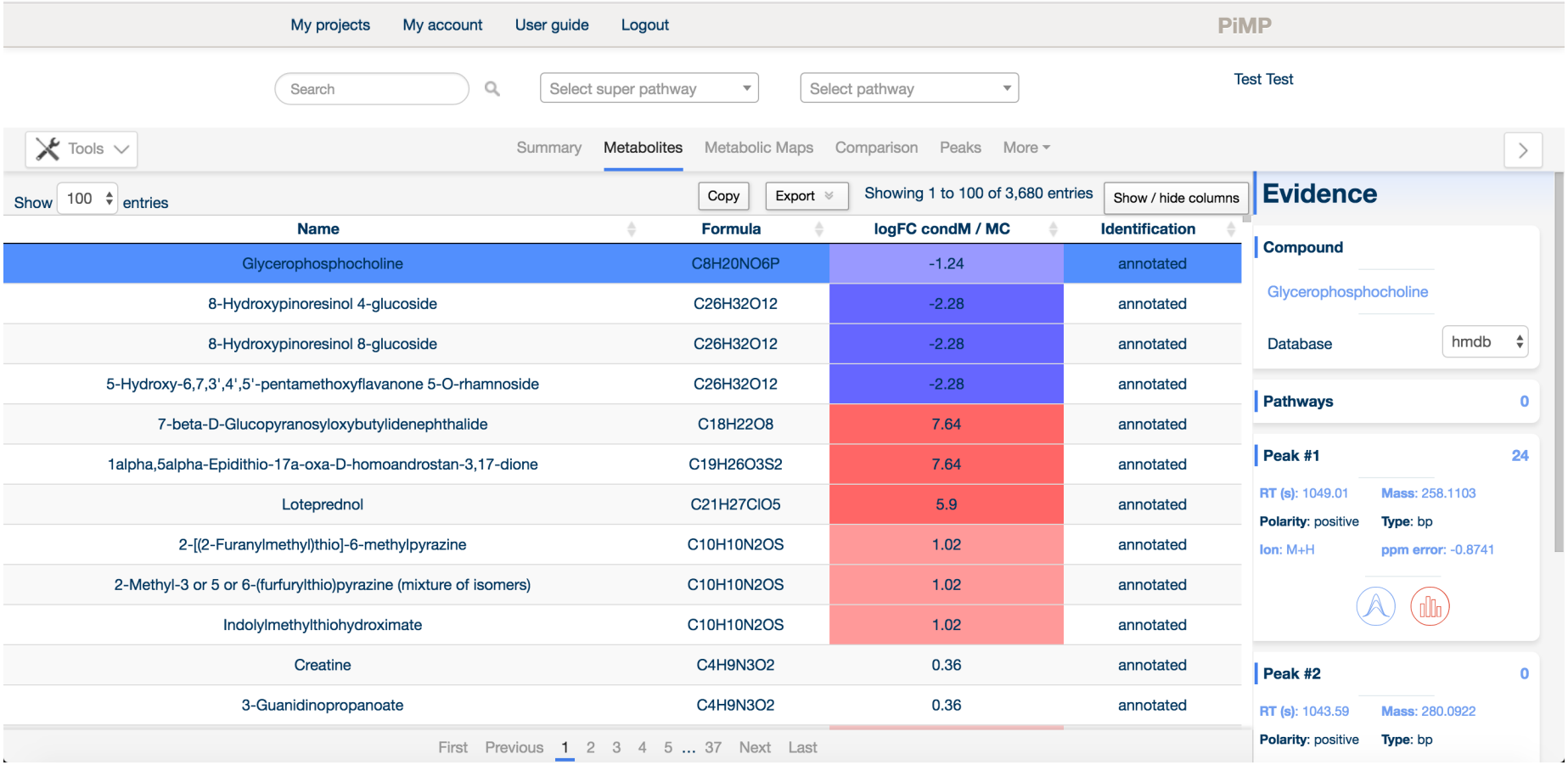

Figure 3.27: Screen shot of the complex organisation of the Metabolite tab with in the center the table, at the top filter, search and export tools, and the evidence panel structured with evidence cards on the right of the window.

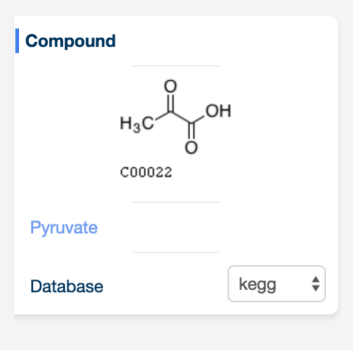

Figure 3.28: Compound card showing the structure of the compound detected, changing the database allows to display the database specific name of the compound and its structure.

card also contains the number of other compounds annotated by this peak (excluding the one being investigated) and an identification flag to show that this feature matches the peak of the standard compound. Two visualisation modules are also embedded within the card allowing the visualisation of the peak set and intensity plots as shown in Figure 3.29.

The last module to support the user in the interpretation of the results is a set of tools to sort, filter and search the data. The sorting tool is directly embedded within the table and allow to sort the results by alphabetical order or high to lowest log fold changes for example. The filtering tool is only based around biological pathways found in KEGG, KEGG organises the pathways into groups forming wider metabolic maps called super-pathways. As shown in Figure 3.30, the user can select a super-pathway which will filter down the pathway selection options to the pathways present in this super-pathway. However, a user can directly select a pathway without narrowing down the selection by choosing a super pathway. The search tool is meant to be used by more experienced users who are looking for something specific. When the user types characters into the search box, the search is launched on the fly on all data available in the table (name, formula, log fold change and identification level). The search box also applies the search on the pathway names. For example, typing ”glycolysis” in the search box would apply the same filter as selecting the glycolysis pathway from the pathway filter selection.

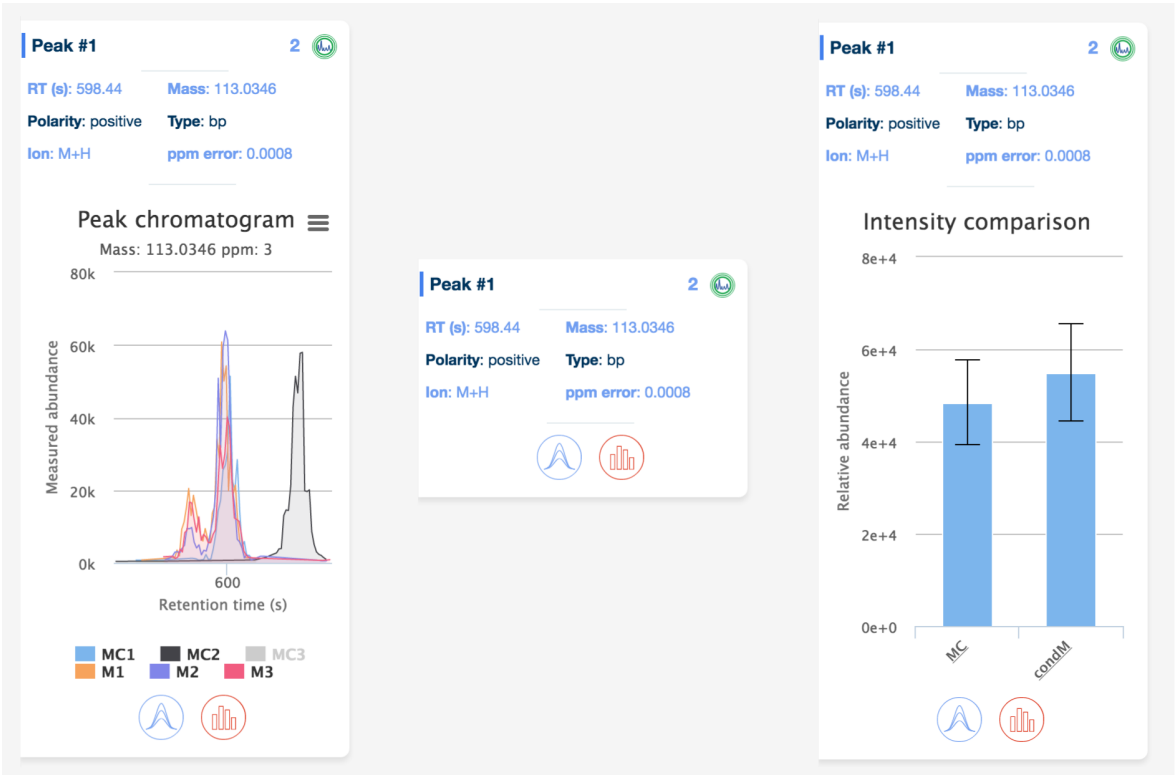

Figure 3.29: General organisation of the peak card. In the centre is default collapsed card giving the essential information about the peak. On the left, the peak button has been clicked in order to display the EIC of the peak. On the right, the bar plot button has been clicked in order to visualise the average intensity per condition.

Figure 3.30: Search tools available at the top of the metabolite tab

3.5 Discussion

The use of LCMS methods for untargeted metabolomics experiment is still in its infancy, while it is a powerful way of getting a better insight on the metabolism of a biological system, the scientific community uses it for a variety of applications. Defined approaches can be used for biomarker discovery or the study of a particular part of metabolism, but untargeted metabolomics is also often used for wider approaches such as hypothesis generation. Indeed, the study of the differences found in the metabolism of a biological system can help to narrow down the area of research to some specific pathways. This implies an iterative process starting from a general hypothesis that some metabolite will show differences when one system is exposed to a specific stress, to a refined and more focused hypothesis. However, limitations intrinsic to the field such as data analysis and interpretation time, and high variability of the system studied create tremendous challenges to make this iterative process straightforward. Overcoming these limitations could have a transformative effect on the field and its impact on the scientific community.

One of the limitations comes from a poor experimental design when metabolomics is used for hypothesis generation, and often the data produced during the first iteration cannot be used again as they do not cover the refined hypothesis or do not have enough replicates to produce significant results. Assisting biologists in designing their experiments and supporting them with, for example, the minimum number of replicates that should be used for their experiment to be valid could reduce the number of iteration. This issue can be in part supported at the data analysis stage guiding the user in structuring its experiment; however, to fully overcome this problem, supports needs to come before data acquisition, when planning and designing the experiment.

The second limitation is the complexity of the data produced by LCMS instruments. Simplifying and streamlining the tasks to perform to process the data by guiding the biologist step by step would make the field more approachable to the scientific community, and eliminate the need for bioinformaticians to generate the results themselves. This would also accelerate significantly the turn around time between data acquisition and interpretation.

Finally, supporting biologists in the interpretation of their data by providing biological context would both accelerate the iterative process and improve the interpretation of the results itself. This would enable biologists to interpret their results fully and therefore have a major impact on their research outcome.

The tools discussed in section 3.2 of this chapter did not offer solutions to these limitations at the beginning of the project. PiMP, the software developed and presented in this chapter was designed to overcome all those limitations and therefore improve the current state of the metabolomics field by making it available to a wider scientific community. The three

research questions were all successfully addressed by the developed web-enabled tool which presents a modular design to allow scalability and responsive feature development while enabling users to collaborate worldwide and interpret their results in biological context. However, the tools available at the start of this work have evolved [42, 61, 78, 44] and several new non-commercial tools have been developed [81, 82, 79, 80] during the development of PiMP. Some of these tools now provide new support to the end users and therefore address some issues introduced in this chapter.

Table 3.2 presents some of the capabilities of the most commonly used non-commercial software. All capabilities presented in this table are crucial to addressing the issues introduced in this chapter to support a wider scientific community in LCMS data analysis for untargeted metabolomics experiments. While some of those tools cover a large amount of the requirements, they still present limitations for biologists with little knowledge in bioinformatics and metabolomics. Some of the tools still require installation on a personal computer which can represent a barrier for users and substantially limits collaborative work [42, 61, 78, 44, 79]. Many software does not provide biological interpretation capabilities or do not cover the entire metabolomics pipeline with visualisation tools to support the end users in performing all required tasks [42, 61, 78, 44, 79, 80, 82, 81]. While most software are designed to support untargeted analysis, some are too technical for users with little knowledge in bioinformatics and mass spectrometry. They also do not present a streamlined pipeline which requires the intervention of the user during the analysis. This forces the user to supervise each step of the analysis and limits the turn around time for iterative approaches [42, 61, 78, 44, 79, 80, 82, 81]. Finally, tools such as MAVEN and MZmine allow external developers to access the code base to integrate their own features to the pipeline; this is a real asset for rapid development when the tool present a modular design as discussed in section 3.4.3 of this chapter.

By providing full support and features presented in Table 3.2, PiMP addresses all research questions introduced in section 3.2. By streamlining the pipeline, providing full visualisation and biological interpretation support, PiMP offer support to non-expert users in the analysis and interpretation of metabolomics datasets (addressing the first research question). Its modularity of data structure and layer separation addresses the second research question by enabling developers to responsively adapt the feature set of the software to the ever expanding requirements in the metabolomics field. The web-based nature of the tool and sharing functionalities also enable worldwide collaboration and therefore overcome the issues related to big data in the field (addressing the third research question).

The first objective which consisted in creating a metabolomics data analysis tool accessible to non experience users has been met by the development of a web enabled software supported by a metabolomics specific data structure and user interface as presented in section 3.4.1 and 3.4.2.

Support & feature / Tool name	PiMP	MAVEN	XCMS	XCMS Online	MetaboAnalyst	MZmine	workflow4metabolomics	IDEOM
Operating syestem independent[1]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Installation free[1]	Yes	No	No	Yes	Yes	No	Yes	No
Full visualisation support[1,4]	Yes	No	No	No	Yes	No	No	No
Untargeted pipeline[1]	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Streamlined pipeline[2]	Yes	No	No	No	No	No	No	No
Open development[3]	Yes	Yes	Yes	No	No	Yes	Yes	No
Biological interpretation[4]	Yes	No	No	Yes	Yes	No	No	Yes
Modular design[3]	Yes	No	No	No	Yes	Yes	Yes	No
Sharing capabilities[5]	Yes	No	No	Yes	No	No	No	No

Table 3.2: Support comparison of non-commercial metabolomics data processing pipelines. Installation free = No installation required to use the software. Biological interpretation = User is provided with some biological context such as pathway (mapping/enrichment) and/or biological network information. Full visualisation support = Visualisation tools support the user for the entire metabolomics pipeline presented in Figure 3.1. Untargeted pipeline = Allows complete processing of raw LCMS data from peak detection to metabolite annotation. Streamlined pipeline = Pipeline designed to limit and centralise user interactions for the data capture. Open development = Source code available to developers (publicly or on request).[12345]: Objectives addressed.

The externalisation and encapsulation of the data analysis pipeline and the development of a new data exchange format respectively detailed in section 3.4.3 and 3.4.5 fulfilled objective 2 while objective 3 was met by the modularity of the architecture and the data structure of the developed tool as described in section 3.4.3 and 3.4.1.

The development of a metabolomics data specific exploration environment which formed objective 4 was addressed and is described in section 3.4.6.

Objective 5 was met by developing user sessions supported by the overall design of the software developed, from modular architecture and data structure to data exchange and visualisation.

While it was demonstrated with PiMP that metabolomics analysis can be made more approachable to users novice to the field, the current version does not yet offer alternative back end pipeline which can limit the possibilities in the parametrisation of the analysis. Although the modularity of the tool provides an opportunity to integrate new pipeline as discussed in section 3.4.3, the alternative software currently offers more choices to expert users in term of pipeline options to analyse their metabolomics datasets.

3.6 Conclusion

As the most recent of the omics technology family, Metabolomics attempts to provide unbiased ways of analysing the small molecules of a biological system. Although the technology shows great potential, it is yet still immature and faces enormous challenges in providing streamlined analysis methods and interpretation of the data in a biological context. Within the past decade, Metabolomics has grown into a powerful tool for biological and biomedical communities for the study of the metabolism and is now used by many applications such as biomarkers discovery or synthetic biology. From medicine to the biosynthesis of fuels or food security, improving a critical tool such as metabolomics could have a tremendous impact. The work presented in this chapter tackle some of the challenges currently faced by the field that limits its expansion. The novel approach to metabolomics data analysis and interpretation and the tools developed as part of this work provide the scientific community with a new environment to overcome challenges such as data interpretation. This allows metabolomics end users to extract more meaningful insight from the biological systems studied with a potential repercussion on all the fields making use of it.

Chapter 4

Extended metabolomics workflow for biological sciences

4.1 Introduction

LCMS based Metabolomics is a fast evolving field [130] which often requires the intervention of many stakeholders. From experiment to design to biological insight, principal investigators, laboratory scientists, mass spectrometry technicians, bioinformaticians, statisticians, biochemists and other scientific field specialists can participate in order to carry a single experiment to completion. Chapter 3 demonstrated that the burden on some of the contributors such as bioinformaticians, statisticians and biochemists can be reduced by improving the support of the analysis and interpretation of the metabolomics data, metabolomics experiments still require the intervention of many specialists. Two main issues can emerge from this multi-contributor study set up: (i) first, the scientist in charge of the study, generally the principal investigator or the lab scientist (biologists, clinicians, PhD student or postdoctoral researcher), often has limited knowledge of LCMS technologies and what they can offer, and therefore risks taking an approach that is not optimal for their study design. (ii) Secondly, the documentation of every step of the study becomes a shared task where every piece of information needs to be recorded and passed on to the next contributor. The careful documentation of a study is imperative in order to allow the scientist reporting the work to provide information about the experiment performed in its entirety and with exactitude. A minimum reporting scheme was proposed by R. Goodacre *et al.* in 2007 [131]. This reporting task represents a real challenge not only because of the number of contributors involved in one study but also because the reporter may not be familiar with the technical approach and protocol followed during the LCMS data acquisition. The issue could be overcome by standardising and unifying the data capture and documentation of metabolomics studies, but also by providing to scientists in charge of a study more information about what a metabolomics

experiment can offer to answer the specific questions their study addresses.

As described in Chapter 3, data analysis capabilities need to continually adapt in order to offer a better understanding of the increasingly complex data generated by LCMS approaches. While LCMS based metabolomics applied to biological science can provide a detailed snapshot of the metabolism state of a biological system, the variable annotation confidence of the compounds found in the system currently poses great limitation [132]. Mass spectrometry technologies allow acquisition of fragmentation data using data dependent or data-independent acquisition (DDA and DIA) to further inform on the structure of molecules analysed [133]. In untargeted metabolomics, taking advantage of these technologies can improve the annotation confidence and provide a better representation of a biological system. However, coupling the analysis of fragmentation data generated by DDA or DIA to a standard untargeted metabolomics pipeline is imperative in order to make this possible. The development of new data analysis features to fit technologies advances is however not the only adaptation challenge that the field is facing. Biological resources are constantly growing, and computational tools applied to systems biology are constantly emerging which create great potential for LCMS based untargeted metabolomics data to be further investigated. Network models, such as genome-scale metabolic reconstruction, for example, represent a promising approach for metabolomics data to be interpreted in [134]. Not only these metabolism models can provide biological context, but they can also serve as a support to expand a study to other omics technologies. Expanding metabolomics pipelines to integrate the analysis of LCMS data within metabolic networks could have a significant impact on the potential biological insight that can be extracted from the data.

4.2 Related work

Laboratory Information Management Systems supporting the recording of laboratory procedures for multi-omics technologies have already emerged at the time this project started. However, no freely available tool offer multi-omics support and only proprietary software can be used at considerable costs. Whilst some data repositories have begun appearing [135] allowing scientist to report their studies thoroughly, these tools do not offer any support for information capture at the time of the study. The support for designing untargeted metabolomics experiments was also non-existent and rely on the scientist in charge of the study to investigate and learn about the technology to design his own study. Other parts of the metabolomics workflow such as fragmentation data analysis are however better supported. Several online repositories allow similarity based comparison of experimental fragmentation data to annotated reference spectra. MassBank [71], for example, was introduced in 2010 and keeps expanding its library of reference spectra and tools available. MAGMa [136] was

first published in 2012 and used for the automatic annotation of a complete metabolite profile of green tea in 2013 [137]. No tool or repository, however, provides a way of coupling standard untargeted metabolomics data analysis to fragmentation data analysis to improve compound identification. Finally, some software offers the possibility to visualise and mine networks, some tools such as Tulip [138] can be used for any type of relational data, others are dedicated to the study of biological networks [139]. They, however, present the same limitations as the fragmentation data analysis tool as they are third party tool which needs to be used on their own. Moreover, performing a network reconstruction from metabolomics data require a deep understanding the underlying data structure and bioinformatics skills. This task can currently only be conducted by scientist expert in the field.

As seen in the previous chapter, the core analysis of LCMS data is supported by many tools offering very divers features with varying coverage of the data analysis pipeline. However, LCMS experiment does not only consists in data processing and many other aspects of LCMS metabolomics studies require specific support. As discussed previously, the existing tools that support these step of LCMS metabolomics workflow are either non free, non LCMS specific and often disparate. The overarching aim of the work presented in this section is therefore to better support important steps of the LCMS workflow in a unified environment. Details on the different steps to support are given below in 4 specific aims and objectives.

The work presented in this chapter will try answer the following questions:

- Can the documentation and data capture of a metabolomics study be unified to facilitate an accurate reporting of the work?
- Is it possible to accurately inform biologists on the potential outcome of a metabolomics experiment according to the system studied?
- Can metabolomics pipeline responsively integrate new analysis capabilities to match the advances in LCMS technologies?
- Can biological network analysis be integrated to metabolomics pipelines to expand the context of interpretation?

Four aims were outlined in order to address these research questions:

1. Support contributors of a metabolomics study in documenting each step of the study within the same environment.
   Objective 1: Develop a tool allowing the easy capture of all necessary information during a metabolomics experiment.

2. Provide biologists with information about the compounds that can be detected in a particular biological system using untargeted metabolomics.
   Objective 2: Integrate a new module to the PiMP platform allowing the mapping of a set of compounds onto metabolic networks prior to analysis.

3. Make use of PiMP modular design to integrate support for fragmentation data analysis as part of the untargeted pipeline.
   Objective 3: Integrate a new django module to PiMP to allow the analysis of fragmentation data alongside MS[1].

4. Expand the PiMP data interpretation environment to allow biological network analysis and visualisation.
   Objective 4: Integrate to PiMP data exploration environment a network visualisation module allowing the visualisation of analyses results in the context of biological networks.

4.3 Supporting study documentation

The work presented in this section addresses the first two objectives outlined in the previous part of this chapter. In chapter 3 was discussed the support of different tasks such as data capture or quality control during the data analysis steps of the metabolomics workflow. However, the purpose of the work presented in this section is to support every step of the workflow upstream to the data analysis. Figure 4.1 shows the area of the workflow for which the work presented here attempt to provide support with.

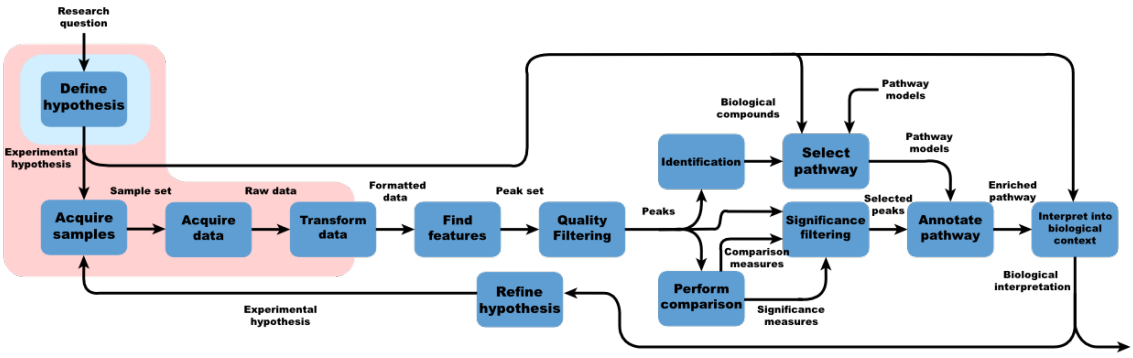

Figure 4.1: Area of limitation in the metabolomics workflow that need support. Highlighted in light blue is the hypothesis definition and study design that need informed guidance. Highlighted in red is sample preparation and data acquisition steps which require a unified documentation and data capture support.

4.3.1 Project management system

The approach taken to unify the documentation across every step of the workflow to the data analysis part is to develop a web-based tool that offers access and support to every contributor involved in a metabolomics experiment capturing all pieces of information necessary to document a study. As the omics technologies are constantly evolving, the developed tool requires a robust and flexible data structure as well as a modular design to be adaptable and scalable. Indeed, as the technologies mature, new information might need to be captured at every step of the workflow. In order to be flexible, the structure of the project management system follows the same design pattern as PiMP presented in chapter 3. The tool presented here was primarily developed to support metabolomics experiment documentation based on the metabolomics workflow shown in Figure 4.1; it was then expanded to support the documentation and data capture of three other omics: genomics, transcriptomics and proteomics.

Data structure

The data structure needs to support two types of users: (i) the principal investigator or lab scientist in charge of the study (collaborator), (ii) the metabolomics technologists, bioinformaticians or other scientists (staff) contributing to the workflow. The main difference between these two user types is that the staff contributors use the tool to record information about the work being done, while the PI of the study uses the tool to access this information. The action that needs to be supported for the PI is therefore limited to accessing the information recorded by the staff users and attached to his projects. The staff users need greater support to be able to capture the information at every step of the workflow. The data structure supports the two different types of user using the table "User", its field "is_staff", and the table "Collaborator". A staff user will, therefore, have an entry in the User table with the field "is_staff" set to True. A collaborator user would have an entry in the User and Collaborator table with the field "is_staff" set to False. This flexible design allows one user to either be staff, collaborator or both.

The "Group" table in the authentication module is used to separate the staff users into the different omics; one user can be involved in one or more omics fields. This information can then be used for various purposes such as knowing what staff user can contribute to a specific omics project.

The "Project" table is the main table that organises the information that needs to be captured for a study. The "Genomics", "Proteomics" and "Metabolomics" table inherit from the "Project" table; they, therefore, share the fields defined by the "Project" table, and also define their own fields in their respective table. As shown in the diagram in Figure 4.2, a project is divided into a set of tasks, each task being assigned to a staff user. The "Task"

table, therefore, contains all information about every task performed during a study such as its status and date of completion. The note field can contain free text entry to give further details about a particular task, and the "Comment" table allows staff users to capture any issue or comments they may have encountered while performing a task.

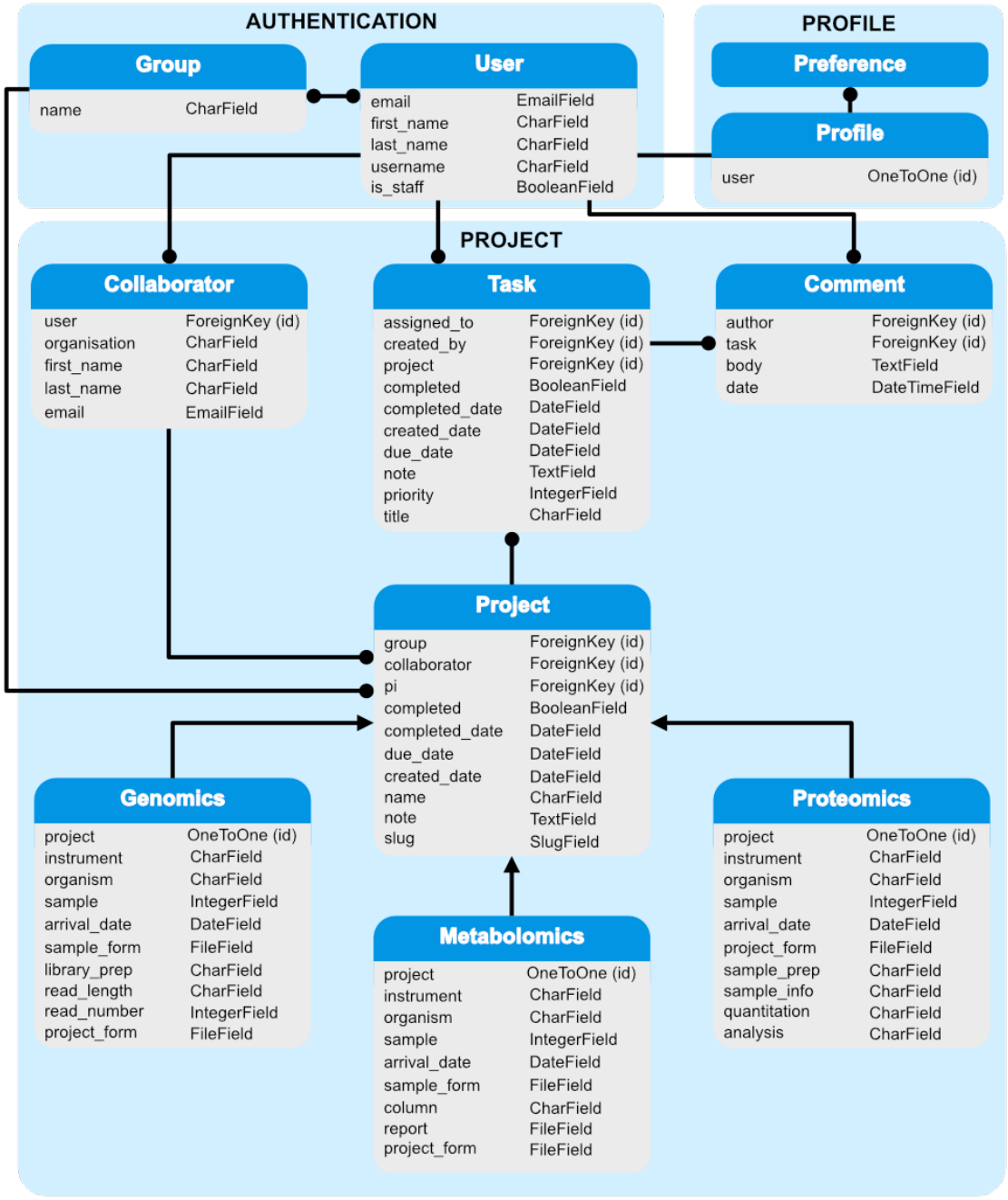

Figure 4.2: Database structure of the project management system. The structure define three modules, each of them used to store different type of data. The authentication module is used to store user's information and authentication details. The profile module store users' preferences. The project module store and organise the information related to experiments. The inheritance design of the project table makes it extendable to any other omics.

Three omics are currently supported and shown in the data structure in Figure 4.1, but the inheritance design enable an easy extension of the structure to other omics. The metabolomics table supports LCMS and GCMS metabolomics experiments, allowing the capture of every component of the workflow such as instruments and columns used, as well as the organism studied, the number of samples or the storage of files related to the experiment such as forms and reports. The proteomics table follows the same pattern as the metabolomics table with different fields specific to proteomics experiments. Finally, the genomics table supports DNA and RNA sequencing experiments allowing the capture of all information related to this omics technology.

As the management system was primarily developed to support LCMS metabolomics, the data captured by the tool is general enough to support other metabolomics laboratories. The genomics and proteomics parts of the system were, however, designed for Glasgow Polyomics platforms only and are too specific to be transferred to other omics laboratories.

Web-enabled tool

The aim of the project management system is to support all contributors of an omics study, and more specifically a metabolomics study, in documenting their experiments within a unified environment. Two main requirements were drawn to address this objective, the tool and data captured needs to be readily available for all contributors, and the users need to be guided during the documentation process. A web-enabled tool following the same MVT design as PiMP presented in chapter 3 was chosen to address the first requirement. The second requirement was addressed by structuring the data capture task. The documentation process requires the capture of different types of information that can be separated into two groups: (i) Static information, (ii) dynamic information. The static information is usually captured at a particular time of the project and is not or rarely changed after. Most of the static information such as the organism studied, the number of samples or the instrument chosen to perform the experiment is set at the beginning of the project, but other information such as quality control reports on the data acquired can be recorded later in the project. The second type of information which is dynamic evolve during the project, this type of information relates to a specific step of the workflow involving an action such as the sample preparation, data acquisition or data analysis.

The two different types of data to capture are reflected in the data structure by simple fields in the project table for the static information, and the task table for dynamic information. The task, therefore, supports the capture of complex information such as time stamp, completion status or the staff member performing the task. Every step of the metabolomics or other omics workflow requiring an action is therefore translated into a task to store all the information related to it.

The user interface is accessible through a web browser and is developed using the same web standard as PiMP. Figure 4.3 shows one page of the user interface.

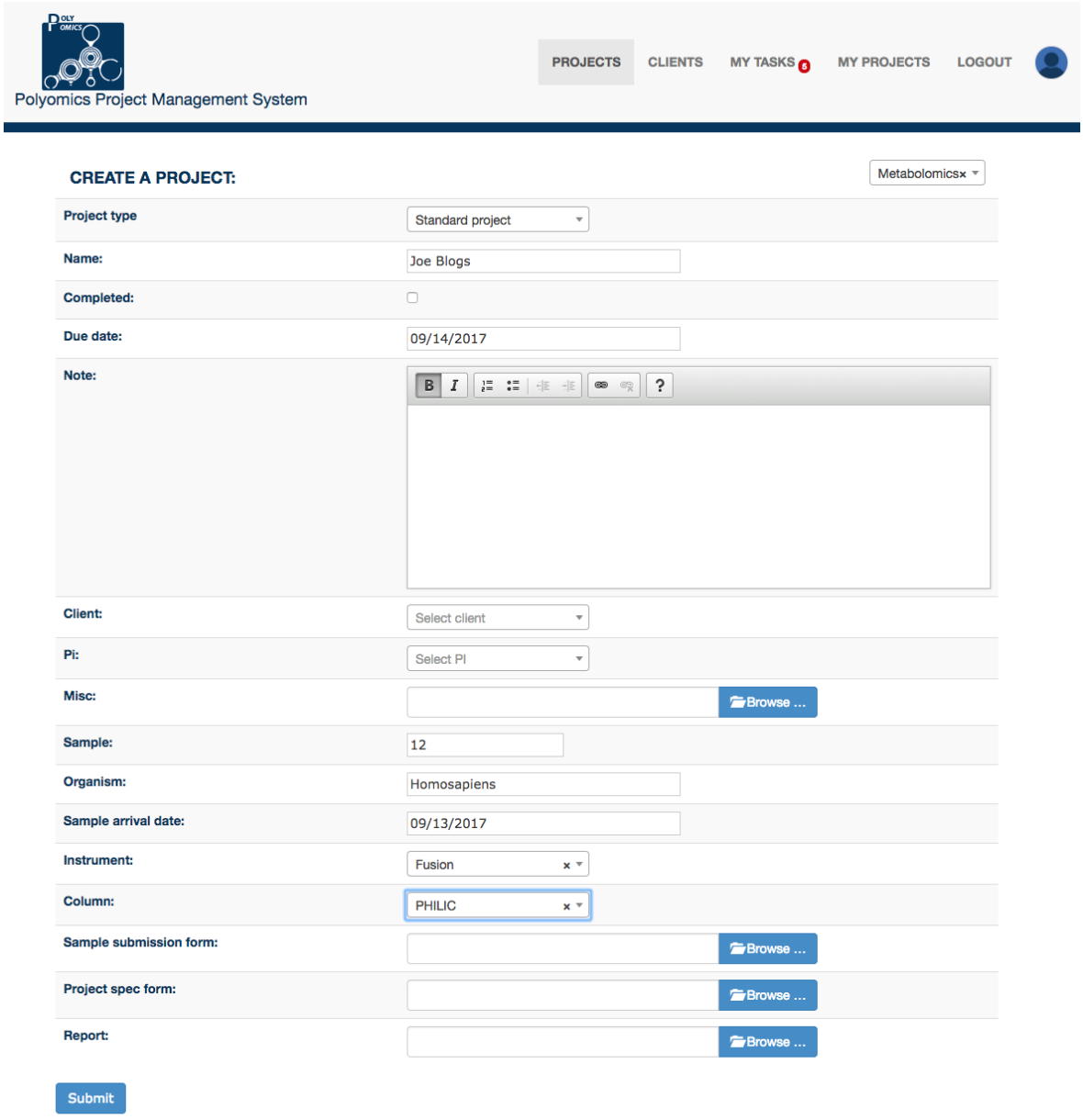

Figure 4.3: Screen shot of the user interface of the management system. This picture shows the form to be filled in to create a new project. The navigation bar at the top allows the user to access the project list, the client list, its assigned tasks, and its account and preferences. The client users only have access to ”my projects” page to visualise the details of their own projects and their progress. The management system user interface is developed using the same web technologies as PiMP (Django, html5, CSS3, javascript)

4.3.2 Biochemical library

Untargeted metabolomics aims to provide a snapshot of the metabolism state of a biological system at a specific time. It is therefore not possible to predict accurately the pool of chemical compounds that will be detected. However, the method chosen and the system studied can be used to provide information on chemicals that can potentially be detected and therefore help biologists designing their experiments. The chemical library presented in this section aims to help biologists understand what part of the metabolism of the system studied can be seen using untargeted metabolomics. The work from this section has been published in *Frontiers in Molecular Biosciences* [140] and tries to inform biologists about the potential outcome of an untargeted metabolomics experiment according to the organism studied (aim number 3). The role of the author was to develop the user interface and controllers on the PiMP side, the communication protocol between the two servers in collaboration with MetExplore developers.

The approach taken here is to use the list of standard compounds run at Glasgow Polyomics routinely for metabolite identification and map them onto genome-scale reconstruction of metabolic networks available in MetExplore [88]. The information returned by this type of approach would allow the user to know the coverage of the standard compounds on a particular metabolic network or organism. PiMP was used to communicate with MetExplore and display the results to the user.

Metabolite identifiers

In order to map metabolites present in Glasgow Polyomics standard compound library onto MetExplore's metabolic networks, the same identifiers need to be used by both tools. Many database specific identifiers can be utilised for metabolites such as KEGG [65], ChEBI [141] and PubChem [66] identifiers. However, those identifiers cannot be used for this type of mapping as they are not commonly used to reference metabolites in metabolic networks, and some metabolites found in metabolic networks are not referenced in any of those databases. While the metabolomics community is currently putting effort in standardising the identification of metabolites using specific identifiers and controlled vocabulary [128], alternative identifiers based on chemical structures can be used to overcome the issues met with the database specific identifiers. MetExplore uses the InChI (IUPAC International Chemical Identifier) and InChIKey identifiers to reference the metabolites in metabolic networks. The InChI identifiers provide a non-ambiguous identification of compounds organised in layers that provide different information about the structure of the molecule. The InChIKey is a hashed version of the InChI forming a 27 uppercase characters identifier. The InchIKey can be calculated from the InChI using a hash algorithm and is the identifier that was selected to

map metabolites from the Glasgow Polyomics chemical library to MetExplore's metabolic networks.

Metabolite mapping and communication protocol

Communication between the network database and the chemical library is necessary to map the metabolites onto the biological networks. The protocol proposed for this communication is based on a dialogue between web services located on the two servers. As illustrated in Figure 4.4, a four steps dialogue process has been created to perform the mapping and receive the results back on the chemical library server to be then presented to the user. The communication process is initiated by the chemical library server (PiMP) informing the network web server that a mapping is requested and providing some specific information. The information provided is the location (URL of chemical library web service) of the list of metabolite to map. As discussed in the previous section, the identifiers used for the metabolites are InChIKey identifiers formatted as a JSON array. Once the call is received by the network web service, the URL passed by the chemical library is used to retrieve the list of metabolites to map by calling the chemical library web service. This second call goes therefore from the network server to the chemical library server. The chemical library server simply returns the list of InChIKey identifiers that need to be mapped as a JSON array. The network server then performs the mapping and returns the results to the chemical library server as a response to its very first call.

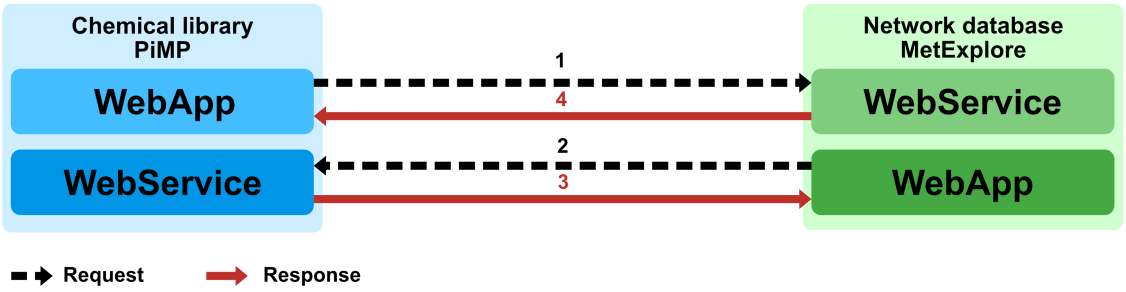

Figure 4.4: 4 steps communication protocol between the chemical library and network database. **1**. PiMP contact the newtork library to request a mapping. **2** and **3**. MetExplore gather the list of InChI using the chemical library webservice. **4**. The result of the mapping is sent back to the chemical library.

Metabolite mapping response

The results are formatted according to the JSON encoding, divided into sections corresponding to individual BioSource (MetExplore's biological network). Each BioSource section contains general information related to the BioSource itself, its name and organism strain,

the source (KEGG, BioCyc, SBML), the version number and MetExplore identifier. The section also contains information related to the network, the total number of metabolites in the network, the total number of metabolites which have an identifier (InChIs or InChIKeys), the total number of unique identifiers present in the network. Finally, each BioSource section contains information related to the mapping, the total number of identifiers from the network mapped in the chemical library and the total number of unique identifiers from the network mapped in the chemical library. Those two numbers may be different from the total number of metabolites in a network as it is based on the compartments (cellular compartment), this means that if a metabolite is present in n compartments it will then be counted x times. If a metabolite is present in different parts of the network within the same compartment, it is considered as one. The mapping information also contains the library and network coverage, respectively the relative number of library identifiers mapped on the network and the relative number of metabolites from the network present in the library. It also contains the percentage of identifiers found in both the library and the network compare to the number of unique identifiers in the network. The last information is a MetExplore mapping id, which allows the user to access the mapping results directly in MetExplore user interface.

Metabolite mapping results

This tool was developed within the context of Glasgow Polyomics (GP) metabolomics platform services; the GP compounds library contains a list of 240 metabolites that are routinely run as standard compounds for identification purposes during the data analysis. GP is involved in a wide range of research areas; it is therefore important for GP users to have access to the coverage information of a maximum number of organisms. Thus, the mapping of GP chemical library is performed on all networks available in MetExplore database. The results of the mapping returned by MetExplore web service is presented as table accessible through PiMP web interface. The table is automatically generated by PiMP Django backend which parses the JSON formatted results and converts it in a Javascript enriched HTML file. The table (presented in Figure 4.5) currently contains almost 60 different metabolic networks with the mapping information attached to it, Javascript functions make the table interactive allowing the user to search, sort and filter it. The name of each network is also clickable to allow the visualisation in a new web browser window of a selected mapping in MetExplore. Figure 4.6 and 4.7 respectively show the metabolite mapping and pathway coverage information as seen in MetExplore. Finally, the tools available in MetExplore allow the user to visualise the entire network or a selected subnetwork as shown in Figure 4.8.

Standard compound library

Show 25 entries Search:

Metabolic network	Number of metabolites in the network	Metabolites found in Polyomics standard compound library	Coverage
Acinetobacter baumannii	680	108	15%
Agrobacterium tumefaciens	881	81	9%
Arabidopsis thaliana	1547	101	6%
Arabidopsis thaliana	1664	172	10%
Bacillus amyloliquefaciens	672	85	12%
Bacillus anthracis	789	83	10%
Bacillus subtilis	1143	145	12%
Bacillus thuringiensis	766	82	10%

Figure 4.5: Glasgow Polyomics standard library table showing the coverage of the first 8 metabolic networks (alphabetically sorted).

	Name	Identifier	Formula	Compartments	Mapping_1
					Identified
1	3-methyl-2-oxopentanoate	M_3mop_m	C6H9O3	m	true
2	3-methyl-2-oxopentanoate	M_3mop_e	C6H9O3	e	true
3	3-methyl-2-oxopentanoate	M_3mop_c	C6H9O3	c	true
4	4-hydroxybenzoate	M_4hbz_m	C7H5O3	m	true
5	4-methyl-2-oxopentanoate	M_4mop_e	C6H9O3	e	true
6	4-methyl-2-oxopentanoate	M_4mop_c	C6H9O3	c	true
7	4-methyl-2-oxopentanoate	M_4mop_m	C6H9O3	m	true
8	5-Methylthioadenosine	M_5mta_c	C11H15N5O3S	c	true

Figure 4.6: Metabolite table as seen in MetExplore showing the compounds mapped in an extra column.

	Name	Identifier	Mapping_1 on Metabolite		
			Coverage	Nb of Mapped	Right tailed Fisher exact test
1	Transport, extracellular	Transport, extracell...	2.13	20	9.67e-8
2	Exchange/demand reaction	Exchange/demand ...	1.48	12	3.52e-3
3	Transport, mitochondrial	Transport, mitocho...	3.01	11	1.10e-5
4	Nucleotide interconversion	Nucleotide intercon...	3.88	8	3.78e-5
5	Pyrimidine catabolism	Pyrimidine cataboli...	7.14	5	7.68e-5
6	Transport, nuclear	Transport, nuclear	3.85	5	1.38e-3
7	Transport, lysosomal	Transport, lysosomal	1.91	4	4.53e-2
8	Valine, leucine, and isoleucine ...	Valine, leucine, and...	5.63	4	1.07e-3
9	Biotin metabolism	Biotin metabolism	6.25	2	1.81e-2
10	Fatty acid oxidation	Fatty acid oxidation	0.21	2	9.92e-1
11	Glycine, serine, alanine and thr...	Glycine, serine, ala...	2.2	2	1.18e-1

Figure 4.7: Pathway table as seen in MetExplore displaying information related to the mapping.

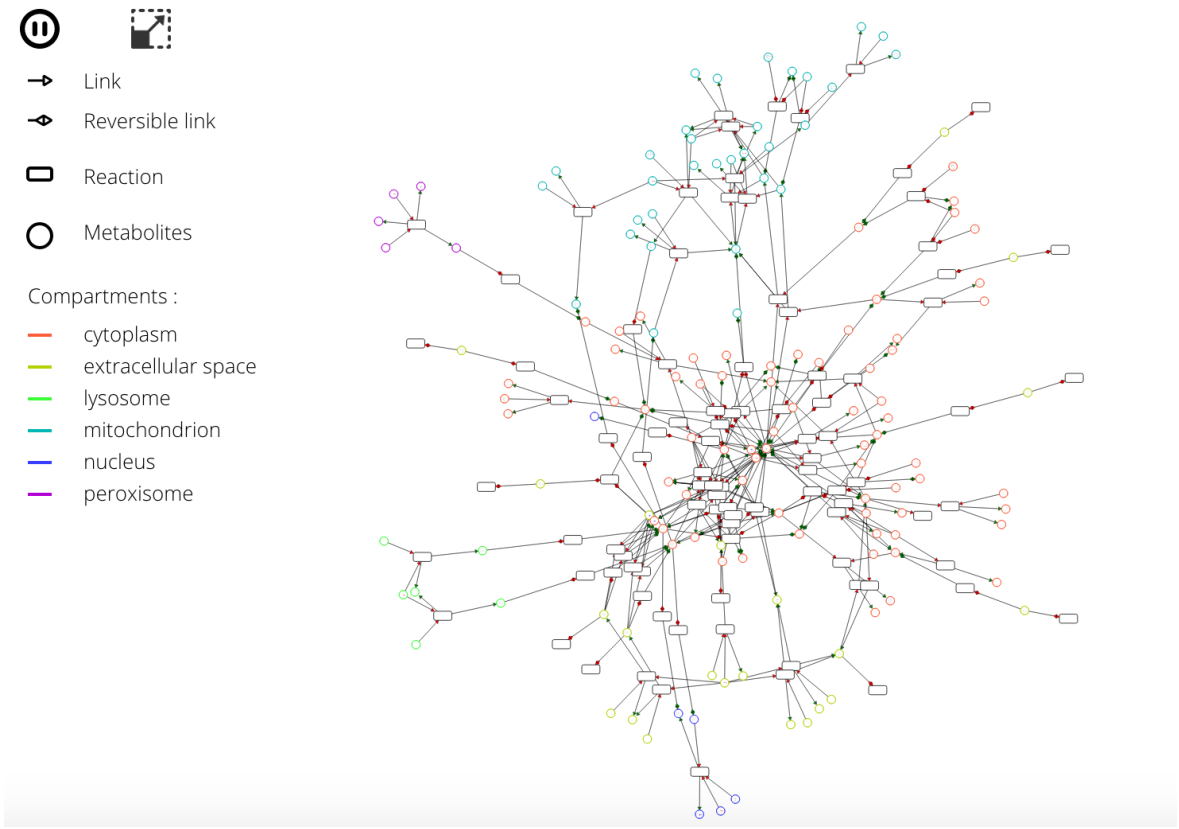

Figure 4.8: Visualisation of the network resulting from the mapping within MetExplore.

4.4 Fragmentation data analysis

> "Fragmentation is when you make big bits of stuff into teeny tiny wee bits."
>
> *Erin D. T. Manson*

Metabolite identification remains the major challenge in metabolomics. Although external standards increase the degree of confidence in the identification of compounds, LCMS based metabolomics provides evidence to support metabolite identification but rarely with absolute certitude. This level of uncertainty is even more applicable to metabolites absent from the external standard compounds and therefore only annotated from the mass. Whilst in some cases putative annotation can be informative enough to guide the biological interpretation of results and extract meaningful insight from a dataset, it often leads to further investigation to ascertain the identity of a compound that might play an important role in the system studied. It is possible to routinely run liquid chromatography tandem mass spectrometry on mass spectrometers [142] (LCMS/MS). MS/MS acquisition produces product ion spectra (Figure 4.9) from which compounds' structural information can be derived. Coupling LCMS to MS/MS acquisition can, therefore, improve the level of confidence with which the compounds are identified and lead to a better interpretation of the results. As shown in Figure 4.10, the identification step is upstream of the biological interpretation in the data analysis pipeline, the quality of compound identification has therefore a critical impact on the biological interpretation.

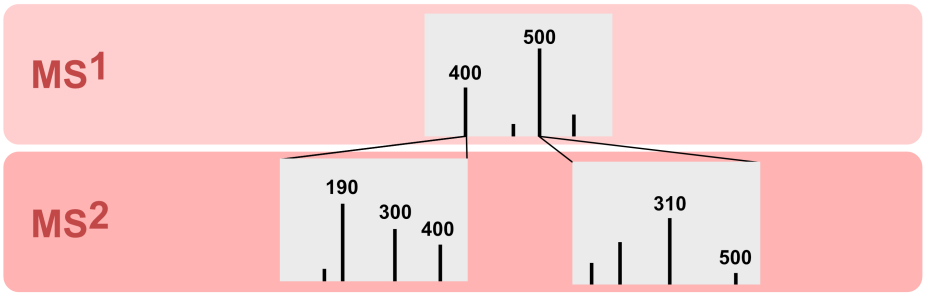

Figure 4.9: Representation of fragmentation spectra produced by tandem MS.

Two type of acquisition can be used to gather MS/MS information, (i) Data-Dependent acquisition (DDA) which is supervised and only fragment precursor ions above a predefined abundance threshold, (ii) Data-Independent acquisition (DIA), however, fragments all ions within a certain m/z window without selection between compounds eluting at the same retention time. However, fragmentation spectra produced by DIA are more complex to analyse as no precursor ion selection is performed [143].

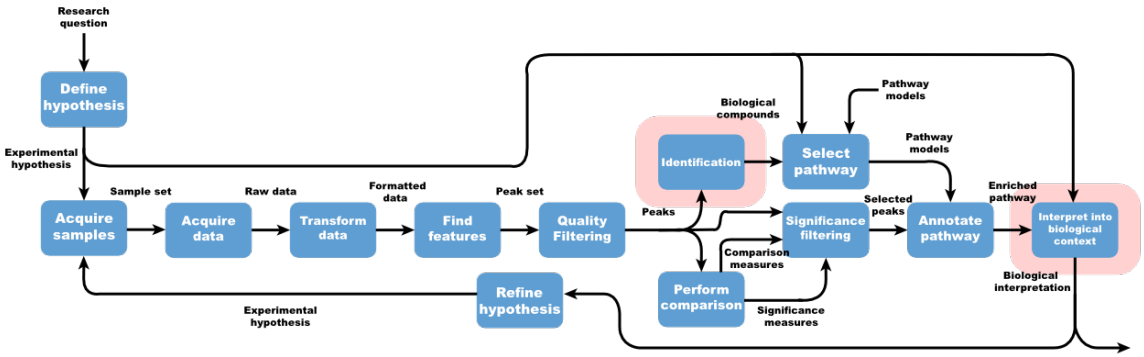

Figure 4.10: Fragmentation helps improving the identification part of the metabolomics workflow and ultimately improves the biological interpretation of the results.

The approach taken here attempt to support DDA fragmentation data analysis as part of the untargeted metabolomics workflow to improve metabolite identification. The tool developed was designed as a flexible internal PiMP plugin to be used as part of the data analysis pipeline in place or as a separate analysis tool. The Fragmentation Annotation Kit (FrAnK) was implemented by Scott J. Greig, Karen McLuskey and Joe Wandy, the role of the author was to coordinate the development and the integration of the tool within PiMP.

As FrAnK is developed for identification purposes, the minimum requirement for the tool to run is the upload of the files corresponding to the pool samples which contain fragmentation data. Indeed, as the pool samples contain every compound present in all experimental samples of a study, it is the only input data required to perform the peak annotation.

4.4.1 Annotation tool and library

Several tandem mass spectra libraries and tools to search them and compare experimental mass spectra to known and annotated fragmentation spectra exist. However, many of them are not adequate for the development of FrAnK. Indeed, some libraries are only accessible through a web browser, others offer a web service alternative but do not allow the submission of several peaks at the same time. Finally, although some tools are programmatically accessible, they do not permit local installation and force the database search to be performed through the web which increases the processing time greatly. The tool chosen to overcome these limitations is MS PepSearch, which was associated with the MassBank [71] and the US National Institute of Science and Technology (NIST) library. The tool and libraries were deployed in a Docker [144] container so they can be installed and used easily on any operating system.

4.4.2 FrAnK architecture and design

As FrAnK was developed as an internal plugin of PiMP, the same design approach was taken. FrAnK is, therefore, a complex Django app that extends the core functionalities of PiMP. As explained in Chapter 3, PiMP is built using three main modules supporting data capture, data analysis and result exploration, each modules being divided into 'apps'. As FrAnK needs to support fragmentation data analysis both independently and as part of the PiMP pipeline, it was designed as a single app supporting data capture, analysis and result exploration. The FrAnK app, therefore, defines its own data models, its own views and templates.

The data processing happens asynchronously using the same task system as PiMP. A python wrapper was created around the annotation tools to allow communication between the task layer and the data analysis pipeline. This means that no exchange format is required to transfer the data output of the annotation tools back to the task and store them permanently in the database.

4.4.3 Data capture and visualisation

Data capture visualisation support was developed so the tool can be used independently from PiMP. The data capture follows the same design as PiMP asking the user to perform a sequence of tasks from file upload to starting the annotation pipeline. Metadata can also be entered to document the experiment and analysis performed.

The results of FrAnK annotation pipeline can be explored in a dedicated data exploration environment and consist of two main pages. The first page display a table containing the list of MS^1 peaks detected with associated information such as the mass, retention time and intensity. Each peak can then be explored further by clicking on the identifier, giving access to the "single peak" page displaying the fragmentation spectra and the different putative annotations respectively shown in Figure 4.11 and Figure 4.12. Extra information is given such as the structure of the compound annotated by the peak and confidence score returned by the annotation tools.

4.4.4 PiMP-FrAnK integration

In order for the user to be able to use FrAnK seamlessly within PiMP, the two tools had to be integrated at different levels. However, as PiMP offers a modular design, only very specific parts of the source code had to be modified with no impact on the rest of the tool. Three main points of communication had to be created for the tools to run in concert. First, a connection between the two different data structures is required to give PiMP access to the fragmentation data. Then, PiMP "data capture" modules need to encapsulate FrAnK data capture to unify

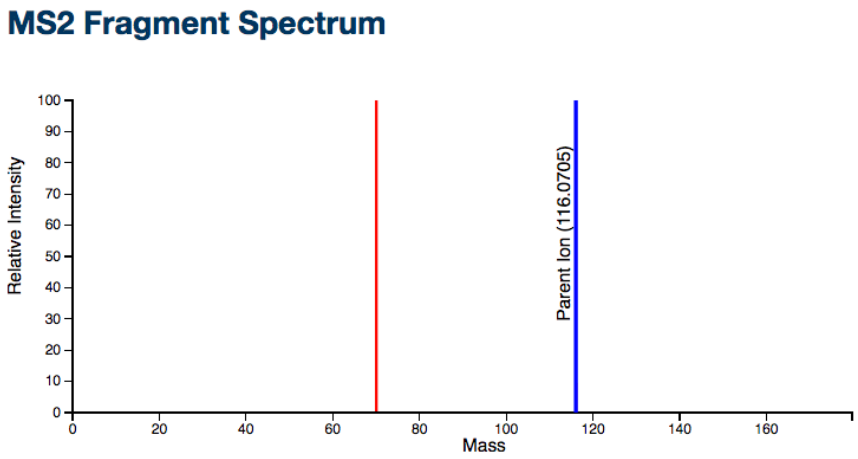

Figure 4.11: Fragment spectrum as seen in FrAnK dedicated data visualisation interface. Courtesy of Karen McLuskey, personal communication.

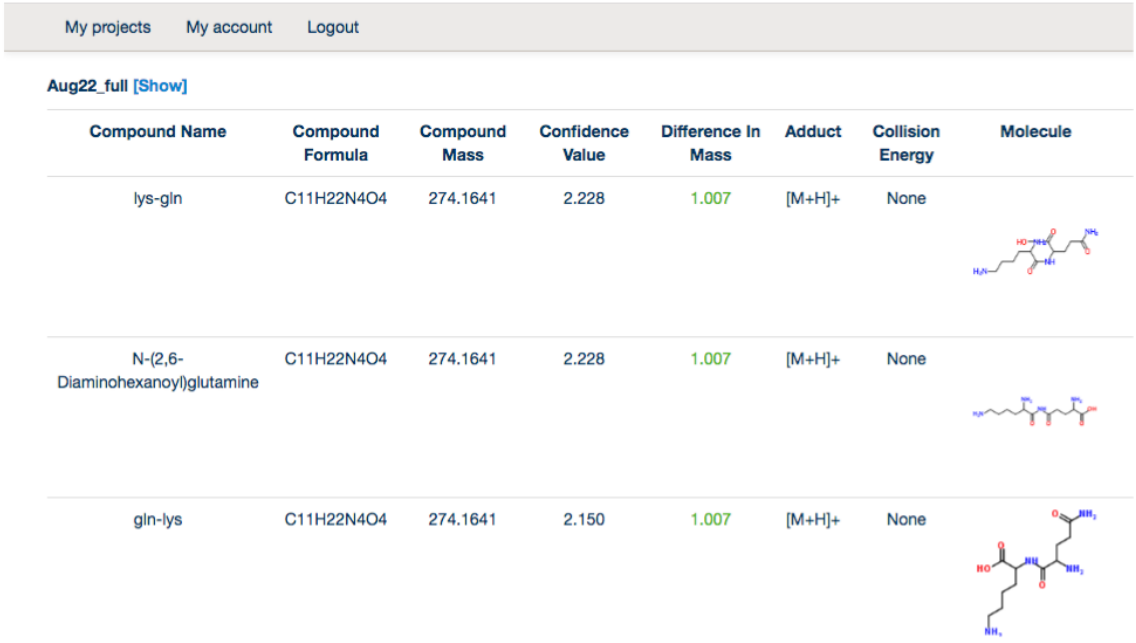

Figure 4.12: FrAnK dedicated annotation page. Courtesy of Karen McLuskey, personal communication.

the data capture as a single task, and finally, PiMP data interpretation module has to present the fragmentation results to the user.

Data structure connection

FrAnK was developed as an integral part of PiMP but also a fully independent Django app. The same technologies were used across the two tools. They, therefore, share a MySQL database, both tools defining its own tables with the exception of the user table which is

shared between the two systems. For the tools to communicate and share data, two joining database tables had to be created to form a cohesive data structure. The first joining table connects an experiment in FrAnK to a PiMP project as shown in Figure 4.13; this high-level join allows PiMP to access FrAnK data tables corresponding to the data capture.

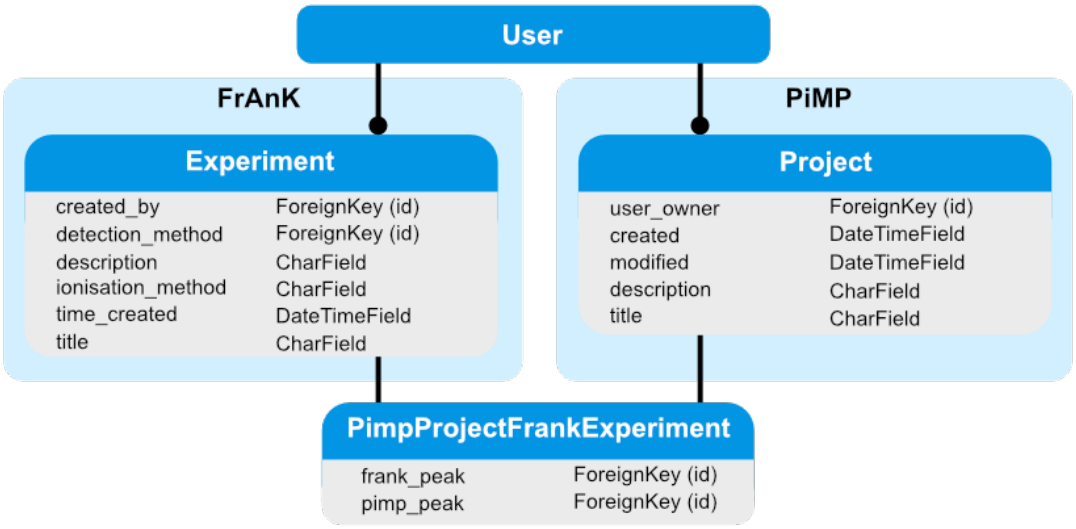

Figure 4.13: Connection between PiMP and FrAnK at the project level of the data structure. The user table is shared between the two tools.

The second joining table was created to unify the results of the two pipelines at the peak level; this joining table, therefore, connects MS^1 peaks from FrAnK to peaks in PiMP (Figure 4.14). This join creates a ”one to one” relationship between MS^1 peaks found in the two systems. This connection gives then an extra layer of information for every peak extracted by the PiMP analysis pipeline which has now access to the fragmentation annotations generated by FrAnK.

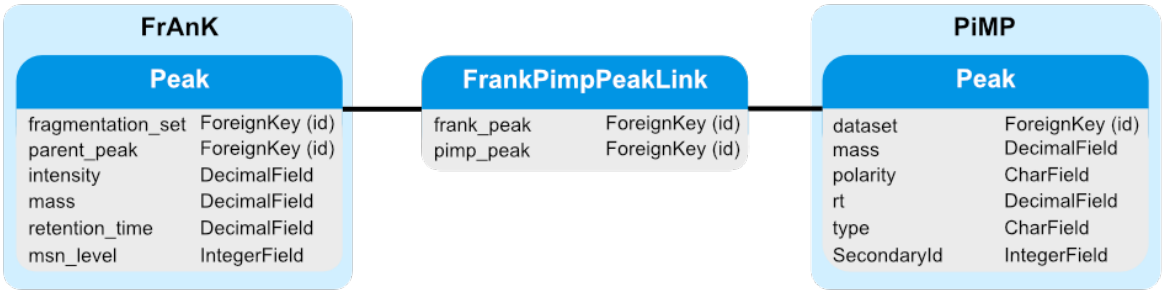

Figure 4.14: Connection between PiMP and FrAnK at the peak level of the data structure. This connection allows to bring together the results of the two data analysis pipelines.

Integrated data analysis pipeline

Whilst LCMS/MS gives in-depth information on the structure of the compounds analysed, the acquisition of several isolated MS/MS spectra for each MS spectrum typically requires a longer duty cycle than for MS alone. This results in a lower number of mass scans acquired per run for the MS^1 as seen in Figure 4.15. As peak detection algorithms' performances are closely related to the peak shapes and therefore the number of time points available to assess if a signal corresponds to a peak or simple noise, the reduction in datapoints reduces the quality of peak detection. Peak detection can be used independently in FrAnK. The Peak detection in PiMP is more robust, applied to MS^1 data only. This can, however, be used as an advantage by using the list of peaks detected in PiMP to feed the fragmentation data analysis. Consequently, when FrAnK is run as part of the PiMP pipeline, no peak detection step is required. This implies that the two data analysis pipeline cannot be run simultaneously as FrAnK requires an input from PiMP data analysis pipeline. As shown in Figure 4.16, the asynchronous task system used to run the data analysis pipelines offers the possibility to chain tasks. The fragmentation data analysis pipeline, when running as part of PiMP, only starts when the untargeted metabolomics data analysis pipeline is complete. The implementation of the modular pipeline architecture is detailed in section 3.4.4 of chapter 3. The connection between the peak entries created by both tools is performed as an extra trivial step during the storage of the results in the database.

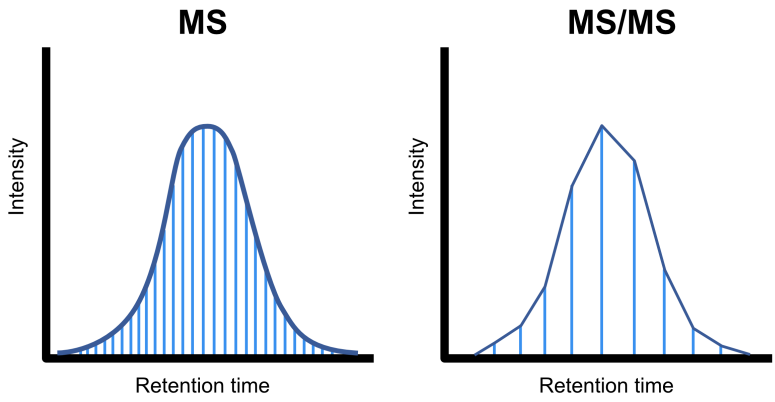

Figure 4.15: Schematic of a MS^1 chromatographic peak comparing MS and MSMS acquisition. A higher number of number of MS^1 mass scans in single MS allows better resolution.

Integrated data visualisation

That last integration step happens in the template layer in order to let users start a fragmentation data analysis as part of PiMP. PiMP data capture template was therefore extended to allow the upload of fragmentation files. When PiMP detects the presence of fragmentation

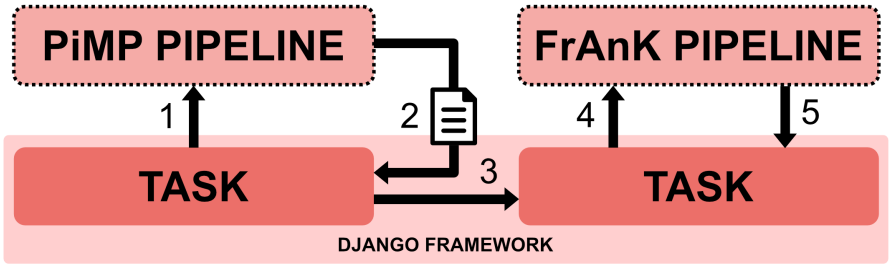

Figure 4.16: Representation of chained pipelines, FrAnK using the results of PiMP as an input, this input being the list of detected peaks.

file, it allows the user to choose to run FrAnK annotation pipeline at the time of starting the analysis using a simple tick box. The upload of fragmentation files is optional to the user who can choose to run the PiMP data analysis pipeline on its own.

The results of the fragmentation data analysis were also integrated into PiMP data exploration environment. The first point of integration is located in the peak tab where an extra column is added to the table to display fragmentation information. When a peak entry is selected, the right panel gives direct access through a link to FrAnK dedicated result page for the selected peak (Figure 4.17).

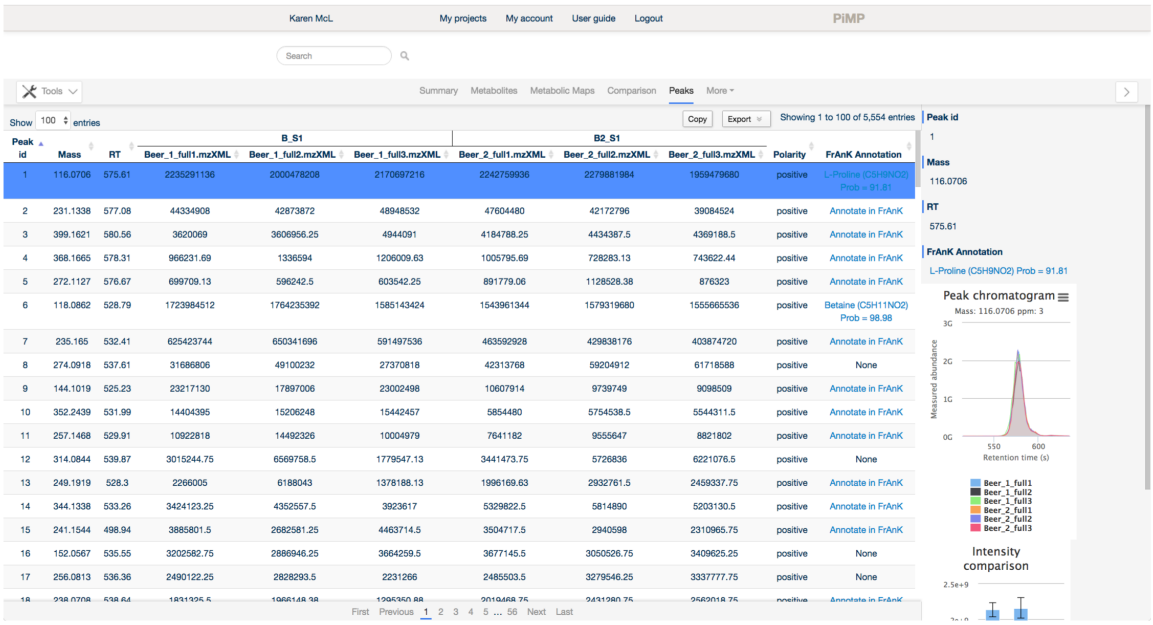

Figure 4.17: Integration of FrAnK information in the peak tab of PiMP data exploration environment. An extra column shows if a fragmentation data annotation exist for every peak. The right panel also shows this information and allow access to FrAnK result page. Courtesy of Karen McLuskey, personal communication.

The fragmentation information was also integrated into the metabolite tab of PiMP data exploration environment. As this tab is central to the interpretation of the results, including

fragmentation information in this part of the user interface can greatly help the user in assessing the quality of peak annotations. The fragmentation information is embedded in the contextual right panel of the metabolite tab by extending the peak card. As shown in Figure 4.18, when a peak has fragmentation information attached to it, an extra line is added to the peak card giving the name of the compound with the highest score provided by the fragmentation analysis. An extra button at the bottom of the card also gives access to FrAnK dedicated result page (opened in a blank page of the browser) to explore the MS^n spectra and the different possible annotations.

Figure 4.18: Peak card from the metabolite tab of PiMP data exploration environment showing fragmentation data. FrAnK best annotation is given and an extra button (in green) gives access the to FrAnK dedicated peak page.

4.5 Biological network analysis

As outlined in the different chapters of this document, the interpretation of the results of the data analysis of an untargeted metabolomics experiment is a complex task. This is due to several factors such as the limited features that the existing software offer but also the complexity of the data. In chapter 3, the issues in supporting the metabolomics users were partly addressed by providing some biological context to the results. Hence, pathways visualisation tools and filters around the biological context were developed to improve the users' experience and help them extracting meaningful information. Also, PiMP was developed using a modular design; the previous section made use of it to extend the features with the integration of an internal plugin to analyse fragmentation data. The work presented here makes use of this modular design to integrate an external plugin to extend the data interpretation capabilities of PiMP by adding biological network analysis support (aim number 4).

4.5.1 Network reconstruction

A similar approach as the chemical library (section 4.3.2) was taken to develop this tool. Several steps are required to present the metabolic network of the system or organism studied to the user; the network also needs to be enriched with the metabolomics data present in PiMP. This multi-step task requires user interaction in order to select the appropriate model and sub-network to visualise; it was therefore developed as a semi-automated tool. Two types of information need therefore to be captured by the user, (i) the network model, (ii) the sub network parts divided in biological pathways. Figure 4.19 shows the form presented to the user for this purpose.

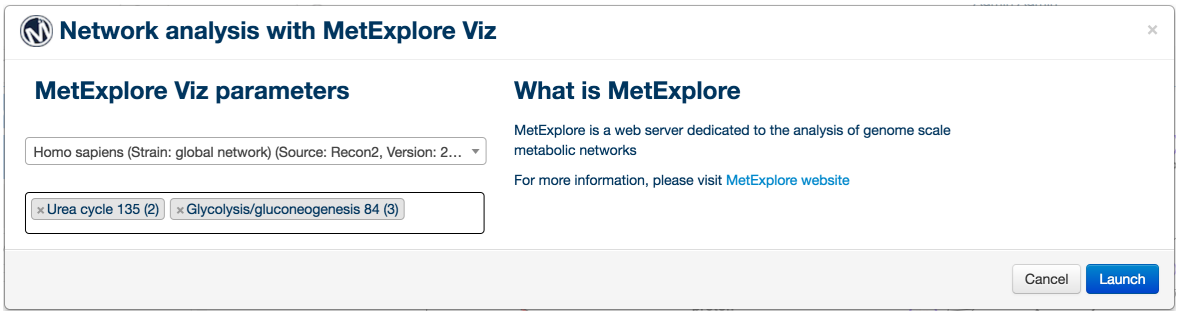

Figure 4.19: Dynamic Network analysis form allowing the user to select the organism and pathways to visualise.

MetExplore's metabolic networks database was chosen to reconstruct the network. The communication is, however, different than the protocol used for the chemical library as it requires user interaction. As PiMP provides InChIKey identifiers for all metabolites present in its database, the metabolite mapping onto MetExplore network was developed using these identifiers. Figure 4.20 shows the communication protocol between PiMP and MetExplore based on REST web service.

The first step is to provide the user with the list of different networks available, therefore, when the user requests a network analysis, PiMP sends a request to MetExplore web service to gather this information. The list of networks is returned to PiMP as a JSON array and presented to the user in a drop-down list (Figure 4.19). Once the user has selected the desired network, a new call is sent to MetExplore to retrieve information regarding the pathways available in this network. This second call to MetExplore web service is sent with the list of metabolites to map on the network, the data returned to PiMP is, therefore, a list of pathway present in the network with both the total number of metabolites present in the pathway and the number of metabolite mapped from the list sent by PiMP. The data sent and received by PiMP is also formatted as JSON arrays. This allows presenting to the user more accurate information to assist with the pathway selection as seen in Figure 4.19. Finally, once the selection of pathways has been performed, a"launch" button allows the user to start the network visualisation; however, a last call is required to reconstruct the network using the user

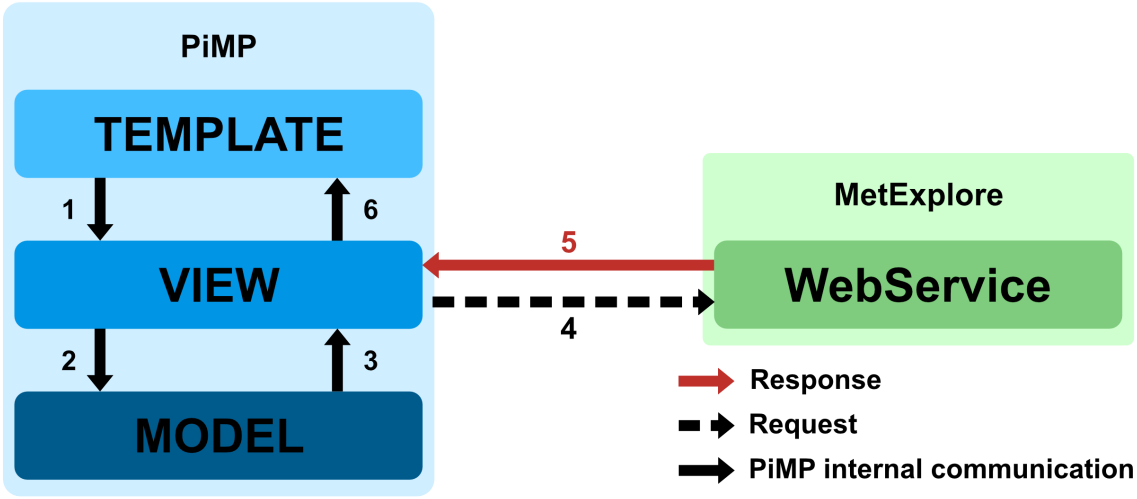

Figure 4.20: Communication between PiMP and MetExplore web service to dynamically perform a metabolite mapping on the model selecting and build the network requested. **1**. Request from the user is sent to the view. **2** and **3**. The view request the list metabolites to map from the model layer. **4**. The list is sent to MetExplore webservice with the selected organism and pathway. **5**. The reconstructed network is sent back to PiMP. **6**. After enrichment of the network with intensity values, the network is sent back to the template for display purposes.

selections. The request sent from PiMP to MetExplore contain the following information, (i) the BioSource id, (ii) the list of pathways, (iii) the list of InChIKey identifiers to map. MetExplore create the network using this information and return it to PiMP as a JSON structure. PiMP then add the intensity values for each mapped metabolite to the network by parsing the JSON structure. The resulting network contains all information necessary to be visualised and explored by the user; it is therefore sent to the PiMP template for visualisation purposes.

4.5.2 Network visualisation

Several software and computing libraries provide user interfaces to visualise, explore and mine biological networks. However, the library used needs to meet specific requirements to be integrated into PiMP. As PiMP uses the web browser to present the data to the user, the network visualisation library needs, therefore, to use web technologies to be integrated into PiMP data exploration environment. MetExploreViz was chosen for this purpose, MetExploreViz is a network visualisation plugin that is used and developed as part of MetExplore. It has been developed using D3.js javascript library [145] which is specifically designed to develop web-enabled data visualisation tool. MetExploreViz integrates network mining and comparison tools as well as most of the common features such as search and export features. As this library is specifically designed for biological network visualisation, it also includes bespoke features to support tasks that biologists may want to perform, visualising

pathways, cell compartments or duplicating highly connected nodes to simplify the network. In metabolic networks, highly connected nodes, also known as side compounds, are compounds that take part in many reactions and have little biological meaning such as water or CO^2. If displayed as one single node, it can create a hairball effect and increase the complexity of the network by generating biologically irrelevant paths between the different compounds of the network, making it difficult to interpret (Figure 4.21). The last benefit of using this javascript library is the input format. Indeed, as the network is reconstructed using MetExplore web service, it is already in the right format and can directly be loaded in the visualisation plugin. This avoids unnecessary network transformation and therefore limits the generation of errors that can happen during the translation of the network to another format.

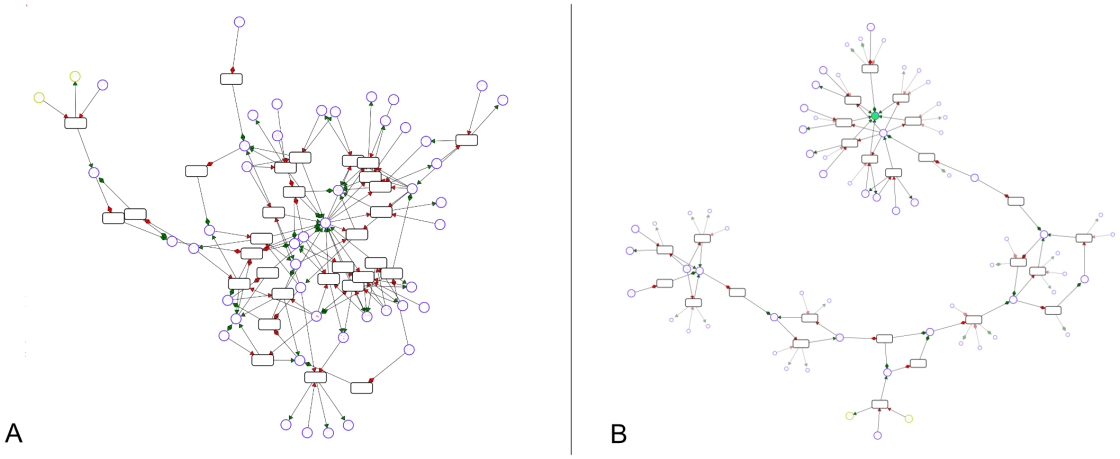

Figure 4.21: Visualisation of the same network before and after duplication of ”side compounds”. **A**. Original network. **B**. Same network after duplicating nodes considered as ”side compounds” to allow a better interpretation.

Once the user has performed all the tasks necessary to reconstruct the network (as explain in the previous section), PiMP template receive the signal from the view layer to start MetExploreViz and load the network. The communication between PiMP and MetExploreViz plugin is directly happening within the template using javascript. Figure 4.21 shows how MetExploreViz network visualisation is integrated into a new tab of PiMP data exploration environment. The user can perform new network reconstruction using the same interface as used initially (Figure 4.19) to load a new network.

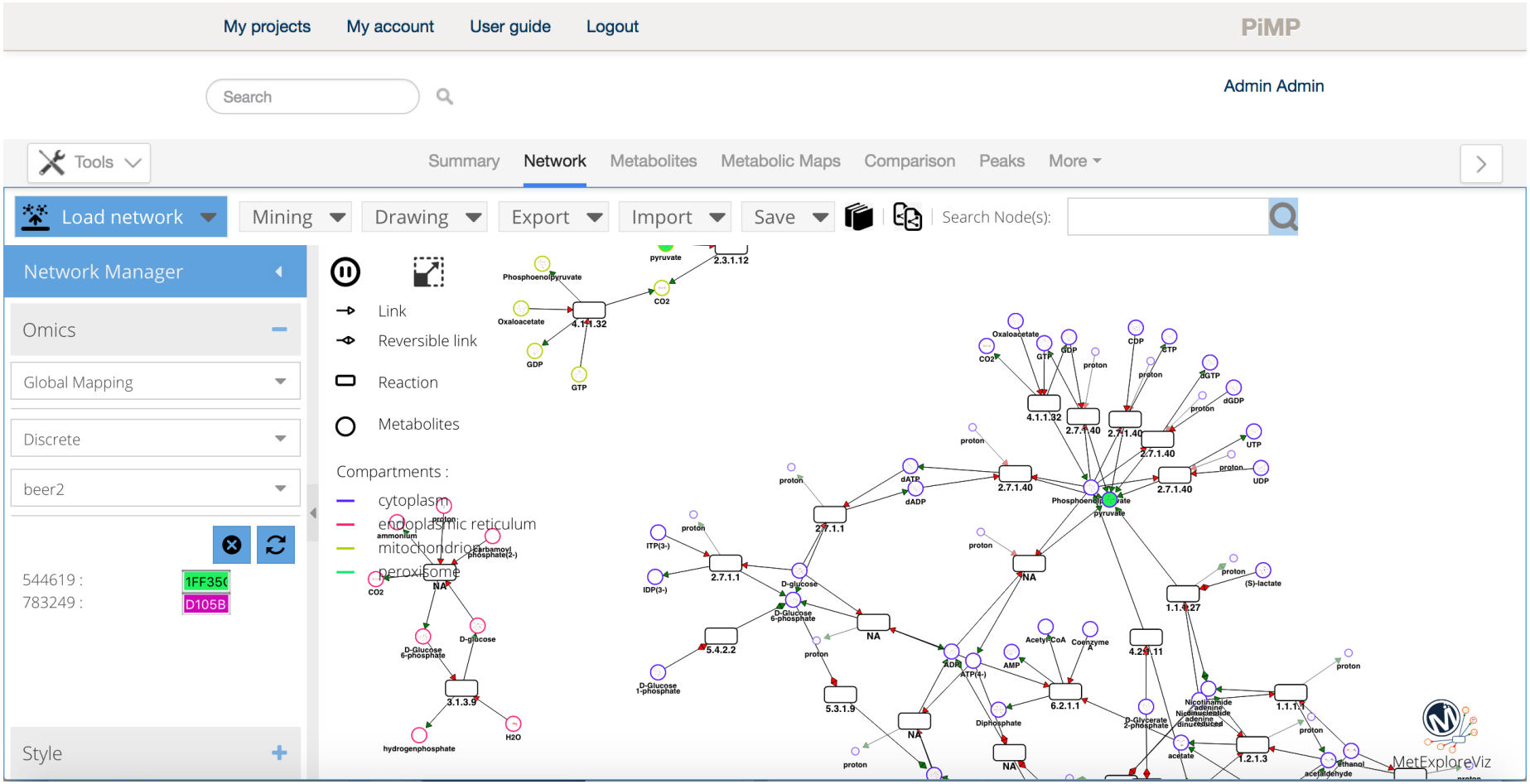

Figure 4.22: Network visualisation as seen in PiMP data exploration environment. A new tab called ”Network” is created on the fly, this new tab embed MetExploreViz javascript plugin and communicate with PiMP to display the reconstructed network.

4.6 Discussion

Chapter 3 exposed some of the major limitations met by untargeted metabolomics experiment with a primary focus on data analysis and interpretation. The proposed solution took the form of a tool that attempts to overcome the different challenges by assisting metabolomics users for each step of the workflow. However, some of these limitations can be better handled before analysis and many other, either unique to LCMS based untargeted metabolomics or more generally to omics technologies cannot find their answers in standard data analysis only.

Amongst these limitations, two can be generalised to many of the omics technologies. Study design remains one of the biggest challenges for biologists to overcome when using a technology that is still in its infancy. Whilst other omics involving sequencing, for instance, have made considerable progress on this part of the workflow due to their systematic use, emerging technologies such as LCMS metabolomics need tailored support to help biologist designing their experiments and understanding its potential outcome.

The second limitation concerns all omics technologies and could be extended to any emerging complex technology. The data capture and documentation of a study have become a real challenge as the technologies used in biological studies have become more complex. Indeed, many contributors are now taking part in the same study when any of the omics technology is involved, and relevant information is not always captured and followed properly along the entire process. Well documented studies are however crucial when it comes to reporting and publishing, and are the foundation of reproducible science. There is, therefore, a pressing need to support the scientists contributing to these complex studies in recording and sharing every information that needs to be reported for the comprehension and reproducibility of the study.

Untargeted LCMS metabolomics also face its own limitations, many of which have been addressed in Chapter 3 by better supporting data analysis. However, standard LCMS data analysis still faces challenges when it comes to unambiguous compound identification. This has major repercussions in the interpretation of the data produced by LCMS metabolomics and consequently on the impact of the field. Improving this part of the metabolomics workflow can, therefore, help to exploit better the potential that the field has to offer.

Finally, many advances in metabolic modelling can now contribute to improving the interpretation of LCMS data in a biological context. The challenge of interpreting metabolomics data was addressed in Chapter 3 by providing pathway information from public databases. However, data interpretation can be taken further by using biological network models as a support to extract meaningful information. Improving this area of the metabolomics workflow can also have a significant impact on all studies using the field to understand biological

systems.

The work presented in this chapter attempted to offer solutions to all these limitations in different manners. Either by providing new dedicated tools, by extending the Polyomics integrated Metabolomics Pipeline making use of its modular design, or by incorporating external tools in this same software.

The first objective which consisted on developing a tool allowing data capture of all important information during the course of metabolomics experiments has been met by the development of a web enabled management system as seen in section 4.3.1 of this chapter.

Many proprietary software offer Laboratory Information Management System (LIMS) to support the task of documenting research and laboratory's operations, however, very few freely available tools exist. In metabolomics, the existing tools are either very specific to the needs of a certain laboratory [146] or limited to one approach [147] (e.g. MetabolomExpress for GCMS). Although metabolomics databases allowing the reporting and publication of metabolomics studies have emerged in the past few years [135, 148, 149], they do not support scientist in capturing information at the time of the experiment. The project management system proposed in this chapter presents a flexible modular design allowing its extension to any omics technology, supporting the recording of information within the same environment by giving access to all contributors of a study to the same tool. The resulting software could even further benefit to biologist by integrating it into PiMP. This can be envisaged as a future work as they have been developed using the same modular technology and design. It would create a complete and unique environment for metabolomics studies documentation, analysis and interpretation. Another improvement that could be considered would be to automate the export of information captured to one of the metabolomics reporting databases; this would facilitate and shorten the process of reporting this information manually using dedicated software provided by those repositories.

Objective 2 was fulfilled by the creation of biochemical library presented in section 4.3.2 allowing the visualisation of a set of compounds in the context of specific metabolic network.

Whilst many facilities and laboratory across the world offer LCMS metabolomics as a service to analyse biological systems, no systematic methods has yet been developed to inform the user what part of the metabolome can be explored depending on the system studied. The automated tool informing on network coverage of a chemical library built by the association of a metabolic network database and the standard compound library run at Glasgow polyomics offer here a new way for biologists to better plan their experiments. They now have the possibility to know with precision the potential outcomes of a metabolomics study. As the necessary tools (web services) have now been developed to automate this task, this could be easily extended to any other laboratory.

Objectives 3 and 4 required the integration of two plugins to respectively enable the analysis

of fragmentation data alongside an untargeted analysis, and extend the PiMP interpretation capabilities to biological networks. These objectives were successfully met by the integration of two new tools to the PiMP platform as detailed in section 4.4 and 4.5

Whilst, as suggested as a further work, a unified and unique tool supporting the complete metabolomics workflow could have a transformative effect on the impact of metabolomics, this impact is still limited by LCMS metabolomics own challenges such as compound identification. Many laboratories have now gathered thousands of fragmentation spectra at different collision energy in the attempt to produce an exhaustive library collection that can be used as a reference to compare experimental spectra. Some tools provide web access through the web browser, others provide an API to automate the mass spectra matching process, but they are all however dedicated to the analysis of fragmentation data only. Some data analysis pipeline have now integrated fragmentation data analysis capabilities such as XCMS, but the analysis process and results still lacks the seamless integration that PiMP and FrAnK discussed in this chapter offer. However, while the tool proposed in this chapter has the advantage to be easy to use as part of the standard metabolomics data analysis workflow, only two external databases are currently used for spectra matching. The next step in the development of this tool would be then to extend the connections to other databases to increase the confidence in the annotation of the peaks. Connecting the tool to in-silico fragmentation tool such as ChemSpider could also be informative to the user when spectra do not present any match in any other database.

This improvement in compound identification brings more confidence in the interpretation of the data. However, the use of external tools is still required to interpret metabolomics data outside the conventional pathway analysis. Network models offer an excellent opportunity to find new potential paths and connections between metabolites that would be easily missed during a pathway analysis. The analysis of biological networks requires the use of specialised software that are not easy to use without prior training. Indeed, while software such as Cytoscape [139] or Tulip [138] provide valuable resources to extract meaningful information from biological networks, they imply that the user is capable of exporting his metabolomics data into a particular format allowing them to reconstruct a biological network, before loading it into the analysis tool. Other software such as MetExplore assists the user by providing the network and requiring as an input a simple list of metabolite with associated data. The tool presented in this chapter takes advantage of this feature in MetExplore and automate every manual step. The resulting tool brings network analysis to any scientist who can focus on the interpretation of the data without having to understand the backend file that describes the investigated network. This metabolic network approach can also offer a way to analyse several omics data simultaneously.

The overall objective of this chapter which consisted in better support certain steps of LCMS metabolomics workflow in a unified platform has been successfully achieved. The inte-

gration of the management system to the data analysis platform (PiMP) leaves, however, opportunities for improvement.

While the tools developed in the first two chapters have revealed themselves useful and have been used in published research work, a real case study is necessary to fully test the platform. The work presented in the next chapter will therefore attempt to better understand biological processes analysing omics data using the tools newly developed.

4.7 Conclusion

Study design, experiment documentation, data interpretation, all have a crucial role in any type of biological sciences. Omics technologies, from the study of genes to the understanding of small molecules of a biological system, share a great potential for understanding better the biology of all living things. But they also share a great complexity in the laboratory operations that they require, in the instrumentation they use and in the data they produced. Understanding a biological system as a whole cannot be achieved by simply repeating experiments and looking at different part of the system. Although trying to understand the various connections involved between the different "omics" layers shows great potential and promise to take the understanding of the interactions between biological molecules to a new level, it can only be enabled by improving the quality, the confidence and the reproducibility of each of these technologies on their own. A better documentation and reporting of the research can help towards that goal, but scientists need support to improve and systematise this process. Indeed, data repositories created for the purpose of thoroughly reporting studies are currently not used to their full potential. For example, since MetaboLights was launched in 2013, 216 studies have been uploaded to the repository and are fully described (as of January the 23rd 2017) , which leaves thousands of metabolomics studies published and simply described in journal articles with no properly formatted and organised documentation. Improving the impact of omics technologies has to go through the process of developing the technologies themselves and the way they are exploited, providing better study documentation to allow control studies and reuse of generated data, better supporting study design and data interpretation to extract meaningful biological insight. This only, can lead to considering every omics layers as one complex system and unleash the full potential of omics approaches.

Chapter 5

Integrative analysis of omics datasets using a network approach

5.1 Introduction

Networks are used in many scientific and non-scientific fields to address a variety of problems. They can also be used to capture the knowledge of a system at a certain point in time. A network representation of a system can be very powerful to understand or highlight the existing connections, sometimes complex, between the different components constitutive of a system. Networks can be used in areas from urban traffic flow management to the study of social interactions. In biology, and more specifically in omics technologies, networks are used for both problems solving and knowledge capture and dissemination. Each omics layer can indeed be represented by a list of components (nodes) connected with each other (edges) forming a network. The nodes generally representing biological components and edges the relationships they have with one another. The omics technologies based on sequencing, genomics and transcriptomics, which are the most established fields of the omics family, have now long taken advantage of network representations to interpret the generated data. Gene interaction networks, gene co-expression networks or gene regulatory networks are some of the most used network representations of genomics and transcriptomics data. Network approaches are also often used to study Proteomics data, protein-protein interaction networks are commonly used for different purposes such as functional module identification. In metabolomics, networks are used for diverse applications, some have been commonly used for years such as metabolic networks for flux balance analyses, others have just been developed such as substructure networks for untargeted metabolomics data exploration [150]. As seen in chapter 4, metabolic networks can also be used in metabolomics to help with data interpretation as they provide an alternative approach from the common pathway analysis. However, the metabolic networks used in this approach also hold information about reac-

tions happening in the biological system; reactions that can be directly or indirectly related to quantitative proteomics or gene expression data. Integrating these types of data together with metabolomics data creating a multi-layer network has the potential to highlight cell processes that cannot be seen in one omics alone. The increased complexity of such integrated network may, however, present a barrier to its interpretation.

The development of an integrative method required the use of a test dataset. In this study, we apply the methodology to attempt understanding the effect of the activation of nicotinic receptors on signalling pathways and the metabolism.

The contribution of nicotine to cancer incipience and growth is subject to thorough investigation by a vast research community [151]. The nicotinic acetylcholine receptors, activated by nicotine, can trigger tumorigenic effects by the activation of signalling pathways [152]. The oral keratinocyte alpha 7 nicotinic receptors, which choline is a selective agonist [153], has been of particular interest to attempt understanding the signaling pathways connected to cancer onset and nicotine carcinogenic mechanisms. These signaling pathways are however still poorly understood, and their comprehension could make them a target for cancer therapy or prevention. The experimental design of this study is detailed in the following sections.

5.2 Related work

Omics technologies are now routinely used for the investigation of biological systems in many biology related fields. From clinical to plant studies [154, 155], the understanding of a system often requires an in-depth examination of the components that it is made of, their states and their relationship with one another. Omics technologies offer the opportunity to observe and annotate genes directly, measure their expression levels, study proteins and their functions in a system, take a snapshot of a metabolism state or measure their fluxes to understand the metabolites flow of a system. These individual applications are however often performed in isolation from the others, and only allow the assessment of one particular part of a system. The combination of these applications could offer the possibility to understand better the relationships that exist between all the components that make a biological system. The previous chapters presented the development of new tools and methods to help biologists in designing, analysing and interpreting metabolomics experiments with an attempt to extend the context of interpretation to the whole system by using genome-scale metabolic models. The work presented in this chapter is a direct continuation of the previous chapter: an attempt to combine metabolomics data to gene expression data to further investigate metabolomics data analysis results in a holistic approach. Many attempts of integrating multi-omics data using different approaches have been made in the past decade with varying degrees of success [156, 25]. None of these studies, however, approach the

subject of automating the integration process. Gomez-Cabrero et al. [157] present an interesting view of the current and future challenges encountered in data integration within the omics context such as data heterogeneity due to the use of many different standards in life sciences. B. Palsson, K. Zengler also raise challenges faced to interpret integrated omics data in a biological context [158]. Overcoming these challenges is essential to generalise omics integration methods and therefore generalise system-wide studies to enhance current omics-related research.

The work presented in this chapter will attempt to answer the following questions:

- Can metabolomics and RNA-seq data be integrated using metabolic network model?
- Can the integration of metabolomics and RNA-seq data be automated?
- Does the interpretation of the integrated network bring added information to the interpretation of metabolomics data on its own?

Four aims were drawn from these research questions; it is implied that metabolomics and RNA-seq data are acquired from the same samples and experiment:

1. Analyse and interpret untargeted metabolomics data using the tools previously developed.
   Objective 1: Analyse and interpret LCMS metabolomics data using exclusively the PiMP platform to highlight metabolic processes involved in the case study.

2. Analyse RNA-seq data and prepare its integration.
   Objective 2: Analyse RNA-seq data and format results for the mapping of genes onto human metabolic networks.

3. Integrate RNA-seq and metabolomics data together.
   Objective 3: Reconstruct a human metabolic network mapping both metabolomics and RNA-seq datasets.

4. Interpret the reconstructed integrated metabolic network
   Objective 4: Attempt to derive new biological insight from the reconstructed network.

5.3 Study design

To develop, assess and validate the network approach presented in this chapter, RNA-seq and metabolomics data of human keratinocytes were acquired. This section describes the study and its objectives.

The oral keratinocyte alpha 7 nicotinic receptor (α7nAChR) has been implicated to play a role in the pathogenesis of oral squamous cell carcinoma (OSCC) and periodontal disease (PD). OSCC is a malignant tumour of the oral epithelium, and over 5000 new cases are diagnosed each year in the UK only. The α7nAChR has been suggested to mediate nicotine-induced abnormal keratinocyte cell cycle progression leading to squamatization and OSCC. PD is a microbially induced chronic inflammatory disease of the oral cavity. Nicotine acting via the α7nAChR mediated "cholinergic anti-inflammatory pathway" is suggested to enhance smokers susceptibility to PD by suppressing oral immune responses resulting in the persistence of oral pathogens and chronic inflammation.

These findings, however, are based on studies applying a reductionist approach investigating isolated signalling pathways, the complete picture of α7nAChR mediated signalling pathways in oral keratinocytes remains unknown.

For the purpose of this study, primary human keratinocytes have been cultured *in vitro* and stimulated with a specific α7nAChR agonist (PHA 543613 HCl). Cells have been stimulated for 24 hours with data collection at three different time point, 4 hours, 9 hours and 24 hours. Unstimulated cells are used as a control with data collection at the same time points. The experiment was conducted in triplicate and samples were prepared for metabolomics and RNA-seq experiments.

5.4 Data acquisition

The samples were prepared by laboratory scientists at Glasgow university dental school, transcriptomics and metabolomics data was then acquired by Glasgow Polyomics. The author was not involved in the data acquisition process, his role was limited to the data analysis.

Primary human keratinocytes were cultured *in vitro* and stimulated with a specific α7nAChR agonist (PHA 543613 HCl). Cells were stimulated for 4, 9 and 24 hours. Unstimulated cells acted as a control. The experiment was repeated three times with cells from three different donors.

At each time point, RNA was harvested using the RNeasy kit (Qiagen, UK). Ribosomal RNA was depleted using RiboMinusTM technology (Life Technologies, UK). The samples were prepared following Illumina standard protocol using the Illumina TruSeq Stranded Total RNA kit. The RNA- seq data was acquired using paired-end RNA sequencing on Illumina NextSeq 500.

Similarly, metabolites were extracted at a ratio of 1:3:1 chloroform:methanol:water.

Hydrophilic interaction liquid chromatography (HILIC) was carried out on a Dionex UltiMate 3000 RSLC system (Thermo Fisher Scientific, Hemel Hempstead, UK) using a ZIC-

pHILIC column (150 mm 4.6 mm, 5 μm column, Merck Sequant)

The column was maintained at 30°C and samples were eluted with a linear gradient (20 mM ammonium carbonate in water, A and acetonitrile, B) over 24 min at a flow rate of 0.3 ml/min as follows:

Time / minutes	%A	%B
0	20	80
15	80	20
15	95	5
17	95	5
17	20	80
24	20	80

Table 5.1: Table describing the elution gradient used for Liquid Chromatography. A = 20 mM ammonium carbonate in water. B = acetonitrile.

The injection volume was 10 μl and samples were maintained at 5°C prior to injection. For the MS analysis, a Thermo Orbitrap QExactive (Thermo Fisher Scientific) was operated in polarity switching mode and the MS settings were as follows:

- Resolution 70,000
- AGC 1×10^6
- m/z range 70 – 1050
- Sheath gas 40
- Auxiliary gas 5
- Sweep gas 1
- Probe temperature 150°C
- Capillary temperature 320°C

For positive mode ionisation: source voltage +3.8 kV, S-Lens RF Level 30.00, S-Lens Voltage -25.00 (V), Skimmer Voltage -15.00 (V), Inject Flatopole Offset -8.00 (V), Bent Flatapole DC -6.00 (V). For negative mode ionisation: source voltage-3.8 kV.

The calibration mass range was extended to cover small metabolites by inclusion of low-mass calibrants with the standard Thermo calmix masses (below *m/z* 138), butylamine ($C_4H_{11}N_1$) for positive ion electrospray ionisation (PIESI) mode (*m/z* 74.096426) and CoF_3 for negative

ion electospray ionisation (NIESI) mode (*m/z* 84.9906726). To enhance calibration stability, lock-mass correction was also applied to each analytical run shown below.

Positive Mode Lock masses:

Number of Lock Masses: 3

Lock Mass #1 (*m/z*): 83.0604

Lock Mass #2 (*m/z*): 149.0233

Lock Mass #3 (*m/z*): 445.1200

Negative Mode Lock masses:

Number of Lock Masses: 1

Lock Mass #1 (*m/z*): 89.0244

5.5 Metabolomics data analysis

As outlined in the previous sections, the experiment contains six experimental biological groups. For quality control purposes, two more groups are routinely run during metabolomics data acquisition, these are the blank samples which correspond to the extraction solvent and allows the identification of contaminants, and the pooled samples to assess the quality of the instrumentation. As this a pilot study, only three replicate samples were acquired for each group. The six experimental groups are presented in the table below:

Condition / Time point	4 hours	9 hours	24 hours
Control	C4	C9	C24
PHA-stimulated	PHA4	PHA9	PHA24

Table 5.2: Table describing each of the experimental groups analysed in the study. Each sample has been collected in 3 biological replicates.

5.5.1 Quality control

Once the data acquisition is performed, the running of the instrument can be assessed using the pooled samples. Pooled samples are acquired every 5th sample throughout the instrument run and therefore representative of the stability and reproducibility of the instrumentation. As shown in Figure 5.1, the instrument shows high reproducibility over time, and no issue can be detected with the instrumentation.

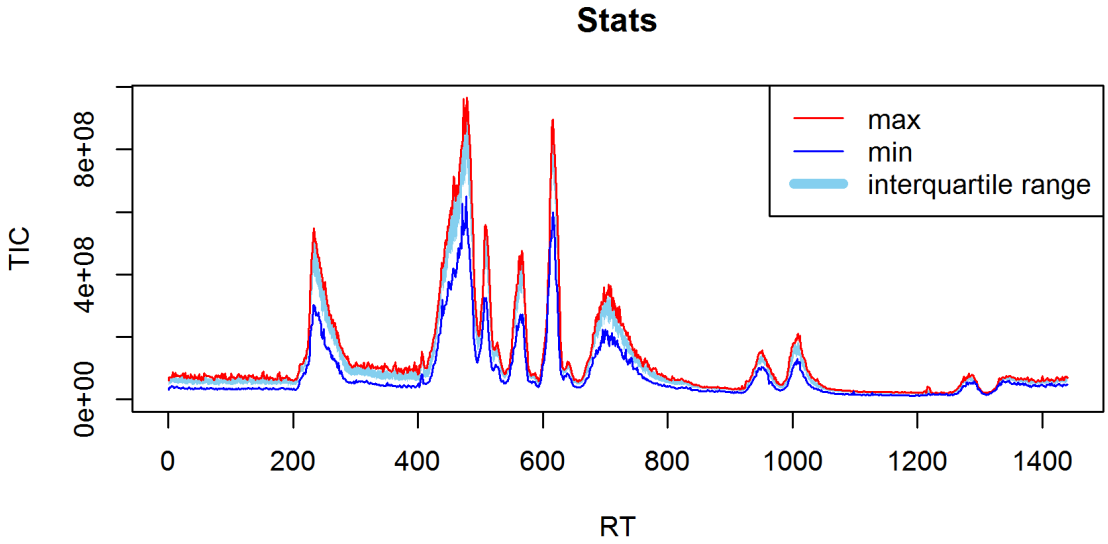

Figure 5.1: Minimum/maximum and interquartile range of the TIC signals for the positive mode pooled samples.

The analysis was performed using PiMP standard untargeted pipeline without fragmentation data. All results and figures presented in this section were generated by the PiMP software. Parameters used for the data processing are as follow:

- Ppm window: 3.0
- Retention time window: 5%
- Rsd filter: 0.80
- Minimum intensity: 5000
- Minimum detection number: 3
- Alignment: CowCoda
- Noise filter: 0.80

A total of 3 592 signals were identified as likely a metabolite. Figure 5.2 shows the unsupervised clustering of experimental samples performed using Principal Component Analysis. As seen in the figure, no clear cluster can be identified from the plot which indicates a high variability between samples within the same biological group.

Seven pairwise comparisons were performed to assess the differences that the metabolites present in each of the biological groups. To evaluate the evolution of the metabolites over time, time points were compared to one another within the same biological class (Control

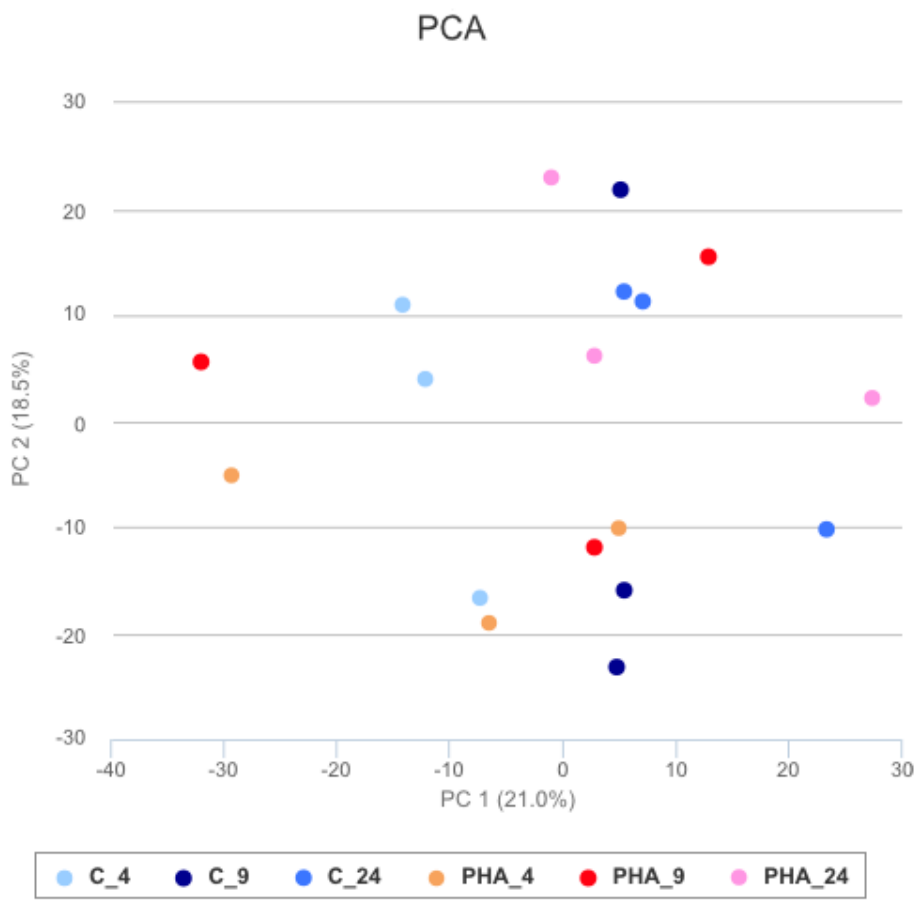

Figure 5.2: Plot of the first two principal components calculated for the experimental groups.

and PHA-activated). Control samples at 4 hours were therefore compared to control samples at 9 and 24 hours. Similarly, PHA-activated samples at 4 hours were compared to PHA activated samples at 9 and 24 hours. Finally, three comparisons were performed across the biological classes for each time point (i.e. Control 4 hours versus PHA-activated 4 hours). Out of 3 592 signals, 58 matched to authentic standards. However, a total of 5 of these identified metabolites were found to be significantly different in a minimum of one comparison (adjusted p-value under 0.05). Table 5.3 shows the list of 5 metabolites with their respective fold changes values ($\log^2$ transformed), the full list of detected standard metabolites is available in Appendix B.1.

The comparisons performed allowed the analysis of the data in two manners. First, a time course analysis of the separate experimental classes (Control and PHA-activated) allows the detection of metabolites changing over time. The trends of these changes can then be compared between the two classes. The analysis of the differences between the two experimental classes at each time point can also be performed to have a global view of the differences induced by the PHA treatment.

Name	Formula	C9/C4	C24/C4	PHA9/PHA4	PHA24/PHA4	PHA4/C4	PHA9/C9	PHA24/C24
Choline phosphate	C5H14NO4P	-0.06	0.09	-0.31	-0.7	-2.67	-2.92	-3.45
Guanine	C5H5N5O	0.94	3.69	1.41	3.16	0.65	1.12	0.11
L-Alanine	C3H7NO2	0.57	1.27	0.24	1.28	0.07	-0.25	0.09
L-Ornithine	C5H12N2O2	1.39	1.84	0.57	1.45	0.07	-0.75	-0.32
Nicotinamide	C6H6N2O	-0.25	-2.58	-0.54	-2.89	0.5	0.2	0.19

Table 5.3: List of standard compounds changing in at least one comparison with an adjusted p-value under 0.05. Significant changes are highlighted in green when a positive change occurs, in red when the changes are negative. Fold changes are $\log^2$ transformed.

5.5.2 Time course analysis

Control class

No significantly changing annotated peaks were found between the first two time points of the control group (Control 4 hours and control 9 hours). Therefore, the analysis focused on the changes appearing between 4 hours and 24 hours. 56 features were found to be significantly changing over time in the control class. Only 30 features were kept for further analysis as 4 were identified against a standard peak and 26 were putatively identified. The volcano plot in Figure 5.3 shows the significance versus fold change of these peaks in y and x-axes respectively.

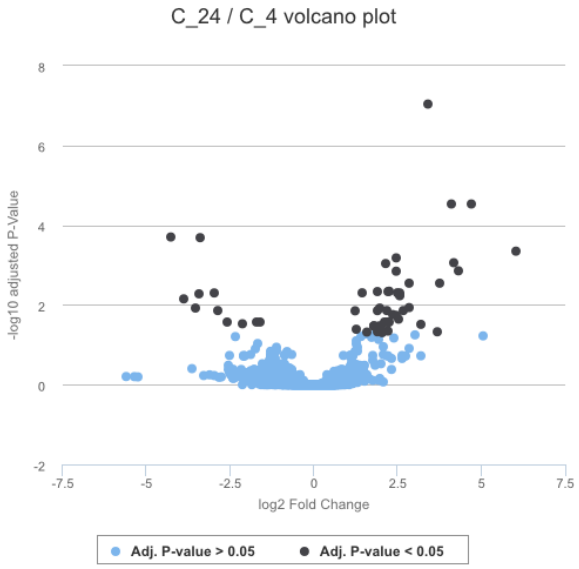

Figure 5.3: Volcano plot showing the changes of the detected peaks in the Control group at 4 and 24 hours. In y axis is the significance, x axis plots the fold change. Significantly changing features are coloured in black.

As mentioned in chapter 3, peaks need to be manually checked as the peak picking algorithm can sometimes consider noise signal as peaks. The peaks were also compared to the blank samples to remove contaminants from the list. These tasks were performed directly in PiMP using the dedicated tools. After filtering, only the four standard matching peaks and one annotated peak were kept. The annotated peak was found to be a lipid with the following chemical formula $C_{51}H_{74}O_2$. The chemical formula was derived from the *m/z* recorded for the base peak and is considered level 2 annotation according to the MSI. The four standard compounds were not used for pathway interpretation as none of them takes part in the same pathway, however, changes in the transcriptome could help further inform and interpret these changes.

PHA activated class

The same process was applied to the PHA-activated group. 18 peaks were found to be significantly different between 4 and 9 hours, 48 between 4 and 24 hours, 3 out of the 48 peaks matched against a standard (Table 5.3). The volcano plots showing these results is presented in Figure 5.4. After filtering to account for contamination, non-annotated peaks and artefacts, only the three standard compounds were kept.

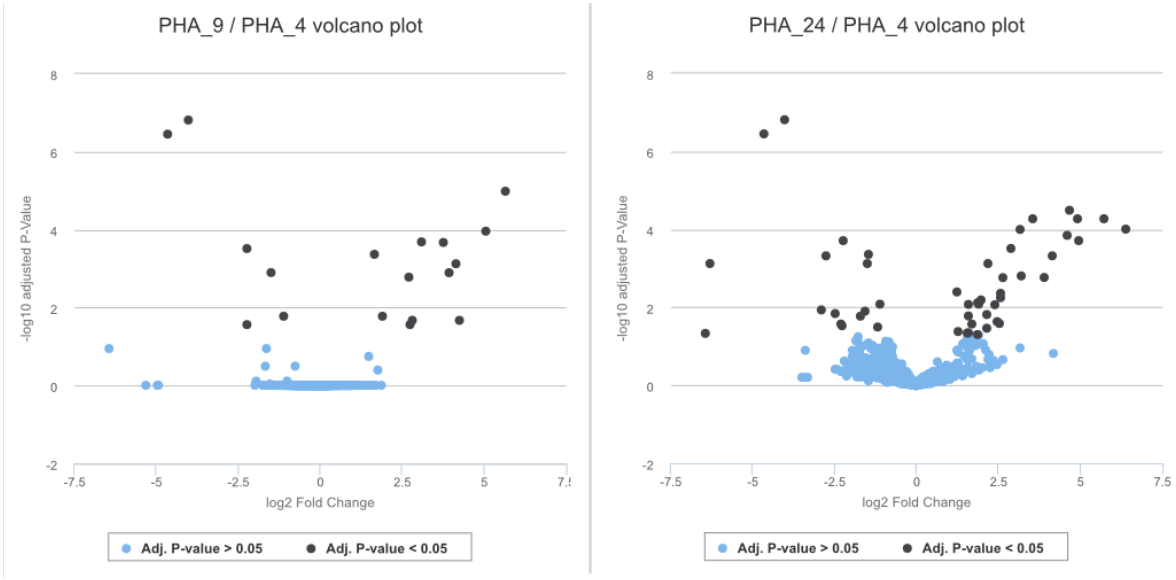

Figure 5.4: Volcano plots showing an overview of the features changing in the pairwise comparisons of the time points in PHA-activated experimental group. It is import to note that while the majority of the changes are not significant, many differences are observed.

Time course comparison

Very few metabolites revealed to be statistically changing over time in both experimental classes, which may indicate that the effect of the drug is well controlled or part of a signalling pathway that has limited effect on the metabolome. Three identified metabolites (L-Alanine, L-Ornithine and Nicotinamide) are common to the two groups; a trend comparison revealed that they change the same way over time with comparable intensities as seen in Figure 5.5. Guanine, which was found to be significantly changing over time in the control group only follow the same pattern, Figure 5.5 shows a higher standard deviation in PHA-activated time points, hence an adjusted p-value over 0.05.

5.5.3 Biological class analysis

The analysis of the time course of the two experimental classes did not reveal any difference in metabolism between the control and PHA-activated groups in the changes of the metabo-

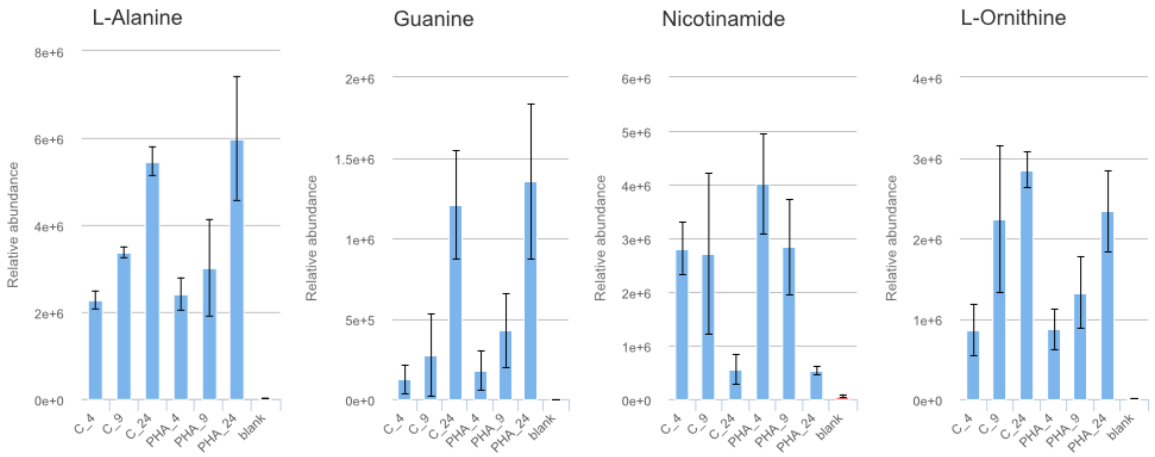

Figure 5.5: Average intensities of the 4 identified metabolites significantly changing over time.

lites levels. A more global approach comparing the two groups directly against each other may, however, reveal differences between the groups across every time points. Each time point of the control group was therefore compared to the corresponding time point of the PHA-activated group.

In total, 59 peaks were significantly different between the control and PHA-activated groups for time point 4 hours, 53 at 9 hours and 63 at 24 hours. Only one compound for which a standard was available: choline phosphate (ChoP), was found to be significantly different across all time points. Volcano plots corresponding to each of the three comparisons are shown in Figure 5.6. The same filtering process as the time course analysis was applied to these peaks and metabolites common to all time points were kept for further investigation.

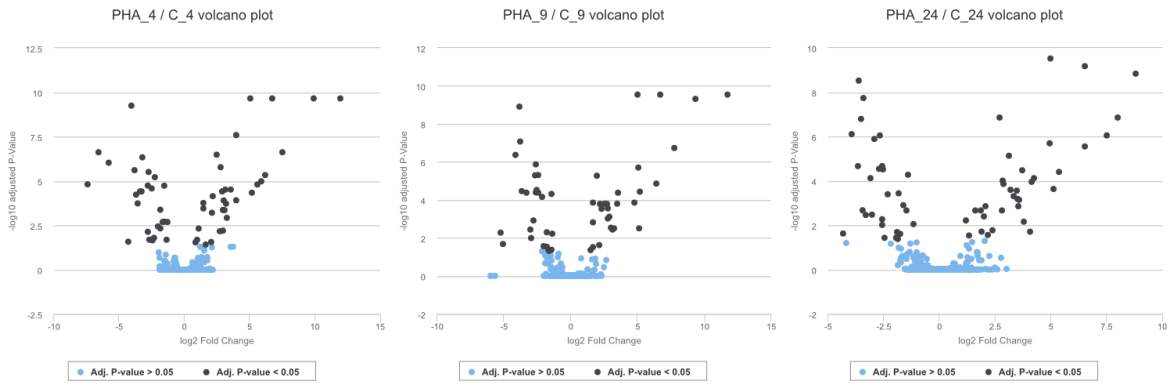

Figure 5.6: Volcano plots showing the differences found between the control and PHA-activated group at each respective time point.

After filtering process, only one annotated compound was found to be relevant. This compound is found in one known metabolic pathway (2-C-Methyl-D-erythritol 4-phosphate in Terpenoid backbone biosynthesis), which however does not allow a pathway analysis. An *in-*

silico analysis of potential reactions connecting this metabolite to choline phosphate could, however, be an avenue to explore. The differences in the levels of the annotated and identified metabolites in the two groups are shown in Figure 5.7.

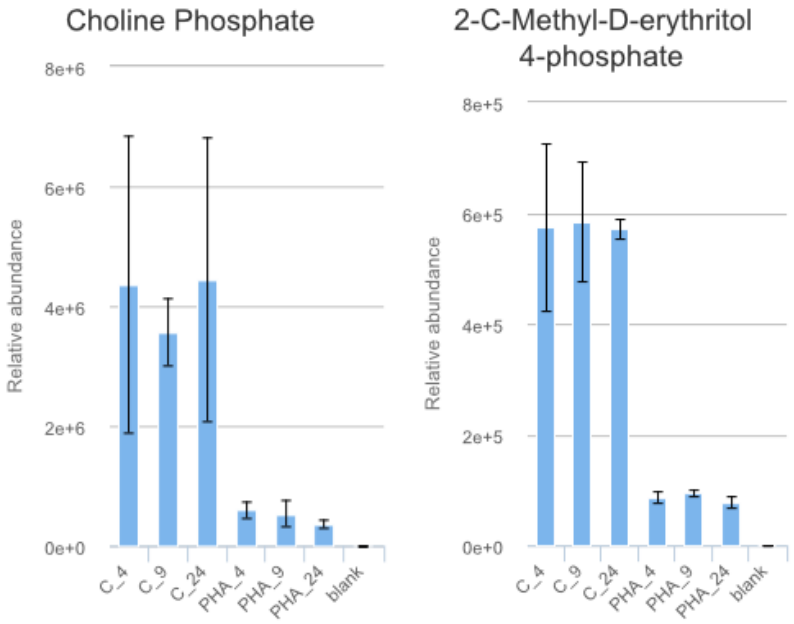

Figure 5.7: Average intensities of the two metabolites significantly different in the two biological groups.

5.5.4 Standard compounds analysis

As mentioned in section 5.5.1 of this chapter, 58 peaks matched against authentic standards. While only five were statistically significant across all comparisons performed, analysing these identified compounds may reveal differences in the experimental groups that were missed due to low statistical power. Each of the metabolites was assessed individually. The vast majority of the peaks can be considered of high quality with no time shift and high reproducibility (shown in Figure 5.8). The metabolites show small differences between the two experimental groups. Although the results were not significant, they could be indicative of a trend and repeating the experiment with more replicates could improve this analysis.

5.6 RNA-seq data analysis

The primary objective of the work presented in this chapter is to expand the interpretation of metabolomics data analysis results by connecting other omics data to further inform about the biological context. The RNA-seq data analysis presented here, therefore, describes the

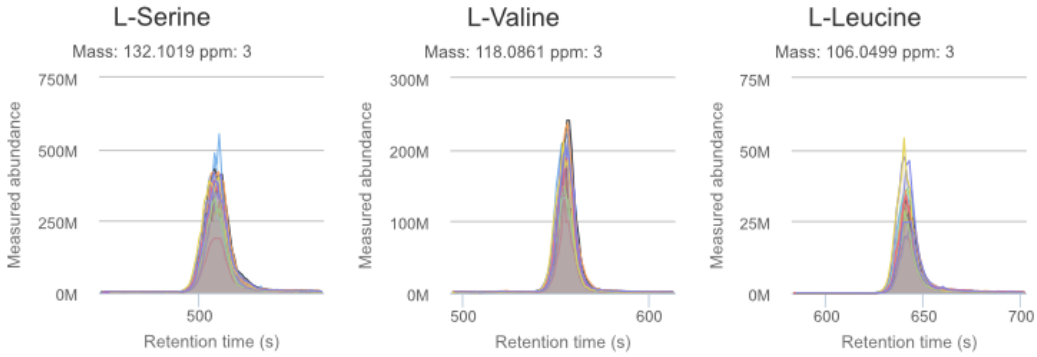

Figure 5.8: Extracted ion chromatogram of three peaks corresponding to standard metabolites (amino acids), showing high quality and reproducibility.

sequential steps performed to extract meaningful data that can potentially be integrated with metabolomics data. The interpretation of the RNA-seq analysis output on its own is not part of this work and therefore not described here. The experimental work was performed the Genomics unit of Glasgow Polyomics, the complete data analysis presented in the next section was carried out by the author.

5.6.1 Data acquisition and analysis pipeline

The samples were prepared using the Illumina TruSeq Stranded Total RNA kit. The RNA-seq data was acquired using paired-end RNA sequencing on Illumina NextSeq 500.

The data analysis consist in a four steps pipeline. First, the adapters were removed from the reads using Cutadapt [159]. Low-quality bases were then trimmed using Sickle [160]. The reads were aligned against the human reference genome GRCh38 using HISAT [161]. The differential expression analysis was then performed using CuffDiff [162].

A quality control step was conducted to assess the quality of the identification of the nucleobases generated by the sequencing. This quality control was done using Phred quality score of a base call which estimates the probability of error and was described by Cock et al. in 2010 [163]. The base call accuracy was found to be higher than 99.9% on average as shown in Figure 5.9.

The same seven comparisons as the metabolomics data analysis were performed resulting in 378 genes differentially expressed in at least one of the comparisons. The data presented below was extracted from the output generated by CuffDiff using script developed in Python.

A time course analysis was performed separately on the two biological classes. The control group showed 204 genes significantly differentially expressed while the PHA-activated group had 233 genes differentially expressed over time.

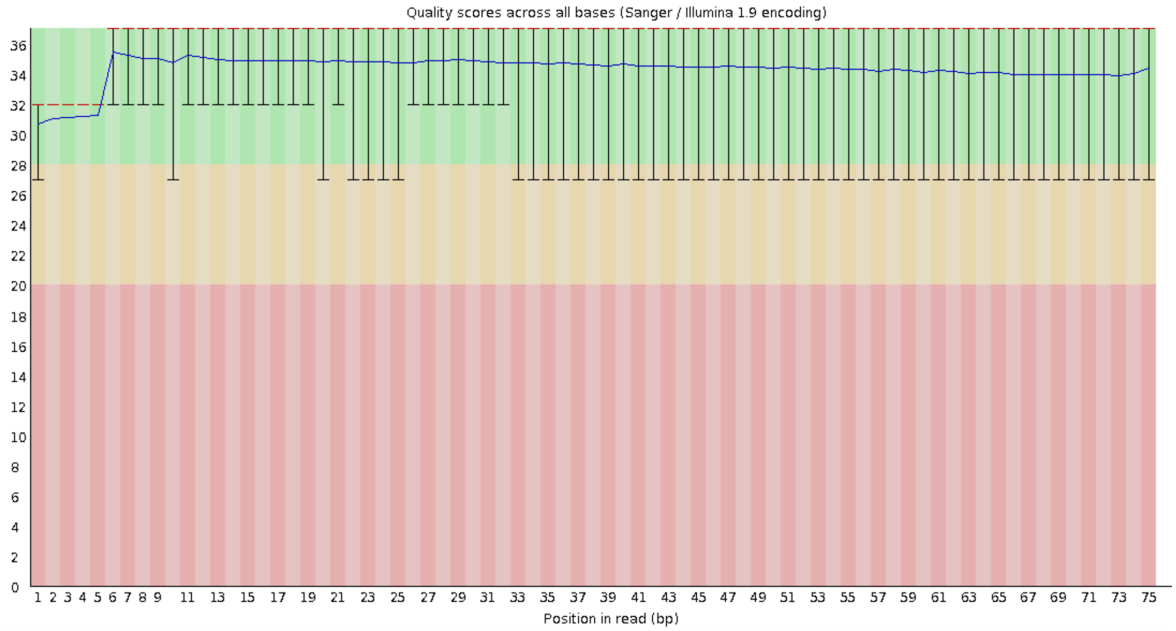

Figure 5.9: Plot showing the average base call accuracy of the reads. It plots the Phred score versus the read position in y and x axes respectively.

The comparison of the respective time points across the two biological groups revealed fewer differences, 25 genes are differentially expressed at 4 hours, 13 genes at 9 hours and 29 at 24 hours.

5.6.2 Gene networks

As the aim of this work is to connect metabolomics data to gene expression data through metabolic network reconstruction, the data available in the reconstructed network could be limited to genes having a role in metabolic regulation. However, many associations and interactions exist between genes with no direct impact on the metabolome. The reconstruction of gene interaction and association networks before building the metabolic network can extend the information available in the final network by adding an extra gene-specific layer. Seven types of gene interaction or association networks were created using GeneMANIA [164] to create this additional layer which could then be plugged into the metabolic network. The Cytoscape app of GeneMANIA which searches through publicly available datasets and databases to create different networks was used for this task. The network was reconstructed using the following interactions or associations:

1. Co-expression
2. Physical Interaction
3. Genetic interaction

4. Shared protein domains
5. Co-localization
6. Pathway
7. Predicted functional relationship

As shown in Figure 5.10, the reconstructed network when visualised as is suffers from "hair-ball effect" which makes it hard to interpret. However, this network is, in fact, a combination of seven different networks represented by different edge colours. Each of these networks can be turned on and off to allow the visualisation of a single or a subset of networks. This represents a vast amount of organised data that can potentially be added to the multi-omics metabolic network as "meta-data" once reconstructed. The network presented in Figure 5.10 also shows log fold changes values as colour scale applied to the nodes.

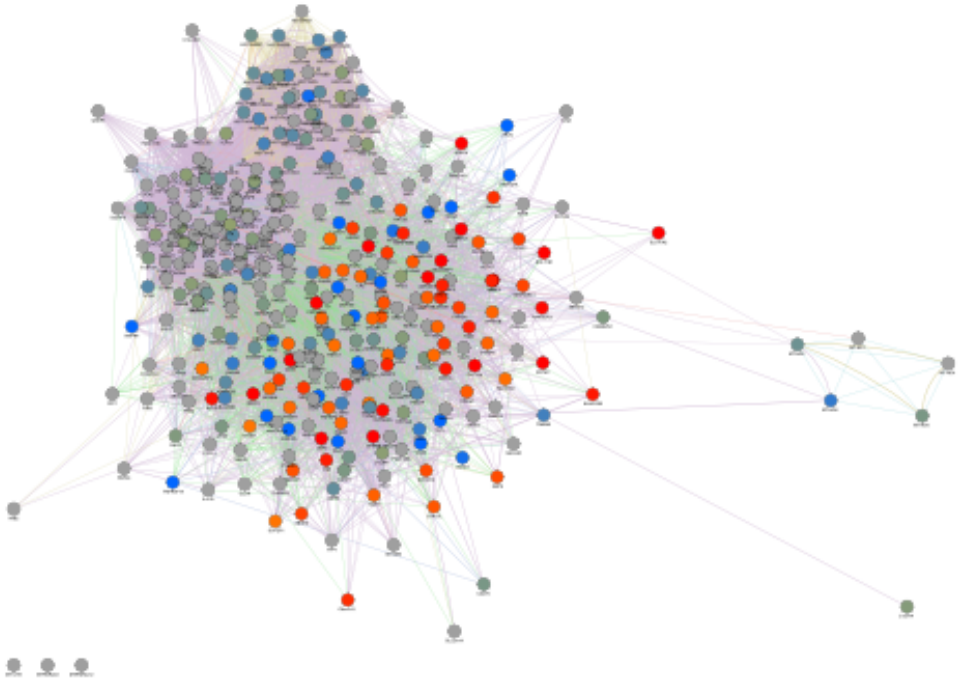

Figure 5.10: Reconstructed and integrated gene interaction and association network. Node colour is based on the log fold changes of the comparison of Control 4 hours against Control 24 hours, no significance threshold was used in this example.

5.7 Integrative analysis

The previous sections of the chapter presented the analysis of two different omics data of the same samples. It was shown that greater changes are detected at the gene expression level than the metabolomics level. However, while very few significant changes can be observed on the metabolomics level, the investigation of standard metabolites revealed many subtle changes. Repeating the experiment with a longer time course and more replicates could give more significant results on the metabolomics level.

The work presented in this section attempt to answer the following research question: "Can metabolomics data be further interpreted when integrated with other omics?"

The subset of metabolomics data selected for this section are the significantly changing metabolites presented in section 5.5.3, and the list of standard metabolites which present subtle changes from which no conclusion can be drawn.

The approach taken to connect metabolomics and gene expression data is based on genome-scale reconstructed metabolic networks. This is considered as the direct extension of the work presented in Chapter 4 on network analysis of metabolomics data.

5.7.1 Multi omics network reconstruction

The network reconstruction approach taken here consists of three main steps; (i) network building from gene expression data, (ii) mapping of metabolomics data on the network, (iii) integrated network interpretation.

The model chosen for this approach is the human genome-scale metabolic reconstruction Recon2.2 [106]. MetExplore, tool presented in previous chapters and partly integrated into PiMP was chosen to build and investigate the network.

Gene mapping

One of the requirement to create a network from gene expression data is to use the same gene identifiers. MetExplore (Recon2.2) uses Entrez gene ids to describe the network, while the output of the RNA-seq data analysis was given using Ensembl identifiers. The identifiers were, therefore, converted from Ensembl to Entrez using DAVID conversion tool [165]. It was found that one Ensembl gene ids [166] can sometimes correspond to several Entrez gene ids [167]. The Entrez gene ids derived from one Ensembl id correspond however to one gene only, they were, therefore, all kept; the list of 378 differentially expressed was then described by a list of 439 Entrez gene ids. 38 genes related to 125 metabolic reactions could be matched to Recon2 model using MetExplore.

The network was reconstructed from this list of reaction by keeping only metabolites directly involved in, at least, one reaction as a substrate or product. This first step of network creation allowed to filter down the list of metabolites present in the model from 5063 to 289. The number of pathways represented in this network is of 41. A pathway is considered represented if one of the reactions present in the network is also part of the pathway, this means that the 41 pathways present generally a very low coverage. This can, however, be indicative of a very controlled effect of the drug. While knowing which pathway is involved in the network is not a specific objective of the network reconstruction, it gives an overview of the different part of the metabolism potentially affected by the differential expression of the detected genes. This list of pathways is available in Appendix C.1.

Figure 5.11 shows the network obtained after application of the filters based on genes previously mentioned. Some highly connected nodes (metabolites) can easily be detected and will need to be duplicated as side compounds for better visualisation. As seen in the network, some reactions are disconnected from the network; it was decided to keep them for the metabolite matching step as they can potentially reveal relevant information.

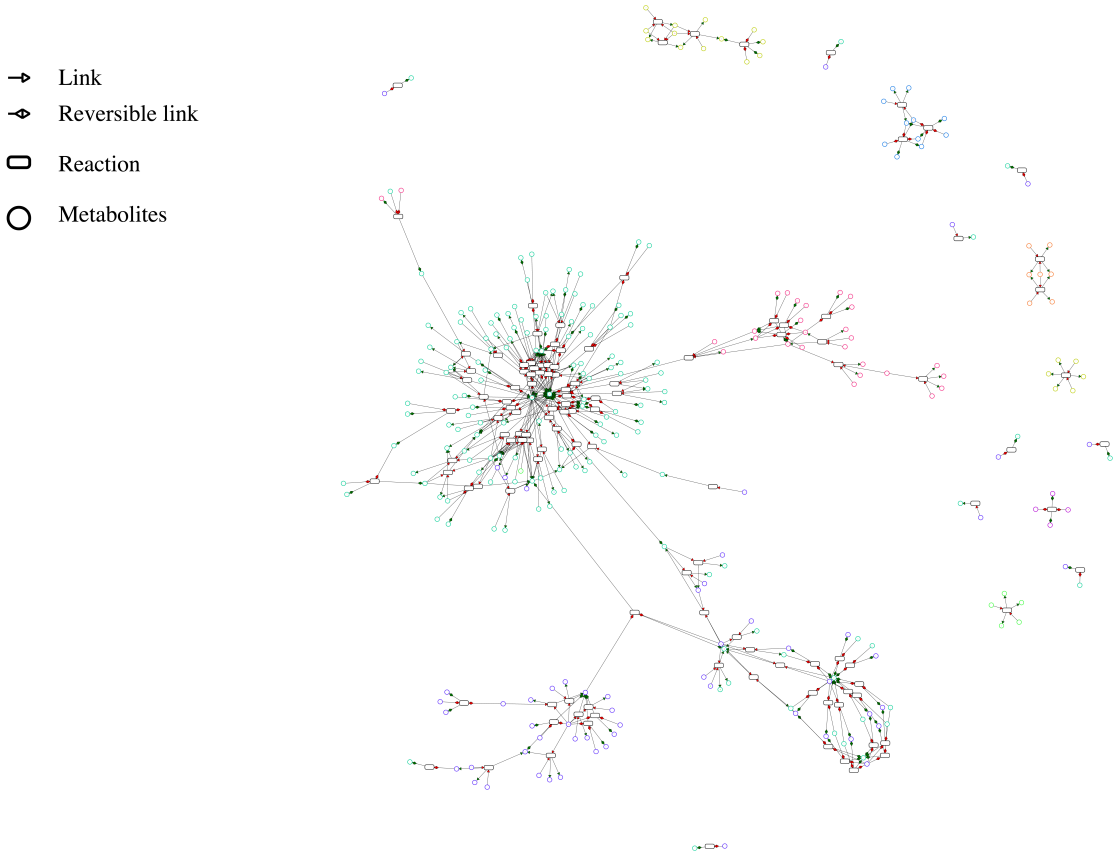

Figure 5.11: Reconstructed metabolic network by mapping the list of genes differentially expressed onto Recon2.2 human metabolic model.

Metabolite mapping

While the web service version of MetExplore allows a mapping based on InChIKey identifiers, this option is not available on the online version. Metabolites were therefore matched using their names, which required manual curation of the metabolites to map on the network. The list of metabolites selected for the mapping was restricted to the standard metabolites with the addition of the annotated metabolites showing a statistically significant difference in the two biological groups, which makes 59 metabolites in total. 21 of these metabolites were found to be present in the network reconstructed from the differentially expressed genes, the list is available in the appendix table D.1. Figure 5.12 shows the final reconstructed network after duplication of "side compounds" for better visualisation. The metabolite nodes considered as side compounds in this network were CO_2, H_2O, NA^+, ADP, Coenzyme A and H^+. The metabolites mapped on the network are highlighted with a red border.

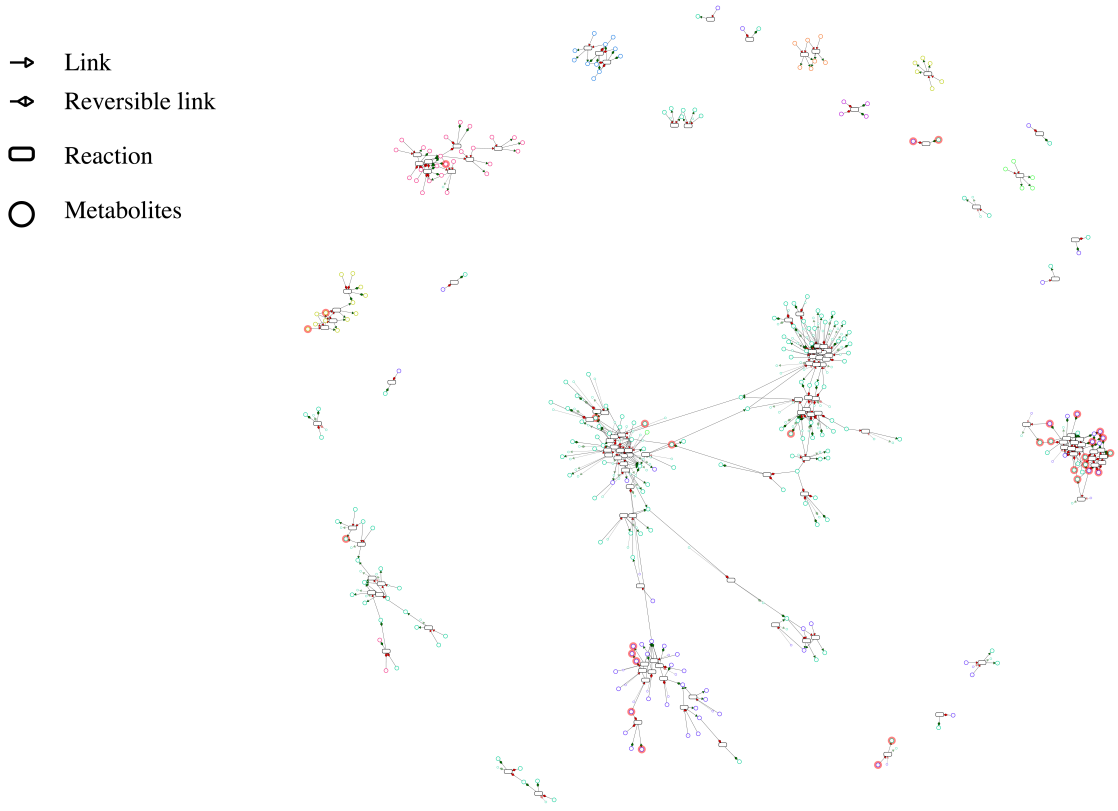

Figure 5.12: Integrated network showing the mapping of metabolites (with red border) onto the previously reconstructed metabolic network. This network has been processed by duplicating side compounds for better readability. The network shows two highly connected clusters at the bottom and the right of the figure.

The mapping allowed the identification of two main clusters showing a high number of connection between the RNA-seq and metabolomics data. The meaning of these clusters is discussed in the next section.

5.8 Discussion

The combined and integrated study of multi-omics data has the potential to offer a holistic approach to understanding biological systems of organisms. Chapter 3 and 4 addressed the limitation met by the most immature technology of the omics family, metabolomics. It addressed the issues in designing, documenting and reporting complex experiments performed in omics fields in an effort to standardise, automate and streamline the tasks assigned to the different contributors of an omics study. The common and general aim of the work presented in those two chapters, however, is to provide the necessary tools to optimise experiment design and fully exploit the results of an omics study, with a particular focus on untargeted metabolomics. Due to the complexity of both the data generated by omics technologies and the biological system studied, it is currently not possible to assert if an omics dataset has been fully mined or still hold unexploited biological insight. With the aim of improving data exploration and mining, the work presented in Chapter 4 attempted to expand the context of interpretation of untargeted metabolomics data using genome-scale metabolic models.

The limitations met in single omics field such as the ones presented in metabolomics can be transferred to the integrative approach. For example, the simple existence of many standards for each omics field makes their integration a great challenge [157]. The differences in variability of the different components of a biological system increase the complexity of designing a study when planning to interrogate several omics layers at once. These issues raise the question whether it possible or not to connect different omics to one another to study them in concert. The work presented in this chapter attempted to answer this question. Another question that was addressed in this chapter was to assess whether the study of two integrated omics gives a better overview of a biological process than the study of a single omics layer.

As presented in this chapter, six annotated or identified metabolites have been found to be significantly changing over time or significantly different across the two biological groups. However, many peaks didn't find any match in the metabolite public databases used in PiMP. Indeed, the last filtering step brought the list of metabolites significantly changing to less than 10% of the initial list, due to a vast majority of unassigned peaks. No fragmentation data was acquired at the time of data collection, but as PiMP now allows the analysis of fragmentation data as part of the standard untargeted metabolomics data analysis pipeline, repeating this experiment with fragmentation data acquisition could produce valuable data to explore. The statistical power issue observed in this experiment could also be further investigated by a repeated study with a higher number of replicates. As the samples studied seem to be highly variable between replicates, it could help identifying potential outliers and have a better confidence on the metabolite levels seen in each group.

From the five metabolites that were found to be either changing over time or be at different

levels in the biological groups, choline phosphate is of particular interest as it is one of the binding targets of C-reactive protein (CRP) which is known to be involved in inflammatory response. The low levels of choline phosphate in PHA-activated cells could indicate a very little abundance of this metabolite in free form, but a high abundance attached to C-reactive protein binding sites. Many studies addressed the question of whether a relationship exists between CRP levels and smoking. In their review [168], S. Tonstad and J. L. Cowan discuss the conflicting results from different studies approaching this relationship. A more recent study [169], however, indicates that while CRP levels in the blood are associated with smoking, it does not correlate with smoking intensity. Current methods of measurement of CRP levels using high-sensitivity C-reactive protein test could be considered to confirm the hypothesis formulated earlier. This could also help understand better the response time of keratinocyte to their stimulation, and the delay to its observation at the metabolome level.

It is apparent that the metabolomics experiment of the system studied suffered from a low statistical power which translated into a low number of compounds that can potentially be interpreted into biological context with confidence. While one of the compounds, choline phosphate, may reflect the immune response triggered by the stimulation of oral keratinocyte α7nAChR, the rest of the dataset couldn't be further exploited with high confidence. However, the trends showed by the metabolites can be informative and are discussed next.

The integration of metabolomics data together with gene expression gave, however, some answers. Indeed, two main clusters can be identified from the reconstructed network. The first one, at the bottom of Figure 5.12 represents the purines metabolism in which can be found AMP and adenine linked together through AMP pyrophosphorylase, the enzyme which takes AMP as a substrate to produce adenine. One of the genes involved in the production of the enzyme was found to be significantly down-regulated over time in both biological conditions. This change is reflected in the levels of adenine in the metabolomics data which are going down over time. An increase of AMP might be expected but is not observed in the metabolomics data (Figure 5.13), the regulation of AMP levels cannot be explained with the data available in this network. Gene expression data also suggests a down-regulation of the production of ribonucleosides over time, data partially reflected in the metabolomics data in the levels of adenosine, cytidine and guanosine, in direct link with this reaction as products (Figure 5.13). While transcription data suggest a general down-regulation of the purine metabolism after a few hours, metabolomics data tends to reflect these changes after 24 hours only. The regulation of nucleotide synthesis has been found to play a significant role in cancer cells and represent a therapeutic target, however, these processes at the metabolic level are still poorly understood [170].

A second cluster disconnected from the main network can be identified on the right of Figure 5.12. This network link the amino acids present in the metabolomics data to amino acids transporter. The expression of the gene coding for the ATP-binding cassette sub-family A

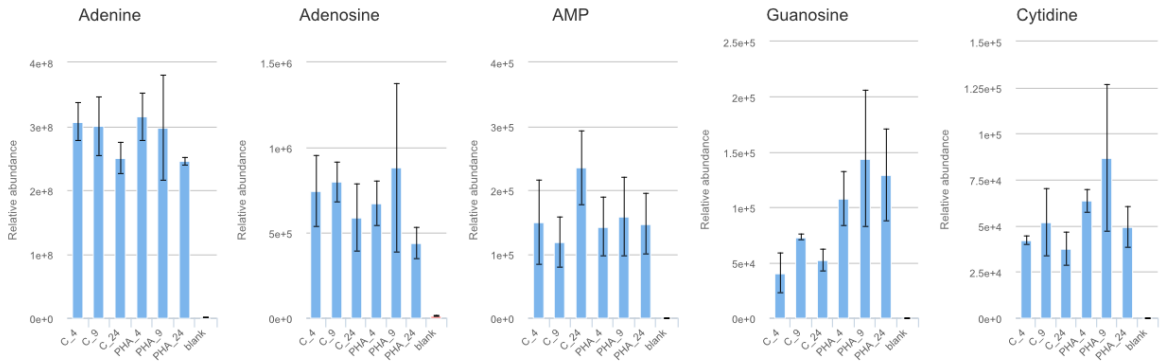

Figure 5.13: Average intensities of metabolites found in the first cluster identified in the network.

member 1, subunit of amino acid transferase, shows a clear up-regulation over time (log fold change of 1.6 in PHA group over 24 hours, 0.9 in Control group but not statistically significant). Many of the connected amino acids were identified during the metabolomics analysis but did not present significant changes. They are however part of the standard compounds showing subtle changes of levels going down over time in PHA activated samples as shown in Figure 5.14. This result shows that the levels observed in the metabolomics results should be further investigated in a separate study as they are not statistically significant. These results support previous studies showing that the expression levels of amino acids transporters are elevated in primary human cancers [171].

Choline phosphate, the main identified metabolite significantly different in the biological classes, has not been matched onto the network and no related genes appear in gene expression data. It, therefore, could not be further interpreted.

The integrative analysis seems to support a potential biological meaning of the subtle diminution of the amino acids levels after 24 hours. This conclusion tends to suggest that a repeated experiment with more replicates and longer time course - to account for the time required for the regulation of genes to be reflected on the metabolome; may be necessary to have a better view at the metabolome levels during the processes operating in the system when stimulated. Many differences were however seen in this dataset but couldn't be credited to any compounds using a standard untargeted metabolomics approach, taking benefit of fragmentation data analysis combined with standard untargeted methods could improve this identification coverage and bring a better understanding of the system. This would, in turn, improve and expand the connections seen in the integrated network and therefore lead to a more in-depth interpretation.

While the first and third research questions have been addressed as an integrated metabolic network approach seem to be achievable to connect two omics datasets together and brings added information on their understanding, the automation of this process (Research question

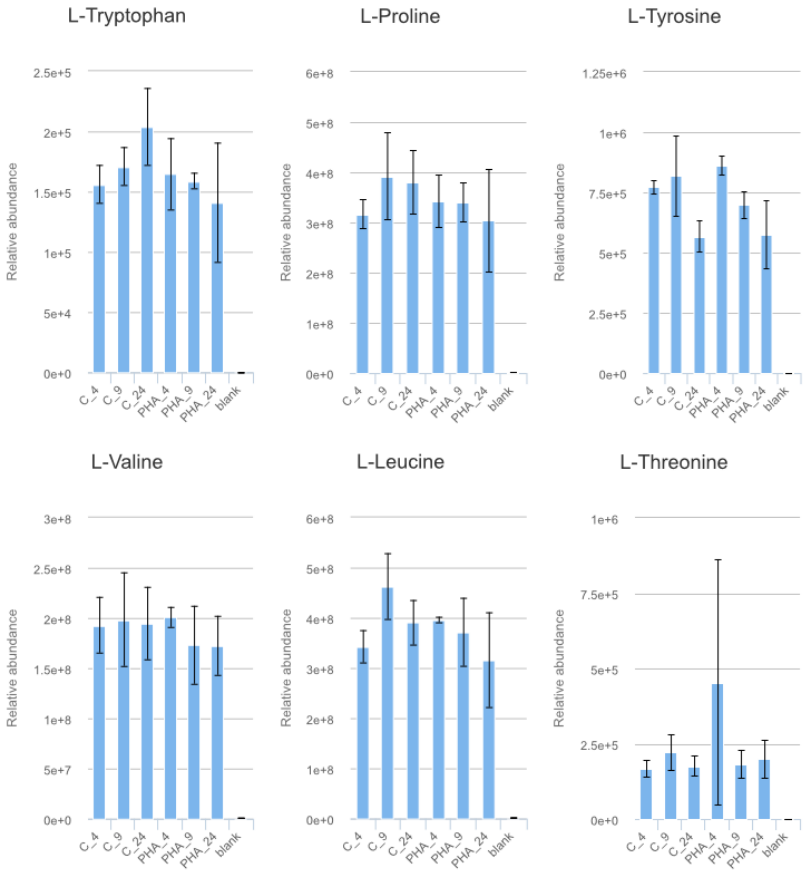

Figure 5.14: Average intensities of amino acids found to be highly connected with gene expression data as seen in the integrated network.

number 2) requires additional work beyond the current state of omics technologies. The issue of no unique standard identifiers was raised many times at several levels, and manual interventions to curate and filter the datasets to proceed to their integration are still key steps that no automated algorithm can currently handle.

During this work, other attempts using different approaches to integrating gene expression and metabolomics data have been reported. R. Cavill et al., for example, present different techniques that can be applied for statistical data integration of metabolomics and gene expression data [172]. These methods, however, do not fully address how the processed and integrated data should be interpreted in a biological context, which is one of the biggest challenges of multi-omics data analysis as B. Palsson and K. Zengler point out in their commentary [158]. The network and knowledge-based approach presented in this chapter attempted to fill this missing step of contextual interpretation to lead to biological knowledge.

5.9 Conclusion

The collection and interpretation of large-scale datasets are powering new discoveries across all disciplines in biomedical science. The recent advances in high-throughput 'omics' technologies such as genomics, metagenomics, transcriptomics, proteomics and metabolomics and improvement in bioinformatics have enabled the investigation of thousands of genes, proteins and metabolites simultaneously. However, while individual datasets are informative and combined analysis of transcriptomic, genomic, proteomic and metabolomic has been found to be very useful for a deeper understanding of key biological processes, greater inferences can be obtained by integrating those datasets that are collected at different levels of biological organization. No standard approach for integrating these datasets is currently available and the existing tools to assist this process are in their infancy. The development and standardisation of new integrative approaches is a crucial step towards providing a holistic perspective of the systems of organisms and can have a major impact beyond the fields of biomedical science.

Chapter 6

General discussion

Nowadays, omics technologies play a major role in the study and the understanding of biological systems. As discussed previously, omics approaches are regularly applied in many biological research fields. However, all omics fields do not have the same maturity, and many challenges are yet to be overcome to use the technologies to their full potential. Metabolomics, the most recent and immature omics is offering great promises for the understanding of biological systems. Indeed, metabolomics forms the link between phenotype and all processes occurring in a biological system, and as such, opens many new possibilities. It also offers the opportunity of finally studying biological systems as a whole, linking genotype to phenotype, environmental stress to direct biological processes offering a holistic view of organisms systems' organisation. However, many limitations still have to be addressed to reach the possibility to systematise and automatise this global approach. The field of metabolomics itself is not yet at its maturity, and many issues from study design to data interpretation in a biological context have yet to be addressed.

Chapter 3 presented the development of a new data analysis and interpretation environment with the aim to expand the reach of the field of metabolomics and improve the biological insight that can be extracted from the dataset. To achieve this objective, three questions were addressed: (i) Can software solutions support non-expert in their metabolomics data analysis? (i) Can software solutions adapt to the rapidly evolving technology requirements? (i) Can world-wide collaboration be enabled in this era of big data?

As discussed in Chapter 3, those three questions were successfully addressed by the development of a semi-automated web-enabled LCMS metabolomics data analysis pipeline. The program integrates a simple step by step data capture pipeline offering the necessary tools to quality control the data, share results directly online and support data interpretation. Moreover, the modular design of the software offers the possibility for developers to expand the functionalities in a responsive manner. Rather than yet another LCMS data analysis pipeline, the developed platform should be considered as the foundation of what could become a com-

munity effort to standardise and unify metabolomics data analysis; offering the possibility to add new blocks for everyone's needs and every advance made on the technology side. The next stage for the PiMP is, therefore, to be easily deployable on any platform, from production server for laboratories treating sensitive information and with specific security requirements, to personal computers development purposes. This transition can be enabled by the Docker container technology and should allow developers around the world work on the development of the same platform in a collaborative effort. One of the limitations discussed in Chapter 3 was the current availability of a single back-end data analysis pipeline; this should also be considered as a future improvement to enlarge the options available to the user. Significant efforts have been made to support the interpretation of the data within the biological context, but this could also be extended by, for example, providing pathway information from databases other than KEGG such as MetCyc. The last short term improvement that should be brought to the platform is the visualisation support of peaks after alignment. Indeed, the results environment currently allow the visualisation of unaligned peaks from the raw files only. This, however, will require careful planning to preserve the modularity of the platform as the alignment is dependent on the back-end pipeline used for the analysis, and therefore, the data format and structure produced. Finally, it was revealed during the development of the fragmentation module presented in Chapter 4, that, while the design of PiMP allows the integration of new features, the modules could have been broken down into smaller units. Limiting each block to a single feature would have made the platform even more modular and ease new modules integration.

In comparison to the current tools, as discussed in Chapter 3, the PiMP platform offer several advantages such as no installation, sharing capabilities or a semi-automated pipeline centralising the data capture at the start. The modularity of the tool is also an asset which is further discussed next.

In their review [173], B. Misra and J. van der Hooft list over one hundred tools, software and databases in an attempt to make researchers aware of the recent development to analyse metabolomics data. While the review is a useful resource, it also shows the sparse state of the current solutions in the field. The diversity of approaches is a necessity for a developing field, but their uncoordinated development can be detrimental to the research. In a fast-evolving field such as metabolomics, it seems imperative for the community to join efforts to organise and standardise the different approaches. This would enable responsive development of new solutions to progress alongside the technology and facilitate systematic control studies.

While the work presented in Chapter 3 was exclusively focused on the support of data analysis, visualisation and interpretation, in Chapter 4 were discussed the support of tasks at various stages of the metabolomics workflow. The work presented in Chapter 4 attempted to provide support for study design and data capture as well as extending the data analysis and interpretation, areas of limitation which require critical support to improve the field. Four

research questions were outlined to address these limitations.

The development of an online tool to support data capture and accurate reporting of the work in a unified manner successfully addressed the first question. The resulting online platform allows data capture from many contributors in a unified environment, logging every piece of information from sample preparation to data analysis. As the developed tool used the same technology as the PiMP software, its modularity allows its extension to support any omics technology. The development of chemical library was developed in complement to the management system, attempting to inform biologist on the potential outcome of an LCMS experiment depending on the organism studied (Second research question). These two platforms bring support to a critical part of omics experiments, study design. In the same way as PiMP, the management system and chemical library should be considered as a framework that can be enriched to support better the different tasks required during a metabolomics study. The first, short-term, improvement to bring to the management system is to plug it to the PiMP program to create a single platform supporting scientists from study design to data interpretation. In a longer term, however, connecting this unified platform to a metabolomics data repository such MetaboLight or Metabolomics Workbench could automate the entire process of publishing data. This would have two immediate effects; more scientists would publish their data along with scientific articles, which would, in turn, enrich the current base knowledge available in these repositories and enable systematic study validation.

The other two questions both related to the improvement of data analysis and interpretation by the extension of PiMP. A fragmentation module was developed and demonstrated the modularity of the PiMP pipeline, and an external tool, MetExploreViz, was integrated into the PiMP platform to expand the data interpretation features. The fragmentation module allowed for the first time the automated analysis of fragmentation data as part of a standard untargeted metabolomics pipeline, bringing better confidence in metabolites annotation and, therefore, better coverage of the metabolome for data interpretation. It is, however, currently limited in the number of external databases used, and could give better confidence enlarging the fragment spectra comparison to other databases. The integration of MetExploreViz and the development of a communication protocol between PiMP and MetExplore web services extension of the data interpretation within the PiMP platform. This new tool has the benefit of eliminating all manual steps required to visualise metabolomics data at a system level, and, therefore, enable scientists with no experience in data formatting to proceed to a network analysis. The current state of the tool is, however, limited to metabolites identified against a standard compound. It will be imperative as a future improvement, to allow users to select annotated compounds (matched by mass only) to have more flexibility and control over the network analysis.

The work presented in Chapter 4 shows that the PiMP platform can be used as the foun-

dation to support the development of new features. It also indicates that the tool still has a great potential for improvement. While this work demonstrated that it is possible to support metabolomics users from study design to data interpretation; unifying the tools together into one unique semi-automated platform, would have an even greater impact on the metabolomics community. From a broader perspective, this approach could be applied to many data analysis software in biological science and other omics, providing simple, standardised platform of analysing data without compromising on the diversity of analysis approaches. While KNIME and Taverna offer the possibility to create data analysis pipelines, Galaxy is currently the main solution used within the metabolomics community for this purpose [174]. These solutions have, however, severe limitations in term of data visualisation and user assistance. The PhenoMeNal consortium, in an effort to standardise metabolomics data analysis, is proposing an alternative approach using docker containers to encapsulate data analysis tools and standardise their usage. This very same approach has been implemented in PiMP and enables the creation of pipelines by linking several docker containers together. A full data analysis pipeline has, however, yet to be implemented as part of the PhenoMeNal e-infrastructure.

Chapter 5 focused on the analysis of metabolomics data in a broader context. The work presented integrates metabolomics and RNA-seq data in an attempt to inform and interpret further metabolomics data analysis results. The approach taken makes use of genome-scale metabolic networks reconstruction to link the two omics together. This work also attempted to assess whether the automation of integrated analysis is possible. The integration resulted in network connecting metabolites to transcripts data which validated the method. However, little biological information could be extracted from this network for several reasons. First, the lack of statistical power of the metabolomics data due the low number of replicates resulted in a small list of metabolites significantly changing between the different groups. The second issue was the number of metabolites with no putative annotations, which could not be included in the integrative analysis. As mentioned in the discussion of chapter 5, the acquisition of fragmentation data could have improved the identification of metabolites, and a higher number of replicates would have given better confidence in the changes observed. Some information could, however, be extracted but should be confirmed by a new experiment. The subtle changes in amino acid contents show a correlation with RNA-seq data, however, not significant. The significant difference of choline phosphate in the two biological groups clearly suggests that the activation of nicotinic receptors has not only an effect on the cascading pathways but also on the metabolisms. The hypothesis was made that this difference could reflect an inflammatory response and would be related to choline phosphate binding to C-reactive proteins. This assumption, however, would require testing using a hs-CRP test. While no real conclusion could be made from this integrative approach, the results suggest that the processes involved in the tumorigenic effects of activated nicotinic receptors

could have an effect on the metabolism.

The approach taken was also found to require too many manual transformations of the data to be automatised at this stage. While efforts are being made towards standardisation of omics data, the lack of standard identifiers at every omics level still represents a barrier to automated holistic approaches.

High-throughput omics technologies have enabled an in-depth investigation of the organisation of the systems of organisms. The potential of the data, is, however, not yet fully exploited. Omics technologies still need to evolve towards an overall standardisation to offer holistic approaches to biological science. The disparity between omics data generation and its in-depth analysis needs to be addressed. As suggested in this commentary [158], one way of improving the analysis and integration of omics data is to increase community-driven development. Another way of improving data analysis and integration is the systematic publication of datasets in dedicated public data repositories. As mentioned in this review [157], the integration of Laboratory Information Management Systems and its standardisation and use in submission to public data repositories can lead to a more efficient use of the data. However, while the development of resources to support scientists in this task is in progress, it will require fundamental changes in the way the scientific community publishes research to make this process of data documentation and publication a standard requirement.

Appendix A

PiMP libraries

A.1 R libraries

- AnnotationDbi
- acepack
- BH
- Biobase
- BiocGenerics
- BiocInstaller
- Biostrings
- bitops
- brew
- caTools
- cluster
- codetools
- colorspace
- curl
- DBI
- devtools
- dichromat
- digest
- doParallel
- evaluate
- fields
- foreach
- foreign
- Formula
- gdata
- ggplot2
- git2r
- gplots
- gptk
- gridExtra

- gtable
- gtools
- Hmisc
- httr
- impute
- IRanges
- iterators
- jsonlite
- KEGGREST
- KernSmooth
- labeling
- lattice
- latticeExtra
- limma
- magrittr
- maps
- MASS
- Matrix
- memoise
- mime
- munsell
- mzR
- mzmatch.R
- nloptr
- nnet
- outliers
- packrat
- plyr
- png
- PiMP
- PiMPDB
- ProtGenerics
- proto
- ptw
- RColorBrewer
- Rcpp
- RCurl
- rJava
- RJSONIO
- RMySQL
- reshape2
- roxygen2
- rpart
- RSQLite
- rstudioapi
- RUnit
- rversions
- R.methodsS3
- R.oo
- R.utils
- R6
- scales

- snow
- spam
- stringi
- stringr
- survival
- S4Vectors
- whisker
- xcms
- XLConnect
- XLConnectJars
- XML
- xml2
- XVector
- yaml
- zlibbioc

A.2 Python libraries

- amqp
- amqplib
- anyjson
- billiard
- celery
- Django
- django-celery
- django-chartit
- django-extensions
- django-registration
- django-jquery-file-upload
- funcsigs
- honcho
- jsonpickle
- kombu
- lxml
- matplotlib
- mock
- MySQL-python
- nose
- numpy
- pbr
- Pillow
- pip
- pyparsing
- PySide
- python-dateutil
- pytz
- rpy2
- scipy
- scikit-learn

- setuptools
- simplejson
- six
- sqlparse
- suds
- virtualenv
- virtualenvwrapper
- wheel

A.3 JavaScript libraries

- Highcharts
- jQuery
- Select2
- Bootstrap 2 and 3
- DataTables
- -prefix-free

Appendix B

List of Standard compounds

Name	Formula	C9/C4	C24/C4	PHA9/PHA4	PHA24/PHA4	PHA4/C4	PHA9/C9	PHA24/C24
(R)-2-Hydroxyglutarate	C5H8O5	0.48	0.63	-0.36	0.09	0.36	-0.48	-0.18
(S)-Malate	C4H6O5	-0.09	0.71	-0.15	0.16	-0.07	-0.12	-0.62
4-Aminobenzoate	C7H7NO2	-0.15	-0.1	-0.24	-0.59	0.14	0.05	-0.35
4-Trimethylammoniobutanoate	C7H15NO2	-0.12	-0.07	-0.64	-0.62	0.34	-0.18	-0.2
5-Oxoproline	C5H7NO3	0.25	0.34	-0.2	-0.12	0.11	-0.34	-0.35
Acetylcholine	C7H15NO2	0.26	-0.09	0.02	-2.14	2.22	1.97	0.17
Adenine	C5H5N5	-0.04	-0.3	-0.14	-0.35	0.03	-0.06	-0.02
Adenosine	C10H13N5O4	0.15	-0.36	0.07	-0.62	-0.11	-0.19	-0.37
AMP	C10H14N5O7P	-0.27	0.76	0.07	0.04	0.01	0.35	-0.71
Betaine	C5H11NO2	0.07	0.37	-0.19	-0.29	0.2	-0.06	-0.47
Choline	C5H13NO	0.6	-0.12	-0.55	-0.17	0.25	-0.9	0.21
Choline phosphate	C5H14NO4P	-0.06	0.09	-0.31	-0.7	-2.67	-2.92	-3.45
Cytidine	C9H13N3O5	0.22	-0.2	0.29	-0.4	0.59	0.66	0.39
D-Gluconic acid	C6H12O7	0.27	0.71	-0.15	-0.16	0.18	-0.23	-0.68
D-Glucosamine	C6H13NO5	-0.21	-0.54	0.03	-0.87	0.25	0.49	-0.07
D-glucose	C6H12O6	0.36	-0.77	-0.48	-1.1	0.33	-0.51	0
D-glucose 6-phosphate	C6H13O9P	0.49	0.55	0.07	0.14	-0.03	-0.45	-0.44

Name	Formula	C9/C4	C24/C4	PHA9/PHA4	PHA24/PHA4	PHA4/C4	PHA9/C9	PHA24/C24
Guanine	C5H5N5O	0.94	3.69	1.41	3.16	0.65	1.12	0.11
Guanosine	C10H13N5O5	1.02	0.52	0.26	0.22	1.55	0.79	1.25
HEPES	C8H18N2O4S	0.18	0.09	-0.22	-0.31	0.17	-0.22	-0.22
Hypoxanthine	C5H4N4O	0.63	2.66	0.48	1.66	0.6	0.45	-0.4
Imidazole-4-acetate	C5H6N2O2	0.08	0.48	-0.3	-0.18	0.08	-0.3	-0.58
Inosine	C10H12N4O5	0.7	-0.01	0.15	0.11	0.93	0.38	1.06
L-Alanine	C3H7NO2	0.57	1.27	0.24	1.28	0.07	-0.25	0.09
L-Arginine	C6H14N4O2	0.46	0.16	-0.2	-0.09	0.16	-0.5	-0.09
L-Aspartate	C4H7NO4	-0.05	-0.76	-0.34	-0.82	-0.09	-0.38	-0.14
L-Carnitine	C7H15NO3	-0.21	-0.36	-0.33	-0.8	0.4	0.27	-0.05
L-Citrulline	C6H13N3O3	0.31	0.49	-0.3	-0.69	0.43	-0.19	-0.76
L-Glutamate	C5H9NO4	0.36	0.11	-0.12	-0.14	0.22	-0.26	-0.04
L-Glutamine	C5H10N2O3	0.11	0.14	-0.16	-0.26	0.11	-0.16	-0.29
L-Histidine	C6H9N3O2	0.06	0.11	-0.1	-0.27	0.13	-0.04	-0.25
L-Leucine	C6H13NO2	0.43	0.19	-0.11	-0.4	0.21	-0.32	-0.38
L-Lysine	C6H14N2O2	0.72	0.35	-0.23	-0.15	0.28	-0.68	-0.22
L-Methionine	C5H11NO2S	0.05	0.07	-0.18	-0.53	0.21	-0.02	-0.39

Name	Formula	C9/C4	C24/C4	PHA9/PHA4	PHA24/PHA4	PHA4/C4	PHA9/C9	PHA24/C24
L-Ornithine	C5H12N2O2	1.39	1.84	0.57	1.45	0.07	-0.75	-0.32
L-Phenylalanine	C9H11NO2	0.21	-0.36	-0.21	-0.39	0.26	-0.16	0.23
L-Proline	C5H9NO2	0.28	0.25	-0.01	-0.25	0.11	-0.18	-0.39
L-Serine	C3H7NO3	0.28	0.38	-0.19	-0.36	0.35	-0.12	-0.39
L-Threonine	C4H9NO3	0.38	0.07	-0.79	-0.67	0.87	-0.3	0.14
L-Tryptophan	C11H12N2O2	0.36	0.34	-0.09	-0.17	0.2	-0.25	-0.31
L-Tyrosine	C9H11NO3	0.29	-0.28	-0.08	-0.68	0.29	-0.08	-0.11
L-Valine	C5H11NO2	0.01	0	-0.25	-0.24	0.07	-0.19	-0.16
Lipoate	C8H14O2S2	0.17	-0.09	-0.14	-0.28	0.04	-0.27	-0.15
Maleic acid	C4H4O4	0.78	-0.11	-0.34	-0.64	0.49	-0.62	-0.04
N(pi)-Methyl-L-histidine	C7H11N3O2	-0.1	0.84	-0.02	0.24	0.58	0.66	-0.01
N-acetyl-L-glutamate	C7H11NO5	0.52	0.72	-0.38	-0.44	0.82	-0.08	-0.33
Nicotinamide	C6H6N2O	-0.25	-2.58	-0.54	-2.89	0.5	0.2	0.19
Orotidine	C10H12N2O8	-0.14	0.19	-0.19	-0.07	-0.11	-0.17	-0.37
Orthophosphate	H3O4P	0.32	0.26	-0.23	-0.18	0.25	-0.3	-0.2
Pantothenate	C9H17NO5	0.2	0.33	-0.26	-0.59	0.37	-0.1	-0.55
Phenolsulfonphthalein	C19H14O5S	-0.13	0.31	-0.17	-0.05	-0.01	-0.05	-0.37

Name	Formula	C9/C4	C24/C4	PHA9/PHA4	PHA24/PHA4	PHA4/C4	PHA9/C9	PHA24/C24
Phosphoenolpyruvate	C3H5O6P	0.09	-0.31	-0.25	-0.72	-0.25	-0.59	-0.66
Phthalate	C8H6O4	0.1	0.04	0.31	-0.24	0.12	0.34	-0.16
Pyruvate	C3H4O3	-0.33	-0.75	-0.42	-1.07	-0.29	-0.38	-0.61
sn-glycero-3-Phosphocholine	C8H20NO6P	0.27	0.9	-0.09	-0.67	0.1	-0.26	-1.47
sn-Glycerol 3-phosphate	C3H9O6P	0.21	1.53	-0.74	0.14	0.19	-0.75	-1.19
Succinate	C4H6O4	0.02	0.32	-0.4	-0.4	0.17	-0.26	-0.54
thymine	C5H6N2O2	-0.45	-1.12	-0.11	-2.29	0.7	1.04	-0.47

Table B.1: List of metabolite identified against standard compounds.

Appendix C

Integrated network pathway list

Name	Gene Coverage (%)	Nb of Mapped
Alanine and aspartate metabolism	5.26	1
Aminosugar metabolism	2.86	1
Arginine and Proline Metabolism	1.89	1
Bile acid synthesis	1.43	1
Cholesterol metabolism	2.04	1
CoA synthesis	3.13	1
Eicosanoid metabolism	1.02	1
Folate metabolism	5	1
Fructose and mannose metabolism	2	1
Glycerophospholipid metabolism	1.04	1
Glycolysis/gluconeogenesis	1.67	2
Glyoxylate and dicarboxylate metabolism	2.5	1
Heme degradation	20	1
Heme synthesis	10	1
Histidine metabolism	3.7	1
Inositol phosphate metabolism	1.37	1
Methionine and cysteine metabolism	1.92	1

Name	Gene Coverage (%)	Nb of Mapped
Miscellaneous	2.5	2
NAD metabolism	5.56	1
Nucleotide interconversion	2.21	4
Oxidative phosphorylation	5.11	7
Pentose phosphate pathway	2.86	1
Phenylalanine metabolism	6.25	1
Propanoate metabolism	5	1
Purine catabolism	5.77	3
Pyrimidine catabolism	4.65	2
Pyruvate metabolism	1.96	1
Selenoamino acid metabolism	2.78	1
Sphingolipid metabolism	1.59	1
Squalene and cholesterol synthesis	20	1
Steroid metabolism	2.33	1
Transport, extracellular	3.57	11
Transport, golgi apparatus	8.33	1
Transport, mitochondrial	1.85	1
Triacylglycerol synthesis	3.7	1
Tryptophan metabolism	1.61	1
Tyrosine metabolism	1.27	1
Unassigned	3.31	4
Urea cycle	8.33	3
Vitamin A metabolism	3.03	1
Vitamin C metabolism	5.88	1

Table C.1: List of pathways covered by the network reconstructed from the differentially expressed genes.

Appendix D

List of metabolites mapped to the metabolic network

Metabolite name	Formula
Guanosine	C10H13N5O5
Hypoxanthine	C5H4N4O
L-citrulline	C6H13N3O3
L-tryptophan	C11H12N2O2
L-tyrosine	C9H11NO3
L-threonine	C4H9NO3
L-proline	C5H9NO2
D-Fructose 6-phosphate	C6H11O9P
L-glutamate	C5H9NO4
L-valine	C5H11NO2
L-aspartate	C4H7NO4
Thymine	C5H6N2O2
AMP	C10H12N5O7P
Xanthine	C5H4N4O2
Cytidine	C9H13N3O5
D-glucose	C6H12O6
adenine	C5H5N5
Adenosine	C10H13N5O4
L-homoserine	C4H9NO3
L-leucine	C6H13NO2
L-alanine	C3H7NO2

Table D.1: List of metabolites found to be in the human metabolic network reconstructed from the RNA-seq data.

Bibliography

[1] M. J. Aardema and J. T. MacGregor, "Toxicology and genetic toxicology in the new era of toxicogenomics: impact of -omics technologies," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 499, no. 1, pp. 13–25, 2002.

[2] J. Vlaanderen, L. E. Moore, M. T. Smith, Q. Lan, L. Zhang, C. F. Skibola, N. Rothman, and R. Vermeulen, "Application of OMICS technologies in occupational and environmental health research; current status and projections." *Occupational and environmental medicine*, vol. 67, no. 2, pp. 136–43, feb 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19933307

[3] C. E. Wheelock, V. M. Goss, D. Balgoma, B. Nicholas, J. Brandsma, P. J. Skipp, S. Snowden, D. Burg, A. D 'amico, I. Horvath, A. Chaiboonchoe, H. Ahmed, S. Ballereau, C. Rossios, K. F. Chung, P. Montuschi, S. J. Fowler, I. M. Adcock, A. D. Postle, S.-E. Dahlén, A. Rowe, P. J. Sterk, C. Auffray, and R. Djukanovi, "Application of 'omics technologies to biomarker discovery in inflammatory lung diseases," *Eur Respir J*, vol. 42, pp. 802–825, 2013. [Online]. Available: http://ow.ly/mjGGcwww.erj.ersjournals.com

[4] X. Zhang, D. Wei, Y. Yap, L. Li, S. Guo, and F. Chen, "Mass spectrometry-based omics technologies in cancer diagnostics," *Mass Spectrometry Reviews*, vol. 26, no. 3, pp. 403–431, may 2007. [Online]. Available: http://doi.wiley.com/10.1002/mas.20132

[5] H. Davies, "A role for omics technologies in food safety assessment," *Food Control*, vol. 21, no. 12, pp. 1601–1610, 2010.

[6] X. Zhang, Y. Yap, D. Wei, G. Chen, and F. Chen, "Novel omics technologies in nutrition research," *Biotechnology Advances*, vol. 26, no. 2, pp. 169–176, 2008.

[7] E. S. Lander, L. M. Linton, and B. e. a. Birren, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, feb 2001. [Online]. Available: http://www.nature.com/doifinder/10.1038/35057062

[8] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, jan 2009. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrg2484

[9] N. L. Anderson and N. G. Anderson, "Proteome and proteomics: New technologies, new concepts, and new words," *Electrophoresis*, vol. 19, no. 11, pp. 1853–1861, aug 1998. [Online]. Available: http://doi.wiley.com/10.1002/elps.1150191103

[10] W. P. Blackstock and M. P. Weir, "Proteomics: quantitative and physical mapping of cellular proteins," *Trends in Biotechnology*, vol. 17, no. 3, pp. 121–127, 1999.

[11] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz, "Systematic functional analysis of the yeast genome," *Trends in Biotechnology*, vol. 16, no. 9, pp. 373–378, 1998.

[12] R. E. Breitbart, A. Andreadis, and B. Nadal-Ginard, "Alternative Splicing: A Ubiquitous Mechanism for the Generation of Multiple Protein Isoforms from Single Genes," *Annual Review of Biochemistry*, vol. 56, no. 1, pp. 467–495, jun 1987. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev.bi.56.070187.002343

[13] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, "Correlation between protein and mRNA abundance in yeast." *Molecular and cellular biology*, vol. 19, no. 3, pp. 1720–30, mar 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10022859

[14] P. H. Tate and A. P. Bird, "Effects of DNA methylation on DNA-binding proteins and gene expression," *Current Opinion in Genetics & Development*, vol. 3, no. 2, pp. 226–231, 1993.

[15] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature Genetics*, vol. 40, no. 12, pp. 1413–1415, dec 2008. [Online]. Available: http://www.nature.com/doifinder/10.1038/ng.259

[16] P. S. Covello and M. W. Gray, "RNA editing in plant mitochondria," *Nature*, vol. 341, no. 6243, pp. 662–666, oct 1989. [Online]. Available: http://www.nature.com/doifinder/10.1038/341662a0

[17] M. Mann and O. N. Jensen, "Proteomic analysis of post-translational modifications," *Nature biotechnology*, vol. 21, no. 3, pp. 255–261, 2003.

[18] M. de Tayrac, S. Le, M. Aubry, J. Mosser, and F. Husson, "Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor

Analysis approach." *BMC Genomics*, vol. 10, no. 1, p. 32, 2009. [Online]. Available: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-32

[19] C. Meng, B. Kuster, A. C. Culhane, and A. Gholami, "A multivariate approach to the integration of multi-omics datasets," *BMC Bioinformatics*, vol. 15, no. 1, p. 162, 2014. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-162

[20] W. Zhang, F. Li, and L. Nie, "Integrating multiple 'omics' analysis for microbial biology: application and methodologies," *Microbiology*, vol. 156, no. 2, pp. 287–301, feb 2010. [Online]. Available: http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.034793-0

[21] K.-A. Le Cao, I. Gonzalez, and S. Dejean, "integrOmics: an R package to unravel relationships between two omics datasets," *Bioinformatics*, vol. 25, no. 21, pp. 2855–2856, nov 2009. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp515

[22] A. D. Polpitiya, W.-J. Qian, N. Jaitly, V. A. Petyuk, J. N. Adkins, D. G. Camp, G. A. Anderson, and R. D. Smith, "DAnTE: a statistical tool for quantitative analysis of -omics data," *Bioinformatics*, vol. 24, no. 13, pp. 1556–1558, jul 2008. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btn217

[23] F. Rohart, B. Gautier, A. Singh, and K.-A. Le Cao, "mixOmics: an R package for 'omics feature selection and multiple data integration," *bioRxiv*, 2017.

[24] A. Fukushima, M. Kusano, H. Redestig, M. Arita, and K. Saito, "Integrated omics approaches in plant systems biology," *Current Opinion in Chemical Biology*, vol. 13, no. 5, pp. 532–538, 2009.

[25] A. R. Joyce and B. Ø. Palsson, "The model organism as a system: integrating 'omics' data sets." *Nature reviews. Molecular cell biology*, vol. 7, no. 3, pp. 198–210, mar 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16496022

[26] O. Fiehn, "Metabolomics–the link between genotypes and phenotypes." *Plant molecular biology*, vol. 48, no. 1-2, pp. 155–71, jan 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11860207

[27] B. E. Winger, K. J. Light-Wahl, R. R. Ogorzalek Loo, H. R. Udseth, and R. D. Smith, "Observation and implications of high mass-to-charge ratio ions from electrospray ionization mass spectrometry," *Journal of the American Society for*

Mass Spectrometry, vol. 4, no. 7, pp. 536–545, jul 1993. [Online]. Available: http://link.springer.com/10.1016/1044-0305(93)85015-P

[28] F. Santos and M. Galceran, "Modern developments in gas chromatographymass spectrometry-based environmental analysis," *Journal of Chromatography A*, vol. 1000, no. 1, pp. 125–151, 2003.

[29] M. Yamashita and J. B. Fenn, "Electrospray ion source. another variation on the free-jet theme," *J. Phys. chem*, vol. 88, no. 20, pp. 4451–4459, 1984.

[30] M. Karas, D. Bachmann, and F. Hillenkamp, "Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules," *Analytical Chemistry*, vol. 57, no. 14, pp. 2935–2939, dec 1985. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ac00291a042

[31] D. Y. Lee, B. P. Bowen, and T. R. Northen, "Mass spectrometry-based metabolomics, analysis of metabolite-protein interactions, and imaging." *BioTechniques*, vol. 49, no. 2, pp. 557–65, aug 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20701590

[32] R. Yost and C. Enke, "Triple quadrupole mass spectrometry for direct mixture analysis and structure elucidation." MICHIGAN STATE UNIV EAST LANSING DEPT OF CHEMISTRY, Tech. Rep., 1979.

[33] "Hydrophilic interaction liquid chromatography (HILIC)–a powerful separation technique." *Analytical and bioanalytical chemistry*, vol. 402, no. 1, pp. 231–47, jan 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21879300

[34] K. Dettmer, P. A. Aronov, and B. D. Hammock, "Mass spectrometry-based metabolomics." *Mass spectrometry reviews*, vol. 26, no. 1, pp. 51–78, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16921475

[35] P. G. A. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, and R. Aebersold, "A common open representation of mass spectrometry data and its application to proteomics research," *Nature Biotechnology*, vol. 22, no. 11, pp. 1459–1466, nov 2004. [Online]. Available: http://www.nature.com/doifinder/10.1038/nbt1031

[36] S. M. Lin, L. Zhu, A. Q. Winter, M. Sasinowski, and W. A. Kibbe, "What is mzXML good for?" *Expert Review of Proteomics*, vol. 2, no. 6, pp. 839–845, dec 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16307524

[37] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Römpp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz, and E. W. Deutsch, "mzML–a community standard for mass spectrometry data." *Molecular & cellular proteomics : MCP*, vol. 10, no. 1, p. R110.000133, jan 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20716697

[38] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M.-Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, and P. Mallick, "A cross-platform toolkit for mass spectrometry and proteomics," *Nature Biotechnology*, vol. 30, no. 10, pp. 918–920, oct 2012. [Online]. Available: http://www.nature.com/doifinder/10.1038/nbt.2377

[39] J. D. Holman, D. L. Tabb, and P. Mallick, "Employing ProteoWizard to Convert Raw Mass Spectrometry Data." *Current protocols in bioinformatics*, vol. 46, pp. 13.24.1–9, jun 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24939128

[40] E. Afgan, D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Grüning, A. Guerler, J. Hillman-Jackson, G. Von Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko, and J. Goecks, "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update," vol. 44, no. W1, pp. W3–W10, jul 2016. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw343

[41] R. Tautenhahn, C. Bottcher, and S. Neumann, "Highly sensitive feature detection for high resolution LC/MS," *BMC Bioinformatics*, vol. 9, no. 1, p. 504, 2008. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-504

[42] Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and G. Siuzdak*, "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification," *Analytical Chemistry*, 2006.

[43] Y. Tikunov, A. Lommen, C. H. R. de Vos, H. A. Verhoeven, R. J. Bino, R. D. Hall, and A. G. Bovy, "A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles." *Plant physiology*, vol. 139, no. 3, pp. 1125–37, nov 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16286451

[44] M. Katajamaa, J. Miettinen, and M. Oresic, "MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data." *Bioinformatics (Oxford, England)*, vol. 22, no. 5, pp. 634–6, mar 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16403790

[45] A. Kassidas, J. F. MacGregor, and P. A. Taylor, "Synchronization of batch trajectories using dynamic time warping," *AIChE Journal*, vol. 44, no. 4, pp. 864–875, apr 1998. [Online]. Available: http://doi.wiley.com/10.1002/aic.690440412

[46] P. H. C. Eilers, "Parametric Time Warping," *Analytical Chemistry*, vol. 76, no. 2, pp. 404–411, jan 2004. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ac034800e

[47] A. van Nederkassel, M. Daszykowski, P. Eilers, and Y. V. Heyden, "A comparison of three algorithms for chromatograms alignment," *Journal of Chromatography A*, vol. 1118, no. 2, pp. 199–210, 2006.

[48] N.-P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping," *Journal of Chromatography A*, vol. 805, no. 1, pp. 17–35, 1998.

[49] J. Listgarten, R. M. Neal, S. T. Roweis, and A. Emili, "Multiple alignment of continuous time series," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, 2005, pp. 817–824. [Online]. Available: http://papers.nips.cc/paper/2721-multiple-alignment-of-continuous-time-series.pdf

[50] C. Christin, A. K. Smilde, H. C. J. Hoefsloot, F. Suits, R. Bischoff, and P. L. Horvatovich, "Optimized Time Alignment Algorithm for LCMS Data: Correlation Optimized Warping Using Component Detection Algorithm-Selected Mass Chromatograms," *Analytical Chemistry*, vol. 80, no. 18, pp. 7012–7021, sep 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18715018

[51] C. Christin, H. C. J. Hoefsloot, A. K. Smilde, F. Suits, R. Bischoff, and P. L. Horvatovich, "Time Alignment Algorithms Based on Selected Mass Traces for

Complex LC-MS Data," *Journal of Proteome Research*, vol. 9, no. 3, pp. 1483–1495, mar 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20070124

[52] J. T. P. And and E. M. Marcotte*, "Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping," 2006.

[53] B. Voss, M. Hanselmann, B. Y. Renard, M. S. Lindner, U. Kothe, M. Kirchner, and F. A. Hamprecht, "SIMA: Simultaneous Multiple Alignment of LC/MS Peak Lists," *Bioinformatics*, vol. 27, no. 7, pp. 987–993, apr 2011. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr051

[54] N. Hoffmann, M. Keck, H. Neuweger, M. Wilhelm, P. Högy, K. Niehaus, and J. Stoye, "Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets," *BMC Bioinformatics*, vol. 13, no. 1, p. 214, 2012. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-214

[55] R. Ballardini, M. Benevento, G. Arrigoni, L. Pattini, and A. Roda, "MassUntangler: A novel alignment tool for label-free liquid chromatographymass spectrometry proteomic data," *Journal of Chromatography A*, vol. 1218, no. 49, pp. 8859–8868, 2011.

[56] R. Smith, D. Ventura, and J. T. Prince, "Novel algorithms and the benefits of comparative validation," *Bioinformatics*, vol. 29, no. 12, pp. 1583–1585, jun 2013. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt176

[57] E. Lange, R. Tautenhahn, S. Neumann, and C. Gröpl, "Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements," *BMC Bioinformatics*, vol. 9, no. 1, p. 375, 2008. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-375

[58] R. Smith, D. Ventura, and J. T. Prince, "LC-MS alignment in theory and practice: a comprehensive algorithmic review," *Briefings in Bioinformatics*, vol. 16, no. 1, pp. 104–117, jan 2015. [Online]. Available: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbt080

[59] M. Berg, M. Vanaerschot, A. Jankevics, B. Cuypers, R. Breitling, and J.-C. Dujardin, "LC-MS metabolomics from study design to data-analysis - using a versatile pathogen as a test case." *Computational and structural biotechnology journal*, vol. 4, p. e201301002, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24688684

[60] R. A. Scheltema, A. Kamleh, D. Wildridge, C. Ebikeme, D. G. Watson, M. P. Barrett, R. C. Jansen, and R. Breitling, "Increasing the mass accuracy of high-resolution LC-MS data using background ions - A case study on the LTQ-Orbitrap," *Proteomics*, vol. 8, no. 22, pp. 4647–4656, 2008.

[61] R. A. Scheltema, A. Jankevics, R. C. Jansen, M. A. Swertz, and R. Breitling, "PeakML/mzMatch: A File Format, Java Library, R Library, and Tool-Chain for Mass Spectrometry Data Analysis," *Analytical Chemistry*, vol. 83, no. 7, pp. 2786–2793, apr 2011. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ac2000994

[62] T. M. Annesley, "Ion suppression in mass spectrometry." *Clinical chemistry*, vol. 49, no. 7, pp. 1041–4, jul 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12816898

[63] C. Kuhl, R. Tautenhahn, C. Böttcher, T. R. Larson, and S. Neumann, "CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets." *Analytical chemistry*, vol. 84, no. 1, pp. 283–9, jan 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3658281

[64] L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W.-M. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. N. Lane, J. C. Lindon, P. Marriott, A. W. Nicholls, M. D. Reily, J. J. Thaden, and M. R. Viant, "Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)." *Metabolomics : Official journal of the Metabolomic Society*, vol. 3, no. 3, pp. 211–221, sep 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24039616

[65] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG." *Nucleic acids research*, vol. 42, no. Database issue, pp. D199–205, jan 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24214961

[66] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant, "PubChem Substance and Compound databases." *Nucleic acids research*, vol. 44, no. D1, pp. D1202–13, jan 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26400175

[67] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu,

F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, and A. Scalbert, "HMDB 3.0–The Human Metabolome Database in 2013." *Nucleic acids research*, vol. 41, no. Database issue, pp. D801–7, jan 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23161693

[68] M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, and S. Subramaniam, "LMSD: LIPID MAPS structure database," *Nucleic Acids Research*, vol. 35, no. Database, pp. D527–D532, jan 2007. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl838

[69] T. Kind and O. Fiehn, "Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm," *BMC Bioinformatics*, vol. 7, no. 1, p. 234, 2006. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-234

[70] D. J. Creek, A. Jankevics, R. Breitling, D. G. Watson, M. P. Barrett, and K. E. V. Burgess, "Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid ChromatographyMass Spectrometry: Improved Metabolite Identification by Retention Time Prediction," *Analytical Chemistry*, vol. 83, no. 22, pp. 8703–8710, nov 2011. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ac2021823

[71] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka, "MassBank: a public repository for sharing mass spectral data for life sciences," *Journal of Mass Spectrometry*, vol. 45, no. 7, pp. 703–714, jul 2010. [Online]. Available: http://doi.wiley.com/10.1002/jms.1777

[72] H. E. Pence and A. Williams, "ChemSpider: An Online Chemical Information Resource," *Journal of Chemical Education*, vol. 87, no. 11, pp. 1123–1124, nov 2010. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ed100697w

[73] F. Fern Andez-Albert, R. Llorach, M. Garcia-Aloy, A. Ziyatdinov, C. Andres-Lacueva, and A. Perera, "Intensity drift removal in LC/MS metabolomics by common variance compensation," vol. 30, no. 20, pp. 2899–2905, 2014.

[74] R. Wehrens, J. A. Hageman, F. van Eeuwijk, R. Kooke, P. J. Flood, E. Wijnker, J. J. B. Keurentjes, A. Lommen, H. D. L. M. van Eekelen, R. D. Hall, R. Mumm, and R. C. H. de Vos, "Improved batch correction in untargeted MS-based metabolomics."

Metabolomics : Official journal of the Metabolomic Society, vol. 12, p. 88, 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27073351

[75] B. A. Ejigu, D. Valkenborg, G. Baggerman, M. Vanaerschot, E. Witters, J.-C. Dujardin, T. Burzykowski, and M. Berg, "Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments." *Omics : a journal of integrative biology*, vol. 17, no. 9, pp. 473–85, sep 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23808607

[76] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[77] D. May, W. Law, M. Fitzgibbon, Q. Fang, and M. McIntosh, "Software platform for rapidly creating computational tools for mass spectrometry-based proteomics." *Journal of proteome research*, vol. 8, no. 6, pp. 3212–7, jun 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2696634

[78] D. J. Creek, A. Jankevics, K. E. V. Burgess, R. Breitling, and M. P. Barrett, "IDEOM: an Excel interface for analysis of LC-MS-based metabolomics data." *Bioinformatics (Oxford, England)*, vol. 28, no. 7, pp. 1048–9, apr 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22308147

[79] M. F. Clasquin, E. Melamud, J. D. Rabinowitz, M. F. Clasquin, E. Melamud, and J. D. Rabinowitz, "LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine." John Wiley & Sons, Inc., mar 2012, pp. 14.11.1–14.11.23. [Online]. Available: http://doi.wiley.com/10.1002/0471250953.bi1411s37

[80] R. Tautenhahn, G. J. Patti, D. Rinehart, and G. Siuzdak, "XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data," *Analytical Chemistry*, vol. 84, no. 11, pp. 5035–5039, jun 2012. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ac300698c

[81] F. Giacomoni, G. Le Corguillé, M. Monsoor, M. Landi, P. Pericard, M. Pétéra, C. Duperier, M. Tremblay-Franco, J.-F. Martin, D. Jacob, S. Goulitquer, E. A. Thévenot, and C. Caron, "Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics." *Bioinformatics (Oxford, England)*, vol. 31, no. 9, pp. 1493–5, may 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25527831

[82] J. Xia, R. Mandal, I. V. Sinelnikov, D. Broadhurst, and D. S. Wishart, "MetaboAnalyst 2.0–a comprehensive server for metabolomic data analysis," *Nucleic*

Acids Research, vol. 40, no. W1, pp. W127–W133, jul 2012. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks374

[83] H. Gowda, J. Ivanisevic, C. H. Johnson, M. E. Kurczy, H. P. Benton, D. Rinehart, T. Nguyen, J. Ray, J. Kuehl, B. Arevalo, P. D. Westenskow, J. Wang, A. P. Arkin, A. M. Deutschbauer, G. J. Patti, and G. Siuzdak, "Interactive XCMS Online: Simplifying Advanced Metabolomic Data Processing and Subsequent Statistical Analyses," *Analytical Chemistry*, vol. 86, no. 14, pp. 6931–6939, jul 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24934772

[84] H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert, and O. Kohlbacher, "OpenMS: a flexible open-source software platform for mass spectrometry data analysis," *Nature Methods*, vol. 13, no. 9, pp. 741–748, aug 2016. [Online]. Available: http://www.nature.com/doifinder/10.1038/nmeth.3959

[85] J. Xia and D. S. Wishart, "MetPA: a web-based metabolomics tool for pathway analysis and visualization," *Bioinformatics*, vol. 26, no. 18, pp. 2342–2344, sep 2010. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq418

[86] T. Yamada, I. Letunic, S. Okuda, M. Kanehisa, and P. Bork, "iPath2.0: interactive pathway explorer." *Nucleic acids research*, vol. 39, no. Web Server issue, pp. W412–5, jul 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21546551

[87] D. P. Leader, K. Burgess, D. Creek, and M. P. Barrett, "Pathos: A web facility that uses metabolic maps to display experimental changes in metabolites identified by mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 25, no. 22, pp. 3422–3426, nov 2011. [Online]. Available: http://doi.wiley.com/10.1002/rcm.5245

[88] L. Cottret, D. Wildridge, F. Vinson, M. P. Barrett, H. Charles, M.-F. Sagot, and F. Jourdan, "MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks." *Nucleic acids research*, vol. 38, no. Web Server issue, pp. W132–7, jul 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20444866

[89] M. Kutmon, M. P. van Iersel, A. Bohler, T. Kelder, N. Nunes, A. R. Pico, and C. T. Evelo, "PathVisio 3: An Extendable Pathway Analysis Toolbox," *PLOS Computational Biology*, vol. 11, no. 2, p. e1004085, feb 2015. [Online]. Available: http://dx.plos.org/10.1371/journal.pcbi.1004085

[90] F. Jourdan, R. Breitling, M. P. Barrett, and D. Gilbert, "MetaNetter: inference and visualization of high-resolution metabolomic networks," *Bioinformatics*, vol. 24, no. 1, pp. 143–145, jan 2008. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm536

[91] A. Karnovsky, T. Weymouth, T. Hull, V. G. Tarcea, G. Scardoni, C. Laudanna, M. A. Sartor, K. A. Stringer, H. V. Jagadish, C. Burant, B. Athey, and G. S. Omenn, "Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data." *Bioinformatics (Oxford, England)*, vol. 28, no. 3, pp. 373–80, feb 2012. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr661

[92] M. Fourment and M. R. Gillings, "A comparison of common programming languages used in bioinformatics," *BMC Bioinformatics*, vol. 9, no. 1, p. 82, 2008. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-82

[93] MATLAB, *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc., 2010.

[94] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[95] Shiny. [Online]. Available: https://shiny.rstudio.com/

[96] Ruby on Rails. [Online]. Available: http://rubyonrails.org/

[97] Django Software Foundation. [Online]. Available: https://www.djangoproject.com/

[98] A. H. Y. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Ménard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A.-M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G. W. Brown, B. Andrews, H. Bussey, and C. Boone, "Global Mapping of the Yeast Genetic Interaction Network," *Science*, vol. 303, no. 5659, 2004.

[99] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules," *Science*, vol. 302, no. 5643, 2003.

[100] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 10, pp. 770–780, oct 2008. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrm2503

[101] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, may 2000. [Online]. Available: http://www.nature.com/doifinder/10.1038/75556

[102] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson, "Global reconstruction of the human metabolic network based on genomic and bibliomic data." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 6, pp. 1777–82, feb 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17267599

[103] J. Förster, I. Famili, P. Fu, B. Ø. Palsson, and J. Nielsen, "Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network." *Genome research*, vol. 13, no. 2, pp. 244–53, feb 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12566402

[104] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?" *Nature Biotechnology*, vol. 28, no. 3, pp. 245–248, mar 2010. [Online]. Available: http://www.nature.com/doifinder/10.1038/nbt.1614

[105] N. D. Price, J. A. Papin, C. H. Schilling, and B. O. Palsson, "Genome-scale microbial in silico models: the constraints-based approach," *Trends in Biotechnology*, vol. 21, no. 4, pp. 162–169, 2003.

[106] I. Thiele, N. Swainston, R. M. T. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, S. G. Thorleifsson, R. Agren, C. Bölling, S. Bordel, A. K. Chavali, P. Dobson, W. B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novère, N. Malys, A. Mazein, J. A. Papin, N. D. Price, E. Selkov, M. I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. H. G. M. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V. Westerhoff, D. B. Kell, P. Mendes, and B. Ø. Palsson, "A community-driven global reconstruction of human metabolism," *Nature Biotechnology*, vol. 31, no. 5, pp. 419–425, mar 2013. [Online]. Available: http://www.nature.com/doifinder/10.1038/nbt.2488

[107] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, F. Yergeau, and J. Cowan, "Extensible Markup Language (XML)," 1998. [Online]. Available: http://www.w3pdf.com/W3cSpec/XML/2/REC-xml11-20060816.pdf

[108] R. Ihaka and R. Gentleman, "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, sep 1996. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10618600.1996.10474713

[109] Celery: Distributed Task Queue. [Online]. Available: http://www.celeryproject.org/

[110] RabbitMQ, Pivotal. [Online]. Available: https://www.rabbitmq.com/

[111] Python Software Foundation. [Online]. Available: https://www.python.org/

[112] MySQL, Oracle. [Online]. Available: https://www.mysql.com/

[113] Nginx, Inc. [Online]. Available: https://www.nginx.com/

[114] Gunicorn. [Online]. Available: http://gunicorn.org/

[115] NumPy. [Online]. Available: http://www.numpy.org/

[116] SciPy. [Online]. Available: https://www.scipy.org/

[117] Scikit-learn. [Online]. Available: http://scikit-learn.org/stable/

[118] rpy2. [Online]. Available: https://rpy2.readthedocs.io/en/version_2.8.x/

[119] S. Beisken, M. Eiden, and R. M. Salek, "Getting the right answers: understanding metabolomics challenges," *Expert Review of Molecular Diagnostics*, vol. 15, no. 1, pp. 97–109, jan 2015. [Online]. Available: http://www.tandfonline.com/doi/full/10.1586/14737159.2015.974562

[120] M. J. Gibney, M. Walsh, L. Brennan, H. M. Roche, B. German, and B. van Ommen, "Metabolomics in human nutrition: opportunities and challenges." *The American journal of clinical nutrition*, vol. 82, no. 3, pp. 497–503, sep 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16155259

[121] I. M. Vincent, R. Daly, B. Courtioux, A. M. Cattanach, S. Biéler, J. M. Ndung'u, S. Bisser, and M. P. Barrett, "Metabolomics Identifies Multiple Candidate Biomarkers to Diagnose and Stage Human African Trypanosomiasis," *PLOS Neglected Tropical Diseases*, vol. 10, no. 12, p. e0005140, dec 2016. [Online]. Available: http://dx.plos.org/10.1371/journal.pntd.0005140

[122] L. H. Stipetic, M. J. Dalby, R. L. Davies, F. R. Morton, G. Ramage, and K. E. V. Burgess, "A novel metabolomic approach used for the comparison of Staphylococcus aureus planktonic cells and biofilm samples," *Metabolomics*, vol. 12, no. 4, p. 75, apr 2016. [Online]. Available: http://link.springer.com/10.1007/s11306-016-1002-0

[123] Y. Gloaguen, F. Morton, R. Daly, R. Gurden, S. Rogers, J. Wandy, D. Wilson, M. Barrett, and K. Burgess, "PiMP my metabolome: An integrated, web-based tool for LC-MS metabolomics data," *Bioinformatics*, aug 2017. [Online]. Available: http://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btx499/4082268/PiMP-my-metabolome-An-integrated-webbased-tool-for

[124] G. J. Patti, O. Yanes, and G. Siuzdak, "Innovation: Metabolomics: the apogee of the omics trilogy," *Nature Reviews Molecular Cell Biology*, vol. 13, no. 4, pp. 263–269, mar 2012. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrm3314

[125] R. M. Salek, C. Steinbeck, M. R. Viant, R. Goodacre, and W. B. Dunn, "The role of reporting standards for metabolite annotation and identification in metabolomic studies." *GigaScience*, vol. 2, no. 1, p. 13, oct 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24131531

[126] T. J. Parr, "Enforcing strict model-view separation in template engines," in *Proceedings of the 13th international conference on World Wide Web*. ACM, 2004, pp. 224–233.

[127] E. J. O'Neil and E. J., "Object/relational mapping 2008," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data - SIGMOD '08*. New York, New York, USA: ACM Press, 2008, p. 1351. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1376616.1376773

[128] R. M. Salek, S. Neumann, D. Schober, J. Hummel, K. Billiau, J. Kopka, E. Correa, T. Reijmers, A. Rosato, L. Tenori, P. Turano, S. Marin, C. Deborde, D. Jacob, D. Rolin, B. Dartigues, P. Conesa, K. Haug, P. Rocca-Serra, S. O'Hagan, J. Hao, M. van Vliet, M. Sysi-Aho, C. Ludwig, J. Bouwman, M. Cascante, T. Ebbels, J. L. Griffin, A. Moing, M. Nikolski, M. Oresic, S.-A. Sansone, M. R. Viant, R. Goodacre, U. L. Günther, T. Hankemeier, C. Luchinat, D. Walther, and C. Steinbeck, "COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access," *Metabolomics*, vol. 11, no. 6, pp. 1587–1597, dec 2015. [Online]. Available: http://link.springer.com/10.1007/s11306-015-0810-y

[129] L. H. Stipetic, G. Hamilton, M. J. Dalby, R. L. Davies, R. M. D. Meek, G. Ramage, D. G. E. Smith, and K. E. V. Burgess, "Draft Genome Sequence of Isolate Staphylococcus aureus LHSKBClinical, Isolated from an Infected

Hip," *Genome Announcements*, vol. 3, no. 2, apr 2015. [Online]. Available: http://genomea.asm.org/content/3/2/e00336-15.abstract

[130] A. Alonso, S. Marsal, and A. JuliÃ, "Analytical Methods in Untargeted Metabolomics: State of the Art in 2015," *Frontiers in Bioengineering and Biotechnology*, vol. 3, p. 23, mar 2015. [Online]. Available: http://www.frontiersin.org/Bioinformatics{_}and{_}Computational{_}Biology/10.3389/fbioe.2015.00023/abstract

[131] R. Goodacre, D. Broadhurst, A. K. Smilde, B. S. Kristal, J. D. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D. B. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjöström, J. Trygg, and F. Wulfert, "Proposed minimum reporting standards for data analysis in metabolomics," *Metabolomics*, vol. 3, no. 3, pp. 231–241, sep 2007. [Online]. Available: http://link.springer.com/10.1007/s11306-007-0081-3

[132] B. P. Bowen and T. R. Northen, "Dealing with the Unknown: Metabolomics and Metabolite Atlases," *Journal of the American Society for Mass Spectrometry*, vol. 21, no. 9, pp. 1471–1476, 2010.

[133] J. J. J. van der Hooft, J. Vervoort, R. J. Bino, and R. C. H. de Vos, "Spectral trees as a robust annotation tool in LCMS based metabolomics," *Metabolomics*, vol. 8, no. 4, pp. 691–703, aug 2012. [Online]. Available: http://link.springer.com/10.1007/s11306-011-0363-7

[134] D. B. Kell, "Metabolomics and systems biology: making sense of the soup," *Current Opinion in Microbiology*, vol. 7, no. 3, pp. 296–307, 2004.

[135] K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. Gonzalez-Beltran, S.-A. Sansone, J. L. Griffin, and C. Steinbeck, "MetaboLights–an open-access general-purpose repository for metabolomics studies and associated metadata," *Nucleic Acids Research*, vol. 41, no. D1, pp. D781–D786, jan 2013. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks1004

[136] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. van Schaik, and J. Vervoort, "Substructure-based annotation of high-resolution multistage MS^n spectral trees," *Rapid Communications in Mass Spectrometry*, vol. 26, no. 20, pp. 2461–2471, oct 2012. [Online]. Available: http://doi.wiley.com/10.1002/rcm.6364

[137] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. J. Bino, and J. Vervoort, "Automatic Chemical Structure Annotation of an LCMS ^{n} Based

Metabolic Profile from Green Tea," *Analytical Chemistry*, vol. 85, no. 12, pp. 6033–6040, jun 2013. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ac400861a

[138] D. Auber, D. Archambault, R. Bourqui, A. Lambert, M. Mathiaut, P. Mary, M. Delest, J. Dubois, and G. Melançon, "The Tulip 3 Framework: A Scalable Software Library for Information Visualization Applications Based on Relational Data," p. 31, 2012.

[139] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome research*, vol. 13, no. 11, pp. 2498–504, nov 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14597658

[140] B. Merlet, N. Paulhe, F. Vinson, C. Frainay, M. Chazalviel, N. Poupin, Y. Gloaguen, F. Giacomoni, and F. Jourdan, "A Computational Solution to Automatically Map Metabolite Libraries in the Context of Genome Scale Metabolic Networks," *Frontiers in Molecular Biosciences*, vol. 3, p. 2, feb 2016. [Online]. Available: http://journal.frontiersin.org/Article/10.3389/fmolb.2016.00002/abstract

[141] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck, "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013." *Nucleic acids research*, vol. 41, no. Database issue, pp. D456–63, jan 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23180789

[142] M. Vogeser and K. Parhofer, "Liquid Chromatography Tandem-mass Spectrometry (LC-MS/MS) - Technique and Applications in Endocrinology," *Experimental and Clinical Endocrinology & Diabetes*, vol. 115, no. 9, pp. 559–570, oct 2007. [Online]. Available: http://www.thieme-connect.de/DOI/DOI?10.1055/s-2007-981458

[143] A. Doerr, "DIA mass spectrometry," *Nature Methods*, vol. 12, no. 1, pp. 35–35, dec 2014. [Online]. Available: http://www.nature.com/doifinder/10.1038/nmeth.3234

[144] D. Merkel, "Docker: lightweight Linux containers for consistent development and deployment," p. 2, 2014. [Online]. Available: http://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment

[145] M. Bostock, V. Ogievetsky, and J. Heer, "D Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, dec 2011. [Online]. Available: http://ieeexplore.ieee.org/document/6064996/

[146] M. Scholz and O. Fiehn, "SetupX - a public study design database for metabolomic projects; Pac Symp Biocomput 2007."

[147] Adam J Carrol, Murray R Badger, and A Harvey Millar, "The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets," *BMC Bioinformatics*, vol. 11, no. 376, dec 2010.

[148] H. Ferry-Dumazet, L. Gil, C. Deborde, A. Moing, S. Bernillon, R. Dominique, M. Nikolski, A. de Daruvar, and D. Jacob, "MeRy-B: a web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles," *BMC Plant Biology*, vol. 11, no. 104, pp. 1471–2229, jun 2011.

[149] "The metabolomics workbench," http://www.metabolomicsworkbench.org/.

[150] J. J. J. van der Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess, and S. Rogers, "Topic modeling for untargeted substructure exploration in metabolomics." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 48, pp. 13 738–13 743, nov 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27856765

[151] S. A. Grando, "Connections of nicotine to cancer," *Nature Reviews Cancer*, vol. 14, no. 6, pp. 419–429, may 2014. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrc3725

[152] H. M. Schuller, "Is cancer triggered by altered signalling of nicotinic acetylcholine receptors?" *Nature Reviews Cancer*, vol. 9, no. 3, pp. 195–205, mar 2009. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrc2590

[153] M. Alkondon, E. F. R. Pereira, W. S. Cartes, A. Maelicke, and E. X. Albuquerque, "Choline is a Selective Agonist of $\alpha 7$ Nicotinic Acetylcholine Receptors in the Rat Brain Neurons," *European Journal of Neuroscience*, vol. 9, no. 12, pp. 2734–2742, dec 1997. [Online]. Available: http://doi.wiley.com/10.1111/j.1460-9568.1997.tb01702.x

[154] K. Urano, Y. Kurihara, M. Seki, and K. Shinozaki, "Omics' analyses of regulatory networks in plant abiotic stress responses," *Current Opinion in Plant Biology*, vol. 13, no. 2, pp. 132–138, 2010.

[155] V. Canuel, B. Rance, P. Avillach, P. Degoulet, and A. Burgun, "Translational research platforms integrating clinical and omics data: a review of publicly available solutions," *Briefings in Bioinformatics*, vol. 16, no. 2, pp. 280–290, mar 2015. [Online]. Available: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbu006

[156] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, "A Sparse PLS for Variable Selection when Integrating Omics Data," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, jan 2008. [Online].

Available: http://www.degruyter.com/view/j/sagmb.2008.7.1/sagmb.2008.7.1.1390/sagmb.2008.7.1.1390.xml

[157] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér, "Data integration in the era of omics: current and future challenges," *BMC Systems Biology*, vol. 8, no. Suppl 2, p. I1, 2014. [Online]. Available: http://www.biomedcentral.com/1752-0509/8/S2/I1

[158] B. Palsson and K. Zengler, "The challenges of integrating multi-omic data sets," *Nature Chemical Biology*, vol. 6, no. 11, pp. 787–789, nov 2010. [Online]. Available: http://www.nature.com/doifinder/10.1038/nchembio.462

[159] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, no. 1, pp. pp. 10–12, 2011.

[160] N. Joshi and J. Fass, "Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]," https://github.com/najoshi/sickle, 2011.

[161] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT: a fast spliced aligner with low memory requirements," *Nature Methods*, vol. 12, no. 4, pp. 357–360, mar 2015. [Online]. Available: http://www.nature.com/doifinder/10.1038/nmeth.3317

[162] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with RNA-seq," *Nature Biotechnology*, vol. 31, no. 1, pp. 46–53, dec 2012. [Online]. Available: http://www.nature.com/doifinder/10.1038/nbt.2450

[163] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." *Nucleic acids research*, vol. 38, no. 6, pp. 1767–71, apr 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20015970

[164] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris, "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research*, vol. 38, no. Web Server, pp. W214–W220, jul 2010. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq537

[165] D. W. Huang, B. T. Sherman, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki, "DAVID gene ID conversion tool." *Bioinformation*, vol. 2,

no. 10, pp. 428–30, jul 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2561161

[166] B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. Fernandez Banet, K. Billis, C. García Girón, T. Hourlier, K. Howe, A. Kähäri, F. Kokocinski, F. J. Martin, D. N. Murphy, R. Nag, M. Ruffier, M. Schuster, Y. A. Tang, J.-H. Vogel, S. White, A. Zadissa, P. Flicek, and S. M. J. Searle, "The Ensembl gene annotation system," *Database*, vol. 2016, p. baw093, jun 2016. [Online]. Available: https://academic.oup.com/database/article-lookup/doi/10.1093/database/baw093

[167] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI." *Nucleic acids research*, vol. 33, no. Database issue, pp. D54–8, jan 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15608257

[168] S. Tonstad and J. L. Cowan, "C-reactive protein as a predictor of disease in smokers and former smokers: a review." *International journal of clinical practice*, vol. 63, no. 11, pp. 1634–41, nov 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19732183

[169] S. Aldaham, J. A. Foote, H.-H. S. Chow, and I. A. Hakim, "Smoking Status Effect on Inflammatory Markers in a Randomized Trial of Current and Former Heavy Smokers," *International Journal of Inflammation*, vol. 2015, pp. 1–6, 2015. [Online]. Available: http://www.hindawi.com/journals/iji/2015/439396/

[170] M. G. Vander Heiden, "Targeting cancer metabolism: a therapeutic window opens," *Nature Reviews Drug Discovery*, vol. 10, no. 9, pp. 671–684, aug 2011. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrd3504

[171] B. C. Fuchs and B. P. Bode, "Amino acid transporters ASCT2 and LAT1 in cancer: Partners in crime?" *Seminars in Cancer Biology*, vol. 15, no. 4, pp. 254–266, 2005.

[172] R. Cavill, D. Jennen, J. Kleinjans, and J. J. Briede, "Transcriptomic and metabolomic data integration," *Briefings in Bioinformatics*, vol. 17, no. 5, pp. 891–901, 2016.

[173] B. B. Misra and J. J. J. van der Hooft, "Updates in metabolomics tools and resources: 2014-2015," *ELECTROPHORESIS*, vol. 37, no. 1, pp. 86–110, jan 2016. [Online]. Available: http://doi.wiley.com/10.1002/elps.201500417

[174] R. J. M. Weber, T. N. Lawson, R. M. Salek, T. M. D. Ebbels, R. C. Glen, R. Goodacre, J. L. Griffin, K. Haug, A. Koulman, P. Moreno, M. Ralser, C. Steinbeck, W. B. Dunn, and M. R. Viant, "Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy."

Metabolomics : Official journal of the Metabolomic Society, vol. 13, no. 2, p. 12, 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28090198