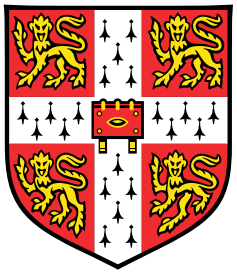
Fit for purpose? A metascientific analysis of metabolomics data in public repositories



Rachel Ann Spicer

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Churchill College July 2018

Declaration

I hereby declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation contains fewer than 60,000 words excluding bibliography, figures, appendices etc.

Rachel Ann Spicer
July 2018

Acknowledgements

Firstly I wish to thank my primary supervisor, Prof. Dr. Christoph Steinbeck for supporting me throughout my PhD. I would also like to thank Dr. Andrew Leach and Dr. Nick Goldman for taking over the administrative side of supervising me after Chris moved to Jena.

Thank you to my thesis advisory committee members Prof. Dame Janet Thornton, Dr. Theodore Alexandrov and Prof. Jules Griffin for their advice and guidance. My special thanks goes to Prof. Jules Griffin for collaborating with me.

Thank you also to everyone in Jules's group who helped me during my time at the Medical Research Council - Human Nutrition Research (MRC-HNR) institute. In particular I wish to thank Dr. Larissa Richardson and Dr. James Smith for their help and guidance during my time at the MRC-HNR.

I also wish to thank the members of the Steinbeck and MetaboLights teams. Their friendship and advice has helped guide me through my PhD. In particular thank you to Dr. Keeva Cochrane, Ken Haug, Dr. Namrata Kale, Dr. Pablo Moreno and Dr. Reza Salek for their helpful discussions and suggestions.

Finally, I wish to thank all my friends and family for their support throughout the PhD program. In particular I wish to thank my Mum, Sally Spicer, for proof reading this thesis.

Abstract

Metabolomics is the study of metabolites and metabolic processes. Due to the diversity of structures and polarities of metabolites, no single analytical technique is able to measure the entire metabolome — instead a varied set of experimental designs and instrumental technologies are used to measure specific portions. This has led to the development of many distinct data analysis and processing methods and software.

There is hope that metabolomics can be utilized for clinical applications, in toxicology and to measure the exposome. However, for these applications to be realised data must be high quality, sufficiently standardised and annotated, and FAIR (Findable, Accessible, Interoperable and Reproducible). For this purpose, it is also important that standardised, FAIR software workflows are available.

There has also recently been much concern over the reproducibility of scientific research, which FAIR and open data, and workflows can help to address. To this end, this thesis aims to assess current practices and standards of sharing data within the field of metabolomics, using metascientific approaches. The types of functions of software for processing and analysing metabolomics data is also assessed.

Reporting standards are designed to ensure that the minimum information required to understand and interpret the results of analysis are reported. However, poor reporting standards are ignored and not complied with. Compliance to the biological context Metabolomics Standards Initiative (MSI) guidelines was examined, in order to investigate their timeliness.

The state of open data within the metabolomics community was examined by investigating how much publicly available metabolomics data there is and where has it been deposited. To explore whether journal data sharing policies are driving open metabolomics data, which journals publish articles that have their underlying data made open was also examined. However, open data alone is not inherently useful: if data is incomplete, lacking in quality or missing crucial metadata, it is not valuable. Conversely, if data are reused, this can demonstrate the worth of public data archiving. Levels of reuse of public metabolomics data were therefore examined.

With greater than 250 software tools specific for metabolomics, practitioners are faced with a daunting task to select the best tools for data collection and analysis. To help educate researchers about what software is available, a taxonomy of metabolomics software tools and a GitHub pages wiki, which provides extensive details about all included software, have been developed.

Table of contents

List of figures

List of tables

Nomenclature

ANOVA	ANalysis Of VAriance
APCI	Atmospheric-Pressure Chemical Ionization
ArMet	Architecture for Metabolomics
BLAST	Basic Local Alignment Search Tool
BMI	Body Mass Index
BML-NMR	Birmingham Metabolite Library
BMRB	Biological Magnetic Resonance Bank
CASMI	Critical Assessment of Small Molecule Identification
CAWG	Chemical Analysis Working Group
CE-MS	Capillary Electrophoresis-Mass Spectrometry
ChEBI	Chemical Entities of Biological Interest
CI	Chemical Ionization
CID	Collision Induced Dissociations
COSMOS	COordination of Standards in MetabOlomicS
COSY	COrrelational SpectroscopY
CRE	Cloud Research Environment
DAC	Data Access Committee
DAD	Diode-Array Detector

dbGaP	Database of Genotypes and Phenotypes
DCI	Data Citation Index
DDI	Data Discovery Index
DDR	Dryad Digital Repository
DESI	Desorption Electrospray Ionization
DIMS	Direct Injection Mass Spectrometry
DNA	Deoxyribonucleic acid
DOI	Digital Object Identifier
DUO	Data Use Ontology
EBI	European Bioinformatics Institute
EGA	European Genome-phenome Archive
EI	Electron Ionization
ELSI	Ethical, Legal and Social Implications
EMBO	European Molecular Biology Organization
ESI	Electrospray Ionization
EU	European Union
FAIR	Findability, Accessibility, Interoperability, and Reusability
FAIR-TLC	Findability, Accessibility, Interoperability, and Reusability-Traceable, Licensed and Connected
FIA-MS	Flow Injection Analysis-Mass Spectrometry
FID	Free Induction Decay
FT-ICR	Fourier Transform-Ion Cyclotron Resonance
FT-IR	Fourier Transform–Infrared

GC-MS	Gas Chromatography-Mass Spectrometry
GCxGC-MS	Two-Dimensional Gas Chromatography-Mass Spectrometry
GNPS	Global Natural Products Social Molecular Networking
GPMDB	Global Proteome Machine and Database
HGP	Human Genome Project
HMDB	Human Metabolome Database
HPLC-MS	High Performance Liquid Chromatography-Mass Spectrometry
HSQC	Heteronuclear Single Quantum Coherence
HSQC-TCOSY	Heteronuclear Single Quantum Coherence-Total COrrelational SpectroscopY
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
HUPO	Human Proteome Organization
IMS-MS	Ion-Mobility Spectrometry-Mass Spectrometry
InChi	International Chemical Identifier
IR	Infrared
ISA	Investigation/Study/Assay
ISA-TAB	Investigation/Study/Assay-Tab Delimited
KEGG	Kyoto Encyclopedia of Genes and Genomes
kNN	k-Nearest Neighbors
LC-HRMS	Liquid Chromatography-High Resolution Mass Spectrometry
LC-MS	Liquid Chromatography-Mass Spectrometry
LIMS	Laboratory Information Management System
LINCS	Library of Integrated Network-based Cellular Signatures

m/z	Mass-to-Charge ratio
MALDI	Matrix-Assisted Laser Desorption Ionization
MAS	Metabolomics Activity Screening
MassIVE	Mass Spectrometry Interactive Virtual Environment
MEDLINE	Medical Literature Analysis and Retrieval System Online
MERIT	MEtabolomics standaRds Initiative in Toxicology
MeRy-B	Metabolomics Repository Bordeaux
MeSH	Medical Subject Headings
MI	Minimum Information
MIAME	Minimum Information About a Microarray Experiment
MIAMET	Minimum Information about a Metabolomics Experiment
MIAPE	Minimum Information About a Proteomics Experiment
MICE	Metabolite Identification Carbon Efficiency
MICE	Multivariate Imputation by Chained Equations
MIE	Metabolite Identification Efficiency
MIECO	Metabolite Identification Evidence Code Ontology
MII	Metabolite Identification Information
MOOSE	Meta-analysis Of Observational Studies in Epidemiology
MRM	Multiple Reaction Monitoring
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MS^n	Multi-Stage Mass Spectrometry
MSI	Metabolomics Standards Initiative

NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Effect Spectroscopy
NOMIS	Normalisation Factor For Each Individual Molecular Species
NSF	National Science Foundation
OmicsDI	Omics Discovery Index
OPLS	Orthogonal Partial Least Squares
OS	Operating System
PC-DFA	Principle Component-Discriminant Function Analysis
PCA	Principle Components Analysis
PhenoMeNal	Phenome and Metabolome aNalysis
PICO	Participants, Interventions, Comparisons and Outcomes
PLOS	Public Library of Science
PLS	Partial Least Squares (Projection to Latent Structures)
PLS-DA	Partial Least Squares-Discriminant Analysis
PMC	PubMed Central
PMCID	PubMed Central ID
PMID	PubMed Unique Identifier
PNAS	Proceedings of the National Academy of Sciences
PQN	Probabilistic Quotient Normalisation
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-analyses
PRISMA-P	Preferred Reporting Items for Systematic reviews and Meta-Analysis Protocols

PROSPERO	Prospective Register of Systematic Reviews
PSI	Protein Standards Initiative
Q-TOF	Quadrupole-Time-Of-Flight
QqQ	Triple Quadrupole
RefMet	Reference list of Metabolite names
ReSpect	RIKEN MSn spectral databases
REST	REpresentational State Transfer
RF	Random Forest
RI	Retention Indices
RRID	Research Resource Identifiers
RUV	Remove Unwanted Variation
SIM	Selected Ion Monitoring
SIMS	Secondary Ion Mass Spectrometry
SMILES	Simplified molecular-input line-entry system
SMPDB	Small Molecule Pathway Database
SMRS	Standard Metabolic Reporting Structure
SRM	Selected Reaction Monitoring
SVM	Support Vector Machine
T2D	Type 2 Diabetes
TCOSY	Total COrrelational SpectroscopY
tMICE	Topological Metabolite Identification Carbon Efficiency
TOF	Time-Of-Flight
UPLC-MS	Ultra Performance Liquid Chromatography-Mass Spectrometry

UV/VIS	Ultraviolet-Visible Spectroscopy
VRE	Virtual Research Environment
XML	Extensible Markup Language

Chapter 1

Introduction

1.1 Introduction to Metabolomics

Metabolomics has been described as the study of the entirety of the endogenous small molecules (metabolites) present within an organism, organ, biological tissue or cell [1]. The vast majority of metabolites are <1500 Daltons in size [2]. The metabolome is composed of both endogenous and exogenous metabolites such as contaminants, drug metabolites and xenobiotics. Metabolites have many functions including catalytic activity, signaling and as a source of energy.

The metabolome is highly dynamic, changing rapidly in response to environmental changes [3]. Many factors including age [4], alcohol intake [5], blood pressure [6], body composition [7], body mass index (BMI) [8, 9], gender [10, 11], nutrition [12, 13] and smoking status [14, 15], have been shown to have significant effects on an individual's metabolome (Figure 1.1). In contrast, the environment has very little effect on the genome [16] .

Whilst gene and protein expression are effected by environmental stressors, these changes are amplified in the metabolome. This means that the metabolome may be a more discriminating indicator of exposures than other omic techniques [17]. Additionally, it is also easier to correlate the metabolome with an individual's phenotype than either the transcriptome or the proteome [18]. The metabolome has thus been dubbed an "intermediate phenotype" [19]. However, this can be conceptually challenging, as the metabolome can be viewed as both a phenotype and as a part of the internal molecular environment.

As well as being affected by environmental exposures, there are also large genetic factor influences on the metabolome, with some genetic loci having 10-60% effect sizes per allele copy [20].

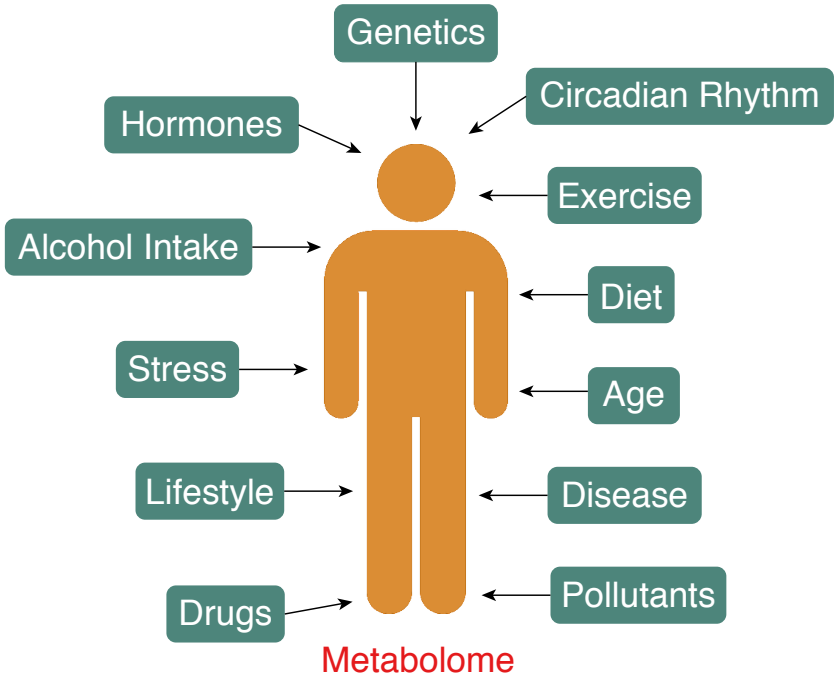


Fig. 1.1 Many factors have a significant effect on the metabolome.

Metabolites are highly structurally diverse and range in polarity from hydrophilic to hydrophobic. Many different classes of metabolites exist, e.g. amino acids, lipids, fatty acids and carbohydrates. Compared to other omics (genomics, transcriptomics and proteomics) the metabolome is thus more challenging to measure — no single analytic technique alone is able to measure the entire metabolome. Instead different, often complementary, analytical techniques are used to measure specific portions of the metabolome. The three most frequently used technologies are: liquid chromatography-mass spectrometry (LC-MS), gas chromatography-mass spectrometry (GC-MS) and nuclear magnetic resonance (NMR). Some alternative and less commonly used platforms include direct injection (DIMS), capillary electrophoresis- (CE-MS), flow injection analysis- (FIA-MS) and ion-mobility spectrometry-(IMS-MS) mass spectrometry, diode-array detector (DAD), Fourier transform–infrared (FT-IR) and Raman spectroscopy.

Mass spectrometry imaging is also becoming increasingly used to measure the spatial distribution of metabolites within tissues [23, 24]. Further analytical techniques are used for mass spectrometry imaging including secondary ion mass spectrometry (SIMS), matrix-assisted laser desorption ionization (MALDI) and desorption electrospray ionization (DESI).

In mass spectrometry samples are ionised and the ions are then sorted by their mass-to-charge ratio (m/z). NMR is based on the principle that nuclei have spin. When placed

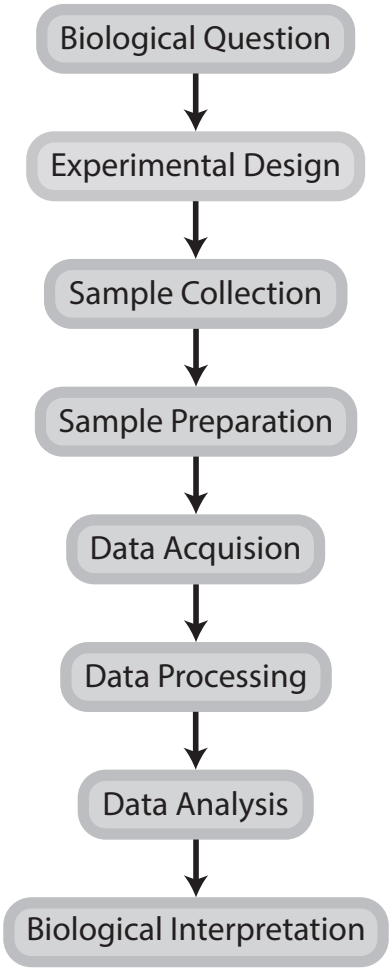


Fig. 1.2 A standard metabolomics workflow, from biological question to interpretation. Other examples of general metabolomics workflows include those presented in Koek *et al.* [21] and León *et al.* [22].

in a magnetic field, NMR active nuclei absorb electromagnetic radiation at a frequency characteristic of the isotope and their chemical environment. The resonant frequency, energy of the radiation absorbed, and the intensity of the signal are proportional to the strength of the magnetic field and their nuclear magnetic moments.

Both MS and NMR use a standard experimental workflow, shown in Figure 1.2. However, the sample preparation and data analysis required for each instrumental technology differs [25]. Both analytical techniques have their advantages: MS is highly sensitive, being able to detect features in the femtomolar range, whereas NMR is more specific and highly reproducible [25]. Whilst NMR is cheaper to run than MS, NMR systems cost more to

maintain and require a lot of space. NMR is a non-destructive technique, allowing samples to be reused, however it requires a greater amount of sample than MS.

There are a greater number of manufacturers of mass spectrometers than NMR spectrometers, and there are more different types of mass spectrometer. Mass spectrometers consist of three main parts: the ion source, mass analyzer and ion detector. Examples of ion source include electron ionization (EI), chemical ionization (CI) and electrospray ionization (ESI), and examples of mass analyzer include quadrupole, time-of-flight (TOF), ion trap and orbitrap. Different types of ion source and mass analyzer are better suited to coupling with different types of chromatography.

Compared to MS, NMR is a far more quantitative technique [25]. The signal intensity of a feature is directly proportional to the molar concentration of the molecule [26]. However, there can be limitations in resolution due to overlapping signals. This is especially problematic with complex mixtures, such as biofluids [27].

There are two main approaches for conducting metabolomics experiments: targeted and untargeted. In targeted approaches, the concentrations of only a small number of selected metabolites are measured. Targeted studies rely on the availability of authentic chemical standards of metabolites of interest, as quantification is performed through the use of internal standards. Untargeted approaches aim to maximise the number of metabolites detected. These are applicable for exploratory studies, as they provide the opportunity to observe both expected and unexpected changes and the ability to detect previously unknown metabolites.

1.2 Metabolomics Data Analysis and Processing Software

The diversity of metabolite classes means that no single analytical technique is able to measure the entire metabolome; rather each is able to measure a specific subset of the metabolome. The mostly widely used technologies are LC-MS, GC-MS, NMR and DIMS [28], but there are also many other technologies used by a smaller number of groups, such as FIA-MS, Raman spectroscopy and chromatography not coupled to mass spectrometry e.g. liquid chromatography-diode array detector (LC-DAD). The data produced by each analytical method requires distinct handling and thus different data analysis and processing tools and workflows. To date more than 250 different software tools specifically designed for the handling of metabolomics data have been published [29–31].

Analysis of metabolomics data typically involves data preprocessing, post-processing (also called pre-treatment), metabolite annotation or identification and statistical analysis. Some studies also include pathway enrichment analysis. However, the order these stages are

conducted in depends on the instrumental technology and approach used for data acquisition (Fig 1.3).

For example, statistical analysis is normally performed prior to metabolite identification in untargeted LC-MS studies as feature identification is challenging. There is low reproducibility between LC-MS measurements of different mass spectrometers due to varying combinations of ionization source and mass analyzer, resulting in shifts in peak intensity and m/z values [32]. It is therefore difficult to identify LC-MS features (including MS/MS), as spectra of the same metabolite acquired under different conditions may not have the same peaks or intensities. Identifying only statistically important metabolites, using a chemometric approach, reduces the amount of labour and time required for metabolite identification. For targeted approaches, metabolite identification does not pose the same problem as only the abundances of metabolites of interest are measured.

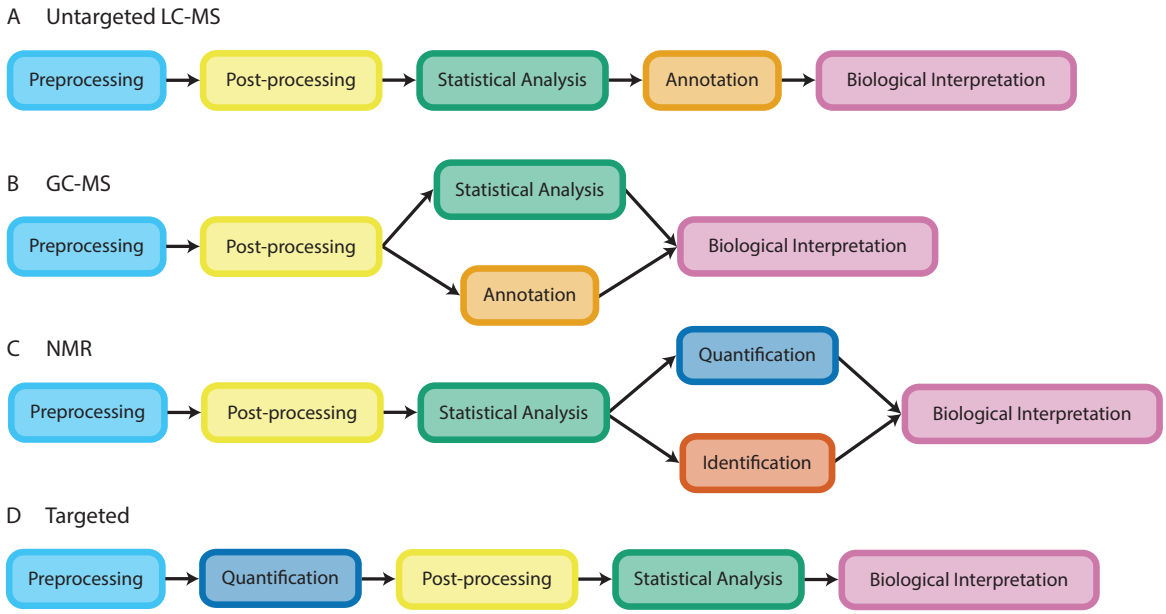


Fig. 1.3 Standard metabolomics data analysis workflows for: A) untargeted LC-MS, B) GC-MS, C) NMR and D) targeted studies.

GC-MS produces spectra that are more reproducible than LC-MS, and that do not have instrument-dependent fragmentation patterns [33]. This has allowed for the creation of a number of databases, including the commercial databases NIST (National Institute of Standards and Technology) v17 reference library[1] and FiehnLib [34], and the publicly available Golm Metabolome Database [35]. The relative ease of feature identification in

[1] https://www.nist.gov/srd/nist-standard-reference-database-1a-v17, accessed 25th March 2018.

GC-MS spectra means that typically metabolite identification is performed directly after data processing, in parallel to statistical analysis. More recently, alternative ionisation sources have been used: GC-ESI-MS and GC-TOF-MS [36], which require different identification methods to traditional GC-EI-MS. Due to the nature of the GC-MS workflow, many individual tools provide both preprocessing and metabolite identification.

In most NMR metabolomics studies, a chemometric approach is used [27]. Datasets are split into hundreds or thousands of buckets. Regions of interest are found using statistical analysis and are then later identified. Alternatively, in untargeted profiling approaches, signals from metabolites are identified and quantified prior to statistical analysis, but this approach is less prevalent.

As well as software that provides individual stages of data analysis, there are also workflows, which provide multiple interconnected tools, encompassing all stages of analysis. This allows users to perform the entirety of their analysis using a single tool and removes problems caused by a lack of interoperability between tools. Workflows also increase data processing and analysis reproducibility.

An example of a workflow provider for metabolomics is PhenoMeNal (Phenome and Metabolome aNalysis)[2]. PhenoMeNal is a H2020 funded e-infrastructure that provides data analysis workflows for clinical metabolomics via a Virtual Research Environment (VRE). It utilizes the Galaxy workflow management system [37], which enables researchers to share workflows, histories and datasets, either publicly or with specific individuals. As of 10th March 2018, the PhenoMeNal App library contains of 61 tools, not including submodules, covering a wide variety of experimental data types e.g. LC-MS, NMR, isotopically labeled data. Users can either design their own workflows or use pre-existing workflows.

1.3 The Importance and Applications of Metabolomics

Since the first occurrences of the term metabolomics, two decades ago, the field has grown substantially and is now beginning to mature [38]. Whilst it remains the smallest of the four main omics, the number of metabolomics research articles published is still growing year-on-year (Figure 1.4).

To date, practical applications of metabolomics that have been developed include the authentication of organic foods [39] and manuka honey [40] and the identification of metabolites that discriminate between different types of whiskeys [41] and wines [42, 43].

[2]https://phenomenal-h2020.eu/home/, accessed 10th March 2018

Metabolomics can be used to study human physiology under both normal and pathological conditions [44]. Metabolic phenotyping has been used in putative molecular biomarker discovery for numerous diseases e.g. rheumatoid arthritis [45, 46], colorectal cancer [47], Type II diabetes [48, 49] and chronic fatigue syndrome [50, 51].

There is also evidence that individual human metabolomes are unique, with multiple studies finding that individuals can be identified based on differences in their metabolic profiles [52, 53], even after a period of up to 7 years [54]. Ideally, metabolic phenotyping data has the potential to play a key role in the development of stratified/precision medicine.

Environmental and lifestyle influences play a critical role in the development of diseases such as cancer [55]; genetics accounts for only a minority of disease etiology for many chronic diseases, including many cancers [56–58]. For certain diseases, only a low amount of the heritability of the disease can be explained despite multiple genetic risk factors being identified. For example, whilst at least 36 genes have been identified as being associated with type 2 diabetes (T2D), only 10% of the heritability of T2D can be explained [59]. As additive genetic variation alone does not elucidate the whole picture in trait heritability, other sources of trait variance, such as epigenetics and exposure to the environment, must be investigated.

The exposome was first defined by Wild [60] as the totality of all "life-course environmental exposures (including lifestyle factors), from the prenatal period onwards". The metabolome has been suggested as a good proxy for measuring exposures and the exposome [55, 61], as it reports a snapshot of the actual physiological status of the cell at any given time [62] and is effected by both internal and external perturbations [63]. Examples of biomarkers of exposure for toxic and carcinogenic compounds that have been identified include: perfluorinated compounds [64], manganese [65] and jet fuel [66].

The metabolome is also correlated with the microbiome, as gut microbiota produce many important metabolites such as vitamins and hormones [67]. Fecal metabolomics can be used in complement to 16S sequencing to provide a functional readout of the activity of the microbiome. There is the potential for the metabolome to be used to elucidate mechanisms by which the microbiome affects health, such as the mechanisms by which the microbiome is associated with obesity [67]. The metabolites that can be detected in blood and feces also differ, so fecal metabolomics may also be used in accompany with blood in more studies.

Whilst animal testing has traditionally been seen as the "gold standard" of toxicology, there is ever increasing need to decrease the amount of it. The Three Rs (3Rs): Replacement, Reduction and Refinement are a framework for performing more ethical animal research, by improving the treatment of laboratory animals and the scientific quality [68]. Aside from the ethical issues of using animals in research, current models do not represent the best analogues

of humans and there is increasing regulation in some areas of the world that prevents animal experimentation. For example, in the European Union (EU) it is illegal to sell cosmetic products and ingredients that have been tested on animals inside the EU (under the directive 2003/15/EC). Of all the omics, metabolomics is the discipline that is most closely connected to classical toxicological endpoints [44]. There is hope that metabolomics can be utilised both for the elucidation of toxicants mechanisms of action and for safety assessments [69].

Metabolomics also has the potential to be used to identify phenotype altering metabolites that could be perturbed to intentionally modulate phenotype [70]. The relatively new technique metabolomics activity screening (MAS) has been used to identify metabolites that modulate a diverse array of biological processes including stem cell differentiation [71], innate immune response [72] and remyelination [73].

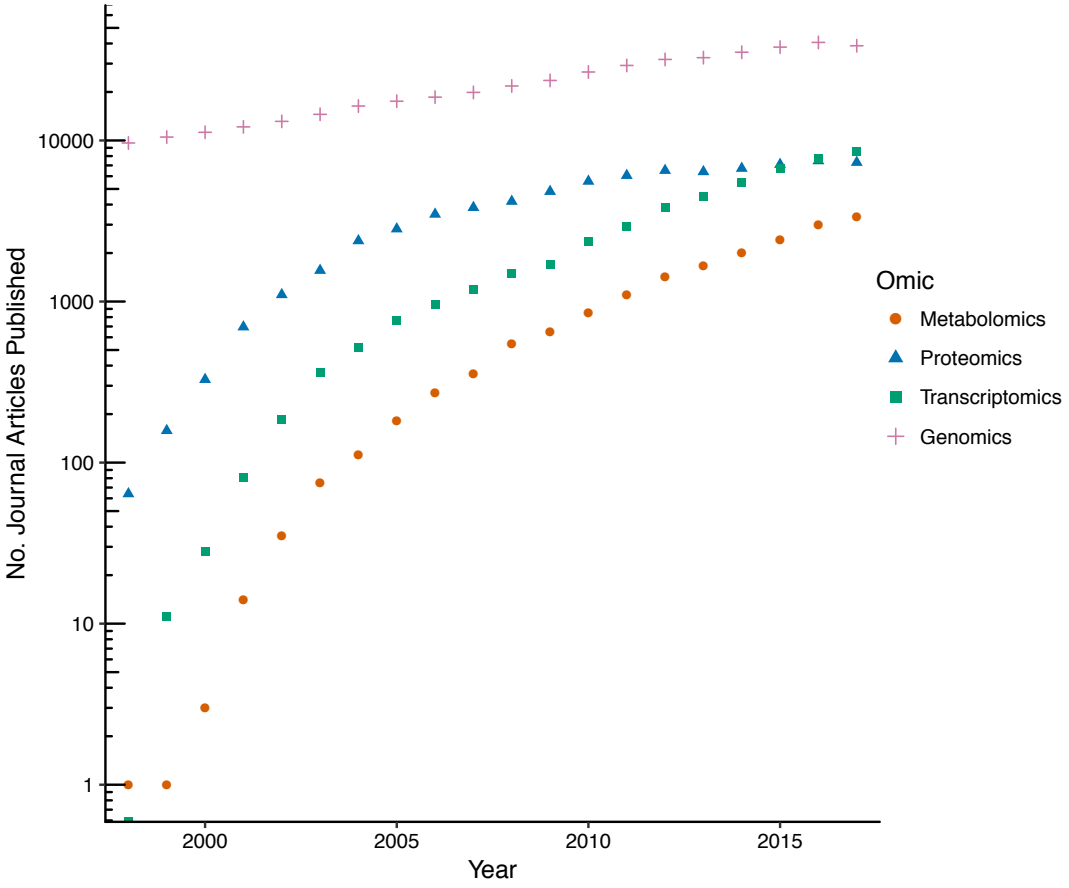


Fig. 1.4 The increase in the number of 'omic' publications since 1998 — the number of journal articles published per year. Metabolomics is shown in orange, proteomics is blue, transcriptomics is green and genomics is pink. Data were obtained from Web of Science by searching articles for 'omic*' OR 'ome*' for each omic i.e. DOCUMENT TYPES: (ARTICLE) AND 'metabolomic*' OR 'metabolome*' etc. Data were collected on the 12th January 2018. It must be noted that this is a very crude analysis, as the search did not include subsections of omics e.g. 'metabonomics', or methods used specifically to measure omics e.g. 'RNA-Seq' (Ribonucleic acid-Sequencing). These terms were not included in an attempt to make a fair comparison between omics, as it would be difficult to compile a list of all such terms for each of the omics. However, it is highly likely that many omics publications have been missed by not including these terms. It would be expected that the number of transcriptomics papers published per year would be much higher than the number of proteomics, as there are many more transcriptomics datasets than proteomics datasets.

1.4 Metascience

The term metascience was first defined in 1938 by C. W. Morris as "the science of science" [74]. Quantifiable scientific methods are applied in order to attempt to understand how current scientific practices affect the accuracy of scientific conclusions. Meta-science has its roots in the philosophy and history of science, however it differs from these disciplines as it requires the use of quantifiable scientific methods. Examples of metascientific research includes citation network analysis [75], examining biases in scientific publishing [76], detection of image manipulation [77] and reporting standards [78]. Systematic reviews and meta-analyses are also examples of metascientific study designs. Ioannidis *et al.* [78] propose catergorising meta-research into five areas: Methods, Reporting, Reproducibility, Evaluation, and Incentives.

Recently there has been much concern about the reproducibility of scientific research [79–81]. Two thirds of researchers are concerned about reproducibility [82] and >50% have been unable to replicate the findings of at least one study [83]. As worries about the lack of reproducibility of research in many fields have grown, metascientific methods have been increasingly used to investigate previously published research.

Large scale attempts to replicate multiple studies in psychology [79], oncology [80] and computer science [81] are additional examples of metascientific research. Metascience can also be used to check for errors in scientific papers: a project to fact check genetic studies by running the published sequence in BLAST, to ensure that the gene of interest is actually being studied, has lead to five retractions as of 15th September 2017 [84].

Another famous example of metascientific research is John Ioannidis's paper "Why Most Published Research Findings Are False" [85], which presents evidence that the conclusions of many scientific articles have insufficient evidence to support them. However, the findings of large studies and well-powered meta-analyses, have a higher probability of being true, and are as close to the unknown 'gold standard' of research certainty as possible.

The work presented in this thesis is one of the first attempts to apply meta-scientific principles to the field of metabolomics. Specifically the state of publicly available (open) metabolomics data is assessed, in order to guide future data sharing policies and reporting standards to maximize the value of the data.

1.5 The Open Science Movement, Open Data and Data Sharing

A systematic review recently developed a unified definition of open science [86]:

"Open science is transparent and accessible knowledge that is shared and developed through collaborative networks."

Open science is composed of many concepts that aim to make science more transparent and reproducible. These include open access, open data, open source and citizen science (Figure 1.5). For a detailed overview of open science concepts see the Open Science Taxonomy[3].

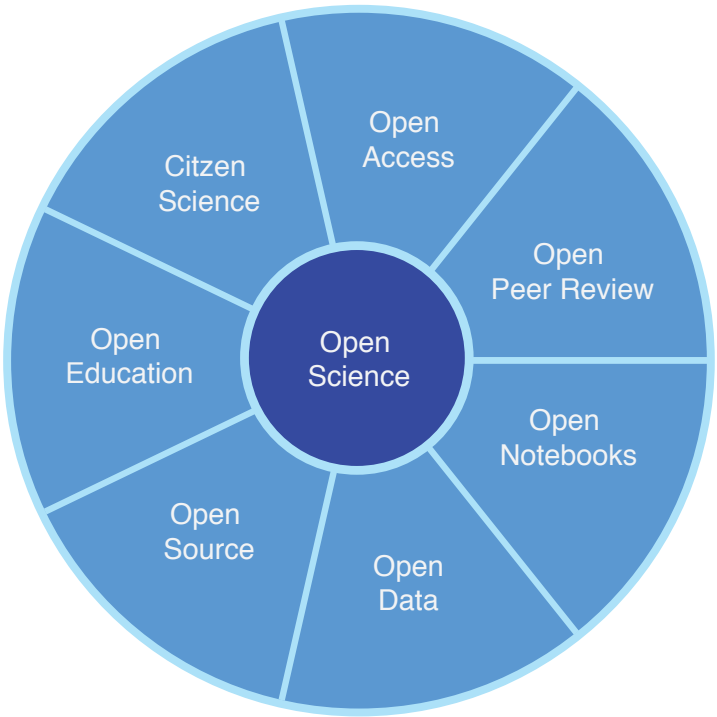


Fig. 1.5 Open Science is comprised of many different individual concepts.

Open data is a single component within the open science framework. It encompasses open administrative data, open government research data and open science research data. The Open Definition[4] defines open data as data that is "free to use, reuse, and redistribute — subject only, at most, to the requirement to attribute and/or share-alike". Under this definition,

[3] https://www.fosteropenscience.eu/foster-taxonomy/open-science, accessed 10th January 2018.
[4] https://opendefinition.org/, accessed 10th January 2018.

for data to be open, it must be downloadable without charge via the Internet. It must also be licensed to allow for reuse [87].

Data sharing refers to researchers making data they have generated available to other investigators. This can be done either by sharing data publicly, in an open manner, on a repository or institutional website, or by sharing data only with researchers who have directly requested it via a research proposal.

Individual researchers, the scientific community, funding bodies and the general public can all benefit from open data. Sharing data openly signals investigators' confidence in their research and encourages collaboration between researchers, which directly profits researchers. Studies that make data publicly available in a repository receive more citations than those without publicly available data [88]. Other scientists can reuse open scientific data for hypothesis generation [89], and to perform further and meta-analysis. The scientific community as a whole gains by improved error identification and reduced risk of fraud. Funding bodies benefit economically from open data, by receiving greater return on their investment [90]. Open data also improves reproducibility, and has been suggested as part of the solution to the replication crisis.

Open data can also help to prevent data loss: 36% of researchers have lost data, with hard drive failure being the most common cause [91]. There are also instances of data being irretrievably lost due to computer theft [92] and fires[5,6]. Data availability also declines with article age [93], with the odds of a dataset being extant falling by 17% per year. Public data repositories address these problems by providing long term storage solutions for data.

Much scientific research is financed by public funds. In 2015, £2.2 billion was invested in research by the UK government and research councils[7] [94], and the US spent $37.8 billion of federal funds on research[8]. Therefore, another advantage of sharing data openly is that it allows the findings of studies funded by the general public to be made directly available to them [95]. Science also profits by engaging the general public in citizen science to complete time consuming projects that cannot be easily automated. Examples of successful citizen science projects include Galaxy Zoo[9] where volunteers assist in the morphological

[5]https://www.theguardian.com/uk-news/2017/apr/28/manchester-christie-cancer-hospital-fire-research-equipment-destroyed, accessed 11th January 2018.

[6]https://www.timeshighereducation.com/news/fire-destroys-phd-work/165585.article, accessed 11th January 2018.

[7]UK gross domestic expenditure on research and development: 2015. https://bit.ly/2y9gs74, accessed 11th January 2018.

[8]https://www.sciencemag.org/news/2017/03/data-check-us-government-share-basic-research-funding-falls-below-50, accessed 11th January 2018.

[9]https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/, accessed 15th May 2018.

Table 1.1 Prominent publishing groups and journals that mandate data sharing. * indicates that a journal states that data sharing is required, but does not state that it is required for publication

Publishing Group/ Journal	Policy
eLife	https://elifesciences.org/inside-elife/6f32c567/supplementary-data
EMBO journal	https://emboj.embopress.org/authorguide#availabilityofpublishedmaterial
Nature Research journals	https://www.nature.com/authors/policies/availability.html
PeerJ	https://peerj.com/about/policies-and-procedures/#data-materials-sharing
PLOS	https://www.plos.org/editorial-publishing-policies
Royal Society journals	https://royalsociety.org/journals/ethics-policies/data-sharing-mining/
Science*	https://www.sciencemag.org/authors/science-journals-editorial-policies

classification of galaxies and Foldit[10], a protein folding game. As of March 2018, these projects have respectively generated 55 and 14 publications. These projects empower participants and increase self-efficacy.

The public sharing of omics data originates from the Bermuda summit of 1996, where members of the Human Genome Project (HGP) consortium agreed upon a set of principles (the "Bermuda Principles")[11] for the public release of DNA sequence data, within 24 hours of its generation. These principles have led to data sharing becoming standard practice within the genomics community [96].

Following pioneering policies such as the Bermuda principles, open data has been increasingly mandated by funding bodies and leading journals. Data sharing is now required or encouraged by >40% of journals [97], being a requirement for publication of 11.9%. Many prominent journals and publishing groups such as Nature Research journals, Public Library of Science (PLOS) and PeerJ now mandate data sharing (Table 1.1). Major research funders including the the European Commission's 8th framework programme "Horizon 2020" [98, 99], the National Institutes of Health (NIH) [100], the National Science Foundation (NSF)[12], Research Councils UK and Wellcome [101], all require that data generated as part of the research they fund be made open.

Despite the benefits of open data, many challenges still remain and must be addressed before data sharing becomes the norm across science. The largest concern of sharing data openly is the harm it can potentially cause. There are rightly concerns over patient privacy and the ethics of sharing public health data [102]. Researchers were able to re-identify

[10]https://fold.it/portal/, accessed 15th May 2018.

[11]https://web.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml, accessed 20th May 2018.

[12]https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf15051, accessed 18th May 2018.

participants of the 1,000 genome project, by combining metadata from multiple sources [103]. Publishing the location of endangered species can also cause more harm than good, as it can lead to increased wildlife poaching [104].

Researchers remain reluctant to share data as they do not receive credit for doing so. Currently open data is not widely considered by grant committees or for tenure. Sharing data via a repository can be very time consuming, and is not considered a priority to already overworked scientists. The lack of standardisation of data formats and of metadata reporting requirements can also hinder public data sharing. A lack of expertise in uploading data to repositories is a barrier to their use [105]. There is also concern about how open data will be used [87]. Researchers are concerned that their data will be misunderstood. Worryingly, the lower the *p*-value (the more statistically significant), the more willing researchers are to publicly share data [106] — hesitance to share data was linked to a greater number of errors reporting statistical results.

Changing researchers' data sharing practices remains a large challenge to widespread open data. Despite being proponents of open data, when surveyed the majority of authors of the minimum information about a proteomics experiment (MIAPE) reporting standards, reported to publishing at least one paper without the accompanying data in the two years following MIAPEs publication [107]. If the biggest supporters of open data do not publicly share their data then it cannot be expected that the rest of the community will do so. Despite the aforementioned obstacles, open data is becoming increasingly supported with 80% of researchers now agreeing that data should be made open as common practice [108].

Attempting to improve the situation there has recently been an increasing amount of research into how to incentivise scientists to share data. Suggestions have included awarding badges to articles that share their underlying data openly [109, 110], improving software infrastructure for data sharing [89] and data citations [111].

1.6 FAIR Principles and Reporting Standards

Alone, open data is insufficient to ensure that data analysis is reproducible and that data can be reused. Instead, sufficient annotation and metadata that describes both the data and how to access it are required. The principles: Findability, Accessibility, Interoperability, and Reusability (FAIR) (Table 1.2) [112] were developed by FORCE11, in order to guide data management and maximize the value of research data. Findable refers to the ability to find data, along with associated metadata, requiring that unique, persistent identifiers be assigned. Accessible means that data must be available, to either download or request access

to, and understandable. For a dataset to be interoperable, formal vocabularies and ontologies must be used to describe it, in order to allow for comparison and combination with other datasets. Reusable refers to the ability of humans and machines to reuse a dataset. There is overlap between the scopes of the four concepts. The FAIR principles apply to both open and restricted access data: data that are open are not necessarily FAIR, and FAIR data may not be open. Instead FAIR aims to ensure that data can be found, understood and reused.

The FAIR principles have been rapidly adopted. The term was first launched in 2014 and the principles were published in 2016 [112]. By 2017, 74% of surveyed researchers said they understand the FAIR concept [113]. However, making data FAIR can be challenging — there is not a strict definition as to what makes data FAIR, and many researchers find the concepts hard to implement in practice. Achieving FAIR data requires adequate reporting standards of data and metadata, suitable data repositories and detailed discipline specific guidelines.

One potential way of achieving FAIR data is the use of reporting standards [114]. Data reporting standards are designed to ensure that the minimum information required to understand and interpret the results of analysis are reported, and that the reporting of metadata is consistent across studies. Additionally, ontologies can be used to formally organise data [115]. Ontologies structure knowledge with rules that describe the relationship between terms, and can be used to formally classify metadata.

Various concepts have emerged that expand upon the original FAIR principles, with the aim of improving them or applying them to specific types of data. The most well-known is FAIR-TLC, which adds the principles Traceable, Licensed and Connected [116]. Traceability refers to data provenance and attribution of contributions, licensed requires that data are made available under an open, permissible license and connected refers to integration between data resources. There is also FAIR-Health for clinical data [117], which includes the additional principles quality and traceability, incentive schemes, and privacy regulation compliance. However, it is worth noting that some of these additional concepts are actually covered by the original FAIR principles: the reusability requirement R1.1 requires data be made available under a clear license.

The FAIR Data Expert Group does not believe that the FAIR principles should be expanded with additional principles, but instead the existing principles should be augmented with additional concepts such as “as open as possible, as closed as necessary” [118]. Whilst FAIR data can be accessible with restrictions, open sharing should be the default for data that does not have privacy or intellectual property concerns. The expert group also believe that the most challenging concepts, interoperability and reusability, must be extended, and FAIR must be more clearly defined. For reusability they suggest additional components be added: timely

Table 1.2 The FAIR Guiding Principles developed by Wilkinson *et al.* (2016) [112].

Principle	Description
Findable	
F1	(meta)data are assigned a globally unique and persistent identifier
F2	data are described with rich metadata (defined by R1 below)
F3	metadata clearly and explicitly include the identifier of the data it describes
F4	(meta)data are registered or indexed in a searchable resource
Accessible	
A1	(meta)data are retrievable by their identifier using a standardized communications protocol
A1.1	the protocol is open, free, and universally implementable
A1.2	the protocol allows for an authentication and authorization procedure, where necessary
A2	metadata are accessible, even when the data are no longer available
Interoperable	
I1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2	(meta)data use vocabularies that follow FAIR principles
I3	(meta)data include qualified references to other (meta)data
Reusable	
R1	meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1	(meta)data are released with a clear and accessible data usage license
R1.2	(meta)data are associated with detailed provenance
R1.3	(meta)data meet domain-relevant community standards

release of data (particularly important for situations such as infectious disease outbreaks), assessability to ensure (meta)data are of sufficient quality, data stewardship and long term sustainable data storage in digital repositories and user responsibility when handling data.

In order for widespread implementation, FAIR requires an extensive technical infrastructure including registries or catalogs of resources (e.g. FAIRsharing[13]), policies and protocols, data management plans, unique identifiers, reporting standards (ideally using ontologies), automated workflows and data repositories. It is especially important that repositories are funded, supported and incentivised, as they play a crucial role by providing data storage. As part of this, it is important repositories and databases be assessed. Two such schemes are CoreTrustSeal[14], which provides accreditation for trusted digital repositories and the Reuseable Data Project[15] where each database receives a star rating based on how well "a resource's data may be build upon, edited, modified, and redistributed".

There have been suggestions for metrics to access whether data and repositories are FAIR, and their level of "FAIRness". Wilkinson *et al.* [119] propose a series of 14 metrics that cover each of the four principles. It has also been recommended that a system based on Tim Berners-Lee's five star open linked data standards[16] be used as a metric.

Whilst the initial focus around FAIR has mostly been data centric, the FAIR principles must additionally be applied to metadata and software used for analysis to ensure reproducible data analysis. With this there is hope that FAIR can help to increase the value of data and benefit the entire scientific community.

However, it must be noted that a major limitation of the FAIR principles is that they do not assess data or metadata quality. This means that even if it dataset is fully findable, accessible and interoperable, it may not be truly reusable if it is too low quality. Therefore, in addition to the FAIR principles, it is important that standardised methods for evaluating the quality of data are also developed.

1.7 Open Metabolomics Data

A series of dedicated repositories to host metabolomics data have been released over the past decade (Table 1.3). There are now >1,500 metabolomics datasets that included data available via these repositories (Figure 1.6). Whilst the amount of publicly available metabolomics data is still lagging behind the other omics (there are >4,700 datasets on ProteomeXchange [120]

[13] https://fairsharing.org/, accessed 7th April 2018.
[14] https://www.coretrustseal.org/, accessed 20th June 2018.
[15] https://reusabledata.org/, accessed 20th January 2018.
[16] https://5stardata.info/en/, accessed 20th January 2018.

and >70,000 on ArrayExpress [121]), the amount of open metabolomics data is growing rapidly (Figure 1.6A).

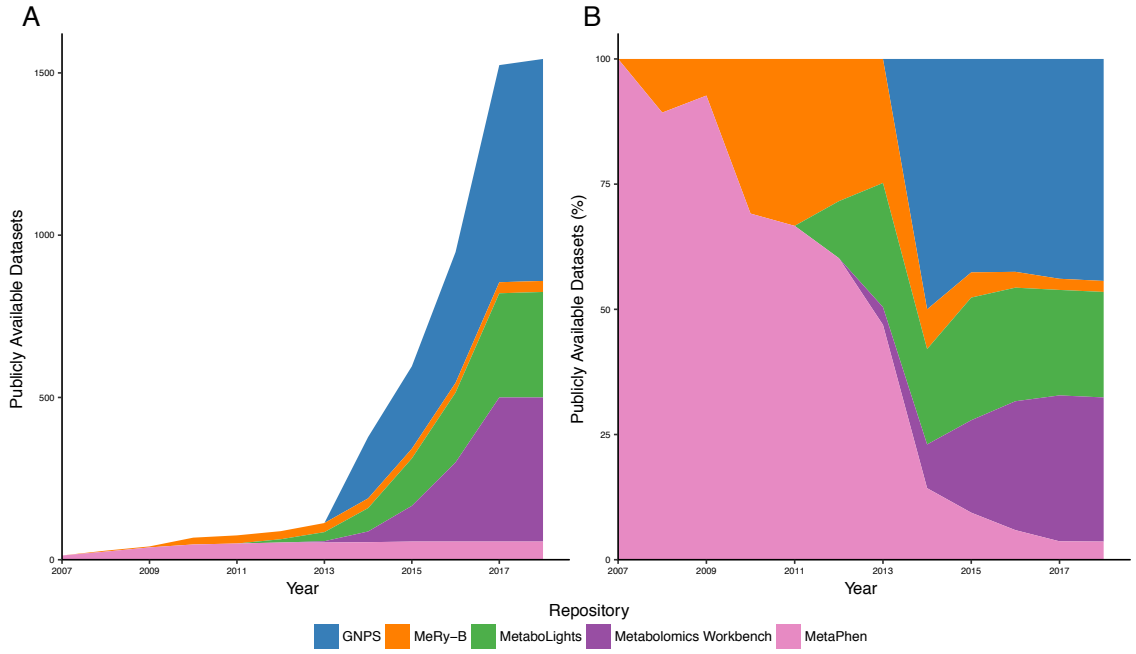


Fig. 1.6 Publicly available metabolomics datasets with raw data available to download, as of 23rd January 2018. A) The total number of publicly available metabolomics datasets and B) the relative percentage of datasets per metabolomics repository. Plots are coloured by repository. Figure is inspired by Figure 5 from Aksenov *et al.* [122].

The first metabolomics repositories to be developed were small and specialised. Both Metabolomics Repository Bordeaux (MeRy-B) [123] and Metabolic Phenotype Database (MetaPhen) [124, 125] were designed to store plant metabolomics data. It was not until 2012 that the first general purpose, global repository for metabolomics data was released. The European Bioinformatics Institute's (EBI) MetaboLights [126] launch was followed the subsequent year by the launch of the United States Government National Institutes of Healths' (NIH) Metabolomics Workbench [127]. These repositories accept all types of metabolomics data, including a wide variety of vendor-specific data formats. More recently the Global Natural Products Social Molecular Networking (GNPS) [128] has also been released. The growth in publicly available metabolomics data has been particularly dramatic since the general purpose metabolomics data repositories were first released. Across the three largest repositories there is >32TB of raw metabolomics data (13.69TB on GNPS, 9.04TB on Metabolomics Workbench and 9.43TB on MetaboLights, as of 10th February 2018).

Table 1.3 Comparison of metabolomics data repositories as of the 26th January 2018. The total number of studies on GNPS was not available.

Repository	Scope	Instrument	No. Studies	No. Public Studies	Analysis	Compound Database
MetaboLights	Global	All	564	326	-	✓
Metabolomics Workbench	Global	All	848	604	✓	✓
MetaPhen	Plant	GC-MS	115	58	✓	✓
MeRy-B	Plant	^{1}H NMR	57	34	✓	✓
GNPS	Natural Products	MS	-	684	✓	✓

Of the two general metabolomics repositories, MetaboLights [126] has a greater focus on curation and more stringent submission guidelines. The MetaboLights repository is comprised of two layers: (i) the repository and (ii) the reference layer. Raw data, along with annotated metadata, are stored in the repository. The reference layer contains a library of 24,957 reference compounds. Both layers can be queried by species, instrumental analysis technology, organism part (e.g. serum, leaf) or study validation status.

Users must submit experimental metadata to MetaboLights in the ISA-TAB format (Investigation/Study/Assay-Tab Delimited), a format for standardised metadata description [129]. When a study has been submitted it undergoes four stages of curation: submitted, in curation, in review and public. When a study is in the submitted stage, the submission is not yet complete and mandatory metadata is missing. Curators will appraise the study when all mandatory fields are complete, looking for any missing optional fields or errors not detected by automatic validation. The submitter is given a read only reviewer link when curation is finished. This can be shared with a journal or collaborators prior to the study becoming public. The study will become public once a user specified publication date has been reached.

MetaboLights is now the recommended metabolomics data repository of a number of journals including: Scientific Data, BioMed Central, PLOS Biology, EMBO Press, Wellcome Open Research, F1000Research, Metabolomics, Frontiers and Metabolites[17]. Along with Metabolomics Workbench [127], it is also part of the Fairsharing project[18], which links databases, standards and policies.

[17]https://www.ebi.ac.uk/metabolights/about, accessed 26th January 2018.
[18]https://fairsharing.org/, accessed 7th April 2018.

Both MetaboLights and Metabolomics Workbench include multiple layers, storing data and metadata from metabolomics studies along with reference libraries of metabolites. However, Metabolomics Workbench provides greater analysis capability than MetaboLights, including the ability to perform exploratory univariate and multivariate statistical analysis on publicly available datasets. There is also the option to compare metabolites between studies (meta-analysis). Additionally, users can perform statistical analysis on their own uploaded dataset. It must be noted, however, that not all publicly available studies in Metabolomics Workbench include raw data (as of 26th January 2018, 444/604 public studies provide raw data).

Another utility provided by Metabolomics Workbench, which is not included in MetaboLights, is a REST (REpresentational State Transfer) service where HTTP (Hypertext Transfer Protocol) requests can be used to access study data and metadata, including metabolite structures and experimental results. On the Metabolomics Workbench website, users can select a specific subset of studies by disease, sample source, species, pathway or metabolite class using bubble plots.

Metabolite structures and annotations are stored on the Metabolomics Workbench Metabolite Database. It contains >61,000 entries collected from public databases including LIPID MAPS [130], Chemical Entities of Biological Interest (ChEBI) [131], Human Metabolome Database (HMDB) [132], BioMagResBank (BMRB) [133], PubChem [134] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [135]. To search for metabolites users can either employ a text based search or, for untargeted MS, m/z values can be used. Metabolomics Workbench also contains the Human Metabolome Gene/Protein Database (MGP), which contains data on >7,300 genes and >15,500 proteins that are related to metabolites.

The Reference list of Metabolite names (RefMet), available on Metabolomics Workbench, aims to provide a standardised reference nomenclature for metabolites identified using spectroscopic techniques in metabolomics experiments. This is important to allow comparison across metabolomics studies. To date RefMet contains 11,681 metabolites, with ~42,000 names, derived from >200 MS and NMR studies.

In order to be able to submit studies to Metabolomics Workbench, users must register and obtain authorization. Following authorization, users must register their study and then submit metadata in the specified format, either via an online form or a supplied excel template. After this raw data and/or supplementary material can be uploaded.

MeRy-B [123] is exclusively a repository for plant metabolomics datasets. It is targeted for proton nuclear magnetic resonance (^{1}H-NMR) data, but also includes GC-MS data. In total it contains 347 GC-MS and 1564 NMR spectra. Under the compounds tab, users can

search for metabolites identified by species, for every species with publicly available data in the repository, along with which pathways metabolites are involved in.

The repository MetaPhen [125] is available as part of MetabolomeExpress [124], a web-based GC-MS metabolomics data analysis platform. The majority of the studies in MetaPhen are of plants and are GC-MS based. MetabolomeExpress contains a variety of tools for the analysis of the publicly available data, providing processing, statistical analysis and visualisation. ResponseFinder can be used to search for metabolites of interest by species and organ/tissue/fluid type. Comparing the results of multiple experiments via meta-analysis can be performed using MetaAnalyser, which plots heatmaps of aligned and clustered data. PhenoMeter allows users to search the reference library via metabolite response patterns.

Natural products mass spectrometry data can be analysed and stored on the GNPS platform [128]. Data are stored in the Mass Spectrometry Interactive Virtual Environment (MassIVE) repository that was originally developed for proteomics studies [136]. MassIVE thus includes both metabolomics and proteomics data. GNPS is unique among metabolomics repositories in providing continuous metabolite identification for MS/MS spectra, with datasets being reanalysed for new identifications once a month. There are >71,000 spectra in the GNPS Public Spectral library, of which 3371 are user submitted. GNPS also includes a series of third party spectral libraries: Massbank [137], RIKEN MSn spectral databases (ReSpect)[19], HMDB [132] and Critical Assessment of Small Molecule Identification (CASMI) [138]. Across the datasets on GNPS, more than 10,000 unique features have been identified. GNPS can be searched by features of interest to identify which datasets they were found in.

The three repositories MeRy-B, MetaboLights and Metabolomics Workbench are all data providers for the MetabolomeXchange consortium[20]. Metabolonote [139], a metadata manager that includes a database of studies along with their experimental metadata, but without any raw or processed data, has recently also become a data provider. The COordination of Standards in MetabOlomicS (COSMOS) consortium [140] founded MetabolomeXchange, and based it on the successful ProteomeXchange [120]. As of 26th January 2018, MetabolomeXchange includes 1056 datasets (324 from MetaboLights, 602 from Metabolomics Workbench, 26 from MeRy-B and 104 from Metabolonote).

Omics Discovery Index (OmicsDI) [141] is a platform that integrates datasets from many different omics repositories, linking different omics assays from the same overall study together. MetaboLights, Metabolomics Workbench, GPNS and MetaPhen are included as the main metabolomics data repositories (MetaPhen datasets are labelled as Metabolome Express

[19] http://spectra.psc.riken.jp/, accessed 27th January 2018.
[20] http://www.metabolomexchange.org/, accessed 26th January 2018.

Table 1.4 General Repositories for data sharing.

Repository	Website
Dryad Digital Repository	https://datadryad.org/
Figshare	https://figshare.com/
GigaDB	http://gigadb.org/site/index
Harvard Dataverse	https://dataverse.harvard.edu/
Open Science Framework	https://osf.io/
Zenodo	https://zenodo.org/

on OmicsDI). OmicsDI aims to support FAIR principles by making omics data more findable, accessible and interoperable. As of 31st January 2018, OmicsDI features 1674 metabolomics datasets including 662 from GNPS, 330 from MetaboLights, 621 from Metabolomics Workbench, and 58 from MetaPhen. Additionally there are also two metabolomics datasets in ArrayExpress [121] and one in the Library of Integrated Network-based Cellular Signatures (LINCS)[21], however this dataset appears to be inaccessible.

As well as via metabolomics specific repositories, datasets may be shared by other means: on personal or institutional websites, or on general, "catch-all" repositories. Examples of general platforms for sharing data include: Dryad Digital Repository (DDR), Figshare and Zenodo (Table 1.4). Datasets shared in this manner may be very hard for other researchers to find, especially those shared on personal websites. It can be challenging to find datasets containing raw data on these general repositories as they also include other types of data e.g. powerpoint presentations, posters, supplementary figures and reports, and it can be difficult to filter searches.

Until recently there was no way to search multiple general repositories simultaneously. DataMed [142] is a project that aims to fill this void and make biomedical datasets findable. It is a data discovery index (DDI) that, as of 31st January 2018, indexes 74 repositories, with 2,336,403 datasets and 15 data types. Currently DataMed is still a prototype and metabolomics data is cataloged as "unspecified". However, Metabolomics Workbench and OmicsDI are already data providers for DataMed and 1,389 studies are returned when searching for "metabolomics".

SciCrunch[22] is a platform for data sharing that, like DataMed, increases the discoverability and accessibility of datasets. Scientific communities can create their own portals on SciCrunch, which can provide access to data sources, tools and relevant literature. Currently

[21] http://lincsportal.ccs.miami.edu/datasets-beta/, accessed 31st January 2018.
[22] https://scicrunch.org/, accessed 31st January 2018.

for metabolomics, the only data source that links to raw data is OmicsDI, however there is potential for this to be expanded in future.

1.8 Meta-Analysis and Systematic Reviews in Metabolomics

A meta-analysis is a type of statistical analysis where the results of multiple studies are combined. Generally, meta-analyses are used as subsets of systematic reviews, and the Cochrane Handbook for Systematic Reviews of Interventions [143] formally defines meta-analysis as "the use of statistical techniques in a systematic review to integrate the results of included studies". Systematic reviews are a type of literature review that are used to gather and critically appraise all empirical evidence around a clearly defined question, synthesizing evidence from across studies. Sometimes systematic reviews are incorrectly referred to as meta-analyses, however the two study types differ in design. Both meta-analyses and systematic reviews are metascientific study designs.

Meta-analyses are used to systematically assess and draw conclusions about a body of research [144], benefiting from higher statistical power than individual studies. In statistical analysis, power is defined as the probability that a test rejects the null hypothesis when the alternative hypothesis is true. The higher the statistical power, the less chance of making a type II error (not rejecting the null hypothesis when it is false). Meta-analyses increase statistical power by shrinking the confidence interval around the weighted average effect size, making it more likely that nonzero population effects are detected [145].

Meta-analyses are highly important for translating research into evidence based medicine. For metabolomics to become used as a clinical tool it is essential that high quality meta-analyses be performed.

Whilst it is sometimes assumed that meta-analysis entail simply the pooling of multiple studies, this can lead to analytical errors. Pooling data can result in the compounding of biases and reduce the ability to detect differences within study groups.

Meta-analyses are typically conducted in a standardised manner, including the following stages:

1. Define the research question, often done using the PICO (Participants, Interventions, Comparisons and Outcomes) model, including eligibility criteria
2. Literature search
3. Select studies (based on inclusion and exclusion criteria)
4. Data collection

5. Standardise data
6. Calculate the overall effect by combining the data (selecting the meta-analysis model e.g. fixed effect or random effects)
7. Examination of sources of bias

A number of guidelines for conducting meta-analysis exist, such as: PRISMA (Preferred Reporting Items for Systematic reviews and Meta-analyses)[23] [146] which is widely applicable and MOOSE (Meta-analysis Of Observational Studies in Epidemiology) [147], which is specifically designed for observational studies. The PRISMA guidelines are the most widely used. More recently, guidelines that expand on the original PRISMA have been released e.g. PRISMA-P (for protocols) [148]. The original PRISMA checklist includes 27 items and PRIMSA-P includes 17.

It is important that meta-analyses be preregistered as preregistration guards against both conscious and unconscious biases. The International Prospective Register of Systematic Reviews (PROSPERO)[24] can be used to register biomedical outcomes. Alternatively, protocols can be posted on preprint servers.

Meta-analyses are typically reported in a standardised manner, including a flow chart (Figure 1.7) describing the experimental design, detailing how many records are included at each stage and using a forest plot to show the number of studies that address the research question, along with their outcomes.

To date there have been very few meta-analyses published about metabolomics research, and the majority of studies have not performed meta-analysis as defined above. This is in contrast to genomics [149] and transcriptomics [150] where many meta-analyses have been performed, however similarly to metabolomics, only a small number of meta-analyses have been conducted about proteomics research [151].

In this research, eight studies involving metabolomics that state they include meta-analyses have been identified, however none of which included a literature review, an essential component of a meta-analysis (Table 1.5). A further three studies that include metabolomics [152–154] that say a meta-analysis was performed but do not state which guidelines were used were also identified. Of these only Park *et al.* (2018) [152] performs a standard meta-analysis. Mehta *et al.* (2017) [153] used a literature search to identify metabolites of interest and then performed targeted analysis to detect only those metabolites.

Worryingly, Goveia *et al.* (2016) [154] performed a "vote counting meta-analysis". Vote counting involves simply comparing the number of positive studies to the number of

[23] http://www.prisma-statement.org/, accessed 6th July 2018.
[24] https://www.crd.york.ac.uk/prospero/, accessed 6th July 2018.

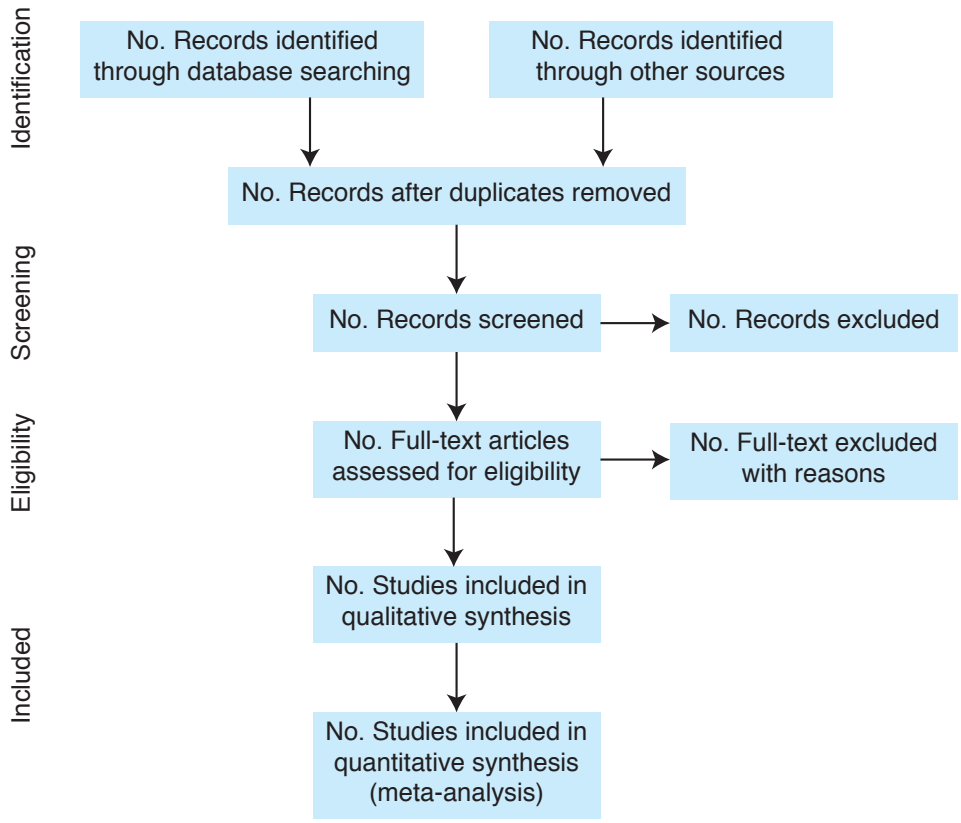


Fig. 1.7 PRISMA Flow Chart, first described in Liberati *et al.* [146].

negative studies and does not consider statistical significance or sample size. There is also no study weighting in a vote counting meta-analysis. In a standard meta-analysis, studies are weighted based on sample size and statistical power. Without weighting or taking sample size into account, a study with 1000 participants and high statistical power will be given equal importance as a low powered study with 10 participants. Whilst there is some evidence that vote-counting can be used to predict biomarker performance [155], vote counting is a statistically flawed procedure, and is not a true meta-analytic technique [156].

In total, three studies that included meta-analyses and metabolomics research and that were pre-registered, were identified: Guasch-Ferré *et al.* (2016) [157] and Okekunle *et al.* (2017) [158] followed the MOOSE guidelines and Siristatidis *et al.* (2017) [159] used the PRISMA guidelines. Guasch-Ferré *et al.* (2016) identified metabolite markers of prediabetes and Type II diabetes, and Okekunle *et al.* (2017) identified amino acids associated with obesity, Type II diabetes and metabolic syndrome. Unlike the other two studies Siristatidis *et al.* (2017) did not attempt to identify metabolomics biomarkers, and instead the study evaluated the effectiveness and safety of metabolomics compared

Table 1.5 Metabolomics studies that state they include meta-analysis, but where no literature review was performed. The cohorts of participants included in these studies are reported.

Year	Cohort(s)	Study
2015	COPDGene	[161]
2017	Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study	[162]
2017	Avon Longitudinal Study of Parents and Children, Northern Finish Birth Cohort 1966 and 1986 (NFBC1966 & NFBC1986)	[163]
2017	European Prospective Investigation into Cancer and Nutrition (EPIC)–InterAct	[164]
2017	-	[165]
2017	Hyperglycemia and Adverse Pregnancy Outcome (HAPO)	[166]
2018	EPIC–Potsdam, EPIC–Heidelberg	[167]
2018	TwinsUK, The Hertfordshire Cohort Study, NFBC1986	[168]

Maga-Nteve *et al.'s* (2017) study [165] was of Balb/cJ mice and did not include any human participants.

to morphology assessment for improving live birth or ongoing pregnancy rates in women undergoing assisted reproductive technologies.

Whilst it is encouraging that systematic reviews including metabolomics have been successfully conducted, the ability to perform meta-analyses in metabolomics is currently hindered by a lack of open data with sufficient metadata. It has not been standard for metabolomics researchers to report all metabolites identified in a study; instead often only metabolites of interest are reported. It is therefore impossible to know whether a metabolite is not detected or simply not reported, resulting in bias in outcome reporting. This leads to bias in statistical analysis that cannot currently be resolved [154]. It is also much harder to reuse datasets that have missing or erroneous metadata, and there is more risk of misinterpretation [160]. Studies that do not report data on a per sample level cannot be used to guide personalised/stratified medicine.

1.9 Research Objectives

Evidence based medicine relies on rigorous meta-analyses with sufficient statistical power. If metabolomics is to become a widely used clinical tool, meta-analyses must be performed to demonstrate that quantitative metabolite measurements are reliable and accurate across

studies. However, meta-analyses require sufficient data and metadata reporting — a lack of open data hinders scientists conducting meta-analyses.

Therefore, the main objective of this thesis is to assess current practices and standards of sharing data within the field of metabolomics, with an outlook towards metabolomics becoming an entrenched tool for diagnosis and classification. To this end, metascientific approaches were used to assess the current reporting standards for metabolomics data and data sharing practices within the community.

Reporting standards are an important constituent in making data FAIR — well written reporting standards ensure the consistency of metadata between datasets, facilitate data reuse and data merger across studies. However, poorly worded standards can impede data reuse if ignored by communities and not adhered to. It is therefore important to assess the extent to which standards are complied with. The Metabolomics Standards Initiative (MSI) guidelines are now a decade old and the majority of dedicated metabolomics repositories purport to comply with them. In chapter 2 compliance of 483 open datasets, from five metabolomics data repositories, to the biological context MSI reporting standards, is examined. Following analysis a series of recommendations as to how the MSI guidelines could be revised and improved are provided.

Increasingly funding bodies, societies and journals are encouraging or mandating data sharing and open data. It is therefore important to assess the effectiveness of these policies at expanding open data sharing. In chapter 3 data sharing practices within the metabolomics community are investigated. The data sharing policies of the journals with publications associated with the most open metabolomics data, and those that publish the most metabolomics research, are explored.

Having identified PLOS ONE as the second largest publisher of metabolomics research, it was surprising to find that <30 PLOS ONE publications were directly linked to open metabolomics data in repositories, considering its pioneering requirements for data sharing. The data availability statements and the levels of data sharing of PLOS ONE metabolomics papers were therefore examined, in order to investigate reasons for the lack of public archiving of data in dedicated repositories. To review the value of publicly sharing metabolomics data, the extent to which public metabolomics data has been reused was also investigated. Finally, ways in which data sharing in metabolomics can be improved are discussed.

The reproducibility of metabolomics analysis depends not only on FAIR data but also on FAIR analysis workflows. There are now >250 software tools specific for metabolomics data, which can make it hard for researchers to find the right tool. In chapter 4 a taxonomy for metabolomics software, in order to formally categorize them, is presented. A GitHub Pages

wiki — https://raspicer.github.io/MetabolomicsTools/ was also developed in order to provide extensive details about all included software.

1.10 Publications

Over the course of my PhD I have authored seven peer reviewed publications: two analyses, two commentaries, a review, a protocol and a book chapter. For four of these publications I was the first author, having conducted the majority of the research and writing.

The analysis "Compliance with minimum information guidelines in public metabolomics repositories" [169] was my first lead author research paper. In this work compliance to the MSI biological context metadata reporting standards was assessed using open data from dedicated metabolomics repositories. I contributed to the publication by designing the methodology, conducting the investigation, performing statistical analysis, and writing and editing the paper.

Complimenting the analysis a commentary, "A decade after the metabolomics standards initiative it's time for a revision" [170], was written. As it was found that the MSI guidelines are not well adhered to, we proposed that MSI guidelines should be revisited and revised, as has been done in other communities, to better reflect the current requirements of the metabolomics community. For this commentary I wrote the manuscript and produced the figure. The analysis and commentary were written simultaneously, and submitted jointly. Both are detailed and expanded upon in chapter 2.

My next first author published research article "A lost opportunity for science: journals promote data sharing in metabolomics but do not enforce it" [171] reviewed the data sharing policies of journals publishing the most metabolomics papers associated with open data and compared these journals' policies to those that publish the most metabolomics papers. It was found that journals that most support data sharing are not necessarily those with the most papers associated to open metabolomics data. I conceptualised this research myself, and composed the methodology, performed the investigation and formal analysis and wrote the article. This research is covered in the section 3.1 of chapter 3.

The review "Navigating freely-available software tools for metabolomics analysis" [29] was written to help guide researchers in choosing software for analysing metabolomics data. Software tools were categorised by the type of instrumental data (i.e. LC–MS, GC–MS or NMR) and the functionality (i.e. pre- and post-processing, annotation, statistical analysis, workflows and other) they are designed for, and an extensive list of the most used tools was compiled. I was the first author of the review and did the majority of the writing and

background research. The review won the 2018 Metabolomics Publications Award in the category of review article, for being the most downloaded review in the journal published over the last year. This review is mentioned in chapter 4, however subsequent work has expanded upon the work initially conducted for the review, so it is not the main focus of the chapter.

The article "MetaboLights: An Open-Access Database Repository for Metabolomics Data" [172] is a protocol, providing a tutorial for using the MetaboLights repository [126], providing instructions for submitting studies to the database and examples of how to query it. I contributed to writing and editing the manuscript.

"The future of metabolomics in ELIXIR" [173] is an opinion article developed from a workshop on the same topic. Metabolite identification was recognised as a key area where computational metabolomics and data management can have the most impact. The article supports the call for metabolomics as a new ELIXIR Use Case. I contributed to writing the document.

For the book chapter "Metabolome Analysis" [174], I contributed to the writing of the section *Data Processing, Workflows and Repositories*, which included the subsections *Workflows in Metabolomics* and *Metabolomics Experiments Databases and Repositories*. These sections cover workflows that can be used to analyse metabolomics data, providing tools for all stages of data analysis, and an overview of dedicated repositories that store metabolomics data.

Chapter 2

Compliance with Reporting Standards in Metabolomics

2.1 Reporting Standards

There are many types of standards in biology; they range from reference reagents to laboratory protocols. Standards have been heralded as a solution to the replication crisis [175]. Data reporting standards are designed to ensure that the minimum information required to understand and interpret the results of analysis are reported.

Well written reporting standards are easy to use and aid researchers in publishing their data [114]. They ensure the consistency of metadata between datasets, and facilitate data reuse and data merger across studies. Reporting standards can help to achieve findable, accessible, interoperable and reusable (FAIR) data [112]. Conversely, poor reporting standards are ignored by the communities they were intended to serve and are not complied with, hindering data reusability [176]. Alternatively, multiple competing standards are developed, leading to different groups using different standards and adding to confusion. As Andrew Tanenbaum said, "The nice things about standards is that there are so many to choose from" [177].

The first omics reporting standards were the Minimum Information About a Microarray Experiment (MIAME) standards [178]. These described the minimum information required for reporting microarray DNA expression studies. Following MIAME, other omics communities began to develop standards. The minimum information about a proteomics experiment (MIAPE) [179] was developed for proteomics and in metabolomics a series of initiatives were formed. The Minimum Information about a Metabolomics Experiment (MIAMET) and the Architecture for Metabolomics consortium (ArMet) [180] provided frameworks for describing plant metabolomics experiments, and the Standard Metabolic Reporting Structure

(SMRS) initiative focused on standardising the reporting of toxicological, *in vitro* animal and NMR-based experiments [181]. In 2005, these earlier efforts by the metabolomics community were built upon and the Metabolomics Standards Initiative (MSI) was formed [182, 183].

The MSI consisted of a series of working groups: Biological context metadata, Chemical analysis, Data processing, Ontology and Data exchange. The Biological context metadata working group consisted of four subgroups: (1) mammalian studies, divided between human and animal studies, (2) plant studies, (3) cell cultures and microbiology, and (4) environmental studies. In 2007, these working groups published a series of reports detailing minimal reporting standard recommendations for each area. Guidelines were produced for ontologies [184], data exchange formats [185], chemical analysis [186], NMR-based experiments [187], data analysis [188], mammalian/ *in vivo* experiments [189], microbial and *in vitro* experiments [190], plant biology [191] and environmental experiments [192] (Figure 2.1). The Mammalian/ *in vivo* report is split into two sets of reporting standards: Mammalian Clinical Trials and Human Studies, and Pre-clinical.

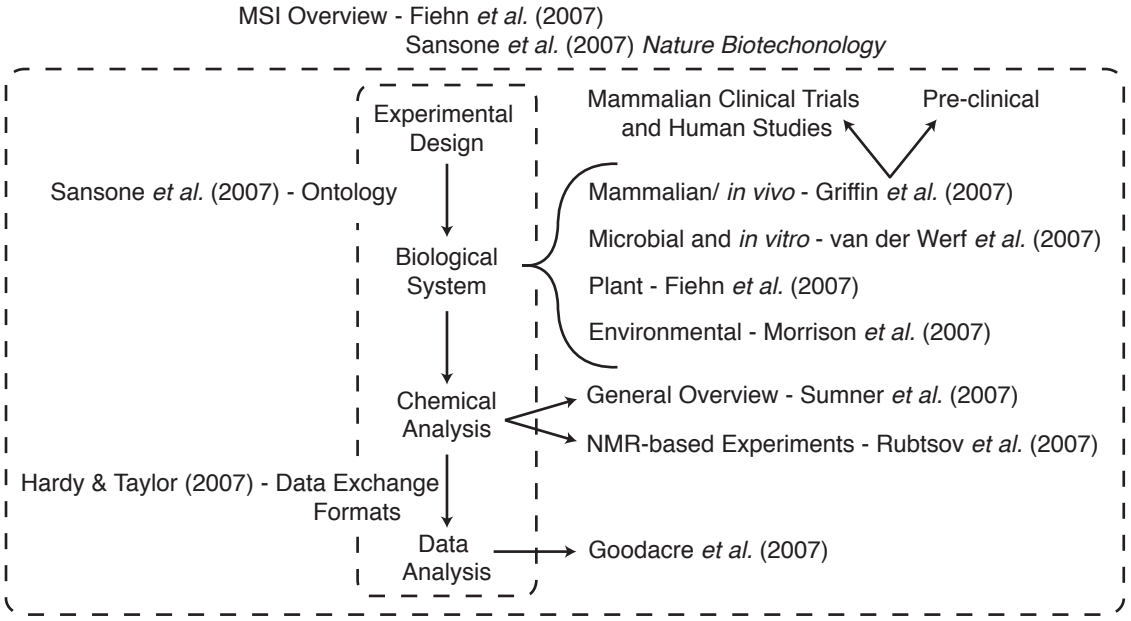


Fig. 2.1 An overview of the sets of reporting standards published by the MSI. Unless otherwise stated, all articles were published in the journal Metabolomics. Figure was adapted from Goodacre (2014) [193] and used with permission.

These guidelines were designed to include all of the descriptive information about an experiment (the metadata) that was considered crucial to the understanding of the data, in

order to enable replication of the experiment and reuse data [182]. Some of the reports consisted of just minimum information (MI) reporting standards, whilst others also included best practice or optional extra standards.

Following the publication of the MSI guidelines, papers detailing how to comply with the mammalian and *in vivo* [194], and plant standards [195], including example studies, were released. These papers acted as case studies, exemplifying how metadata should be reported.

In the years immediately following the release of the MSI guidelines, they were criticised for a lack of practical applications [196, 197]. At the time no metabolomics data repositories existed, and the recommendations could only be used to guide journal submission guidelines and data collection in individual laboratories. The journal *Metabolomics* encourages authors to make submitted manuscripts as compliant with the MSI guidelines as possible [198]. It was thought that release of general purpose, cross platform and cross species repositories would lead to greater use and compliance with the MSI reporting standards [197].

This research aims to examine how complied with the MSI guidelines are, in order to assess whether they reflect the current needs of the metabolomics community. A subset of guidelines were selected for investigation: the biological context metadata reporting standards. The number of metadata mandated by each minimal and best practice reporting standard was quantified. The level of compliance to the standards by currently available public datasets in five metabolomics data repositories was assessed. Successes and failures of the current MSI guidelines were identified, and the chapter is concluded by suggesting improvements to the MSI guidelines that could be implemented in future revisions.

2.2 Methods

2.2.1 Dataset Selection

Of the five main repositories of metabolomics data, four have been developed to fulfill the MSI guidelines for minimum metadata reporting: MetaboLights [126], Metabolomics Workbench [127], MeRy-B [123] and MetaPhen [124, 125] (for a more detailed description of each repository see chapter 1.7). GNPS [128] does not aim to comply with the MSI guidelines and has only minimal requirements for reporting experimental metadata. Table 2.1 provides an overview of each of the repositories, at the time this analysis was conducted, in regards to their scope, number of studies and the number of different species included. As GNPS has minimal requirements for reporting experimental metadata, it is assumed that studies in GNPS will not follow the MSI guidelines.

Table 2.1 Comparison of metabolomics data repositories as of the 7th March 2017. The number of species refers to the number of different species across publicly available studies only, and does not include species in private studies. The total number of studies on GNPS was not available.

Repository	Scope	No. Studies	No. Public Studies	No. Species
MetaboLights	All Metabolomics	435	236	77
Metabolomics Workbench	All Metabolomics	505	365	40
MetaPhen	Plants - GC-MS focused	115	58	17
MeRy-B	Plants - ^{1}H NMR	54	30	17
GNPS	Natural Products - MS	-	430	234

In order to identify a subset of publicly available datasets, to which compliance with the MSI standards would be tested, the instrumental analysis type (Figure 2.2) and species (Figure 2.3) of every study were identified. It was found that GNPS only requires users to submit the type of mass spectrometer used for analysis, and has no dedicated field for reporting the type of chromatography used. Therefore, the type of chromatography used for GNPS studies could not be determined and GNPS was not included in Figure 2.2. Accordingly, only the species included in datasets was used to select a subset of studies for analysis. Across repositories a total of 347 unique species were identified.

The three species that were most prevalent across the five repositories' publicly available studies were selected for analysis: *Homo sapiens*, *Mus musculus* and *Arabidopsis thaliana* (Figure 2.3). These species provide a good coverage across experimental areas with different metadata requirements, covering four out of the five biological experimental areas of the MSI standards. The non-human species *M. musculus* and *A. thaliana* are additionally both extensively studied model organisms.

Some *H. sapiens* and *M. musculus* studies did not fit within the scope of any of the existing MSI standards. These included studies of intra-laboratory differences, the development of new experimental techniques, sample collection conditions and studies of within sample variation. Studies that did not have an applicable MSI biological context metadata guideline were excluded from further analysis.

Every study from the four repositories including the selected species was categorized by applicable biological context metadata standards. Some studies contain multiple assays, which can include multiple species or both cell lines and clinical research. Multiple sets of MSI reporting standards can therefore be applicable to a single study.

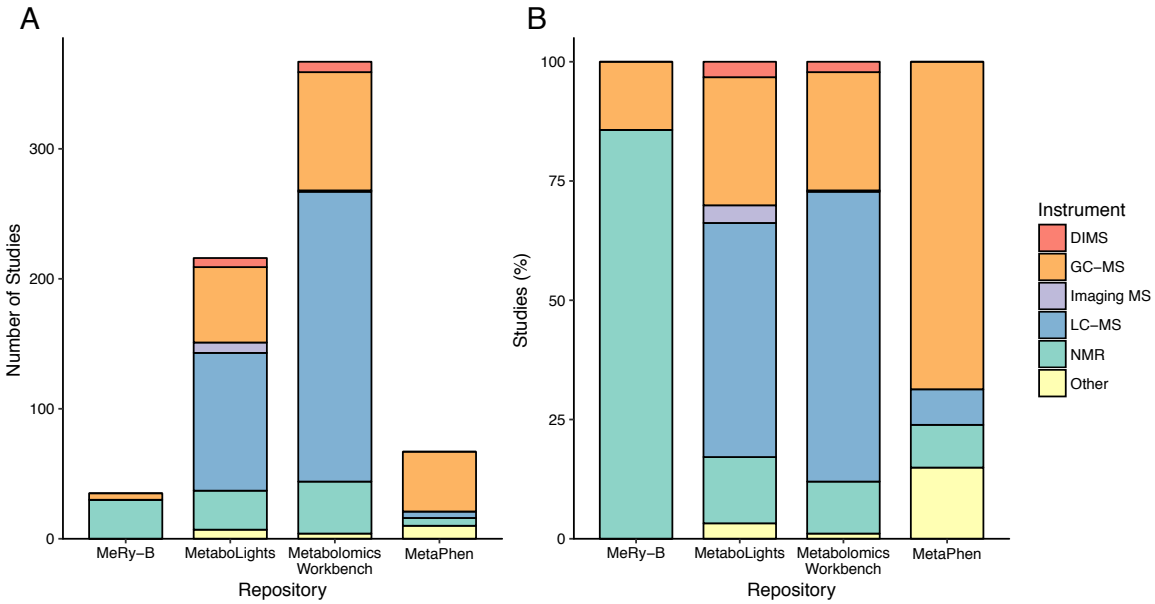


Fig. 2.2 Distribution of analytical techniques across metabolomics data repositories. The (A) frequency and (B) percentage of studies including different analytical techniques in the metabolomics data repositories: MeRy-B, MetaboLights, Metabolomics Workbench and MetaPhen. GNPS is not included as it has no requirements for reporting type of chromatography used, although it is exclusively a repository for MS data.

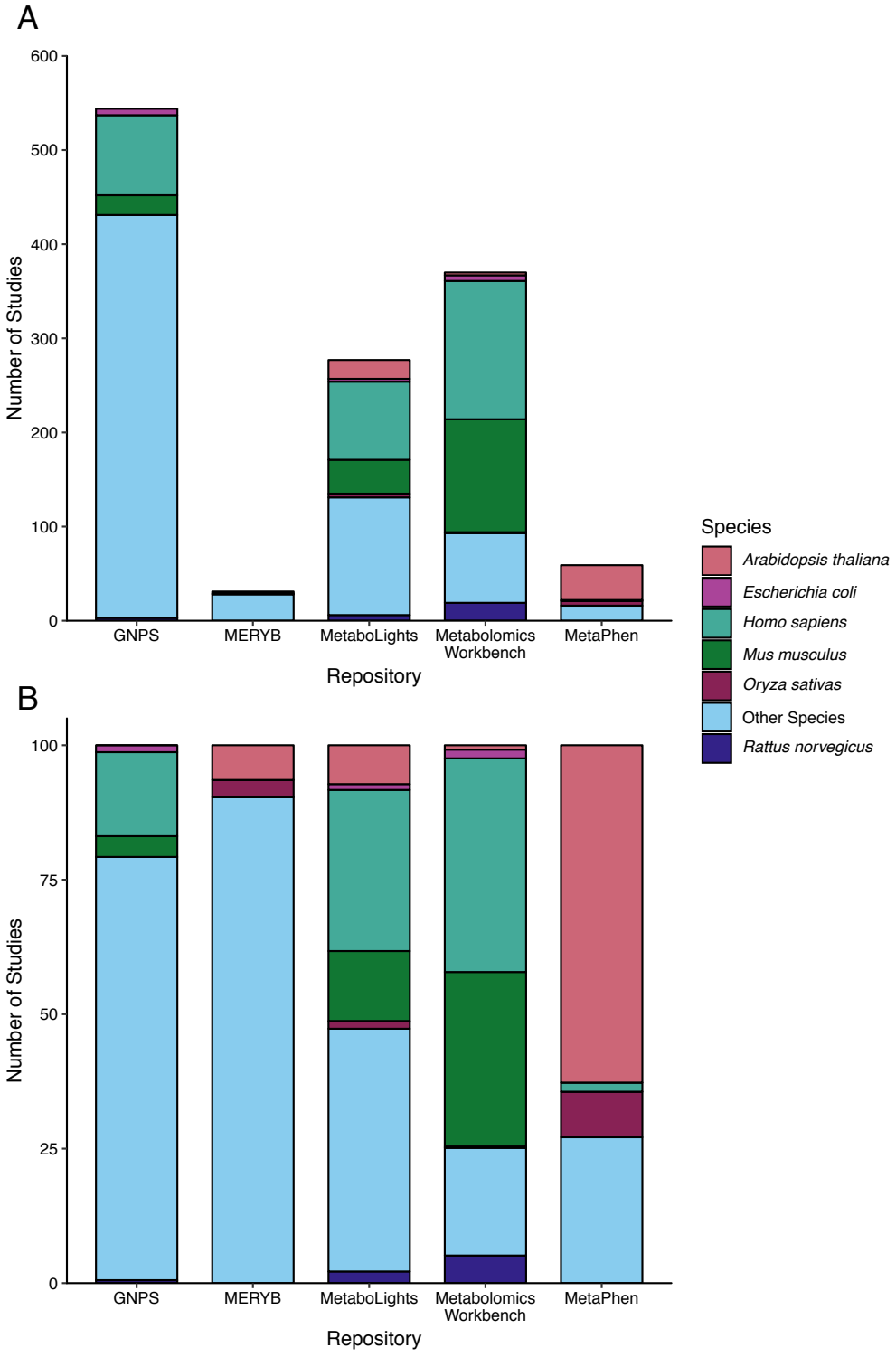


Fig. 2.3 Distribution of species across metabolomics data repositories. The (A) frequency and (B) percentage of studies including different species in the metabolomics data repositories: GNPS, MeRy-B, MetaboLights, Metabolomics Workbench and MetaPhen. For a species to be plotted as an individual band it must have been found in a minimum of ten studies across the repositories. Species found in <10 studies across all repositories are reported as "Other Species".

2.2.2 MSI Guidelines

Every MSI biological system report provided minimum reporting standards required for that class of experiment, as decided upon by the MSI subgroup. As well as the minimal reporting standards, the mammalian clinical trials and human studies (abbreviated to clinical), pre-clinical and environmental reports also included recommended further information. Additional best practice reporting standards were also included in the Microbial and *in vitro* report (abbreviated to *in vitro*). For brevity, throughout the remained of this chapter all optional and best practice reporting standards will be referred to as optional reporting standards.

The number of metadata mandated by each minimal and optional reporting standard was quantified (Table 2.2), with some guidelines being combined to obtain a binary list (e.g. the organ and cell type standards in the plant guidelines were combined to a single item, as studies usually include only one of these biosources). The pre-clinical reporting guidelines contained the highest number of minimal reporting standards and the *in vitro* guidelines suggested the most optional.

Compliance to the environmental set of reporting standards was not examined, as none of the repositories contained a sufficient number of environmental studies to enable testing.

Table 2.2 The number of minimal and optional reporting standards for each biological experimental type. The plant guidelines contain only minimal reporting standards as there are no additional optional reporting standards. Table is reproduced from Spicer *et al.* (2017) [169].

Standard	Minimal	Optional
Environmental	22	38
Mammalian Clinical trials and human studies	22	33
Microbial and *in vitro*	15	39
Plant	20	-
Pre-clinical	30	16
Total	109	116

2.2.3 Metadata Scoring

Following classification, every study was examined for its compliance with the reporting of each metadata included in the MSI guideline. All metadata were classified manually. For metadata to be considered reported, the metadata must have been either directly included

in the repository, along with the data, or in a publication accessible by direct link from the study page.

Some metadata were also recorded as "implicit", where the metadata were not reported on a per sample basis, instead being reported as descriptive statistics for the overall study. For example if sex were reported as 60% male and 40% female, but the sex of individual samples was not reported, this was recorded as implicit metadata.

Once compliance to each metadata reporting standard had been assessed for each individual study, the percentage of studies the metadata was reported in was calculated. The distribution of the percentage of studies reporting each minimal and optional standard was found for each set of standards for every repository.

Homo sapiens

Two of the MSI biological metadata reporting standards are applicable to *H. sapiens* studies: Mammalian Clinical Trials and Human Studies, and Microbial and *in vitro*. Human clinical trials were classified as Mammalian Clinical Trials and Human Studies and *H. sapiens* cell line studies were categorized as Microbial and *in vitro* studies. There are also *H. sapiens* studies that are neither of these experiment types and these were classified as other.

There are 83 public *H. sapiens* studies in MetaboLights as of the 7th March 2017, 147 with the species "human" in Metabolomics Workbench, 85 *H. sapiens* or human studies in GNPS and one *H. sapiens* study in MetaPhen. As there was only a single *H. sapiens* study in MetaPhen, it was excluded from further analysis. Following classification there were 58 clinical, 18 *in vitro* and 7 other *H. sapiens* studies in MetaboLights. Human Metabolomics Workbench studies consisted of 99 clinical, 45 *in vitro* and 3 other. There were 60 Clinical, 3 *in vitro* and 22 studies classified as other in GNPS.

Mus musculus

The appropriate reporting standards for use with *M. musculus* studies can be either Clinical, Microbial and *in vitro* or Pre-clinical. However, there are also *M. musculus* studies that are not covered by the existing reporting standards. In this analysis these studies are classified as other. Across the repositories there are currently no examples of *M. musculus* clinical studies.

As of 7th March 2017, there were 33 *M. musculus* studies in MetaboLights, 120 in Metabolomics Workbench and 21 in GNPS. Only studies categorized as pre-clinical were included in this work, as compliance with the Microbial and *in vitro* guidelines was assessed

using *H. sapiens* cell line studies. There were 29 *M. musculus* pre-clinical studies in MetaboLights, 91 in Metabolomics Workbench and 18 in GNPS.

Arabidopsis thaliana

The Plant reporting standards are broad and are applicable to all *A. thaliana* studies that were publicly available in the four repositories as of 7th March 2017. There were 2 *A. thaliana* studies from MeRy-B, 20 from MetaboLights, 3 from Metabolomics Workbench and 37 from MetaPhen.

2.2.4 Statistical Analysis

A Shapiro–Wilk test [199] was used to test the normality of the data. The Shapiro-Wilk test statistic is as follows:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.1}$$

Where n is the number of observations, $x_{(i)}$ are the ordered sample values, $\bar{x}$ is the sample mean and a_i are the tabulated coefficients.

As the distribution of the data was found to be highly skewed, and the sample size is relatively small, non-parametric statistical tests were used for analysis.

For analysing differences within repositories (MetaboLights, Metabolomics Workbench and GNPS) Kruskal-Wallis tests [200] were used. Kruskal-Wallis tests were also used for comparing differences between repositories.

The Kruskal-Wallis test statistic is as follows:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} \frac{r_i^2}{n_i} - 3(N+1) \tag{2.2}$$

Where N is the total number of observations across all groups, n_i is the number of observations in group i and r_i is the rank of observations in group i.

It must be noted that Kruskal-Wallis tests can be inaccurate if there are less than five observations per group. This means that the p-value for comparing compliance to the plant MSI reporting standards may be inaccurate, as respectively only 2 and 3 studies were included from MeRy-B and Metabolomics Workbench.

If a significant difference was found between groups with a Kruskal-Wallis test, a Dunn post-hoc test [201] was then used. Benjamini-Hochberg correction (false discovery rate) [202] was used to correct Dunn tests for multiple comparisons.

The Dunn test statistic is as follows:

$$z_{A,B} = \frac{\bar{W}_A - \bar{W}_B}{\sigma_{A,B}} \tag{2.3}$$

Where $\bar{W}_A$ is the mean of the joint ranks for group A, $\bar{W}_B$ is the mean of the joint ranks for group B and $\sigma_{A,B}$ is:

$$\sigma_{A,B} = \sqrt{\left[\frac{N(N+1)}{12} - \frac{\sum_{s=1}^{r} \tau_s^3 - \tau_s}{12(N-1)}\right]\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}; \tag{2.4}$$

Where N is the total number of observations in all groups, r is the number of tied ranks across all groups and τ_s is the number of observations across all groups with tied rank.

Mann-Whitney U tests [203] were used to analyse differences between minimal and optional reporting standards.

The Mann-Whitney U test statistic for is as follows:

$$U = min\,(U_1, U_2) \tag{2.5}$$

Where:

$$U_1 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_1 \tag{2.6}$$

$$U_2 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_2 \tag{2.7}$$

Where n_1 is the sample size of group 1, n_2 is the sample size of group 2, R_1 is the sum of the ranks for observations from group 1 and R_2 is the sum of the ranks for observations from group 2.

2.2.5 Data and Code Availability

Raw data used for the analysis in Spicer *et al.* (2017) [169] (that did not include studies from GNPS) is available on Figshare: https://dx.doi.org/10.6084/m9.figshare.c.3803764.v1. The classification of *H. sapiens* and *M. musculus* studies can be found in the files **HumanStudyClassification.xlsx** and **MouseStudyClassification.xlsx**, respectively. Raw data rating each metadata as either reported in accordance with the MSI guideline, or not reported in accordance to the guideline are available in **HumanClinicalRawData.xlsx** for *H. sapiens* clinical studies, **HumanCellRawData.xlsx** for *H. sapiens* cell line studies, **MousePreclinicalRawData.xlsx** for *M. musculus* cell line studies, and **ArabidopsisPlantRawData.xlsx** for *A. thaliana* studies.

All of the analysis in Spicer *et al.* (2017) [169] was performed using R version 3.3.2, using the packages ggplot2 2.2.1, xlsx 0.5.7 and FSA 0.8.13. The code used for analysis is available at https://github.com/RASpicer/Compliance_MSI_Guidelines.

Code and data for reproducing the other figures used in this thesis chapter, which were not included in the paper, can be found at https://github.com/RASpicer/Compliance_Metabolomics_ReportingStandards. Analysis was performed using R version 3.5.1 with ggplot2 version 3.0.0.

For a full list of studies used in this analysis, along with identifiers and links, see Appendix B.

2.3 Compliance to the MSI Standards within and between Repositories

Studies of three species (*H. sapiens*, *M. musculus* and *A. thaliana*) that were prevalent across five metabolomics data repositories were selected in order to test compliance to the MSI biological systems reporting standards. Compliance to four of the five MSI biological context metadata guidelines (clinical, *in vitro*, plant and pre-clinical) was evaluated. Of the repositories analysed, only the general purpose MetaboLights [126] and Metabolomics Workbench [127] contained studies of all three of the selected species. Neither of the plant focused repositories MetaPhen [124, 125] or MeRy-B [123] contained any *M. musculus* studies, however MetaPhen contained a single *H. sapiens* study. The natural products repository GNPS contained *H. sapiens* and *M. musculus* studies, but no *A. thaliana* studies. In total compliance of 483 studies to the MSI reporting standards was examined.

Analysis of the data shows that there are no reporting standards that are fully complied with in every publicly available study (Tables 2.6 - 2.12). The overall rate of compliance varies from 0–97%. However, some reporting standards are complied with by every study within a repository (Tables 2.8, 2.9 & 2.11). Reporting standards relating to biosource had the highest percentage compliance across all of the guidelines.

The results of the Kruskal-Wallis tests comparing compliance to the MSI reporting standards within and between metabolomics repositories are shown respectively in Tables 2.3 and 2.4. The results of the Dunn post-hoc tests are shown in Appendix C. The results of the Mann-Whitney *U* tests comparing compliance between minimal and optional reporting standards within repositories are shown in Table 2.5.

Table 2.3 Kruskal-Wallis tests comparing compliance to the MSI reporting standards within metabolomics repositories. H = test statistic, df = degrees of freedom, p-value = probability value.

Repository	Reporting Standards	H	df	p-value
MetaboLights	Minimal	9.37	3	0.025*
MetaboLights	Optional	7.63	2	0.022*
Metabolomics Workbench	Minimal	32.05	3	5.11×10^{-7}*
Metabolomics Workbench	Optional	10.89	2	0.0043*
GNPS	Minimal	5.52	2	0.063
GNPS	Optional	2.56	2	0.28

* indicates significant values

The *in vitro* minimal reporting standards were complied with significantly less (Kruskal-Wallis $H = 9.37$, $df = 3$, $p = 0.025$; Table 2.3) than the pre-clinical and plant minimum reporting standards in MetaboLights (Figure 2.4A). In Metabolomics Workbench the plant minimum reporting standards were adhered to at a significantly higher rate than the three other guidelines: *in vitro*, clinical and pre-clinical (Kruskal-Wallis $H = 32.05$, $df = 3$, $p = 5.11 \times 10^{-7}$; Figure 2.4B; Table 2.3). There was no significant difference (Kruskal-Wallis $H = 5.52$, $df = 2$, $p = 0.063$; Table 2.3) between the compliance to the *in vitro*, clinical and pre-clinical guidelines in GNPS (Figure 2.5A).

Across the three repositories, the *in vitro* minimal guidelines were complied with at the lowest rate. Conversely, the optional *in vitro* reporting standards had significantly higher compliance than the optional clinical and pre-clinical guidelines in Metabolomics Workbench (Kruskal-Wallis $H = 10.89$, $df = 2$, $p = 0.0043$; Figure 2.6B; Table 2.3) and significantly greater compliance than the clinical in MetaboLights (Kruskal-Wallis $H = 7.63$, $df = 2$, p

Table 2.4 Kruskal-Wallis tests comparing compliance to the MSI reporting standards between metabolomics repositories. H = test statistic, df = degrees of freedom, p-value = probability value.

Repository	Reporting Standards	H	df	p-value
MetaboLights, Metabolomics Workbench, MeRy-B, MetaPhen	Plant	8.38	3	0.039*
MetaboLights, Metabolomics Workbench	Minimal	2.55	1	0.11
MetaboLights, Metabolomics Workbench	Optional	9.02	1	0.0027*
GNPS, MetaboLights	Minimal	56.89	1	4.60×10^{-14}*
GNPS, MetaboLights	Optional	58.50	1	2.03×10^{-14}*
GNPS, Metabolomics Workbench	Minimal	51.60	1	6.80×10^{-13}*
GNPS, Metabolomics Workbench	Optional	31.19	1	2.35×10^{-8}*

* indicates significant values

= 0.022; Figure 2.6A; Table 2.3). However, there was no significant difference between compliance to the different optional reporting standards in GNPS (Kruskal-Wallis $H = 2.56$, $df = 2$, $p = 0.28$; Figure 2.5B; Table 2.3).

There was significantly greater compliance to the minimal reporting standards compared to the optional for the clinical and pre-clinical guidelines within the MetaboLights (Mann-Whitney U test, clinical: $U = 118.5$, $p = 2.02 \times 10^{-5}$, pre-clinical: $U = 94$, $p = 7.59 \times 10^{-4}$; Table 2.5) and Metabolomics Workbench repositories (Mann-Whitney U test, clinical: $U = 78.5$, $p = 1.62 \times 10^{-7}$, pre-clinical: $U = 107.5$, $p = 0.0022$; Table 2.5). However, no significant difference was found between the compliance with the minimal and optional *in vitro* reporting standards in either repository (Mann-Whitney U test, MetaboLights: $U = 341$, $p = 0.33$, Metabolomics Workbench: $U = 268$, $p = 0.62$; Figure 2.7A–B; Table 2.5). The minimal clinical reporting standards were complied with significantly more that the optional in GNPS (Mann-Whitney $U = 280.5$, $p = 0.0047$; Figure 2.7C; Table 2.5). There was no significant difference between the minimal and optional *in vitro* reporting standards (Mann-Whitney $U = 315$, $p = 0.28$; Table 2.5) or between the minimal and optional pre-clinical guidelines (Mann-Whitney $U = 179$, $p = 0.053$; Table 2.5).

It was also found that some of the clinical studies in the repositories did not fully report the metadata collected in the study. Instead the reporting standards were partially complied with by the reporting of "implicit" metadata. Rather than reporting, e.g. gender, as a factor for each individual sample, the metadata were reported as descriptive statistics in

Table 2.5 Mann-Whitney *U* tests comparing compliance between minimal and optional reporting standards within metabolomics repositories. *U* = test statistic, *p*-value = probability value.

Repository	Reporting Standards	*U*	*p*-value
MetaboLights	Clinical	118.5	2.02×10^{-5}*
MetaboLights	*in vitro*	341.0	0.33
MetaboLights	Pre-clinical	94.0	7.59×10^{-4}*
Metabolomics Workbench	Clinical	78.5	1.62×10^{-7}*
Metabolomics Workbench	*in vitro*	268.0	0.62
Metabolomics Workbench	Pre-clinical	107.5	0.0022*
GNPS	Clinical	280.5	0.15
GNPS	*in vitro*	315.0	0.28
GNPS	Pre-clinical	179.0	0.053

* indicates significant values

the corresponding publication to the study. For example gender is reported as "Gender (male/female): 15/29" in the associated publication [204] of MTBLS218[1], but the gender of individual samples is not reported. Gender was reported as implicit metadata in 32.76% of clinical studies in MetaboLights and in 6.93% clinical Metabolomics Workbench studies. Ethnicity and disease status were also both reported as implicit metadata in a number of studies (Table 2.6).

The plant guidelines were complied with at significantly different rates across the four repositories that contained *A. thaliana* studies (Kruskal-Wallis $H = 8.38$, $df = 2$, $p = 0.039$; Table 2.4). Studies in Metabolomics Workbench complied with the plant reporting standards significantly more than those in MeRy-B (Figure 2.8).

There was no significant difference between MetaboLights and Metabolomics Workbench in the rates of compliance to the minimal reporting standards (Kruskal-Wallis $H = 2.55$, $df = 1$, $p = 0.11$; Table 2.4), however the optional reporting standards were significantly more complied with in MetaboLights than Metabolomics Workbench (Kruskal-Wallis $H = 9.02$, $df = 1$, $p = 0.0027$; Table 2.4). Both MetaboLights (Kruskal-Wallis $H = 56.89$, $df = 1$, $p = 4.60 \times 10^{-14}$; Table 2.4) and Metabolomics Workbench (Kruskal-Wallis $H = 51.60$, $df = 1$, $p = 6.80 \times 10^{-13}$; Table 2.4) complied with the MSI minimum reporting standards at a higher rate than GNPS. There was also significantly higher compliance to the optional guidelines in MetaboLights (Kruskal-Wallis $H = 58.50$, $df = 1$, $p = 2.03 \times 10^{-14}$; Table 2.4)

[1] https://www.ebi.ac.uk/metabolights/MTBLS218, accessed 24th May 2017.

and Metabolomics Workbench (Kruskal-Wallis $H = 31.12$, $df = 1$, $p = 2.35 \times 10^{-8}$; Table 2.4), compared to GNPS.

Unlike the other repositories, Metabolomics Workbench has no requirement for sharing raw metabolomics data. Users are simply required to share a list of annotated metabolites, along with *m/z*-retention time values or binned chemical shift ranges for untargeted MS or NMR respectively. Only 67.95% (248/365) public studies in Metabolomics Workbench have associated raw data. In this analysis 61.61% human clinical, 71.11% human cell line, 74.73% *M. musculus* preclinical and 33.33% *A. thaliana* studies included raw data.

Between the repositories there are also notable differences in the percentage of studies that include associated publications. In MetaboLights, 96.55% of clinical, 93.10% of preclinical, 94.44% of *in vitro* and 75% of plant studies have an associated publication. 12.12% of the clinical studies in Metabolomics Workbench have an associated publication, and 29.67% of preclinical, 22.22% of *in vitro* and 33.33% of plant studies do. All of the studies on MetaPhen have an association journal article (see section 3.1). 50% of *A. thaliana* studies on MeRy-B have an associated publication. No GNPS clinical or *in vitro* studies have an associated publication and 16.67% of preclinical studies do.

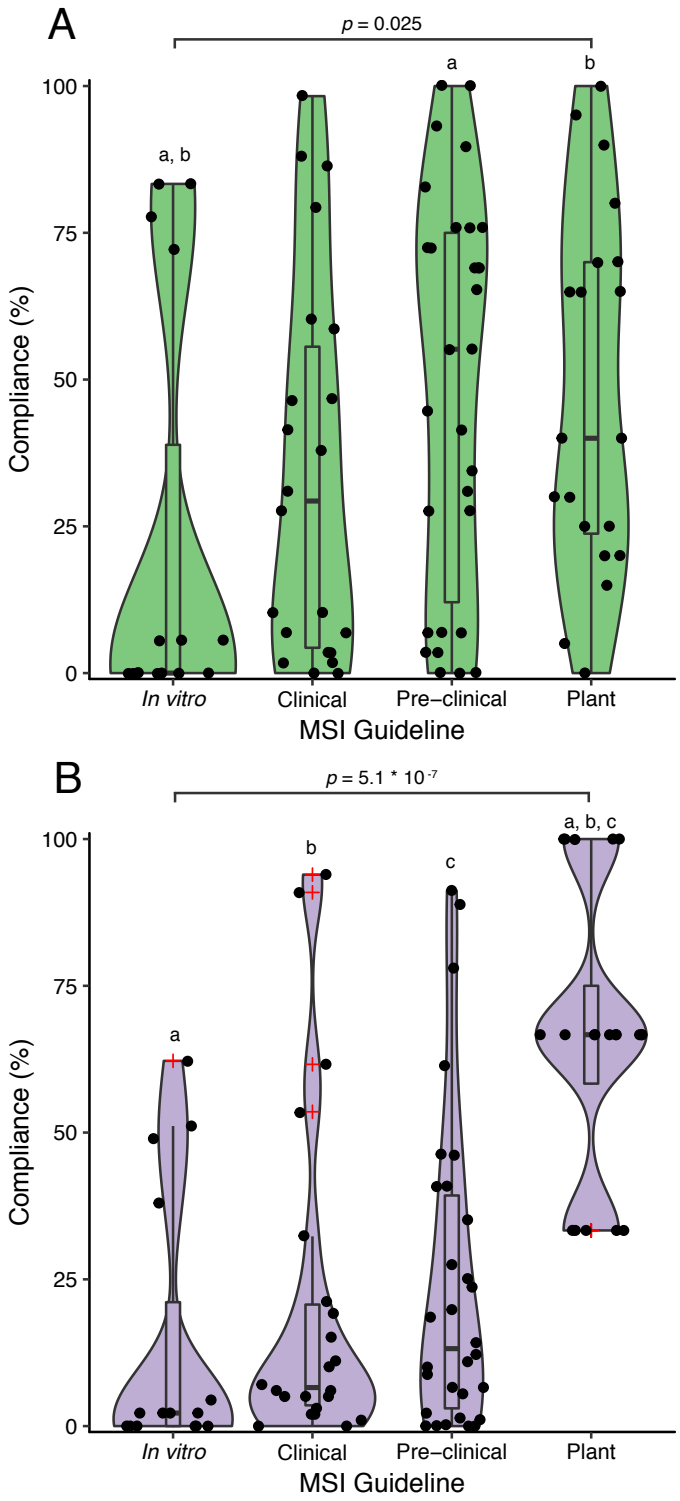


Fig. 2.4 Compliance to the MSI minimum reporting standards. Combined violin and dot plots showing the percentage compliance with the MSI minimum reporting standards within the (A) MetaboLights and (B) Metabolomics Workbench repositories. Red '+' indicate outliers; each 'dot' represents compliance to a single reporting standard. Letters denote significant differences in compliance (Kruskal Wallis tests, followed by Dunn post-hoc test with Benjamini-Hochberg correction). Figure was adapted from Spicer *et al.* (2017) [169].

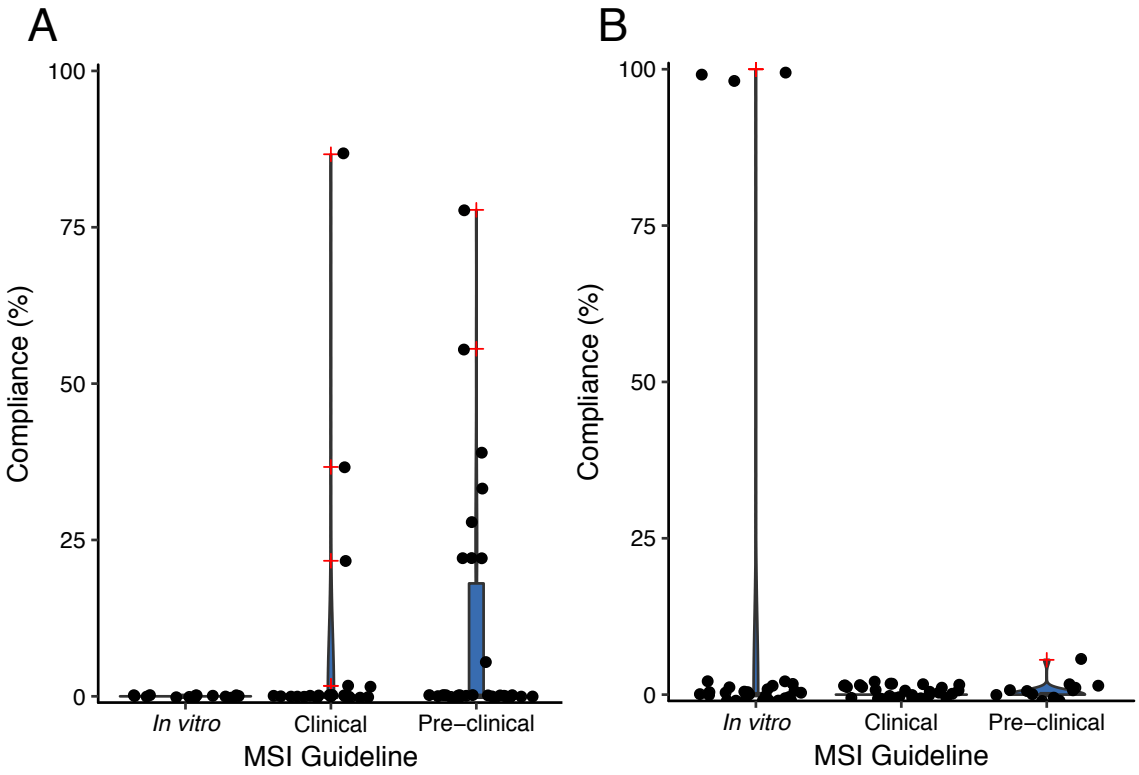


Fig. 2.5 Compliance to the MSI reporting standards in the GNPS Repository. Combined violin and dot plots showing the percentage compliance with the MSI (A) minimum and (B) optional standards. Red '+' indicate outliers; each 'dot' represents compliance to a single reporting standard. No significant difference between compliance to either the minimal or optional standards was found using Kruskal Wallis tests.

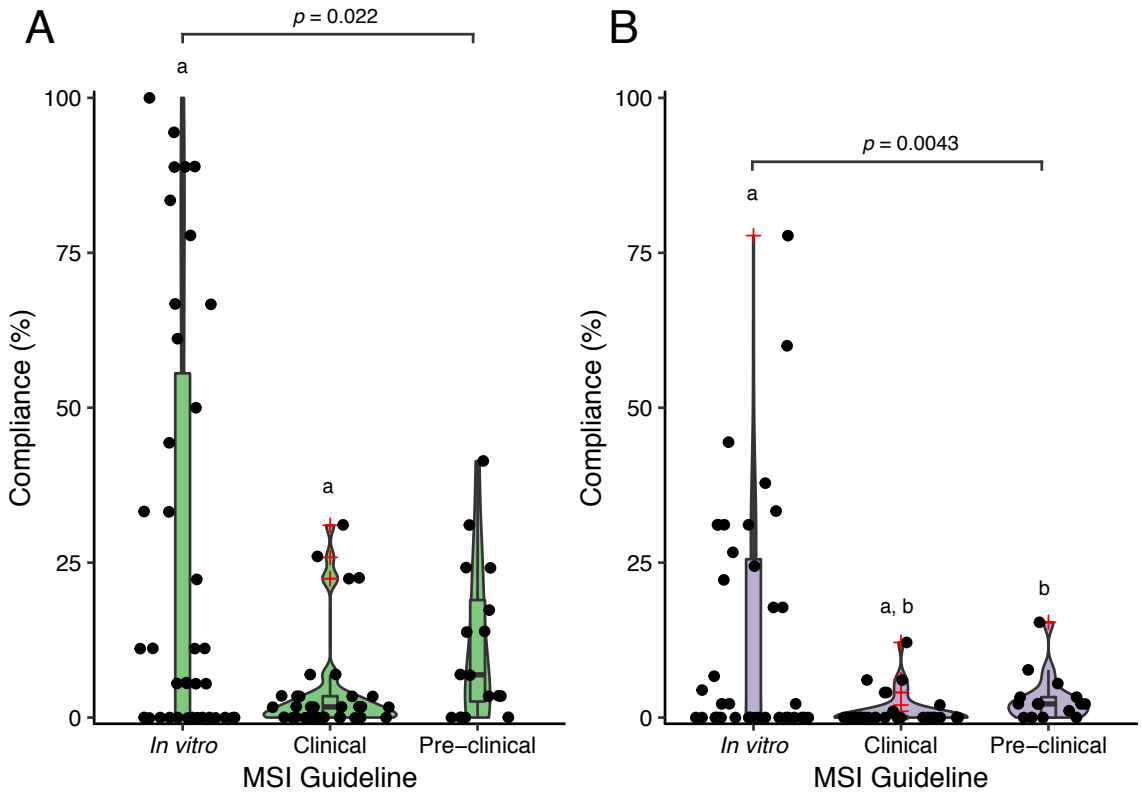


Fig. 2.6 Compliance to the MSI optional reporting standards. Combined violin and dot plots showing the percentage compliance with the MSI optional reporting standards within the (A) MetaboLights and (B) Metabolomics Workbench repositories. Red '+' indicate outliers; each 'dot' represents compliance to a single reporting standard. Letters denote significant differences in compliance (Kruskal Wallis tests, followed by Dunn post-hoc test with Benjamini-Hochberg correction). Figure was adapted from Spicer *et al.* (2017) [169].

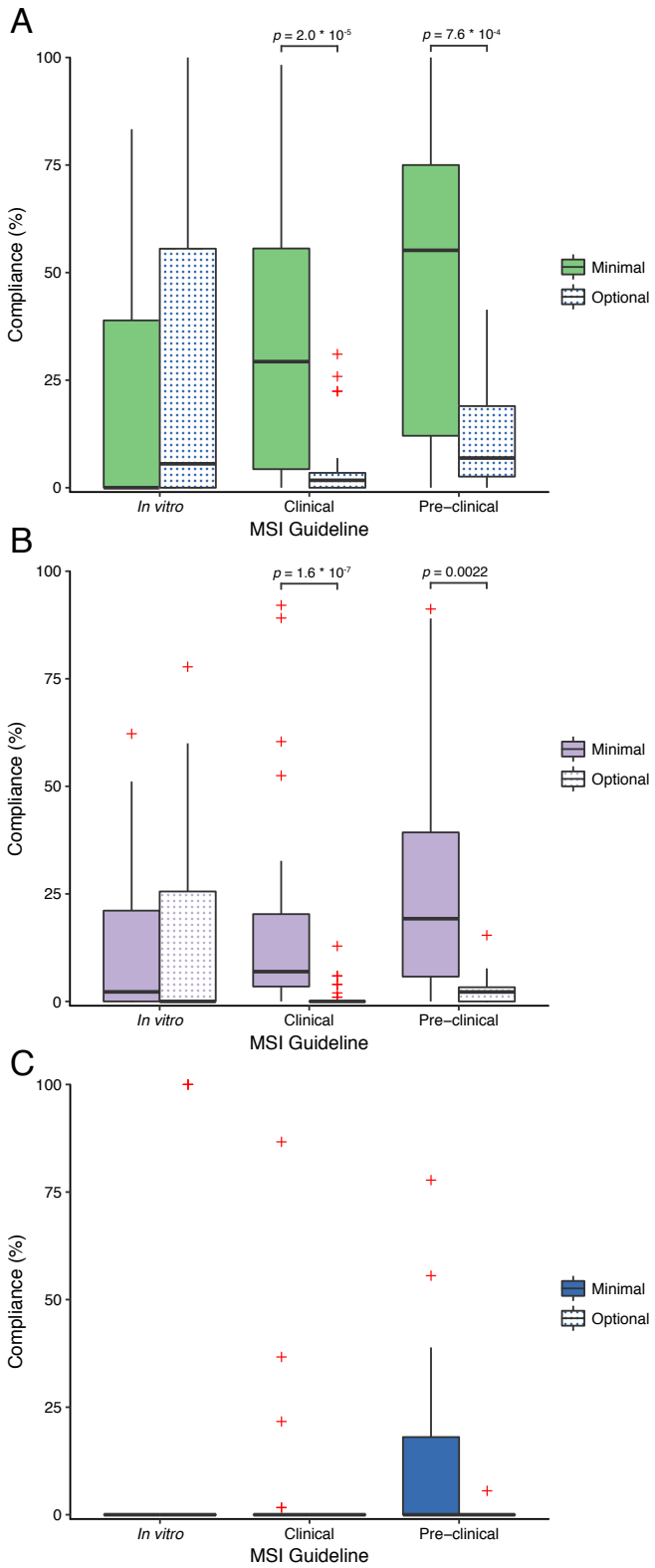


Fig. 2.7 Comparison of compliance to minimal and optional reporting standards. Box-and-whisker plots showing the percentage compliance to the minimal and optional reporting standards in (A) MetaboLights, (B) Metabolomics Workbench and (C) GNPS. Minimal reporting standards are indicated with filled bars, optional reporting standards are indicated with a dot pattern and red '+' indicate outliers. Mann- Whitney *U* tests were used to assess significance. Figure was adapted from Spicer *et al.* (2017) [169].

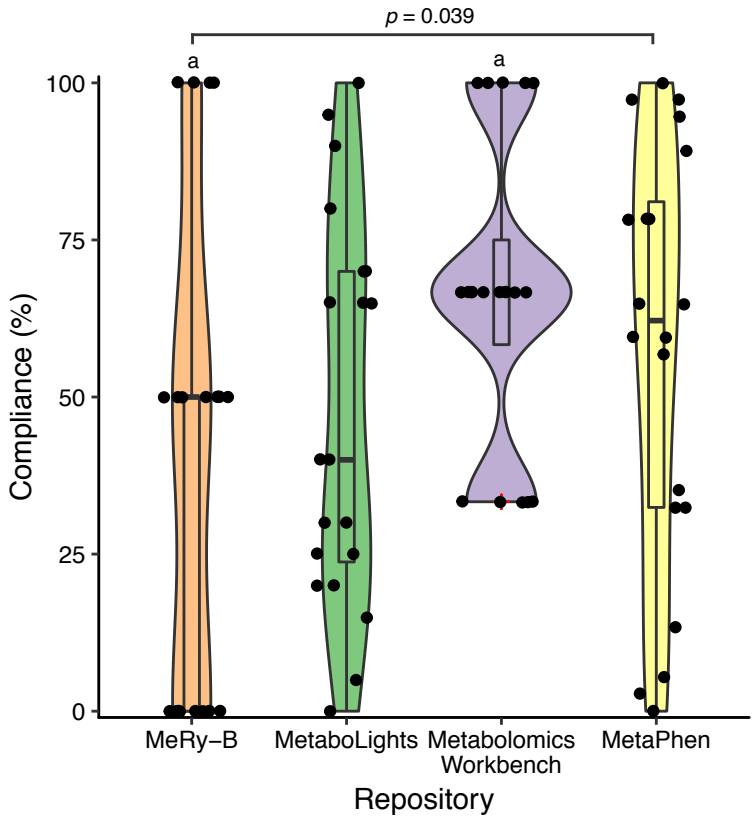


Fig. 2.8 Compliance to Plant MSI Reporting Standards. Combined violin and dot plots showing the percentage compliance of A. thaliana studies to the plant MSI minimum reporting standards within the MeRy-B, MetaboLights, Metabolomics Workbench and MetaPhen repositories. Red '+' indicate outliers; each "dot" represents compliance to a single reporting standard. Letters denote significant differences in compliance (Kruskal Wallis, Dunn post-hoc test with Benjamini-Hochberg correction). Figure was adapted from Spicer *et al.* (2017) [169].

Table 2.6 The percentage of *Homo sapiens* studies in each repository that comply with each mammalian clinical trials and human studies minimal reporting standard. Reported refers to metadata that are directly reported in either the study or an associated publication. Implicit refers to metadata that were reported either in a table of descriptive statistics or controlled for during statistical analysis but were not reported for each individual sample.

Minimal Reporting Standard	Compliance (%)					
	MetaboLights ($n = 58$)		Metabolomics Workbench ($n = 99$)		GNPS ($n = 60$)	
	Reported	Implicit	Reported	Implicit	Reported	Implicit
Biofluid or Tissue	98.28	-	90.91	-	86.67	-
Ethical Approval	87.93	-	15.15	-	0.00	-
Number of Groups	86.21	-	93.94	-	36.67	-
Age Range	79.31	-	19.19	-	0.00	-
Gender	46.55	32.76	3.03	6.93	1.67	0.00
Disease Status	60.34	5.17	61.62	3.96	21.67	5.00
Sample Storage Temperature	58.62	-	53.54	-	0.00	-
Weight range and Height and/or BMI	46.55	-	32.32	-	0.00	-
Inclusion Criteria	41.38	-	5.05	-	0.00	-
Exclusion Criteria	37.93	-	7.07	-	0.00	-
Fasting Status	31.03	-	21.21	-	0.00	-
Volume or Quantity of Collection	27.59	-	4.95	-	0.00	-
Ethnicity	6.90	13.79	6.06	1.98	1.67	0.00
Anticoagulant	10.34	-	10.10	-	0.00	-
Trial Type	10.34	-	1.01	-	0.00	-
Location of Collection	2.02	-	1.98	-	0.00	-
Treatment	3.45	-	11.11	-	0.00	-
Treatment Dose	3.45	-	6.06	-	0.00	-
Treatment Duration	1.72	-	5.05	-	0.00	-
Treatment Route	1.72	-	2.02	-	0.00	-
Bacteriostatic Agent	0.00	-	0.00	-	0.00	-
Treatment Vehicle	0.00	-	0.00	-	0.00	-
Mean	33.93	17.24	20.52	4.29	6.74	1.67
Median	29.31	13.79	6.57	3.96	0.00	0.00

Table 2.7 The percentage of *Homo sapiens* studies in each repository that comply with each microbial and *in vitro* minimal reporting standard.

Minimal Reporting Standard	Compliance (%)		
	MetaboLights (n = 18)	Metabolomics Workbench (n = 45)	GNPS (n = 3)
Metabolism Quenching Method	83.33	62.22	0.00
Harvesting Method	83.33	48.89	0.00
Metabolite Extraction	77.78	51.11	0.00
Sample Storage	72.22	37.78	0.00
Normalisation by Cell Number	5.56	2.22	0.00
Cell Integrity	5.56	0.00	0.00
Stability	5.56	0.00	0.00
Temperature from Sampling to Quenching	0.00	4.44	0.00
Extracellular Metabolites Discriminated	0.00	2.22	0.00
Recovering from Extraction	0.00	2.22	0.00
Time Until Quenching	0.00	2.22	0.00
Detection Limit	0.00	0.00	0.00
Quality Control	0.00	0.00	0.00
Sample Clean-up	0.00	0.00	0.00
Sample Storage Duration	0.00	0.00	0.00
Mean	22.22	14.22	0.00
Median	0.00	2.22	0.00

Table 2.8 The percentage of *Mus musculus* studies in each repository that comply with each pre-clinical minimal reporting standard.

Minimal Reporting Standard	Compliance (%)		
	MetaboLights (n = 29)	Metabolomics Workbench (n = 91)	GNPS (n = 18)
Biofluid or Tissue	100.00	89.01	77.78
Number of Groups	75.86	91.21	55.56
Strain	100.00	61.54	22.22
Treatment	75.86	78.02	38.89
Sex	93.10	46.15	22.22
Collection Time	89.66	40.66	33.33
Age at Collection or Euthanization	68.97	46.15	27.78
Sample Storage Temperature	65.52	35.16	0.00
Age at Study Start	75.86	24.18	22.22
Treatment Duration	72.41	27.47	0.00
Animal Supplier	82.76	10.99	0.00
Treatment Dose	72.41	19.78	0.00
Diet	44.83	40.66	0.00
Treatment Route	68.97	14.29	0.00
ad lib or Restricted Diet	68.97	6.59	0.00
Light Cycle	55.17	9.89	0.00
Treatment Vehicle	55.17	6.59	0.00
Euthanasia Method	41.38	18.68	0.00
Tissue Processing	27.59	25.27	0.00
Group or Individual Housing	31.03	5.49	0.00
Fasting Status	27.59	8.79	0.00
Weight Range	34.48	0.00	0.00
Volume or Quantity of Sample Collection	6.90	12.09	0.00
Anticoagulant	6.90	0.00	0.00
Tap or Purified Water	6.90	0.00	0.00
Collection Frequency	3.45	1.10	0.00
Collection Method	3.45	1.10	0.00
Location of Sample Collection	0.00	2.20	0.00
Bacteriostatic Agent	0.00	0.00	0.00
Collection Duration	0.00	0.00	0.00
Mean	48.51	24.10	10.19
Median	55.17	13.19	0.00

Table 2.9 The percentage of *Arabidopsis thaliana* studies in each repository that comply with each plant minimal reporting standard. Table is reproduced from the supplementary material of Spicer *et al.* (2017) [169].

Minimal Reporting Standard	Compliance (%)			
	MeRy-B (n = 2)	MetaboLights (n = 20)	Metabolomics Workbench (n = 3)	MetaPhen (n = 37)
Organ or Cell Type	100.00	100.00	100.00	94.59
Genotype	100.00	70.00	66.67	100.00
Plant Growth Stage	100.00	65.00	100.00	78.38
Metabolism Quenching Method	100.00	30.00	100.00	64.86
Growth Support	50.00	95.00	66.67	97.30
Light	50.00	90.00	66.67	97.30
Date(s) of Plant Establishment	50.00	80.00	66.67	89.19
Temperature	50.00	70.00	66.67	78.38
Biosource Amount	50.00	65.00	66.67	13.51
Nutrients Regime	50.00	40.00	66.67	64.86
Sample Storage	50.00	40.00	66.67	32.43
Harvest Method	50.00	30.00	100.00	35.14
Harvest Time, Date	0.00	65.00	100.00	78.38
Treatment	0.00	25.00	33.33	59.46
Treatment Time	0.00	25.00	33.33	56.76
Humidity	0.00	20.00	66.67	32.43
Treatment Dose	0.00	20.00	33.33	59.46
Growth Location	0.00	15.00	33.33	0.00
Plot Design	0.00	5.00	66.67	5.41
Watering Regime	0.00	0.00	33.33	2.70
Mean	40.00	47.50	66.67	57.03
Median	50.00	40.00	66.67	62.16

Table 2.10 The percentage of *Homo sapiens* studies in each repository that comply with each mammalian clinical trials and human studies optional reporting standard.

Optional Reporting Standard	Compliance (%)		
	MetaboLights (n = 58)	Metabolomics Workbench (n = 99)	GNPS (n = 60)
Arterial or Venous Blood	31.03	4.04	0.00
Speed of Centrifugation	25.86	6.06	0.00
Temperature of Centrifugation	22.41	6.06	0.00
Time of Centrifugation	22.41	4.04	0.00
Smoking Status	6.90	12.12	0.00
Time from Collection to Freezing	6.90	2.02	0.00
Drug Consumption	3.45	1.01	0.00
Alcohol Consumption	3.45	0.00	0.00
Hemoglobin	3.45	0.00	0.00
Platelets	3.45	0.00	0.00
White Blood Count	3.45	0.00	0.00
Creatinine	1.72	0.00	0.00
Diet	1.72	0.00	0.00
Hemocrit	1.72	0.00	0.00
Malnutrition	1.72	0.00	0.00
Mid Flow or Total Urine	1.72	0.00	0.00
Potassium	1.72	0.00	0.00
Sample Storage Duration	1.72	0.00	0.00
Sodium	1.72	0.00	0.00
Metal Exposure	0.00	0.00	0.00
Albumin	0.00	0.00	0.00
ALP	0.00	0.00	0.00
ALT	0.00	0.00	0.00
Bilirubin	0.00	0.00	0.00
Glucose	0.00	0.00	0.00
-GT	0.00	0.00	0.00
HDL Cholesterol	0.00	0.00	0.00
Hemolysis	0.00	0.00	0.00
LDL Cholesterol	0.00	0.00	0.00
Total Cholesterol	0.00	0.00	0.00
Total Protein	0.00	0.00	0.00
Triglycerides	0.00	0.00	0.00
Urea	0.00	0.00	0.00
Mean	4.44	1.07	0.00
Median	1.72	0.00	0.00

Table 2.11 The percentage of *Homo sapiens* studies in each repository that comply with each microbial and *in vitro* best practice reporting standard.

Best Practice Reporting Standard	Compliance (%)		
	MetaboLights (n = 18)	Metabolomics Workbench (n = 45)	GNPS (n = 3)
Cell Type	100.00	77.78	100.00
Treatment	88.89	60.00	100.00
Treatment Dose	88.89	44.44	0.00
Medium or Substrate	94.44	37.78	0.00
Medium or Substrate Concentration	88.89	31.11	0.00
Treatment Time	77.78	33.33	100.00
Treatment Vehicle	83.33	26.67	0.00
Growth Container	61.11	24.44	0.00
Cell Supplier	66.67	17.78	0.00
Medium or Substrate Supplier	66.67	17.78	0.00
CO_2	50.00	31.11	0.00
Temperature	44.44	22.22	0.00
Harvesting Time	22.22	31.11	0.00
Isotopic Labelling	11.11	31.11	0.00
Replicates	33.33	6.67	0.00
Subculturing and Splitting Protocols	33.33	0.00	0.00
Inoculation Size	11.11	2.22	0.00
Growth Support	11.11	0.00	0.00
Immortalized or Transformed	11.11	0.00	0.00
pO_2	5.56	4.44	0.00
pH	5.56	2.22	0.00
Additional -omics Datasets	5.56	0.00	0.00
Growth Container Supplier	5.56	0.00	0.00
Growth Support Supplier	5.56	0.00	0.00
Humidity	0.00	2.22	0.00
Evaporation	0.00	0.00	0.00
Gas Composition	0.00	0.00	0.00
Growth Configuration	0.00	0.00	0.00
Growth Rate	0.00	0.00	0.00
Harvesting Cell Density	0.00	0.00	0.00
Harvesting Depletion of Nutrients	0.00	0.00	0.00
Harvesting Growth Phase	0.00	0.00	0.00
Marker of Differentiated Stage	0.00	0.00	0.00
Number of Generations Until Harvesting	0.00	0.00	0.00
Number of Culture Passages	0.00	0.00	0.00
Pretreatment	0.00	0.00	0.00
Pretreatment Time	0.00	0.00	0.00
Stabilization Time	0.00	0.00	0.00
Stirrer Speed	0.00	0.00	0.00
Mean	28.32	13.60	7.32
Median	11.11	2.22	0.00

Table 2.12 The percentage of *Mus musculus* studies in each repository that comply with each pre-clinical optional reporting standard.

Optional Reporting Standard	Compliance (%)		
	MetaboLights ($n = 29$)	Metabolomics Workbench ($n = 91$)	GNPS ($n = 18$)
Use of Anesthesia	41.38	7.69	0.00
Environmental Enrichment: Temperature	31.03	2.20	0.00
Acclimation Duration to Experimental Facility	24.14	3.30	0.00
Germ-free or Conventional Housing	24.14	2.20	5.56
Fasting Duration	17.24	5.49	0.00
Bedding Type	6.90	15.38	0.00
Environmental Enrichment: Humidity	13.79	2.20	0.00
Anesthesia Time	13.79	0.00	0.00
Anesthesia Dose	6.90	2.20	0.00
Cage Cleaning Frequency	3.45	0.00	0.00
Cage Type	3.45	0.00	0.00
Inclusion Criteria	3.45	0.00	0.00
Additional Phenotypic Model	0.00	3.30	0.00
Sample Storage Duration	0.00	2.20	0.00
Temperature of Collection Tube	0.00	1.10	0.00
Body Weights or Food Consumption	0.00	0.00	0.00
Mean	11.85	2.95	0.35
Median	6.90	2.20	0.00

2.4 Evaluation of MSI Biological Metadata Standards

The level of compliance to the different sets of reporting standards varies greatly across public repositories, from 0–97%. However, the majority of studies have very low compliance with the relevant MSI standard of their respective application domain.

Overall, across repositories, the plant minimal reporting standards were complied with at the highest rate and the microbial and *in vitro* standards the lowest. The greater rate of compliance to the plant guidelines may be in part be due the exactness of their wording. Whilst all of the MSI biological context subgroup reports encourage the use of ontologies, the plant guidelines detail precisely which ontologies and taxonomies should be used to describe each part of the biosource, including species, genotype, organ and cell type [191]. The specificity of the plant reporting standards makes them easy to use and may contribute to the higher compliance to them.

Levels of compliance to the minimal and best practice in the microbial and *in vitro* reporting standards are similar, with the minimal standards having very low median compliance compared to the other minimal reporting standards. This may stem from how minimal and best practice reporting standards are defined by the sub-working group. The minimal reporting standards include only metabolomics specific factors, whilst all other general aspects, as well as additional factors specific to metabolomics experiments, are included in the best practice reporting standards [190]. General experimental factors, such as cell type and treatment are included as best practice and not minimal reporting standards. This is in contrast to all other sub-groups' guidelines, where equivalent reporting standards are included as minimal rather than best practice. This discrepancy between standards may be confusing to researchers.

The lower levels of compliance to the microbial and *in vitro* minimal reporting standards, compared to the other guidelines, may result from researchers not understanding what they are expected to report. The standards are ambiguous and it is unclear precisely what needs to be communicated in order to comply with them. The stability reporting standard asks "What is known about the stability of (specific) metabolites during quenching, extraction and sample preprocessing?" [190]. This is cryptic and hard for a practitioner to interpret.

Differences in the rates of partial adherence with guidelines by reporting "implicit metadata" between MetaboLights and Metabolomics Workbench may be accounted for by the percentage of studies included in the analysis that had an associated publication. Only 12% of human clinical studies in Metabolomics Workbench had an associated publication, in contrast to 96% of those in MetaboLights that did. This is also true for mouse pre-clinical

studies, where 30% of Metabolomics Workbench studies had an associated publication, whereas 93% of MetaboLights studies did.

The more stringent submission criteria to MetaboLights than to Metabolomics Workbench may result in greater compliance to the MSI guidelines in specific instances e.g. gender in clinical *H. sapiens* studies (Table 2.6) and age at study start in pre-clinical *M. musculus* studies (Table 2.8).

The significant difference in compliance to the plant reporting standards between repositories may be due to low sample size biasing the Kruskal-Wallis test. There are only two *A. thaliana* studies in MeRy-B and three in Metabolomics Workbench, less than the five per group required for Kruskal-Wallis p-values to be accurate. It is also worth noting that the paper accompanying one of the studies in Metabolomics Workbench specifically serves as a demonstration of how to adhere to the MSI guidelines [195], so it would be expected that this study would fully comply with them.

As GNPS has very minimal metadata reporting requirements, and does not aim to fulfil MSI guidelines, it is unsurprising that GNPS studies have very low conformity with them. The only metadata that users are required to report are species, instrument, keywords, principle investigator and post-translational modification. The last of these is not applicable to metabolomics data and is required as MassIVE is primarily a proteomics data repository. Whilst only 20.9% of minimal reporting standards are complied with by any of the examined GNPS studies, any metadata that researchers choose to report is of interest, as doing so is voluntary and can be very time consuming. It can be assumed that researchers' consider any metadata reported as essential for the understanding of the study. Most of the metadata reported in GNPS studies relates to biosource or treatment. Biosource and treatment metadata are also highly reported in other repositories, signifying their importance.

These results indicate that the MSI biological context guidelines do not fulfill the current needs of the metabolomics community. I believe that the MSI guidelines should be revisited and revised. Consideration should be given to the minimum amount of metadata required to be able to a) repeat an experiment and b) re-analyse the data, to enable the maximum amount of information that can be extracted from the data. The community, along with data curators, industry representatives, publishers and funders should be consulted. In the following sections I shall discuss further problems with the MSI guidelines (Section 2.5), developments in reporting standards for metabolite identification (Section 2.6.1), data exchange formats (Section 2.6.2), and toxicology (Section 2.6.3), and suggest improvements for future revisions (Section 2.7).

2.5 Other Criticisms of Existing MSI Standards

One of the largest problems with the MSI standards is the lack of unified description as to their use [196]. Indeed scientists must choose which biological context metadata guidelines are best suited for describing their research. This could be a disheartening prospect and may put researchers off using the standards.

Researchers may also find the sheer number of sets of standards that they are expected to comply with daunting. For NMR metabolomics experiments, in theory researchers should comply with the relevant biological metadata context guidelines, the chemical analysis working group (CAWG) guidelines [186], the NMR guidelines [187] and the data analysis guidelines [188]. This may seem overwhelming, and researchers may simply abandon attempting to use the standards at this point.

Despite the seemingly large number of standards, there is actually much overlap between the different guidelines: all of the biological context metadata subgroups reporting standards also including sample processing methods. This redundancy between the biological context metadata subgroups and the CAWG reporting standards [186] inflates the total number of reporting standards. However, it is simply caused by the duplication of standards in multiple guidelines. There is a particularly high amount of duplication between the microbial and *in vitro* biology [190] and CAWG guidelines, where the majority of minimal *in vitro* reporting standards are also included in the CAWG minimal reporting standards.

Reporting only age and weight (and Height and/or BMI for clinical studies) ranges, and not the age and weight corresponding to each individual sample, is required by the minimal mammalian/ *in vivo* standards [189]. This means that a study with publicly available data can fully comply with the reporting standards but not be reusable. Age and weight (or BMI) differences are frequently controlled for using a nested case-control study design [205, 206]. However, if this crucial metadata is not reported for individual samples further research and meta-analysis using the data is limited.

Since the MSI guidelines were written, there is now far greater understanding of the metabolome, and what effects it. Factors such as gender [10, 11, 207], age [4], BMI [8, 9], smoking status [15] and nutrition [13] have all been identified as having significant effects on the human metabolome. At the time the original mammalian/ *in vivo* reporting standards were published, it was unclear whether confounding factors were study dependent [189], and thus the guidelines around these areas are hazy. Whilst it appears that the intention of the MSI was to require practitioners to fully report all studied or controlled-for factors, this is not clearly stated.

Complementary to this analysis, Considine *et al.* (2018) [208] investigated compliance to the data analysis MSI reporting standards [188]. They performed a systematic review of the reporting of data pre-treatment and data analysis (algorithm selection, univariate analysis and multivariate analysis) of metabolomics biomarker discovery studies. Of the 27 identified journal articles, only one study had the potential to be reproducible, but, as it did not contain linked executable code and data, it was classified as partially reproducible. No study provided a clear description of exactly which analysis steps were performed and in which order. All but one of the analysed studies had large omissions in reporting of statistical analysis — only one study provided sufficient detail that an analysis workflow was able to be extrapolated.

This research specifically focused on the studies that were most likely to adhere to the MSI standards, those who chose to share their data openly, on an MSI compliant repository (with publicly available studies on GNPS assumed not to comply with the standards). Considine *et al.* [208] performed an unbiased systematic review; none of the studies included in their research shared data openly on a repository. If a systematic review of compliance to the MSI biological context metadata standards were to be conducted, it is likely that a far lower level of compliance would be found than in this research, as the studies that were selected for inclusion in this study were likely to be the most open.

The research by Considine *et al.* [208] suggests that, like the biological context metadata standards, the data analysis MSI standards are confusing and unclear. They echo our call [170] for revision of the MSI guidelines.

The ultimate responsibility for enforcing compliance to reporting standards belongs to journals and data hosts [209]. If repositories do not comply with a standard, it will not be successful [210]. From this research, it is clear that even repositories that were designed to facilitate compliance with the MSI guidelines, do not in actuality enforce adherence to the reporting standards. However, if a standard is out of date and not beneficial to the community, then full compliance should not be required.

Compliance to an entire set of reporting standards is very time-consuming. It has not yet become universal practice for grant awarders to provide funding specifically to allow researchers time to ensure data is in a usable format, with sufficient metadata, appropriate for sharing. Researchers may feel that the effort required to comply with reporting standards is wasted time, or be put off sharing their data by ambiguous wording or difficulty in obtaining metadata. Unless experiments and technological setup are carefully planned to take into account later data and metadata submission procedure into account, it is cumbersome to convert lab book entries into standardised formats, such as ISA (Investigation/Study/Assay) [129].

Repositories, therefore, have a trade-off between attracting users to submit their data and enforcing reporting standards to ensure deposited data is informative. Especially during their infancy repositories will not vigorously enforce compliance with standards out of fear of annoying submitters. However, data alone is not inherently reusable and must be accompanied by rich metadata to be so [211]. As a database matures it may be found that much of the data it contains is not reusable due to insufficient metadata.

To date, metabolomics repositories have chosen to position themselves at different points on an ease-of-use versus reusability spectrum, and enforce compliance with standards to varying degrees. Of the three largest repositories, GNPS focuses the most on attracting users and least on collecting metadata, which can hinder reusability. Conversely, MetaboLights has the highest metadata requirements, but also a lower number of submissions, likely to be because of these more stringent requirements. Despite four metabolomics repositories stating they aim to follow the MSI guidelines, across repositories there is no agreement as to required metadata.

A well written reporting standard is easy to use, aids researchers in publishing their data and ensures consistency between datasets [114]. The current MSI standards are badly worded and have failed to achieve this. The community has mostly ignored the guidelines, and this has hindered the reproducibility of metabolomics research [208].

2.6 Areas with Advancement

However, in some areas, flaws of the existing MSI reporting standards have been identified and steps have been taken to improve the standards. In the following section I shall discuss these areas and new initiatives relating to reporting standards that have been developed.

2.6.1 Metabolite Identification

The current MSI metabolite identification criteria is comprised of four levels (Table 2.13) [186]. For a metabolite to be considered identified (Level 1) it must be matched to an authentic chemical standard that has been analysed in the same laboratory, using two orthogonal analytical techniques [212]. If no chemical reference standards are available, the highest level of annotation possible is Level 2, putatively annotated compounds. Level 2 annotation can be achieved by matching spectra by similarity to spectral libraries and/or based upon physicochemical properties. Features annotated at Level 3 are not unambiguously identifiable as a specific metabolite and are instead annotated with a putatively characterized compound

Table 2.13 Levels of metabolite identification under the currently used MSI criteria proposed by the CAWG in 2007 [186].

Level	Classification
Level 1	Identified Compound
Level 2	Putatively Annotated Compound
Level 3	Putatively Characterised Compound Class
Level 4	Unknown Compound

class, using the same techniques as for Level 2 annotation. Unknown compounds are categorised as Level 4. These are typically reported as m/z and retention time for MS.

Similarly to the findings for compliance to the MSI biological context metadata standards in this chapter, Salek *et al.* [212] found poor reporting of the levels of confidence in metabolite identification. Of 20 randomly selected metabolomics studies published in 2013, only 6 defined how metabolites were annotated or identified, with one including relevant metadata. None of the selected articles used the MSI criteria to report metabolite identification.

A problem with the MSI criteria is that the system is coarse, and the levels are not precisely defined [213]. In many cases identification confidence fits in between the MSI proposed levels [214]. To address this issue, Schymanski *et al.* [214] proposed alternative metabolite identification criteria for high resolution-mass spectrometry (HR-MS) that also included minimum data requirements for each level (Table 2.14). Whilst Level 1 remains unchanged compared to the original MSI criteria, levels 2–5 are different. Probable structure (Level 2) annotation requires either a library spectrum match (2a) or diagnostic evidence (2b). When there is evidence for more than one candidate structure, with inadequate information to narrow identification down to a single structure, identification confidence is Level 3, tentative candidate(s). Annotation of unequivocal molecular formula (Level 4) requires the use of spectral information for unambiguous assignment. If an exact mass (m/z) can be measured, but there is no information to assign even a molecular formula, a feature can be identified to confidence Level 5.

The lack of use of the metabolite identification guidelines and advancement within the community, led the Metabolite Identification task group of the Metabolomics Society to reassess the MSI reporting standards [215]. They proposed three potential revisions: 1) amending the original classification to include sub levels, similarly to Schymanski *et al.* [214], 2) a quantitative scoring system instead of the current system, or 3) a quantitative scoring system to enhance existing MSI levels. The community was called upon to provide comments on the proposed revisions.

Table 2.14 Levels of metabolite identification proposed by Schymanski *et al.* (2014) [214]. MS^2 is used to represent all kinds of MS fragmentation.

Level	Classification	Minimum Data Requirements
Level 1	Confirmed structure	MS, MS^2, RT, Reference Standard
Level 2	Probable structure	
	Level 2a) Library	MS, MS^2, Library MS^2
	Level 2b) Diagnostic	MS, MS^2, Experimental Data
Level 3	Tentative candidate(s)	MS, MS^2, Experimental Data
Level 4	Unequivocal molecular formula	MS isotope/adduct
Level 5	Exact mass	MS

Sumner *et al.* [216] proposed a unified identification point system, based on the the sum of all the data types supporting the identification (Table 2.15). If matching to a standardised reference compound was used, the score would be doubled and the sum would be multiplied by 1.5 if matching to literature values or other externally generated spectral libraries due to lower confidence. They suggest using a minimum score of 5.0 for a confident metabolite identification. Sumner *et al.* also suggested an alternative scoring system using alphanumeric identifier strings that summarise the data types used for an identification. This would be similar to SMILES (Simplified molecular-input line-entry system) or InChi (International Chemical Identifier) chemical identifier codes, which provide standardised methods for encoding molecular information in strings.

The requirement of comparison to an authentic reference standard for a Level 1 metabolite identification under the original MSI criteria has been criticised as being too strict for NMR-based studies [213]. Features can be identified with high confidence using database matching to authentic reference compounds [213, 217], not requiring spectra from an authentic reference standard to be analysed using the same NMR spectrometer.

Everett proposed new indices for NMR metabolite identification: metabolite identification efficiency (MIE) and metabolite identification carbon efficiency (MICE) [213]. Both of these methods rely on gathering multiple pieces of metabolite identification information (MII), such as the number of carbon, hydrogen, oxygen, nitrogen and sulfur atoms, number of chiral centres and number of proton chemical shifts. For MIE the number of pieces of metabolite information is then divided by the number of heavy atoms in the metabolite and for MICE it is divided by the number of carbon atoms in the metabolite. These methods have the disadvantage of requiring researchers to manually judge how well the MII pieces represent the entire metabolite molecular structure.

Table 2.15 Quantitative scoring system for metabolite identification proposed by Sumner *et al.* (2014) [216].

Data Type	Score
IR absorbance spectrum	0.5
UV absorbance spectrum	0.5
Retention time (±2.5%)	1.0
High resolution retention time (±0.5%, $W_{1/2} < 10$ s)	1.5
High resolution retention index (±0.5%, RI ± 25, $W_{1/2} < 10$ s)	2.0
Nominal mass of parent ion	0.5
Accurate mass of parent ion (<5 ppm)	1.0
Molecular formula based upon accurate m/z and isotope pattern	1.0
Confident EI Spectral match to commercial library	1.0
Tandem mass spectrum	1.5
Accurate mass tandem mass spectrum	2.0
^{1}H 1D NMR	2.0
^{1}H 2D NMR	3.0
^{1}H-^{13}C 2D NMR	4.0

The MICE approach has since been expanded upon to include topological analysis: topological metabolite identification carbon efficiency (tMICE) [218]. This method improves upon MICE by measuring the MII for each separate molecular topology element of the metabolite structure, removing the need for manual judgment of how well the MII measured represented the entire molecular structure of the metabolite.

Schober *et al.* [219] proposed a pattern-based taxonomy for describing the evidence supporting identification of a metabolite. The overall evidence naming pattern of the taxonomy is structured as: MolecularStructureElement [*annotation relation*] AssayOutcome *used_in* AssertionMethod. This method would allow multiple pieces of evidence to be described for each metabolite, providing greater granularity. The authors have released an ontology for implementing this taxonomy: Metabolite Identification Evidence Code Ontology (MIECO)[2].

The PhenoMeNal consortium proposed Metabolite identification as the Elixir Metabolomics Use Case [173]. This proposal was recently accepted[3], and the PhenoMeNal consortium will work towards developing analysis workflows that include metabolite identification.

Work by the Metabolite Identification task group remains ongoing. To date, no official new reporting standards for metabolite identification have been announced and their design is

[2]https://github.com/DSchober/MIECO, accessed 10th February 2018.

[3]https://www.elixir-europe.org/news/elixir-establish-new-use-cases-proteomics-metabolomics-and-galaxy, accessed 10th February 2018.

still under discussion. However, of all the areas covered by the original MSI guidelines, the best progress towards updating the reporting standards has been in metabolite identification.

2.6.2 Data Exchange Formats

The COSMOS initiative [140] built on work by the MSI [185], to further improve data exchange formats for raw MS and NMR data. The mzML [220] and mzQuantML [221] formats, which were originally developed by the Human Proteome Organization's (HUPO) Protein Standards Initiative (PSI) [179], were extended so that they would meet the requirements of reporting metabolomics MS experiments.

COSMOS and subsequently PhenoMeNal has been instrumental in developing the NMR Markup Language (nmrML), an open exchange format for NMR spectral data [222]. A web-based converter is available to convert the proprietary formats Agilent/Varian, Bruker and JEOL to nmrML[4]. So far MetaboLights [126] has adopted nmrML as a storage format, and work is underway for it to become an accepted format for Metabolomics Workbench [127].

The mzTab format is designed to serve as a lightweight, tab-delimited supplement to existing XML-based file formats (mzML, mzQuantML), providing a comprehensive summary of proteomics experimental results [223]. The MSI has worked with the PSI to add support for metabolomics. In the next planned release (version 2.0.0) full support for metabolomics will be provided[5].

2.6.3 MERIT

A deeper understanding of molecular events that underpin toxicity could be achieved by using metabolomics [69]; metabolomics has the potential to play a key role in improving chemical risk assessment. Harmonisation of standards and ontologies was identified as key requirements in order for metabolomics to be used for hazard assessment [63]. In light of this, the MEtabolomics standaRds Initiative in Toxicology (MERIT)[6] was recently launched. The initial aim of MERIT is to define best practice and minimal reporting standards for the application of metabolomics to regulatory toxicology. These will include best practice guidelines for the acquisition, processing and analysis of metabolomics data and minimal reporting standards including the use of appropriate quality assurance and quality control.

[4]http://nmrml.org/, accessed 12th February 2018.

[5]https://github.com/HUPO-PSI/mzTab, accessed June 25th 2018.

[6]http://www.ecetoc.org/topics/standardisation-metabolomics-assays-regulatory-toxicology/, accessed 2nd August 2017.

2.7 Suggested Improvements

The need for data sharing and reuse is now very well established [112] and has been adopted and promoted by many major funders [99–101] and learned societies. Untargeted metabolomics holds the promise of correlating complex patterns — molecular phenotypes — of concentration changes and occurrences of metabolites with aspects of the exposome of an organism. In order to fulfill this promise and to enable discovery of these patterns, we will inevitably need more metadata on the exposome than most individual researchers will initially envision when designing an experiment.

A number of initiatives to establish computational e-infrastructures for metabolomics have also recently been funded, such as PhenoMeNal[7] and MetaboFlow[8]. These computational infrastructures aim to provide well-tested and reproducible workflows for metabolomics, where data, often from public repositories, is handed from one workflow node to the next for different steps of data processing. Such computational workflows critically depend on the public availability of data, well defined and open data formats and compliance with a given set of minimum information standards.

As FAIR data sharing is not an end in itself but a means to enable better science and new scientific discoveries, our results indicate a need for a second round of MSI consultations where the existing standards are critically revisited and revised. Consideration should be given to the minimum amount of metadata required to be able to a) repeat an experiment, b) re-analyse data and c) reuse data to the fullest potential. The >1,500 publicly available metabolomics datasets available in open access repositories can be used to help guide this process.

Across the MSI guidelines single, unified definitions of "minimal" and "best practice" reporting standards should be described [196]. These definitions must be unambiguous, so that it is completely clear that minimal metadata are essential to report. It would also address the inconsistency between what constitutes a minimal reporting standard between the biological context metadata subgroups' guidelines. If, like the existing guidelines, the updated versions were to be written across a series of reports, it is essential that an overview document be produced, detailing clearly how the different standards fit together, and serving as an overall guide. Which standards should be followed when reporting metadata relating to a biological system must be apparent.

I propose that for reporting metadata relating to biological systems, a *Mandatory*, *Required* and *Optional* system be adopted. This is similar but not identical to the MUST,

[7] https://phenomenal-h2020.eu, accessed 16th February 2018.
[8] https://www.metaboflow.org/, accessed 16th February 2018.

SHOULD and MAY requirements terminology[9]. *Mandatory* metadata should be reported for all studies; *required* metadata should be reported for all studies where it has been used or is present e.g. treatment, and reporting of *optional* metadata is down to the researcher's discretion. The proposed system is presented in Appendix D and is available online at: https://doi.org/10.6084/m9.figshare.5788947.

At minimum, publicly available studies should provide sufficient metadata to enable other researchers to reproduce the results found in the original study. For clinical studies this often includes factors that are confounding, and are controlled for during statistical analysis (e.g. age, gender and BMI). It must be clear to scientists that it is necessary for them to report this information. A lack of reproducibility is hindering many areas of science [80, 81, 224, 225], and publishing open data with sufficient metadata can help to solve this.

Ontologies can help to ensure that metadata are reported in a consistent manner across studies, and are therefore comparable between studies [211]. This research found higher compliance with the MSI plant guidelines that specifically stated which ontology should be used, compared to the other reporting standards. When the MSI guidelines are updated, care must be given to ensure that as many of the standards as possible precisely state which ontology should be used for reporting them.

Whilst some use of ontologies is currently required by repositories for reporting metadata, this is at a far lower level than would be required by updated MSI guidelines, should all standards be ontologised. Ferreira *et al.* [226] applied Metadata Analyser [227], a tool for automatically measuring metadata quality based on the amount of metadata annotations that use ontology concepts, and their semantic specificity, to the entire MetaboLights dataset. They found that coverage values ranged from 0.03 to 0.46, with an average of 0.27, showing that metadata are still often reported without using appropriate ontologies.

A simpler alternative to using ontologies for reporting comparable metadata would be Medical Subject Headings (MeSH) [160]. MeSH are a manually curated, hierarchically-organized terminology for indexing and cataloging biomedical information.

The level of granularity metadata should be reported with must also be considered. There are a huge number of different cancers, encompassing many different diseases and etiologies, yet many studies will only report organ of origin e.g. breast or lung cancer, and not the specific type, e.g. invasive lobular carcinoma or small-cell lung carcinoma. It is also important that researchers are precise when reporting times, dates and ages — "date of blood draw" is a far more informative description than "date". Lack of specificity in reporting

[9]https://tools.ietf.org/html/rfc2119, accessed 16th February 2018.

metadata may hinder reuse, however researchers may be less likely to comply with guidelines requiring them to report in more detail.

Repositories have a trade-off between attracting users to submit their data and enforcing reporting standards to ensure deposited data is informative. This could partially be addressed by systematically capturing experimental metadata via LIMS systems during experiments as proposed by Rocca-Serra *et al.* [114], or by tools such as mzML2ISA [228] and biocrates2isatab[10] that can generate partially filled ISA reporting templates for MetaboLights.

It should also be noted that many of the most successful reporting standards were written using a bottom up approach, with a community identifying a need for standardisation [210]. However, imposed reporting standards usually fail [229]. When revising the MSI standards this must be considered and the community must be engaged with. In other communities, such as MIAPE within proteomics [179], continuous revisions have been adopted. This allows guidelines to be rapidly changed, and standards to adapt to new advances in technology or increased biochemical understanding. The metabolomics community should follow the Human Proteome Organization's (HUPO) lead and incorporate continuous revisions into the updated standards.

The metabolomics community should also follow in the footstep of proteomics and harmonise submission to MetabolomeXchange. Since its initial release, submission guidelines to ProteomeXchange have been updated [230]. Unlike ProteomeXchange, MetabolomeXchange does not currently use unified identifiers, nor a two tier submission system, of "complete" and "partial" submissions. Complete submissions include raw mass spectral data, experimental metadata, and processed peptide and protein identification results that can be parsed by the repository, allowing identified peptides to be directly linked to the raw mass spectra. Partial submissions still contain raw data, and experimental metadata, but results are in formats that cannot be parsed by the repository, so raw spectra cannot be directly linked to results. The complete and partial submission system should be adopted by MetabolomeXchange, and global identifiers should be provided.

2.8 Conclusion

The Metabolomics Standards Initiative (MSI) guidelines were first published in 2007. These guidelines provided reporting standards for all stages of metabolomics analysis: experimental design, biological context, chemical analysis and data processing. Since 2012, a series of public metabolomics databases and repositories, which accept the deposition of

[10]https:// github.com/ISA-tools/isa-api, accessed 17th February 2018.

metabolomic datasets, have arisen. In this study, the compliance of 483 public datasets, from five metabolomics data repositories, to the biological context MSI reporting standards were evaluated. None of the reporting standards were complied with in every publicly available study, although adherence rates varied greatly, from 0 to 97%. The plant minimum reporting standards were the most complied with and the microbial and *in vitro* were the least. These results indicate the need for reassessment and revision of the existing MSI reporting standards. The community, along with data curators, industry representatives, publishers and funders should be consulted. Consideration should be given to the minimum amount of metadata required to be able to a) repeat an experiment, b) re-analyse the data, and c) reuse data to the fullest potential. The data that are now publicly available in a number of open-access repositories around the globe will be a treasure trove for guiding the selection of minimal reporting standards.

Chapter 3

Data Sharing and Reuse in Metabolomics

As the open sharing of the results of scientific research is being increasingly encouraged or mandated by funding bodies, journals and societies, it is important to assess how effective the policies of these institutions are at boosting the amount of open data. This research aimed to explore practices of data sharing within the metabolomics community. Firstly, the data sharing policies of the journals with publications associated with the most publicly available metabolomics data were reviewed, and the journals that published the most metabolomics research were identified.

Having identified PLOS ONE as the second largest publisher of metabolomics research (and the largest by number of papers indexed in PubMed) it was surprising to find that <30 PLOS ONE publications were directly linked to open metabolomics data in repositories. Considering the stringent data sharing policy of the journal, which has been in place since March 2014 [231], potential reasons for the lack of public archiving of data in dedicated repositories were investigated further. The data availability statements and the levels of data sharing of PLOS ONE metabolomics papers were therefore manually examined and classified.

In the final portion of this chapter, the value of publicly sharing metabolomics data is reviewed, by investigating the frequency of data reuse. The chapter is concluded by reviewing challenges to open metabolomics data, and how its sharing can be improved.

Similarly to chapter 2, the methods used in this research include the manual classification of publicly available metabolomics datasets. Research in both chapters involves analysing whether studies fulfill a set of criteria — the MSI reporting standards and the data sharing categorisation of PLOS ONE articles.

3.1 Journal Data Sharing Policies in Metabolomics

As discussed previously (see chapter 1.5), funding bodies, journals and societies are increasingly encouraging or mandating data sharing: >40% of journals now encourage or require data sharing [97]. However, policies requiring data sharing must be enforced in order to be effective. In psychology it has been found that journal data sharing policies are effective at promoting data sharing [232]. Journal mandates also appear to increase data sharing in evolutionary biology [233].

To date, however, no studies have assessed the effectiveness of journal data sharing policies in metabolomics. This research aims to evaluate the effectiveness of the data sharing policies of the journals with publications associated with the most publicly available metabolomics data.

3.1.1 Methods

The initial aim of this research was to extract all publications directly linked to metabolomics datasets. Metabolomics studies with open data available via a dedicated repository (GNPS [128], MetaboLights [126], Metabolomics Workbench [127], MetaPhen [124, 125] and MeRy-B [123]) were examined for directly linked publications, as of 5th September 2017.

A full list of MetaboLights studies was downloaded from the public EBI server via File Transfer Protocol (FTP). From this publications associated with public studies and the Digital Object Identifiers (DOIs) of those associated publications were extracted. The Metabolomics Workbench REST service was used to download .mwtab files, which contain metadata about Metabolomics Workbench studies. The .mwtab files were searched for the field "Publications" and for DOIs. To identify any publications that had been missed by automatic extraction, datasets from both MetaboLights and Metabolomics Workbench were then manually reviewed. GNPS, MetaPhen and MeRy-B datasets were manually inspected in order to extract associated publications, along with their DOIs.

Following the identification of all publications directly linked to publicly available metabolomics data, the title, authors, publication year and journal were extracted for each publication. If PMID (PubMed ID) and PMCID (PubMed Central ID) were available these were also recorded. The number of papers published in each journal, and the number of papers published per year were then quantified.

Next, an attempt was made to estimate the total number of metabolomics journal articles. Initially searching PubMed using the MeSH "metabolomics" or "metabolome", was considered to obtain this estimate. MeSH are manually applied to journal articles indexed by

PubMed. However it was found that neither of the MeSH terms was ever applied to articles published by the Metabolomics journal, one of the largest publishers of metabolomics research. Searching Web of Science for “metabolomics” or “metabolome” was also considered, however at the time the research was conducted the Metabolomics journal did not appear to be indexed by Web of Science.

Instead PubMed was searched for:

“(“metabolomics”[MeSH Terms] OR “metabolomics”[All Fields]) OR (“metabolome” [MeSH Terms] OR “metabolome”[All Fields]) AND Journal Article[ptyp]”.

As well as journal articles this search also returned review articles and systematic reviews. To remove reviews and systematic reviews, PubMed was sequentially searched for:

“("metabolomics"[MeSH Terms] OR "metabolomics"[All Fields]) OR ("metabolome" [MeSH Terms] OR "metabolome" [All Fields]) AND Review[ptyp]”,

“("metabolomics"[MeSH Terms] OR "metabolomics"[All Fields]) OR ("metabolome" [MeSH Terms] OR "metabolome"[All Fields]) AND systematic[sb])”, and

“("metabolomics"[MeSH Terms] OR "metabolomics"[All Fields]) OR ("metabolome" [MeSH Terms] OR "metabolome"[All Fields]) AND (systematic[sb] OR Review[ptyp]).” All of these PubMed search histories were downloaded using FLink[1]. Publications tagged as reviews and/or systematic reviews where then removed from further analysis.

Next, the data sharing policies of the ten journals with the most publications directly linked to open metabolomics datasets, and the ten journals that publish the most metabolomics journal articles were investigated.

Data and Code Availability

The datasets generated during and/or analysed during the current study, along with the analysis code are available on the GitHub repository: https://github.com/RASpicer/Metabolomics_Data_Sharing, under the GNU General Public License v3.0. All of the analysis was performed using R version 3.3.2., along with the package ggplot2 version 2.2.1.

3.1.2 Open Metabolomics Data Linked to Publications

GNPS [128], MetaboLights [126], Metabolomics Workbench [127], MetaPhen [124, 125] and MeRy-B [123] datasets were searched for directly linked publications. MetaboLights and Metabolomics Workbench were first searched automatically, and then manually reviewed for additional associated publications that had been missed by the automatic search. GNPS,

[1] https://www.ncbi.nlm.nih.gov/Structure/flink/flink.cgi, accessed 5th September 2017

MetaPhen and MeRy-B were manually searched to identify directly linked publications. In total 368 unique journal articles are directly linked to open metabolomics data. There are 465 direct links from datasets to journal articles.

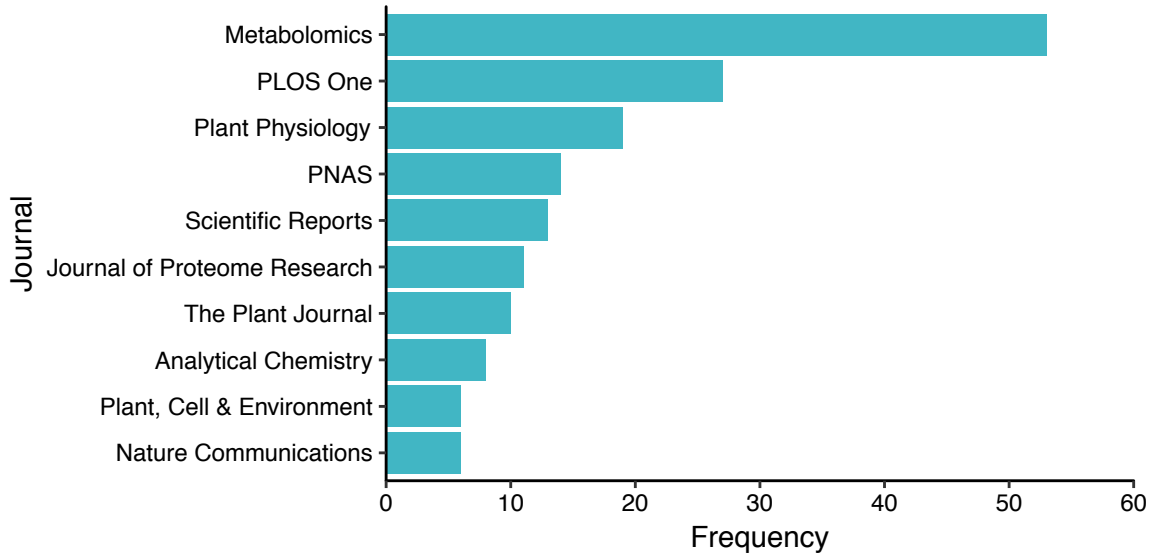


Fig. 3.1 The ten journals with the highest frequency of publications directly linked from a publicly available metabolomics study, in a dedicated repository (MetaboLights, Metabolomics Workbench, MetaPhen, MeRy-B and GNPS), as of the 5th September 2017. Figure reproduced from Spicer and Steinbeck (2018) [171].

Automatic search returned 242 MetaboLights studies that had an associated publication with a DOI, however some studies had multiple associated publications, and 282 DOIs directly linked to studies were identified, with 207 unique publications. Further manual checking increased the total studies with associated publications identified to 252. 231 unique publications were associated to MetaboLights datasets. At the time the study was conducted there were 285 publicly available studies in MetaboLights, meaning 88.42% of datasets had an associated publication.

All of the available (396) .mwtab files were downloaded from the Metabolomics Workbench REST service. 32 studies included the Publications field in their .mwtab file. An additional 9 studies had associated DOIs that were not in the Publications field; 41 studies with associated publications were identified via automatic search. However, it was later found that .mwtab files are not automatically updated if the study description is updated on the website, meaning that automatic search of the .mwtab files missed associated publications e.g. ST000004 had an associated publication that was missed.

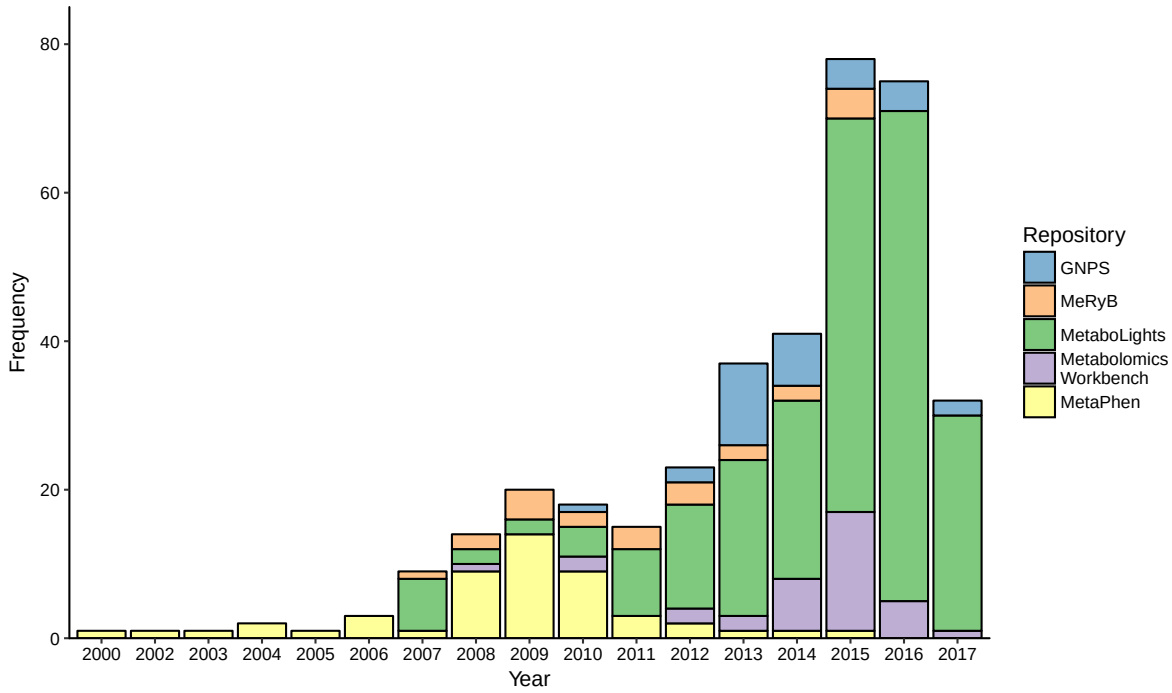


Fig. 3.2 The year of publication of journal articles linked to open metabolomics data. The frequency of journal articles published in each year is coloured by the repository the journal article is linked to, with GNPS in blue, MeRyB in orange, MetaboLights in green, Metabolomics Workbench in purple and MetaPhen in yellow. Some journal articles are linked to studies in multiple repositories so are included twice. Data were collected on the 5th September 2017.

Manual inspection of all Metabolomics Workbench studies found that there were 408 public studies in total, as some studies did not have an associated .mwtab file. Only 240 of these studies included raw data. 60 studies that included raw data had an associated publication and 15 studies without raw data did. There were 36 unique papers linked to studies with raw data. In total 18.38% of studies had an associated publication (75/408) and 25% of studies with raw data available had an associated publication (60/240).

100% of the studies (58) on MetaPhen had an associated publication. 50 unique publications were directly linked to MetaPhen datasets, with 5 publications being associated to more than one study. There is one publication that is associated to both MetaboLights datasets and a MetaPhen study. 27 studies in MeRy-B had an associated publication (84.38%). In total 23 unique publications were linked to MeRy-B datasets. One publication was linked to both a MetaboLights and a MeRy-B study. There are 31 unique publications linked on GNPS datasets. Only 7.07% of (42/594) studies had an associated paper.

In total 820 open metabolomics datasets that included raw data were not linked to any publications. 48 were studies from MetaboLights, 219 from Metabolomics Workbench, 5 from MeRy-B and 553 from GNPS. 37 of the studies in GNPS are re-uploads of MetaboLights studies. Despite 94.6% (35) of these studies having at least one association on MetaboLights, only one study (2.7%) had a linked publication on GNPS.

The 368 publications directly associated with open metabolomics datasets are published by 146 journals. Ten journals published 45.4% of these journal articles (Figure 3.1), with 58 (14.4%) being published in the Metabolomics journal.

The number of journal articles published that have open metabolomics data appears to be broadly increasing (Figure 3.2). The year with the highest frequency of journal articles published that were linked to publicly available metabolomics was 2015, followed closely by 2016. MetaboLights has the highest frequency of publications linked to studies, however all MetaPhen studies are linked to at least one publication.

3.1.3 An Estimate of the Number of Metabolomics Publications

In total 21,230 publications were returned when searching PubMed for “("metabolomics" [MeSH Terms] OR "metabolomics"[All Fields]) OR ("metabolome"[MeSH Terms] OR "metabolome"[All Fields]) AND Journal Article[ptyp]”. FLink is only able to download 10,000 PubMed search results at a time, so the 21,230 article details were downloaded as three .csv files and then recombined. 3,498 of these publications were review articles, 228 were systematic reviews and 3,608 were tagged as both reviews and systematic reviews. There were 118 publications tagged only as systematic reviews and not as reviews. After subtracting reviews and systematic reviews 17,614 primary research articles remained.

Only a very rough estimate of the total number of metabolomics journal articles can be obtained by searching PubMed like this. The MEDLINE (Medical Literature Analysis and Retrieval System Online) database that is searched using PubMed contains >28,000,000 journal article references, however there are only ~4,700,000 full text manuscripts archived on PubMed Central (PMC). Therefore, despite including metabolomics research, articles that do not contain the words “metabolome” or “metabolomics” in their title or abstract may not be returned by the PubMed search used. Also non-metabolomics papers may be returned or metabolomics papers may not be indexed in MEDLINE. Only 25.8% (304/1178) of articles published in Metabolomics are indexed in PubMed.

It must additionally be noted that PubMed was not searched for the term "metabonomics" during this research. Metabonomics is considered to be a subset of metabolomics [234]

and is defined as “the study of the multiparametric metabolic responses of a particular living organism to pathogenic stimuli or genetic modification”. Metabonomics was initially excluded from this research as there is no metabonomics MeSH Term. Following the change of experimental design to search [All Fields], as well as [MeSH Terms], metabonomics could have been included, however, it was unintentionally not reincluded. There are 1480 articles returned when searching PubMed for "metabonomics" OR "metabonome", that were published before 5th September 2017, the time period used in the initial research. These articles may have also been missed by the PubMed search used in this study.

Of the publications directly associated with open metabolomics data, 44% were not returned when searching PubMed for “("metabolomics"[MeSH Terms] OR "metabolomics"[All Fields]) OR ("metabolome"[MeSH Terms] OR "metabolome"[All Fields]) AND Journal Article[ptyp]”. 31.8% of these publications were indexed on MEDLINE but not returned and 12.6% were not indexed. Despite these limitations, this method of identification of metabolomics journal articles is sufficient for this analysis, which aims only to identify the journals that publish the most metabolomics research. It can be assumed metabolomics research is included in the majority of articles returned when searching PubMed for “metabolome” OR “metabolomics”.

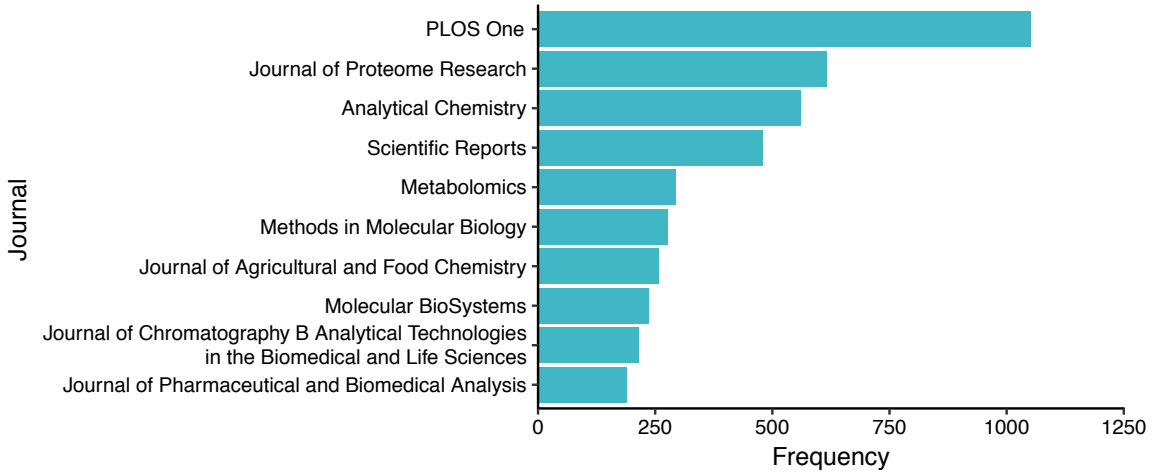


Fig. 3.3 The ten journals with the highest frequency of primary journal articles when searching PubMed for ("metabolomics"[MeSH Terms] OR "metabolomics"[All Fields]) OR ("metabolome"[MeSH Terms] OR "metabolome"[All Fields]) AND Journal Article[ptyp], after reviews and systematic reviews are removed, as of the 5th September 2017. Figure reproduced from Spicer and Steinbeck (2018) [171].

Publications returned when searching for the aforementioned criteria in PubMed were published in 1,995 journals. However, 780 journals had published only a single piece of

metabolomics research. The ten journals that published the highest number of metabolomics articles returned by PubMed are shown in Figure 3.3. These journals published 23.65% of metabolomics research (4,170 articles); 56.45% of metabolomics articles (9,954) were published in 100 journals.

3.1.4 Journal Data Sharing Policies

None of the journals that publish the most metabolomics research have very high data sharing rates via metabolomics repositories (Table 3.1). Methods in Molecular Biology has no publications linked to publicly available metabolomics data.

Five of the journals that publish the most metabolomics research also have the highest number of journal articles linked to open metabolomics data: PLOS ONE, Journal of Proteome Research, Analytical Chemistry, Metabolomics and Scientific Reports (Table 3.1). Other journals that publish the most articles linked to open metabolomics data have all published at least 50 journal articles including metabolomics research (Table 3.1).

PLOS ONE's data sharing statement specifically states that "All data and related metadata underlying the findings reported in a submitted manuscript should be deposited in an appropriate public repository, unless already provided as part of the submitted article" and recommends MetaboLights [126] and Metabolomics Workbench [127] as metabolomics repositories. It also states that data sharing is a requirement for publication. However, publications with human data are subject to different data sharing guidelines that emphasize the need to protect participants privacy, and do not require sharing via a repository. Also it is stated that if the standard within a field is to share processed data, researchers do not need to share raw data.

Metabolomics and Plant Physiology require that authors make materials available to investigators for non-commercial research purposes (Table 3.1). Metabolomics specifically requires raw data sharing, and suggests that users deposit their data in a repository (data sharing on demand complies with the journal's policy). It encourages the use of data availability statements.

The Plant Journal has published guidelines for the reporting of metabolomics data [235]. However these guidelines detail how metabolomics experiments should be reported within a manuscript, and there are no requirements for sharing raw data. The Journal of Agricultural and Food Chemistry also has specific requirements for reporting spectroscopic data within the text of a manuscript, but no raw data sharing requirement.

Two of the journals with the highest number of metabolomics papers on PubMed have no data sharing or data availability statement policies: Analytical Chemistry and Methods in Molecular Biology (Table 3.1). Plant, Cell & Environment also has no statement on data sharing. The Journal of Proteome Research encourages users to deposit proteomics data in ProteomeXchange [120], however, unsurprisingly has no specific policy on metabolomics data.

Where possible, Molecular BioSystems and the Journal of Chromatography B, require data sharing. The Journal of Pharmaceutical and Biomedical Analysis encourages data sharing. All three journals suggest sharing by a repository.

Table 3.1 The journals that publish the most metabolomics research and the journals that publish the most articles directly associated with open metabolomics data, and the data sharing policies of the journals. Equals signs (=) are used to represent journals that are equally ranked because they have the same number of articles linked to open data.

Journal	PubMed Ranking	No. PubMed Publications	Open Data Ranking	No. Articles Linked to Open Data	Data Availability Statement	Data Sharing	Data Sharing via Repository
Journals with the most metabolomics publications and the most publications linked to open metabolomics data							
Analytical Chemistry	3	561	8	8	None	None	None
Journal of Proteome Research	2	615	6	11	None	Encouraged	Encouraged
Metabolomics	5	293*	1	53	Encouraged	Required	Encouraged
PLOS ONE	1	1,050	2	27	Required	Required	Required
Scientific Reports	4	480	5	13	Required	Required	Encouraged
Journals with the most publications linked to open data							
Nature Communications	53	54	=9	6	Required	Required	Encouraged
Plant, Cell & Environment	36	70	=9	6	None	None	None
Plant Physiology	13	176	3	19	None	Required	Required if available
PNAS	14	165	4	14	Required	Required	Encouraged
The Plant Journal	30	84	7	10	None	Within Article Text	None
Journals with the most metabolomics publications							
Journal of Agricultural and Food Chemistry	7	257	=28	2	None	Within Article Text	None
Journal of Chromatography B	9	215	=48	1	Encouraged	Required	Encouraged
Journal of Pharmaceutical and Biomedical Analysis	10	188	=28	2	Encouraged	Encouraged	Encouraged
Methods in Molecular Biology	6	276	-	0	None	None	None
Molecular BioSystems	8	235	=21	3	None	Required	Encouraged

* Whilst Metabolomics has only 293 journal articles indexed on PubMed, it has published >1,000 articles.

3.1.5 Potential Reasons for a Lack of Data Sharing in Metabolomics

Just how little open metabolomics data is available via dedicated repositories is surprising, given the number of journal articles published in the field. It is especially surprising given some of the long withstanding strict data sharing policies of some of the journals that publish the most metabolomics research. PLOS ONE's current data sharing policy has been in place since March 2014 [231] and Springer Nature have had their policy since September 2016 [236]. Since 2015 PLOS ONE has published ~400 metabolomics papers and since 2017 Scientific Reports has published >140. Despite not requiring data sharing via a dedicated repository, articles published by the Metabolomics journal share data in dedicated repositories at a higher rate than those in PLOS ONE.

Although PLOS ONE recommends using a dedicated, field-specific repository, users may remain unaware of dedicated metabolomics repositories and instead publish their data in general repositories such as Dryad Digital Repository[2], figshare[3] or Zenodo[4]. Alternatively, as PLOS ONE's data sharing policy specifically states "authors do not need to submit the raw data collected during an investigation if the standard in the field is to share data that have been processed", researchers may feel that metabolomics is one such field where sharing only preprocessed data or an annotated list of identified metabolites is sufficient, rather than raw spectral data.

Another possibility is that journals such as PLOS ONE and Scientific Reports publish a higher percentage of clinical research or other studies with human participants. Due to concerns of patient privacy and consent, both journals have different data sharing requirements for clinical studies compared to those for studies including nonhuman subjects. Only summary rather than raw data must be reported for clinical studies.

An additional concern, that may also be the cause of the perceived lack of data sharing in PLOS ONE and Scientific Reports, is the number of publicly available metabolomics studies with raw data that have no associated publication: >800. Whilst some of this data will have no associated publication, e.g. because the researcher was unable to get the study published, or if the dataset is purely indented for training purposes, associated journal articles probably exist for much of this open data, however there is no direct link between the data and the literature. This hinders the reuse of data, as papers are likely to contain more detailed experimental design descriptions and additional metadata, and data alone are insufficient for reanalysis [211].

[2]https://datadryad.org/, accessed 11th October 2017.
[3]https://figshare.com/, accessed 11th October 2017.
[4]https://zenodo.org/, accessed 11th October 2017.

3.2 Metabolomics Data Sharing in PLOS ONE

PLOS ONE is a mega journal that publishes research across all scientific disciplines, focusing on the quality of research rather than the novelty of results. Since 2015 PLOS ONE has published >75,000 journal articles and has published the highest number of metabolomics papers indexed on PubMed (see section 3.1.3). It is a pioneer of mandating open data, having required public data sharing since March 2014 [231]. Previous research found that this policy change was successful with the data sharing rate of studies published in PLOS ONE that used the population genetic analysis tool, STRUCTURE[5], increasing from 11.7% in 2012–2013 [237] to 40% in 2014 [238]. A recent study [239] found that overall 15% of PLOS ONE studies, published since the introduction of its data sharing policy, shared data via a repository. All of the components mentioned above make PLOS ONE an ideal candidate to investigate for this study, to see if its unique combination of popularity and policy towards data sharing have a measurable effect on public data in metabolomics.

Unfortunately, whilst PLOS ONE has published ~400 metabolomics papers since 2015, only 12 PLOS ONE journal articles published during the same time period that were directly linked to raw metabolomics data hosted in dedicated repositories, were identified. This research aims to better understand the reasons for the lack of linkage between metabolomics journal articles published in PLOS ONE and open data on metabolomics specific repositories. Whether researchers share raw data by means other than metabolomics specific repositories, or if PLOS ONE publishes a high percentage of clinical studies, where raw data cannot be made publicly available, is investigated. If there is evidence that rather than openly sharing raw data, researchers instead share processed data, and consider this to be sufficient, is also examined.

In the previous research in this chapter (section 3.1), a PubMed search was used to identify metabolomics research. However, this search may also have returned publications that did not include metabolomics. In this study, the previously used method is improved upon by manual assessment of whether publications contain metabolomics research or not.

3.2.1 Raw Metabolomics Data

Raw data is most commonly used to refer to unprocessed data [240]. In metabolomics this refers to raw spectral data. Whilst the "rawest" types of metabolomics data are data in vendor specific proprietary formats, data in these formats are at risk of becoming obsolete and of data rot, as support for older formats and software is discontinued [241]. Therefore, metabolomics

[5]https://web.stanford.edu/group/pritchardlab/structure.html, accessed 5th April 2018.

data that has been minimally processed, in order to convert it from a commercial to an open format, is also usually considered to be raw data. During this process raw mass spectrometry data may additionally be converted from profile mode to centroid (depending on the type of mass spectrometer used, data may initially be collected as profile or centroid).

The sharing of raw data is important as multiple studies have found that using different software to preprocess data results in different features being detected [242–244]. Peak lists have at minimum undergone preprocessing, and may have also been post-processed prior to statistical analysis. Whilst there is such potential to gain additional information from raw metabolomics data, it is crucial that it is publicly shared.

In this study raw metabolomics data is defined as raw spectral data in either proprietary or open formats. Table 3.2 lists examples of file formats used to store raw metabolomics spectra.

Table 3.2 A non-exhaustive list of raw data formats used in metabolomics.

File Format	Instrument	Type	Vendor	Instrument Software
.imzML	Imaging MS	Open	-	-
.mzML	MS	Open	-	-
.mzXML	MS	Open	-	-
.mzData	MS	Open	-	-
.netCDF	MS	Open	-	-
.d	MS	Proprietary	Agilent	MassHunter
.D	MS	Proprietary	Agilent	ChemStation
.d	MS	Proprietary	Bruker	Compass
.BAF	MS	Proprietary	Bruker	Compass
.FID	MS	Proprietary	Bruker	Compass
.YEP	MS	Proprietary	Bruker	Compass
.jpf	MS	Proprietary	JEOL	MassCenter
.wiff	MS	Proprietary	SCIEX	Analyst
.lcd	MS	Proprietary	SHIMADZU	LCMSsolution
.RAW	MS	Proprietary	Thermo Fisher	Xcalibur
.raw	MS	Proprietary	Waters	MassLynx
.nmrML	NMR	Open	-	-
.dx	NMR	Proprietary	Bruker	TopSpin
.jdx	NMR	Proprietary	Bruker	TopSpin
.fid	NMR	Proprietary	Varian	VNMR

3.2.2 Methods

Changes in publication policies are not usually smooth, and it usually takes time for them to be more broadly complied with [238]. As PLOS ONE's data sharing policy was updated in March 2014 [231], it was felt that by the beginning of 2015 most of the initial teething problems would have been fixed, and the vast majority of publications would include a data sharing statement. In this research a subset of the 1,050 PLOS ONE papers that were identified in the previous research were examined (see section 3.1.1 for full methods by which papers were identified). Of these journal articles, 426 had been published since the beginning of 2015, and the data sharing statements and level of data sharing of these studies were investigated.

Firstly, all selected PLOS publications were manually reviewed in order to assess whether they included a) metabolomics research and b) "primary" metabolomics research. In this context "primary" research refers to a study that generates new raw data, rather than reusing data, using simulated data or performing meta-analysis ("secondary" analysis). This is due to differences in the types of data produced and data sharing guidelines between these study types. For the purpose of this research, computational studies that either generated simulated data or reuse published data, were not considered to be "primary" metabolomics research, and were excluded from subsequent stages of analysis. Only primary metabolomics publications were included in subsequent stages of analysis, including computational studies where new raw data was generated.

Publications were next assessed to ascertain whether or not they included clinical human research. Following this, the data sharing statements of the journal articles were classified into 7 levels (Table 3.3), including six classifications of data availability statement ("A–F") and one level ("G") indicating no data availability statement was present. Data availability statements were categorized by where they stated data was available: "A" in the paper and its supporting information, "B" in the paper (when an article had no supplementary material), "C" in a repository, "D" in the paper and its supporting information, but another type of omics data was shared via a repository and "E" on request. Statements classified as "F" pronounced that data could not be made publicly available. If the data availability statement mentioned a repository this was recorded. Next whether the raw metabolomics data associated with the study was publicly available to download, or not, was documented.

The level of data sharing of the study was then recorded (Table 3.4). If raw data was available, the study was classified as level 1, with the subcategories a, b and c denoting whether the data was available in a metabolomics specific repository e.g. MetaboLights, a general repository e.g. Figshare, or in another location, such as on an institutional website.

If a study provided a peak list or signal intensity matrix of the relative quantifications or concentrations on a per sample basis, it was categorised as level 2 (Figure 3.4). Level 3 indicates that data were presented as a table of either all identified or only statistically significantly changed metabolites, level 4 is a figure of the raw spectra and level 5 is a figure showing statistically significantly changed metabolites in e.g. a heat map or a bar chart.

Studies that do not include publicly available raw data can have multiple classifications of both level 4 and either level 2, 3 or 5 (Figure 3.5). Raw data (Level 1) can be processed and used to generate any of the subsequent types of data: peak lists or signal intensity matrices (Level 2), tables (Level 3), figures of spectra (Level 4) and figures of metabolites (Level 5), therefore studies of level 1 do not need multiple classifications. Studies cannot be classified as levels 2 and 3, 2 and 5 or 3 and 5, because level 2 is inherently more raw than levels 3 and 5, and level 3 is more raw than level 5. Tables and figures of metabolites can be produced from a peak list or signal intensity matrix, but cannot be used to generate a figure of the raw spectra. From some tables (level 3), it is possible to generate some figures (level 5), but not all contain sufficient information.

The accession numbers of studies with data available on dedicated metabolomics repositories were compared to those identified in the initial research (section 3.1.2), to identify any studies where data was publicly shared but not linked to the associated publication. As links to publications may have been added to studies in repositories since the initial research was conducted, all identified studies were then manually reviewed to see if the studies had been updated with links to publications. Studies were also checked for missing links from the publication to the open data on a dedicated repository. Full instructions as to how publications were classified can be found in Appendix E.

Statistical Analysis

Following classification, the percentage of studies with each type of data sharing statement and their levels of data sharing were then calculated. A Pearson's chi-squared (χ^2) test [245] was used to evaluate the likelihood that the observed differences between the relationships of data availability statements and levels of data sharing arose due to chance. The Pearson's χ^2 test statistic is as follows:

$$\chi^2 = \sum^{n} \frac{(O_i - E_{ij})^2}{E_{ij}} \tag{3.1}$$

Where n is the number of cells in the table, i is the row index, j is the columns index, O_{ij} is the actual number of observations of ij and E_{ij} is the expected (theoretical) frequency of ij. The larger the χ^2 test statistic, the stronger the evidence against the null hypothesis [246].

To find which cells' contributed the most to the χ^2 test statistic and p-value, standardized residuals were calculated. These shows which cells' observed values deviate the most from the expected values, under the assumed model. Standardized adjusted residuals are adjusted for row and column totals. The standardized adjusted Pearson residual for a two way table is as follows:

$$\frac{O_{ij} - \hat{E}_{ij}}{\sqrt{\hat{E}_{ij}(1 - p_{i+})(1 - p_{+j})}} \tag{3.2}$$

Where $\hat{E}_{ij}$ are estimated expected frequencies and p is the population proportion.

Table 3.3 Data availability statement classification levels.

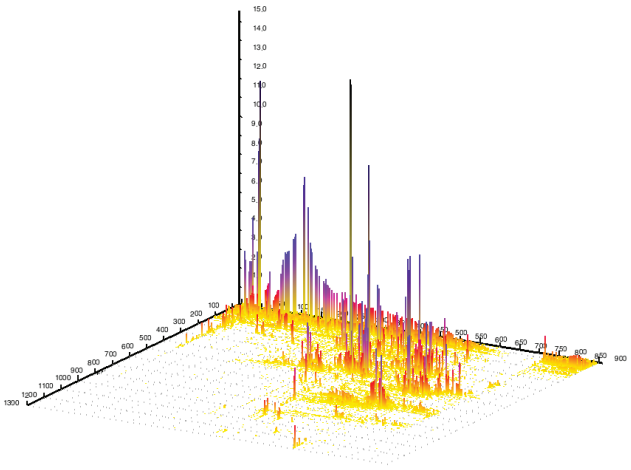
Level	Description
A	All relevant data are within the paper and its Supporting Information files. This classification includes typos and variations such as "All relevant data are within the paper and Supporting Information files." or "All relevant data are within the paper and supplement."
B	All relevant data are within the paper. This type of statement is used when a paper has no supplementary material.
C	Metabolomics data available in a repository.
D	All relevant data are within the paper and its Supporting Information files and data for another type of omics is available in a repository or just a repository specific for a different type of omics data is mentioned.
E	Data available on request.
F	Data cannot be made publicly available.
G	No data availability statement.

Table 3.4 Data sharing classification levels. If a study does not have raw data publicly available, it can be classified as both level 4 and either level 2, 3 or 5. It cannot be classified as 2 and 3, 2 and 5 or 3 and 5.

Level	Description
1	Raw data available (either commercial format e.g. Thermo .raw, Agilent .d or as an open format e.g. .mzML, .mzXML).
1a	Raw data available in a specific metabolomics mepository (e.g. MetaboLights, Metabolomics Workbench).
1b	Raw data available in a general repository (e.g. Figshare, Dryad, Zenodo).
1c	Raw data available by other means e.g. in supplementary material or on an institutional website.
2	Peak list or signal intensity matrix containing relative quantifications, concentrations, etc., on a per sample level (usually in .csv or .xlsx format).
3	Table of metabolites, including all identified metabolites or only statistically significantly changed metabolites.
4	Figure of spectra.
5	Figure showing statistically significantly changed metabolites (such as a scatter plot, bar chart, heat map, etc.)

1 - Raw Spectra

.mzML, .mzXML, .nrmML, .RAW, .FID, .d, .WIFF, etc.



2 - Peak List or Signal Intensity Matrix

.csv, .xlsx

	Sample 1	Sample 2	...	Sample *n*
Metabolite 1	93651	98565	...	94003
Metabolite 2	2461	2662	...	6536
...	...	...	...	...
Metabolite *n*	60489	53020	...	79580

3 - Table of Metabolites

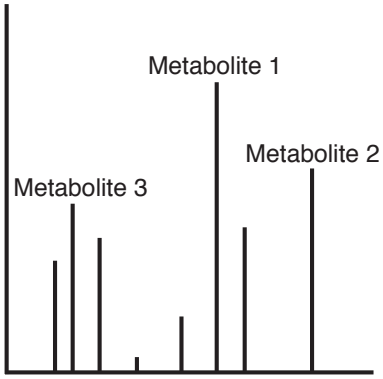
.pdf, .png

	Condition 1	Condition 2
Metabolite 1	145258	197862
Metabolite 2	23364	20612
...	...	...
Metabolite *n*	38094	25787

Fig. 3.4 The different formats of data that are required for each data sharing level.

4 - Figure of Spectra

.pdf, .png



5 - Figure of Statistically Significantly Changed Metabolites

.pdf, .png

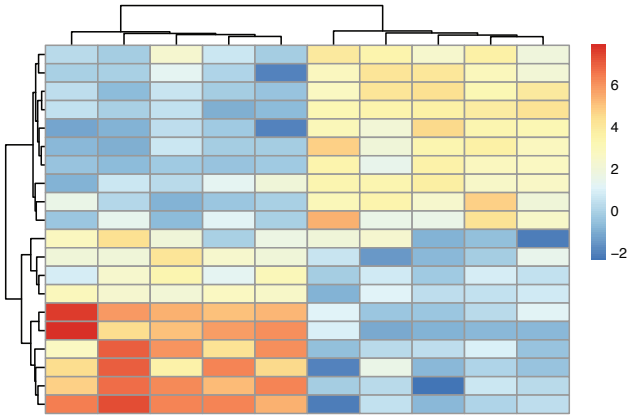


Fig. 3.4 The different formats of data that are required for each data sharing level.

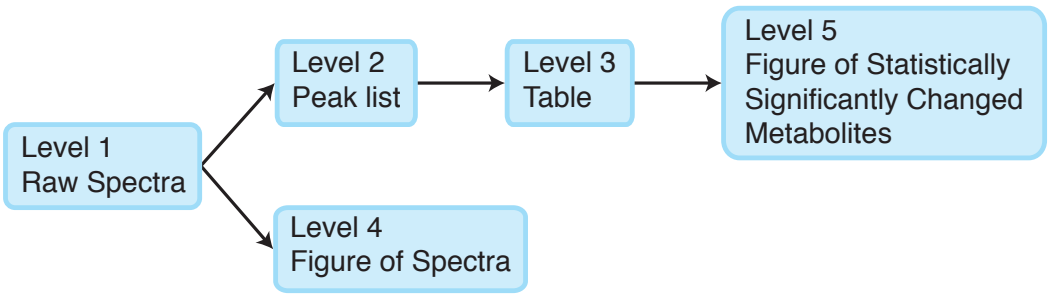


Fig. 3.5 The hierarchy of rawness of metabolomics data sharing levels.

3.2.3 Results

Of the 426 PLOS ONE studies identified, 372 (87.3%) were classified as including metabolomics analysis when manually reviewed. 359 of these were primary metabolomics studies and 13 were secondary metabolomics studies. 35.4% (127) of primary metabolomics studies included clinical human research.

The most common form of data availability statement used was type A, "all relevant data are within the paper and its supporting information files", with 65.2% of studies including this statement type (Figure 3.6A). The second most frequent data availability statement type was C, which states data are in a repository (13.4%), followed by type E, data are available on request (8.1%), type B, "all relevant data are within the paper" (7.0%) and type D (5.0%), where a repository is mentioned for a different type of omics data, along with "all relevant data are within the paper and its supporting information files". Four studies (1.1%) were missing data availability statements and a single study stated that data could not be made publicly available due to ethical concerns (type F). All but one study (96.6%) with type E data availability statements were clinical human studies. The type F study was also a clinical human study.

Only 8.08% (29) of the examined metabolomics PLOS ONE studies openly shared raw data by any means (Figure 3.6B). The most popular way of sharing raw data was in a dedicated metabolomics repository (1a — 69.0%); the second most frequent means was in general repositories (1b — 20.7%). Three studies shared raw data in the supplementary material of the journal article (1c — 10.4%). It must be noted that one study [247] shared raw data in the Dryad repository as well as GNPS. As Dryad was mentioned in the data availability statement but GNPS was not, it was classed as 1b.

Raw data was most commonly shared in the MetaboLights repository (Figure 3.8), with raw data also being shared in the Metabolomics Workbench, Dryad, Figshare and Zenodo repositories. Whilst 14 studies stated that raw data were available in the MetaboLights repository, the data corresponding to two journal articles [248, 249] were not publicly available. Three studies [250–252] that shared data in Metabolomics Workbench also had missing or incorrect links to their raw data, however the correct accession numbers were able to be manually identified.

Two journal articles made raw data publicly available on a dedicated repository but were missing links to it, and in nine instances open data was missing links to the associated publication (Table 3.5). A journal article [253] that had raw data publicly available on a dedicated repository, but was not returned by the search in order to identify metabolomics studies (section 3.1.1), was also identified.

The relationship between data statement classification and data sharing level was found to be significant (Pearson's $\chi^2 = 237.77$, $df = 36$, $p = 1.45 \times 10^{-31}$). Noticeably, the majority of studies that shared raw data had a data availability statement classified as type C, although there were also studies with type A statements that shared raw data. Studies of type C are highly positively correlated with statements of class 1a and 1b and negatively correlated with type 3 (Figure 3.7). Studies that have type A statement ("all relevant data are within the paper and its supporting information files") were negatively correlated with sharing raw data in a repository (class 1a and 1b) and studies with the statement type B ("all relevant data are within the paper") are positively correlated with sharing figures of statistically significantly changed metabolites, class 5.

However, only 54.2% of studies that stated metabolomics data were shared in a repository shared raw data. Most other studies (29.2%) shared a peak list (level 2), with a lower number of studies providing only tables or figures. Of the repositories where non-raw metabolomics data were shared, Figshare was most frequently used, followed by Metabolomics Workbench (Figure 3.8). Non-raw data was also shared on other repositories including the Open Science Framework and institutional repositories.

Overall, metabolomics data was most frequently shared as a table of metabolites, level 3 (Figure 3.6B). Sharing data as a peak list, level 2 was the second most common. A total of 53 studies shared data only in figures (at level 5 or at level 4 and 5).

Data and Code Availability

The datasets generated and analysed during the current study, along with the analysis code are available on the Open Science Framework: https://doi.org/10.17605/osf.io/hn5xg, under the MIT License. The majority of the analysis was performed using R version 3.3.2., along with the packages ggplot2 2.2.1., tidyr 0.7.2, dplyr 0.7.4, corrplot 0.84 and knitr 1.17.

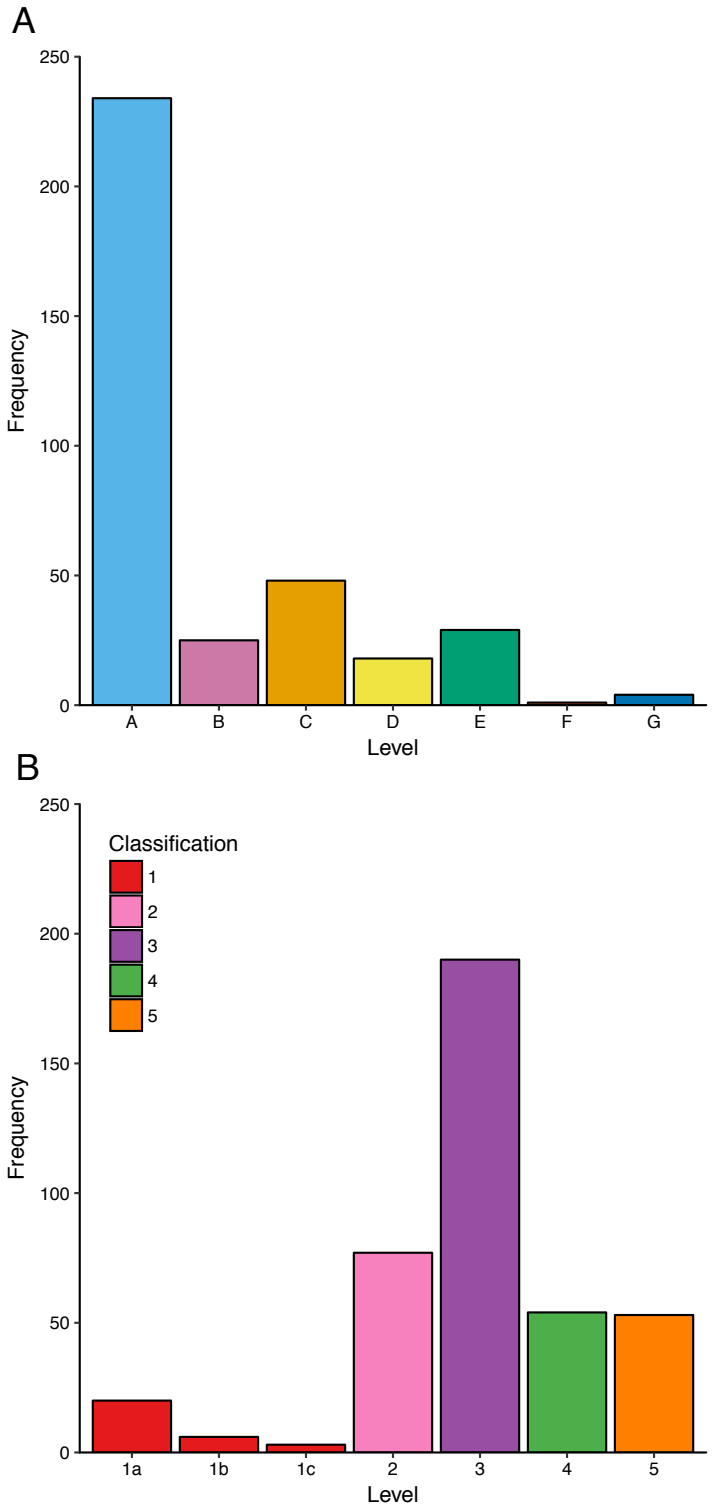


Fig. 3.6 Bar charts showing the frequency of A) data availability statement levels and B) data sharing levels in primary metabolomics studies, as described in Tables 3.3 and 3.4, respectively. Studies can receive multiple data sharing levels if they are not classified as level 1, but can have only a single data availability statement level.

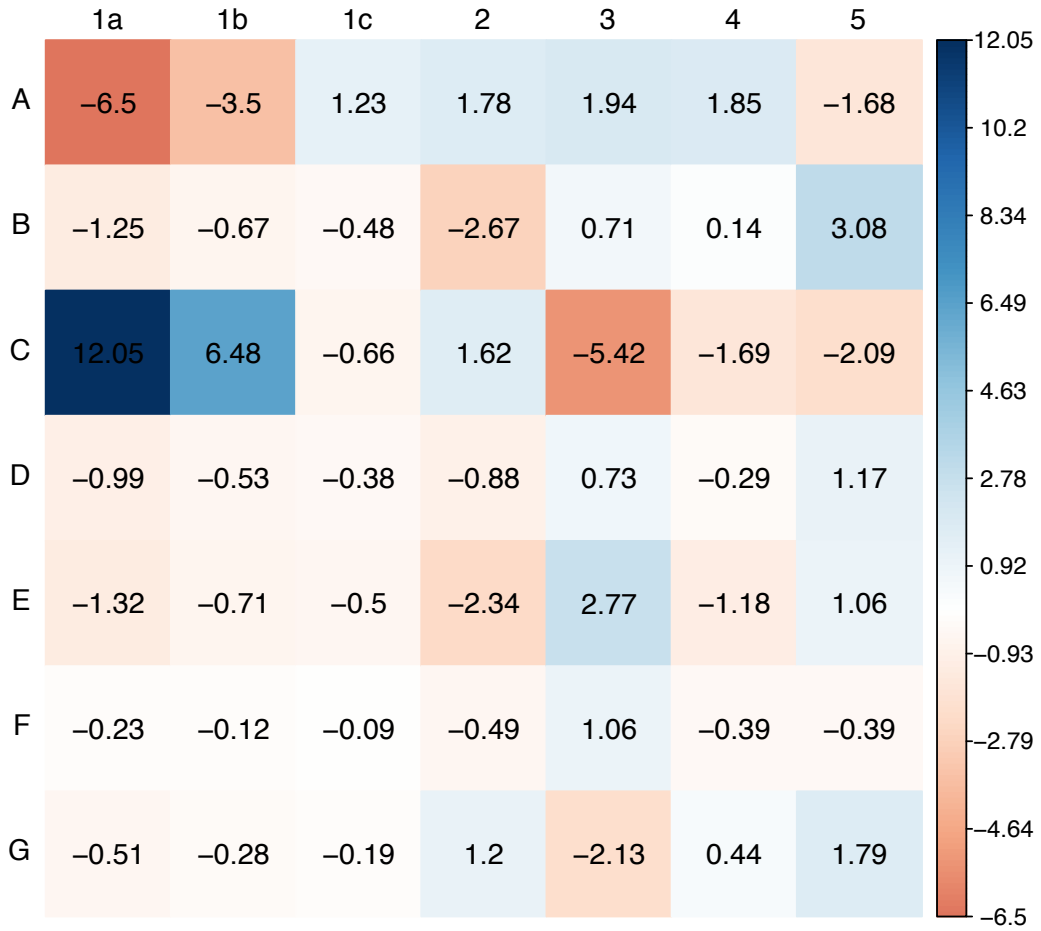


Fig. 3.7 A heat map representation of the χ^2 residuals correlation matrix for data statement classification and data sharing level. Statement classification levels are described in Table 3.3 and data sharing levels are in Table 3.4. Maximum positive correlation and negative correlation are respectively indicated in blue and red.

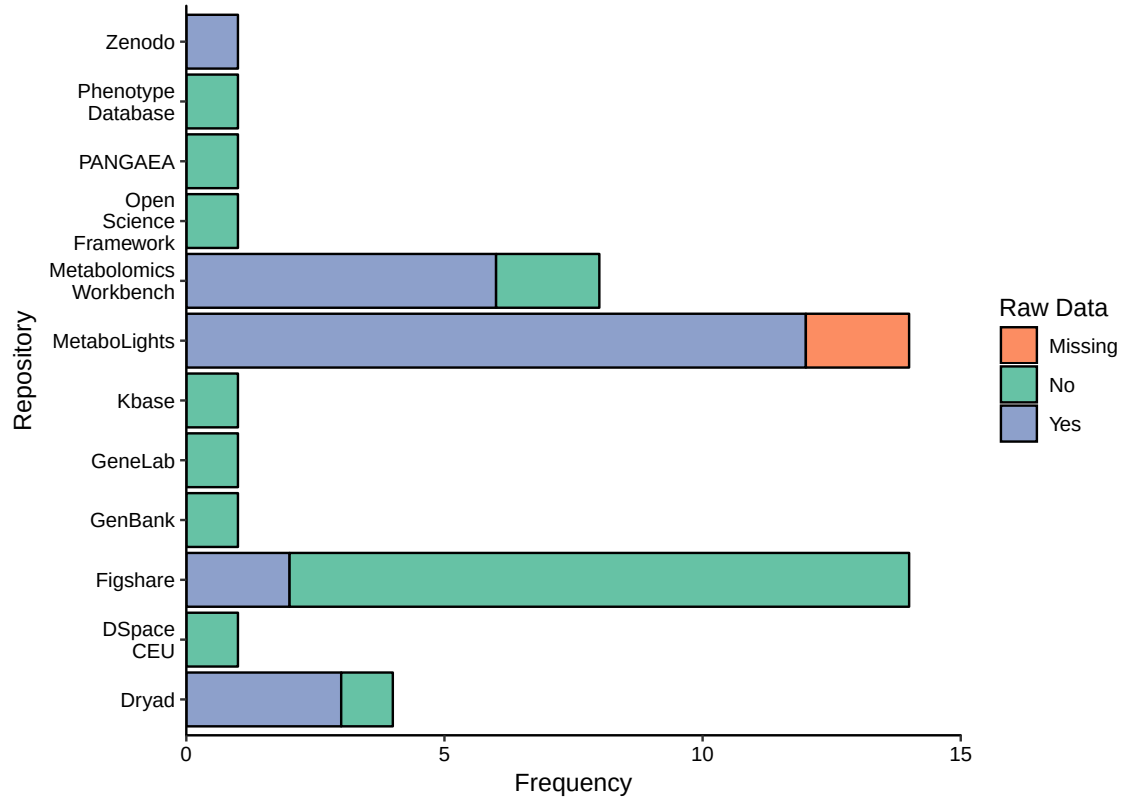


Fig. 3.8 The frequency of PLOS ONE studies linked to public data on each repository. Studies are coloured in blue if raw data is available, green if it is not, and orange if the publication states that data is available at a given accession number, but no data is publicly available.

Table 3.5 Studies with missing links from the published journal article to associated open data or from open data to journal article. Publications that link only to repositories and not specific accession numbers, and linking incorrect accession numbers are also highlighted.

Repository	Accession Number	Publication	Missing Link(s)
MetaboLights	MTBLS165	[254]	Publication to Data
MetaboLights	MTBLS309	[255]	Publication to Data
MetaboLights	MTBLS123	[256]	Data to Publication
Metabolomics Workbench	ST000121	[257]	Data to Publication
Metabolomics Workbench	ST000194	[258]	Data to Publication
Metabolomics Workbench	ST000223	[250]	Data to Publication, Publication only links to repository not specific Accession No.
Metabolomics Workbench	ST000316	[251]	Data to Publication, Publication only links to repository not specific Accession No.
Metabolomics Workbench	ST000405	[259]	Data to Publication
Metabolomics Workbench	ST000465	[252]	Data to Publication, Incorrect Accession No. linked in Publication
Metabolomics Workbench	ST000578	[260]	Data to Publication
Metabolomics Workbench	ST000590	[261]	Data to Publication

3.2.4 Exploring Reasons for the Lack of Data Sharing in PLOS ONE

The results of this research show that <10% of PLOS ONE metabolomics articles publicly share raw data, despite the journal's pioneering policy on data sharing. This lack of open data is concerning as it indicates that either the data sharing policy is not being enforced, there is a general lack of understanding as to what constitutes "raw data", or that processed data alone is considered sufficient to underlie findings. As only a minority of the included studies did not have a data availability statement, this indicates that it is reasonable to consider PLOS ONE's data availability requirements as established.

To explore whether ethical considerations prevented open data sharing in the included studies, data availability statements were examined, and studies that included clinical human research were identified. A relatively large number of the studied articles did include clinical human research, however only a minority of papers stated that data could not be made publicly available due to ethical or legal restrictions, and in all but one case data were available on request. The majority of clinical studies had data availability statements of class A, "All relevant data are within the paper and its Supporting Information files". This suggests that ethical considerations are not the primary reason for a lack of open data associated with metabolomics PLOS ONE papers.

An alternative hypothesis for the lack of open data in dedicated repositories linked to PLOS ONE was that raw data may be instead shared by other means, such as in general repositories, in supplementary material or on other websites. Many scientists will use metabolomics within their research, but it will not be their primary research focus, and they therefore may not be aware of the existence of metabolomics specific repositories such as MetaboLights [126] or Metabolomics Workbench [127]. However, it was found that only a minority of studies openly shared raw data by means other than in dedicated metabolomics repositories. This indicates that sharing raw data by other methods is also unlikely to be the reason for not sharing data in a repository.

Instead, there is evidence that the current standard in metabolomics is to share processed rather than raw data. PLOS ONE has the second highest number of journal articles directly linked to open metabolomics data on a dedicated repository (section 3.1.2) and a strong policy in support of data sharing, yet raw data is publicly shared by <10% of articles. The most common formats data are publicly shared as are tables of metabolites and peak lists. As it appears that the community standard is to share processed data, authors are not required to make raw data public in order to comply with PLOS ONE's data availability policy.

It must be noted that a limitation of this research is that articles were categorised by only a single rater. This means there is the risk of pervasive reviewer biases in classification.

A superior experimental design would be to have every article reviewed by multiple raters with a high inter-rater reliability. Unfortunately for this research it was not possible to have multiple raters. An additional disadvantage of the experimental design is that only articles published during a limited, specific time period (January 2015 - September 2017) were reviewed. The findings of this limited snapshot of time are not generalisable to other time points.

The findings of this research are akin to those in chapter 2. If reporting standards or data sharing policies exist but are not enforced or are poorly worded they do little to aid standardisation or facilitate a cultural shift to open data sharing. For metabolomics to fulfill its potential it is essential that measurements are reliable and accurate across studies. In this regard, the acceptance of sharing only processed data is concerning due to the lack of consistency between different data analysis methods for untargeted metabolomics. It has been found that not all analysis methods result in the same metabolites being detected in a study. Despite Gürdeniz *et al.* [262] being able to extract the same biological information using different analysis software, the majority of features detected in the study were unique to the software used. Myers *et al.* [243] found many false positive peaks when peaks were identified by MZmine 2 or XCMS only, and not by both. Baran [244] reanalysed data from 5 publicly available studies and identified at least 50 relevant compounds that were not reported in the original results for each study.

Whilst such discrepancy exists between the metabolites detected using different preprocessing software, sharing raw data is of paramount importance. A peak list may appear to be raw data, as it contains the relative abundances for every metabolite in every sample, however raw data refers to data that has not been processed. At minimum data in a peak list has undergone peak detection, but may also have been "cleaned" (normalised, filtered, etc.) and it is impossible to revert a peak list to the original raw spectra.

Considine *et al.* [208] recently found that all data analysis steps in metabolomics are insufficiently reported. Although the reporting of data analysis has not been examined in this research, it is likely that many of the included studies will not have adequately reported these crucial stages of analysis in order to allow for their replication.

However, there is evidence that researchers are attempting to share data openly, but are unaware of the importance of sharing raw data. In nearly half of the studies that shared data via a repository, only processed and not raw data was shared. Any sharing of data via a repository, whether raw or processed, requires time and effort, and demonstrates that authors are attempting to be open. If researchers do not understand how important open raw data is, improved training on the concept is required.

An additional concern is the relatively high number of studies (~50) that shared data only in figures. Although figures are readable to humans, they are not easily machine readable and thus it is difficult to automatically extract data from them. All of the full text articles in PLOS ONE are available to download as XML (extensible markup language) files, which includes any tables in the main body of the text, allowing computers to easily process them [263]. Tables will therefore also be searched by any text mining performed on these XML files. However, figures will not be searched (neither will any files in the supplementary material).

The poor linking from data to publications and vice versa is also problematic. A fundamental problem of the current system of dissemination of scientific results via journal articles is that they are static documents, and it is usually not possible to update them with new information such as where data underlying the article is located, if the data is deposited after the journal article. However, it is usually relatively simple to update the descriptions of open data in repositories to include links to associated publications, and currently missing links could be added in future.

Both articles that stated raw data were available on MetaboLights, but the data were not publicly available, were published in 2017. This means that the data may still be under review and will be published on the repository in the future.

Whilst MetaboLights has been a PLOS ONE recommended repository for omics data since June 2015, Metabolomics Workbench has only been recommended since 2018. It would therefore be expected that more PLOS ONE studies would link to data in Metabolights than Metabolomics Workbench.

Field-specific repositories have stricter submission requirements and more detailed guidelines than general repositories. It is therefore to be expected that a greater percentage of studies in dedicated metabolomics repositories publicly share raw data than those in general repositories. Researchers who share data in metabolomics specific repositories may also have greater domain specific knowledge and be more aware of the importance of sharing raw data. Given that raw data is most commonly shared in repositories, it is unsurprising that studies with a data availability statement that referenced a metabolomics or general repository (type C) were the most likely to make raw data publicly available.

PLOS ONE is a forerunner in promoting open scientific practices. If the rates of public sharing of raw data that underlie journal articles published in PLOS ONE are so low, it is likely that they will be even worse in other journals that have less data availability requirements. The metabolomics community must learn from the practices of other omics, in order to increase the amount of open data.

3.3 Data Reuse in Metabolomics

Data sharing is a means to an end, and alone is insufficient; if shared data is incomplete, lacking in quality or missing crucial metadata, it is not valuable. Instead, data reuse is evidence of the value of open data [264]. If a dataset is reused this demonstrates that its sharing was worthwhile and justifies the cost of hosting the data.

The value of open data has been demonstrated in transcriptomics and proteomics [265, 266]. Publicly available transcriptomics data has been reused in experimental and computational studies, meta-analyses, resources, software tools, and method and ontology development [265]. Four main ways proteomics data can be utilised have been identified: use, reuse, reprocess and repurpose [266]. Use refers to direct use, such as looking up information about a given protein in a knowledge-base such as UniProt; in reuse data are reused to conduct new experiments, that can produce new insight or in spectral libraries. Data can be reprocessed using different software or parameters with the aim of extracting new biological information, but having the same goal as the original experiment. Repurposed data are used for the analysis of questions entirely different from those of the original study.

However, measuring the reuse of data objectively is challenging. There are multiple ways a researcher may cite a given dataset — citing either the original paper the dataset was described in, a data descriptor (whose sole purpose is to describe the dataset) or the dataset itself. One tool that has been developed to track data citations is Thomson Reuters' Data Citation Index (DCI)[6], which links datasets to citations received from papers indexed in the Web of Science. The DCI indexes major data repositories such as Gene Expression Omnibus, UniProt and Protein Data Bank.

Only 18.3% of the datasets indexed on the DCI have received a citation [267]. He and Nahar (2016) [268] found that 84% of data citations on Dryad were self-citing (the journal article and the data had the same title). There were only 86 non-self citations of the 7185 datasets present on Dryad Digital Repository by the end of 2014. However, both of the preceding studies found that biological research was one of the areas with the highest data reuse [267, 268].

Where data citation has been adopted, using only publisher supplied accession numbers misses many data citations; many more citations can be extracted using text mining [269], especially in supplementary material [270]. Authors may be told by publishers to not include data citations in the main body of the text and instead only in the supplementary material. Bousfield *et al.* (2016) [271] caution against the use of simple metrics, such as citation counts

[6] http://wokinfo.com/products_tools/multidisciplinary/dci/, accessed 10th April 2018

from DCI, for measuring research impact as they fail to capture many examples of data reuse. Approximately 8,000 patents contained data citations in 2014.

As metabolomics is still a young field, to date there have been no attempts to systematically measure the reuse of publicly available metabolomics data. The only research that discusses data reuse in metabolomics, references only three studies [122], or discusses only the authors' own experiences of reusing metabolomics data [272]. This research aims to identify all examples of reuse of metabolomics data (at the time the study was conducted), which repositories reused data are stored in and the frequency of reuse per study.

3.3.1 Methods

To find publications that reused open metabolomics data methods from Rung & Brazma's 2013 analysis [265] were adapted. They analysed papers published in a single year (2011) that were either tagged with "ArrayExpress" or cited any of the five ArrayExpress publications in the Nucleic Acid Research journal database issues. Papers that did not directly reuse the data were filtered, leaving 90 papers.

In this research journal articles that cited any of the publications of the metabolomics data repositories: GNPS [128], MetaboLights [126, 172, 273, 274], Metabolomics Workbench [127], MetaPhen [124, 125] and MeRy-B [123], were examined. Europe PMC was also searched for the repositories ("Global Natural Product Social Molecular Networking", "MetaboLights", "Metabolomics Workbench", "MetabolomeExpress", "MetaPhen", "Metabolomics Repository Bordeaux" and "MeRy-B") and for accession numbers: "MTBLS*", "ST000*" and "MSV000*". The term "GNPS" was not searched, as GNPS is a widely used acronym in multiple areas of science, including as an abbreviation of gold nanoparticles.

The full-text of papers returned by these searches were then manually examined to see if they actually reused metabolomics data, or were a primary study generating data, or if they simply referenced a repository. Studies that reused publicly available metabolomics data were then classified as either: "Biological Study", "Metadata", "Methods", "Resource" or "Software". These categories were adapted from Rung & Brazma's analysis [265]. The studies were then examined to see whether they reused data produced by the same set of authors, submitters or study owners.

Finally, the frequency of studies published per year per repository and the average time until data reuse were investigated.

3.3.2 Results

As of 15th February 2018, a total of 33 instances of reuse of metabolomics data available via a dedicated repository have been identified (Table D.4). There was a large increase in the frequency of data reuse in 2017, with 19 publications reusing open metabolomics data (Figure 3.9). Data from MetaboLights were reused at the highest rate. 47.06% of the studies that reused data shared at least one author with the authors, submitter or owner of the original study. There were 12 studies classified as "Methods" that reused data, 10 "Software", 6 "Resource", 4 "Metadata" and 2 "Biological Studies".

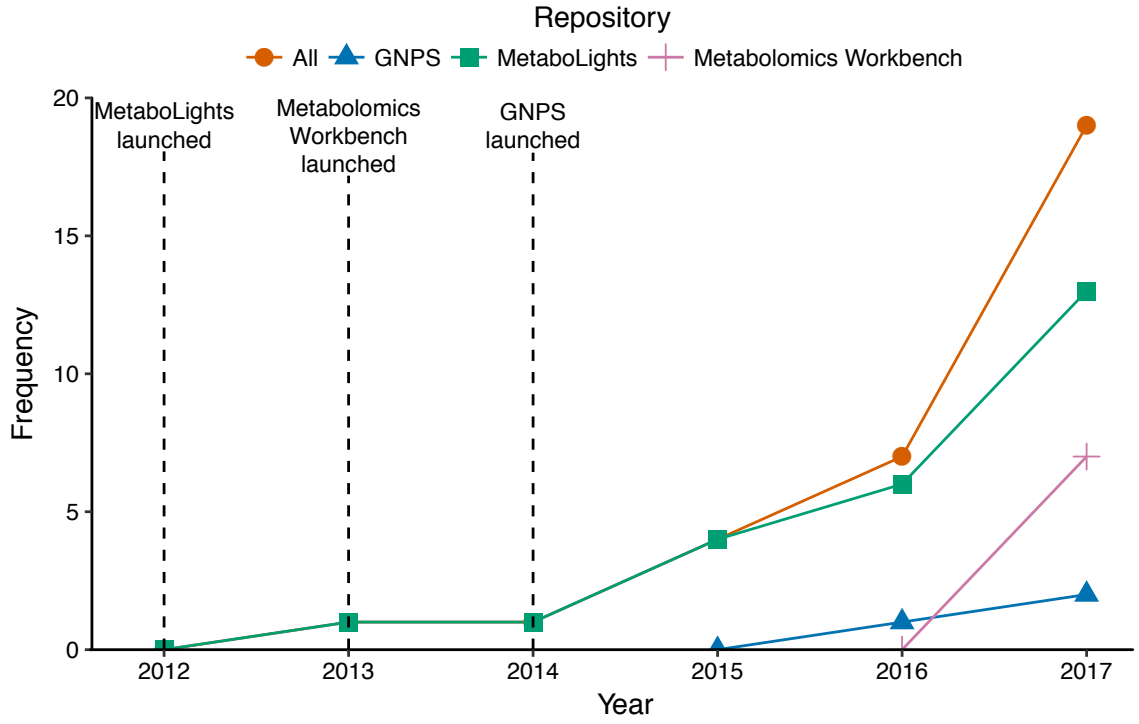


Fig. 3.9 The number of articles that reuse metabolomics data over time, as of 15th February 2018. Data reuse across all repositories is shown in orange, GNPS reuse is shown in blue, MetaboLights reuse is shown in green and Metabolomics Workbench is shown in purple. Some articles reuse data from multiple repositories (both MetaboLights and Metabolomics Workbench), so the total number of articles that reuse metabolomics data per year is not the sum of the articles that reuse data from each repository per year. The launch year of each repository is also highlighted.

In total, across repositories, 515 unique studies have been reused. However, 68.7% (354) of these studies were reused exclusively in my publication [169] (see Chapter 2), and only 161 were reused in the other 32 journal articles that have reused publicly available

metabolomics data. The studies MTBLS36[7], "Metabolic differences in ripening of Solanum lycopersicum 'Ailsa Craig' and three monogenic mutants", and MTBLS93[8], "Large-scale non-targeted serum metabolomics in the Swedish Twin Registry", have been reused at the highest frequency, both having been reused 4 times. 145 MetaboLights studies were reused a total of 195 times, 246 Metabolomics Workbench studies were reused 265 times and 85 GNPS studies were reused 90 times. Excluding reuse in my research [169], 54 MetaboLights studies were reused a total of 73 times and 25 Metabolomics Workbench studies were reused 28 times (my published research reused no GNPS studies). To the best of my knowledge, the only published example of reuse of data from MetaPhen and MeRy-B is in my research article [169].

However, it must be noted that some of the studies that reuse open metabolomics data from many studies do not state precisely which studies are reused. Inacio *et al.* (2017) [227] and Ferreira *et al.* (2017) [226] reused all of the datasets in MetaboLights that were present at a certain time point; Mohimani *et al.* (2016) [275] reused 201 GNPS studies and Gurevich *et al.* (2018) [276] reused 120 GNPS studies. Chen *et al.* (2017) [277] reused studies from MetaboLights and Metabolomics Workbench, but do not state which studies.

Of the studies that have been reused, 78.7%, have been reused only a single time, and excluding my research [169], this increases to 85.7%. The majority of metabolomics studies have never been reused (Figure 3.10). However, studies from MetaboLights have been reused at the highest rate, with 43.7% being reused at least once including Spicer *et al.* [169] (Figure 3.10A), and 16.4% excluding Spicer *et al.* [169] (Figure 3.10B).

On average, studies are reused 1.81 years after their public data release (Figure 3.11B). There has been a large increase in the total number of studies with publicly available data since 2015; nearly double the number of studies were released in 2017 than in 2016 (Figure 3.11A). The average year of publication of studies is 2016.

[7]https://www.ebi.ac.uk/metabolights/MTBLS36, accessed 12th April 2018.
[8]https://www.ebi.ac.uk/metabolights/MTBLS93, accessed 12th April 2018.

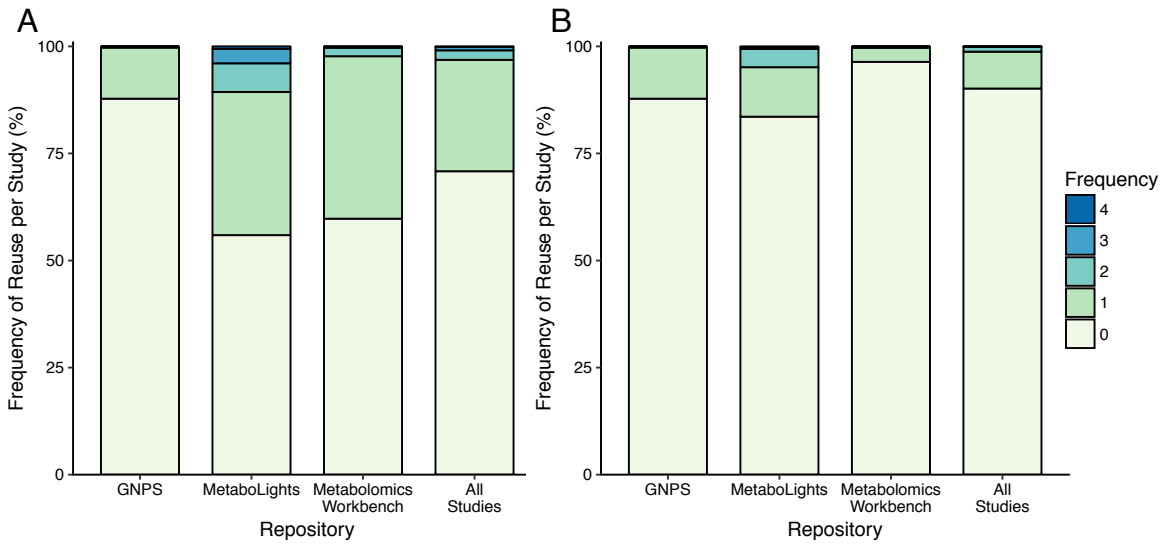


Fig. 3.10 The percentage of studies reused at each frequency: 0, 1, 2, 3, or 4 times, as of 15th February 2018, A) including reuse in all studies, B) excluding reuse by Spicer *et al.* (2017) [169].

Table 3.6 Studies that reuse publicly available metabolomics data, as of 15th February 2018. The table shows the repository(-ies) that were the source of the data, the year the article that reused the data was published, classification of how the data were reused and the article reference.

Title	Journal	Repository(-ies)	Year	Classification	Reference
Predicting Network Activity from High Throughput Metabolomics	PLOS Computational Biology	MetaboLights	2013	Methods	[278]
The Risa R/Bioconductor package: integrative data analysis from experimental metadata and back again	BMC Bioinformatics	MetaboLights	2014	Software	[279]
PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems	Analytical Chemistry	MetaboLights	2015	Software	[280]
The influence of scaling metabolomics data on model classification accuracy	Metabolomics	MetaboLights	2015	Methods	[281]
Joint Analysis of Dependent Features within Compound Spectra Can Improve Detection of Differential Features	Frontiers in Bioengineering and Biotechnology	MetaboLights	2015	Methods	[282]
BiNChE: A web tool and library for chemical enrichment analysis based on the ChEBI ontology	BMC Bioinformatics	MetaboLights	2015	Resource	[283]

Title	Journal	Repository(-ies)	Year	Classification	Reference
Approaches to sample size determination for multivariate data: Applications to PCA and PLS-DA of omics data	Journal of Proteome Research	MetaboLights	2016	Methods	[284]
Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data	GigaScience	MetaboLights	2016	Software	[285]
Effect of Insulin Resistance on Monounsaturated Fatty Acid Levels: A Multi-cohort Non-targeted Metabolomics and Mendelian Randomization Study	PLOS Genetics	MetaboLights	2016	Biological Study	[286]
Non-targeted metabolomics combined with genetic analyses identifies bile acid synthesis and phospholipid metabolism as being associated with incident type 2 diabetes	Diabetologia	MetaboLights	2016	Biological Study	[287]
Partial least squares with structured output for modelling the metabolomics data obtained from complex experimental designs: A study into the Y-block coding	Metabolites	MetaboLights	2016	Methods	[288]
Dereplication of peptidic natural products through database search of mass spectra	Nature Chemical Biology	GNPS	2016	Methods	[275]
DES-ncRNA: A knowledgebase for exploring information about human micro and long noncoding RNAs based on literature-mining	RNA Biology	MetaboLights	2017	Resource	[289]

Title	Journal	Repository(-ies)	Year	Classification	Reference
DES-TOMATO: A Knowledge Exploration System Focused On Tomato Species	Scientific Reports	MetaboLights	2017	Resource	[290]
MsPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics	Analytical Chemistry	MetaboLights	2017	Software	[291]
NOREVA: normalization and evaluation of MS-based metabolomics data	Nucleic Acids Research	MetaboLights	2017	Software	[292]
Untargeted metabolomics suffers from incomplete data analysis	Metabolomics	MetaboLights, Metabolomics Workbench	2017	Methods	[244]
Joint Bounding of Peaks Across Samples Improves Differential Analysis in Mass Spectrometry-Based Metabolomics	Analytical Chemistry	MetaboLights	2017	Software	[293]
LiverWiki: a wiki-based database for human liver	BMC Bioinformatics	MetaboLights, Metabolomics Workbench	2017	Resource	[277]
mzML2ISA & nmrML2ISA: generating enriched ISA-Tab metadata files from metabolomics XML data	Bioinformatics	MetaboLights	2017	Software	[228]

Title	Journal	Repository(-ies)	Year	Classification	Reference
Compliance with minimum information guidelines in public metabolomics repositories	Scientific Data	MetaboLights, Metabolomics Workbench, MetaPhen, MeRy-B	2017	Metadata	[169]
Mass Spectral Feature List Optimizer (MS-FLO): A Tool To Minimize False Positive Peak Reports in Untargeted Liquid Chromatography-Mass Spectroscopy (LC-MS) Data Processing	Analytical Chemistry	Metabolomics Workbench	2017	Software	[294]
xMSannotator: an R package for network-based annotation of high-resolution metabolomics data	Analytical Chemistry	Metabolomics Workbench	2017	Software	[295]
Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies	BMC Bioinformatics	Metabolomics Workbench	2017	Methods	[296]
Significance estimation for large scale untargeted metabolomics annotations	Nature Communications	GNPS	2017	Methods	[297]
Metadata Analyser: measuring metadata quality	Advances in Intelligent Systems and Computing	MetaboLights	2017	Metadata	[227]

Title	Journal	Repository(-ies)	Year	Classification	Reference
Proposal for a common nomenclature for fragment ions in mass spectra of lipids	PLOS ONE	MetaboLights	2017	Metadata	[298]
Assessing Public Metabolomics Metadata, Towards Improving Quality	Journal of Integrative Bioinformatics	MetaboLights	2017	Metadata	[226]
Molecular structures enumeration and virtual screening in the chemical space with RetroPath2.0	Journal of Cheminformatics	MetaboLights	2017	Resource	[299]
Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets	Scientific Reports	Metabolomics Workbench	2017	Resource	[300]
Meta-mass shift chemical profiling of metabolomes from coral reefs	PNAS	GNPS	2017	Methods	[301]
Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data	BMC Bioinformatics	MetaboLights, Metabolomics Workbench	2018	Methods	[302]
Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra	Nature Microbiology	GNPS	2018	Methods	[276]

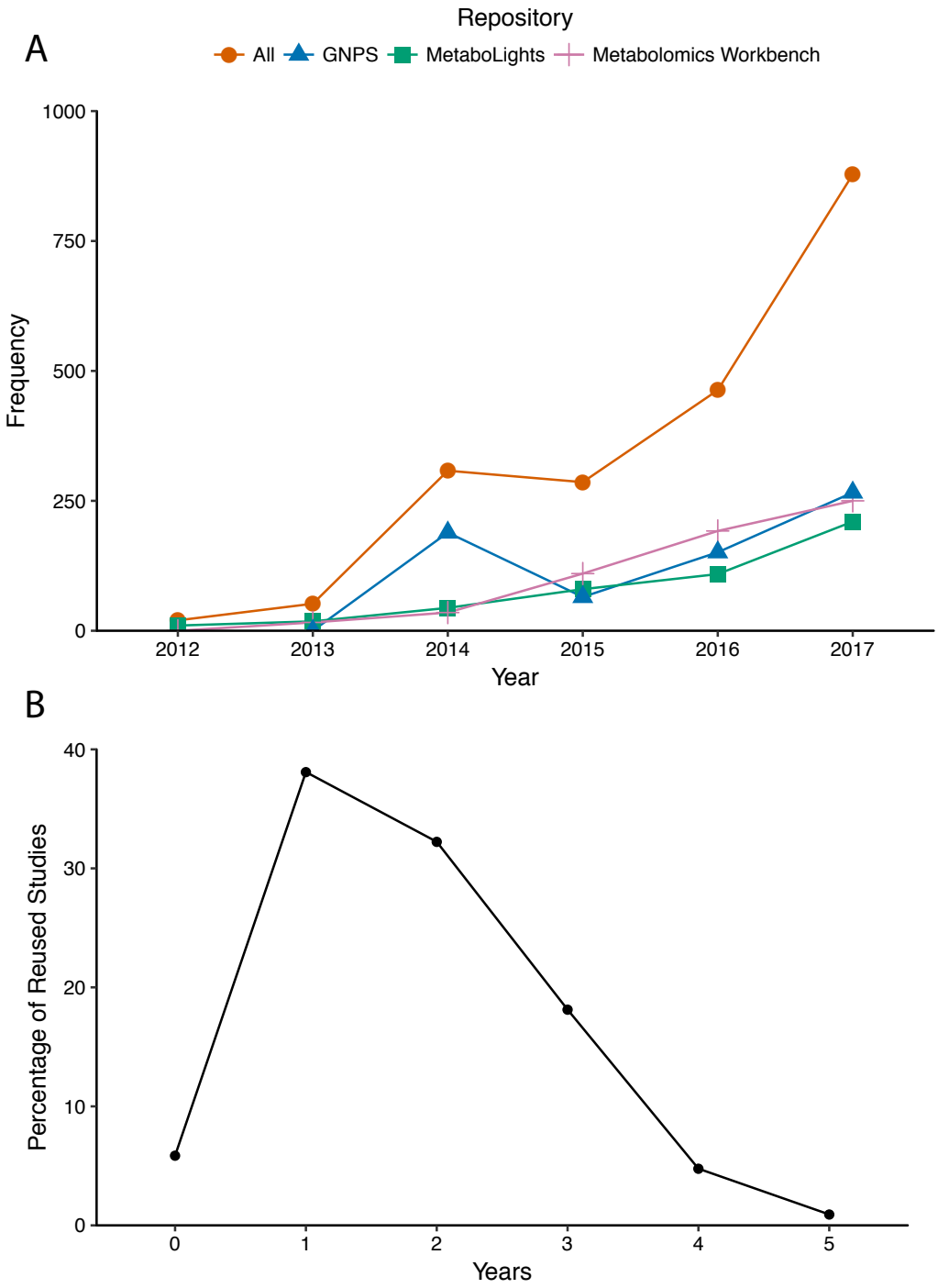


Fig. 3.11 The frequency of publicly available studies and time until data reuse. A) The frequency of metabolomics studies released per year. The total number of studies released per year is shown in orange, GNPS studies are in blue, MetaboLights are in green and Metabolomics Workbench are in purple. The spike in GNPS studies published in 2014 is likely due to researchers using GNPS to reanalyse old datasets soon after its launch. B) The time from initial data release to the publication of data reuse, as of 15th February 2018.

Data and Code Availability

The datasets generated during and/or analysed during the current study, along with the analysis code are available on the GitHub repository: https://github.com/RASpicer/Data_Reuse_Metabolomics, under the MIT License. The majority of the analysis was performed using R version 3.3.2., along with the packages ggplot2 version 2.2.1., reshape2 1.4.3, dplyr 0.7.5, tidyr 0.8.1 and knitr 1.20.

3.3.3 Discussion

These findings indicate that there is greater reuse of metabolomics data than previously assumed: Aksenov *et al.* [122] asserted that "there are fewer than a handful of papers in which data have been reused". Whilst the metabolomics community still lags behind other omics in terms of reuse, Rung and Brazma [265] identified only 90 papers published in a single year (2011) that reused data from ArrayExpress. In 2013, at the time Rung & Brazma's analysis was conducted, ArrayExpress contained data from >30,000 experiments and nearly 1,000,000 assays [121]. This is far larger than the ~1,700 metabolomics datasets identified in this analysis.

The greater amount of reuse of data from MetaboLights is unsurprising given that it is the oldest general purpose metabolomics data repository, being launched in 2012. There was also a big increase in the number of publications that reused metabolomics data in 2017, which may indicate that researchers are now becoming increasingly aware of the existence of data repositories, or are more confident in the quality of the available data.

Studies most frequently reused metabolomics data for software or methods development. Some computational research groups do not have access to laboratories, therefore high quality publicly available data is essential to allow for testing when developing new algorithms or software implementations. Conversely, only two biological studies that reused open data were performed, and both of these were published by the same group as the original research. Whilst more researchers now share raw data on public repositories, sharing sufficient metadata is still rare. The lack of biological studies that reuse data may be due to deficiencies in published biological context metadata (see Chapter 2). The groups who published biological studies that reused publicly available data may have access to additional metadata that were not made public.

One of the most common reasons researchers cite for not sharing data is the fear that they will be able to generate less publications from their data, and other researchers will scoop them [303]. Some journals have gone so far as to dub researchers who reuse data

as "research parasites" [304]. However, in this research, it was found that nearly 50% of publications that reused metabolomics data shared at least one author with the original study. This demonstrates that, in metabolomics, fears of being unable to publish new research using previously publicly shared data are unfounded.

Nearly half of all metabolomics datasets were published in 2016 or 2017. As it takes on average 1.8 years for a dataset to be reused, it would not be expected that many of these studies would have been reused yet. There is hope that many more of these datasets will be reused over the coming years. The oldest general purpose metabolomics repository is only 6 years old; if the increase in reuse continues at the current rate then the amount of data reuse in metabolomics may begin to catch up with the high frequency in transcriptomics and proteomics.

3.4 Challenges to Metabolomics Data Sharing and Reuse

Poorly annotated data with inadequate metadata is not reusable [211]. Tenenbaum *et al.* (2018) [272] reused data from four metabolomics studies of Alzheimer's disease and found that supplied metadata were human readable but not machine readable, hindering reuse. Insufficient metadata to enable data reuse is not a problem unique to metabolomics, it is also a obstacle in proteomics. Extensive annotation about the biological samples is not required by the Minimum Information About a Proteomics Experiment (MIAPE) guidelines, and insufficient metadata remains a major barrier to performing meta analysis [305].

Currently, raw metabolomics data is often shared in proprietary vendor formats, rather than in open formats e.g. .mzML [220]. This can limit reuse, as libraries that can be used to convert data to open formats run only on Microsoft Windows operating systems [241]. There has also been no adoption of formats that allow for the standardised sharing of processed data such as mzIdentML [306] and mzQuantML [221], which are widely used in the proteomics community.

A major challenge to the reuse of metabolomics data is the lack of direct comparability between datasets. The most widely used analytical technique in metabolomics is untargeted LC-MS, which suffers from the limitation of instrumental parameter dependent fragmentation (e.g. collision energy and resolution) [307]. This makes it very difficult to combine the results of multiple metabolomics studies, although this has been achieved for NMR metabolomics data [308].

There are legal barriers that can prevent sharing of data: privacy concerns and ethical issues. These are especially important to consider in regards to clinical human data. Alterna-

tively, researchers may be impeded from data sharing by not owning the intellectual property rights to data used in their studies.

Researchers fear that openly sharing data means that they will be able to generate less publications from it [303]. This problem is exacerbated by a majority of scientists feeling that they have ownership of any data they generate in the course of their research [309]. Whilst the finding of this research suggest that members of the metabolomics community have been able to publish additional papers from openly available data, the sample size was small and results may not be generalisable.

Data citations have not been widely adopted in the field of metabolomics. Of the studies identified in this research that reused publicly available metabolomics data, the majority that used data from >100 studies did not state precisely which studies had been reused. It also remains challenging to measure the number of data citations.

3.5 The Way Forward in Metabolomics Data Sharing

The Ethical, Legal and Social Implications (ELSI) of sharing data from research involving human participants should always be considered and protecting patient privacy must be a priority. However, except potentially in the case of rare diseases, there are currently no known means of identifying a patient from their metabolic profile. This is especially true for large cohort studies that include many patients with the same disease. There is a far greater risk of patient identification from genetic data than metabolomic data. However, repositories for sensitive genomics data have been developed, the European Genome-phenome Archive (EGA) [310] and the NCBI's Database of Genotypes and Phenotypes (dbGaP) [311]. Both of these repositories allow controlled access to datasets that cannot be made publicly available. Researchers can identify studies containing data relevant to their research by searching for diseases, variants, technology and sample type (metadata), and then apply for access to the study. The main difference between EGA and dbGaP is who authorises access to the data. In EGA each study has a Data Access Committee (DAC), whom users submit data access requests to. For dbGaP, users must submit requests to access data to the NIH institute that sponsored the study. As well as sensitive genomics data, these repositories could be used to store a greater amount of clinical metabolomics data. To date there are 12 metabolomics studies on dbGaP and none on EGA.

Both controlled access repositories could be improved for searching metabolomics data, by including the ability to search by metabolites or pathways of interest. As recommended

by the H2020 PhenoMeNal guidelines[9] for encoding data terms of use in the ISA format, metadata describing terms of use, consent availability and additional ancillary information in the repository should be encoded using the Data Use Ontology (DUO)[10]. EGA is planning to adopt DUO within the near future but has not yet done so.

The connection between the publication review process at journals and the deposition of data to public repositories (such as MetaboLights or Metabolomics Workshop) must also be improved. Potential methods to enhance this connection include Research Resource Identifiers (RRID) and project preregistration. RRID are unique, persistent identifiers that can be used for referencing a research resource, such as software, organisms or cell lines. Publishers could use RRID to link publications to data. Following the generation of the experimental design, projects can be preregistered — outlining what data and analysis will be performed prior to observing the research outcomes. Examples of repositories that allow project preregistration include the European Bioinformatics Institute's BioSamples [312] and the National Center for Biotechnology Information's BioProject [313].

In proteomics continuous identification is becoming common: data are routinely reprocessed with updated protein databases [266] in GPMDB (Global Proteome Machine and Database) [314] and PeptideAtlas [315], allowing for the identification of additional proteins. For metabolomics, currently the only repository that provides continuous identification is GNPS [128]. Reprocessing data should be practiced more widely by the metabolomics community, considering that using different data analysis methods can result in the identification of different metabolites [242–244].

Data are valuable research outputs in their own right. However, in order to give data their full credit, a cultural shift is required to citing data themselves, rather than citing journal articles. The metabolomics community must move away from accepting non-interoperable summary tables as an acceptable way of disseminating data and towards requiring raw data sharing. At the absolute minimum which metabolites were identified within a study should be reported. Sharing both raw and processed metabolomics data, in the style of the system used by ProteomeXchange [230] could also be adopted. For MS metabolomics, the data exchange formats used in proteomics could be adapted.

The importance of sharing raw data, specifically, must be highlighted. Further training is required to raise awareness as to a) what raw data is and b) why it is important to share it.

[9] https://phenomenal-h2020.eu/home/, accessed 18th April 2018.
[10] https://www.ebi.ac.uk/ols/ontologies/duo, accessed 18th April 2018.

3.6 Conclusion

Metabolomics is still lagging behind other omics in regards to data sharing. In transcriptomics and proteomics the value of open data has been demonstrated, and there are many examples of data reuse [265, 266]. In metabolomics, whilst journal policies exist that require data sharing, they are not necessarily strongly enforced, and a large majority of publications do not share their underlying data openly. This research shows that open metabolomics data is increasingly reused, which indicates its importance.

Data sharing has become the standard in more mature communities such as genomics. There is hope that time will lead to a similar situation in metabolomics. Funders and journals must move to requiring open data and must insist it be deposited in dedicated repositories, specifically developed to fulfil the needs of the community. Greater awareness must be made to the value of sharing raw data. To allow controlled access to sensitive metabolomics datasets, repositories such as EGA and dbGaP should be more widely utilised. There must also be greater effort to improve the linking of data to publications and vice versa.

Chapter 4

Improving the Discoverability of Metabolomics Tools

4.1 Metabolomics Data Analysis and Processing Software

The variety of analytical techniques used to measure specific portions of the metabolome produce different types of data. This in turn means that multiple software tools are required to analyse each distinct type of data e.g. NMR and MS. There are also many stages of data processing and analysis, with the majority of software tools being designed to accept a single instrumental data type and having functionality that covers only one or two stages of the data analytic pipeline (see section 1.2). In total there are now more than 250 different software tools specifically designed for processing and analysing metabolomics data [29–31].

The majority of metabolomics practitioners' day-to-day activities now consists of a combination of wet and dry lab work and only half have dedicated bioinformatics support [28]. Which tools and methods to use for data analysis can be confusing for both experienced practitioners and those new to working in this field; the sheer number of tools makes the prospect of finding the right one daunting. The FAIR principle of findability can also be applied to scientific software and code and it is important that users are made aware of the range of tools available for data analysis and that those tools are discoverable and usable.

A number of resources that provide lists of metabolomics tools do exist. OMICtools [316], is a manually curated metadatabase of tools for the analysis of omics data, containing both commercial and open source software. Whilst it provides a lot of useful information about software beyond its functionality, including computer skills required, licensing, programming languages and interfaces, it does not contain other information that a user will require when deciding which tools to use, such as accepted input formats or whether a tool is maintained.

As it is a commercial database, users can view only 5 pages per month without signing up to the website, which is closed source.

Ms-utils[1] provides a list of tools for mass spectrometry data analysis, but it is mainly focused on proteomics. The Fiehn lab website[2] and the metabolomics society webpage[3] also contain lists of metabolomics software. Additionally, there is the Elixir run bio.tools[4], which is a portal for bioinformatics resources, including metabolomics. However, again these are not comprehensive lists, and are not updated to include the mostly recently released tools.

This work aims to address the limitations of the previously mentioned databases and improve the discoverability of metabolomics software tools by producing a) a taxonomy of types of tools and b) a GitHub Pages hosted wiki (https://raspicer.github.io/MetabolomicsTools/) that provides lists of tools as classified by the taxonomy. An accompanying review article, Spicer *et al.* (2017) [29], was also produced, where the functionality provided by the most popular metabolomics software tools was discussed.

4.2 Metabolomics Tools Taxonomy

Taxonomies are systems of classification, many of which have a hierarchical structure. During my initial investigation of metabolomics software tools, I was frustrated by the difficulty in finding tools that performed specific tasks, and began to manually categorize the software I identified. At the behest of colleagues I worked to improve this system, with the aim of making it easier for researchers to find suitable tools for their analyses. This classification system was then reviewed by experts within the field, including the PhenoMeNal partners, Tim Ebbels (Faculty of Medicine, Department of Surgery & Cancer, Imperial College London) and Steffen Neumann (Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry), and improved as per their recommendations.

The current version of the metabolomics software tools taxonomy is visualised in Figure 4.1. An interactive version implemented on Coggle is also available on the Metabolomics Tool Wiki[5], which includes direct links to Metabolomics Tool Wiki pages.

[1] https://www.ms-utils.org/, accessed 25th March 2018.

[2] http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics, accessed 25th March 2018.

[3] http://metabolomicssociety.org/resources/metabolomics-software, accessed 25th March 2018.

[4] https://bio.tools/, accessed 7th May 2018

[5] https://coggle.it/diagram/WpbhUR0nGfCNFhej/t/metabolomics-software-tools/7a2ecb682f32801b2daa75171a1adf16661a2479f2a8ba107d5b26977b386aec, accessed 13th March 2018.

A potential use of the taxonomy is as a training resource, to educate scientists new to the field about metabolomics data analysis. This could be for new doctoral students or principle investigators looking to move into metabolomics from other areas.

The taxonomy classifies software by three main categories: *Approaches, Functionality* and *Instrument Data Type.* Approaches consists of different kinds of experimental designs, including *Targeted* and *Untargeted Metabolomics*, measuring only specific proportions of the metabolome such as *Lipidomics*, *Isotopic Labelling Analysis* and *Multiomics*. Functionality categorizes tools by which portion(s) of the metabolomics data analysis workflow they provide: *Preprocessing*, *Annotation*, *Post-processing*, *Statistical Analysis*, *Pathway Analysis*, *Workflows*, *Optimisation* and *Other Tools*. Instrumental Data Types is split hierarchically first into general categories i.e. *Spectrometry*, *Spectroscopy* and *Chromatography*, which are then divided into subcategories for specific instrument types e.g. *Mass Spectrometry*, *Raman Spectroscopy*, and then further into specific instrumental data types e.g. ^{1}H-^{1}H *TOCSY NMR* or properties e.g. *Centroid MS*.

In the following sections 4.2.1, 4.2.2 and 4.2.3, the categories and sub-catergoies will be discussed in further detail. All software can be assigned a functionality, however software designed for the certain stages of data analysis: post-processing, statistical analysis and pathway analysis, can be appropriate for all metabolomics data, and therefore may not be assigned an approach or instrumental data type.

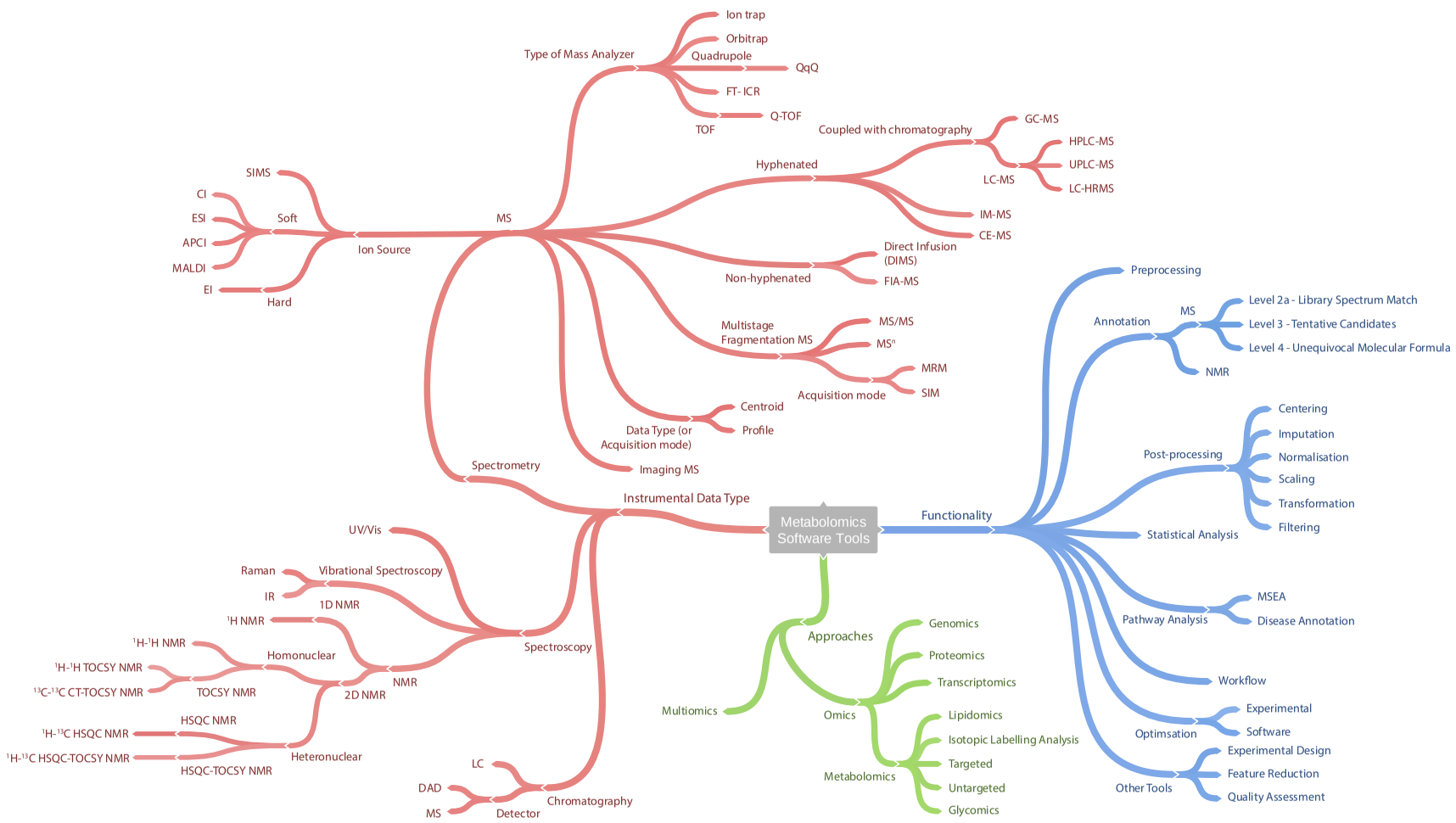


Fig. 4.1 Metabolomics Tools Taxonomy. Software is classified by three categories: Approaches, Instrument Data Type and Functionality. The taxonomy was visualised using Coggle.

4.2.1 Functionality

Preprocessing Preliminary data processing. In MS metabolomics this typically consists of peak picking, deconvolution, peak matching and peak alignment across samples, but can also include baseline correction, noise reduction and smoothing, depending on the algorithm used. Signals are generated as free induction decay (FID) in NMR metabolomics, and must be first transformed into frequency spectra prior to any subsequent analysis [27]. This means that the preprocessing of NMR metabolomics data differs from MS, with the first stages consisting of zero-filling, apodization, Fourier transformation and phase correction [317]. Later stages of baseline correction, deconvolution, binning and peak alignment are the same as for MS, although the precise algorithms used may vary. Software for preprocessing is therefore usually specifically targeted at either MS or NMR data.

Annotation Metabolite annotation refers to the process by which structural information or molecular formulas are inferred in order to identify features. It is the most time consuming stage of metabolomics analysis for many practitioners [28]. The challenges of metabolite annotation vary between analytical techniques: only limited structural information can be obtained from mass spectrometry, whilst NMR has limitations in resolution due to overlapping signals. To address these challenges dedicated software is required, and annotation software is divided into *MS* and *NMR*. The Schymaski criteria [214] are then used to further subdivide *MS* software into *Level 2a — Library Spectrum Match*, *Level 3 — Tentative Candidates* and *Level 4 — Unequivocal Molecular Formula*. These criteria were chosen over the original MSI criteria as they provide clearer classification of metabolite annotation assignment confidence (see Chapter 2).

MS

Level 2a — Library Spectrum Match Dedicated spectral libraries that contain spectra acquired using authenticated chemical standards, for ESI-MS/MS, MS^n and GC-MS spectra, are searched against to provide features with metabolite annotation via library spectrum matching. Examples of databases used include HMDB [132], mzCloud[6], METLIN [318], MassBank [137], and NIST[7]. Software can perform automatic database matching, allowing the user to search multiple MS/MS databases simultaneously.

[6] https://www.mzcloud.org/, accessed 27th March 2018.
[7] https://www.nist.gov/srd/nist-standard-reference-database-1a-v17, accessed 25th March 2018.

Level 3 — Tentative Candidates When it is not possible to unambiguously assign a metabolite to a feature, it can instead be annotated with tentative candidates. Metabolite databases (HMDB [132], METLIN [318], KEGG [135], etc.) can be either searched manually or automatically using dedicated software tools to annotate features with tentative candidate metabolites. MS/MS data is not necessarily required for the assignment of tentative metabolite candidates.

Level 4 — Unequivocal Molecular Formula When there is insufficient evidence to assign a structure to a feature, but adequate information to unambiguously assign a molecular formula. Software tools are able to assign features with adduct, isotope, neutral loss and fragment information. Molecular formula annotation is appropriate for low quality MS/MS data and MS data lacking retention time information.

NMR Identification of NMR metabolomics data features as metabolites. Like with MS/MS data, spectral libraries that contain NMR spectra from authentic chemical standards of metabolites are used for metabolite identification. Examples include HMDB [132], Biological Magnetic Resonance Data Bank (BMRB) [133] and Birmingham Metabolite Library (BML-NMR) [319]. NMR is inherently a far more quantitative technique than MS [320], with consistent chemical shifts [217]. Software for NMR metabolite identification usually also provide quantification.

Post-processing Additional data wrangling, prior to statistical analysis (alternatively called data pretreatment). Post-processing encompasses *Filtering*, *Imputation*, *Normalisation*, *Scaling*, *Centering* and *Transformation*.

Filtering Applying thresholds to parameters such as signal-to-noise ratio or the minimum percentage of samples a feature must be detected in (consensus features) to remove features which are not found in a minimum number of samples [321].

Imputation Replacing missing values with substitute values. Methods for imputation include: zero, mean, median, random forest, half minimum, k-nearest neighbors (kNN), and Multivariate Imputation by Chained Equations (MICE) [322, 323].

Normalisation The abundance of features is adjusted to reduce unwanted variation between samples (i.e. not true biological variation) [324]. Examples include probabilistic quotient normalisation (PQN), normalisation factor for each individual molecular species (NOMIS) and remove unwanted variation (RUV) [292].

Scaling Each feature is divided by a scaling factor, changing the emphasis to different aspects of the data. Examples include autoscaling, Pareto scaling and Vast scaling [281].

Centering Converting the abundance to fluctuations around zero instead of around the mean to remove the offset [325].

Transformation Nonlinear conversions of the data that change the emphasis to different aspects of the data [325]. Log transformations are the most commonly used in metabolomics.

Statistical analysis Following post-processing, metabolomics data from all analytical techniques will be in the form of a matrix of signal intensities, and can be explored to discover underlying patterns and trends. The unsupervised method principal components analysis (PCA) is generally used as an initial exploratory technique. The supervised methods *partial least squares (PLS) regression* (also called projection to latent structures), *partial least squares - discriminant analysis (PLS-DA)* and *orthogonal partial least squares (OPLS)* are also used, however these techniques have been criticized as they can lead to overfitting [326, 327], although validation techniques can be used to evaluate this. More recently other methods are being more widely used as alternatives to PLS-DA [328]: *principal component-discriminant function analysis (PC-DFA)*, *support vector machine (SVM)* and *random forest (RF)*. Univariate analyses are also applied, with *analysis of variance (ANOVA)* and *t-tests*, along with their non-parametric equivalents, being the most widely used [28]. These statistical techniques are widely used across many fields and are therefore implemented in many general statistical analysis software applications that are not specifically designed for metabolomics analysis, as well as metabolomics specific packages.

Pathway analysis This is used to investigate which metabolic pathways are present at a higher frequency, "enriched", between different experimental conditions [329]. Enriched pathways are identified by searching against pathway databases such as KEGG [135], MetaCyc [330] and SMPDB (The Small Molecule Pathway Database) [331]. Pathway analysis is important for investigating biological functionality. Many pathway analysis applications are not specific for metabolomics, and are instead applicable to multiple types of omics data.

Workflows Software comprised of multiple interconnected tools, that encompass multiple, or all, stages of data analysis. They are designed so that the majority, or the entirety of data analysis can be performed using a single tool, rather than having to use separate tools for each stage of analysis. The scope of workflows varies greatly, with some including a lot of in-house software and others being workflow management systems, combining existing tools into workflows.

Optimisation Tools designed to improve parameters and optimise feature selection, using either **Experimental Optimisation** or **Software Optimisation**. Experimental Optimisation requires researchers to perform experiments in a specified way.

Other tools Tools that do not fit into any of the previously mentioned categories. These tools do not have a standardised place in the analysis workflow; where they fit into the workflow depends on their particular functionality. Examples include tools for **Experimental Design**, **Quality Assessment** and **Feature Reduction**.

4.2.2 Approaches

Omics Some techniques of post-processing, statistical and pathway analysis are applicable to all types of omics experiment, and therefore tools that provide these may have been specifically designed for the analysis of multiple different kinds of omics data. Included omics are: **Metabolomics**, **Genomics**, **Transcriptomics** and **Proteomics**.

Multiomics Datasets from multiple different kinds of omics e.g. *metabolomics* and *genomics* are combined into a single dataset. Functions of software tools for multiomics datasets may be to combine individual omics datasets or pathway analysis.

Targeted or Untargeted

Targeted The abundance of only specific metabolites of interest are measured in targeted approaches. Tools for analysing targeted data typically require the list of measured metabolites as part of data input.

Untargeted The abundance of as many metabolites as possible are measured, within the given input parameters e.g. m/z range. Novel metabolites can potentially be detected in untargeted studies.

Metabolomics Sub-types The metabolome is comprised of many types of molecule, such as *amino acids*, *nucleotides*, *lipids* and *glycans*. Some studies aim to measure only

one of these specific subsections of the metabolome e.g. **Lipidomics** and **Glycomics**. Tools have been developed to specifically process these types of data.

Isotopic Labelling Analysis The passage of isotopes e.g. ^{13}C through a system (a cell, metabolic pathway or reaction) can be tracked, allowing metabolic flux to be measured with absolute quantification. Software tools for isotopic labelling analysis often require a specific experimental design: two samples that differ only in the isotopic labelling of one specific element, with one sample being labelled and the other not.

4.2.3 Instrumental Data Type

Spectrometry "The measurement of electromagnetic radiation as a means of obtaining information about physical systems and their components"[8].

MS An analytical technique where samples are ionised and the ions are then sorted by their mass-to-charge ratio (m/z). Mass spectrometry (MS) is the most widely used analytical technique in metabolomics [28], so much software has been designed for the analysis of MS metabolomics data. Software tools generally specify the subtype of MS data they are designed for, and the subsequent taxa expand upon these potential subtypes.

Ion Source The analyte is ionised by the ion source to produce gas phase ions, which are suitable for resolution in the mass analyzer. There are two main types of ionisation: *Soft* and *Hard*. *Secondary-ion mass spectrometry (SIMS)* is a specific type of mass spectrometry where "secondary ions" are measured. Not all types of ionisation are appropriate for use with every chromatography type, and some software expressly states the ionisation type of data it is designed for.

Hard ionisation High quantities of energy are imparted onto the analyte, leaving the molecule in an excited state. To relax, bonds are ruptured to remove the excess energy, which results in a large degree of fragmentation. Electron ionisation **(EI)** is the most commonly used type of hard ionisation. EI is generally used with GC-MS, however there are instances of it also being used with LC-MS [332].

Soft ionisation Fragmentation is minimised by soft ionisation, as less residual energy is imparted onto the analyte. Examples include electrospray ionisation **(ESI)**, chemical ionisation **(CI)**, atmospheric-pressure chemical ionisation **(APCI)** and matrix-assisted laser desorption/ionisation **(MALDI)**. It is most common for

[8] https://goldbook.iupac.org/html/S/S05848.html, accessed April 16th 2018.

LC-MS metabolomics experiments to use ESI. Whilst ESI is considered a soft ionisation technique, some metabolites will fragment with neutral losses, which is utilised for MS/MS.

SIMS The surface of the sample is sputtered by bombarding it with a high-energy beam of primary ions. This results in the emission of secondary ions, which are then analysed by the mass analyzer. SIMS is being increasingly used in imaging metabolomics.

Type of Mass Analyzer Ions are separated by their mass to charge ratios and then targeted onto detectors by mass analyzers. There are multiple different types including **Ion trap**, **Orbitrap**, **Quadrupole**, triple quadrupole **(QqQ)**, Fourier-transform ion cyclotron resonance **(FT-ICR)**, time-of-flight **(TOF)** and quadrupole-time-of-flight **(Q-TOF)**. Some software is specifically designed for handling data gathered using a particular kind of mass analyzer.

Data Type There are two main types of MS data: *centroid* or *profile*. Some mass spectrometers allow for data to be specified at acquisition, others will only collect data in profile mode. Data can be converted from profile mode to centroid by using software such as msconvert (available as part of the proteowizard platform) [333].

Profile Peaks are displayed as curves, which are constituted of a collection of signals acquired over multiple scans. It is easier to distinguish true peaks from noise when data is in profile mode.

Centroid Peaks are represented as discrete m/z values with zero line width. Centroid data has significantly smaller file sizes than profile. Some algorithms, such as centWave [334], which is implemented as part of XCMS [335] require data to be in the centroid format.

Multistage Fragmentation MS Multiple stages of mass spectrometry selection are used in order to fragment precursor ions into product ions. Precursor ions (MS1) are first selected and then fragmented (MS2) to generate products in tandem MS (**MS/MS**). Sequential mass spectrometry (**MS^n**) involves multiple stages of fragmentation, and can only be performed by certain trapping instruments (quadrupole ion-trap and FT-ICR). Tandem MS is very important for enabling metabolite annotation: without MS/MS data it can be hard to annotate beyond a molecular formula. Two frequently used modes of acquiring MS/MS data in metabolomics are: *multiple reaction monitoring (MRM)* and *selected ion monitoring (SIM)*.

Acquisition Mode

MRM Specific collision induced dissociations (CID) reactions can be monitored with MRM. It is also known as selected reaction monitoring (SRM).

SIM Only a limited m/z range is detected by the mass spectrometer. Compared to full scan mode, where a wide range of m/z values are detected, SIM is a more sensitive technique, as every m/z is scanned for a longer period of time.

Non-hyphenated

DIMS The sample is introduced directly into the mass analyzer in direct infusion mass spectrometry (DIMS), without the use of any chromatography or other separation techniques. A chip-based nanoelectrospray ion source is used. It is the simplest form of ESI-MS used in metabolomics, and is most commonly used for lipidomics analysis.

FIA-MS Flow injection analysis-mass spectrometry (FIA-MS), is a type of DIMS. It is a more high-throughtput method than LC-MS, however suffers from the limitation of providing only m/z, being unable to separate isomers.

Hyphenated The coupling of a separation technique to mass spectrometry. In metabolomics is it more common to use hyphenated MS than DIMS. Hyphenated MS can be split into *chromatography based approaches* and those which do not use chromatography such as: *Capillary electrophoresis-mass spectrometry (CE-MS)* and *ion-mobility spectrometry-mass spectrometry (IMS-MS)*.

CE-MS Capillary electrophoresis is a liquid separation method, where differences in intrinsic electrophoretic mobility are used to separate compounds. CE-MS is particularly appropriate for analysing polar and charged metabolites [336].

IMS-MS In ion-mobility spectrometry-mass spectrometry (IMS-MS), ions are separated by their mobility through an inert gas [337]. More compounds can be detected using IMS-MS compared to using MS alone.

Coupled to Chromatography A subtype of hyphenated MS, where chromatography is coupled to mass spectrometry. MS coupled to chromatography is the most widely used type of hyphenation in metabolomics. Analytes are separated based on their relative affinity for the stationary phase of the column: those with a lower affinity will elute first. The most regularly used types are *liquid chromatography-mass spectrometry (LC-MS)* and *gas chromatography-mass spectrometry (GC-MS)*.

GC-MS In GC-MS, reproducible fragmentation patterns are produced, when ionisation is performed at a fixed voltage, and are not instrument dependent [33]. Retention times are generally converted into instrument independent retention indices (RI), for comparison to existing databases for compound identification. Whilst GC-MS is a robust analytical technique, it is only able to detect volatile and thermally stable compounds and those that can be rendered volatile by chemical derivatization. This means that despite GC-MS's greater reproducibility and established metabolite libraries for metabolite identification, it is used less frequently than LC-MS. **GCxGC-MS**, where a pair GC columns are connected via a modulator, is also becoming an increasingly used analytical technique, which additionally requires dedicated software [338].

LC-MS A large variety of metabolites can be detected using LC-MS, in both positive and negative charge ionisation mode. Because of its suitability for untargeted metabolomics, LC-MS is the most widely used analytical technique. There is also the greatest amount of software specifically designed for the analysis of metabolomics LC-MS data. More recently variations of LC-MS have emerged including high performance liquid chromatography-mass spectrometry **(HPLC-MS)**, ultra performance liquid chromatography-mass spectrometry **(UPLC-MS)** and liquid chromatography-high resolution mass spectrometry **(LC-HRMS)**.

Imaging MS Mass spectrometry imaging is used to measure the spatial distribution of metabolites within tissues [339]. Specialised software is required in order to enable visualisation of imaging MS data.

Spectroscopy "The study of physical systems by the electromagnetic radiation with which they interact or that they produce"[9].

NMR Nuclear magnetic resonance (NMR) is an analytical technique based on the principle that all nuclei have spin. Certain isotopes of chemical elements have a magnetic moment. The magnetic moments of nuclei are measured by subjecting the analyte to a strong magnetic field from which nuclei absorb energy.

1D One dimensional NMR data has a frequency axis of chemical shifts in ppm and an axis of intensities.

[9] https://goldbook.iupac.org/html/S/S05848.html, accessed April 16th 2018.

^{1}H NMR The most commonly used type of NMR in metabolomics is 1D ^{1}H NMR, due to high abundance of ^{1}H in nature and low relaxation time [340]. However, a challenge is that 1D ^{1}H NMR spectra can contain overlapping peaks, which some software has attempted to overcome.

2D Two dimensional NMR contains two frequency axes of chemical shifts and one axis of intensities. In 2D NMR magnetization transfer is measured. This can either be through bonds of the same type of nucleus *homonuclear*, bonds of different types of nucleus *heteronuclear* or through space e.g. *nuclear overhauser effect spectroscopy (NOESY)*.

Homonuclear Two nuclei of the same type are correlated with each other, via magnetization transfer, through J-coupling of nuclei connected by up to a few bonds. Some software does not specify which homonuclear analysis they are designed for only the type of nucleus i.e. **^{1}H-^{1}H NMR**, whilst others specify the type e.g. *total correlation spectroscopy (TOCSY)*.

TOCSY A powerful variant of *correlation spectroscopy (COSY)* that can detect small couplings. The most common type used in metabolomics is **^{1}H-^{1}H TOCSY NMR** but **^{13}C-^{13}C constant-time (CT) TOCSY NMR** has also been used.

Heteronuclear Two nuclei of different types are correlated with each other. Variants used in metabolomics include *heteronuclear single quantum coherence (HSQC)* and *heteronuclear single quantum coherence-total correlation spectroscopy (HSQC-TOCSY)*.

HSQC Magnetization of the proton is transfered to the second nucleus, and then back to the proton, improving sensitivity. In metabolomics carbon is most frequently used as the second nucleus: **^{1}H-^{13}C HSQC NMR**.

HSQC TOCSY A hybrid inverse experiment that consists of an initial HSQC pulse train followed by a TOCSY spin lock. The most commonly used variant in metabolomics is **^{1}H-^{13}C HSQC TOCSY**.

Vibrational spectroscopy A non-destructive method of compound identification where infrared or near-infrared is used to create vibrations in chemical species, in order to measure the vibrational energy of compounds. It is used to collectively refer to *Infrared (IR)* and *Raman* spectroscopy, which differ in the types of vibrations and transitions that they measure. Some software has been designed specifically for the analysis of these data types.

IR A beam of light within the infrared spectrum is used to excite bonds and the absorption and transmission of lights due to vibrations of the bonds is then measured. It depends upon a change in dipole moment, rather than the actual polarisation, and is sensitive to hetero-nuclear functional group vibrations and polar bonds.

Raman Inelastic (or Raman) scattering is used to probe molecular vibrations of analytes in order to identify molecular fingerprints, depending on a change in polarizability. Compared to IR, it can use a much wider range of wavelengths, but measures only a narrow type of vibration. It is sensitive to homo-nuclear molecular bonds.

UV/Vis Ultraviolet–visible spectroscopy (UV/Vis or UV-Vis) is an absorption or reflectance spectroscopy in the ultraviolet-visible spectral region. Whilst it is not currently widely used in metabolomics, there is software designed for handling UV/Vis metabolomics data.

Chromatography An analytical technique that is used to separate mixtures by the different partition coefficients of their components. In metabolomics sometimes liquid chromatography is used without being coupled to mass spectrometry; instead other detectors are used, e.g. diode-array detection (DAD). Therefore specific software is required for handling this type of metabolomics data.

DAD A type of UV detector, where a photodiode array is used to record UV/Vis absorption spectra of samples passing through LC. Like with mass spectral detection, variations of LC-DAD, such as **HPLC-DAD** and **UPLC-DAD**, can be used.

4.3 Metabolomics Tools Wiki

In complement to the metabolomics tools taxonomy the **Metabolomics Tools Wiki** was created. It is available at: https://raspicer.github.io/MetabolomicsTools/. The wiki aims to provide extensive details about all included software, to enable researchers to more easily find a tool that suits their needs. All tools are classified by their functionality, and where relevant, instrumental data types and approaches.

The wiki was originally implemented using Markdown and was available at https://github.com/RASpicer/MetabolomicsTools/wiki, however GitHub Markdown tables suffer the major limitation of not being sortable without the user manually installing a plugin[10]. Therefore, the wiki was migrated to GitHub pages. The wiki is now written in HTML, with CSS styling. JavaScript is used to automatically extract tables directly from a Google Sheet (https://goo.gl/XpoJKi), which contains all of the information about the tools. The code for this was adapted from https://github.com/crunchprank/google-sheets-to-html, accessed 1st March 2018.

The main page features links to tables of tools, split into the main categories of the taxonomy: *Approaches, Functionality* and *Instrumental Data Type* (Figure 4.2), as well as two additional categories: *Language* and *Software Type*, which were added after feedback from the community. Next an interactive Coggle implementation of the taxonomy is embedded. Where a dedicated page of the wiki exists for that particular tool type, e.g. *Annotation*, the taxonomy provides a direct link to this page. All taxa with blue text have a dedicated page. The taxonomy is followed by a list of nomenclature, containing the full forms of all of the abbreviations used in the taxonomy. Next there are instructions for adding new tools to the wiki (Figure 4.3), including links to all tools already featured on the wiki. Practitioners are instructed to add to tools to the wiki using the aforementioned Google Sheet. The Google Sheet can also be used to update details about tools already on the wiki.

In total there are 43 tables of specific types of software tools on the wiki e.g. *LC-MS Tools*, *Statistical Analysis Tools* and *R Packages*, of which 41 have the same page format. They each feature a header linking to the main page, followed by the title that indicates the type of tools included on the page and then a table including all the tools on the wiki of that particular type. Figure 4.4 shows the *Workflows* page as an example of a "table of tools" page. The tables feature links to the pages of individual tools on the wiki, as well as links to the tools' websites. Details of the tool's *Functionality*, *Instrumental Data Type*, *Approaches*, *Software Type*, *Interface*, *Operating System (OS)*, *Language* and *Last Update Date* are also

[10] https://github.com/Mottie/GitHub-userscripts/wiki/GitHub-sort-content, accessed 3rd 2017

included. All of the tables are sortable by every parameter they are comprised of. Initially the table of tools pages included only functionality, instrumental data type and approaches, but after responses from the metabolomics community, software type, interface, OS, language and last update date were also added.

Compared to the other table of tools pages, the *All Tools* and *Obsolete Tools* pages differ by including additional details about tools in their tables, however, this decreases the readability of the tables and increases the time taken for them to load. These tables are designed only to serve as a reference of all the tools included on the wiki, for developers looking to see if their software are included, and are not intended for the majority of users of the wiki.

For every tool on the wiki, we have aimed to compile the tools version, web address, description, functionality, instrumental data type, approaches, computer skills, software type, interface, OS, programming language written in, dependencies, license, associated publications, PMID (PubMed unique identifier), accepted data input formats — both open and proprietary, and dates of publication and most recent update. An example of an individual tool page, for *MetaboAnalyst*, is illustrated in Figure 4.5. Whilst I have aimed to assemble as much information about each software tool as possible, much information remains missing as the version and license the software is available under are often not stated.

Obsolete tool pages differ from extant tool pages by featuring a red box across the top of the page, which states that the tool is obsolete (Figure 4.6). On the wiki, only a single table, *Obsolete Tools*, contains links to the individual obsolete tools pages, so that users will not be accidentally directly towards software that is no longer available.

As of the 21st March 2018, a total of 168 extant tools are available on the wiki, along with 13 obsolete tools. Of these 73% of tools are designed to analysis MS data and 18.45% NMR.

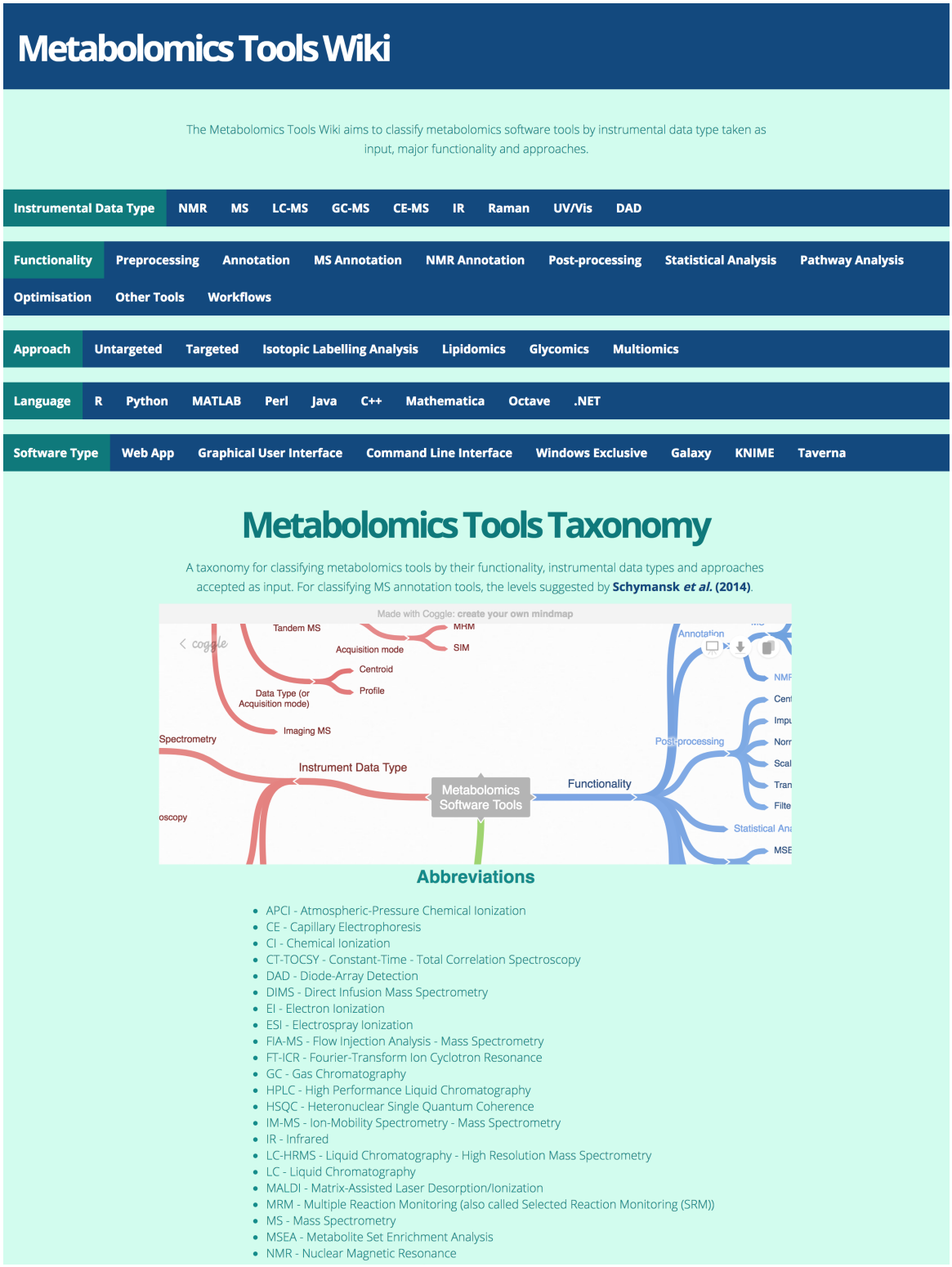


Fig. 4.2 Screenshot of the Metabolomics Tools Wiki Main Page taken on 13th March 2018.

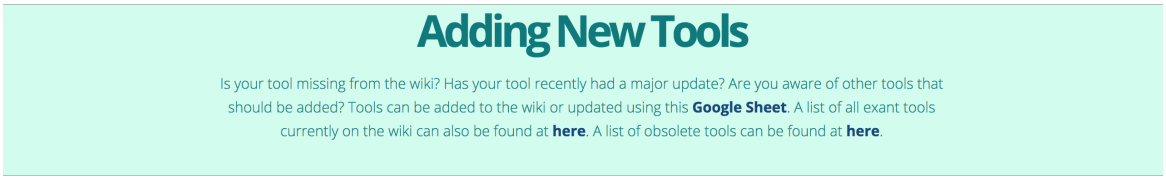
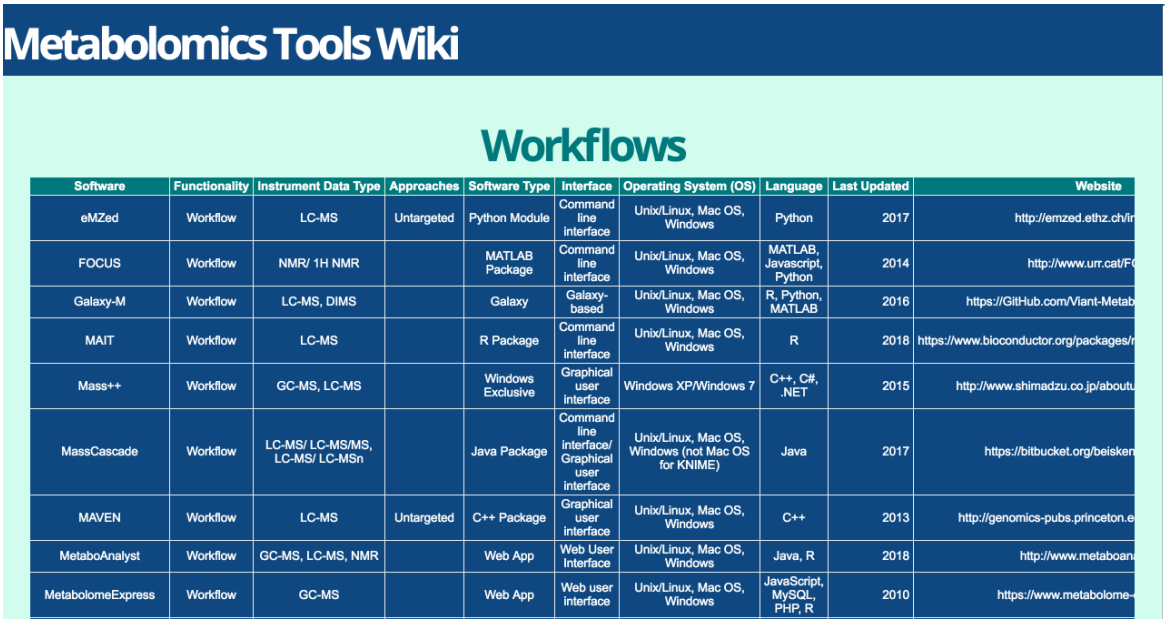


Fig. 4.3 Screenshot of the instructions for adding new tools on the Metabolomics Tools Wiki Main Page taken on 13th March 2018.



Software	Functionality	Instrument Data Type	Approaches	Software Type	Interface	Operating System (OS)	Language	Last Updated	Website
eMZed	Workflow	LC-MS	Untargeted	Python Module	Command line interface	Unix/Linux, Mac OS, Windows	Python	2017	http://emzed.ethz.ch/ir
FOCUS	Workflow	NMR/ 1H NMR		MATLAB Package	Command line interface	Unix/Linux, Mac OS, Windows	MATLAB, Javascript, Python	2014	http://www.urr.cat/F
Galaxy-M	Workflow	LC-MS, DIMS		Galaxy	Galaxy-based	Unix/Linux, Mac OS, Windows	R, Python, MATLAB	2016	https://GitHub.com/Viant-Metab
MAIT	Workflow	LC-MS		R Package	Command line interface	Unix/Linux, Mac OS, Windows	R	2018	https://www.bioconductor.org/packages/r
Mass++	Workflow	GC-MS, LC-MS		Windows Exclusive	Graphical user interface	Windows XP/Windows 7	C++, C#, .NET	2015	http://www.shimadzu.co.jp/aboutu
MassCascade	Workflow	LC-MS/ LC-MS/MS, LC-MS/ LC-MSn		Java Package	Command line interface/ Graphical user interface	Unix/Linux, Mac OS, Windows (not Mac OS for KNIME)	Java	2017	https://bitbucket.org/beisken
MAVEN	Workflow	LC-MS	Untargeted	C++ Package	Graphical user interface	Unix/Linux, Mac OS, Windows	C++	2013	http://genomics-pubs.princeton.e
MetaboAnalyst	Workflow	GC-MS, LC-MS, NMR		Web App	Web User Interface	Unix/Linux, Mac OS, Windows	Java, R	2018	http://www.metaboan
MetabolomeExpress	Workflow	GC-MS		Web App	Web user interface	Unix/Linux, Mac OS, Windows	JavaScript, MySQL, PHP, R	2010	https://www.metabolome-

Fig. 4.4 Screenshot of the Metabolomics Tools Wiki Workflows Table taken on 13th March 2018.

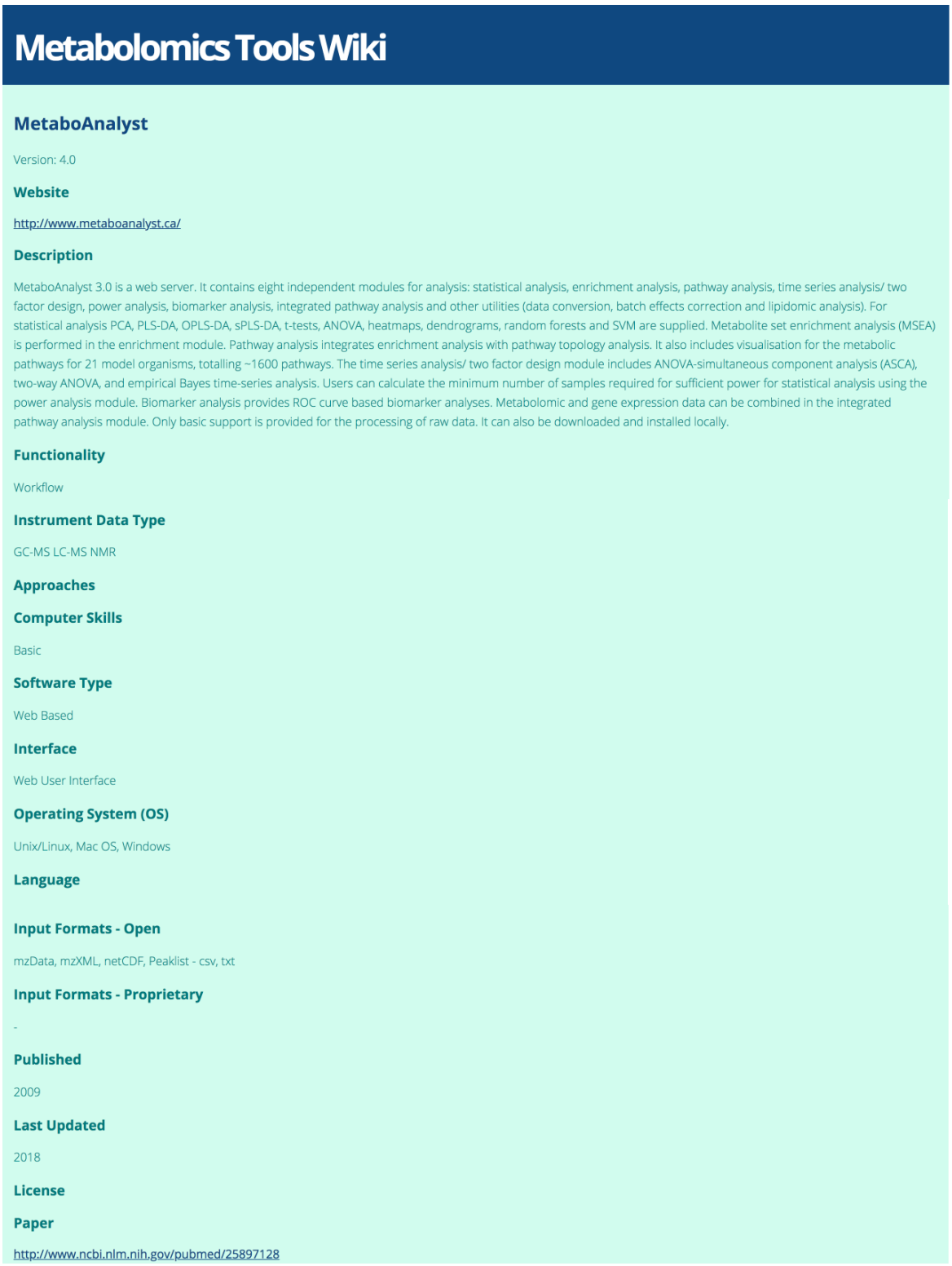


Fig. 4.5 Screenshot showing an example of an individual tool page (for MetaboAnalyst) on the Metabolomics Tools Wiki Main Page taken on 13th March 2018.

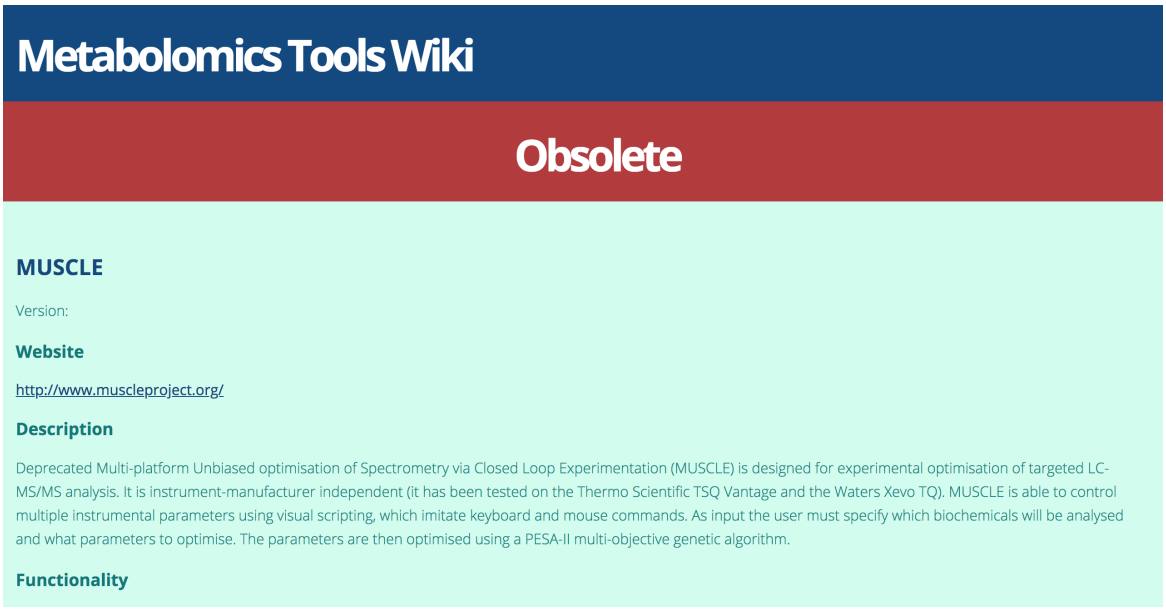


Fig. 4.6 Screenshot illustrating an example of an obsolete metabolomics software tool, MUSCLE, on the Metabolomics Tools Wiki taken on 13th March 2018.

4.4 Additional Use of the Metabolomics Tools Taxonomy

PhenoMeNal (Phenome and Metabolome aNalysis)[11], is a H2020 funded e-infrastructure that provides data analysis workflows for clinical metabolomics via a Virtual Research Environment (VRE): the "PhenoMeNal Cloud Research Environment" (CRE). The main workflow system used for PhenoMeNal is Galaxy [37], which allows for the integration of multiple software tools into complete analytical workflows.

All of the tools that are currently available in the Galaxy workflows via the CRE are displayed in the PhenoMeNal App Library (https://portal.phenomenal-h2020.eu/app-library, accessed 10th March 2018). The App Library currently contains 61 tools, as of 10th March 2018 (Figure 4.7). The metabolomics tools taxonomy is used to categorised tools in the PhenoMeNal App Library. The taxonomy was implemented as an AngularJS tree object by PhenoMeNal developers.

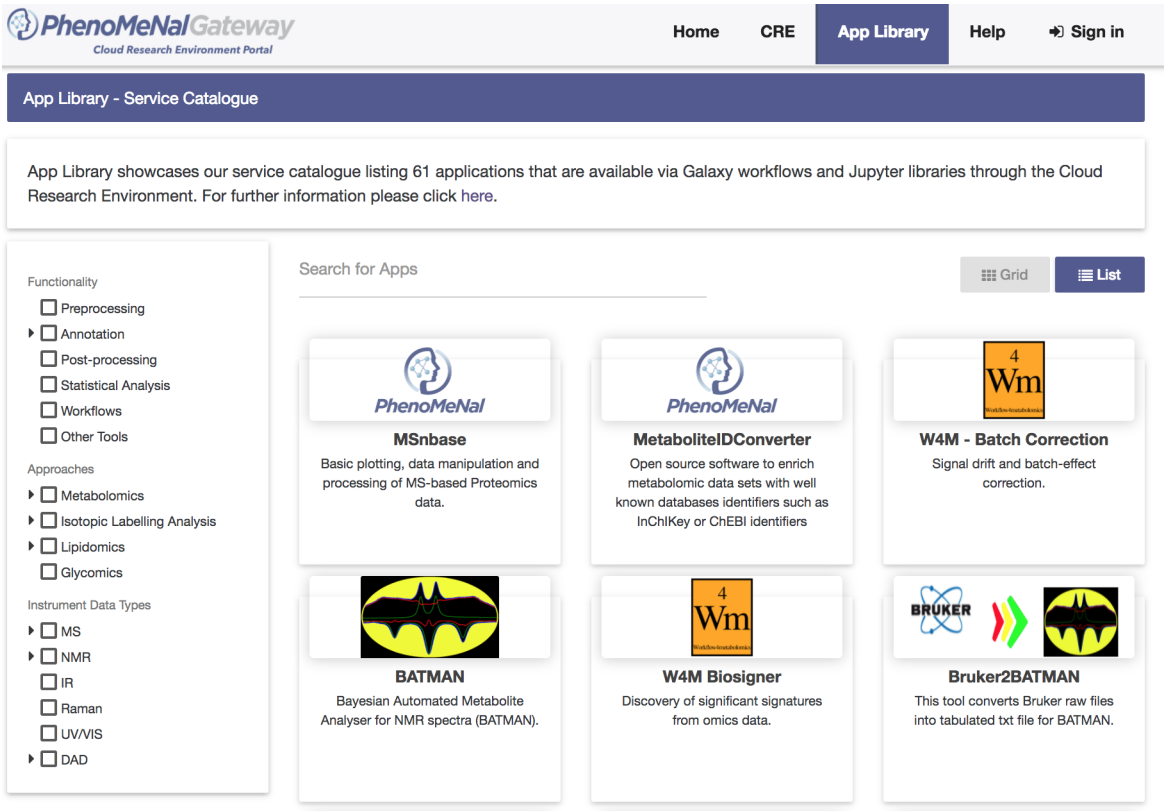


Fig. 4.7 Screenshot of the PhenoMeNal App Library taken on the 13th March 2018, showing that the taxonomy is used for filtering Apps.

[11] https://phenomenal-h2020.eu/home/, accessed 10th March 2018

4.5 Discussion

Scientific software is critical for research: 91% of researchers rate software as important for their own research and on average spend 40% of their time using software [341]. In the metabolomics community, the majority of practitioners work in both wet and dry labs, collecting data and then analysing it [28]. Only half have dedicated bioinformatics support [28].

The discoverability of scientific software has been highlighted as a metric of good software development [342]. Software should be easy to find by its functionality, without its exact name being known. In metabolomics finding software can be challenging, as depending on the search terms used, the majority of results are often for expensive commercial software, which can be cost prohibitive to many researchers. The metabolomics tools taxonomy produced in this work aims to make it easier for 1) researchers to find suitable software tools for their needs and 2) for tool developers to classify their tools functionality. To date the taxonomy has been used for categorising tools on the Metabolomics Tools Wiki and the PhenoMeNal App library.

This research aimed to create a platform describing software for metabolomics data analysis. The Metabolomics Tools Wiki also includes comprehensive descriptions of all included software, and all software is categorised using the metabolomics tools taxonomy. Anyone can manually add new tools to the Metabolomics Tools Wiki to inform the community about them.

Currently in metabolomics there is a major problem of a lack of interoperability between tools for the different steps of data analysis: the output of one tool is not always an acceptable input format for other tools for the subsequent stages of analysis. There can also be incompatibility between dependencies and required software versions. By reporting accepted input formats the Wiki aims to help address the lack of compatibility between tools.

However, a major limitation of the Metabolomics Tools Wiki is that I shall be unable to maintain it long term. Whilst the wiki is far more up to date than the review, "Navigating freely-available software tools for metabolomics analysis" [29], it is still missing many tools, as new tools are constantly being published and the descriptions of some existing tools may become out of date as tools are updated. It has been suggested that a more sustainable long term solution would be to transfer the data from the wiki to bio.tools (https://bio.tools/, accessed 7th May 2018), an Elixir run open-source registry of biological software tools. As of 7th May 2018, bio.tools contains 92 metabolomics tools, however ~10% of these are databases. Nevertheless, a problem with this solution is that bio.tools does not provide the same level of

granularity of tools classification as the Metabolomics Tools Wiki, with metabolomics tools being labelled only as "Metabolomics" and "Command-line tool", "Desktop application", "Web application" and/or "Library".

An inherent disadvantage of the review paper format is that it can provide only a static snapshot as to the state of the field at the time it was composed. Capped page or word counts also limit the amount of information that can be conveyed in many review papers. Our review, Spicer *et al.* [29] included only the most widely used tools (those with ≥ 50 citations on Web of Science (as of 08/09/16) or the use of the tool being reported in the recent Metabolomics Society survey [28]). If another survey were to be conducted today the tools that researchers use may have changed, and more tools would be highly cited e.g. GNPS [128] was cited 13 times on Web of Science in 2016, but has been cited 135 times as of 27$^{\text{th}}$ March 2018.

Popularity is also not necessarily a good proxy for quality [343] and many novel bioinformatics algorithms are not directly compared to other existing algorithms [344]. A high number of citations also do not mean that a resource is necessarily discoverable [345]. Older software is likely to have acquired a higher number of citations, however it may be out of date[12] and no longer maintained. In the Metabolomics Society survey that was conducted in 2016 [28], researchers still reported using MSFACTs for GC-MS data analysis, a software tool that has not been updated since 2003 (and which is now no longer available to download).

Our review also suffers from the limitation of not providing a direct comparison of tools, in order to produce a ranked list of the "best" tools to perform a specific data analysis task e.g. peak picking. Whilst there have been some reviews comparing the accuracy of peak picking of a number of software for LC-MS and GC-MS, these reviews have mostly focused on commercial software [242], have not optimised software parameters [346] or have not used MS/MS data [347]. Recently there have been more studies comparing software published [243, 348] but these have compared only 2 and 4 tools, respectively. There is also the Metabolomics Research Group data analysis study[13], where any researchers were invited to identify metabolites from the same set of mice LC-MS samples, however, at the time of writing this thesis, the full results of this research have not been published. It would therefore be beneficial for systematic reviews to be conducted, which compare larger numbers (5-10) of freely available tools designed for specific tasks in metabolomics data analysis, using benchmarked datasets containing only known metabolites, such as MTBLS59[14] [349] and

[12] https://www.the-scientist.com/?articles.view/articleNo/51260/title/Scientists-Continue-to-Use-Outdated-Methods/, accessed 10$^{\text{th}}$ May 2018.

[13] https://abrf.org/research-group/metabolomics-research-group-mrg accessed 19$^{\text{th}}$ December 2018.

[14] https://www.ebi.ac.uk/metabolights/MTBLS59, accessed 10$^{\text{th}}$ May 2018.

MTBLS79[15] [350]. This is especially important for NMR-based metabolomics where no such review has yet been conducted.

The Critical Assessment of Small Molecule Identification (CASMI) [138] competition provides an alternative means of directly comparing metabolomics software. Teams compete in a series of challenges with the aim of correctly identifying as many small molecules as possible. To date there have been five competitions, running in 2012, 2013, 2014, 2016 and 2017[16]. In 2017 the categories were "Best Structural Identification on Natural Products", "Best Automatic Structural Identification — *In Silico* Fragmentation Only", "Best Automatic Structural Identification — Full Information" and "Best Automatic Candidate Ranking". The Best Automatic Structural Identification and Best Automatic Candidate Ranking categories directly compare tools for small molecule identification. In previous years winners have included MS-FINDER [351], IOKR [352], CSI:FingerID [353] and CFM-ID [354]. Additional CASMI style competitions could directly compare software designed specifically for other stages of data processing, e.g. peak picking, to find which performed optimally.

As well as comparing data processing and analysis methods, it is also important to examine the reproducibility of measurements between laboratories. If different instruments detect metabolites at dissimilar levels this must be accounted for during data processing. Over the last several years a number of ring trials have been conducted to compare the measurement of metabolites between laboratories [355–358]. These studies have found varying levels of replicability and have highlighted metabolites that are majors sources of irreproducibility. Being aware of which metabolites are particularly challenging to measure accurately will allow researchers to correct for their levels during data processing.

The FAIR principles [112] were originally written for data and not software, however it is also important to consider the findability, accessibility, interoperability, and usability of software. The expanded FAIR-TLC [117] principle of licensing is also highly relevant. If software are not discoverable they will not be used, and a lack of licensing will prevent their use. As with metadata reporting and data sharing (see chapters 2 and 3, respectively) it is also highly important that there is standardisation of data formats, so that the inputs and outputs of software are interoperable.

The primary means of disseminating scientific knowledge is via journal articles, which is how the majority of scientific software are reported. However, as journal articles are static

[15]https://www.ebi.ac.uk/metabolights/MTBLS79, accessed 10th May 2018.

[16]http://www.casmi-contest.org/, accessed 10th May 2018.

objects, if the website address of a software tool is updated this cannot be easily changed in the journal article, and may prevent the tool from being findable.

A final point of consideration is that mass spectrometry software developers overrate the value of their tools, whilst users still feel many of their needs that are unfulfilled [359]. This can lead to a glut of software with very similar functionalities, but nothing that does precisely what the user requires, leading the user to either manually analysing data or spending significant amounts of time attempting to use general-purpose software. A goal of this research is to increase the visibility of niche software, and hopefully begin to address this problem by highlighting what software tools are already available.

4.6 Conclusion

There are now >250 software tools specific for metabolomics. The sheer number can make it a daunting task for practitioners to select tools for data analysis. I have therefore produced a taxonomy of metabolomics software tools, in order to formally categorize them. I have also developed a GitHub Pages wiki — https://raspicer.github.io/MetabolomicsTools/, that aims to provide extensive details about all included software, such as instrumental technology data designed for, operating system, accepted data input formats, programming language written in, dependencies and dates of publication and most recent update.

Chapter 5

Conclusions

5.1 Summary

Metabolomics has been defined as "the study of the entire set of small molecules present within an organism, organ, biological tissue or cell" [1]. No single analytical technique is able to measure the entire metabolome, due the diversity of structures and polarities of metabolites, and instead specific fractions are measured using complementary techniques such as MS and NMR.

There is hope that metabolomics can be applied to measure exposures, to toxicology and precision medicine. However, for metabolomics to be utilised for these applications, rigorous meta-analyses must be performed to ensure that measurements are reliable and accurate across studies. For meta-analyses to be performed sufficiently annotated data must be available. Even when datasets cannot be made publicly available (e.g. clinical data), it is important that data is FAIR, to facilitate its reuse.

Reporting standards can be an important component in making data reusable by ensuring metadata are reported in a consistent manner across studies; however poorly worded reporting standards can hinder reuse. The Metabolomics Standards Initiative (MSI) guidelines were first published a decade ago and included reporting standards for all stages of metabolomics analysis: experimental design, biological context, chemical analysis and data processing. Over the subsequent decade a series of public metabolomics repositories have arisen, to which the MSI guidelines can be applied.

In chapter 2 the compliance of 483 public datasets, from five dedicated metabolomics repositories, with the MSI biological context metadata reporting standards, were investigated. None of the reporting standards were fully complied with in every publicly available study, although compliance rates varied greatly, from 0 to 97%. The *in vitro* reporting standards

were the least complied with and the plant were the most. The findings of this study, along with complementary research by Considine *et al.* [208], indicate that the MSI reporting standards are no longer fit for purpose and should be revisited and revised. For this the community, data curators, publishers and funders should all be consulted. The open data that are already available in repositories can also serve to guide the selection of updated or new minimal reporting standards.

Compared to the other omics, metabolomics is lagging behind in regards to data sharing. The value of open data has been clearly demonstrated in transcriptomics and proteomics, and there are many examples of data reuse. In chapter 3 data sharing within the metabolomics community and the reuse of open metabolomics datasets were examined. Whilst journal policies exist that require data sharing in metabolomics, these are not necessarily strongly enforced, and the majority of publications do not have their underlying data publicly shared. However, the metabolomics data that is shared openly is being increasingly reused, indicating its importance.

The PLOS publishing group has been a pioneer of mandating open data. Since March 2014, all articles have been required to share data and include data availability statements. PLOS ONE is also one of the largest publishers of metabolomics research. It was therefore surprising to find only 12 publicly available datasets in metabolomics repositories, published since 2015, directly linked to PLOS ONE articles, when ~400 metabolomics papers had been published in the same time period. Potential reasons for this discrepancy were therefore investigated.

It was found that only a minority of metabolomics studies openly shared raw data (8.08%), and >50 shared data only as figures. Whilst 35.4% of articles included human data, potentially preventing data sharing due to privacy concerns, only a single paper stated that data could not be made available due to ethical concerns, with an additional 8.1% of studies stating data are available on request. This suggests that ethical concerns are not the primary reason for a lack of open data sharing. Instead the findings suggest that sharing only processed data is the standard within the metabolomics community, meaning that raw data sharing is not required in order to comply with PLOS ONE's data sharing guidelines. There also appears to be a lack of understanding as to what actually constitutes raw data.

More than 250 software tools that are specific for metabolomics data now exist. The sheer number can make it a daunting task for practitioners to select tools for data analysis, and it can also be difficult to find appropriate tools for more niche tasks. Chapter 4 details a taxonomy that has been developed in order to formally categorize metabolomics software tools, along with a GitHub Pages wiki. The wiki aims to provide extensive details about

all included software, such as instrumental technology data designed for, operating system, accepted data input formats, programming language written in, dependencies and dates of publication and most recent update.

5.2 Conclusions

Currently, the data reporting practices within the metabolomics community are insufficient to enable reuse of the bulk of studies. The majority of studies do not openly share data: if the estimate of the total number of metabolomics studies in chapter 3 (17,614) were correct, then only 2% of metabolomics studies have deposited their data in a dedicated metabolomics repository. Whilst this estimate of the total number of metabolomics papers is inaccurate, sharing data in a dedicated metabolomics repository appears to be the most popular method of openly sharing metabolomics data. The number of metabolomics studies with open data is therefore far lower than for other omics.

In more mature communities such as genomics, data sharing has become the standard, despite greater concerns about participant re-identification than for metabolomics. The research in this thesis suggests that, at least in PLOS ONE, the main reason for not sharing raw metabolomics data is not ethical concerns. Instead, it appears that the accepted community standard is to share only processed data in tables or figures. This is concerning given the disparity of outputs between different preprocessing software. The importance of open raw data must accordingly be highlighted to the field. At the very minimum, all studies should report which metabolites were identified or annotated within the study.

Where sensitive metabolomics datasets cannot be openly shared, the data can still be made FAIR by being deposited in controlled access repositories such as European Genome-phenome Archive (EGA) and Database of Genotypes and Phenotypes (dbGaP). These repositories allow researchers to identify studies of interest by searching for diseases, variants, technology and sample type (metadata), and to then apply for access to the study.

There are also challenges relating to the description of datasets, as a lack of appropriate metadata can prevent studies from being reusable, or greatly increase the time taken to do so. Openly sharing data can be a very time consuming process, and doing so with sufficient metadata takes even longer. However, open data is only valuable if it is reusable. As funding bodies are increasingly mandating open data, grants should include exclusively allocated time to enable researchers to prepare their data for sharing.

To date, data citations have not been widely adopted by the metabolomics community. The majority of studies that reused >100 metabolomics datasets did not cite each individual

dataset. It may be unrealistic to include such a large number of citations in the body of a journal article, however data citations can instead be included as supplementary material. Data are valuable research outputs in their own right, and should be cited as such.

Understanding of the metabolome and the factors that affect it has greatly increased since the MSI standards were first published a decade ago. Reporting standards for metabolomics must be updated to reflect this. There is currently ongoing work by the Metabolite Identification task group of the Metabolomics Society to update the metabolite identification reporting standards, new data exchange formats have been released and MERIT has been launched. This progress is promising, however reporting standards for other stages of metabolomics analysis must also be updated. At minimum, a sufficient amount of metadata to enable replication of an experiment should be reported. Ideally consideration would also be given to the minimum amount of metadata required to be able to reuse data to the fullest potential. Ontologies or MeSH should be used to ensure metadata is reported consistently across studies.

Scientific software is critical to metabolomics research; it is highly impractical to process and analyze complex metabolomics data without it. However, currently many metabolomics software tools are not FAIR. Researchers struggle to find the correct tool for their needs, and instead end up wasting time either manually analysing data or attempting to use general purpose software, without the functionally to meet their specific needs. There is also a lack of interoperability between software tools, with incompatible input and output formats, and dependencies. Additionally, not all freely available metabolomics software is correctly licensed, preventing its reuse — if software does not have a license then it cannot be reused. There must be efforts by metabolomics software developers to make their tools more FAIR. It is also important the newly published algorithms and software are directly compared to established ones, to enable researchers to better decide which tool is right for them. This could be done by reviews, or CASMI style competitions to directly compare software for different stages of processing e.g. peak picking.

5.3 Recommended Future Research

A commonly cited reason for not openly sharing data is fear of not being able to generate as many publications from it, and being scooped. This fear does not appear to be justified for metabolomics, where nearly 50% of publications that reused metabolomics data shared at least one author with the original study. However, to date, the effect of open data on citation rates in metabolomics has not been investigated. This would be an interesting area

of future research as open data has been found to increase citation rates of gene expression microarray journal articles [88]. An increase of citation rate would entice more metabolomics researchers to make their data public.

Over recent years there has been much concern about the reproducibility of scientific research. In a systematic review Considine *et al.* [208] found that only a single metabolomics biomarker discovery study had the potential to be reproducible (although data and code were not available so this was not tested). It would therefore be interesting to investigate whether it is possible to repeat data analysis for metabolomics research, as has been done in other fields [224, 360, 361]. Whilst data must be either requested or openly available for such an analysis, there are now >1,500 publicly available metabolomics datasets, which should be sufficient to enable this investigation. The findings of such research could be important for updating the MSI minimal data analysis reporting standards.

5.4 Closing Remarks

Metabolomics has the potential to be utilised in many areas, and to become a valuable clinical tool. However for this to come to fruition, a cultural shift within the metabolomics community must occur and data sharing must become routine. I hope that research conducted for this thesis can help to guide journal data sharing policies and future revisions to the MSI guidelines, such as the ongoing work by the Metabolite Identification task group.

References

The bibliography follows the citation style of the journal Nature.

1. Fiehn, O. Metabolomics - The link between genotypes and phenotypes. *Plant Molecular Biology.* **48(1-2),** 155–171. doi:10.1023/A:1013713905833 (2002).
2. Wishart, D. S. Current progress in computational metabolomics. *Briefings in Bioinformatics.* **8(5),** 279–293. doi:10.1093/bib/bbm030 (2007).
3. Jordan, K. W. *et al.* Metabolomic Characterization of Human Rectal Adenocarcinoma With Intact Tissue Magnetic Resonance Spectroscopy. *Diseases of the Colon & Rectum.* **52(3),** 520–525. doi:10.1007/DCR.0b013e31819c9a2c (2009).
4. Yu, Z. *et al.* Human serum metabolic profiles are age dependent. *Aging Cell.* **11(6),** 960–967. doi:10.1111/j.1474-9726.2012.00865.x (2012).
5. Jaremek, M. *et al.* Alcohol-induced metabolomic differences in humans. *Translational Psychiatry.* **3,** e276. doi:10.1038/tp.2013.55 (2013).
6. Menni, C. *et al.* Metabolomic Identification of a Novel Pathway of Blood Pressure Regulation Involving Hexadecanedioate. *Hypertension.* **66(2),** 422–429. doi:10.1161/hypertensionaha.115.05544 (2015).
7. Jourdan, C. *et al.* Body fat free mass is associated with the serum metabolite profile in a population-based study. *PLOS ONE.* **7(6),** e40009. doi:10.1371/journal.pone.0040009 (2012).
8. Moore, S. C. *et al.* Human metabolic correlates of body mass index. *Metabolomics.* **10(2),** 259–269. doi:10.1007/s11306-013-0574-1 (2014).
9. Wahl, S. *et al.* Multi-omic signature of body weight change: results from a population-based cohort study. *BMC Medicine.* **13,** 48. doi:10.1186/s12916-015-0282-y (2015).
10. Lawton, K. A. *et al.* Analysis of the adult human plasma metabolome. *Pharmacogenomics.* **9(4),** 383–397. doi:10.2217/14622416.9.4.383 (2008).
11. Krumsiek, J. *et al.* Gender-specific pathway differences in the human serum metabolome. *Metabolomics.* **11(6),** 1815–1833. doi:10.1007/s11306-015-0829-0 (2015).
12. Altmaier, E. *et al.* Questionnaire-based self-reported nutrition habits associate with serum metabolism as revealed by quantitative targeted metabolomics. *European Journal of Epidemiology.* **26(2),** 145–156. doi:10.1007/s10654-010-9524-7 (2011).
13. Menni, C. *et al.* Targeted metabolomics profiles are strongly correlated with nutritional patterns in women. *Metabolomics.* **9(2),** 506–514. doi:10.1007/s11306-012-0469-6 (2013).

14. Wang-Sattler, R. *et al.* Metabolic profiling reveals distinct variations linked to nicotine consumption in humans - First results from the KORA study. *PLOS ONE.* **3(12),** e3863. doi:10.1371/journal.pone.0003863 (2008).

15. Xu, T. *et al.* Effects of smoking and smoking cessation on human serum metabolite profile: results from the KORA cohort study. *BMC Medicine.* **11,** 60. doi:10.1186/1741-7015-11-60 (2013).

16. Everett, J. R. Pharmacometabonomics in humans: a new tool for personalized medicine. *Pharmacogenomics.* **16(7),** 737–754. doi:10.2217/pgs.15.20 (2015).

17. Lankadurai, B. P., Nagato, E. G. & Simpson, M. J. Environmental Metabolomics: an emerging approach to study organisms responses to environmental stressors. *Environmental Reviews.* **21(3),** 180–205. doi:10.1139/er-2013-0011 (2013).

18. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: The apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology.* **13(4),** 263–269. doi:10.1038/nrm3314 (2012).

19. Gieger, C. *et al.* Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum. *PLOS Genetics.* **4(11),** e1000282. doi:10.1371/journal.pgen.1000282 (2008).

20. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature.* **477(7362),** 54–60. doi:10.1038/nature10354 (2011).

21. Koek, M. M., Jellema, R. H., van der Greef, J., Tas, A. C. & Hankemeier, T. Quantitative metabolomics based on gas chromatography mass spectrometry: Status and perspectives. *Metabolomics.* **7(3),** 307–328. doi:10.1007/s11306-010-0254-3 (2011).

22. León, Z., García-Cañaveras, J. C., Donato, M. T. & Lahoz, A. Mammalian cell metabolomics: Experimental design and sample preparation. *Electrophoresis.* **34(19),** 2762–2775. doi:10.1002/elps.201200605 (2013).

23. Inglese, P. *et al.* Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer. *Chemical Science.* **8(5),** 3500–3511. doi:10.1039/C6SC03738K (2017).

24. Hansen, R. L. & Lee, Y. J. High-Spatial Resolution Mass Spectrometry Imaging: Toward Single Cell Metabolomics in Plant Tissues. *The Chemical Record.* **18(1),** 65–77. doi:10.1002/tcr.201700027 (2018).

25. Emwas, A.-H. M. The Strengths and Weaknesses of NMR Spectroscopy and Mass Spectrometry with Particular Focus on Metabolomics Research. *Methods in Molecular Biology.* **1277,** 161–193. doi:10.1007/978-1-4939-2377-9_13 (2015).

26. Bharti, S. K. & Roy, R. Quantitative H NMR spectroscopy. *Trends in Analytical Chemistry.* **35,** 5–26. doi:10.1016/j.trac.2012.02.007 (2012).

27. Ellinger, J. J., Chylla, R. A., Ulrich, E. L. & Markley, J. L. Databases and Software for NMR-Based Metabolomics. *Current Metabolomics.* **1(1),** doi:10.2174/2213235X11301010028 (2013).

28. Weber, R. J. *et al.* Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy. *Metabolomics.* **13,** 12. doi:10.1007/s11306-016-1147-x (2017).

29. Spicer, R., Salek, R. M., Moreno, P., Cañueto, D. & Steinbeck, C. Navigating freely-available software tools for metabolomics analysis. *Metabolomics.* **13(9),** 106. doi:10.1007/s11306-017-1242-7 (2017).

30. Misra, B. B., Fahrmann, J. F. & Grapov, D. Review of emerging metabolomic tools and resources: 2015–2016. *Electrophoresis.* **38(18),** 2257–2274. doi:10.1002/elps.201700110 (2017).

31. Misra, B. B. New tools and resources in metabolomics: 2016-2017. *Electrophoresis.* **39(7),** 909–923. doi:10.1002/elps.201700441 (2018).

32. Zhou, B., Xiao, J. F., Tuli, L. & Ressom, H. W. LC-MS-based metabolomics. *Molecular BioSystems.* **8(2),** 470–481. doi:10.1007/978-1-61737-985-7_13 (2012).

33. Garcia, A. & Barbas, C. in (ed Metz, T. O.) 191–204 (Humana Press, Totowa, NJ, 2011). doi:10.1007/978-1-61737-985-7_11.

34. Kind, T. *et al.* FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Analytical Chemistry.* **81(24),** 10038–10048. doi:10.1021/ac9019522 (2009).

35. Kopka, J. *et al.* GMD@CSB.DB: The Golm metabolome database. *Bioinformatics.* **21(8),** 1635–1638. doi:10.1093/bioinformatics/bti236 (2005).

36. Lei, Z., Huhman, D. V. & Sumner, L. W. Mass spectrometry strategies in metabolomics. *Journal of Biological Chemistry.* **286(29),** 25435–25442. doi:10.1074/jbc.R111.238691 (2011).

37. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research.* **44(W1),** W3–W10. doi:10.1093/nar/gkw343 (2016).

38. Kell, D. B. & Oliver, S. G. The metabolome 18 years on: a concept comes of age. *Metabolomics.* **12(9),** 148. doi:10.1007/s11306-016-1108-4 (2016).

39. Cubero-Leon, E., De Rudder, O. & Maquet, A. Metabolomics for organic food authentication: Results from a long-term field study in carrots. *Food Chemistry.* **239(555),** 760–770. doi:10.1016/j.foodchem.2017.06.161 (2018).

40. Jandrić, Z. *et al.* Discrimination of honey of different floral origins by a combination of various chemical parameters. *Food Chemistry.* **189,** 52–59. doi:10.1016/j.foodchem.2014.11.165 (2015).

41. Kew, W., Goodall, I., Clarke, D. & Uhrín, D. Chemical Diversity and Complexity of Scotch Whisky as Revealed by High-Resolution Mass Spectrometry. *Journal of the American Society for Mass Spectrometry.* **28(1),** 200–213. doi:10.1007/s13361-016-1513-y (2017).

42. Schmidtke, L. M., Blackman, J. W., Clark, A. C. & Grant-Preece, P. Wine metabolomics: Objective measures of sensory properties of semillon from GC-MS profiles. *Journal of Agricultural and Food Chemistry.* **61(49),** 11957–11967. doi:10.1021/jf403504p (2013).

43. Roullier-Gall, C., Witting, M., Gougeon, R. D. & Schmitt-Kopplin, P. High precision mass measurements for wine metabolomics. *Frontiers in Chemistry.* **2,** 102. doi:10.3389/fchem.2014.00102 (2014).

44. Ramirez, T. *et al.* Metabolomics in toxicology and preclinical research. *ALTEX.* **30(2),** 209–225. doi:10.14573/altex.2013.2.209 (2013).

45. Priori, R. *et al.* 1H-NMR-Based Metabolomic Study for Identifying Serum Profiles Associated with the Response to Rtanercept in Patients with Rheumatoid Arthritis. *PLOS ONE.* **10(11),** e0138537. doi:10.1371/journal.pone.0138537 (2015).

46. Surowiec, I., Ärlestig, L., Rantapää-Dahlqvist, S. & Trygg, J. Metabolite and Lipid Profiling of Biobank Plasma Samples Collected Prior to Onset of Rheumatoid Arthritis. *PLOS ONE.* **11(10),** e0164196. doi:10.1371/journal.pone.0164196 (2016).

47. Ritchie, S. a. *et al.* Reduced levels of hydroxylated, polyunsaturated ultra long-chain fatty acids in the serum of colorectal cancer patients: implications for early screening and detection. *BMC Medicine.* **8,** 13. doi:10.1186/1741-7015-8-13 (2010).

48. Wang-Sattler, R. *et al.* Novel biomarkers for pre-diabetes identified by metabolomics. *Molecular Systems Biology.* **8,** 615. doi:10.1038/msb.2012.43 (2012).

49. Menni, C. *et al.* Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach. *Diabetes.* **62(12),** 4270–4276. doi:10.2337/db13-0570 (2013).

50. Armstrong, C. W., McGregor, N. R., Lewis, D. P., Butt, H. L. & Gooley, P. R. Metabolic profiling reveals anomalous energy metabolism and oxidative stress pathways in chronic fatigue syndrome patients. *Metabolomics.* **11(6),** 1626–1639. doi:10.1007/s11306-015-0816-5 (2015).

51. Naviaux, R. K. *et al.* Metabolic features of chronic fatigue syndrome. *Proceedings of the National Academy of Sciences of the United States of America.* **113(37),** E5472–80. doi:10.1073/pnas.1607571113 (2016).

52. Assfalg, M. *et al.* Evidence of different metabolic phenotypes in humans. *Proceedings of the National Academy of Sciences of the United States of America.* **105(5),** 1420–1424. doi:10.1073/pnas.0705685105 (2008).

53. Bernini, P. *et al.* Individual human phenotypes in metabolic space and time. *Journal of Proteome Research.* **8(9),** 4264–4271. doi:10.1021/pr900344m (2009).

54. Yousri, N. A. *et al.* Long term conservation of human metabolic phenotypes and link to heritability. *Metabolomics.* **10(5),** 1005–1017. doi:10.1007/s11306-014-0629-y (2014).

55. Rattray, N. J. W. *et al.* Beyond genomics: understanding exposotypes through metabolomics. *Human Genomics.* **12(1),** 4. doi:10.1186/s40246-018-0134-x (2018).

56. Buck Louis, G. M. & Sundaram, R. Exposome: Time for transformative research. *Statistics in Medicine.* **31(22),** 2569–2575. doi:10.1002/sim.5496 (2012).

57. Rappaport, S. M. Genetic factors are not the major causes of chronic diseases. *PLOS ONE.* **11(4),** e0154387. doi:10.1371/journal.pone.0154387 (2016).

58. Theodoratou, E., Timofeeva, M., Li, X., Meng, X. & Ioannidis, J. P. Nature, Nurture, and Cancer Risks: Genetic and Nutritional Contributions to Cancer. *Annual Review of Nutrition.* **37,** 293–320. doi:10.1146/annurev-nutr-071715-051004 (2017).

59. Herder, C. & Roden, M. Genetics of type 2 diabetes: Pathophysiologic and clinical relevance. *European Journal of Clinical Investigation.* **41(6),** 679–692. doi:10.1111/j.1365-2362.2010.02454.x (2011).

60. Wild, C. P. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers and Prevention.* **14(8),** 1847–1850. doi:10.1158/1055-9965.EPI-05-0456 (2005).

61. Rappaport, S. M. & Smith, M. T. Environment and disease risks. *Science.* **330(6003),** 460–461. doi:10.1126/science.1192603 (2010).

62. Bundy, J. G., Davey, M. P. & Viant, M. R. Environmental metabolomics: A critical review and future perspectives. *Metabolomics.* **5,** 3. doi:10.1007/s11306-008-0152-0 (2009).

63. European Food Safety Authority. Modern methodologies and tools for human hazard assessment of chemicals. *EFSA Journal.* **12(4),** 3638. doi:10.2903/j.efsa.2014.3638 (2014).

64. Wang, X. *et al.* Serum metabolome biomarkers associate low-level environmental perfluorinated compound exposure with oxidative /nitrosative stress in humans. *Environmental Pollution.* **229,** 168–176. doi:10.1016/j.envpol.2017.04.086 (2017).

65. Baker, M. G., Simpson, C. D., Lin, Y. S., Shireman, L. M. & Seixas, N. The Use of Metabolomics to Identify Biological Signatures of Manganese Exposure. *Annals of Work Exposures and Health.* **61(4),** 406–415. doi:10.1093/annweh/wxw032 (2017).

66. Pleil, J. D., Stiegel, M. A. & Sobus, J. R. Breath biomarkers in environmental health science: exploring patterns in the human exposome. *Journal of Breath Research.* **5(4),** 046005. doi:10.1088/1752-7155/5/4/046005 (2011).

67. Zierer, J. *et al.* The fecal metabolome as a functional readout of the gut microbiome. *Nature Genetics.* **50(6),** 790–795. doi:10.1038/s41588-018-0135-7 (2018).

68. Russell, W. & Burch, R. *The Principles of Humane Experimental Technique* (Methuen and Co., Ltd., 1959).

69. Loizou, G. D. Animal-free chemical safety assessment. *Frontiers in Pharmacology.* **7,** 218. doi:10.3389/fphar.2016.00218 (2016).

70. Guijas, C., Montenegro-Burke, J. R., Warth, B., Spilker, M. E. & Siuzdak, G. Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nature Biotechnology.* **36(4),** 316–320. doi:10.1038/nbt.4101 (2018).

71. Panopoulos, A. D. *et al.* The metabolome of induced pluripotent stem cells reveals metabolic changes occurring in somatic cell reprogramming. *Cell Research.* **22(1),** 168–177. doi:10.1038/cr.2011.177 (2012).

72. Gil-de-Gomez, L. *et al.* A Phosphatidylinositol Species Acutely Generated by Activated Macrophages Regulates Innate Immune Responses. *The Journal of Immunology.* **190(10),** 5169–5177. doi:10.4049/jimmunol.1203494 (2013).

73. Beyer, B. A. *et al.* Metabolomics-based discovery of a metabolite that enhances oligodendrocyte maturation. *Nature Chemical Biology.* **14(1),** 22–28. doi:10.1038/nchembio.2517 (2018).

74. Morris, C. W. *Foundations of the theory of signs* (The University of Chicago Press, Chicago, Ill., 1938).

75. Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I. & Matsushima, K. Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. *Technological Forecasting and Social Change.* **78(2),** 274–282. doi:10.1016/j.techfore.2010.07.006 (2011).

76. Dickersin, K. The existence of publication bias and risk factors for its occurrence. *JAMA.* **263(10),** 1385–1389 (1990).

77. Bucci, E. M. Automatic detection of image manipulations in the biomedical literature. *Cell Death & Disease.* **9(3),** 400. doi:10.1038/s41419-018-0430-3 (2018).

78. Ioannidis, J. P., Fanelli, D., Dunne, D. D. & Goodman, S. N. Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLOS Biology.* **13(10),** e1002264. doi:10.1371/journal.pbio.1002264 (2015).

79. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature.* **483(7391),** 531–533. doi:10.1038/483531a (2012).

80. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science.* **349(6251),** aac4716. doi:10.1126/science.aac4716 (2015).

81. Collberg, B. Y. C. & Proebsting, T. A. Repeatability in Computer Systems Research. *Communications of the ACM.* **59(3),** 62–69. doi:10.1145/2812803 (2016).

82. Reality check on reproducibility. *Nature.* **533(7604),** 437. doi:10.1038/533437a (2016).

83. Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M. & Zwelling, L. A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic. *PLOS ONE.* **8(5),** e63221. doi:10.1371/journal.pone.0063221 (2013).

84. Phillips, N. Tool spots DNA errors in paper. *Nature.* **551(7681),** 422–423. doi:10.1038/nature.2017.23003 (2017).

85. Ioannidis, J. P. A. Why most published research findings are false. *PLOS Medicine.* **2(8),** e124. doi:10.1371/journal.pmed.0020124 (2005).

86. Vicente-Saez, R. & Martinez-Fuentes, C. Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research.* **88,** 428–436. doi:10.1016/j.jbusres.2017.12.043 (2018).

87. Molloy, J. C. The Open Knowledge Foundation: Open Data Means Better Science. *PLOS Biology.* **9(12),** e1001195. doi:10.1371/journal.pbio.1001195 (2011).

88. Piwowar, H. A. & Vision, T. J. Data reuse and the open data citation advantage. *PeerJ.* **1,** e175. doi:10.7717/peerj.175 (2013).

89. Rowhani-Farid, A., Allen, M. & Barnett, A. G. What incentives increase data sharing in health and medical research? A systematic review. *Research Integrity and Peer Review.* **2,** 4. doi:10.1186/s41073-017-0028-9 (2017).

90. Piwowar, H. A., Vision, T. J. & Whitlock, M. C. Data archiving is a good investment. *Nature.* **473(7347),** 285. doi:10.1038/473285a (2011).

91. Digitial Science *et al. The State of Open Data Report 2017* tech. rep. (2017). doi:10.6084/m9.figshare.5481187.v1.

92. Sills, J. Editorial expression of concern. *Science.* **354(6317),** 1242. doi:10.1126/science.aah6990 (2016).

93. Vines, T. H. *et al.* The availability of research data declines rapidly with article age. *Current Biology.* **24(1),** 94–97. doi:10.1016/j.cub.2013.11.014 (2014).

94. *UK gross domestic expenditure on research and development: 2015* tech. rep. (2017).

95. Darch, P. T. & Knox, E. J. Ethical perspectives on data and software sharing in the sciences: A research agenda. *Library & Information Science Research.* **39(4),** 295–302. doi:10.1016/j.lisr.2017.11.008 (2017).

96. Contreras, J. L. Bermuda's Legacy: Policy, Patents, and the Design of the Genome Commons. *Minnesota Journal of Law, Science & Technology.* **12,** 61–125 (2011).

97. Vasilevsky, N. A., Minnier, J., Haendel, M. A. & Champieux, R. E. Reproducible and reusable research: are journal data sharing policies meeting the mark? *PeerJ.* **5,** e3208. doi:10.7717/peerj.3208 (2017).

98. European Commission. *H2020 Programme Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020* tech. rep. (2016).

99. Guedj, D. & Ramjoué, C. European Commission Policy on Open-Access to Scientific Publications and Research Data in Horizon 2020. *Biomedical Data Journal.* **1(1),** 11–14. doi:10.11610/bmdj.01102 (2015).

100. National Institutes of Health. *Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research* tech. rep. February (2015).

101. *Concordat On Open Research Data* tech. rep. (2016), 1–23.

102. Van Panhuis, W. G. *et al.* A systematic review of barriers to data sharing in public health. *BMC Public Health.* **14,** 1144. doi:10.1186/1471-2458-14-1144 (2014).

103. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science.* **339(6117),** 321–324. doi:10.1126/science.1229566 (2013).

104. Lindenmayer, B. D. & Scheele, B. Do not publish. *Science.* **356(6340),** 800–801. doi:10.1126/science.aan1362 (2017).

105. Federer, L. M., Lu, Y.-L., Joubert, D. J., Welsh, J. & Brandys, B. Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. *PLOS ONE.* **10(6),** e0129506. doi:10.1371/journal.pone.0129506 (2015).

106. Wicherts, J. M., Bakker, M. & Molenaar, D. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS ONE.* **6(11),** e26828. doi:10.1371/journal.pone.0026828 (2011).

107. Credit where credit is overdue. *Nature Biotechnology.* **27(7),** 579. doi:10.1038/nbt0709-579 (2009).

108. Ross-Hellauer, T., Deppe, A. & Schmidt, B. Survey On Open Peer Review: Attitudes And Experience Amongst Editors, Authors And Reviewers. *PLOS ONE.* **12(12),** e0189311. doi:10.1371/journal.pone.0189311 (2017).

109. Kidwell, M. C. *et al.* Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology.* **14(5),** e1002456. doi:10.1371/journal.pbio.1002456 (2016).

110. Rowhani-Farid, A. & Barnett, A. G. Badges for sharing data and code at Biostatistics: an observational study [version 2; referees: 2 approved]. *F1000Research.* **7,** 90. doi:10.12688/f1000research.13477.1 (2018).

111. Ball, A. & Duke, M. *Data Citation and Linking* tech. rep. (2011).

112. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data.* **3,** 160018. doi:10.1038/sdata.2016.18 (2016).

113. Grootveld, M., Leenarts, E., Jones, S., Hermanns, E. & Fankhauser, E. *OpenAIRE survey about Horizon 2020 template for Data Management Plans 2017* tech. rep. (2018). doi:10.5281/zenodo.1120245.

114. Rocca-Serra, P. *et al.* Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics.* **12,** 14. doi:10.1007/s11306-015-0879-3 (2016).

115. Bard, J. B. L. & Rhee, S. Y. Ontologies in Biology: Design, Applications and Future Challenges. *Nature Reviews Genetics.* **5(3),** 213–222. doi:10.1038/nrg1295 (2004).

116. Haendel, M. *et al.* FAIR-TLC: Metrics To Assess Value Of Biomedical Digital Repositories: Response To RFI NOT-OD-16-133. doi:10.5281/zenodo.203295 (2016).

117. Holub, P. *et al.* Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. *Biopreservation and Biobanking.* **16(2),** 97–105. doi:10.1089/bio.2017.0110 (2018).

118. Hodson, S. *et al.* Turning FAIR data into reality: interim report from the European Commission Expert Group on FAIR data (Version Interim draft). doi:10.5281/zenodo.1285272 (2018).

119. Wilkinson, M. D. *et al.* A design framework and exemplar metrics for FAIRness. *Scientific Data.* **5,** 180118. doi:10.1038/sdata.2018.118 (2018).

120. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology.* **32(3),** 223–226. doi:10.1038/nbt.2839 (2014).

121. Rustici, G. *et al.* ArrayExpress update-trends in database growth and links to data analysis tools. *Nucleic Acids Research.* **41(Database issue),** D987–D990. doi:10.1093/nar/gks1174 (2013).

122. Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical analysis of biology by mass spectrometry. *Nature Reviews Chemistry.* **1(7),** 0054. doi:10.1038/s41570-017-0054 (2017).

123. Ferry-Dumazet, H. *et al.* MeRy-B: a web knowledgebase for the storage , visualization, analysis and annotation of plant NMR metabolomic profiles. *BMC Plant Biology.* **11,** 104. doi:10.1186/1471-2229-11-104 (2011).

124. Carroll, A. J., Badger, M. R. & Harvey Millar, A. The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC Bioinformatics.* **11,** 376. doi:10.1186/1471-2105-11-376 (2010).

125. Carroll, A. J. *et al.* PhenoMeter: A Metabolome Database Search Tool Using Statistical Similarity Matching of Metabolic Phenotypes for High-Confidence Detection of Functional Links. *Frontiers in Bioengineering and Biotechnology.* **3,** 106. doi:10.3389/fbioe.2015.00106 (2015).

126. Haug, K. *et al.* MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research.* **41(Database issue),** 781–786. doi:10.1093/nar/gks1004 (2013).

127. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research.* **44(Database issue),** D463–D470. doi:10.1093/nar/gkv1042 (2016).

128. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology.* **34(8),** 828–837. doi:10.1038/nbt.3597 (2016).

129. Rocca-Serra, P. *et al.* ISA software suite: Supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics.* **26(18),** 2354–2356. doi:10.1093/bioinformatics/btq415 (2010).

130. Sud, M. *et al.* LMSD: LIPID MAPS structure database. *Nucleic Acids Research.* **35(Database issue),** D527–32. doi:10.1093/nar/gkl838 (2007).

131. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research.* **44(D1),** D1214–D1219. doi:10.1093/nar/gkv1031 (2016).

132. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research.* **46(D1),** D608–D617. doi:10.1093/nar/gkx1089 (2017).

133. Ulrich, E. L. *et al.* BioMagResBank. *Nucleic Acids Research.* **36(Database issue),** D402–D408. doi:10.1093/nar/gkm957 (2008).

134. Kim, S. *et al.* PubChem substance and compound databases. *Nucleic Acids Research.* **44(D1),** D1202–D1213. doi:10.1093/nar/gkv951 (2016).

135. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research.* **40(Database issue),** D109–D114. doi:10.1093/nar/gkr988 (2012).

136. Perez-Riverol, Y., Alpi, E., Wang, R., Hermjakob, H. & Vizcaíno, J. A. Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. *Proteomics.* **15(5-6),** 930–950. doi:10.1002/pmic.201400302 (2015).

137. Horai, H. *et al.* MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry.* **45(7),** 703–714. doi:10.1002/jms.1777 (2010).

138. Schymanski, E. L. *et al.* Critical Assessment of Small Molecule Identification 2016: automated methods. *Journal of Cheminformatics.* **9,** 22. doi:10.1186/s13321-017-0207-1 (2017).

139. Ara, T. *et al.* Metabolonote: a wiki-based database for managing hierarchical metadata of metabolome analyses. *Frontiers in Bioengineering and Biotechnology.* **3,** 38. doi:10.3389/fbioe.2015.00038 (2015).

140. Salek, R. M. *et al.* COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. *Metabolomics.* **11(6),** 1587–1597. doi:10.1007/s11306-015-0810-y (2015).

141. Perez-Riverol, Y. *et al.* Discovering and linking public omics data sets using the Omics Discovery Index. *Nature Biotechnology.* **35(5),** 406–409. doi:10.1038/nbt.3790 (2017).

142. Chen, X. *et al.* DataMed – an open source discovery index for finding biomedical datasets. *Journal of the American Medical Informatics Association.* **25(3),** 300–308. doi:10.1093/jamia/ocx121 (2018).

143. Higgins, J. P. T. & Green, S. in *The Cochrane Collaboration* Table 7.7.a: Formulae for combining groups (2011).

144. Haidich, A. B. Meta-analysis in medical research. *Hippokratia.* **14(Suppl 1),** 29–37 (2010).

145. Cohn, L. D. & Becker, B. J. How Meta-Analysis Increases Statistical Power. *Psychological Methods.* **8(3),** 243–253. doi:10.1037/1082-989X.8.3.243 (2003).

146. Liberati, A. *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLOS Medicine.* **6(7),** e1000100. doi:10.1371/journal.pmed.1000100 (2009).

147. Stroup, D. F. *et al.* Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting. *Jama.* **283(15),** 2008–2012. doi:10.1001/jama.283.15.2008 (2000).

148. Shamseer, L. *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ.* **349,** i4086. doi:10.1136/bmj.g7647 (2015).

149. Panagiotou, O. A., Willer, C. J., Hirschhorn, J. N. & Ioannidis, J. P. A. The Power of Meta-Analysis in Genome Wide Association Studies. *Annual Reviews of Genomics and Human Genetics.* **14,** 441–465. doi:10.1146/annurev-genom-091212-153520 (2013).

150. Vinga, S. Global Meta-Analysis of Transcriptomics Studies. *PLOS ONE.* **9(2),** e89318. doi:10.1371/journal.pone.0089318 (2014).

151. Martens, L. & Vizcaíno, J. A. A Golden Age for Working with Public Proteomics Data. *Trends in Biochemical Sciences.* **42(5),** 333–341. doi:10.1016/j.tibs.2017.01.001 (2017).

152. Park, J. E., Lim, H. R., Kim, J. W. & Shin, K. H. Metabolite changes in risk of type 2 diabetes mellitus in cohort studies: A systematic review and meta-analysis. *Diabetes Research and Clinical Practice.* **140,** 216–227. doi:10.1016/j.diabres.2018.03.045 (2018).

153. Mehta, K. Y. *et al.* Metabolomic biomarkers of pancreatic cancer: a meta-analysis study. *Oncotarget.* **8(40),** 68899–68915. doi:10.18632/oncotarget.20324 (2017).

154. Goveia, J. *et al.* Meta-analysis of clinical metabolic profiling studies in cancer: challenges and opportunities. *EMBO Molecular Medicine.* **8(10),** 1134–1142. doi:10.15252/emmm.201606798 (2016).

155. Rikke, B. A., Wynes, M. W., Rozeboom, L. M., Barón, A. E. & Hirsch, F. R. Independent validation test of the vote-counting strategy used to rank biomarkers from published studies. *Biomarkers in medicine.* **9(8),** 751–61. doi:10.2217/bmm.15.39 (2015).

156. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of research synthesis. *Nature.* **555(7695),** 175–182. doi:10.1038/nature25753 (2018).

157. Guasch-Ferré, M. *et al.* Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis. *Diabetes Care.* **39(5),** 833–846. doi:10.2337/dc15-2251 (2016).

158. Okekunle, A. P. *et al.* Abnormal circulating amino acid profiles in multiple metabolic disorders. *Diabetes Research and Clinical Practice.* **132,** 45–58. doi:10.1016/j.diabres.2017.07.023 (2017).

159. Siristatidis, C., Sertedaki, E. & Vaidakis, D. Metabolomics for improving pregnancy outcomes in women undergoing assisted reproductive technologies. *Cochrane Database Systematic Review.* **5,** CD011872. doi:10.1002/14651858.CD011872.pub2 (2017).

160. Bhandary, P., Seetharam, A. S., Arendsee, Z. W., Hur, M. & Wurtele, E. S. Raising orphans from a metadata morass: A researcher's guide to re-use of public 'omics data. *Plant Science.* **267,** 32–47. doi:10.1016/j.plantsci.2017.10.014 (2018).

161. Bowler, R. P. *et al.* Plasma Sphingolipids Associated with Chronic Obstructive Pulmonary Disease Phenotypes. *American Journal of Respiratory and Critical Care Medicine.* **191(3),** 275–284. doi:10.1164/rccm.201410-1771OC (2015).

162. Playdon, M. C. *et al.* Identifying biomarkers of dietary patterns by using metabolomics. *American Journal of Clinical Nutrition.* **105(2),** 450–465. doi:10.3945/ajcn.116.144501 (2017).

163. Ferreira, D. L. S. *et al.* Association of pre-pregnancy body mass index with offspring metabolic profile : Analyses of 3 European prospective birth cohorts. *PLOS Medicine.* **14(8),** e1002376. doi:10.1371/journal.pmed.1002376 (2017).

164. Imamura, F. *et al.* A combination of plasma phospholipid fatty acids and its association with incidence of type 2 diabetes : The EPIC-InterAct case-cohort study. *PLOS Medicine.* **14(10),** e1002409. doi:10.1371/journal.pmed.1002409 (2017).

165. Maga-Nteve, C., Vasilopoulou, C. G., Constantinou, C., Margarity, M. & Klapa, M. I. Sex-comparative study of mouse cerebellum physiology under adult-onset hypothyroidism: The significance of GC–MS metabolomic data normalization in meta-analysis. *Journal of Chromatography B.* **1041-1042,** 158–166. doi:10.1016/j.jchromb.2016.12.016 (2017).

166. Lowe Jr, W. L. *et al.* Maternal BMI and Glycemia Impact the Fetal Metabolome. *Diabetes Care.* **40(7),** 902–910. doi:10.2337/dc16-2452 (2017).

167. Floegel, A. *et al.* Serum metabolites and risk of myocardial infarction and ischemic stroke: a targeted metabolomic approach in two German prospective cohorts. *European Journal of Epidemiology.* **33,** 55–66. doi:10.1007/s10654-017-0333-0 (2018).

168. Metrustry, S. J. *et al.* Metabolomic signatures of low birthweight: Pathways to insulin resistance and oxidative stress. *PLOS ONE.* **13(3),** e0194316. doi:10.1371/journal.pone.0194316 (2018).

169. Spicer, R. A., Salek, R. & Steinbeck, C. Compliance with minimum information guidelines in public metabolomics repositories. *Scientific Data.* **4,** 170137. doi:10.1038/sdata.2017.137 (2017).

170. Spicer, R. A., Salek, R. & Steinbeck, C. A decade after the metabolomics standards initiative it's time for a revision. *Scientific Data.* **4,** 170138. doi:10.1038/sdata.2017.138 (2017).

171. Spicer, R. A. & Steinbeck, C. A lost opportunity for science: journals promote data sharing in metabolomics but do not enforce it. *Metabolomics.* **14(1),** 16. doi:10.1007/s11306-017-1309-5 (2018).

172. Kale, N. S. *et al.* MetaboLights: An Open-Access Database Repository for Metabolomics Data. *Current Protocols in Bioinformatics.* **53(March),** 14.13.1–14.13.18. doi:10.1002/0471250953.bi1413s53 (2016).

173. Van Rijswijk, M. *et al.* The future of metabolomics in ELIXIR [version 2; referees: 3 approved]. *F1000Research.* **6(ELIXIR),** 1649. doi:10.12688/f1000research.12342.2 (2017).

174. Dayalan, S., Xia, J., Spicer, R. A., Salek, R. & Roessner, U. in *Reference Module in Life Sciences* (Elsevier, 2018). doi:10.1016/B978-0-12-809633-8.20251-3.

175. Global Biological Standards Institute (GBSI). *The Case for Standards in Life Science Research: Seizing Opportunities at a Time of Critical Need.* tech. rep. (2013).

176. Ball, C. A. Are we stuck in the standards? *Nature Biotechnology.* **24(11),** 1374–1376. doi:10.1038/nbt1106-1374 (2006).

177. Tanenbaum, A. S. *Computer Networks* (Prentice Hall, 1996).

178. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics.* **29(4),** 365–71. doi:10.1038/ng1201-365 (2001).

179. Taylor, C. F. *et al.* The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology.* **25(8),** 887–893. doi:10.1038/nbt1329 (2007).

180. Jenkins, H. *et al.* A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology.* **22(12),** 1601–1606. doi:10.1038/nbt1041 (2004).

181. Lindon, J. C. *et al.* Summary recommendations for standardization and reporting of metabolic analyses. *Nature Biotechnology.* **23(7),** 833–838. doi:10.1038/nbt0705-833 (2005).

182. Fiehn, O. *et al.* The metabolomics standards initiative (MSI). *Metabolomics.* **3(3),** 175–178. doi:10.1007/s11306-007-0070-6 (2007).

183. Sansone, S.-A. *et al.* The Metabolomics Standards Initiative. *Nature Biotechnology.* **25(8),** 846–848. doi:10.1038/nbt0807-846b (2007).

184. Sansone, S.-A. *et al.* Metabolomics standards initiative: ontology working group work in progress. *Metabolomics.* **3(3),** 249–256. doi:10.1007/s11306-007-0069-z (2007).

185. Hardy, N. W. & Taylor, C. F. A roadmap for the establishment of standard data exchange structures for metabolomics. *Metabolomics.* **3(3),** 243–248. doi:10.1007/s11306-007-0071-5 (2007).

186. Sumner, L. W. *et al.* Proposed minimum reporting standards Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics.* **3(3),** 211–221. doi:10.1007/s11306-007-0082-2 (2007).

187. Rubtsov, D. V. *et al.* Proposed reporting requirements for the description of NMR-based metabolomics experiments. *Metabolomics.* **3(3),** 223–229. doi:10.1007/s11306-006-0040-4 (2007).

188. Goodacre, R. *et al.* Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics.* **3(3),** 231–241. doi:10.1007/s11306-007-0081-3 (2007).

189. Griffin, J. L. *et al.* Standard reporting requirements for biological samples in metabolomics experiments: mammalian/ in vivo experiments. *Metabolomics.* **3(3),** 179–188. doi:10.1007/s11306-007-0077-z (2007).

190. Van der Werf, M. J. *et al.* Standard reporting requirements for biological samples in metabolomics experiments: microbial and in vitro biology experiments. *Metabolomics.* **3(3),** 189–194. doi:10.1007/s11306-007-0080-4 (2007).

191. Fiehn, O. *et al.* Minimum reporting standards for plant biology context information in metabolomic studies. *Metabolomics.* **3(3),** 195–201. doi:10.1007/s11306-007-0068-0 (2007).

192. Morrison, N. *et al.* Standard reporting requirements for biological samples in metabolomics experiments: environmental context. *Metabolomics.* **3(3),** 203–210. doi:10.1007/s11306-007-0067-1 (2007).

193. Goodacre, R. Water, water, every where, but rarely any drop to drink. *Metabolomics.* **10,** 5–7. doi:10.1007/s11306-013-0618-6 (2014).

194. Griffin, J. L., Atherton, H. J., Steinbeck, C. & Salek, R. M. A Metadata description of the data in "A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human.". *BMC Research Notes.* **4,** 272. doi:10.1152/physiolgenomics.00194.2006 (2011).

195. Fiehn, O. *et al.* Quality control for plant metabolomics: Reporting MSI-compliant studies. *Plant Journal.* **53(4),** 691–704. doi:10.1111/j.1365-313X.2007.03387.x (2008).

196. Griffin, J. L. & Steinbeck, C. So what have data standards ever done for us? The view from metabolomics. *Genome Medicine.* **2(6),** 2–4. doi:10.1186/gm159 (2010).

197. Salek, R. M., Haug, K. & Steinbeck, C. Dissemination of metabolomics results: Role of MetaboLights and COSMOS. *GigaScience.* **2,** 8. doi:10.1186/2047-217X-2-8 (2013).

198. Goodacre, R. An overflow of... what else but metabolism! *Metabolomics.* **6,** 1–2. doi:10.1007/s11306-010-0201-3 (2010).

199. Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika.* **52(3-4),** 591–611. doi:10.2307/2333709 (1965).

200. Kruskal, W. H. & Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association.* **47(260),** 583–621. doi:10.1080/01621459.1952.10483441 (1952).

201. Dunn, O. J. Multiple Comparisons Using Rank Sums. *Technometrics.* **6(3),** 241–252. doi:10.1080/00401706.1964.10490181 (1964).

202. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society.* **57(1),** 289–300. doi:10.2307/2346101 (1995).

203. Mann, H. & Whitney, D. On a test of whether one of two random variables is stochastically larger than the other. *Statistics.* **18(1),** 50–60. doi:10.1021/acs.jpcc.5b07268 (1947).

204. Narath, S. H. *et al.* An Untargeted Metabolomics Approach to Characterize Short-Term and Long-Term Metabolic Changes after Bariatric Surgery. *PLOS ONE.* **11(9),** e0161425. doi:10.1371/journal.pone.0161425 (2016).

205. Mapstone, M. *et al.* Plasma phospholipids identify antecedent memory impairment in older adults. *Nature Medicine.* **20(4),** 415–418. doi:10.1038/nm.3466 (2014).

206. Samino, S. *et al.* Metabolomics reveals impaired maturation of HDL particles in adolescents with hyperinsulinaemic androgen excess. *Scientific Reports.* **5,** 11496. doi:10.1038/srep11496 (2015).

207. Cai, X. *et al.* Untargeted Lipidomic Profiling of Human Plasma Reveals Differences due to Race, Gender and Smoking Status. *Metabolomics: Open Access.* **4,** 131. doi:10.4172/2153-0769.1000131 (2014).

208. Considine, E. C., Thomas, G., Boulesteix, A. L., Khashan, A. S. & Kenny, L. C. Critical review of reporting of the data analysis step in metabolomics. *Metabolomics.* **14,** 7. doi:10.1007/s11306-017-1299-3 (2018).

209. Chervitz, S. A. *et al.* Data Standards for Omics Data: The Basis of Data Sharing and Reuse. *Methods in Molecular Biology (Clifton, N.J.)* **719,** 31–69. doi:10.1007/978-1-61779-027-0 (2011).

210. Brazma, A., Krestyaninova, M. & Sarkans, U. Standards for systems biology. *Nature Reviews Genetics.* **7(8),** 593–605. doi:10.1038/nrg1922 (2006).

211. Kind, T. & Fiehn, O. What are the obstacles for an integrated system for comprehensive interpretation of cross-platform metabolic profile data? *Bioanalysis.* **1(9),** 1511–1514. doi:10.4155/bio.09.141 (2009).

212. Salek, R. M., Steinbeck, C., Viant, M. R., Goodacre, R. & Dunn, W. B. The role of reporting standards for metabolite annotation and identification in metabolomic studies. *GigaScience.* **2,** 13. doi:10.1186/2047-217X-2-13 (2013).

213. Everett, J. R. A New Paradigm for Known Metabolite Identification in Metabonomics/Metabolomics: Metabolite Identification Efficiency. *Computational and Structural Biotechnology Journal.* **13,** 131–144. doi:10.1016/j.csbj.2015.01.002 (2015).

214. Schymanski, E. L. *et al.* Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science and Technology.* **48(4),** 2097–2098. doi:10.1021/es5002105 (2014).

215. Creek, D. J. *et al.* Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics.* **10(3),** 350–353. doi:10.1007/s11306-014-0656-8 (2014).

216. Sumner, L. W. *et al.* Proposed quantitative and alphanumeric metabolite identification metrics. *Metabolomics.* **10(6),** 1047–1049. doi:10.1007/s11306-014-0739-6 (2014).

217. Dona, A. C. *et al.* A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Computational and Structural Biotechnology Journal.* **14,** 135–153. doi:10.1016/j.csbj.2016.02.005 (2016).

218. Sanchon-Lopez, B. & Everett, J. R. New Methodology for Known Metabolite Identification in Metabonomics/Metabolomics: Topological Metabolite Identification Carbon Efficiency (tMICE). *Journal of Proteome Research.* **15(9),** 3405–3419. doi:10.1021/acs.jproteome.6b00631 (2016).

219. Schober, D., Salek, R. M. & Neumann, S. Towards standardized evidence descriptors for metabolite annotations. *CEUR Workshop Proceedings.* **1692(2013),** 1–5 (2016).

220. Turewicz, M. & Deutsch, E. W. Spectra, Chromatograms, Metadata: mzML - The Standard Data Format for Mass Spectrometer Output. *Methods in Molecular Biology (Clifton, N.J.)* **696,** 179–203. doi:10.1007/978-1-60761-987-1_11 (2011).

221. Walzer, M. *et al.* The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Molecular & Cellular Proteomics.* **12(8),** 2332–2340. doi:10.1074/mcp.O113.028506 (2013).

222. Schober, D. *et al.* NmrML: A Community Supported Open Data Standard for the Description, Storage, and Exchange of NMR Data. *Analytical Chemistry.* **90,** 649–656. doi:10.1021/acs.analchem.7b02795 (2018).

223. Griss, J. *et al.* The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Molecular & Cellular Proteomics.* **13(10),** 2765–2775. doi:10.1074/mcp.O113.036681 (2014).

224. Ioannidis, J. P. A. *et al.* Repeatability of published microarray gene expression analyses. *Nature Genetics.* **41(2),** 149–155. doi:10.1038/ng.295 (2009).

225. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery.* **10(9),** 712–713. doi:10.1038/nrd3439-c1 (2011).

226. Ferreira, J. D., Inácio, B., Salek, R. M. & Couto, F. M. Assessing Public Metabolomics Metadata, Towards Improving Quality. *Journal of Integrative Bioinformatics.* **14(4),** doi:10.1515/jib-2017-0054 (2017).

227. Inácio, B., Ferreira, J. D. & Couto, F. M. Metadata Analyser: measuring metadata quality. In: (eds) 11th International Conference on Practical Applications of Computational Biology & Bioinformatics. PACBB 2017. Advances in Intelligent Systems and Computing. **616.** doi:10.1007/978-3-319-60816-7_24 (2017).

228. Larralde, M. *et al.* mzML2ISA & nmrML2ISA: generating enriched ISA-Tab metadata files from metabolomics XML data. *Bioinformatics.* **33(16),** 2598–2600. doi:10.1093/bioinformatics/btx169 (2017).

229. Quackenbush, J. *et al.* Top-down standards will not serve systems biology. *Nature.* **440(7080),** 24. doi:10.1038/440024a (2006).

230. Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition. *Nucleic Acids Research.* **45(D1),** D1100–D1106. doi:10.1093/nar/gkw936 (2017).

231. Bloom, T., Ganley, E. & Winker, M. Data Access for the Open Access Literature: PLOS's Data Policy. *PLOS Medicine.* **11(2),** e1001797. doi:10.1371/journal.pmed.1001607 (2014).

232. Nuijten, M. B. *et al.* Journal Data Sharing Policies and Statistical Reporting Inconsistencies in Psychology. *Collabra: Psychology.* **3(1),** 1–22. doi:10.1525/collabra.102 (2017).

233. Thelwall, M. & Kousha, K. Do journal data sharing mandates work? Life sciences evidence from Dryad. *Aslib Journal of Information Management.* **69,** 36–45. doi:10.1108/ajim-09-2016-0159 (2017).

234. Ramsden, J. J. Metabolomics and Metabonomics. In: Bioinformatics: An Introduction. Computational Biology, vol 10. Springer, London, 1–6. doi:10.1007/978-1-84800-257-9_16 (2009).

235. Fernie, A. R. *et al.* Recommendations for Reporting Metabolite Data. *The Plant Cell.* **23(7),** 2477–2482. doi:10.1105/tpc.111.086272 (2011).

236. Where are the data? *Nature.* **537(7619),** 138. doi:10.1038/537138a (2016).

237. Vines, T. H. *et al.* Mandated data archiving greatly improves access to research data. *FASEB Journal.* **27(4),** 1304–1308. doi:10.1096/fj.12-218164 (2013).

238. Van Noorden, R. Confusion over publisher's pioneering open-data rules. *Nature.* **515(7528),** 478–478. doi:10.1038/515478a (2014).

239. Federer, L. M. *et al.* Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLOS ONE.* **13(5),** e0194768. doi:10.1371/journal.pone.0194768 (2018).

240. Lavrakas, P. J. *Encyclopedia of survey research methods* doi:10.4135/9781412963947 (SAGE Publications Ltd, Thousand Oaks, CA, 2008).

241. Deutsch, E. W. File Formats Commonly Used in Mass Spectrometry Proteomics. *Molecular & Cellular Proteomics.* **11(12),** 1612–1621. doi:10.1074/mcp.R112.019695 (2012).

242. Rafiei, A. & Sleno, L. Comparison of peak-picking workflows for untargeted liquid chromatography/ high-resolution mass spectrometry metabolomics data analysis. *Rapid Communications in Mass Spectrometry.* **29,** 119–127. doi:10.1002/rcm.7094 (2015).

243. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Analytical Chemistry.* **89(17),** 8689–8695. doi:10.1021/acs.analchem.7b01069 (2017).

244. Baran, R. Untargeted metabolomics suffers from incomplete data analysis. *Metabolomics.* **13,** 107. doi:10.1101/143818 (2017).

245. Pearson, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science.* **50(302),** 157–175. doi:10.1080/14786440009463897 (1900).

246. Agresti, A. *An Introduction to Categorical Data Analysis: Second Edition* doi:10.1002/0470114754 (2007).

247. Boudreau, P. D. *et al.* Expanding the Described Metabolome of the Marine Cyanobacterium Moorea producens JHB through Orthogonal Natural Products Workflows. *PLOS ONE.* **10(7),** e0133297. doi:10.1371/journal.pone.0133297 (2015).

248. Abela, L. *et al.* Plasma metabolomics reveals a diagnostic metabolic fingerprint for mitochondrial aconitase (ACO2) deficiency. *PLOS ONE.* **12(5),** e0176363. doi:10.1371/journal.pone.0176363 (2017).

249. Rzeznik, M. *et al.* Identification of a discriminative metabolomic fingerprint of potential clinical relevance in saliva of patients with periodontitis using 1H nuclear magnetic resonance (NMR) spectroscopy. *PLOS ONE.* **12(8),** e0182767. doi:10.1371/journal.pone.0182767 (2017).

250. Sandlers, Y. *et al.* Metabolomics Reveals New Mechanisms for Pathogenesis in Barth Syndrome and Introduces Novel Roles for Cardiolipin in Cellular Function. *PLOS ONE.* **11(3),** e0151802. doi:10.1371/journal.pone.0151802 (2016).

251. Suh, D. H. *et al.* Comparison of Metabolites Variation and Antiobesity Effects of Fermented versus Nonfermented Mixtures of Cudrania tricuspidata, Lonicera caerulea, and Soybean According to Fermentation In Vitro and In Vivo. *PLOS ONE.* **11(2),** e0149022. doi:10.1371/journal.pone.0149022 (2016).

252. Batova, A. *et al.* Englerin A induces an acute inflammatory response and reveals lipid metabolism and ER stress as targetable vulnerabilities in renal cell carcinoma. *PLOS ONE.* **12(3),** e0172632. doi:10.1371/journal.pone.0172632 (2017).

253. Rempe, C. S. *et al.* Computational ranking of yerba mate small molecules based on their predicted contribution to antibacterial activity against methicillin-resistant staphylococcus aureus. *PLOS ONE.* **10(5),** e0123925. doi:10.1371/journal.pone.0123925 (2015).

254. Zhang, T. *et al.* Changes in the Milk Metabolome of the Giant Panda (Ailuropoda melanoleuca) with Time after Birth - Three Phases in Early Lactation and Progressive Individual Differences. *PLOS ONE.* **10(12),** e0143417. doi:10.1371/journal.pone.0143417 (2015).

255. Misra, B. B., De Armas, E., Tong, Z. & Chen, S. Metabolomic Responses of Guard Cells and Mesophyll Cells to Bicarbonate. *PLOS ONE.* **10(12),** e0144206. doi:10.1371/journal.pone.0144206 (2015).

256. Witowski, N. *et al.* A Four-Compartment Metabolomics Analysis of the Liver, Muscle, Serum, and Urine Response to Polytrauma with Hemorrhagic Shock following Carbohydrate Prefeed. *PLOS ONE.* **10(4),** e0124467. doi:10.1371/journal.pone.0124467 (2015).

257. Overmyer, K. A., Thonusin, C., Qi, N. R., Burant, C. F. & Evans, C. R. Impact of Anesthesia and Euthanasia on Metabolomics of Mammalian Tissues: Studies in a C57BL/6J Mouse Model. *PLOS ONE.* **10(2),** e0117232. doi:10.1371/journal.pone.0117232 (2015).

258. Aslam, M. N. *et al.* Calcium Reduces Liver Injury in Mice on a High-Fat Diet: Alterations in Microbial and Bile Acid Profiles. *PLOS ONE.* **11(11),** e0166178. doi:10.1371/journal.pone.0166178 (2016).

259. Gelaye, B. *et al.* Maternal Early Pregnancy Serum Metabolomics Profile and Abnormal Vaginal Bleeding as Predictors of Placental Abruption: A Prospective Study. *PLOS ONE.* **11(6),** e0156755. doi:10.1371/journal.pone.0156755 (2016).

260. Uppal, K. *et al.* Plasma metabolomics reveals membrane lipids, aspartate/asparagine and nucleotide metabolism pathway differences associated with chloroquine resistance in Plasmodium vivax malaria. *PLOS ONE.* **12(8),** e0182819. doi:10.1371/journal.pone.0182819 (2017).

261. Laíns, I. *et al.* Human plasma metabolomics in age-related macular degeneration (AMD) using nuclear magnetic resonance spectroscopy. *PLOS ONE.* **12(5),** e0177749. doi:10.1371/journal.pone.0177749 (2017).

262. Gürdeniz, G., Kristensen, M., Skov, T. & Dragsted, L. O. The Effect of LC-MS Data Preprocessing Methods on the Selection of Plasma Biomarkers in Fed vs. Fasted Rats. *Metabolites.* **2,** 77–99. doi:10.3390/metabo2010077 (2012).

263. Dai, H. J., Chang, Y. C., Tzong-Han Tsai, R. & Hsu, W. L. New challenges for biological text-mining in the next decade. *Journal of Computer Science and Technology.* **25(1),** 169–179. doi:10.1007/s11390-010-9313-5 (2009).

264. Pasquetto, I. V., Randles, B. M. & Borgman, C. L. On the Reuse of Scientific Data. *Data Science Journal.* **16,** 8. doi:10.5334/dsj-2017-008 (2017).

265. Rung, J. & Brazma, A. Reuse of public genome-wide gene expression data. *Nature Reviews Genetics.* **14(2),** 89–99. doi:10.1038/nrg3394 (2013).

266. Vaudel, M. *et al.* Exploring the potential of public proteomics data. *Proteomics.* **16(2),** 214–225. doi:10.1002/pmic.201500295 (2016).

267. Robinson-García, N., Jiménez-Contreras, E. & Torres-Salinas, D. Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology.* **67(12),** 2964–2975. doi:10.1002/asi.23529 (2016).

268. He, L. & Nahar, V. Reuse of scientific data in academic publications: An investigation of Dryad Digital Repository. *Aslib Journal of Information Management.* **68(4),** 478–494. doi:10.1108/ajim-01-2016-0008 (2016).

269. Kafkas, Ş., Kim, J. H. & McEntyre, J. R. Database Citation in Full Text Biomedical Articles. *PLOS ONE.* **8(5),** e63184. doi:10.1371/journal.pone.0063184 (2013).

270. Kafkas, Ş., Kim, J. H., Pi, X. & McEntyre, J. R. Database citation in supplementary data linked to Europe PubMed Central full text biomedical articles. *Journal of Biomedical Semantics.* **6,** 1. doi:10.1186/2041-1480-6-1 (2015).

271. Bousfield, D. *et al.* Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources [version 1; referees: 3 approved]. *F1000Research.* **5(ELIXIR),** 160. doi:10.12688/f1000research.7911.1 (2016).

272. Tenenbaum, J. D. & Blach, C. Best practices and lessons learned from reuse of 4 patient-derived metabolomics datasets in Alzheimer's disease. *Proceedings of the Pacific Symposium.* **23,** 280–291. doi:10.1142/9789813235533_0026 (2018).

273. Steinbeck, C. *et al.* MetaboLights: Towards a new COSMOS of metabolomics data management. *Metabolomics.* **8(5),** 757–760. doi:10.1007/s11306-012-0462-0 (2012).

274. Salek, R. M. *et al.* The MetaboLights repository: curation challenges in metabolomics. *Database.* **2013,** bat029. doi:10.1093/database/bat029 (2013).

275. Mohimani, H. *et al.* Dereplication of peptidic natural products through database search of mass spectra. *Nature Chemical Biology.* **13(1),** 30–37. doi:10.1038/nchembio.2219 (2016).

276. Gurevich, A. *et al.* Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nature Microbiology.* **3(3),** 319–327. doi:10.1038/s41564-017-0094-2 (2018).

277. Chen, T. *et al.* LiverWiki: a wiki-based database for human liver. *BMC Bioinformatics.* **18,** 452. doi:10.1186/s12859-017-1852-0 (2017).

278. Li, S. *et al.* Predicting Network Activity from High Throughput Metabolomics. *PLOS Computational Biology.* **9(7),** e1003123. doi:10.1371/journal.pcbi.1003123 (2013).

279. González-Beltrán, A., Neumann, S., Maguire, E., Sansone, S.-A. & Rocca-Serra, P. The Risa R/Bioconductor package: integrative data analysis from experimental metadata and back again. *BMC Bioinformatics.* **15(Suppl 1),** S11. doi:10.1186/1471-2105-15-S1-S11 (2014).

280. Stanstrup, J., Neumann, S. & Vrhovšek, U. PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Analytical Chemistry.* **87(18),** 9421–9428. doi:10.1021/acs.analchem.5b02287 (2015).

281. Gromski, P. S., Xu, Y., Hollywood, K. A., Turner, M. L. & Goodacre, R. The influence of scaling metabolomics data on model classification accuracy. *Metabolomics.* **11(3),** 684–695. doi:10.1007/s11306-014-0738-7 (2015).

282. Trutschel, D., Schmidt, S., Grosse, I. & Neumann, S. Joint Analysis of Dependent Features within Compound Spectra Can Improve Detection of Differential Features. *Frontiers in Bioengineering and Biotechnology.* **3,** 129. doi:10.3389/fbioe.2015.00129 (2015).

283. Moreno, P. *et al.* BiNChE: A web tool and library for chemical enrichment analysis based on the ChEBI ontology. *BMC Bioinformatics.* **16,** 56. doi:10.1186/s12859-015-0486-3 (2015).

284. Saccenti, E. & Timmerman, M. E. Approaches to sample size determination for multivariate data: Applications to PCA and PLS-DA of omics data. *Journal of Proteome Research.* **15(8),** 2379–2393. doi:10.1021/acs.jproteome.5b01029 (2016).

285. Davidson, R. L., Weber, R. M. J., Liu, H., Sharma-Oates, A. & Viant, M. R. Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. *GigaScience.* **5,** 10. doi:10.1186/s13742-016-0115-8 (2016).

286. Nowak, C. *et al.* Effect of Insulin Resistance on Monounsaturated Fatty Acid Levels: A Multi-cohort Non-targeted Metabolomics and Mendelian Randomization Study. *PLOS Genetics.* **12(10),** e1006379. doi:10.1371/journal.pgen.1006379 (2016).

287. Fall, T. *et al.* Non-targeted metabolomics combined with genetic analyses identifies bile acid synthesis and phospholipid metabolism as being associated with incident type 2 diabetes. *Diabetologia.* **59(10),** 2114–2124. doi:10.1007/s00125-016-4041-1 (2016).

288. Xu, Y., Muhamadali, H., Sayqal, A., Dixon, N. & Goodacre, R. Partial least squares with structured output for modelling the metabolomics data obtained from complex experimental designs: A study into the Y-block coding. *Metabolites.* **6(4),** 38. doi:10.3390/metabo6040038 (2016).

289. Salhi, A. *et al.* DES-ncRNA: A knowledgebase for exploring information about human micro and long noncoding RNAs based on literature-mining. *RNA Biology.* **14(7),** 963–971. doi:10.1080/15476286.2017.1312243 (2017).

290. Salhi, A. *et al.* DES-TOMATO: A Knowledge Exploration System Focused On Tomato Species. *Scientific Reports.* **7(1),** 5968. doi:10.1038/s41598-017-05448-0 (2017).

291. Lawson, T. N. *et al.* MsPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics. *Analytical Chemistry.* **89(4),** 2432–2439. doi:10.1021/acs.analchem.6b04358 (2017).

292. Li, B. *et al.* NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Research.* **45(W1),** W162–W170. doi:10.1093/nar/gkx449 (2017).

293. Myint, L., Kleensang, A., Zhao, L., Hartung, T. & Hansen, K. D. Joint Bounding of Peaks Across Samples Improves Differential Analysis in Mass Spectrometry-Based Metabolomics. *Analytical Chemistry.* **89(6),** 3517–3523. doi:10.1021/acs.analchem.6b04719 (2017).

294. DeFelice, B. C. *et al.* Mass Spectral Feature List Optimizer (MS-FLO): A Tool To Minimize False Positive Peak Reports in Untargeted Liquid Chromatography-Mass Spectroscopy (LC-MS) Data Processing. *Analytical Chemistry.* **89(6),** 3250–3255. doi:10.1021/acs.analchem.6b04372 (2017).

295. Uppal, K., Walker, D. I. & Jones, D. P. xMSannotator: an R package for network-based annotation of high-resolution metabolomics data. *Analytical Chemistry.* **89(2),** 1063–1067. doi:10.1021/acs.analchem.6b01214 (2017).

296. Shah, J. S. *et al.* Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. *BMC Bioinformatics.* **18,** 114. doi:10.1186/s12859-017-1547-6 (2017).

297. Scheubert, K. *et al.* Significance estimation for large scale untargeted metabolomics annotations. *Nature Communications.* **8(1),** 1494. doi:10.1038/s41467-017-01318-5 (2017).

298. Pauling, J. K. *et al.* Proposal for a common nomenclature for fragment ions in mass spectra of lipids. *PLOS ONE.* **12(11),** e018839. doi:10.1371/journal.pone.0188394 (2017).

299. Faulon, J.-L. Molecular structures enumeration and virtual screening in the chemical space with RetroPath2.0. *Journal of Cheminformatics.* **9(1),** 64. doi:10.1186/s13321-017-0252-9 (2017).

300. Barupal, D. K. & Fiehn, O. Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. *Scientific Reports.* **7(1),** 14567. doi:10.1038/s41598-017-15231-w (2017).

301. Hartmann, A. C. *et al.* Meta-mass shift chemical profiling of metabolomes from coral reefs. *Proceedings of the National Academy of Sciences of the United States of America.* **114(44),** 11685–11690. doi:10.1073/pnas.1710248114 (2017).

302. Marco-Ramell, A. *et al.* Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. *BMC Bioinformatics.* **19,** 1. doi:10.1186/s12859-017-2006-0 (2018).

303. Smith, R. & Roberts, I. Time for sharing data to become routine: the seven excuses for not doing so are all invalid [version 1; referees: 2 approved, 1 approved with reservations]. *F1000Research.* **5,** 781. doi:10.12688/f1000research.8422.1 (2016).

304. Longo, D. L. & Drazen, J. M. Data Sharing. *New England Journal of Medicine.* **374(3),** 276–277. doi:10.1056/nejme1516564 (2016).

305. Griss, J., Perez-Riverol, Y., Hermjakob, H. & Vizcaíno, J. A. Identifying novel biomarkers through data mining-A realistic scenario? *Proteomics - Clinical Applications.* **9(3-4),** 437–443. doi:10.1002/prca.201400107 (2015).

306. Jones, A. R. *et al.* The mzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results. *Molecular & Cellular Proteomics.* **11(7),** M111.014381. doi:10.1074/mcp.M111.014381 (2012).

307. Aretz, I. & Meierhofer, D. Advantages and pitfalls of mass spectrometry based metabolome profiling in systems biology. *International Journal of Molecular Sciences.* **17(5),** 632. doi:10.3390/ijms17050632 (2016).

308. Rueedi, R. *et al.* Genome-Wide Association Study of Metabolic Traits Reveals Novel Gene-Metabolite-Disease Links. *PLOS Genetics.* **10(2),** e1004132. doi:10.1371/journal.pgen.1004132 (2014).

309. Berghmans, S. *et al. Open data: The researcher perspective* tech. rep. (2017), 48. doi:10.17632/bwrnfb4bvh.1.

310. Lappalainen, I. *et al.* The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics.* **47(7),** 692–695. doi:10.1038/ng.3312 (2015).

311. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics.* **39(10),** 1181–1186. doi:10.1038/ng1007-1181 (2007).

312. Faulconbridge, A. *et al.* Updates to BioSamples database at European Bioinformatics Institute. *Nucleic Acids Research.* **42(Database issue),** D50–D52. doi:10.1093/nar/gkt1081 (2014).

313. Barrett, T. *et al.* BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. *Nucleic Acids Research.* **40(Database issue),** D57–D63. doi:10.1093/nar/gkr1163 (2012).

314. Craig, R., Cortens, J. P. & Beavis, R. C. Open Source System for Analyzing, Validating, and Storing Protein Identification Data research articles. *Journal of Proteome Research.* **3(6),** 1234–1242. doi:10.1021/pr049882h (2004).

315. Deutsch, E. W., Lam, H. & Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Reports.* **9(5),** 429–434. doi:10.1038/embor.2008.56 (2008).

316. Henry, V. J., Bandrowski, A. E., Pepin, A. S., Gonzalez, B. J. & Desfeux, A. OMICtools: an informative directory for multi-omic data analysis. *Database.* **2014,** bau069. doi:10.1093/database/bau069 (2014).

317. Smolinska, A., Blanchet, L., Buydens, L. M. C. & Wijmenga, S. S. NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta.* **750,** 82–97. doi:10.1016/j.aca.2012.05.049 (2012).

318. Smith, C. A. *et al.* METLIN: a metabolite mass spectral database. *Therapeutic Drug Monitoring.* **27(6),** 747–751. doi:10.1097/01.ftd.0000179845.53213.39 (2005).

319. Ludwig, C. *et al.* Birmingham Metabolite Library: A publicly accessible database of 1-D 1H and 2-D 1H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics.* **8,** 8. doi:10.1007/s11306-011-0347-7 (2012).

320. Emwas, A.-H. *et al.* Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: a review. *Metabolomics.* **11(4),** 872–894. doi:10.1007/s11306-014-0746-7 (2015).

321. Alonso, A., Marsal, S. & Julià, A. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in Bioengineering and Biotechnology.* **3,** 3–23. doi:10.3389/fbioe.2015.00023 (2015).

322. Armitage, E. G., Godzien, J., Alonso-Herranz, V., López-Gonzálvez, Á. & Barbas, C. Missing value imputation strategies for metabolomics data. *Electrophoresis.* **36(24),** 3050–3060. doi:10.1002/elps.201500352 (2015).

323. Wei, R. *et al.* Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports.* **8(1),** 663. doi:10.1038/s41598-017-19120-0 (2018).

324. Wu, Y. & Li, L. Sample normalization methods in quantitative metabolomics. *Journal of Chromatography A.* **1430,** 80–95. doi:10.1016/j.chroma.2015.12.007 (2016).

325. Van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K. & van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics.* **7,** 142. doi:10.1186/1471-2164-7-142 (2006).

326. Westerhuis, J. A. *et al.* Assessment of PLSDA cross validation. *Metabolomics.* **4(1),** 81–89. doi:10.1007/s11306-007-0099-6 (2008).

327. Szymańska, E., Saccenti, E., Smilde, A. K. & Westerhuis, J. A. Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics.* **8(Suppl 1),** 3–16. doi:10.1007/s11306-011-0330-3 (2012).

328. Gromski, P. S. *et al.* A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta.* **879,** 10–23. doi:10.1016/j.aca.2015.02.012 (2015).

329. Booth, S. C., Weljie, A. M. & Turner, R. J. Computational Tools for the Secondary Analysis of Metabolomics Experiments. *Computational and Structural Biotechnology Journal.* **4(5),** e201301003. doi:10.5936/csbj.201301003 (2013).

330. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research.* **44(D1),** D633–D639. doi:10.1093/nar/gkx935 (2016).

331. Jewison, T. *et al.* SMPDB 2.0: Big improvements to the small molecule pathway database. *Nucleic Acids Research.* **42(Database issue),** D478–D484. doi:10.1093/nar/gkt1067 (2014).

332. Tsizin, S. *et al.* Comparison of electrospray LC–MS, LC–MS with Cold EI and GC–MS with Cold EI for sample identification. *International Journal of Mass Spectrometry.* **422,** 119–125. doi:10.1016/j.ijms.2017.09.006 (2017).

333. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology.* **30(10),** 918–920. doi:10.1038/nbt.2377 (2012).

334. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics.* **9,** 504. doi:10.1186/1471-2105-9-504 (2008).

335. Benton, H. P., Wong, D. M., Trauger, S. A. & Siuzdak, G. XCMS2: Processing tandem mass spectrometry data for metabolite identification and structural characterization. *Analytical Chemistry.* **80(16),** 6382–6389. doi:10.1021/ac800795f (2008).

336. Ramautar, R., Somsen, G. W. & de Jong, G. J. CE–MS for metabolomics: Developments and applications in the period 2014–2016. *Electrophoresis.* **38(1),** 190–202. doi:10.1002/elps.201600370 (2017).

337. Hinz, C., Liggi, S. & Griffin, J. L. The potential of Ion Mobility Mass Spectrometry for high-throughput and high-resolution lipidomics. *Current Opinion in Chemical Biology.* **42,** 42–50. doi:10.1016/j.cbpa.2017.10.018 (2018).

338. Winnike, J. H. *et al.* Comparison of GC-MS and GCxGC-MS in the analysis of human serum samples for biomarker discovery. *Journal of Proteome Research.* **14(4),** 1810–1817. doi:10.1021/pr5011923 (2015).

339. Oetjen, J. *et al.* Benchmark datasets for 3D MALDI- and DESI-imaging mass spectrometry. *GigaScience.* **4,** 20. doi:10.1186/s13742-015-0059-4 (2015).

340. Kruk, J. *et al.* NMR Techniques in Metabolomic Studies: A Quick Overview on Examples of Utilization. *Applied Magnetic Resonance.* **48,** 1–21. doi:10.1007/s00723-016-0846-9 (2017).

341. Hannay, J. E. *et al.* How do scientists develop and use scientific software? *Proceedings of the 2009 ICSE Workshop on Software Engineering for Computational Science and Engineering, SECSE 2009,* 1–8. doi:10.1109/secse.2009.5069155 (2009).

342. Artaza, H. *et al.* Top 10 metrics for life science software good practices [version 1; referees: 2 approved]. *F1000Research.* **5(ELIXIR),** 2000. doi:10.12688/f1000research.9206.1 (2016).

343. Joppa, L. N. *et al.* Troubling Trends in Scientific Software Use. *Science.* **340,** 814–816. doi:10.1126/science.1231535 (2013).

344. Smith, R., Ventura, D. & Prince, J. T. Novel algorithms and the benefits of comparative validation. *Bioinformatics.* **29(12),** 1583–1585. doi:10.1093/bioinformatics/btt176 (2013).

345. Hwang, L., Fish, A., Soito, L., Smith, M. & Kellogg, L. H. Software and the Scientist: Coding and Citation Practices in Geodynamics. *Earth and Space Science.* **4(11),** 670–680. doi:10.1002/2016ea000225 (2017).

346. Coble, J. B. & Fraga, C. G. Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery. *Journal of Chromatography A.* **1358,** 155–164. doi:10.1016/j.chroma.2014.06.100 (2014).

347. Lange, E., Tautenhahn, R., Neumann, S. & Gröpl, C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics.* **9,** 375. doi:10.1186/1471-2105-9-375 (2008).

348. Li, Z. *et al.* Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. *Analytica Chimica Acta.* **1029,** 50–57. doi:10.1016/j.aca.2018.05.001 (2018).

349. Franceschi, P., Masuero, D., Vrhovsek, U., Mattivi, F. & Wehrens, R. A benchmark spike-in data set for biomarker identification in metabolomics. *Journal of Chemometrics.* **26,** 16–24. doi:10.1002/cem.1420 (2012).

350. Kirwan, J. A., Weber, R. J., Broadhurst, D. I. & Viant, M. R. Direct infusion mass spectrometry metabolomics dataset: A benchmark for data processing and quality control. *Scientific Data.* **1,** 140012. doi:10.1038/sdata.2014.12 (2014).

351. Vaniya, A., Samra, S. N., Palazoglu, M., Tsugawa, H. & Fiehn, O. Using MS-FINDER for identifying 19 natural products in the CASMI 2016 contest. *Phytochemistry Letters.* **21,** 306–312. doi:10.1016/j.phytol.2016.12.008 (2017).

352. Brouard, C. *et al.* Fast metabolite identification with Input Output Kernel Regression. *Bioinformatics.* **32(12),** i28–i36. doi:10.1093/bioinformatics/btw246 (2016).

353. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences of the United States of America.* **112(41),** 12580–5. doi:10.1073/pnas.1509788112 (2015).

354. Allen, F., Pon, A., Wilson, M., Greiner, R. & Wishart, D. CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Research.* **42(Web Server issue),** W94–W99. doi:10.1093/nar/gku436 (2014).

355. Martin, J. C. *et al.* Can we trust untargeted metabolomics? Results of the metabo-ring initiative, a large-scale, multi-instrument inter-laboratory study. *Metabolomics.* **11(4),** 807–821. doi:10.1007/s11306-014-0740-0 (2015).

356. Djekic, D., Pinto, R., Vorkas, P. A. & Henein, M. Y. Replication of LC–MS untargeted lipidomics results in patients with calcific coronary disease: An interlaboratory reproducibility study. *International Journal of Cardiology.* **222,** 1042–1048. doi:10.1016/j.ijcard.2016.07.214 (2016).

357. Siskos, A. P. *et al.* Interlaboratory Reproducibility of a Targeted Metabolomics Platform for Analysis of Human Serum and Plasma. *Analytical Chemistry.* **89(1),** 656–665. doi:10.1021/acs.analchem.6b02930 (2017).

358. Bowden, J. A. *et al.* Harmonizing lipidomics: NIST interlaboratory comparison exercise for lipidomics using SRM 1950–Metabolites in Frozen Human Plasma. *Journal of Lipid Research.* **58(12),** 2275–2288. doi:10.1194/jlr.M079012 (2017).

359. Smith, R. Conversations with 100 Scientists in the Field Reveal a Bifurcated Perception of the State of Mass Spectrometry Software. *Journal of Proteome Research.* **17(4),** 1335–1339. doi:10.1021/acs.jproteome.8b00015 (2018).

360. Hothorn, T. & Leisch, F. Case studies in reproducibility. *Briefings in Bioinformatics.* **12(3),** 288–300. doi:10.1093/bib/bbq084 (2011).

361. Gilbert, K. *et al.* Recommendations for utilizing and reporting population genetic anlyses: the reproducibility of genetic clustering using the program STRUCTURE. *Molecular Ecology.* **21(20),** 4925–4930. doi:10.1111/j.1365-294X.2012.05754.x (2012).

Appendix A

Code and Data for Introductory Figures

Code and data for reproducing figures 1.4 and 1.6 can be found at: https://github.com/RASpicer/thesis_introduction_figures. Analysis was performed using R version 3.5.1 with the packages: ggplot2 version 3.0.0, reshape2 version 1.4.3. and plyr version 1.8.4..

Appendix B

Data Citations for Compliance with MSI Reporting Standards

Pieter Dorrestein (2014) GNPS - Topobiographical molecular analysis of human skin. GNPS, MSV000078556.

Pieter Dorrestein (2014) GNPS-Lung_cysticFibrosis_tandemMS. GNPS, MSV000078565.

Pieter Dorrestein and Forest Rohwer (2016) GnPS CF and non-CF sputa UPLC-MS/MS. GNPS, MSV000078586.

Forest Rohwer and Pieter Dorrestein (2014) GNPS CF and non-CF sputa UPLC-MS/MS. GNPS, MSV000078589.

Pieter Dorrestein (2014) GNPS_mouse_embryo_nanoDESI_PNAS2013_Hsu. GNPS, MSV-000078590.

Pieter Dorrestein (2014) GNPS - 3D molecular analysis of skin surface_Volunteer 3-man_ Volunteer 4-Woman. MS/MS Data. GNPS, MSV000078622.

Pieter Dorrestein (2014) GNPS_mouse_embryo_nanoDESI_PNAS2013_Hsu – raw Thermo files. GNPS, MSV000078637.

Pieter Dorrestein (2014) GNPS_mouse embryo tissue section nanoDESI data dependent collection. GNPS, MSV000078638.

Pieter Dorrestein and Forest Rohwer (2014) GNPS_A549 Epithelial cell Bacteria Interaction NanoDESI-MS/MS. GNPS, MSV000078642.

Pieter Dorrestein (2014) GNPS - 3D molecular analysis of skin surface_Time_Course_ Volunteer 5-man-Face. MS/MS Data. GNPS, MSV000078683.

Forest Rohwer and Pieter Dorrestein (2014) GNPS Longitudinal CF Sputum Through Exacerbation UPLC-MS/MS. GNPS, MSV000078719.

Forest Rohwer and Pieter Dorrestein (2014) GNPS HPLC CF Sputa Ex v St May 2014 maxis. GNPS, MSV000078726.

Michael Katze, Ralph Baric, Amy Sims, Katrina Waters, Dick Smith and Thomas Metz (2014) GNPS_Lipidomics charactarization of SARS infected 2B4 cells positive ionization. GNPS, MSV000078780.

Michael Katze, Ralph Baric, Amy Sims, Katrina Waters, Dick Smith and Thomas Metz (2014) GNPS_Lipidomics charactarization of SARS infected 2B4 cells negative ionization. GNPS, MSV000078781.

Sarkis K. Mazmanian (2014) GNPS_Autism_MIA_mouse_model_20140808. GNPS, MSV-000078801.

Pieter Dorrestein (2014) GNPS - Trace detection of skin molecules on objects_Volunteer 6. MS/MS Data. GNPS, MSV000078816.

Pieter Dorrestein (2014) GNPS - Trace detection of skin molecules on objects_Volunteers 1-11_MS/MS Data. GNPS, MSV000078832.

Scott Kelly (2014) GNPS_Periodontitis. GNPS, MSV000078894.

Forest Rohwer and Pieter Dorrestein (2014) GNPS CF sputum HPLC CF v non-CF maxis. GNPS, MSV000078903.

Forest Rohwer and Pieter Dorrestein (2014) GNPS UPLC May 2014 CF Sputum. GNPS, MSV000078922.

Pieter Dorrestein (2014) GNPS_CysticFibrosis_LungTissue. GNPS, MSV000078938.

Pieter Dorrestein (2014) GNPS - Trace detection of skin molecules on objects_Volunteers 12-39_MS/MS Data. GNPS, MSV000078993.

Pieter Dorrestein (2015) GNPS - Molecular analysis of human brain tissues. GNPS, MSV000079040.

Forest Rohwer and Pieter Dorrestein (2015) GNPS YW CF Sputum Samples Exacerbation Longitudinal. GNPS, MSV000079104.

R. C. Hunter (2015) GNPS CF sinusitis samples. GNPS, MSV000079105.

Pieter Dorrestein (2015) GNPS IBD Fecal Sample Solvent Test. GNPS, MSV000079115.

Pieter Dorrestein (2016) GNPS-CF Fecal study. GNPS, MSV000079134.

Pieter Dorrestein (2015) GNPS - 150702_Office_Human_Profiling. GNPS, MSV000079181.

Thomas Metz (2015) GNPS - Mouse lung metabolome response to H5N1 (and mutants) and H1N1 Influenza viruses. GNPS, MSV000079327.

Pieter Dorrestein (2015) GNPS - Swab background analysis of catchall buccal swabs. GNPS, MSV000079341.

Maria Gloria Dominguez-Bello and Pieter Dorrestein (2015) GNPS_Amazon skin and environmental samples. GNPS, MSV000079389.

Pieter Dorrestein (2015) GNPS-Cystic Fibrosis_Human lung tissue. GNPS, MSV000079398.

Nicola Zamboni (2015) GNPS - Mouse BXD liver metabolomics. GNPS, MSV000079411.

Pieter Dorrestein and Forest Rohwer (2016) GNPS Old Sputum Longitudinal. GNPS, MSV000079443.

Pieter Dorrestein and Forest Rohwer (2016) GNPS CF1 4y longitudinal. GNPS, MSV00007-9444.

Pieter Dorrestein and Rob Knight (2016) Fermentation Festival Samples GNPS. GNPS, MSV000079485.

Rob Knight and Pieter Dorrestein (2016) GNPS - High/Low Biomass Swab/Extraction Study - C18-LC-MS/MS - Positive polarity. GNPS, MSV000079523.

Yoshihiro Kawaoka, Katrina Waters, Richard Smith and Thomas Metz (2016) GNPS - Mouse lung lipidome response to a wild type infectious clone of H7N9 Influenza virus and mutants. GNPS, MSV000079542.

Pieter Dorrestein (2016) GNPS - Metabolites of psoriasis vs healthy skin_Pilot study. GNPS, MSV000079558.

Pieter Dorrestein (2017) GNPS - Beauty products and skin microbiome. GNPS, MSV000079-559.

Pieter Dorrestein (2016) GNPS-Cystic Fibrosis_Breath Condensate. GNPS, MSV000079575.

Rob Knight (2016) GNPS_American_84_UK_86_Gut. GNPS, MSV000079598.

James McKerrow (2016) GNPS_Determinants of Chagas disease progression. GNPS, MSV000079615.

Tom Conrad, Pieter Dorrestein, Forest Rohwer and Rob Knight (2016) GNPS Rapid Response CF94. GNPS, MSV000079621.

Pieter Dorrestein (2016) GNPS-Cystic Fibrosis Lung Tissue Patient1. GNPS, MSV000079652.

UCSD Systems Mass Spectrometry class (2016) GNPS - System Wide MS course. GNPS, MSV000079697.

UCSD Systems Mass Spectrometry class (2016) GNPS - System Wide MS course (QqQ). GNPS, MSV000079699.

UCSD Systems Mass Spectrometry class (2016) GNPS System Wide MS course Group 3 Person 1 through 3. GNPS, MSV000079707.

UCSD Systems Mass Spectrometry class (2016) GNPS-System Wide MS Course Group 1 Person 1.1 through 1.4. GNPS, MSV000079710.

UCSD Systems Mass Spectrometry class (2016) GNPS-System Wide MS Course Group 4 Person 4.1 through 4.4. GNPS, MSV000079712.

UCSD Systems Mass Spectrometry class (2016) GNPS-System Wide MS Course Person Group 2. GNPS, MSV000079718.

UCSD Systems Mass Spectrometry class (2016) GNPS-System Wide MS Course Group 5 Person 1 through 2. GNPS, MSV000079719.

Brigid Boland (2016) GNPS - IBD Biobank Stool Samples - C18-LC-MS/MS Positive Polarity. GNPS, MSV000079777.

Larry Smarr (2016) GNPS - Human Stool Study from Human Stool. GNPS, MSV000079778.

Pieter Dorrestein (2016) GNPS - Mouse dataset. GNPS, MSV000079787.

Pieter Dorrestein (2016) GNPS - Human dataset. GNPS, MSV000079788.

Pieter Dorrestein (2016) GNPS - Trace detection_10 volunteers_4 months later_Time2. GNPS, MSV000079825.

Purna Kashyap (2016) GNPS - Individualized responses of gut microbiota to dietary intervention modeled in humanized mice. GNPS, MSV000079922.

Pieter Dorrestein (2016) GNPS 3D Germ Free and Specific Pathogen Free Mice. GNPS, MSV000079949.

Charles Ansong (2016) GNPS - Lipidomics reveals dramatic lipid compositional changes in the maturing postnatal lung. GNPS, MSV000080000.

Pieter Dorrestein (2016) GNPS - Forensic study, swabs of hands and objects, 80 volunteers. GNPS, MSV000080030.

Pieter Dorrestein (2016) GNPS - Metabolites of psoriasis vs healthy skin, volonteer 2 to 5. GNPS, MSV000080031.

Pieter Dorrestein (2016) GNPS - Global Stool Dataset - C18-LC-MS/MS Positive polarity. GNPS, MSV000080050.

Christina Chambers (2016) GNPS - Breast Milk metabolomics - Solvent Screen. GNPS, MSV000080074.

Christina Chambers (2017) GNPS - Breast Milk metabolomics - Solvent Screen - take 2. GNPS, MSV000080117.

David Relman (2016) GNPS -3DMouth - A spatial analysis of the human mouth metabolites. GNPS, MSV000080167.

David Relman (2016) GNPS -3DMouth - Triple Quad dataset. GNPS, MSV000080170.

Pieter Dorrestein and Rob Knight (2017) GNPS_AMG_500_09202016. GNPS, MSV00008-0179.

Gabriel Haddad (2017) GNPS - Mouse Blood Metabolites. GNPS, MSV000080180.

Gabriel Haddad (2017) GNPS - Mouse Mesenteric Lymph Node Metabolites. GNPS, MSV000080181.

Gabriel Haddad (2017) GNPS - Mouse Lung Tissue Extract. GNPS, MSV000080182.

Gabriel Haddad (2017) GNPS - Mouse Fecal Metabolites. GNPS, MSV000080183.

Pieter Dorrestein and Rob Knight (2017) GNPS_AMG_Veg_500_09222016. GNPS, MSV0-00080186.

Pieter Dorrestein and Rob Knight (2017) GNPS_AMG_500_antibiotics_09222016. GNPS, MSV000080187.

Michael Fischbach (2016) GNPS_hair_follicles_propimycin_detection. GNPS, MSV00008-0335.

Pieter Dorrestein (2017) GNPS Rapid Response non-CF1. GNPS, MSV000080377.

Cliff Kapono (2017) GNPS_Surferbiome_Ireland. GNPS, MSV000080442.

Cliff Kapono (2017) GNPS_Surferbiome_London. GNPS, MSV000080443.

Cliff Kapono (2017) GNPS_Surferbiome_Morocco. GNPS, MSV000080444.

Michael Karin (2017) GNPS_Steatohepatitis progresses to cancer via immunosuppression of HCC. GNPS, MSV000080558.

Chris Turck (2017) GNPS - Turck - Mouse Sleep Disruption Study and Intervation. GNPS, MSV000080574.

Reza Salek and Jules Griffin (2012) A metabolomic study of urinary changes in type 2 diabetes in human compared to the control group. MetaboLights, MTBLS1.

Christoph Böttcher and Steffen Neumann (2012) Comparative LC/MS-based profiling of silver nitrate-treated Arabidopsis thaliana leaves of wild-type and cyp79B2 cyp79B3 double knockout plants. MetaboLights, MTBLS2.

Joachim Kopka (2015) Metabolomic profiles in long day, after bolting, single leaf analysis, Arabidopsis thaliana. MetaboLights, MTBLS7.

Joachim Kopka (2015) Metabolomic profiles in long day, before bolting, single leaf analysis, Arabidopsis thaliana. MetaboLights, MTBLS11.

Joachim Kopka (2015) Metabolomic profiles in short day, after bolting, single leaf analysis, Arabidopsis thaliana. MetaboLights, MTBLS12.

Joachim Kopka (2015) Metabolomic profiles in short day, before bolting, single leaf analysis, Arabidopsis thaliana. MetaboLights, MTBLS13.

Habtom Ressom (2013) Utilization of Metabolomics to Identify Serum Biomarkers for Hepatocellular Carcinoma in Patients with Liver Cirrhosis. MetaboLights, MTBLS17.

Habtom Ressom (2013) LC-MS Based Serum Metabolomics for Identification of Hepatocellular Carcinoma Biomarkers in Egyptian Cohort. MetaboLights, MTBLS19.

Christophe Junot and Aurélie Roux (2013) Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a LTQ-Orbitrap mass spectrometer. MetaboLights, MTBLS20.

Diego Sanchez and Joachim Kopka (2015) Mining for metabolic responses to long-term salt stress: a case study on Arabidopsis thaliana Col-0 (A). MetaboLights, MTBLS22.

Wolfram Gronwald (2012) Analysis of human urine reveals metabolic changes related to the development of acute kidney injury following cardiac surgery. MetaboLights, MTBLS24.

Harald Köfeler and Friedrich Spener (2014) Lipidomic analysis of lipid droplets from murine hepatocytes reveals distinct signatures for nutritional stress. MetaboLights, MTBLS26.

Emily Mackay, Aalim Weljie and Oliver Bathe (2016) Fatty acid synthesis in colorectal cancer: characterization of lipid metabolism in serum, tumour, and normal host tissues. MetaboLights, MTBLS27.

Ewy Mathe, Andrew Patterson, Majda Haznadar, Soumen Manna, Kristopher Krausz, Elise Bowman, Peter Shields, Jeffrey Idle, Philip Smith, Anami Katsuhiro, Dickran Kazandjian, Frank Gonzalez and Curtis Harris (2014) Non-invasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. MetaboLights, MTBLS28.

Christopher Newgard, Bert O'Malley and Brian York (2013) Tissue- and Pathway-Specific Metabolomic Profiles of the Steroid Receptor Coactivator (SRC) family. MetaboLights, MTBLS30. Alan Aderem (2013) Lipid mediators of inflammation in BALF 3-19 days after infection with influenza. MetaboLights, MTBLS31.

Alan Aderem (2013) Lipid mediators of inflammation in BALF 6-19 days after infection with influenza. MetaboLights, MTBLS32.

Alan Aderem (2013) Lipid mediators of inflammation in BALF 3-11 days after infection with influenza. MetaboLights, MTBLS33.

Alan Aderem (2013) Lipid mediators of inflammation in BALF 3-13 days after infection with influenza. MetaboLights, MTBLS34.

Miyako Kusano and Atsushi Fukushima (2014) Unbiased characterization of genotype-dependent metabolic regulations by metabolomic approach in Arabidopsis thaliana. MetaboLights, MTBLS40.

Diego Sanchez and Joachim Kopka (2015) Mining for metabolic responses to long-term salt stress: a case study on Arabidopsis thaliana Col-0 (B). MetaboLights, MTBLS41.

Diego Sanchez and Joachim Kopka (2015) Mining for metabolic responses to long-term salt stress: a case study on Arabidopsis thaliana Col-0 (C). MetaboLights, MTBLS42.

Atsushi Fukushima (2014) Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach. MetaboLights, MTBLS45.

Leonardo Tenori (2015) The cardiovascular risk of healthy individuals studied by NMR metabonomics of plasma samples. MetaboLights, MTBLS46.

Atsushi Fukushima (2014) Metabolomic Characterization of Knock-Out Mutants in Arabidopsis - Development of a Metabolite Profiling Database for Knock-Out Mutants in Arabidopsis (MeKO). MetaboLights, MTBLS47.

Ryo Nakabayashi and Kazuki Saito (2013) Metabolome profiling using LC-MS in Arabidopsis overaccumulating and lacking flavonoids. MetaboLights, MTBLS57.

Koen van de Wetering (2013) LC-MS metabolic profiling of mouse plasma and cell culture medium in relation to ABCC6 expression. MetaboLights, MTBLS61.

Thomas Metz (2014) A statistical analysis of the effects of urease pre-treatment on the measurement of the urinary metabolome by gas chromatography–mass spectrometry. MetaboLights, MTBLS71.

Howard Federoff (2014) Plasma Lipidomics for the Identification of Anticedent Memory Impairment. MetaboLights, MTBLS72.

Stephan Schmidt, Steffen Neumann, Diana Trutschel and Ivo Grosse (2014) Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data. MetaboLights, MTBLS74.

Harald Köfeler and Friedrich Spener (2014) The impact of genetic stress by ATGL deficiency on the lipidome of lipid droplets from murine hepatocytes. MetaboLights, MTBLS81.

Romanas Chaleckis, Masahiro Ebe, Tomáš Pluskal, Itsuo Murakami, Hiroshi Kondoh and Mitsuhiro Yanagida (2014) Unexpected similarities between the Schizosaccharomyces and human blood metabolomes, and novel human metabolites (Blood fraction). MetaboLights, MTBLS87.

Romanas Chaleckis, Masahiro Ebe, Tomáš Pluskal, Itsuo Murakami, Hiroshi Kondoh and Mitsuhiro Yanagida (2014) Unexpected similarities between the Schizosaccharomyces and human blood metabolomes, and novel human metabolites (Blood plasma and RBC fractions). MetaboLights, MTBLS88.

Andrea Ganna, Samira Salihovic, Erik Ingelsson and Lars Lind (2014) Large-scale non-targeted serum metabolomics in the Prospective Investigation of the Vasculature in Uppsala Seniors. MetaboLights, MTBLS90.

Mika Hilvo (2014) Monounsaturated fatty acids in serum triacylglycerols are associated with response to neoadjuvant chemotherapy in breast cancer patients. MetaboLights, MTBLS92.

Frederik Walter (2014) Large-scale non-targeted serum metabolomics in the Swedish Twin Registry. MetaboLights, MTBLS93.

Thomas Metz (2014) Decreased abundance of type III secretion systeminducing signals in Arabidopsis mkp1 enhances resistance against Pseudomonas syringae. MetaboLights, MTBLS95.

Thomas Metz (2014) A Multi-Omic View of Host-Pathogen-Commensal Interplay in Salmonella-Mediated Intestinal Infection. MetaboLights, MTBLS96.

Zerihun Dame and David Wishart (2015) The Human Saliva Metabolome. MetaboLights, MTBLS100.

Frank Schmidt (2015) Comparative analysis of the adaptation of Staphylococcus aureus to internalization by different types of human non-professional phagocytic host cells (NMR assay). MetaboLights, MTBLS102.

Sara Samino and Oscar Yanes (2015) Long-term health risks in PCOS occur prematurely with serum markers of oxidative stress impacting HDL maturation through oxidation of methionine residues in apolipoprotein A1. MetaboLights, MTBLS103.

Michael Kennedy and Bo Wang (2014) NMR based metabolomics study of Y2 receptor activation by neuropeptide Y in the SK-N-BE2 human neuroblastoma cell line. MetaboLights, MTBLS104.

Habtom Ressom (2015) GC-MS based plasma metabolomics for HCC biomarker discovery in Egyptian cohort. MetaboLights, MTBLS105.

Wang Dezhen, Qiu Jing, Zhang Ping, Deng Nian, Wang Xinru, Wang Yao and Zhu Wentao (2015) Large-scale non-targeted plasma metabolomics in the Uppsala Longitudinal Study of Adult Men. MetaboLights, MTBLS124.

Itsuo Murakami, Romanas Chaleckis, Tomas Pluskal, Ken Ito, Kousuke Hori, Masahiro Ebe, Mitsuhiro Yanagida and Hiroshi Kondoh (2014) Distribution of RESV and its metabolite peaks in mouse tissues after oral and skin administration. MetaboLights, MTBLS125.

Itsuo Murakami, Romanas Chaleckis, Tomáš Pluskal, Ken Ito, Kousuke Hori, Masahiro Ebe, Mitsuhiro Yanagida and Hiroshi Kondoh (2014) Absorption efficiency of RESV through mouse skin using 3 bases in different tissues. MetaboLights, MTBLS126.

Itsuo Murakami, Romanas Chaleckis, Tomáš Pluskal, Ken Ito, Kousuke Hori, Masahiro Ebe, Mitsuhiro Yanagida and Hiroshi Kondoh (2014) Resveratrol metabolism in HepG2 (human

hepatocytes), HaCaT (human keratinocytes), and C2C12 (mouse myoblasts). MetaboLights, MTBLS127.

Sanjay Swarup, Amit Rai and Shivshankar Umashankar (2016) Coordinate Regulation of Metabolite Glycosylation and Stress Hormone Biosynthesis by TT8 in Arabidopsis. MetaboLights, MTBLS129.

Christina Ranninger, Marc Rurik, Alice Limonciel, Paul Jennings, Oliver Kohlbacher and Christian Huber (2015) Metabolome analysis via an HPLC-ESI-MS-based experimental and computational pipeline for chronic nephron toxicity profiling. MetaboLights, MTBLS140.

Sam Ansari, Stephanie Boue and Kim Ekroos (2015) Comprehensive systems biology analysis of a seven-month cigarette smoke inhalation study in C57BL/6 mice. MetaboLights, MTBLS143.

Hemi Luan (2015) Pregnancy-Induced Metabolic Phenotype Variations in Maternal Plasma. MetaboLights, MTBLS146.

Kathleen Vermeersch and Mark Styczynski (2014) Distinct metabolic responses of an ovarian cancer stem cell line using gas chromatography mass spectrometry. MetaboLights, MTBLS150.

Kathleen Vermeersch and Mark Styczynski (2015) ovarian serous adenocarcinoma cell line. MetaboLights, MTBLS152.

Martina Vermathen (2015) 1H HR-MAS NMR Based Metabolic Profiling of Cells in Response to Treatment with a Hexacationic Ruthenium Metallaprism as Potential Anticancer Drug. MetaboLights, MTBLS156.

Susann Mönchgesang, Nadine Strehmel, Stephan Schmidt, Lore Westphal, Franziska Taruttis, Erik Müller, Siska Herklotz, Steffen Neumann and Dierk Scheel (2016) Natural variation of root exudates in Arabidopsis thaliana-linking metabolomic and genomic data. MetaboLights, MTBLS160.

Christopher Armstrong and Paul Gooley (2015) NMR-based metabolic profiling of Chronic Fatigue Syndrome patients. MetaboLights, MTBLS161.

Michael Eiden and Julian Griffin (2015) Mechanistic insights revealed by lipid profiling in monogenic insulin resistance syndromes. MetaboLights, MTBLS162.

Mesut Bilgin and Andrej Shevchenko (2015) Quantitative profiling of endocannabinoids in lipoproteins by LC–MS/MS. MetaboLights, MTBLS163.

Simona Cristescu and Phil Brown (2016) Searching for metabolic changes in urine headspace composition as an effect of strenuous walking. MetaboLights, MTBLS164.

Wang Dezhen, Qiu Jing, Zhang Ping, Deng Nian, Wang Xinru, Wang Yao and Zhu Wentao (2015) 1H NMR and LC–MS/MS based urine metabolomic investigation of the subacute effects of HBCD in mice. MetaboLights, MTBLS166.

Stephan Schmidt, Steffen Neumann, Diana Trutschel and Ivo Grosse (2015) Metabolite Profiling of wildtype and overexpression Arabidopsis thaliana. MetaboLights, MTBLS169.

Veronica Ghini, Florian Unger, Leonardo Tenori, Paola Turano, Hartmut Juhl and Kerstin David (2015) Metabolomics profiling of pre-and post-anesthesia plasma samples obtained via Ficoll separation. MetaboLights, MTBLS172.

Xinzhu Wang, James West, Andrew Murray and Julian Griffin (2015) Comprehensive Metabolic Profiling of Age-Related Mitochondrial Dysfunction in the High-Fat-Fed ob/ob Mouse Heart. MetaboLights, MTBLS173.

Ewa Gralka and Paola Turano (2015) Multi-omic profiles of human nonalcoholic fatty liver disease tissue highlight heterogenic phenotypes. MetaboLights, MTBLS174.

Janina Oetjen and Theodore Alexandrov (2016) Benchmark datasets for 3D MALDI- and DESI-Imaging Mass Spectrometry. MetaboLights, MTBLS176.

Sean Palecek and Vijesh Bhute (2015) Metabolic responses induced by DNA damage and poly (ADP-ribose) polymerase (PARP) inhibition in MCF-7 cells. MetaboLights, MTBLS177.

Myung-Hee Nam and Kyoungwon Cho (2015) UPLC-Q/TOF-MS based metabolomics of Human Serum. MetaboLights, MTBLS178.

Carsten Kuhl, Christoph Böttcher and Steffen Neumann (2015) Supplemental data for CAMERA, an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. MetaboLights, MTBLS188.

Aiping Li and Zhenyu Li (2017) Comparison of Two Different Astragali Radix by 1H NMR based Metabolomic Approach. MetaboLights, MTBLS189.

Jessica Danaher and Matthew Cooke (2015) The use of metabolomics to monitor simultaneous changes in metabolic variables following supramaximal low volume high intensity exercise. MetaboLights, MTBLS191.

Koen van de Wetering and Robert Jansen (2015) ATP-binding Cassette Sub-family C Member 5 (ABCC5) Functions as an Efflux Transporter of Glutamate Conjugates and Analogs. MetaboLights, MTBLS197.

Maria Ulaszewska (2017) Urinary metabolomic profiling to identify biomarkers of a flavonoid-rich and flavonoid-poor fruits and vegetables diet in adults: the FLAVURS trial. MetaboLights, MTBLS198.

Farshad Farshidfar and Omair Sarfaraz (2017) Temporal characterization of serum metabolite signatures in lung cancer patients undergoing treatment. MetaboLights, MTBLS200.

Jordi Capellades and Oscar Yanes (2015) geoRge: A Computational Tool To Detect the Presence of Stable Isotope Labeling in LC/MS-Based Untargeted Metabolomics. MetaboLights, MTBLS213.

Lei Zhou and Chan Eric (2015) Global Metabonomic and Proteomic Analysis of Human Conjunctival Epithelial Cells (IOBA-NHC) in Response to Hyperosmotic Stress. MetaboLights, MTBLS214.

Chun-Xue Zhou, Dong-Hui Zhou, Hany Elsheikha, Yu Zhao, Xun Suo and Xing-Quan Zhu (2017) Metabolomic Profiling of Mice Serum during Toxoplasmosis Progression Using Liquid Chromatography-Mass Spectrometry. MetaboLights, MTBLS216.

Chun-Xue Zhou, Dong-Hui Zhou, Hany Elsheikha, Guang-Xue Liu, Xun Suo and Xing-Quan Zhu (2017) Global Metabolomic Profiling of Mice Brains following Experimental Infection with the Cyst-Forming Toxoplasma gondii. MetaboLights, MTBLS217.

Christoph Magnes, Harald Sourij and Sophie Narath (2015) An untargeted metabolomics approach to characterize short-term and long-term metabolic changes after bariatric surgery. MetaboLights, MTBLS218.

Michael Rutzler and Peter Spegel (2015) Phenotypic analysis of AQP9 KO in the C57Bl6 db/db background. MetaboLights, MTBLS219.

Timothy Zacharewski and Rance Nault (2017) Targeted metabolomic analysis in hepatic extracts, serum, and urine of female mice gavaged with TCDD every 4 days for 28 days. MetaboLights, MTBLS225.

Mark Styczynski and Suganthagunthalam Dhakshinamoorthy (2015) Metabolomics identifies the intersection of phosphoethanolamine with menaquinone- triggered apoptosis in an in vitro model of leukemia. MetaboLights, MTBLS226.

Manfred Wuhrer and Stephanie Holst (2017) N-glycosylation Profiling of Colorectal Cancer Cell Lines Reveals Association of Fucosylation with Differentiation and Caudal Type Homebox 1 (CDX1)/Villin mRNA Expression. MetaboLights, MTBLS227.

Christina Ranninger, Lukas Schmidt, Marc Rurik, Alice Limonciel, Paul Jennings, Oliver Kohlbacher and Christian Huber (2016) Improving global feature detectabilities through scan range splitting for untargeted metabolomics by high-performance liquid chromatography-Orbitrap mass spectrometry. MetaboLights, MTBLS233.

Erhan Kenar and Oliver Kohlbacher (2015) Automated Label-free Quantification of Metabolites from Liquid Chromatography–Mass Spectrometry Data (Plasma). MetaboLights, MTBLS234.

Ulrich Gunther, Jacob Bjerrum, Yulan Wang, Fuhua Hao and Christian Ludwig (2017) Metabonomics of human fecal extracts characterize ulcerative colitis, Crohn's disease and healthy individuals. MetaboLights, MTBLS237.

Susana Sanchez-Tena, Marta Cascante, Ulrich Gunther and Michelle Thompson (2017) Maslinic acid-enriched diet decreases intestinal tumorigenesis in Apc(Min/+) mice through transcriptomic and metabolomic reprogramming. MetaboLights, MTBLS240.

Bernd Schultes and Claudio Luchinat (2017) Metabolomic fingerprint of severe obesity is dynamically affected by bariatric surgery in a procedure-dependent manner. MetaboLights, MTBLS242.

Min He and Slavik Koval (2015) Collagen induced arthritis in dba/1j mice associates with oxylipin changes in plasma. MetaboLights, MTBLS243.

Maria Vinaixa (2017) Positional Enrichment by Proton Analysis (PEPA): a One-Dimensional 1H-NMR Approach for 13C Stable Isotope Tracer Studies in Metabolomics. MetaboLights, MTBLS247.

Johannes Cornelius Schoeman and Jun Hou (2016) Oxylipin profilling of clincal phase defined chronic hepatitis B samples. MetaboLights, MTBLS253.

Romanas Chaleckis, Hiroshi Kondoh, Mitsuhiro Yanagida, Itsuo Murakami and Junko Takada (2016) Individual variability in human blood metabolites identifies age-related differences - constant blood metabolite levels during 24h in 4 individuals. MetaboLights, MTBLS264.

Romanas Chaleckis, Hiroshi Kondoh, Mitsuhiro Yanagida, Itsuo Murakami and Junko Takada (2016) Individual variability in human blood metabolites identifies age-related differences (30 persons, whole blood data). MetaboLights, MTBLS265.

Romanas Chaleckis, Hiroshi Kondoh, Mitsuhiro Yanagida, Itsuo Murakami and Junko Takada (2016) Individual variability in human blood metabolites identifies age-related differences (30 persons, plasma data). MetaboLights, MTBLS266.

Romanas Chaleckis, Hiroshi Kondoh, Mitsuhiro Yanagida, Itsuo Murakami and Junko Takada (2016) Individual variability in human blood metabolites identifies age-related differences (30 persons, RBC data). MetaboLights, MTBLS267.

James McKenzie and Zoltan Takats (2016) Metabolic phenotyping of ex-vivo breast samples by DESI mass spectrometry imaging. MetaboLights, MTBLS273.

Christina Ranninger, Marc Rurik, Oliver Kohlbacher and Christian Huber (2016) Multi-omics toxicity profiling of engineered nanomaterials. MetaboLights, MTBLS277.

Johannes Cornelius Schoeman and Jun Hou (2016) Lipidomics characterisation of clincal phase defined chronic hepatitis B samples. MetaboLights, MTBLS279.

Johannes Cornelius Schoeman and Jun Hou (2016) Amino acid and acyl-carnitine profilling of clincal phase defined chronic hepatitis B samples. MetaboLights, MTBLS280.

James McKenzie and Zoltan Takats (2016) 3D DESI mass spectrometry imaging of 51 serial sections from a human liver metastasis. MetaboLights, MTBLS282.

James McKenzie and Zoltan Takats (2016) Analysis of colorectal adenocarcinoma tissue samples by DESI mass spectrometry imaging. MetaboLights, MTBLS289.

Matej Oresic (2016) Genome-scale study reveals reduced metabolic adaptability in patients with non-alcoholic fatty liver disease. MetaboLights, MTBLS298.

Justin van der Hooft, Karl Burgess, Michael Barrett and Sandosh Padmanabhan (2016) Molecular networking coupled to high-resolution mass spectrometry fragmentation reveals a multitude of urinary antihypertensive drug metabolites. MetaboLights, MTBLS307.

Andrew Palmer and Theodore Alexandrov (2017) FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry imaging MS. MetaboLights, MTBLS313.

Saskia Decuypere (2016) Towards improving point-of-care diagnosis of non-malaria febrile illness: a metabolomics approach. MetaboLights, MTBLS315.

Julian Griffin (2017) Mechanistic insights revealed by lipid profiling in monogenic insulin resistance syndromes (Mouse data). MetaboLights, MTBLS327.

Vijesh Bhute and Sean Palecek (2016) The Poly (ADP-Ribose) Polymerase inhibitor Veliparib and Radiation Cause Significant Cell Line Dependent Metabolic Changes in Breast Cancer Cells. MetaboLights, MTBLS337.

Susann Mönchgesang, Nadine Strehmel, Diana Trutschel, Lore Westphal, Steffen Neumann and Dierk Scheel (2016) Plant-to-plant variability in root metabolite profiles of 19 Arabidopsis thaliana accessions is substance-class dependent. MetaboLights, MTBLS338.

Nadine Strehmel, Susann Mönchgesang, Siska Herklotz, Sylvia Krüger, Jörg Ziegler and Dierk Scheel (2016) Piriformospora indica stimulates root metabolism of Arabidopsis thaliana. MetaboLights, MTBLS341.

Tom Metz (2016) Relative influence of early life exposure, host genetics and diet on gut microbiome and metabolome. MetaboLights, MTBLS345.

Yuen Kwok-Yung (2016) Lipid metabolites as potential diagnostic and prognostic biomarkers for acute community acquired pneumonia. MetaboLights, MTBLS354.

Sam Ansari, Bjoern Titz and Terhi Vihervaara (2017) Chronic Obstructive Pulmonary Disease (COPD) Biomarker Identification Study: Serum Lipidomics. MetaboLights, MTBLS358.

Emmanuel Minet, Manuja Kaluarachchi, Claire Boulange, Isabel Garcia-Perez and John Lindon (2017) Multiplatform serum metabolic phenotyping combined with pathway mapping to identify biochemical differences in smokers. MetaboLights, MTBLS364.

Charles Brenner (2016) Nicotinamide Riboside is Uniquely Bioavailable in People and Mice. MetaboLights, MTBLS368.

Emmanuel Minet, Manuja Kaluarachchi, Claire Boulange, Isabel Garcia-Perez and John Lindon (2017) Multiplatform serum metabolic phenotyping combined with pathway mapping to identify biochemical differences in smokers. MetaboLights, MTBLS374.

Mitsuharu Matsumoto and Hiroyuki Yamamoto (2017) Appropriate taxonomic classification of intestinal microbiome by 16S rRNA gene targeting metagenomic approach. MetaboLights, MTBLS376.

Andrew Palmer, Prasad Phapale and Theodore Alexandrov (2016) FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry imaging MS. MetaboLights, MTBLS378.

James McKenzie and Zoltan Takats (2017) Desorption electrospray ionisation mass spectrometry imaging of esophageal lymph mode metastases. MetaboLights, MTBLS385.

Etienne Thevenot and Christophe Junot (2017) Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses ('Sacurine' data set). MetaboLights, MTBLS404.

Mogens Johannsen (2016) Untargeted Metabolomics Reveals a Mild Impact of Remote Ischemic Conditioning on the Plasma Metabolome and α-Hydroxybutyrate as a Possible Cardioprotective Factor and Biomarker of Tissue Ischemia. MetaboLights, MTBLS414.

Veronica Ghini, Mattia Di Nunzio, Leonardo Tenori, Veronica Valli, Francesca Danesi, Francesco Capozzi, Luchinat Claudio and Alessandra Bordoni (2017) Evidence of a DHA signature in the lipidome and metabolome of human hepatocytes. MetaboLights, MTBLS419.

Tokuwa Kanno, A. Mason, Jocelyn Choo, Nur Masirah Zain, Lex Leong, Guy Abell, Julie Keeble, Kinneth Bruce and Geraint Rogers (2017) Divergent relationships between faecal microbiota and metabolome following distinct antibiotic-induced disruptions. MetaboLights, MTBLS422.

Alessia Vignoli, Leonardo Tenori, Angelo Di Leo and Luchinat Claudio (2017) Serum metabolomic profiles identify ER-positive early breast cancer patients at increased risk of disease recurrence in a multicentre population. MetaboLights, MTBLS424.

Erin Baker and Tom Metz (2017) SPE-IMS-MS: An automated platform for sub-sixty second surveillance of endogenous metabolites and xenobiotics in biofluids. MetaboLights, MTBLS427.

Tobias Kind (2013) Fatb Induction Experiment (FatBIE). Metabolomics Workbench, ST000001, V1.

Amber Hartman and Oliver Fiehn (2013) Intestinal Samples II pre/post transplantation. Metabolomics Workbench, ST000002, V1.

Eoin Fahy (2013) Lipidomics studies on NIDDK / NIST human plasma samples. Metabolomics Workbench, ST000004, V1.

Chuck Burant (2013) Mixed meal tolerance. Metabolomics Workbench, ST000009, V1.

Venkat Keshamouni (2013) Lung Cancer Cells 4. Metabolomics Workbench, ST000010, V1.

Simon Thompson (2013) African Metabolomics. Metabolomics Workbench, ST000011, V1.

Tobias Kind (2013) Mutation Study. Metabolomics Workbench, ST000013, V1.

Scott McDonnell (2013) NPM-ALK metabolic regulation. Metabolomics Workbench, ST000016, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(DuraSTAT-Cecal). Metabolomics Workbench, ST000026, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(DuraSTAT-Liver). Metabolomics Workbench, ST000027, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(DuraSTAT-Urine). Metabolomics Workbench, ST000028, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(TranSTAT-Cecal). Metabolomics Workbench, ST000029, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(TranSTAT-Liver). Metabolomics Workbench, ST000030, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(TranSTAT-Serum). Metabolomics Workbench, ST000031, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(NOD-Cecal). Metabolomics Workbench, ST000032, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(NOD-Liver). Metabolomics Workbench, ST000033, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(NOD-Serum). Metabolomics Workbench, ST000034, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(EstroSTAT-Liver). Metabolomics Workbench, ST000035, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(EstroSTAT-Serum). Metabolomics Workbench, ST000036, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(EstroSTAT-Urine). Metabolomics Workbench, ST000037, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(VGSTAT-Cecal). Metabolomics Workbench, ST000038, V1.

Susan Sumner (2015) Metabolomics Involved in Early Life Antibiotic Exposures(VGSTAT-Liver). Metabolomics Workbench, ST000039, V1.

Maureen Kachman (2014) High PUFA diet in humans. Metabolomics Workbench, ST000041, V1.

Maureen Kachman (2014) BALF Control vs ALI by RPLC-MS. Metabolomics Workbench, ST000042, V1.

Maureen Kachman (2014) MDA-MB-231 cells and p38 gamma knockdown. Metabolomics Workbench, ST000043, V1.

Maureen Kachman (2014) Pilot experiment looking for the existence of certain molecules in pancreatic cancer cells. Metabolomics Workbench, ST000044, V1.

Sreekumaran Nair (2014) Plasma metabolomics: Comparison of non-diabetic controls with T1D patients. Metabolomics Workbench, ST000045, V1.

Ronald Petersen (2014) Identification of altered metabolic pathways in Alzheimer's disease, mild cognitive impairment and cognitively normals using Metabolomics (plasma). Metabolomics Workbench, ST000046, V1.

Ronald Petersen (2014) Identification of altered metabolic pathways in Alzheimer's disease, mild cognitive impairment and cognitively normals using Metabolomics (CSF). Metabolomics Workbench, ST000047, V1.

Susan Sumner (2015) Environmental impact on metabolomics and food allergy. Metabolomics Workbench, ST000056, V1.

Oliver Fiehn (2017) Combined Metabolomics and Lipidomics of Type 1 Diabetes (GCMS). Metabolomics Workbench, ST000057, V1.

Oliver Fiehn (2014) Metabolite changes associated with methionine stress sensitivity of cancer (GC TOF MS analysis). Metabolomics Workbench, ST000058, V1.

Oliver Fiehn (2014) Metabolic Profiling of Visceral and Subcutaneous Adipose Tissue from Colorectal Cancer Patients: GC-TOF MS analyis of subcutaneous and visceral adipose tissue samples. Metabolomics Workbench, ST000061, V1.

Oliver Fiehn (2014) Biomarkers for Depression in Human Plasma in a Population Sample. Metabolomics Workbench, ST000062, V1.

Oliver Fiehn (2014) Biomarkers for Depression in Human Cerebrospinal Fluid in a Population Sample. Metabolomics Workbench, ST000063, V1.

Oliver Fiehn (2014) Metabolic Profiling of Visceral and Subcutaneous Adipose Tissue from Colorectal Cancer Patients: GC-TOF MS analyis of serum samples. Metabolomics Workbench, ST000065, V1.

Susan Sumner (2015) Metabolomic profiling of influenza: a 2009 pandemic H1N1 influenza in lean and obese mice (via blood and tissue). Metabolomics Workbench, ST000069, V1.

Susan Sumner (2015) Metabolomic profiling of influenza: a 2009 pandemic H1N1 influenza in lean and obese mice (via tissue). Metabolomics Workbench, ST000070, V1.

Susan Sumner (2015) Metabolomic profiling of influenza: a 2009 pandemic H1N1 influenza in lean and obese mice (via Urine). Metabolomics Workbench, ST000071, V1.

Susan Sumner (2015) Genetic effects of high fat diet on mouse fecal metabolomics. Metabolomics Workbench, ST000074, V1.

Oliver Fiehn (2014) Combined Metabolomics and Lipidomics of Type 1 Diabetes (QTOF). Metabolomics Workbench, ST000075, V1.

Jun Yan (2014) A549 Cell Study. Metabolomics Workbench, ST000076, V1.

Oliver Fiehn (2014) Metabolite changes associated with methionine stress sensitivity of cancer (CSH QTOF MS analysis). Metabolomics Workbench, ST000077, V1.

Oliver Fiehn (2014) Metabolic Profiling of Visceral and Subcutaneous Adipose Tissue from Colorectal Cancer Patients: UHPLC-QTOF MS analyis of subcutaneous and visceral adipose tissue samples. Metabolomics Workbench, ST000081, V1.

Oliver Fiehn (2014) Metabolic Profiling of Visceral and Subcutaneous Adipose Tissue from Colorectal Cancer Patients: UHPLC-QTOF MS analyis of serum samples. Metabolomics Workbench, ST000082, V1.

Thomas Metz (2014) A Multi-Omic View of Host-Pathogen-Commensal Interplay in Salmonella-Mediated Intestinal Infection. Metabolomics Workbench, ST000083, V1.

Oliver Fiehn (2014) A study of changes in lipid metabolism of ovarian cancer cells co-cultured with adipocytes: GC-TOF MS analysis. Metabolomics Workbench, ST000087, V1.

Oliver Fiehn (2014) A study of changes in lipid metabolism of ovarian cancer cells co-cultured with adipocytes: UHPLC-QTOF MS analysis. Metabolomics Workbench, ST000089, V1.

Maureen Kachman (2014) Comparison of caloric restriction vs medications that prolong life (Untargeted). Metabolomics Workbench, ST000090, V1.

Sreekumaran Nair (2014) Quantitative Metabolomics by 1H-NMR and LC-MS/MS Confirms Altered Metabolic Pathways in Diabetes. Metabolomics Workbench, ST000091, V1.

Thomas Metz (2014) A statistical analysis of the effects of urease pre-treatment on the measurement of the urinary metabolome by gas chromatography-mass spectrometry. Metabolomics Workbench, ST000092, V1.

Pratima Karnik (2014) Metabolomics Analysis of Frontal Fibrosing Alopecia. Metabolomics Workbench, ST000093, V1.

John Newman (2015) Dysfunctional lipid metabolism underlies the effect of perinatal DDT exposure on the development of metabolic syndrome. Metabolomics Workbench, ST000095, V1.

John Newman (2015) A study of changes in lipid metabolism of ovarian cancer cells co-cultured with adipocytes: UPLC-QTRAP MS analysis. Metabolomics Workbench, ST000096, V1.

Arthur Edison (2015) NMR analysis of DMD mouse serum. Metabolomics Workbench, ST000099, V1.

Susan Sumner (2016) Factors for Epigenetic Silencing of Lung Cancer Genes. Metabolomics Workbench, ST000104, V1.

Maureen Kachman (2015) SCOR Metabolomics. Metabolomics Workbench, ST000105, V1.

Maureen Kachman (2015) IWMS Study 1:Weight comparison of obese and lean patients. Metabolomics Workbench, ST000106, V1.

Teresa Fan (2015) SIRM Analysis of human P493 cells under hypoxia in [U-13C/15N] labeled Glutamine medium (Both positive and ion mode FTMS). Metabolomics Workbench, ST000110, V1.

Teresa Fan (2015) Study of biological variation in PC9 cell culture. Metabolomics Workbench, ST000111, V1.

Teresa Fan (2015) SIRM Analysis of human P493 cells under hypoxia in [U-13C/15N] labeled Glutamine medium (Positive ion mode FTMS). Metabolomics Workbench, ST000113, V1.

Teresa Fan (2015) SIRM Analysis of human P493 cells under hypoxia in [U-13C] labeled Glucose medium. Metabolomics Workbench, ST000114, V1.

Sreekumaran Nair (2015) Impact of insulin deprivation and treatment on sphingolipid distribution in different muscle subcellular compartments of streptozotocin-diabetic C57Bl/6 mice. Metabolomics Workbench, ST000115, V1.

Charles Evans (2015) Impact of anesthesia and euthanasia on metabolomics of mammalian tissues: studies in a C57BL/6J mouse model. Metabolomics Workbench, ST000121, V1.

Nilesh Gaikwad (2014) Perinatal DDT causes dysfunctional lipid metabolism underlying metabolic syndrome. Metabolomics Workbench, ST000122, V1.

Jianzhi Hu (2015) 1H NMR Metabolomics Study of Metastatic Melanoma in C57BL/6J Mouse Spleen. Metabolomics Workbench, ST000133, V1.

Meden Isaac-Lam (2015) Monitoring In Vitro Response of Selenium-Treated Prostate Cells by 1H NMR Spectroscopy. Metabolomics Workbench, ST000134, V1.

Adrew Patterson (2015) LCMS analysis of Bile Acids. Metabolomics Workbench, ST000135, V1.

Andreea Geamanu (2015) Metabolomics in sarcoidosis. Metabolomics Workbench, ST000137, V1.

Rhonda Cooper-DeHoff (2016) Targeted LC MS of acylcarnitines: TLCMS. Metabolomics Workbench, ST000138, V1.

Candice Ulmer (2016) Cell Rinsing Solution Comparison. Metabolomics Workbench, ST000140, V1.

Teresa Fan (2015) H1299 13C-labeled Cell Study. Metabolomics Workbench, ST000142, V1.

Maureen Kachman (2015) Primary T Cell Baseline (Donor 5) - II. Metabolomics Workbench, ST000144, V1.

Maureen Kachman (2015) Primary T Cell Noxa Knockdown (Donor 8). Metabolomics Workbench, ST000145, V1.

Maureen Kachman (2015) Noxa regulation of malate aspartate shuttle. Metabolomics Workbench, ST000146, V1.

Maureen Kachman (2015) 13C targeted metabolomics. Metabolomics Workbench, ST000147, V1.

Sreekumaran Nair (2015) High Insulin Combined With Essential Amino Acids Stimulates Skeletal Muscle Mitochondrial Protein Synthesis While Decreasing Insulin Sensitivity in Healthy Humans. Metabolomics Workbench, ST000149, V1.

Mary Cloud Ammons (2015) Association of Metabolic Profile and Microbiome in Chronic Pressure Ulcer Wounds. Metabolomics Workbench, ST000150, V1.

Maureen Kachman (2015) Primary T Cell Noxa Knockdown (Donor 8)-II. Metabolomics Workbench, ST000153, V1.

Maureen Kachman (2015) Use of Aspartate Dehydrogenase by cancer cells. Metabolomics Workbench, ST000154, V1.

Maureen Kachman (2016) Human fecal bile acid profiles before and after fecal transplant. Metabolomics Workbench, ST000158, V1.

Maureen Kachman (2015) U13C-Glutamine and U13C-Glucose Flux Analysis (MFA SiHa B16F10). Metabolomics Workbench, ST000159, V1.

Maureen Kachman (2015) Basic Metabolism Studies (wt and TPA null). Metabolomics Workbench, ST000160, V1.

Maureen Kachman (2015) Plasma Nucleotide/adenosine concentrations (Human AxP Batch 3). Metabolomics Workbench, ST000161, V1.

Karan Uppal (2015) Metabolomics of 50 healthy humans and common marmosets (MetabNet). Metabolomics Workbench, ST000163, V1.

Xin Li (2016) Metabolomic analysis of normal and diabetic mouse bone marrow under PBS or metformin treatment. Metabolomics Workbench, ST000164, V1.

Yong-Mei Cha (2015) Cardiac Resynchronization Therapy Induces Adaptive Metabolic Transitions in the Metabolomic Profile of Heart Failure. Metabolomics Workbench, ST000166, V1.

Sreekumaran Nair (2015) Effect of Insulin Sensitizer Therapy on Amino Acids and Their Metabolites. Metabolomics Workbench, ST000168, V1.

Maureen Kachman (2015) Baby Mice Hyperoxia treatment (G64-N.H-7d14d). Metabolomics Workbench, ST000169, V1.

Maureen Kachman (2015) cell metabolomics (metabolic phenotypes of a clock mutant mouse). Metabolomics Workbench, ST000171, V1.

Maureen Kachman (2015) THP1 Human Monocyte cells Project A (part I). Metabolomics Workbench, ST000172, V1.

Maureen Kachman (2015) cell and liver metabolomics. Metabolomics Workbench, ST000176, V1.

Maureen Kachman (2016) Bone Marrow Metabolomics (part I). Metabolomics Workbench, ST000182, V1.

Maureen Kachman (2016) Murine apyrase. Metabolomics Workbench, ST000188, V1.

Maureen Kachman (2016) Hirschprung Enterocolitis SCFA. Metabolomics Workbench, ST000189, V1.

Maureen Kachman (2015) TRF Microbiome Study NA/FA/FT/CDKO 4-14. Metabolomics Workbench, ST000192, V1.

Maureen Kachman (2015) BAF60a LKO Liver specific knockout Bile acid. Metabolomics Workbench, ST000193, V1.

Maureen Kachman (2016) Lanthanide-mineral induced alteration of bile acid metabolism in a murine model of steatohepatitis. Metabolomics Workbench, ST000194, V1.

Maureen Kachman (2016) ACSL5 Whole Body Bile Acids. Metabolomics Workbench, ST000195, V1.

Maureen Kachman (2015) Murine gastrointestinal bile acid profiles before and after antibiotics. Metabolomics Workbench, ST000196, V1.

Maureen Kachman (2015) Liver and Plasma metaboites for 13C-glucose load in wild type, LIRKO and LIRFKO mice. Metabolomics Workbench, ST000198, V1.

Maureen Kachman (2015) IDH1 and Glioma knockdown idh1 (part II). Metabolomics Workbench, ST000199, V1.

Maureen Kachman (2015) Bone Marrow Metabolomics (part II). Metabolomics Workbench, ST000201, V1.

Maureen Kachman (2015) THP1 Human Monocyte cells Project A (part II). Metabolomics Workbench, ST000202, V1.

Maureen Kachman (2015) Germfree vs Conventional Swiss webster mice Studies. Metabolomics Workbench, ST000203, V1.

Maureen Kachman (2016) Mice inoculation with human microbiota (AMY1.1SCFA). Metabolomics Workbench, ST000206, V1.

Maureen Kachman (2016) Mice inoculation with human microbiota (MouseAMY1.1SCFA. Cecal). Metabolomics Workbench, ST000207, V1.

Maureen Kachman (2015) Murine gut bile acid analysis. Metabolomics Workbench, ST000212, V1.

Maureen Kachman (2015) Germfree vs Conventional Swiss webster mice Studies (part II). Metabolomics Workbench, ST000213, V1.

Maureen Kachman (2015) Liver and Plasma metaboites for 13C-glucose load in wild type, LIRKO GTT 1 mice. Metabolomics Workbench, ST000215, V1.

Maureen Kachman (2015) Role of Microbiome in Psoriatic Arthritis (SCFA in PsA). Metabolomics Workbench, ST000218, V1.

Susan Sumner (2016) Small cell lung cancer metabolome (part II). Metabolomics Workbench, ST000220, V1.

William Gonsalves (2016) Normal plasma cells,Low proliferation multiple myeloma and High proliferation multiple myeloma cells. Metabolomics Workbench, ST000221, V1.

Christoph Borchers (2015) Bile acid targeted metabolomics of the small intestine in malnourished and control mice. Metabolomics Workbench, ST000222, V1.

Susan Sumner (2016) Metabolic Aberrations in Barth Syndrome. Metabolomics Workbench, ST000223, V1.

Christoph Borchers (2015) Vitamin targeted metabolomics of the small intestine in malnourished and control mice. Metabolomics Workbench, ST000224, V1.

Andrew Patterson (2015) Aryl Hydrocarbon Receptor Activation by Persistent Organic Pollutants Impacts Gut Microbiota-Host Metabolic Homeostasis in Mice. Metabolomics Workbench, ST000225, V1.

Andrew Patterson (2015) Metabolomics Reveals that Aryl Hydrocarbon Receptor Activation by Environmental Chemicals Induces Systemic Metabolic Dysfunction in Mice (Part I). Metabolomics Workbench, ST000226, V1.

Yoshiki Murakami (2015) Comprehensive analysis of transcriptome and metabolome in Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma. Metabolomics Workbench, ST000230, V1.

Yoshiki Murakami (2015) Comprehensive analysis of transcriptome and metabolome in Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma (part II). Metabolomics Workbench, ST000231, V1.

Christoph Borchers (2015) Untargeted metabolomic analysis of the small intestinal content of malnourished mice. Metabolomics Workbench, ST000232, V1.

Andrew Patterson (2015) Metabolomics Reveals that Aryl Hydrocarbon Receptor Activation by Environmental Chemicals Induces Systemic Metabolic Dysfunction in Mice (Part II). Metabolomics Workbench, ST000233, V1.

Qiande Liang (2015) Quick Comparison of Urine Metabolites in Human and SD Rats of Different Sex by Untargeted UPLC-TOFMS and In-house Software Platform. Metabolomics Workbench, ST000236, V1.

Gianrico Farrugia (2016) Whole unconditioned medium (Defined culture media, M199),Whole M1 medium,Whole M2 medium. Metabolomics Workbench, ST000242, V2.

Maureen Kachman (2016) Acyl-Carnitine Analysis in mouse soleus muscle. Metabolomics Workbench, ST000245, V1.

Candice Ulmer (2016) Lipid Extraction Efficiency Comparison. Metabolomics Workbench, ST000246, V1.

Loic Deleyrolle (2016) Metabolic heterogeneity in Glioblastoma. Metabolomics Workbench, ST000248, V1.

Franz Schaub (2016) Measurement of free amino acid (AA) in response to MYC. Metabolomics Workbench, ST000249, V1.

John Koethe (2016) The Role of Obesity and Adipocytes in Immune Activation on Antiretroviral Therapy. Metabolomics Workbench, ST000250, V1.

Susan Sumner (2016) The role of microbial metabolites in experimental liver disease. Metabolomics Workbench, ST000254, V1.

John Newman (2015) NIH WCMC Pilot & Feasibility Project: Metabolomics of Neonatal Pulmonary Hypertension in human. Metabolomics Workbench, ST000255, V1.

John Newman (2015) NIH WCMC Pilot & Feasibility Project: Metabolite changes associated with weight loss. Metabolomics Workbench, ST000257, V1.

Jianzhi Hu (2015) 1H NMR metabolomics study of spleen from C57BL/6 mice exposed to gamma radiation. Metabolomics Workbench, ST000261, V1.

Jatinder Lamba (2016) Serum samples from Mla patients responsive or not responsive to chemotherapy. Metabolomics Workbench, ST000270, V1.

Michael Wolfgang (2016) Acyl-carnitine analysis (plasma). Metabolomics Workbench, ST000272, V1.

Michael Weiden (2016) Short-chain fatty acid analysis in bronchoalveolar lavage fluid (BAL SCFA). Metabolomics Workbench, ST000273, V1.

Bivin Thomas (2016) HIF 1 alpha type 2 cells metabolomics. Metabolomics Workbench, ST000274, V1.

Andrea Calvert (2016) IDH1 and Glioma knockdown idh1. Metabolomics Workbench, ST000276, V1.

Yanhua Zhao (2016) Viral Effect on Metabolism (part I). Metabolomics Workbench, ST000278, V1.

Yanhua Zhao (2016) Viral Effect on Metabolism (part II). Metabolomics Workbench, ST000279, V1.

Michelle Wynn (2016) Pilot Study 13C flux effects when RhoC or RhoA perturbed (13C BCs). Metabolomics Workbench, ST000282, V1.

Patrick Robichaud (2016) Plasma Nucleotide/adenosine concentrations (Human AxP Batch 4). Metabolomics Workbench, ST000283, V1.

Haiwei Gu (2015) Colorectal Cancer Detection Using Targeted Serum Metabolic Profiling. Metabolomics Workbench, ST000284, V1.

Haiwei Gu (2015) NMR-based Metabolomics for CRC Diagnosis. Metabolomics Workbench, ST000285, V1.

Sean Forbes (2016) Energetics in dystrophic muscle. Metabolomics Workbench, ST000287, V1.

Neha Bhise (2016) Metabolomic profiling of AML Cell Lines. Metabolomics Workbench, ST000288, V1.

Haiyan Liu (2016) LC-MS Based Approaches to Investigate Metabolomic Differences in the Urine of Young Women after Drinking Cranberry Juice or Apple Juice. Metabolomics Workbench, ST000291, V1.

Haiyan Liu (2016) LC-MS Based Approaches to Investigate Metabolomic Differences in the Plasma of Young Women after Drinking Cranberry Juice or Apple Juice. Metabolomics Workbench, ST000292, V1.

Jesse Gregory (2016) Vitamin B6 Effects on one-carbon metabolism. Metabolomics Workbench, ST000293, V1.

Maureen Kachman (2016) Metabolic analysis of Human and Mouse Lung Fiboblasts. Metabolomics Workbench, ST000295, V1.

Maureen Kachman (2016) Chronic mild stress and Lactobacillus experiments on mice. Metabolomics Workbench, ST000296, V1.

Maureen Kachman (2016) Analysis of steroid metabolites in psoriasis. Metabolomics Workbench, ST000298, V1.

Maureen Kachman (2016) Conjugated linoleic acid (CLA) study in LDLR-/- mice. Metabolomics Workbench, ST000299, V1.

Maureen Kachman (2016) GBM Cell Lines Reproducibility Pilot Study. Metabolomics Workbench, ST000301, V1.

Maureen Kachman (2016) Isocitrate dehydrogenase-1/Glioma Fluxomics Study. Metabolomics Workbench, ST000302, V1.

Maureen Kachman (2016) Human fecal bile acid profiles before and after fecal transplant (Part 2). Metabolomics Workbench, ST000303, V1.

Maureen Kachman (2016) Early in life exposure studies on human and mouse samples. Metabolomics Workbench, ST000304, V1.

Susan Sumner (2016) Metabolomics Approach to Identify Molecules and Pathways Involved in the Development of Atherosclerotic Coronary Artery Disease. Metabolomics Workbench, ST000306, V1.

Laurence Morel (2016) TC and B6 untreated plasma in lupus-prone mice lipidomics (part-II). Metabolomics Workbench, ST000310, V1.

Laurence Morel (2017) TC and B6 untreated plasma in lupus-prone mice. Metabolomics Workbench, ST000311, V1.

Maureen Kachman (2016) Muscle Clock knock out mice metabolic changes (iMSBmal1-Exp1). Metabolomics Workbench, ST000313, V1.

Susan Sumner (2016) NSAID treatment alters the metabolomics profile of liver, kidney, lung, and heart in an experimental mouse model of heat stroke. Metabolomics Workbench, ST000314, V1.

Susan Sumner (2016) Metabolomics and Childhood Obesity: A Pilot and Feasibility Study With Multiple Phenotypic Anchors. Metabolomics Workbench, ST000315, V1.

Dong Ho Suh (2016) Comparison of Metabolites Variation and Antiobesity Effects of a Mixture of Cudrania tricuspidata, Lonicera caerulea, and the Soybean According to Fermentation in vitro and in vivo. Metabolomics Workbench, ST000316, V1.

Oliver Fiehn (2016) Single treatment gene impact on Arabidopsis metabolites. Metabolomics Workbench, ST000320, V1.

Oliver Fiehn (2016) Effects of LGG on current drinkers gut metabolism. Metabolomics Workbench, ST000321, V1.

Oliver Fiehn (2016) Metabolomic effects of metformin on mouse liver, intestine, and serum. Metabolomics Workbench, ST000325, V1.

Oliver Fiehn (2016) Minimal change disease and focal segmental sclerosis in urine. Metabolomics Workbench, ST000329, V1.

Oliver Fiehn (2016) Effects of Zinc on GI tract metabolites (Part 1: Esophagus). Metabolomics Workbench, ST000330, V1.

Oliver Fiehn (2016) Effects of Zinc on GI tract metabolites (Part 2: Prostate). Metabolomics Workbench, ST000331, V1.

Oliver Fiehn (2016) Effects of Giardia intestinalis on mice GI tract. Metabolomics Workbench, ST000332, V1.

Brittany Lee (2017) Targeted LC/MS of urine from boys with DMD and controls. Metabolomics Workbench, ST000336, V1.

John Seal (2017) Metabolomics Approach to Allograft Assessment in Liver Transplantation. Metabolomics Workbench, ST000337, V1.

Anna Mathew (2016) Gut microbiome-derived metabolites modulate intestinal epithelial cell damage and mitigate graft-versus-host disease. Metabolomics Workbench, ST000338, V1.

Oliver Fiehn (2016) Metformin effects on liver and kidney tissue. Metabolomics Workbench, ST000340, V1.

Hyung-Ok Lee (2017) Methylation in mouse lymphomas. Metabolomics Workbench, ST000341, V1.

Oliver Fiehn (2016) Renal metabolic pathways indicating ischemic or inflammatory changes. Metabolomics Workbench, ST000342, V1.

Oliver Fiehn (2016) Metabolites detected from human bronchoalveolar lavage. Metabolomics Workbench, ST000346, V1.

Daesung Shin (2016) Metabolic profiling reveals biochemical pathways and potential biomarkers associated with the pathogenesis of Krabbe disease. Metabolomics Workbench, ST000352, V1.

Oliver Fiehn (2016) Metabolite comparison of mouse gastric tissue and glands. Metabolomics Workbench, ST000354, V1.

Guoxiang Xie (2016) GC/MS and LC/MS metabolomics profiling for breast cancer plasma data and control plasma data. Metabolomics Workbench, ST000355, V1.

Guoxiang Xie (2016) GC/MS and LC/MS metabolomics profiling for breast cancer serum data and control serum data. Metabolomics Workbench, ST000356, V1.

Teresa Fan (2016) Distinctly perturbed metabolic networks underlie differential tumor tissue damages induced by immune modulator b-glucan in a two-case ex vivo non-small cell lung cancer study. Metabolomics Workbench, ST000367, V1.

Oliver Fiehn (2016) Investigation of metabolomic blood biomarkers for detection of adenocarcinoma lung cancer. Metabolomics Workbench, ST000368, V1.

Oliver Fiehn (2016) Investigation of metabolomic blood biomarkers for detection of adenocarcinoma lung cancer (part II). Metabolomics Workbench, ST000369, V1.

Maureen Kachman (2016) Primary T Cell Noxa Knockdown (Donor 8) III. Metabolomics Workbench, ST000370, V1.

Maureen Kachman (2016) Colon cancer metabolomics via stool sample. Metabolomics Workbench, ST000371, V1.

Maureen Kachman (2016) Mice exercise metabolomics (part I). Metabolomics Workbench, ST000374, V1.

Maureen Kachman (2016) Mice exercise metabolomics (part II). Metabolomics Workbench, ST000375, V1.

Maureen Kachman (2016) Kidney choline acetyltransferase (ChAT)-3. Metabolomics Workbench, ST000376, V1.

Oliver Fiehn (2016) Temporal metabolomic responses of cultured HepG2 liver cells to high fructose and high glucose exposures. Metabolomics Workbench, ST000379, V1.

Oliver Fiehn (2016) Temporal metabolomic responses of cultured HepG2 liver cells to high fructose and high glucose exposures (part II). Metabolomics Workbench, ST000380, V1.

Oliver Fiehn (2016) Urinary Metabolites in IC/PBS Diagnosis (part I). Metabolomics Workbench, ST000381, V1.

Oliver Fiehn (2016) Urinary Metabolites in IC/PBS Diagnosis (part II). Metabolomics Workbench, ST000382, V1.

Oliver Fiehn (2016) Plasma Metabolomic Profiles Reflective of Glucose Homeostasis in Non-Diabetic and Type 2 Diabetic Obese African-American Women. Metabolomics Workbench, ST000383, V1.

Oliver Fiehn (2016) Investigation of metabolomic blood biomarkers for detection of adenocarcinoma lung cancer (training set). Metabolomics Workbench, ST000385, V1.

Oliver Fiehn (2016) Investigation of metabolomic blood biomarkers for detection of adenocarcinoma lung cancer (test/validation). Metabolomics Workbench, ST000386, V1.

Oliver Fiehn (2016) Changes in the metabalome and lipidome in response to exercise training. Metabolomics Workbench, ST000387, V1.

Oliver Fiehn (2016) Serum phosphatidylethanolamine levels distinguish benign from malignant solitary pulmonary nodules and represent a potential diagnostic biomarker for lung cancer (part I). Metabolomics Workbench, ST000388, V2.

Oliver Fiehn (2016) Serum phosphatidylethanolamine levels distinguish benign from malignant solitary pulmonary nodules and represent a potential diagnostic biomarker for lung cancer (part II). Metabolomics Workbench, ST000389, V2.

Oliver Fiehn (2016) Metabolomic markers of altered nucleotide metabolism in early stage adenocarcinoma (part I). Metabolomics Workbench, ST000390, V2.

Oliver Fiehn (2016) Metabolomic markers of altered nucleotide metabolism in early stage adenocarcinoma (part II). Metabolomics Workbench, ST000391, V2.

Oliver Fiehn (2016) Systemic Metabolomic Changes in Blood Samples of Lung Cancer Patients Identified by Gas Chromatography Time-of-Flight Mass Spectrometry. Metabolomics Workbench, ST000392, V2.

Oliver Fiehn (2016) Lung Cancer Plasma Discovery. Metabolomics Workbench, ST000396, V2.

Oliver Fiehn (2016) Long-term neural and physiological phenotyping of a single human. Metabolomics Workbench, ST000397, V2.

Oliver Fiehn (2016) Metabolic profiling of maternal urine can aid clinical management of Gestational Diabetes Mellitus (GDM). Metabolomics Workbench, ST000398, V1.

Darren Creek (2016) Metabolomics-based elucidation of active metabolic pathways in erythrocytes and HSC-derived reticulocytes. Metabolomics Workbench, ST000403, V1.

Oliver Fiehn (2016) Role of HVCN1 in B cell malignancies. Metabolomics Workbench, ST000404, V1.

Oliver Fiehn (2016) Metabolic profiling during ex vivo machine perfusion of the human liver (part I). Metabolomics Workbench, ST000412, V1.

Oliver Fiehn (2016) Metabolic profiling during ex vivo machine perfusion of the human liver (part III). Metabolomics Workbench, ST000413, V1.

Oliver Fiehn (2016) Impact Of High Sugar Diet On L-Arginine Metabolism In The Lung. Metabolomics Workbench, ST000419, V1.

Sreekumaran Nair (2016) Type 1 Diabetes poor glycemic control versus control samples. Metabolomics Workbench, ST000421, V1.

Sreekumaran Nair (2016) Type 1 Diabetes good glycemic control and controls samples. Metabolomics Workbench, ST000422, V1.

Monte Willis and Amro Ilaiwy (2016) Non targeted metabolomics of gastrocnemius tissue samples obtained from 20 month old (old) mice- Both Sham and after inducing lung injury (part I). Metabolomics Workbench, ST000425, V1.

Monte Willis and Amro Ilaiwy (2016) Non targeted metabolomics of gastrocnemius tissue samples obtained from 6 month old (adult) mice- Both Sham and after inducing lung injury (part II). Metabolomics Workbench, ST000426, V1.

Monte Willis and Amro Ilaiwy (2016) targeted metabolomics of gastrocnemius tissue samples obtained from 6 month old (adult) mice- Both Sham and after inducing lung injury (part I). Metabolomics Workbench, ST000427, V1.

Monte Willis and Amro Ilaiwy (2016) Targeted metabolomics of gastrocnemius tissue samples obtained from 20 month old (old) mice- Both Sham and after inducing lung injury (part II). Metabolomics Workbench, ST000428, V1.

Michelle Mielke (2016) Quantitative measurements of vitamin D in T1D poor control, good control, and controls. Metabolomics Workbench, ST000432, V1.

Sreekumaran Nair (2016) Plasma sphingolipid changes with autopsy-confirmed Lewy body or Alzheimer's pathology. Metabolomics Workbench, ST000433, V1.

Michelle Mielke (2016) Quantitative measurements of free fatty acid in T1D poor control, good control, and controls. Metabolomics Workbench, ST000434, V1.

Sreekumaran Nair (2016) Quantitative measurements of amino acids in T1D poor control, good control, and controls. Metabolomics Workbench, ST000435, V1.

Susan Sumner (2016) Characterizing commonalities and differences between the breast and prostate cancer metabotypes in African-American cohorts (part I). Metabolomics Workbench, ST000438, V1.

Susan Sumner (2016) Characterizing commonalities and differences between the breast and prostate cancer metabotypes in African-American cohorts. Metabolomics Workbench, ST000439, V1.

Susan Sumner (2016) Metabotypes of Subjects with Adverse Reactions Following Vaccination . Metabolomics Workbench, ST000440, V1.

Stewart Delisha (2016) Metabolomics Analysis of Triple Negative Breast Cancer (BCa) Cell Lines. Metabolomics Workbench, ST000442, V1.

Sreekumaran Nair (2016) Quantitative measurements of TCA cycle metabolites in T1D poor control, good control, and controls. Metabolomics Workbench, ST000443, V1.

Livia Da Costa (2016) Follicular fluid lipidomics reveals lipid alterations by LH addition during IVF cycles. Metabolomics Workbench, ST000445, V1.

Robert Naviaux (2016) Metabolic features of chronic fatigue syndrome. Metabolomics Workbench, ST000450, V1.

Monte Willis and Amro Ilaiwy (2016) The alpha-1A adrenergic receptor agonist A61603 reduces cardiac polyunsaturated fatty acid-Heart raw data. Metabolomics Workbench, ST000451, V1.

Monte Willis and Amro Ilaiwy (2016) The alpha-1A adrenergic receptor agonist A61603 reduces cardiac polyunsaturated fatty acid-Serum raw data. Metabolomics Workbench, ST000452, V1.

Ayse Batova (2016) Uniquely Tumor-Selective Englerin A Profoundly Alters Lipid Metabolism in Renal Cell Carcinoma inducing ER-Stress and an Acute Inflammatory Response. Metabolomics Workbench, ST000465, V1.

Jaapna Dhillon (2016) A metabolomics approach to document compliance with long-term almond consumption. Metabolomics Workbench, ST000477, V1.

Sreekumaran Nair (2016) Amino Acid Quantifcation of obese patients on a 16 week caloric restriction from Plasma. Metabolomics Workbench, ST000483, V1.

Line Engelbrechtsen (2016) Weight loss and weight maintenance obtained with or without GLP-1 analogue treatment decrease branched chain amino acid levels. Metabolomics Workbench, ST000502, V1.

Travis Nemkov (2017) Assessing metabolic changes to maternal rat liver tissue during and post-pregnancy (part II). Metabolomics Workbench, ST000510, V1.

Darren Creek (2017) Metabolomics-based elucidation of active metabolic pathways in erythrocytes and HSC-derived reticulocytes (part II). Metabolomics Workbench, ST000539, V1.

Sreekumaran Nair (2017) Triple Quadrupole Mass Spectrometer to measure low abundance isotope enrichment in individual muscle proteins. Metabolomics Workbench, ST000542, V1.

Sreekumaran Nair (2017) High Resolution orbittrap Mass Spectrometer to measure low abundance isotope enrichment in individual muscle proteins. Metabolomics Workbench, ST000543, V1.

Jim Whelan (2008) Combined Moderate High Light and Drought Stress: Mitochondrial Alternative Oxidase 1a KO Arabidopsis. MetaPhen, MEX1.

A. Harvey Millar (2008) Timecourse: Treatment of Arabidopsis cell suspension cultures with rotenone. MetaPhen, MEX2.

Barry J. Pogson (2009) Metabolomic characterisation of drought-tolerant alx8 and fry1-1 Arabidopsis mutants. MetaPhen, MEX5.

A. Harvey Millar (2009) Metabolic response of wildtype Arabidopsis thaliana ecotype Columbia root to menadione-induced oxidative stress. MetaPhen, MEX6.

A. Harvey Millar (2009) Arabidopsis Mitochondrial ETC Complex I Mutants - ndusf4 and complemented line - metabolome during the day and night. MetaPhen, MEX10.

A. Harvey Millar (2009) Metabolomic characterisation of ndusf4 and ndufa1 mitochondrial electron transport chain Complex I mutants at the ends of two successive nights. MetaPhen, MEX11.

Julia Bailey-Serres (2008) Anoxic treatment of Arabidopsis seedlings_Branco-Price (2008a). MetaPhen, MEX15.

Peter Geigenberger (2008) Responses of Arabidopsis seedlings to mild decreases in external oxygen concentration_van Dongen (2009a). MetaPhen, MEX16.

Mark Stitt (2002) Responses of arabidopsis seeds and silique walls to decreases in external oxygen concentration gibon (2002a). MetaPhen, MEX18.

A. Harvey Millar (2010) Arabidopsis mMDH DKO - Diurnal_Tomaz (2010a). MetaPhen, MEX27.

A. Harvey Millar (2010) Arabidopsis mMDH DKO and Complementation - Day_Tomaz (2010b). MetaPhen, MEX28.

Adam James Carroll (2015) Arabidopsis thaliana photorespiration mutants. MetaPhen, MEX36.

Adam James Carroll (2015) Metabolic phenotyping of unknown photorespiratory mutants isolated by chlorophyll fluorescence screen of EMS mutant population. MetaPhen, MEX42.

Adam James Carroll (2015) Response of Arabidopsis thaliana ecotype Wassilewskija-2 (cold tolerant) and Cape Verde Islands-1 (cold susceptible) to cold treatment. MetaPhen, MEX44.

Lee James Sweetlove (2006) Response to menadione induced oxidative stress in cultured heterotrophic cells of Arabidopsis thaliana ecotype Lansberg erecta. MetaPhen, MEX45.

Toru Fujiwara (2010) Metabolic responses to molybdenum deficiency in shoots of Columbia wild type and MOT1 molybdenum transporter knockout Arabidopsis thaliana by GC-EI-TOF-MS, CE-ESI-TOF-MS and UHPLC. MetaPhen, MEX53.

Charles L. Guy (2004) Timecourse metabolic response of Arabidopsis thaliana ecotype Columbia wild type to cold stress. MetaPhen, MEX68.

Charles L. Guy (2004) Timecourse metabolic response to heat stress in Arabidopsis thaliana wild type Columbia. MetaPhen, MEX69.

Claudia Jonak (2008) Response of Arabidopsis thaliana wildtype Col-0 to abscisic acid treatment. MetaPhen, MEX70.

Claudia Jonak (2008) Time course salt treatment on Arabidopsis thaliana ecotype Columbia wild type. MetaPhen, MEX71.

Eiichiro Fukusaki (2006) Time course salt stress treatment on Arabidopsis Col-0 wildtype T78 cells by GC-EI-Q-MS and LC-ESI-Q-MS. MetaPhen, MEX72.

Jacques Bourguignon (2006) Metabolic response of Arabidopsis photosynthetic cells to caesium stress. MetaPhen, MEX74.

Jacques Bourguignon (2006) Metabolic response of Arabidopsis photosynthetic cells to potassium. MetaPhen, MEX75.

Lee James Sweetlove (2008) Metabolic response of wildtype Arabidopsis thaliana ecotype Columbia root to menadione-induced oxidative stress. MetaPhen, MEX76.

Michael K. Udvardi (2009) Metabolic profiling of AtMyb41-overexpressor lines under non-stressed and salt stressed conditions. MetaPhen, MEX77.

Alain Bouchereau (2008) Metabolomic investigation of the freezing-tolerant Arabidopsis mutant esk1. MetaPhen, MEX82.

Kazuko Yamaguchi-Shinozaki (2009) Metabolomic analysis of Arabidopsis thaliana lines overexpressing transcription factor DREB1A or DREB2A. MetaPhen, MEX86.

Victoria Nikiforova (2005) Metabolic response of Arabidopsis thaliana wildtype Columbia genotype G1 to sulfur deficiency. MetaPhen, MEX90.

Carole Deleu (2010) Metabolic phenotyping of wild-type and pop2 (At3g22200) 4-aminobutyrate transaminase mutant Arabidopsis thaliana plants under non-stress and salt stress conditions. MetaPhen, MEX93.

Jacques Bourguignon (2006) Metabolic response of Arabidopsis thaliana wild type Col-0 suspension cells to increasing concentrations of Cadmium treatment (2, 5, 20, 50, 200 uM) for 24 hours. MetaPhen, MEX96.

Dirk Inze (2009) Metabolic response to prolonged mild osmotic stress in Arabidopsis leaves of different developmental stages measured by GC-MS. MetaPhen, MEX97.

Jiping Chen (2009) The metabolic responses of Arabidopsis thaliana ecotype Wassilewskija (Ws) to cadmium exposure. MetaPhen, MEX98.

Kazuo Shinozaki (2008) Dehydration responses in detached aerial parts of wild-type and NCED3 (AT3G14440) knockout Arabidopsis thaliana plants by GC-MS and CE-MS. MetaPhen, MEX99.

Wout Boerjan (2012) Metabolic phenotypes of Arabidopsis thaliana mutants with single gene mutation in the lignin biosynthesis pathway characterised by GC-MS and UPLC-MS. MetaPhen, MEX101.

Chihiro K. Watanabe (2010) Metabolic phenotyping of shoots and roots of Arabidopsis thaliana alternative oxidase 1a (aox1a) mutant plants under low nitrogen stress. MetaPhen, MEX104.

Holger Hesse (2010) Metabolic responses of Arabidopsis thaliana plants to sulphur depletion, high light and combined stresses. MetaPhen, MEX106.

Atsushi Fukushima (2014) MeKO: Metabolic phenotypes of Arabidopsis mutants with mutation in gene with known or unknown functions. MetaPhen, MEX115.

Marina Gromova and Claude Roby (2010) Arabidopsis thaliana hydrophilic metabolome. MeRy-B, A05001.

MetaboP-Arabidopsis-photoperiode. MeRy-B, A06001.

Appendix C

Dunn Post-hoc Tests for Compliance with MSI Guidelines

Table C.1 Dunn Post Hoc Test, with Benjamini-Hochberg correction, comparing compliance with the MSI minimal reporting standards in the MetaboLights repository.

Reporting Standard 1	Reporting Standard 2	Z	*p*-value	*q*-value
Clinical	*in vitro*	-1.53	0.13	0.25
Clinical	Pre-clinical	-1.33	0.19	0.28
Clinical	Plant	-1.16	0.24	0.29
in vitro	Pre-clinical	-2.80	0.0052*	0.031*
in vitro	Plant	-2.55	0.011*	0.032*
Pre-clinical	Plant	0.04	0.97	0.97

* indicates significant values

Table C.2 Dunn Post Hoc Test, with Benjamini-Hochberg correction, comparing compliance with the MSI minimal reporting standards in the Metabolomics Workbench repository.

Reporting Standard 1	Reporting Standard 2	Z	*p*-value	*q*-value
Clinical	*in vitro*	-1.20	0.23	0.27
Clinical	Pre-clinical	-0.42	0.67	0.67
Clinical	Plant	-4.34	1.41×10^{-5}*	4.23×10^{-5}*
in vitro	Pre-clinical	-1.65	0.099	0.15
in vitro	Plant	-5.11	3.26×10^{-7}*	1.95×10^{-6}*
Pre-clinical	Plant	-4.24	2.26×10^{-5}*	4.52×10^{-5}*

* indicates significant values

Table C.3 Dunn post-hoc test, with Benjamini-Hochberg correction, comparing compliance with the MSI minimal reporting standards in the GNPS repository.

Reporting Standard 1	Reporting Standard 2	Z	*p*-value	*q*-value
Clinical	*in vitro*	-1.58	0.11	0.17
in vitro	Pre-clinical	-2.34	0.019*	0.057
Clinical	Pre-clinical	-0.76	0.49	0.45

* indicates significant values

Table C.4 Dunn post-hoc test, with Benjamini-Hochberg correction, comparing compliance with the MSI optional reporting standards in the MetaboLights repository.

Guideline 1	Guideline 2	Z	*p*-value	*q*-value
Clinical	*in vitro*	2.60	0.0095*	0.028*
Clinical	Pre-clinical	-1.95	0.051	0.077
in vitro	Pre-clinical	0.07	0.95	0.95

* indicates significant values

Table C.5 Dunn post-hoc test, with Benjamini-Hochberg correction, comparing compliance with the MSI optional reporting standards in the Metabolomics Workbench repository.

Reporting Standard 1	Reporting Standard 2	Z	*p*-value	*q*-value
Clinical	*in vitro*	3.086	0.0020*	0.0060*
Clinical	Pre-clinical	-2.36	0.018*	0.027*
in vitro	Pre-clinical	0.038	0.97	0.97

* indicates significant values

Table C.6 Dunn post-hoc test, with Benjamini-Hochberg correction, comparing compliance with MSI plant reporting standards between metabolomics repositories.

Repository 1	Repository 2	Z	*p*-value	*q*-value
MeRy-B	MetaboLights	-0.69	0.49	0.49
MeRy-B	Metabolomics Workbench	-2.75	0.0060*	0.036*
MetaboLights	Metabolomics Workbench	-2.05	0.040*	0.12
MeRy-B	MetaPhen	-1.53	0.13	0.25
MetaboLights	MetaPhen	-0.84	0.40	0.48
Metabolomics Workbench	MetaPhen	1.21	0.23	0.34

* indicates significant values

Appendix D

Suggested Metadata Classification

A suggested update to the Metabolomics Standards Initiative (MSI) biological context metadata reporting standards. The existing standards are classified as either *Mandatory* (should be reported for all studies), *Required* (should be reported for all studies where used e.g. treatment) or *Optional* (reporting is down to the authors discretion). The standards are also classified as whether they should be reported on a per sample or per study basis.

Table D.1 Suggested update to the MSI Mammalian Clinical Trials and Human Studies Reporting Standards. Metadata that are recommended further information under the existing MSI standards are in italics.

MSI Classification - Clinical			Mandatory/ Required/ Optional	Per Study/ Per Sample
Experimental Subject Descriptors	Weight range and Height and/or BMI		R	Sample
	Trial Type		R	Study
	Diet		O	Sample
	Smoking Status		O	Sample
	Drug Consumption		O	Sample
	Alcohol Consumption		O	Sample
	Metal Exposure		O	Sample
	Malnutrition		O	Sample
Experimental Design	Number of Groups		M	Study
	Inclusion Criteria		R	Study
	Exclusion Criteria		R	Study
	Treatments/ Fasting	Fasting Status	R	Sample
		Treatment	R	Sample
		Treatment Dose	R	Sample
		Treatment Duration	R	Sample
		Treatment Route	R	Study
		Treatment Vehicle	R	Study
End Points	*Urea*		O	Sample
	Creatinine		O	Sample
	Hemoglobin		O	Sample
	Hemocrit		O	Sample
	Glucose		O	Sample
	Total Cholesterol		O	Sample
	HDL Cholesterol		O	Sample
	LDL Cholesterol		O	Sample
	Triglycerides		O	Sample
	Total Protein		O	Sample

MSI Classification - Clinical			Mandatory/ Required/ Optional	Per Study/ Per Sample
End Points	*Albumin*		O	Sample
	Sodium		O	Sample
	Potassium		O	Sample
	Platelets		O	Sample
	Bilirubin		O	Sample
	ALT		O	Sample
	ALP		O	Sample
	-GT		O	Sample
	White Blood Count		O	Sample
Metabolomics-related sample collection	Biofluid or Tissue		O	Sample
	Volume or Quantity of Collection		M	Study
	Sample Storage Temperature		R	Study
	Sample Storage Duration		O	Study
	Blood	Location of Collection	O	Study
		Time from Collection to Freezing	O	Study
		Anticoagulant	R	Study
		Speed of Centrifugation	O	Study
		Temperature of Centrifugation	O	Study
		Time of Centrifugation	O	Study
		Arterial or Venous Blood	O	Study
		Haemolysis	O	Study
	Urine	*Mid Flow/ Total Urine*	O	Study
		Bacteriostatic Agent	O	Study

Table D.2 Suggested update to the MSI Microbial and *in vitro* Reporting Standards. Metadata that are best practice and not mandatory under the existing MSI standards are in italics. Metadata in the category *Other* were not included in the original MSI guidelines, however it was felt that these were important and should be included in the updated guidelines.

MSI Classification - *in vitro*		Mandatory/ Required/ Optional	Per Study/ Per Sample
Minimal Reporting Standards	Harvesting Method	M	Study
	Time Until Quenching	O	Study
	Temperature from Sampling to Quenching	O	Study
	Quenching Method	M	Study
	Cell Integrity	O	Study
	Extracellular Metabolites Discriminated	O	Study
	Metabolite Extraction	M	Study
	Recovering from Extraction	O	Study
	Normalisation (Cell No.)	O	Study
	Sample Clean-up	O	Study
	Sample Storage	R	Study
	Sample Storage (Duration)	O	Study
	Quality Control	O	Study
	Detection Limit	O	Study
	Stability	O	Study
Experimental Design	*Replicates*	R	Study
	Additional -omics Datasets	R	Study
Biosource	*Supplier*	M	Study
	Cell Type	M	Sample
	Immortalized or transformed	O	Study
Growth Environment	*Growth container*	M	Study
	Growth container (Supplier)	O	Study
	Growth Support	R	Study
	Growth Support (Supplier)	O	Study
	Growth configuration	O	Study
	Subculturing and splitting protocols	R	Study
	Inoculation size	R	Study

MSI Classification - *in vitro*		Mandatory/ Required/ Optional	Per Study/ Per Sample
Growth Environment	*Medium/ Substrate*	M	Study
	Medium/ Substrate (Supplier)	M	Study
	Medium/ Substrate (Concentration)	M	Study
	Temperature	R	Study
	pH	R	Study
	Gas Composition	R	Study
	Humidity	R	Study
	CO_2	R	Study
	Stirrer Speed	R	Study
	Evaporation	R	Study
	pO_2	R	Study
	Growth Rate	R	Study
Treatment	*Treatment*	R	Sample
	Treatment Dose	R	Sample
	Treatment Vehicle	R	Sample
	Treatment Time	R	Sample
	Pretreatment	O	Sample
	Pretreatment Time	O	Sample
	Labeling	R	Sample
Harvesting	*Harvesting Growth Phase*	O	Sample
	Harvesting Time	M	Sample
	No. Generations Until Harvesting	O	Study
	Stabilization Time	O	Study
	Number of Culture Passages	O	Study
	Marker of Differentiated Stage	O	Study
	Harvesting Cell Density	O	Study
	Harvesting Depletion of nutrients	O	Study
Other	Sample Storage (Temperature)	R	Study
	Harvesting Volume	M	Study

Table D.3 Suggested update to the MSI Mammalian Pre-clinical Studies Reporting Standards. Metadata that are recommended further information under the existing MSI standards are in italics.

MSI Classification - Mammalian Preclinical			Mandatory/ Required/ Optional	Per Study/ Per Sample
Experimental Subject Descriptors	Strain		M	Sample
	Additional Phenotypic Model		O	Sample
	Age at Study Start		M	Sample
	Age at Collection or Euthanization		M	Sample
	Animal Supplier		M	Study
	Weight Range		R	Sample
Husbandry	Housing	Group or Individual Housing	O	Study
		Germ-free or Conventional Housing	O	Study
		Bedding Type	O	Study
		Cage Cleaning Frequency	O	Study
		Cage Type	O	Study
		Environment Enrichment: Temperature	R	Study
		Environment Enrichment: Humidity	R	Study
	Light Cycle		R	Study
	Feed	Diet	M	Study
		ad lib or Restricted Diet	M	Study
	Water	Tap or Purified Water	O	Study
	Veterinary treatments if any	*Use of Anesthesia*	R	Study
		Anesthesia Time	R	Study
		Anesthesia Dose	R	Study
	Acclimation	*Acclimation Duration to Experimental Facility*	R	Study
Experimental-	*Inclusion Criteria*		R	Study

MSI Classification - Mammalian Preclinical			Mandatory/ Required/ Optional	Per Study/ Per Sample
Experimental Design	Number of Groups	Number of Groups	M	
		Sex	M	Sample
	Treatments	Treatment	R	Sample
		Treatment Dose	R	Sample
		Treatment Duration	R	Sample
		Treatment Route	R	Sample
		Treatment Vehicle	R	Sample
	Fasting	Fasting Status	R	Study
		Fasting Duration	R	Study
	End Points	Euthanasia method	R	Study
		Biofluid or Tissue	M	Sample
		Tissue Processing	R	Study
		Body Weights or Food Consumption	O	Sample
Metabolomics-related sample collection	Volume or Quantity of Collection		M	Study
	Sample Storage Temperature		R	Study
	Sample Storage Duration		O	Study
	Collection Time		M	Sample
	Blood	Location of Collection	O	Study
		Anticoagulant	O	Study
	Urine	Collection Method	O	Study
		Collection Frequency	O	Study
		Collection Duration	O	Study
		Bacteriostatic Agent	O	Study
		Temperature of Collection Tube	O	Study

Table D.4 Suggested update to the MSI Plant Reporting Standards.

MSI Classification - Plant		Mandatory/ Required/ Optional	Per Study/ Per Sample
Biosource	Genotype	M	Sample
	Organ/ Cell Type	M	Sample
	Biosource Amount	M	Study
Growth Environment	Growth Support (Soil/Agar/Media)	M	Study
	Growth Location	O	Study
	Plot Design	O	Study
	Light	R	Study
	Humidity	R	Study
	Temperature	R	Study
	Watering Regime	O	Study
	Nutrients Regime	R	Study
	Timing/ Dates	M	Sample
Treatment	Treatment	R	Sample
	Treatment Dose	R	Sample
	Treatment Time	R	Sample
Harvest	Harvest Time/Date	M	Sample
	Plant Growth Stage	O	Study
	Metabolism quenching method	R	Study
	Harvest Method	R	Study
	Sample Storage	R	Study

Table D.5 Suggested update to the MSI Environmental Reporting Standards. Metadata that are recommended further information under the existing MSI standards are in italics.

MSI Classification - Environmental Context			Mandatory/ Required/ Optional	Per Study/ Per Sample
Sample		Organism Taxonomy	M	Sample
		Common Name	O	Sample
		Genotype	M	Sample
		Ecotype	M	Sample
		Sample Composition	O	Study
		Sample Type	M	Sample
		Phenotypic Characteristics	O	Sample
		Weight	R	Sample
		Age	R	Sample
		Sex	R	Sample
		Development Stage	R	Sample
Environment	Description of Any Field Environment	Geographic Location	R	Study
		Altitude/Depth	R	Study
		Habitat	R	Study
		Weather Type	R	Study
		Humidity	R	Study
		Precipitation	R	Study
		Wind Speed and Direction	R	Study
		Lunar/Solar Phase	R	Study
		Pollutant Concentration	R	Study
	Description of Any Laboratory Equipment	Laboratory Address	M	Study
		Laboratory Contact details	M	Study
	Description of Terrestrial Environment	Inclination and Aspect	M	Study
		Substrate Type	R	Study
		Substrate Temperature	R	Study

MSI Classification - Environmental Context			Mandatory/ Required/ Optional	Per Study/ Per Sample
Environment	Description of Terrestrial Environment	*Substrate pH*	O	Study
		Substrate Organic Context	O	Study
	Description of Aquatic Environment	Submerged/Emerged	R	Study
		Water Temperature	R	Study
		Tidal Phase	R	Study
		pH	R	Study
		Salinity	R	Study
		Dissolved (in)organic content	R	Study
	Description of Atmospheric Environment	Atmospheric Temperature	R	Study
		Atmospheric pressure	R	Study
		(In)organic content	O	Study
	Description of Biotic Environment	Description of host organism	R	Study
		Relationship of organism(s) to host	R	Study
		pH	R	Study
		Temperature	R	Study
Process	Description of capture/sampling of sample or organism(s)	*Capture Method*	R	Study
		Reason for Capture	O	Study
		Other Capture Parameters	O	Study
	Description of storage/ preservation of sample(s)	*Storage/Preservation Medium*	R	Study
		Reason for Storage/ Preservation	O	Study
		Temperature	O	Study
	Description of-	*Type of Housing*	R	Study

MSI Classification - Environmental Context			Mandatory/ Required/ Optional	Per Study/ Per Sample
Process	Description of maintenance of organism(s)	*Reason for Maintenance*	O	Study
		Feeding Regime	R	Study
		Cage Dimensions	O	Study
	Description of transportation of samples or organism(s)	*Means of Transport*	R	Study
		Reason for Transportation	O	Study
		Other Transportation Parameters	O	Study
	Description of acclimation of organism(s)	*Type of Housing*	R	Study
		Reason for acclimation	O	Study
		Other acclimation parameters	O	Study
	Description of general manipulation of sample or organism(s)	Manipulation type	R	Study
		Description of manipulation procedure	O	Study
		Reason for manipulation	O	Study
		Other manipulation parameters	O	Study

Appendix E

Instructions for Coding PLOS ONE Papers

1. **Does the paper include metabolomics?** Score as **Yes = 1, No = 0** (If a publication is not rated as including metabolomics, do not perform any of the subsequent analysis steps for that papers and instead move to the next publication.)

2. If the paper includes metabolomics, **is the included metabolomics 'primary' research?** This refers to whether the study is generating new metabolomics data, rather than reusing data or performing meta-analysis. This means that publications that include multiple assays of different omics (e.g. genomics, transcriptomics, proteomics) are still rated as primary metabolomics research. Score as **Yes = 1, No = 0** (If a publication is not rated as primary metabolomics research, do not perform any of the subsequent analysis steps for that papers and instead move to the next publication.)

3. Is the research a **clinical human study?** Score as **Yes = 1, No = 0**

4. Classification of the data availability statement

    A All relevant data are within the paper and its Supporting Information files. (Including typos and variations) e.g. 'All relevant data are within the paper and Supporting Information files.' or 'All relevant data are within the paper and supplement.'

    B All relevant data are within the paper. (This type of statement is used when a paper has no supplementary material.)

C Data are in a repository.

D All relevant data are within the paper and its Supporting Information files & other -omics data in a Repository.

E Data available on request.

F Data cannot be made publicly available.

G No statement.

5. If a repository is included in the data availability statement, **record the repository**. (If there is no repository, move to 7)

6. Is there **raw data** available? Score as **Yes = 1, No = 0** (As well as on repositories raw data may also be present in supplementary material or institutional website)

7. **Classification of level of data sharing.** If studies do not have raw data, they can be classified as both 4 and either 2, 3 or 5. They cannot be classified as 2 & 3, 2 & 5 or 3 & 5. Raw data can be used to generate any of the subsequent classifications: peak lists, tables, figures of spectra and figures of metabolites. Data from a peak list can be used to generate tables or figures of metabolites, but cannot be used to generate a figure of the raw spectra. This is why it is possible to have multiple classifications. Figures are considered less raw, as they are less machine readable.

   1 Raw Data available (either commercial format e.g. Thermo .raw, Agilent .d or as an open format e.g. .mzML, .mzXML)

      a Specific Metabolomics Repository (e.g. MetaboLights, Metabolomics Workbench)

      b General Repository (e.g. Figshare, Dryad, Zenodo)

      c Other e.g. supplementary material, institutional website

   2 Peak list, containing relative quantifications, concentrations, etc., on a per sample level (usually in .csv or .xlsx format)

   3 Table of metabolites, either all identified metabolites or only statistically significantly changed metabolites

   4 Figure of spectra

   5 Figure showing statistically significantly changed metabolites (such as a scatter plot, bar chart, heat map, etc.)

8. If data is classified as 1a or 1b record the **link or accession number**.

Appendix F

Permissions of Use

This appendix contains information regarding permissions obtained for the use of Figure 2.1, which is an original of another publication. For the use of Figure 2.1, I obtained a license for electronic and paper distribution to be used solely within this thesis, through the Rightslink service. License number 4355360456390, dated May 24th 2018. The license requires that the following text is included in the thesis:

"Reprinted by permission from Springer Nature: Metabolomics, R. Goodacre, Water, water, every where, but rarely any drop to drink, vol. 10, no. 1, pp. 5-7, Copyright 2014".