Automated processing of untargeted ^{1}H-NMR metabolomics data of urine from treated TB patients

J van der Westhuizen

orcid.org/0000-0001-8483-3939

Dissertation accepted in partial fulfilment of the requirements for the degree *Master of Science in Biochemistry* at the North-West University

Supervisor: Prof. SW Mason
Co-supervisor: Dr. M van Wyk

Graduation July 2022

28710665

ACKNOWLEDGEMENTS

"It doesn't matter who you are, or where you come from. The ability to triumph begins with you. Always." – Oprah Winfrey

What a journey this has been! Loads of ups and downs, wanting to give up, but pushing through instead. I could never have done this alone.

Firstly, thank you Prof Mason & Dr Mari for your guidance, patience, and invaluable input into not only this dissertation, but into me as a biochemist. I have learnt from the two of you immensely and your contribution to my life will never be forgotten.

Secondly, I want to thank my family & friends for all their support through this experience. Specifically, I would like to thank you, my Kylie, for all the pep-talks and wiping off the discouraged tears when I felt like giving up. Without you I would never have gotten through this, and I can't wait to marry you. Thank you, Grandad, for always being interested in my work, even when you didn't understand a word of it. You read my honours dissertation from cover to cover and I'm sure you will do the same for this one. Your support means the world to me, and I hope I make you proud. Mommy: thank you will never be enough. I will never have enough words to convey my appreciation to you for absolutely everything. Thank you for being my best friend and for making everything possible, and for making me believe I can do it even before I consider it. You always just know what is best and I am eternally grateful to you. So here it is – the master's degree I was convinced I didn't want, for YOU.

Lastly, I would like to acknowledge and thank the National Research Foundation (NRF) for the funding in the form of a full master's scholarship that made this all possible, financially.

Please COVID-19, may I finally be allowed to graduate?

ABSTRACT

Metabolomics is becoming an increasingly popular field of study. An analytical platform commonly used for metabolomics studies is proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy. The main challenge of ^{1}H-NMR is that the interpretation of the spectral output is time-consuming and somewhat difficult, requiring not only a trained analyst, but one with experience. For this reason, the automation of ^{1}H-NMR metabolomics data processing is gaining attention, leading to the invention of various software tools/applications. One such software tool is BAYESIL, a fully automated and quantitative tool focussing on ^{1}H-NMR. However, BAYESIL has not yet been rigorously tested to ascertain robustness and accuracy in urine. This software tool is the focus of this study, where it was compared to the typical method currently used at the Centre for Human Metabolomics (CHM) at the North-West University (NWU). The aim of this study was to determine whether the automation of the processing of untargeted ^{1}H-NMR metabolomics data has the capability to obtain a comprehensive and comparable metabolic profile of treated and untreated tuberculosis (TB) patient urine samples which have been specifically selected for this study due to their complexity.

TB urine samples were analysed by a ^{1}H-NMR spectrometer, along with healthy control samples. The data matrices containing the results of each method were subjected to the same statistical analysis via the online tool, MetaboAnalyst, to identify important metabolites and obtain a metabolite profile of the samples. While working with MetaboAnalyst, a few areas were identified where a novice user easily makes mistakes. Tips on how to navigate around these areas were included in Chapter 4 as guideposts to help future novices along the way. The metabolite profiles obtained by each method were then compared. The main findings of this study were that the automated BAYESIL method seems to struggle with the identification of some metabolites (such as those arising from TB disease or medication) due to its limited library. BAYESIL also had some difficulty with the coverage of the spectra and identified metabolites with a lower confidence and lower specificity. On the other hand, the manual method was able to comfortably identify metabolites relating to TB medication and identified metabolites with higher specificity and higher confidence.

These results indicate that the BAYESIL method is currently better suited for less detailed research, such as shotgun metabolic profiling. If a more detailed representation of the metabolism is desired, it is recommended to use the manual method or even combine these methods.

Keywords: ^{1}H-NMR; metabolomics; profiling; BAYESIL; automated; processing

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Dissertation structure and research outputs

This dissertation is written in article format. The research output from this study is a peer-reviewed scientific article submitted for publication in the journal *Metabolomics* (IF: 4.29) and is included in Chapter 4 of this dissertation. This research was also presented at the 27th congress of the South African Society of Biochemistry and Molecular Biology (SASBMB) held from 23 – 26 January 2022 as an e-poster, where a 2nd place award was won for best poster. See Annexure E for the certificate that was awarded for this achievement.

Author contributions

The primary author of this dissertation and the associated publication, and main experimental investigator is J van der Westhuizen. J van der Westhuizen was responsible for the planning of the project, all sample analysis and data analysis as well as the writing of this dissertation and the publication associated with this study. Prof. S Mason served as supervisor, and supervised all aspects of this study, including the project design, planning, sample analysis, and writing of this dissertation, and the publication resulting from this study. Dr. M van Wyk served as co-supervisor, and supervised aspects relating to project design, planning, statistical analysis and critical feedback on the dissertation, and the associated publication.

As indicated above, I declare that my role in this study is a representation of my actual contribution, and I hereby give my consent that this work may be published as part of the MSc dissertation of J van der Westhuizen.

Miss J. van der Westhuizen

Prof S. Mason

Dr M. van Wyk

CHAPTER 1 INTRODUCTION

The purpose of the first chapter is to introduce the topic of this dissertation to the reader and provide sufficient background information that is integral to the understanding of the information to follow. This chapter will be broken down into the background, problem statement and lastly, the aims and objectives.

1.1 Background

Metabolomics is a growing field that focuses on the study of small molecules that provide insights into the link between genotype and phenotype (Schrimpe-Rutledge *et al*., 2016). Untargeted metabolomics approaches are usually focused on generating hypotheses and collecting as much information as possible about all present metabolites (Schrimpe-Rutledge *et al*., 2016). These untargeted approaches, especially proton nuclear magnetic resonance (^{1}H-NMR)-based methods, have many applications. One of these applications is highlighted by a study conducted by Silva and colleagues (2019), where untargeted urinary ^{1}H-NMR-based metabolomics approaches were studied as a potential way of discovering biomarkers to be used in breast cancer detection. This study, among others, shows the vast potential and strengths of untargeted ^{1}H-NMR metabolomics methods.

In order to use these methods to discover biomarkers, large-scale studies with thousands of samples are required, making it a complex, time-consuming, and sometimes expensive process (Miggiels *et al*., 2019). For this reason, new advances in technology are rapidly being developed to make this process cheaper, less time-consuming and more comprehensive (Miggiels *et al*., 2019). These technological advances include BQuant - a probabilistic approach towards the identification of ^{1}H-NMR metabolites introduced by Zheng *et al*. (2011), an R package for automated metabolite analysing presented by Hao *et al*. (2012), a web server application called MetaboHunter introduced by Tulpan *et al*. (2011) and BAYESIL – a fully automated ^{1}H-NMR spectral profiling technique (Ravanbakhsh *et al*., 2015), which is the focus of this study.

The BAYESIL system will be discussed in greater detail in this dissertation, where it will also be compared to the currently used in-house processing method of untargeted ^{1}H-NMR metabolomics data, applied to treated and untreated TB patients' urine samples.

1.2 Problem statement

In-house processing of untargeted ^{1}H-NMR metabolomics data requires an NMR metabolomics specialist and a biostatistician working in synergy. This in-house processing can be time-consuming and is vulnerable to human fatigue and error.

Therefore, the research questions are as follows: Can the processing of untargeted ^{1}H-NMR metabolomics data of complex urine samples (treated TB patients) be automated using online tools? What are the pitfalls?

BAYESIL is a fully automated ^{1}H-NMR metabolite profiling method that is free to use and easily accessible online. MetaboAnalyst is a web-based tool suitable for comprehensive metabolomics data analysis and interpretation. Combining these online platforms could potentially automate the processing of untargeted ^{1}H-NMR metabolomics data, which is greatly needed; however, automated systems are still in their infancy. For this reason, the automated processing methods still need to be tested against the established in-house metabolomics methods currently used. Furthermore, robustness needs to be tested by means of complex biological samples, such as urine samples collected from treated and untreated TB patients. Parallel determination of the metabolic effect of TB treatment through urinary profiling via both the in-house and automated method needs to be compared, and limitations of automation identified.

1.3 Aims and objectives

This study aimed to determine if automation of the processing of untargeted ^{1}H-NMR metabolomics data had the capability to obtain a comprehensive and comparable metabolic profile of treated and untreated TB patient urine samples.

Objectives to be met to achieve this aim are listed below:

1) <u>The manual method</u>

    1.1 Manually prepare a spectral binned data matrix of the raw ^{1}H-NMR output.
    1.2 Subject this data matrix to statistical analysis via MetaboAnalyst.
    1.3 Identify ^{1}H-NMR spectral regions (bins) that differentiate between experimental groups within this data matrix and assign metabolite names to them.
    1.4 Calculate absolute concentrations (µM) of these differentiating metabolites.
    1.5 Subject quantified differentiating metabolites to further statistical analysis to obtain a comparable metabolite profile.

2) <u>The automated BAYESIL method</u>

    2.1 In parallel, upload raw ^{1}H-NMR spectral data to BAYESIL website and combine metabolite profile data into a single data matrix of metabolite names and concentrations.

2.2 Upload the BAYESIL dataset to the MetaboAnalyst website and perform the same pre-processing and statistical analysis as performed on the manually generated dataset.

3) Compare the outputs of the in-house and automated processing methods and identify any differences and what they imply, and pitfalls for novice users.

1.4 References

Hao, J., Astle, W., De Lorio, M. & Ebbels, T.M. 2012. BATMAN – an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics*, 28(15):2088-2090. DOI: 10.1093/bioinformatics/bts308

Miggiels, P., Wouters, B., van Western, G.J.P., Dubbelman, A. & Hankemeier, T. 2019. Novel technologies for metabolomics: More for less. *Trends in Analytical Chemistry*, 120, art. #115323. DOI: 10.1016/j.trac.2018.11.021

Ravanbakhsh, S., Liu, P., Bjordahl, T.C., Mandal, R., Grant, J.R., Wilson, M., Eisner, R., Sinelnikov, I., Hu, X., Luchinat, C., Greiner, R. & Wishart, D.S. 2015. Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS ONE*, 10(5), e0124219. DOI: 10.1371/journal.pone.0124219

Schrimpe-Rutledge, A.C., Codreanu, S.G., Sherrod, S.D. & McLean, J.A. 2016. Untargeted metabolomics strategies – challenges and emerging directions. *Journal of the American Society for Mass Spectrometry*, 27(12):1897-1905. DOI: 10.1007/s13361-016-1469-y

Silva, C.L., Olival, A., Perestrelo, R., Silva, P., Tomas, H. & Camara, J.S. 2019. Untargeted urinary ^{1}H NMR-based metabolomic pattern as a potential platform in breast cancer detection. *Metabolites*, 9(11), art. #269. DOI: 10.3390/metabo9110269

Tulpan, D., Léger, S., Belliveau, L., Culf, A. & Čuperlović-Culf, M. 2011. MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. *BMC Bioinformatics*, 12(1):400. DOI: 10.1186/1471-2105-12-400

Zheng, C., Zhang, S., Ragg, S., Raftery, D. & Vitek, O. 2011. Identification and quantification of metabolites in 1H-NMR spectra by Bayesian model selection. *Bioinformatics*, 27(12):1637-1644. doi:10.1093/bioinformatics/btr118

CHAPTER 2 LITERATURE STUDY

This chapter provides an overview of the literature relevant to this study and aims to provide the reader with sufficient background knowledge on the topic and related concepts. To set this chapter in motion, the field of metabolomics will be discussed as well as the significance and applications thereof. Following this, a very brief overview of TB and its metabolism will be given, and a short description of sample types used in TB studies. The subsequent parts of this chapter will cover metabolite profiling, the ^{1}H-NMR platform, and automation of metabolomics. Following this, the focus will be shifted to the two online tools around which the automated processing of this study is centred, namely: BAYESIL and MetaboAnalyst.

2.1 The field of metabolomics

2.1.1 What is it?

Metabolomics is a fast-growing field that involves the identification of all metabolites present in an organism, the quantification of their concentrations/abundances, as well as the detection of any changes in these concentrations/abundances as a response to the environment, drugs or any other stimulus, in order to provide information that can be used to elucidate the physiological state of the organism (Beckonert *et al*., 2007; Larive *et al*., 2015; Markley *et al*., 2017; Silva *et al*., 2020; Yanes *et al*., 2011). The field of metabolomics is one of the 'omics' sciences, including proteomics, genomics, and transcriptomics. These other 'omics' sciences are known to be high-throughput sciences, while metabolomics is relatively low-throughput (Ravanbakhsh *et al*., 2015), meaning that valuable information is processed at a moderate to slow rate.

There are two main approaches associated with metabolomics: targeted and untargeted analyses. Targeted analyses focus on a specific metabolite or a small group of pre-defined metabolites and are usually affiliated with studies driven by a specific goal or hypothesis (Yanes *et al*., 2011). These analyses are generally more precise and more quantitative (i.e., concentration data is generated), making targeted analyses more reproducible and more time consuming (Lipfert *et al*., 2019). On the other hand, untargeted analyses, such as this study, are concerned with identifying and quantifying as many metabolites as possible, which can reveal the biochemical pathways that are affected by a specific change, whether they have been previously explored or not (Yanes *et al*., 2011). Untargeted approaches take less time, are more open-ended and can be applied in a variety of different situations but are usually less quantitative (i.e., abundance data is generated) and not as reproducible (i.e., metabolite names are not always assigned with certainty) (Lipfert *et al*., 2019).

2.1.2 Significance & applications

The uniqueness of metabolomics lies in its ability to provide information about the interaction between genes and the environment (Ravanbakhsh *et al*., 2015), allowing metabolomics to become a valuable asset to many types of research that involve the integrated use of various analytical platforms. Some applications of metabolomics include the monitoring of drugs and other medical interventions, diagnosis of diseases, and the analysis of biochemical pathways in the body, as well as the effect of external stressors such as diet, aging, mutations, and lifestyle on these pathways (Markley *et al.*, 2017).

Since metabolomics has the capability to obtain a so-called “profile” of all metabolites in a system, as will be covered in more detail in Chapter 2.3, it has been applied extensively to study metabolites that are affected by the interaction between hosts and pathogens, allowing the identification of biomarkers that can be used to diagnose infectious diseases caused by pathogens (Izquierdo-Garcia *et al*., 2020). The information obtained from metabolomic studies provides a much-needed understanding of the biological processes occurring in an organism, the most valuable application of metabolomics in recent times is the discovery of new biomarkers for diseases (Silva *et al*., 2020), including the disease of interest to this study, i.e., TB.

2.2 Tuberculosis (TB)

2.2.1 TB metabolism

TB is an infection caused by the obligate pathogen *Mycobacterium tuberculosis*, whose only known host is human beings (Rhee, 2013; Warner, 2014). For this reason, the metabolism of this microbe has evolved in such a way that it functions in both the physiology and pathogenesis of its human-dependent life cycle (Rhee, 2013; Warner, 2014). For this reason, the metabolism of *M. tuberculosis* has been the topic of many research studies, to gain insight into how this pathogen functions and possibly discover new treatment options for this infectious disease (Izquierdo-Garcia *et al*., 2020). Since the metabolism has such a vast effect on this pathogen, metabolomic studies have been applied to study the progression of TB, obtain metabolic profiles for this disease and to monitor the responses to TB treatment (Izquierdo-Garcia *et al*., 2020).

One such study was conducted by Combrink and colleagues in 2019 and is titled ‘Time-dependent changes in urinary metabolome before and after intensive phase tuberculosis therapy: a pharmacometabolomics study’. The main findings of this study were: reduced levels of oxidative stress during treatment, various enzymes were time-dependently induced or inhibited because of altered oxidative stress levels as well as in response to drugs, an upregulation of the urea cycle, and

alterations to insulin production (Combrink *et al*., 2019). As stated by Combrink and colleagues (2019), the results obtained in their study provide valuable information regarding the mechanism of action of TB drugs as well as their metabolism and side effects.

2.2.2 Sample types

The most widely used sample type for TB studies is sputum, but recently a great deal of effort has been made to discover new sample types of similar utility, such as exhaled breath, blood, and urine (Peter *et al*., 2010). Urine has become a very attractive option for many reasons: it is in ample supply, it is easily obtained in a non-invasive manner, it is easily stored and processed, and it is sterile, thus lowering the risk of infection to those that handle it during collection and analysis (Izquierdo-Garcia *et al*., 2020; Peter *et al*., 2010). Urine has proven to be a feasible alternative to sputum in a study conducted by Izquierdo-Garcia and colleagues (2020) where they succeeded in developing an NMR-based metabolomics approach for the search of biomarkers to aid in the diagnosis of TB, using urine as their sample type.

For this study, treated TB patient urine samples were used, specifically for their complexity. What is meant by this is that these samples do not only contain the usual metabolites from typical metabolic processes. Since these samples belong to patients that have been infected with TB, there will be differences in the metabolites that are detected as well as their concentrations brought about by the TB infection, as well as differences caused by drug intervention during TB treatment. This complexity is desired so that the accuracy and reliability of online data processing tools can be tested, since the samples used in biological studies are rarely straightforward and of simple composition.

2.3 Metabolite profiling

The term ‘metabolite profiling’ refers to the process of obtaining and scrutinizing a dataset of metabolite concentration/abundances for a specific biological sample or organism (Ravanbakhsh *et al*., 2015). This is useful because diseases (and medications) can cause changes to these metabolite concentrations, which will be picked up in the biofluid samples being analysed (Ravanbakhsh *et al*., 2015). The metabolites that are usually affected by a certain disease are called biomarkers of the disease and are very useful for diagnostic purposes and the development of new treatment options (Silva *et al*., 2020).

Since metabolites are naturally occurring molecules, they are very diverse in chemical and structural properties, making metabolite profiling quite an arduous process (Yanes *et al*., 2011). Often, analytical platforms available to obtain metabolic profiles are used in conjunction. Some of these can detect more metabolites than others (e.g., GC-TOF-MS), while others detect less metabolites

with a higher level of confidence (e.g., ^{1}H-NMR spectroscopy). For this reason, it is important to consider the goal of a specific study before determining the experimental and measurement design (van der Westhuizen, 2020). For this study, the ^{1}H-NMR spectroscopy platform was used.

2.4 ^{1}H-NMR spectroscopy

2.4.1 Introduction

First reported in 1963 by Jungnickel and Forbes, ^{1}H-NMR spectroscopy is now a well-known analytical technique that is usually used to determine the structures of molecules being analysed (Bharti & Roy, 2012). The term ^{1}H-NMR spectroscopy can be broken down into two parts:

1. NMR - refers to the scientific technique in which certain types of nuclei can selectively absorb high-frequency radio waves when subjected to a strong magnetic field (The Editors of Encyclopaedia Britannica, 1998). The type of nucleus being used in this specific application is ^{1}H (protons); hence, ^{1}H-NMR.

2. Spectroscopy – refers to the field of science in which the ability of matter to absorb and release light and other forms of radiation is studied (The Editors of Encyclopaedia Britannica, 1999).

There are other forms of NMR, each with their own applications, that use alternative nuclei such as ^{31}P, ^{15}N, and ^{13}C, but because these nuclei are not as naturally abundant as ^{1}H, they are less suitable for use in metabolomics (Bharti & Roy, 2012; Markley *et al*., 2017).

As mentioned by Markley *et al*. (2017), NMR methods have both one-dimensional (1D), and two-dimensional (2D) approaches. In 1D approaches, which are the most widely used, the signals are binned and analysed or fitted to known signal patterns to identify metabolites present in the sample, whereas 2D approaches such as ^{1}H-^{1}H COSY are used when more accurate metabolite identification is desired (Markley *et al.*, 2017). According to Ludwig and Viant (2010), 2D NMR spectroscopy methods make the spectra easier to interpret by lowering the amount of peak crowding by means of spreading the overlapping resonances into a second dimension.

2.4.2 Principle of the technique

Without diving too far into the complex world of quantum mechanics, this section of the dissertation aims to briefly explain exactly how NMR spectroscopy works. Firstly, it is important to know that all charged nuclei possess a quality allowing them to spin and these spinning nuclei tend to act like tiny individual bar magnets or magnetic moments with nuclear spins that can be aligned in any direction (Keeler, J., 2002). When these spinning nuclei are placed within an externally

applied magnetic field, their magnetic moments align with the direction of the applied field, resulting in bulk magnetization of the sample (Keeler, J., 2002). A pulse of radiofrequency radiation is then applied, causing the magnetization to tilt away from the z-axis and oscillate around the direction of the magnetic field in a specific type of motion referred to as Larmor precession, which is detected during an NMR analysis (Keeler, J., 2002).

To detect what is known as the 'free induction decay' (FID) signal, the sample is placed in a small wire coil so that when the magnetization crosses the coil, it induces a current which is amplified and recorded as a time-dependent signal (Keeler, J., 2002). Eventually, this signal starts to decay and lose its strength due to the tendency of the magnetization, like everything else in nature, to return to its equilibrium position in a process known as relaxation – hence the name of the signal is 'free induction decay' (Keeler, J., 2002). The result of this is a time-dependent signal that needs to be converted to a frequency-dependent spectrum of chemical shifts, which is made possible by a mathematical process called Fourier transformation, which is performed automatically by the NMR software (Keeler, J., 2002).

A metabolite that is detectable by ^{1}H-NMR will contain at least one proton linked to a carbon, and these protons each produce at least one peak (Zheng *et al*., 2011). The number, positions, and height ratios of the peaks in the spectrum produced by a metabolite are unique to that specific metabolite, like a fingerprint is unique to every human being, and is dependent on molecular structure (Zheng *et al*., 2011). These peak patterns are unique to each metabolite and can be saved as spectral library databases for later use in identifying metabolites in a sample by matching the obtained spectra and reference spectra in the database (Zheng *et al*., 2011). An example of an online database that stores ^{1}H-NMR reference spectra is the human metabolome database (HMDB; www.hmdb.ca).

2.4.3 Typical ^{1}H-NMR metabolomics method

2.4.3.1 Sample preparation

Most other analytical methods used in metabolomic studies require complex sample preparation techniques. For example, mass spectrometry requires ionization and sometimes derivatization of compounds (Markley *et al.*, 2017; Wishart, 2019). One of the most significant positive attributes of ^{1}H-NMR spectroscopy is that very little, if any, sample preparation is needed, depending on the biofluid that is used for the analysis (Markley *et al.*, 2017). Biofluids such as urine require very little sample preparation. Serum and plasma require extra steps to be taken to remove interfering compounds such as proteins and lipids, but these steps are not complex and usually only require the addition of a reagent such as methanol or the inclusion of centrifugal filtration (Markley *et al.*, 2017).

2.4.3.2 Signal generation

To obtain a spectrum of peaks, a useable signal must be generated from the sample. After placing the samples in the 500 MHz ^{1}H-NMR spectrometer, a signal is detected. This signal is produced by the movement of protons and their magnetic moments within an applied magnetic field. This movement is as a result of an applied radiofrequency pulse which excites the protons to a higher energy level, as explained in section 2.4.2 above. When these protons return to their equilibrium state, energy is released and detected as the free induction decay signal. To reiterate, Fourier transformation is then performed to convert the signal as a function of time to a frequency domain function, resulting in a spectrum that displays peak intensity in relation to chemical shift (Zheng *et al*., 2011).

2.4.3.3 Spectral & data processing steps

Before the spectra obtained from ^{1}H-NMR analysis can be seen as useful, a variety of spectral and data processing steps must be carried out. Figure 2-1 below, taken from Emwas *et al*. (2018), provides a useful summary of the steps that are typically followed when working with urinary ^{1}H-NMR data.

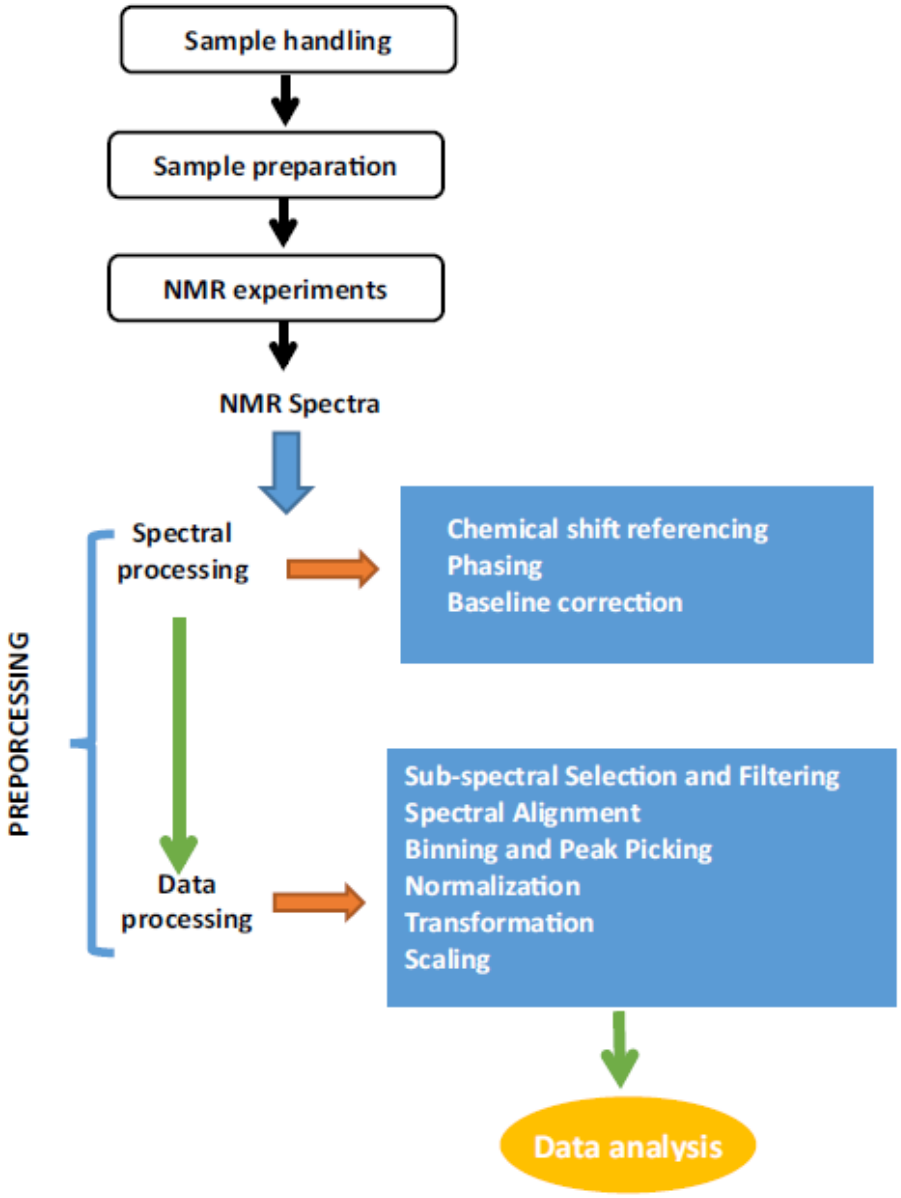


Figure 2-1: **Summary of spectral processing and data processing steps on urinary ^{1}H-NMR-data (Emwas *et al*., 2018).**

2.4.3.3.1 Spectral processing

The first important spectral processing step is chemical shift referencing, which involves the use of an internal standard such as trimethylsilyl propanoic acid (TSP) or sodium trimethylsilyl propane sulfonate (DSS) and is useful for compound identification, alignment of peaks and quantification of metabolite concentrations (Emwas *et al*., 2018). The next step is one of the most important spectral processing tools and is called phasing, which is usually carried out automatically by the spectrometer and aims to enhance the symmetry of all peaks in the spectrum (Emwas *et al*.,

2018). The third and final spectral processing step is crucial for removing spectral artefacts that arise from various technical faults such as electronic distortions and insufficient digital filtering, and is called baseline correction (Emwas *et al*., 2018). According to Emwas *et al*. (2018), baseline correction is usually a semi-automatic process and involves the use of software packages such as Chenomx NMR suite, NMRPipe or MNova by MestreLab Inc.

2.4.3.3.2 Data processing

The ^{1}H-NMR spectra obtained from urine samples are usually complex and contain thousands of peaks that tend to overlap and complicate the identification and quantification of metabolites. For this reason, a few data processing steps can be followed to simplify these spectra (Emwas *et al*., 2018).

The first technique is called sub-spectral selection and is a filtering method in which the parts of the spectrum that are not informative are discarded, such as the region of 0.00 to 0.60 ppm where no metabolite signals exist and the so-called water region from 4.50 to 4.90 ppm (Emwas *et al*., 2018).

The second important technique is spectral alignment, where peak positions in multiple spectra are shifted to allow the peaks of the same compounds to be directly aligned/overlaid to simplify the comparison and quantification of peaks across multiple spectra (Emwas *et al*., 2018).

Another important data processing technique worth mentioning is binning, which is usually applied to lower the resolution and account for any small misalignments that may be present (Lipfert *et al*., 2019). The process of binning is usually applied in untargeted metabolomics studies and involves the division of the spectra into smaller regions of set widths (0.01–0.02 ppm) to find the location of peaks and their spectral segments (Emwas *et al*., 2018; Lipfert *et al*., 2019). Typically, this will result in 500-1000 bins per spectrum which are then presented for statistical analysis and metabolite identities are then allocated to the most noteworthy bins (Emwas *et al*., 2018). Many parameters, such as sample pH, affect the position of peaks in a ^{1}H-NMR spectrum and could cause unwanted peak shifts (Emwas *et al*., 2018). For this reason, it is important to be cautious when binning NMR spectral data to prevent peaks from being divided or from falling into the wrong bins (Lipfert *et al*., 2019). According to Emwas *et al*. (2018), binning is most often carried out using an automatic algorithm.

The next step in data processing of NMR data is usually normalization, which aims to correct for concentration differences and dilution effects to allow the comparison across samples (Emwas *et al*., 2018). Normalization in urine samples can be carried out either numerically (relative to total

spectrum intensity) or physiologically (relative to creatinine). The latter is usually the preferred method of normalization for urine samples (Emwas *et al*., 2018).

The last data processing steps are scaling and transformation, which involve numerical functions that reduce the effect of varying orders of magnitude between metabolites and aim to enhance normality (Emwas *et al*., 2018). A commonly used scaling method is called centering. Centering adjusts all the values in the data so that they vary around a mean of zero to avoid highly concentrated metabolites affecting the other metabolites that are present in lower concentrations (Emwas *et al*., 2018). In turn, log transformations are often performed to improve normality by reducing the asymmetry when observed data mimic a skewed distribution.

2.4.3.4 Metabolite identification & quantification

After the data has been prepared and processed into a more functional format, metabolite annotation and identification can be carried out. There are a variety of different software tools available that automate this process and eliminate some of the complexity associated with NMR studies, such as MetaboAnalyst and BAYESIL used in this study. Other tools that are available to aid in this process include commercial packages such as NMR Suite by Chenomx, Bruker AMIX, and MNova by MestreLab (Emwas *et al*., 2018). After the identities of the metabolites have been found, their absolute concentrations can be calculated with the aid of the known concentration of the internal standard that was used.

The product of this entire process should therefore be a list of names of the metabolites that are present in the sample as well as their concentrations. This metabolite profile could be used for further biological interpretation if that was the objective of the specific study.

2.4.4 Advantages & limitations of NMR

2.4.4.1 Advantages

The NMR analytical platform carries many advantages over other techniques, such as being completely non-destructive, highly reproducible, robust, and reliable (Beckonert *et al*., 2007; Zheng *et al*., 2011). It is also important to note that a single spectrum obtained from an NMR experiment contains a unique richness of information, making it possible to simultaneously identify a wide range of structurally diverse metabolites and even novel compounds (Beckonert *et al*., 2007; Lipfert *et al.*, 2019; Wishart, 2019).

NMR spectroscopy is quantitative and unbiased, allowing it to have a wide range of diverse applications (Lipfert *et al.*, 2019; Wishart, 2019). Very little sample preparation is needed for an NMR experiment, meaning that no prior chromatographic separation is needed (Silva *et al*., 2020).

One of the main strong points of NMR spectroscopy is that it can detect metabolites at low concentrations and allows the identification of compounds with identical masses, since the detection of compounds in NMR is not based on their masses (Markley *et al*., 2017; Silva *et al*., 2020). NMR analysis is faster than most other platforms, taking only a few minutes to obtain a standard NMR spectrum and allows for the analysis of hundreds of samples per day on only one spectrometer (Beckonert *et al*., 2007).

The last important advantage that is relevant to mention is that NMR spectroscopy is easily adaptable to automation, which is explored further in this study (Lipfert *et al.*, 2019; Wishart, 2019).

2.4.4.2 Limitations

As with any other analytical platform, NMR is not perfect and has a few limitations. For example: in comparison to some other methods, such as MS-based metabolomics, NMR is less sensitive, more expensive to setup, and takes up more space in a laboratory environment (Wishart, 2019). The databases, spectral libraries, and software packages currently available for NMR-based metabolomics are not as extensive as those available for MS-based methods, which somewhat limits the applications of NMR (Wishart, 2019). Another important limitation is that NMR-based metabolomics is mostly restricted to 1D NMR, which has a lower resolution than the 2D techniques inherently used by most MS-based methods (Wishart, 2019). However, 2D spectra can be obtained via additional, albeit time–consuming, analysis on the NMR.

2.5 Automation of ^{1}H-NMR spectral data profiling

Recently there has been an increase in the amount of interest in the field of metabolomics, which has revealed a considerable weakness, i.e., metabolomics is not fully automated yet and still requires some manual interventions (Ravanbakhsh *et al*., 2015). This is mostly attributed to the fact that the analytical techniques used in the field of metabolomics, such as NMR spectroscopy, GC-MS, and LC-MS, were developed to identify pure compounds and not complex mixtures of metabolites (Ravanbakhsh *et al*., 2015). Most biological samples, especially the complex samples used in this study, contain hundreds and even thousands of metabolites, some being more abundant than others (Zheng *et al*., 2011). For this reason, the ^{1}H-NMR spectra obtained from metabolomics studies usually contain thousands of peaks that tend to overlap substantially and correspond to various metabolites (Ravanbakhsh *et al*., 2015; Zheng *et al*., 2011). This makes it increasingly difficult to interpret the spectra and identify and quantify the metabolites present in the mixture to obtain a metabolite profile.

To quantify metabolite concentrations, the intensity or integral area of their corresponding peaks in ^{1}H-NMR spectra is used, which is very challenging to determine if there is substantial overlapping of peaks (Jung *et al*., 2016). Therefore, it is apparent that the use of software programs is unavoidable, especially for the deconvolution of peaks (Jung *et al*., 2016). An example of a software program commonly used for peak deconvolution of an NMR spectrum is the Chenomx NMR Suite.

In the effort to fully automate the processing of ^{1}H-NMR metabolomics data from urine, a few semi-automated software packages have been developed. One such semi-automated method is called Signature Mapping (SigMa) and was presented by Khakimov *et al*. (2020). This new method was developed to increase the efficiency and lower the turnaround time for processing ^{1}H-NMR metabolomics data obtained from complex human urine samples into a table of noteworthy metabolites (Khakimov *et al*., 2020). Khakimov *et al*. (2020) explained that the SigMa approach divides the ^{1}H-NMR spectra into three significant variables, namely signature signals of known metabolites, spin systems of unknown metabolites and bins that represent complex unresolved spectral regions containing signals from more than one metabolite. SigMa also provides visualization, normalization, editing and alignment of data, an algorithm that automatically carries out peak selection, and a SigMa library of chemical shifts to detect the metabolites present in the urine samples (Khakimov *et al*., 2020). To quantify metabolites, two different approaches are used depending on the variable being quantified. Multivariate curve resolution modelling is used for the signature signals and the unknown spin systems, while integration or summation of values is used to quantify the bins (Khakimov *et al*., 2020). Although this system has achieved good results thus far, it is still very new and requires more time to allow for optimization.

Filntisi *et al*. (2017) mention a few other semi-automated methods that are currently available, such as the BATMAN package which is a probabilistic method based on Markov chain Monte Carlo and Metropolis-Hastings's block updates, an algorithm that fits spectra automatically called AutoFit, and MetaboHunter software that matches peaks on a ^{1}H-NMR spectrum to reference spectra in widely available databases (Filntisi *et al*., 2017).

The main concerns around these software tools are highlighted by Cardoso and colleagues (2021), where it is made clear that most of them concentrate on the quantification of metabolites, which tends to limit the actual number of metabolites that these tools can identify. Another obstacle highlighted by Cardoso *et a*l. (2021) is that some of these software tools, such as BAYESIL, have only been found to be reliable when identical conditions to those stipulated by them have been used during sample preparation and analysis. To combat these obstacles, Cardoso and colleagues (2021) designed and tested new algorithms that aimed to correctly annotate metabolites present in 1D ^{1}H-NMR samples. The best performing algorithm of their study is called

NMRFinder, which matches the sample with every compound in the reference library, regardless of the conditions used during data acquisition (Cardoso *et al*., 2021). Their study suggests that NMRFinder is a reliable algorithm to use for metabolite annotation and that it could possibly be a very valuable tool in the future to aid in metabolite identification (Cardoso *et al*., 2021). Although NMRFinder has achieved favourable results so far, it is still a very new technique that has been around for less than a year, whereas BAYESIL has been well-established for a few years, for this reason BAYESIL is used in this study.

Hence, ^{1}H-NMR spectral data profiling automation is topical, with numerous new online software rapidly emerging. For my MSc study, two well-established online metabolomics applications were used: BAYESIL, for identifying and quantifying compounds and MetaboAnalyst, for statistical analysis. These applications will be discussed in the following paragraphs.

2.5.1 BAYESIL

First proposed by Ravanbakhsh *et al*. (2015), BAYESIL is an online, fully automated spectral profiling application for ^{1}H-NMR-based metabolomics that can carry out spectral processing and spectral profiling of 1D ^{1}H-NMR spectra automatically without the need for manual intervention. According to Ravanbakhsh *et al*. (2015), the BAYESIL system is based on a statistical method called the sequential Monte Carlo inference method and uses probabilistic graphical modelling to determine the metabolic profile that is the most probable for the complex mixture being analysed.

Starting from the raw spectrum that is uploaded onto the BAYESIL website, all spectral processing steps are carried out automatically by the software including zero-filling, Fourier transformation, phasing, baseline correction, smoothing, chemical shift referencing and deconvolution (Ravanbakhsh *et al*., 2015). By doing this, any differences in results that are caused by differing opinions and bias among analysts are avoided, allowing an increase in the reproducibility, uniformity, and consistency of spectral profiling (Ravanbakhsh *et al*., 2015). According to Lipfert and colleagues (2019), BAYESIL has an accuracy of 95-100% when it comes to fitting the spectra; however, this accuracy is only achieved by following the BAYESIL sample preparation protocol. BAYESIL also carries out absolute quantification of the metabolites present in the sample by using a known concentration of internal standard (DSS or TSP) (Ravanbakhsh *et al*., 2015). This process takes approximately 7 minutes per spectrum (Lipfert *et al*., 2019), which is considerably faster than doing it manually, which usually takes a few hours.

The BAYESIL system was initially developed for use on serum and cerebrospinal fluid (CSF) samples and has produced good results when applied to these biofluids, obtaining metabolic profiles within 10% of those generated by a human counterpart (Ravanbakhsh *et al*., 2015). In a

study conducted by Filntisi *et al*. (2017), the BAYESIL system was tested on different biological samples and produced good results when applied to amniotic fluid samples.

Although BAYESIL has proved its potential and value, it is still a relatively new method and remains limited in its applications, requiring more research to ascertain its reliability and accuracy (van der Westhuizen, 2020).

2.5.2 MetaboAnalyst

MetaboAnalyst is an online application (https://www.metaboanalyst.ca) that is free to use and has a range of different functions, mainly for data analysis, such as statistical analysis, discovery, and analysis of biomarkers, as well as the analysis of metabolic pathways (Selegato *et al.*, 2019; Gowda & Raftery, 2019). MetaboAnalyst is recommended by Emwas *et al*. (2018) as a valuable software tool for ^{1}H-NMR metabolomics, specifically for multivariate analysis, as well as the annotation of metabolites and biological interpretation of data. A variety of different data types can be processed by the MetaboAnalyst software, from compound concentration tables to binned data and lists of peaks from NMR or MS analyses (Xia *et al*., 2009). The MetaboAnalyst software includes a wide range of different analysis options for normalization, identification of features, clustering, and classification of data (Xia *et al*., 2009).

MetaboAnalyst (version 5.0), as introduced by Pang *et al*. (2021), was used in this MSc study. This version focuses on narrowing the gap between raw data and useable information for global metabolomics based on high-resolution mass spectrometry and includes improvements to the web interface, graphics, overall performance and user experience (Pang *et al*., 2021).

Figure 2-2 below, taken from Xia *et al*. (2009), is a diagram outlining the workflow used by MetaboAnalyst and the different data processing options that are available. As explained by Xia *et al*. (2009), the input data is first converted to a data matrix that is compatible with the MetaboAnalyst software. After that, algorithms are applied to carry out normalization, analysis, and annotation of the data. Once statistical analysis is complete, a full PDF report, processed data, and high-resolution images can be downloaded (Xia *et al*., 2009).

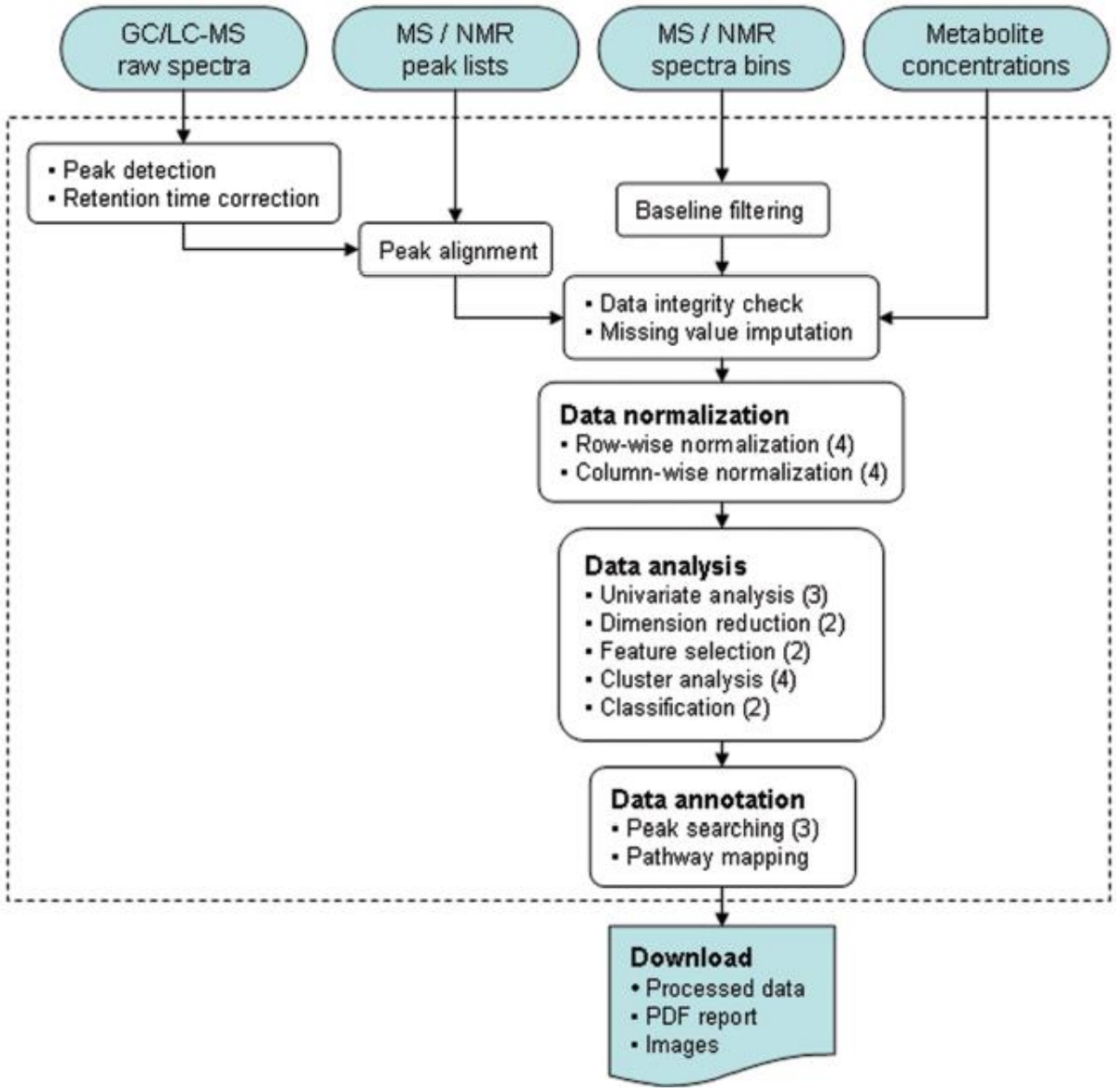


Figure 2-2: **A diagram illustrating MetaboAnalyst workflow and data processing options available on the website (Xia *et al*., 2009).**

2.6 References

Beckonert, O., Keun, H.C., Ebbels, T.M., Bundy, J., Holmes, E., Lindon, J.C. & Nicholson, J.K. 2007. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum, and tissue extracts. *Nature Protocols*, 2(11):2692-2703. DOI: 10.1038/nprot.2007.376

Bharti, S.K. & Roy, R. 2012. Quantitative 1H-NMR spectroscopy. Trends in Analytical Chemistry, 35:5-26. DOI: 10.1016/j.trac.2012.02.007

Cardoso, S., Cabral, D., Maraschin, M. & Rocha, M. 2021. NMRFinder: a novel method for 1D 1H-NMR metabolite annotation. *Metabolomics*, 17(2):1-12. DOI: 10.1007/s11306-021-01772-9

Combrink, M., Du Preez, I., Ronacher, K., Walzl, G. and Loots, D.T. 2019. Time-dependent changes in urinary metabolome before and after intensive phase tuberculosis therapy: a pharmacometabolomics study. *Omics: a journal of integrative biology*, *23*(11):560-572.

Emwas, A.H., Saccenti, E., Gao, X., McKay, R.T., Martins dos Santos, V.A.P., Roy, R. & Wishart, D.S. 2018. Recommended strategies for spectral processing and post-processing of 1D 1H-NMR data of biofluids with a particular focus on urine. *Metabolomics*, 14(3):31. https://doi.org/10.1007/s11306-018-1321-4

Filntisi, A., Fotakis, C., Asvestas, P., Matsopoulos, G.K., Zoumpoulakis, P. & Cavouras, D. 2017. Automated metabolite identification from biological fluid 1H-NMR spectra. *Metabolomics*, 13(12):146. DOI 10.1007/s11306-017-1286-8

Gowda, G.A.N and Raftery, D. 2019. Overview of NMR spectroscopy-based metabolomics: opportunities and challenges. In: Gowda, G.A.N. & Raftery, D., eds. *NMR-Based Metabolomics*. New York: Springer. pp 3-14.

Izquierdo-Garcia, J.L., Comelia-Del-Barrio, P., Campos-Olivas, R., Prat-Aymerich, C., De Souza-Galvao, M.L., Jimenez-Fuentes, M.A., Ruiz-Manzano, J., Stojanovic, Z., Gonzalez, A., Serra-Vidal, M., Garcia-Garcia, E., Muriel-Moreno, B., Millet, J.P., Molina-Pinargote, I., Casas, X., Santiago, J., Sabria, F., Martos, C., Herzmann, C., Ruiz-Cabello, J. & Dominguez, J. 2020. Discovery and validation of an NMR-based metabolomic profile in urine as TB biomarker. *Scientific Reports*, 10(1), art.#22317. DOI: 10.1038/s41598-020-78999-4

Jung, Y.S., Hyeon, J. & Hwang, G. 2016. Software-assisted serum metabolite quantification using NMR. *Analytica Chimica Acta*, 934:194-202. https://doi.org/10.1016/j.aca.2016.04.054

Jungnickel, J.L. & Forbes, J.W. 1963. Quantitative measurement of hydrogen types by integrated nuclear magnetic resonance intensities. *Analytical Chemistry*, 35(8): 938-942.

Keeler, J. 2002. Understanding NMR Spectroscopy. 1st ed. Chichester, West Sussex: John Wiley & Sons, Inc.

Khakimov, B., Mobaraki, N., Trimigno, A., Aru, V. & Engelsen, S.B. 2020. Signature mapping (SigMa): An efficient approach for processing complex human urine 1H NMR metabolomics data. *Analytica Chimica Acta*, 1108: 142-151. DOI: 10.1016/j.aca.2020.02.025

Larive, C.K., Barding, G.A., Jr. & Dinges, M.M. 2015. NMR spectroscopy for metabolomics and metabolic profiling. *Analytical Chemistry*, 87(1):133-146. DOI: 10.1021/ac504075g

Lipfert, M., Rout, M.K., Berjanskii, M. & Wishart, D.S. 2019. Automated Tools for the Analysis of 1D-NMR and 2D-NMR Spectra. In: Gowda, G.A.N. & Raftery, D., eds. *NMR-Based Metabolomics*. New York: Springer. pp 429-449

Ludwig, C. and Viant, M.R. 2010. Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochem Anal* 21(1): 22-32. DOI: 10.1002/pca.1186

Markley, J.L., Brüschweiler, R., Edison, A.S., Eghbalnia, H.R., Powers, R., Raftery, D. & Wishart, D.S. 2017. The future of NMR-based metabolomics. *Current Opinion in Biotechnology*, 43:34-40. http://dx.doi.org/10.1016/j.copbio.2016.08.001

Pang, Z., Chong, J., Zhou, G., de Lima Morais, D.A., Chang, L., Barrette, M., Gauthier, C., Jacques, P.E., Li, S. & Xia, J. 2021. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research*. DOI: 10.1093/nar/gkab382

Peter, J., Green, C., Hoelscher, M., Mwaba, P., Zumla, A. & Dheda, K. 2010. Urine for the diagnosis of tuberculosis: current approaches, clinical applicability, and new developments. *Current Opinion in Pulmonary Medicine*, 16(3):262-270. DOI:10.1097/MCP.0b013e328337f23a

Ravanbakhsh, S., Liu, P., Bjordahl, T.C., Mandal, R., Grant, J.R., Wilson, M., Eisner, R., Sinelnikov, I., Hu, X., Luchinat, C., Greiner, R. & Wishart, D.S. 2015. Accurate, fully automated NMR spectral profiling for metabolomics. *PLoS ONE*, 10(5), e0124219. DOI: 10.1371/journal.pone.0124219

Rhee, K. 2013. Minding the gaps: metabolomics mends functional genomics. *EMBO reports*, 14(11):949-950. https://doi.org/10.1038/embor.2013.155

Selegato, D.M., Pilon, A.C. and Neto, F.C. 2019. Plant metabolomics using NMR spectroscopy. In: Gowda, G.A.N. & Raftery, D., eds. *NMR-Based Metabolomics*. New York: Springer. pp 345-362.

Silva, R.A., Pereira, T.C.S., Souza, A.R. & Ribeiro, P.R. 2020. 1H-NMR based metabolite profiling for biomarker identification. *Clinica Chimica Acta*, 502:269-279. https://doi.org/10.1016/j.cca.2019.11.015

The Editors of Encyclopaedia Britannica. 1998. Nuclear magnetic resonance (NMR). In: Britannica academic. https://academic-eb-com.nwulib.nwu.ac.za/levels/collegiate/article/nuclear-magnetic-resonance/56443 Date of access: 7 May 2021.

The Editors of Encyclopaedia Britannica. 1999. Spectroscopy. In: Britannica academic. https://academic-eb-com.nwulib.nwu.ac.za/levels/collegiate/article/spectroscopy/110407 Date of access: 7 May 2021.

van der Westhuizen, J. 2020. Metabolite profiling of serum samples using ^{1}H-NMR spectral data: comparing an online, automated method (BAYESIL) vs current CHM manual method. Potchefstroom: North-West University. (Mini-dissertation – BSc Hons).

Warner, D.F. 2014. Mycobacterium tuberculosis metabolism. *Cold Spring Harbor Perspectives in Medicine*, 5(4), a021121. DOI: 10.1101/cshperspect.a021121

Wishart, D.S. 2019. NMR metabolomics: a look ahead. *Journal of Magnetic Resonance*, 306:155-161. https://doi.org/10.1016/j.jmr.2019.07.013

Xia, J., Psychogios, N., Young, N. and Wishart, D.S. 2009. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic acids research*, 37(suppl_2), pp. W652-W660. DOI: 10.1093/nar/gkp356

Yanes, O., Tautenhahn, R., Patti, G.J. & Siuzdak, G. 2011. Expanding coverage of the metabolome for global metabolite profiling. *Analytical Chemistry*, 83(6):2152-2161. DOI: 10.1021/ac102981k

Zheng, C., Zhang, S., Ragg, S., Raftery, D. & Vitek, O. 2011. Identification and quantification of metabolites in 1H-NMR spectra by Bayesian model selection. *Bioinformatics*, 27(12):1637-1644. doi:10.1093/bioinformatics/btr118

CHAPTER 3 STUDY DESIGN

The purpose of this chapter is to provide the reader with an overview of the design that was followed in this study to reach the objectives as outlined in Chapter 1.3. A detailed description of the methods is given in a manuscript submitted to the journal *Metabolomics* (IF: 4.29) – see Chapter 4. To avoid redundant repetition, Chapter 3 will therefore give an overall description of the design of the study and what was done during this MSc.

3.1 Sample description and ethics requirements

The samples used in this study were anonymised urine samples collected longitudinally from 46 TB-positive patients at different intervals during treatment, and 30 healthy control samples taken from individuals that were not TB-positive. These samples were specifically selected for this study based on their complexity, owing to both the disease and the medication metabolites expected to be present. The performance of the automated BAYESIL system has already been established with less complex samples, but it is unknown whether BAYESIL will be able to perform well when applied to such highly complex biological samples, hence the aim of this study.

The samples were obtained from a biorepository and ethics approval for the use of these samples has been granted by the Ethics Committee of the University of Stellenbosch (No.99/039). This MSc project falls under a larger study titled “Metabolomic investigations of tuberculosis for the purpose of improved characterization, diagnostics, and treatment”, which has NWU HREC approval under the following ethics number: NWU-00127-11-A1. A sub-study NWU HREC approval was obtained for this MSc study: NWU-00127-11-A1-02 (see Annexure A for NWU HREC approval letter).

3.2 Training and workshops

Before the commencement of the practical work for this study, I completed advanced NMR training – from the 6th of June 2021 to the 11th of June 2021. During this training, I re-familiarized myself with the theory and practise of NMR analysis. I gained valuable practical experience by working in the laboratory and following the standard operating procedures (SOPs) typically used for the NMR analysis of mock urine and serum samples. I was trained extensively in using the NMR spectrometer including loading of samples, calibration and running of the machine, quantitative assessment of spectrum quality and exporting of raw data. At the end of the training, I created a quantified data matrix of several metabolites from the mock samples analysed during the week of training. Annexure B contains the raw data and the calculations determining my intra-day and inter-day repeatability for both the urine and serum protocols. My coefficient of variance

percentage (CV) for the urine analysis ranged between 2.10% and 3.22%. I turn, my CV ranged between 0.69% and 3.19% for the serum analysis. These low CV values indicate good repeatability, making me sufficiently competent to carry out the practical aspect of this MSc.

I also attended an NMR-based metabolomics workshop hosted by Metabolomics South Africa in partnership with the University of South Africa (UNISA) and DIPLOMICS – from the 7th of September 2021 to the 9th of September 2021. This workshop was facilitated by Professor Gerhard Prinsloo from UNISA and included an introduction to NMR-based metabolomics, the experimental layout that is followed by their laboratory, and analysis and interpretation of NMR metabolomics data.

3.3 Untargeted ^{1}H-NMR metabolomics approach

Silva *et al*. (2020) mentioned that metabolomics usually follows one of two approaches: untargeted or targeted. Untargeted analysis is designed to detect and identify as many compounds as possible and involves the classification of the metabolome without any prior knowledge of the samples or which metabolites are present (Yanes *et al*., 2011). On the other hand, targeted analysis focuses on specific metabolites of interest.

To determine whether a targeted or untargeted approach should be followed, the purpose of the study had to be carefully considered. Since the aim of this study is centred around data processing, an untargeted ^{1}H-NMR metabolomics approach was decided to be best, to allow the collection of as much information as possible.

3.4 Sample preparation and analysis

BAYESIL is a fully automated ^{1}H-NMR metabolite profiling method that is free to use and easily accessible online. During my honour's degree in 2020, I worked with BAYESIL and tested its performance on quality control serum samples. Although BAYESIL does not have an algorithm designed specifically for urine, it is still the best established automated ^{1}H-NMR metabolite profiling method available currently. For this reason, I chose to test BAYESIL on these complex samples.

For the sample preparation, the protocol prescribed by BAYESIL was followed, with some minor alterations (described in Chapter 4.1.3.3) and included adjusting the pH of every sample to between 6.8 and 7.4, centrifugation at 12 000 *g*, the addition of NMR buffer and D_2O to the sample supernatant and lastly, further centrifugation at 12 000 *g*. From the final supernatant, the sample is transferred to a 5 mm glass NMR tube and randomly loaded into the NMR spectrometer to avoid confounding, with a pooled QC sample inserted at the beginning, middle and end of each

of the six batches, to evaluate the precision and reliability of the analysis. This pooled QC was created by combining a small volume (~70µL) from each sample. Once the samples were loaded, they were analysed by the ^{1}H-NMR spectrometer, following the parameters provided by BAYESIL. The sample preparation and spectral collection method is described in more detail in the manuscript provided in Chapter 4.

3.5 Data processing

For the purposes of this study, there were two data processing methods, namely: 1) The automated BAYESIL method which yields a metabolite profile containing metabolite names and absolute concentrations, as determined by the BAYESIL software, and 2) The manual binning method which yields a binned data matrix. In the following paragraphs, these two methods will be explained.

The BAYESIL system is fully automated, allowing it to perform spectral processing and profiling on the raw spectral data files. This system makes use of several phasing and baseline correction approaches to allow the automatic processing of raw spectra obtained from one dimensional ^{1}H-NMR analyses (Ravanbakhsh *et al*., 2015). According to Ravanbakhsh *et al*. (2015), BAYESIL does this by dividing the spectrum into blocks and using a probabilistic graphical model to represent the statistical relationship between these blocks. To find the most probable metabolic profile, BAYESIL applies an approximate inference method to this model as a substitute for spectral profiling (Ravanbakhsh *et al*., 2015). BAYESIL then uses a known concentration of a reference compound such as DSS or TSP to calculate the absolute concentrations of the detected metabolites (Ravanbakhsh *et al*., 2015). The result of the BAYESIL method is therefore a list of metabolite names, the certainty with which each name was assigned, and their absolute concentrations.

The spectral processing functions that are carried out automatically by BAYESIL include zero-filling, Fourier and Hilbert transformations, phasing, baseline correction, smoothing, chemical shift referencing and reference deconvolution (Ravanbakhsh *et al*., 2015). Figure 3-1 on the next page (taken from Ravanbakhsh *et al*., 2015), is a diagram indicating the spectral processing steps followed by the BAYESIL system.

The manual binning method that was followed is the standard ^{1}H-NMR binning method used at the CHM of the NWU. This method involves spectral processing using a program called Bruker Topspin (V3.5), which carries out chemical shift referencing, phasing and baseline correction. After the spectra were processed, there were some post-processing steps that had to be conducted. Firstly, the non-informative regions of the spectra were discarded, as well as the water region (4.69 – 4.90) of the spectra. After this, the spectral data were divided into bins with a set

interval width of 0.02 ppm and the data was normalized to the methylene moiety of creatinine at 4.05 ppm.

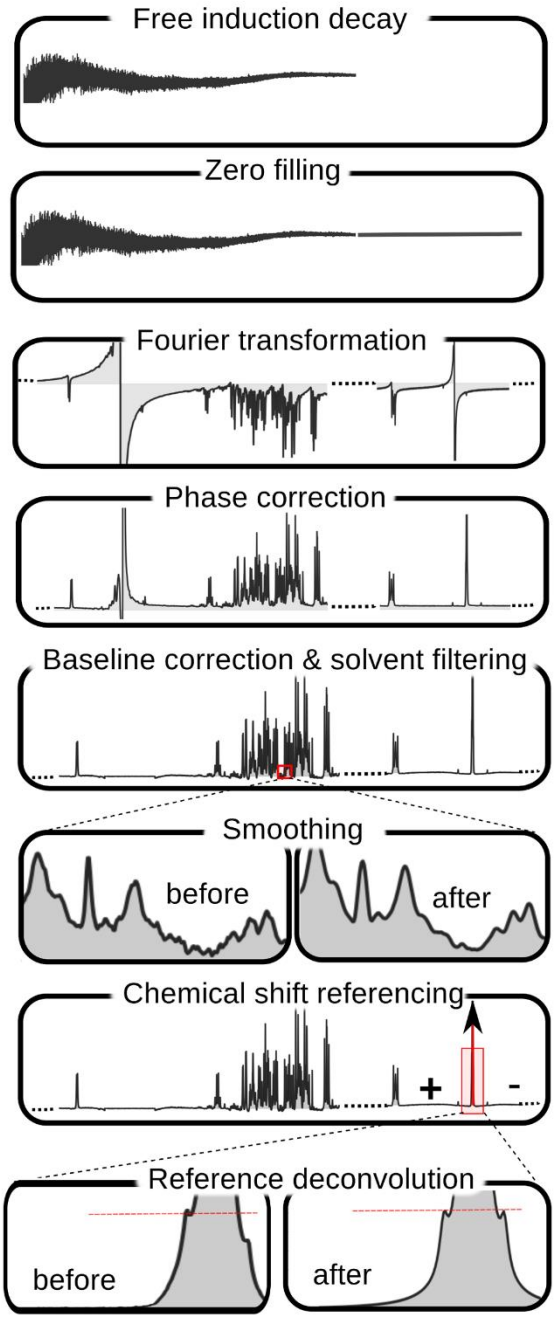


Figure 3-1: **Spectral processing steps in BAYESIL (taken from Ravanbakhsh *et al*., 2015)**

The next part of manual processing included noise removal and zero filtering. Firstly, the noise regions in the data matrix were identified. These noise regions are the parts of the spectra where there are no discernible peaks present, and hence, do not contribute to the information in a sample. To identify these noise regions, a software program called Bruker Amix (V3.9.14) was used to overlap the spectra obtained from all the samples. A qualitative assessment was then conducted and the bins that were considered noise in the spectra were identified.

Based upon an in-house method, a noise threshold was set based on the variability within identified noise regions with the aim to maximize data retention. All values below this threshold were set to 0. The manual data processing method yielded a binned data matrix, meeting all the requirements for upload onto MetaboAnalyst for statistical analysis.

3.6 Statistical analysis

The statistical analyses that were conducted during this study involved only one statistical tool namely, MetaboAnalyst, which is a very user-friendly and well-established online metabolomics suite (Chong *et al*., 2019). MetaboAnalyst is one of, if not the most used data analysis tool in Metabolomics. According to the 'User stats" tab on the MetaboAnalyst website (https://www.metaboanalyst.ca/MetaboAnalyst/docs/UserStats.xhtml), MetaboAnalyst has been used by more than 450,000 researchers across the world since the year 2017. A limitation of MetaboAnalyst is, however, that it does not have many effect size options. The role of effect sizes is being recognized more and more, with the American Psychological Association (APA) manual for publication published by Fidler (2010) strongly recommending the reporting and interpretation of effect sizes, due to their strong statistical power. MetaboAnalyst, however, provides only a fold change, which is not fully recognised as a measure of practical importance as it does not take variation into account, and basic model goodness-of-fit statistics.

Since there is a lack of standardization in data processing/statistical analysis, a biostatistician was consulted before statistical analysis was conducted, to determine which statistical tests would be best and what the cut-off criteria should be, based on their knowledge and experience.

A total of six data sets were subjected to statistical analysis via MetaboAnalyst. These data sets included the following three main datasets 1) Binned spectral data obtained from the manual method, with the aim of determining which bins were statistically significant; 2) The metabolite profile of the urine samples obtained from BAYESIL, to determine which of the identified metabolites were statistically significant; and 3) A list of quantified important metabolites, which correspond to the significant bins obtained from dataset 1, subjected to statistical analysis to obtain a metabolite profile that is more comparable to the BAYESIL metabolite profile. Each of the three main data sets were divided into two further datasets to compare BAYSIL's ability in the presence

of disease and treatment. These included one dataset for the comparison of healthy controls versus week 0 and one for the comparison of week 0 versus week 2. All these datasets followed the same steps on MetaboAnalyst, which will be discussed in the forthcoming paragraphs.

Once each dataset was uploaded, the first important step was data filtering, where the option to filter features if their relative standard deviations (RSDs) are higher than 30% in the QC samples was selected, as well as interquartile range (IQR) filtering. Excluding variables based on QC RSD values is a sensible thing to do since these variables could not be consistently measured. The IQR filter removes variables that showed too little variation to differ significantly between groups, so reducing the number of univariate tests required and consequently the impact of the false discovery rate correction.

Following this was normalization consisting of three categories: sample normalization, data transformation and data scaling. The binned data had already been normalized to creatinine, as part of data processing. The reason for this being that although the flow of urine does not remain constant throughout the day, the total creatinine output is usually always the same, making this a useful variable to use for normalization to account for dilution differences (Spierto *et al*., 1997). The BAYESIL data had also already been normalized by the software, so sample normalization was not repeated in MetaboAnalyst.

Transformation of the data is necessary to reduce the skew of data distributions that commonly occur in metabolomics data. In so doing, the validity of statistical tests assuming normality are improved. Reporting untransformed descriptive statistics is also not considered an accurate representation as these are not the values compared in the statistical tests used. The CHM prefers only to transform and not also scale data for univariate analysis because means associated with scaled data can be negative, which is unconventional in the clinical setting. Square root transformation was selected due to its ability to produce a similar pattern to log transformation, while also having the ability to work with zero values while having a positive effect on the heteroscedasticity of the data (van den Berg *et al*., 2006).

When multivariate tests were to be conducted, the data was square root transformed as well as auto scaled, since scaling cannot be ignored in the multivariate setting as differences in the orders of magnitude between metabolites will place greater importance on metabolites in higher abundance, which may not reflect biological importance. Auto scaling was chosen because it allows the comparison of metabolites based on the relationship between them, with all metabolites being equivalent to one another (van den Berg *et al*., 2006).

For the univariate analysis of each dataset, t-tests were performed, and fold changes were calculated, which were displayed as volcano plots. Since it is difficult to test all the underlying assumptions of the univariate tests for each metabolite, more robust methods are preferred. The assumption of homoscedasticity can be circumvented by using a variation on the t-test that can accommodate heteroscedasticity, which is why unequal group variances were assumed. When performing hypothesis testing at a univariate level across multiple compounds, the Type I error rate (α) does not remain constant. As explained by Bhandari, P. (2021), a Type I error occurs when you mistakenly consider something as statistically significant, when it is the result of random chance. The variable alpha (α) is a measure of how likely you are to make this mistake. With each additional test performed there is an increase in the likelihood of a false discovery (i.e., finding a p-value ≤ α). To improve the validity of findings, the rate at which such false discoveries are made over an increasing number of hypothesis tests needs to be controlled, which is made possible by the option to select an 'FDR-adjusted p value' on MetaboAnalyst. Correcting for multiple testing was then achieved controlling the False Discovery Rate (FDR).

For the chemometric (multivariate) analysis of each dataset, principal component analysis (PCA) and the partial least squares – discriminant analysis (PLS-DA) were carried out. PCA models are used to summarize variation in the data and visualize how the largest proportion of variability relates to the groups of interest. PLS-DA does the same but is explicitly informed of the groups represented in the data. The availability of this information tends to unduly inflate the power of a PLSDA model (overfit), so these models must always be validated to assess the severity of overfit. The most basic and lenient validation procedure involves leaving a single sample out repeatedly and rebuilding the model each time (LOOCV or leave-one-out cross-validation). LOOCV is complemented by the preferred performance metric, which in this instance is prediction accuracy.

For each dataset, MetaboAnalyst generated an analysis report containing all the important information. This report was downloaded, along with high-resolution images of the result from each statistical test.

Through the consultation and guidance of a biostatistician, specific rules were established as cut-off criteria to determine which bins/metabolites were significant. With the first dataset, the bins were selected as significant if they met the following criteria:

1) All latent variables (from PLS-DA) must be greater than or equal to 1, only if the PLS-DA model is validated.
2) The absolute value of the logarithm of the fold change (from the volcano plot) must be greater than 1.
3) FDR-adjusted p-value from the t-test must be smaller than or equal to 0.05.

The list of significant bins was then used to obtain metabolite identities by using two semi-automated software systems, namely Chenomx profiler (V8.2) and Bruker Amix (V3.9.14). Chenomx profiler was used to determine possible metabolite annotations for the peaks that fall within the significant bins, based on matches with pure compound libraries. Bruker Amix, together with pure compound spectral libraries, was used to confirm these possible annotations, and 2D correlation spectroscopy (COSY) and 2D 1H J-resolved (JRES) NMR spectroscopy were used to confirm these identities.

These significant metabolites were then quantified to obtain their absolute concentrations. The quantification process involved first determining the relative metabolite concentrations (mmol/mol creatinine) to account for dilution differences. The internal standard (TSP) with known concentration was then used to calculate the absolute value of creatinine per sample. The previously calculated relative concentration was then multiplied by the absolute concentration of creatinine (mM) per sample to obtain the absolute concentration of each metabolite (uM). The absolute quantified manual data was combined into dataset 3, which is now more comparable to the BAYESIL data as both contain absolute concentration values.

For datasets 2 and 3, the criteria were relaxed slightly to increase the number of metabolites selected, as these datasets form the core of the comparison between BAYESIL and the manual method. Metabolites were selected as significant if they met these criteria:

- Latent variable 1 or 2 must be greater than or equal to 1 – this is the only criteria that was relaxed (“and” changed to “or”).
- The absolute value of the logarithm of the fold change (from the volcano plot) must be greater than or equal to 1.
- The FDR-adjusted t-test p-value must be smaller than or equal to 0.05.

3.7 Guideposts

During this MSc, a lot was learned, and many mistakes were made. Most of the time, these mistakes were made due to gaps in knowledge from being a novice in the field of untargeted ^{1}H-NMR metabolomics data analysis. Through the guidance and assistance of my supervisors and other experts in this field, I picked up some tips along the way. These tips have been included as guideposts in Chapter 4 of this dissertation. These guideposts include discoveries/tips/assumptions that I have articulated during my MSc. The hope is that these guideposts can help the next novice navigate the world of untargeted ^{1}H-NMR metabolomics data processing more smoothly, by implementing the developed process and guidelines as a formal, peer-reviewed published SOP.

3.8 References

Bhandari, P. 2021. Type I & type II errors| differences, examples, visualizations. Scribbr. https://www.scribbr.com/statistics/type-i-and-type-ii-errors/. Date of access: 4 Mar. 2022.

Chong, J., Wishart, D. S., & Xia, J. 2019. Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis. *Current Protocols in Bioinformatics*, 68, e86. doi: 10.1002/cpbi.86

Fidler, F. 2010. *The American Psychological Association Publication Manual sixth edition: implications for statistics education*. Contributed paper at the 8th International Conference on Teaching Statistics, Slovenia. http://icots.info/8/cd/pdfs/contributed/ICOTS8_C156_FIDLER.pdf. Date of access: 4 Mar. 2022.

Ravanbakhsh, S., Liu, P., Bjordahl, T.C., Mandal, R., Grant, J.R., Wilson, M., Eisner, R., Sinelnikov, I., Hu, X., Luchinat, C., Greiner, R. & Wishart, D.S. 2015. Accurate, fully automated NMR spectral profiling for metabolomics. *PLoS ONE*, 10(5), e0124219. DOI: 10.1371/journal.pone.0124219

Silva, R.A., Pereira, T.C.S., Souza, A.R. & Ribeiro, P.R. 2020. 1H-NMR based metabolite profiling for biomarker identification. *Clinica Chimica Acta*, 502:269-279. https://doi.org/10.1016/j.cca.2019.11.015

Spierto, F., Hannon, W., Gunter, E. & Smith, S. 1997. Stability of urine creatinine. *Clinica Chimica Acta*, 264(2):227-232.

van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K. & van der Werf, M.J. 2006. Centering, scaling and transformations: improving the biological information content of metabolomics data. BMC Genomics, 7:142. DOI: 10.1186/1471-2164-7-142

Yanes, O., Tautenhahn, R., Patti, G.J. & Siuzdak, G. 2011. Expanding coverage of the metabolome for global metabolite profiling. *Analytical Chemistry*, 83(6):2152-2161. DOI: 10.1021/ac102981k

CHAPTER 4 A NOVICE'S GUIDE TO PROCESSING UNTARGETED ^{1}H NMR METABOLOMICS DATA

4.1 A novice's guide to processing untargeted ^{1}H NMR metabolomics data

Jessica van der Westhuizen, Mari van Reenen and Shayne Mason*

Human Metabolomics, Faculty of Natural and Agricultural Sciences, North-West University, Potchefstroom, South Africa.

Author details:

J van der Westhuizen: jessicavanderwesthuizen152@gmail.com (https://orcid.org/0000-0001-8483-3939)

M van Reenen: van.reenen.mari@gmail.com (https://orcid.org/0000-0002-5856-3258)

S Mason (*corresponding author): nmr.nwu@gmail.com (https://orcid.org/0000-0002-2945-5768)

NOTE:

Submitted to the journal *Metabolomics* (IF: 4.29) on 9 March 2022. Latest update (16 March 2022) from the online submission tracking system indicates that 3 Reviewers have been invited (i.e., it us under peer-review), see Annexure D for a screenshot indicating this. The guidelines for submission to this journal can be found at: https://www.springer.com/journal/11306/submission-guidelines and are included as Annexure F of this dissertation.

4.1.1 Abstract

Introduction: The processing of untargeted ^{1}H NMR metabolomics data is difficult, especially when extracted from complex biological matrices; hence, there is an increasing need for automated ^{1}H NMR spectral metabolite profiling. Here, the respective abilities of two approaches to generate metabolite profiles from complex urine samples are compared.

Methodology: Samples were analysed to obtain ^{1}H NMR spectral data. Two data processing methods were used: an in-house manual method and an automated method using the open-source software BAYESIL. The open-source software MetaboAnalyst was used for all statistical analyses. These approaches were compared to understand the risk/reward of high-throughput processing. The groups were purposefully selected to compare performance in the presence of exogenous metabolites from pathogens and medication.

Results: Comparing healthy controls with untreated tuberculosis cases, 14 and 8 important metabolites were identified via the BAYESIL and quantified manual methods, respectively. Comparing untreated with treated tuberculosis cases identified 22 and 16 important metabolites via the BAYESIL software and quantified manual methods, respectively.

Discussion and conclusion: Expert knowledge is essential when a comprehensive metabolic profile is desired, novel/exogenous metabolites are of interest, or the biological matrix is complex. By comparison, BAYESIL was less time-consuming and required little user expertise – it is recommended for novice NMR metabolomics analysts, with guideposts suggested here. BAYESIL's metabolite library is still limited and identification was often only approximate. The BAYESIL software is therefore not yet capable of producing comprehensive and accurate metabolic profiles from urine, but is well-suited to obtain a bird's-eye view of the corresponding metabolism responsible.

Keywords: proton nuclear magnetic resonance (^{1}H NMR) spectroscopy; Untargeted urinary metabolomics; Data processing; BAYESIL; Automated; MetaboAnalyst

4.1.2 Introduction

The processing of ^{1}H NMR metabolomics data is not fully automated and therefore still requires some manual interventions (Ravanbakhsh *et al*., 2015). This is mostly because the analytical tool used – NMR spectroscopy – was initially developed for the purpose of identifying single, pure chemical compounds and not complex biological mixtures of metabolites (Ravanbakhsh *et al*., 2015). Most biological samples contain hundreds of metabolites, of widely varying abundance, which correspond to hundreds, or even thousands, of peaks in the ^{1}H NMR spectra and which can overlap substantially (Ravanbakhsh *et al*., 2015; Zheng *et al*., 2011). This makes it increasingly difficult to interpret the spectra and identify and quantify the metabolites present in a metabolite profile. To simplify the data processing and interpretation of untargeted ^{1}H NMR metabolomics data, numerous examples of online software are emerging to automate this process. For this study, two well-established online tools for processing ^{1}H NMR metabolomics data were assessed: BAYESIL – for the automated identification and quantification of metabolites – and MetaboAnalyst, for statistical analysis.

First reviewed by Ravanbakhsh *et al*. (2015), BAYESIL (https://bayesil.ca) is a free online, fully automated spectral profiling application for ^{1}H NMR-based metabolomics data that can carry out the processing and profiling of 1D ^{1}H NMR spectra automatically without the need for manual intervention. Ravanbakhsh *et al.* (2015) showed that this software is well-suited for serum and cerebrospinal fluid, but did not test it on urine. MetaboAnalyst is an online application (https://www.metaboanalyst.ca), also free to use, and has a range of different functions, mainly for data manipulation, such as statistical analysis, the revelation of biomarkers, as well as the analysis of metabolic pathways (Selegato *et al*., 2019; Gowda & Raftery, 2019). MetaboAnalyst is recommended by Emwas *et al*. (2018) as a valuable software tool for ^{1}H NMR metabolomics, specifically for multivariate analysis, as well as for the biological interpretation of data. Hence, both BAYESIL and MetaboAnalyst were selected as the online tools for our study as they are so well-established.

For this study, comparisons were made between urine samples from patients found to be either positive or negative (healthy) for tuberculosis (TB). Sample groups were purposely selected to compare critically both analytical methods. These samples are considered complex because their metabolite profiles will not only contain metabolites with concentrations differing by orders of magnitude, but also foreign metabolites from medication and/or the TB pathogen. Comparisons of this nature are common in untargeted ^{1}H NMR metabolomics investigations.

Combining the use of BAYESIL and MetaboAnalyst could potentially automate the processing of untargeted ^{1}H NMR metabolomics data, which is greatly needed because of the growing demands for accurate and convenient diagnosis; however, automated systems are still in their infancy. For

this reason, automated processing methods still need to be tested against the established in-house standard operating procedure (SOP) metabolomics methods currently used. Furthermore, unless these procedures are properly trained and experienced, the use of automated data processing tools can be quite complex, even daunting. Hence, user-friendly guidelines are needed for novice users of untargeted ^{1}H NMR metabolomics data. Thus, the aim of this study was the parallel assessment of both our "in-house" SOP and automated data processing methods of complex urinary ^{1}H NMR metabolomics data, with guideposts offered by a novice (J. vdW) for novice users, at identified pitfalls.

4.1.3 Materials and Methods

4.1.3.1 Sample description and ethics requirements

The samples used in this study were anonymized urine samples collected from 46 TB-positive patients before treatment (week 0), after 2 weeks of treatment, and 30 samples from healthy controls who were TB negative. All samples were obtained from a biorepository; ethics approval for the use of these samples was granted by the Ethics Committee of the University of Stellenbosch (No. 99/039). This research also received ethical approval from the Health Research Ethics Committee of North-West University under the following ethics number: NWU-00127-11-A1-02. A pooled quality control (QC) sample (consisting of ~70 µL from each sample) was inserted at the beginning, middle and end of each batch to evaluate the precision of the analysis. Samples were randomized to analytical batches in a representative manner.

4.1.3.2 Instrumentation and materials

The instruments used in this study were the following: an Orto Alresa UNICEN 21 centrifuge; a Dragonlab D3024 high-speed microcentrifuge; a Dragonlab MX-S vortex mixer; a Sentron pH meter with a 9270-010 MicroFET pH probe; and a Bruker Avance™ III HD 500 MHz NMR spectrometer. Materials and consumables used in the laboratory during this study included 5 mm glass NMR tubes, Eppendorf® microcentrifuge tubes, pipettes of various sizes and tips, as well as nitrile gloves. The chemicals used in this study, mainly for the preparation of the NMR buffer, were the following: distilled water, 2-chloropyrimidine-5-carboxylic acid, potassium phosphate monobasic (KH_2PO_4), trimethylsilyl propanoic acid (TSP) and sodium azide, and deuterium oxide (D_2O).

4.1.3.3 Buffer preparation

A 100 mL buffer solution, used to prepare the urine sample for NMR analysis, was collected according to the recipe [see Table S1, available in the online supplementary information (SI)]

provided by BAYESIL on their website (http://bayesil.ca/), with some minor modifications described below. These modifications were required for practical reasons but were inconsequential in the identification and quantification of metabolites.

The original recipe provided by BAYESIL calls for the use of potassium phosphate dibasic (K_2HPO_4). This was substituted for potassium phosphate monobasic (KH_2PO_4), which was already available in the laboratory. The necessary calculations were made to ensure that the correct amount was added to keep the desired concentration, as indicated in Table S1. Another minor change to the recipe included the use of trimethylsilyl propanoic acid instead of sodium trimethylsilyl propane sulfonate (DSS) as the first internal standard. Lastly, the BAYESIL recipe allows for a choice between sodium formate and 2-chloropyrimidine-5-carboxylic acid. In this study, the latter was chosen and used as the second internal standard to aid in the optimization of phase correction carried out by BAYESIL during analysis. The last (modified) component added to this buffer was sodium azide, whose presence prevents bacterial growth. After the addition of all these reagents, the buffer was filled up to 100 mL with sterilized water and mixed thoroughly under vortex into a homogenous solution.

It is important to note that NMR as an analytical tool is pH sensitive. For this reason, it is vital to ensure that the buffer solution has a final pH of exactly 7.00. To achieve this, small amounts of NaOH or HCl solutions were added to adjust the pH. The urine buffer was then left to stabilize overnight and micro-adjustments to the pH, if needed, were made the next day.

4.1.3.4 Sample preparation

All samples were prepared according to the protocol suggested by BAYESIL. Prior to preparation, all samples were stored at –20ºC and then allowed to thaw at room temperature before use. From each sample, 1 mL was transferred to an Eppendorf® microcentrifuge tube for further use. The pH of each sample was checked and adjusted, if needed, so that it fell within the range of 6.8–7.4. After this, the Eppendorf® tubes containing 1 mL of each sample were centrifuged at 12 000 *g* for 5 minutes. In a new microcentrifuge tube, 60 µL of the previously prepared NMR buffer and 70 µL of D_2O were added, along with 570 µL of the urine supernatant. This combination was then mixed under vortex for a few seconds and further centrifuged at 12 000 *g* for another 5 minutes. From this supernatant, 540 µL was taken and transferred to a 5 mm glass NMR tube. These tubes were then randomly loaded into the NMR spectrometer to avoid confounding with the order of analysis, with a pooled QC sample inserted at the beginning, middle and end of each batch.

4.1.3.5 Spectral collection

It is important to note that NMR samples are sensitive to both pH and temperature. The NMR buffer solution that was added to each sample functioned to maintain the pH at ~7.00, while the temperature was controlled by the temperature control unit of the NMR spectrometer, which keeps the probe constant at 300 K (~ 27ºC) during analysis of the samples. The ^{1}H NMR parameters that were used to collect the spectra in this study were identical for both the manual and BAYESIL methods and were as follows: a sweep width of 12 ppm, an acquisition time of 4 seconds, mixing time of 100 milliseconds, 10 milliseconds recycle delay and a saturation delay of 990 milliseconds. The pre-saturation pulse power was calibrated to provide a field strength of 60–80 Hz, to prevent signal saturation near the region of water resonance (~ 4.78 ppm). The transmitter offset and saturation pulse was positioned on the water resonance region and standard phase cycling was used to suppress water resonance and unwanted spectral artefacts. All excitation pulses were calibrated to correspond to a 90º flip angle. Four steady-state scans were collected prior to data acquisition (dummy scans = 4), which were at 128 transients and a receiver gain of 64 (number of scans = 128), yielding a run time of 15 minutes and 45 seconds per sample.

4.1.3.6 Manual (binning) data processing

The data analysis steps that were carried out during this study are summarised below and in the accompanying figures. For the standard manual (binning) method, the spectra collected had to be processed. This was conducted using the software program Bruker Topspin (V3.5) and involved three steps. The first step was chemical shift referencing, where everything was made relative to an internal standard (in this case, TSP). This step is important for the identification of metabolites, alignment of peaks, and multivariate statistical analysis (Emwas *et al*., 2018). The second step – phasing – was used to adjust the NMR spectra to improve peak symmetry. The final spectral processing step – baseline correction – is vital for the removal of spectral artefacts (e.g., electronic distortions, poor filtering, or errors during sampling) that interfere with the NMR spectrum (Emwas *et al*., 2018).

Once the spectra were processed, the data were subjected to further processing, known as data post-processing. During post-processing, non-informative regions (where neither distinct peaks nor metabolite signals were present) were removed from the data, spectra were divided into bins of width 0.02 ppm (yielding approximately 500 bins), and the data were normalized relative to creatinine at 4.05 ppm to account for dilution differences. The binned data matrix was then exported as a text file and converted to a Microsoft® Excel sheet for data clean-up and statistical analysis.

In Microsoft® Excel, the water region was identified and removed (spectral region 4.69 – 4.89 ppm). Noise filtering was also conducted in Microsoft® Excel, by qualitatively identifying the bins that correspond to noise regions in the spectra. The standard deviations of these noise bins were then used to calculate the noise threshold. All values in the binned data matrix that fell below the noise threshold were replaced with a value of 0, by using the IF function in Microsoft® Excel. The binned data matrix was then saved in comma separated value (.csv) format, as a requirement for uploading onto MetaboAnalyst. See Guidepost #1 for a novice tip.

Guidepost #1

Conversion of the data matrix into comma separated value (.csv) format converts the document into a delimited text file where the values are separated by a comma. Check that the binned data matrix has point (.) decimals and not comma (,) decimals when converting the file into a comma separated format (.csv), otherwise the binned data will be converted incorrectly. This can be checked by reviewing the global decimal setting in Excel under Options and converting numbers stored as text to numbers if need be.

After the data were processed through MetaboAnalyst, a list of important bins was obtained based on pre-defined selection criteria, namely: fold change analysis with a threshold of 1; the t-test false discovery rate (FDR)-adjusted p-value had to be less than or equal to 0.05; and the PLSDA VIPs were greater than or equal to 1 for all selected latent variables. To obtain the metabolite identities that correspond to these significant bins, two semi-automated software systems were used, namely, Chenomx profiler (V8.2) and Bruker Amix (V3.9.14). Chenomx profiler identifies possible metabolite annotations for the peaks that fall within the significant bins, based on matches with libraries of pure compounds. Bruker Amix, together with pure compound spectral libraries, was used to confirm these possible annotations, and 2D correlation spectroscopy (COSY) and 2D ^{1}H J-resolved (JRES) NMR spectroscopy was used to confirm these identities.

The next step was to calculate the concentrations of identifiable important metabolites. All metabolite concentrations were made relative (as mmol/mol creatinine) to the methylene moiety of creatinine at 4.05 ppm to account for dilution differences and then multiplied by the absolute concentration of creatinine (mM) per sample to obtain the absolute concentrations of the metabolites (μM) to allow the data to be compared to the BAYESIL results. The quantified list of metabolites was then subjected to another round of statistical analysis via MetaboAnalyst, to determine which metabolites were selected again based on the same criteria indicated above. This list was used for a direct comparison with the BAYESIL list of important variables. Figure 1 summarises the standard manual (binning) method used at the Centre for Human Metabolomics at North-West University (CHM of NWU).

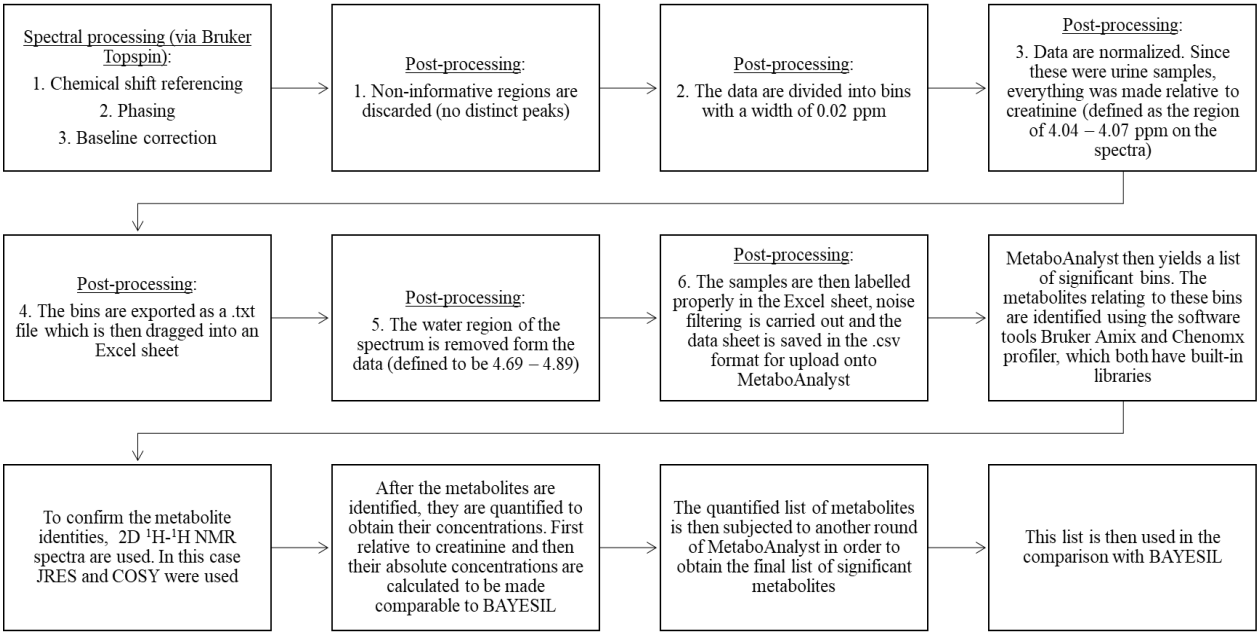


Figure 4-1: Schematic representation of the standard manual (binning) untargeted ^{1}H NMR metabolomics data processing method.

4.1.3.7 Automated (BAYESIL) data processing

For the automated BAYESIL method, the raw NMR data files (FID files) were uploaded directly onto the BAYESIL website (http://bayesil.ca/). The software carries out all spectral processing steps automatically, including zero-filling, Fourier and Hilbert transformation, phasing, and baseline correction. The reader is referred to Ravanbakhsh *et al*. (2015), who describe BAYESIL's spectral processing algorithms and the principles and rationale behind this spectral profiling method in greater detail. Post-processing and zero filtering of the BAYESIL data are also not necessary, since the program is essentially a metabolite profiling tool and the resulting output is a list of metabolite names (albeit with a certain percentage certainty) and their concentrations (µM). Once uploaded onto BAYESIL, the samples are placed in a queue and take approximately 10 minutes per sample once they reach the front of the queue. Figure 2 summarises the automated BAYESIL method and can be used as a step-by-step guide for first-time users. The resultant outputs from the program were separated files per sample, which then had to be manually combined into a single data matrix consisting of metabolite names as columns and sample names given as rows, with each corresponding cell containing the associated concentration value. This metabolite concentration matrix was then uploaded onto MetaboAnalyst, where statistical analysis was conducted to determine which metabolites differ notably among groups.

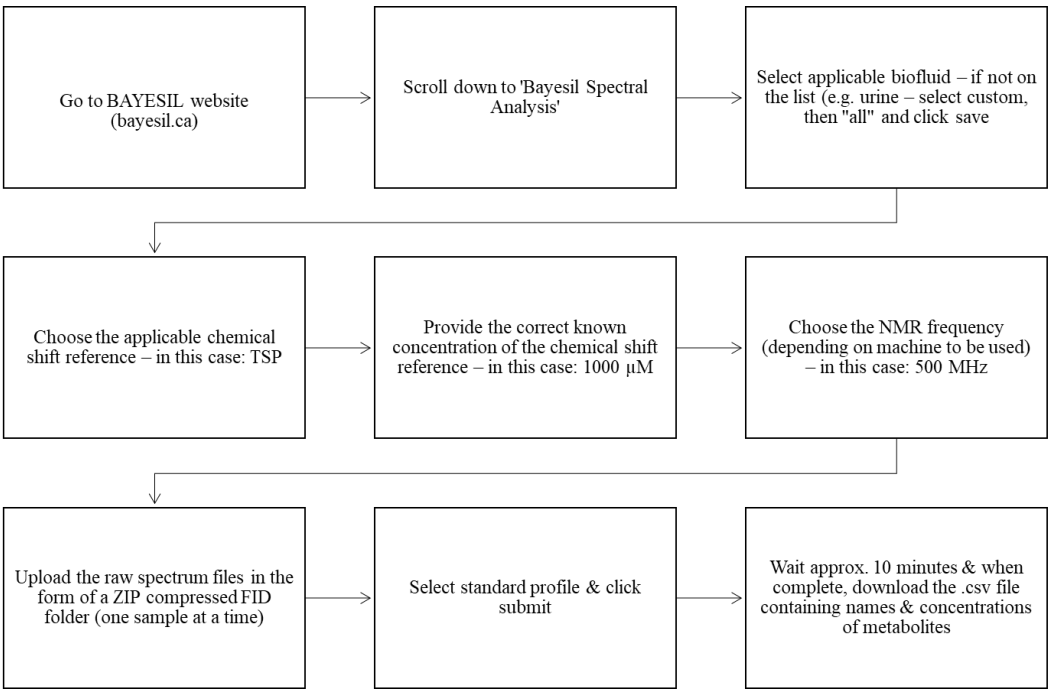


Figure 4-2: Schematic summary of the automated BAYESIL data processing method.

4.1.3.8 MetaboAnalyst – statistical analyses

The online metabolomics suite, MetaboAnalyst (V5.0), was used for statistical analysis of both data sets from the manual and automated methods. Figure 3 is a summary of how MetaboAnalyst was used in this study – it is sufficiently universal to be used by a novice wanting to compare two groups. First, to use MetaboAnalyst, one must navigate to the website (https://www.metaboanalyst.ca/), where the option 'Statistical Analysis [one factor]' is selected. The data file can now be uploaded, followed by the selection of the correct data type (e.g., spectral bins or concentrations).

To upload data onto MetaboAnalyst, specific requirements need to be met. For this study, the data sets that were used were spectral bins and concentration tables. According to the MetaboAnalyst website, these data types must meet the following criteria:

1. Data must be uploaded in comma separated value (.csv) format or tab delimited text (.txt).

2. Sample or feature names must be unique and may contain only common English letters, underscores, and numbers. No Latin/Greek letters.

3. Class labels must immediately follow sample names.

4. If the data are a time series, the time-point group must be labelled as "Time".

5. Data values should contain only numeric and positive values. Missing values should be left empty or indicated with "NA".

6. No spaces between numbers (e.g., 1800 instead of 1 800).

After uploading the data, MetaboAnalyst performs various data integrity checks, which verify that all the requirements are met (See Guidepost #2 for a novice tip).

Guidepost #2

It is important to assess the information output by MetaboAnalyst after the data integrity check, to ensure that it corresponds with the uploaded data. Important things to check include:

- The number of rows
- The number of columns
- The proportion of missing values
- Group sizes.

The next step is data filtering. Here, the option to filter features if their relative standards deviations (RSDs) are more than 30% in the QC samples was selected, along with interquartile range (IQR) filtering. Excluding variables based on QC RSD values is sensible as these variables could not be measured consistently. The IQR filter removes variables that showed too little variation to differ significantly among groups, thereby reducing the number of univariate tests required and consequently the impact of the false discovery rate correction.

Once the data have been filtered, MetaboAnalyst automatically moves to the next step, referred to as normalization. This step includes normalization, transformation, and scaling of the data, with various options for each. Since the data had already been normalized to creatinine, as mentioned before, this part was skipped and only transformation and scaling were of interest.

Guidepost #3 (just to be safe)

At the time of this analysis, a big limitation of MetaboAnalyst was that paired and unpaired data cannot be uploaded in a single document. For this reason, the data were split into data sets, since the HC vs W0 group is unpaired and the W0 vs W2 group is paired.

For univariate statistical analysis, the data were split into two data sets including the healthy control versus week 0 (HC vs W0) group and week 0 versus week 2 (W0 vs W2) group, respectively. The entries in these data sets were only transformed and not scaled. The CHM prefers not to scale data at this point as the means associated with scaled data can be negative, which is unconventional in the clinical setting. At the same time, transformations are required to reduce the skew of distributions, for which metabolomics data are known. In so doing, the validity of the statistical test assuming normality is improved. Reporting untransformed descriptive statistics is also not considered an accurate representation as these are not the values compared in the statistical tests

used. The type of transformation that was used is called square root transformation, which produces a similar pattern to log transformation but holds an advantage due to its ability to work with zero values and has a positive effect on the heteroscedasticity of the data (van den Berg *et al*., 2006). Once transformed, the data can be subjected to univariate statistical tests. The univariate tests that were used in our study included fold change analysis with a threshold of 1, independent/dependent t-test with the assumption of unequal group variance and false discovery rate (FDR) adjusted p values, and a volcano plot with an FDR p-value cut-off of 0.05 and assumed unequal group variance. The univariate statistics selected are considered robust and more powerful than their non-parametric counterparts; a novice is likely to have some experience with these.

Guidepost #4

When conducting statistical tests on MetaboAnalyst, it is important to know exactly which settings to select. Settings such as equal or unequal group variances, whether to use a raw p value or an FDR adjusted p value, parametric or non-parametric tests, the advantages/disadvantages/effect of different transformation and scaling methods, etc. can have a dramatic impact on the statistical result. It is therefore suggested to consult with a biostatistician for guidance on the selection of statistical methods.

Before continuing with multivariate tests, the normalization step of MetaboAnalyst needs to be revisited. Scaling cannot be ignored in the multivariate setting as differences in the orders of magnitude between metabolites will place greater importance on metabolites in higher abundance, which may not reflect biological importance. This time, both square root transformation and auto scaling were selected. According to van den Berg *et al*. (2006), auto scaling allows for the comparison of metabolites based on the relationship between them, with the importance of all metabolites being equivalent. The transformed and scaled data were subsequently subjected to principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA), for each group comparison. For this, the data editor tool of MetaboAnalyst was used (see Guidepost #5).

Guidepost #5

MetaboAnalyst has a built-in feature called “data editor”, which allows the user to select which groups in the data to include for further analysis. This feature may be useful when the data file contains various groups, and the user wishes to compare specific groups, while ignoring others. The data editor feature works by temporarily removing/adding groups as the user desires. The same can be achieved to cases.

Because PCA and PLS-DA are the most well-known multivariate models in the field of metabolomics, even a novice should have encountered them before. PCA models are used to summarize variation in the data and visualize how the largest proportion of variability relates to the groups of interest. PLS-DA does the same but is explicitly informed of the groups present in the data. Any PLS-DA model must be validated as this method is prone to fit the given data well, but not previously unseen data, a characteristic known as overfit. When validating the PLS-DA result, one must determine which latent variables/components are important for separation of the groups. This can be achieved by visual inspection – that is, which components must be viewed to reveal separation between the groups. Once this has been assessed and separation is visible, cross-validation can be carried out. To assess the severity of overfit, the stability of the model – how sensitive the model's performance is to slight alterations in the data – is accessed as well as its statistical significance. Under the cross-validation tab, search for the number of components determined to be important. We suggest assessing stability by setting cross-validation to 'LOOCV' and the performance measure to accuracy. The next important aspect is statistical significance; to accommodate this, go to the permutation tab and select the test statistic as 'prediction accuracy during training'. The maximum number of permutations should be selected (2000). The PLS-DA validation criteria are given in Guidepost #6. From the PLS-DA result, an Excel sheet containing the important features can be downloaded. If the PLS-DA model is considered stable and significant, the VIP score (a ranking assigned to each variable according to its importance in the models) can be used to select important compounds/metabolites.

Guidepost #6

The PLS-DA validation criteria are as follows:

- Accuracy > 0.8 (80%)
- Q2 > 0.6
- R2 > 0.8.

Once statistical analysis has been completed, MetaboAnalyst generates an analysis report containing the results and explanations of all the steps that were followed. This report can be downloaded as a .pdf document.

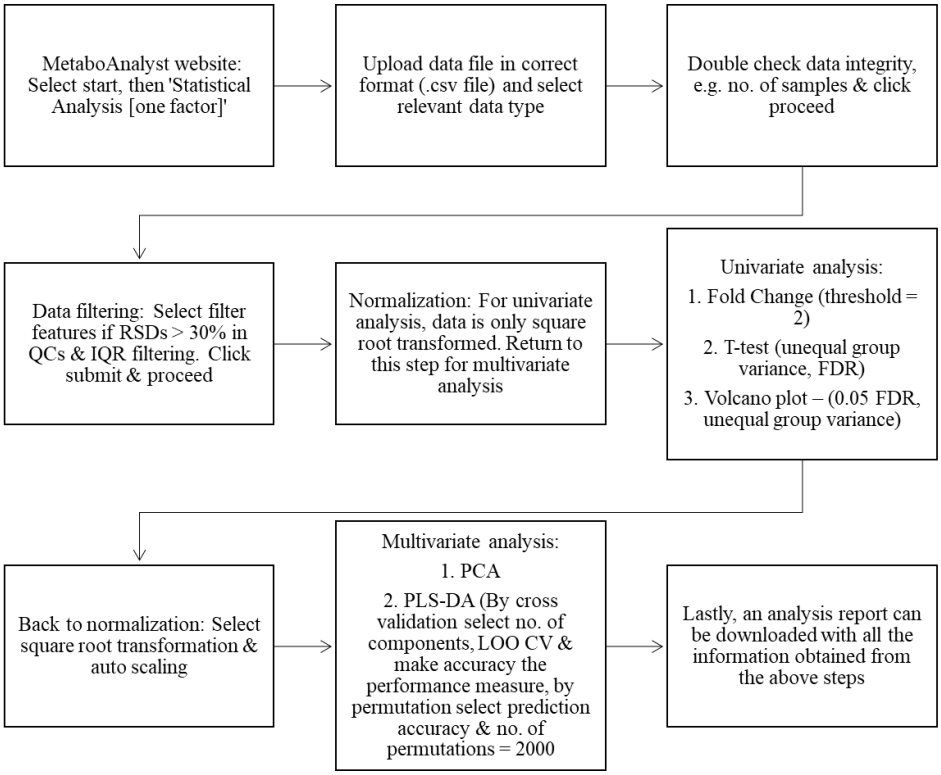


Figure 4-3: A summarised guide on how to use the online metabolomics suite MetaboAnalyst (V5.0) for a basic two-group comparison.

4.1.4 Results

4.1.4.1 Multivariate model performance

Table 1 summarizes the results obtained from analysing the data for the BAYESIL method and the 'quantified manual' method using the MetaboAnalyst platform. The 'quantified manual' method refers to the data that were analysed using MetaboAnalyst after the bins had been identified and quantified, to make the manual result more comparable to the BAYESIL result. This table includes information about the PCA, as well as PLS-DA plots, with a grading scale indicating how well the groups were separated in the different comparisons (scores plots are available in the SI).

Table 4-1: A summary of the multivariate MetaboAnalyst statistical results for both methods and comparisons, showing the performance of the multivariate models (PCA and PLS-DA).

		PCA	PLS-DA				
Method	**Group**	**Separation**	**Separation**	**Stability**		**Performance**	
				R^2	**Q^2**	**p-value**	**Accuracy**
BAYESIL	HC vs W0	1	2	0.75	0.67	< 0.0005	92%
Quantified manual	HC vs W0	2	2	0.77	0.72	< 0.0005	97%
BAYESIL	W0 vs W2	1	2	0.67	0.54	< 0.0005	86%
Quantified manual	W0 vs W2	2	2	0.83	0.80	< 0.0005	95%

Grading scale for PCA and PLS-DA separation: 1 = significant overlap; 2 = some separation.

None of the groups showed complete separation for the PCA analysis. The only group to reveal complete separation was the PLS-DA analysis of the binned manual HC vs W0 group (see Table S2). The PLS-DA results for the BAYESIL groups did not perform as well as the 'quantified manual' groups. The BAYESIL HC vs W0 group had an R^2 of 0.75, which is below the validation criterion of 0.8 (given in Guidepost #6). The BAYESIL W0 vs W2 group had the lowest accuracy value of 86% as well as an R^2 value of 0.67 and a Q^2 value of 0.54, both below the validation criterion, indicating a result with insufficient stability. The 'quantified manual' HC vs W0 group had an R^2 value of 0.77 and a Q^2 value of 0.72 but validates on all other criteria with the highest overall accuracy value of 97%. All statistical diagrams (PCA, PLS-DA, volcano plots, t-tests, and fold

changes) for the ‘quantified manual’ and BAYESIL methods are given in the supplementary information.

4.1.4.2 Healthy controls versus week 0 (HC vs W0) group comparison

Table 2 shows the metabolites that were identified as important for the healthy controls versus week 0 group, with the corresponding predetermined statistical rules, and provides a confidence score (0–10, where 0 is the lowest and 10 is the highest confidence) per metabolite for each method, indicating the confidence in identification. The manual method has a confidence score of 10 for each metabolite, since their identities were confirmed by two-dimensional ^{1}H NMR methods. The statistical rules that were used to determine if metabolites are significant, as indicated in Tables 2 and 3, are the following:

- The absolute value of the logarithm of the fold change must be greater than or equal to 1.
- The FDR-adjusted t-test p-value must be no greater than 0.05.
- The PLS-DA VIP score for latent variables showing group separation must be equal to at least 1.

It is important to note that for the HC vs W0 comparison, the BAYESIL method revealed a total of 90 metabolite identifications before statistical analysis, whereas the ‘quantified manual’ method had 22; the lists of metabolites that were not identified as significant are available in Table S3 of the supplementary information.

Table 4-2: **The list of metabolites that were identified as important for differentiating between the HC vs W0 group, along with the statistical rules that were used to determine the significance of metabolites and a confidence score from 0–10 indicating the confidence in identification of these metabolites.**

			Rule for significance					
	Confidence in ID		|logFC| ≥ 1		t-test FDR p-value ≤ 0.05		PLSDA VIP ≥ 1 for LV 1 or 2	
Metabolite	**Auto***	**Q. Man**	**Auto***	**Q. Man**	**Auto***	**Q. Man**	**Auto***	**Q. Man**
2-Aminoadipic acid	N/A	10		×		✓		✓
3-Hydroxyisovaleric acid	6	10	✓	×	✓	✓	×	✓
Acetoacetate	8	N/A	✓		✓		✓	
Anhydro glucose	N/A	10		×		×		✓
Betaine	10	N/A	×		×		✓	
Formic acid	9	10	×	×	✓	✓	✓	✓
Glucose	10	10	×	×	×	×	×	✓
Histidine	7	N/A	✓		✓		✓	
Isopropanol	10	N/A	✓		✓		✓	
Lactic acid	10	N/A	×		×		✓	
Leucine	10	N/A	✓		✓		✓	
Methanol	10	10	✓	✓	✓	✓	×	✓
Methylmalonic acid	–	N/A	×		×		✓	
N-AcetylX	N/A	10		×		×		✓
Phenylalanine	10	10	×	×	×	✓	✓	✓
Proline	10	N/A	×		×		✓	
Pyroglutamic acid	10	N/A	✓		✓		✓	
Succinic acid	8	N/A	×		×		✓	

*Auto = Automated BAYESIL method. Q. Man = quantified manual method. "– "= no confidence score provided by BAYESIL (assumed to be 0). N/A = metabolites that were not identified by that method and therefore a confidence score could not be given. N-AcetylX is the N-acetyl moiety (CH_3) of an unidentified metabolite. The key for understanding this table is as follows: green indicates that the metabolite was identified as important according to the statistical rule; red indicates that the metabolite was identified by the method but was not significant according to the rule; and orange indicates that the metabolite was not identified by the method.

Table 2 shows that BAYESIL had a relatively low confidence score for the identification of 3-hydroxyisovaleric acid and histidine. Moreover, BAYESIL did not provide a confidence score for methylmalonic acid, which was therefore assumed to be 0. The statistical rule that identified the most metabolites as significant is the rule for the PLS-DA analysis – 12 metabolites for BAYESIL and 8 metabolites for the 'quantified manual' method. The univariate statistical rule of |logFC| > 1 identified 7 significant metabolites for BAYESIL and 1 significant metabolite for the 'quantified manual' method. The univariate statistical rule of t-test FDR p-value ≤ 0.05 identified 8 significant metabolites for BAYESIL and 5 significant metabolites for the 'quantified manual' method. Thus, when assessing a diseased state (W0) vs a control group (HC), the BAYESIL method identified more metabolites; however, only 16% were found to be important. For the 'quantified manual' method, 36% of identified metabolites were found to be important.

4.1.4.3 Week 0 versus week 2 (W0 vs W2) group comparison

Table 3 provides the list of metabolites that were identified as important in differentiating between the W0 vs W2 group, with a confidence score per metabolite. All the annotations, the associated key, and the statistical rules in Table 3 are identical to Table 2. Before statistical analysis, the BAYESIL method listed a total of 90 metabolite identities, whereas the 'quantified manual' method revealed 28. Table S4 in the supplementary information provides the list of metabolites that were not found to be significant.

Table 4-3: The W0 vs W2 group comparison, the statistical rules that were applied, and which metabolites were identified as important as a result of these rules, as well as a confidence score (0–10) to indicate confidence in identification for each metabolite. Metabolites boxed in red represent TB medications

			Rule for significance					
	Confidence in ID		|logFC| ≥ 1		t-test FDR p-value ≤ 0.05		PLSDA VIP ≥ 1 for LV 1 or 2	
Metabolite	**Auto***	**Q. Man**	**Auto***	**Q. Man**	**Auto***	**Q. Man**	**Auto***	**Q. Man**
1-Methylnicotinamide	N/A	10		✓		✓		✓
2-Aminoadipic acid	N/A	10		×		✓		×
2-Hydroxybutyrate	9	N/A	×		✓		✓	

2-Hydroxyisovaleric acid	10	N/A	×		×		✓			
2-Pyridinylformamidoacetic acid	N/A	10		✓		✓		✓		
3-Hydroxybutyric acid	10	10	×	×	×	✓	×	×		
3-Hydroxyphenylacetic acid	N/A	10		×		✓		×		
5-Aminopentanoic acid	-	N/A	✓		✓		×			
Acetoacetate	8	N/A	×		✓		✓			
Acetyl isoniazid	N/A	10		✓		✓		✓		
Alanine	10	N/A	×		✓		✓			
Butyric acid	-	N/A	✓		✓		✓			
Citric acid	10	N/A	×		✓		✓			
Creatinine	10	N/A	×		✓		✓			
Dimethylamine	10	N/A	×		×		✓			
Ethambutol	N/A	10		✓		✓		✓		
Formic acid	9	10	×	×	✓	✓	✓	×		
Glutamine	10	N/A	×		✓		×			
Glycine	10	N/A	×		✓		✓			
Hypoxanthine	10	10	×	×	×	×	✓	×		
Isobutyric acid	10	N/A	✓		✓		×			
Isonicotinic acid	N/A	10		✓		✓		✓		
Lactic acid	10	10	×	×	×	✓	×	×		
Leucine	10	10	✓	✓	✓	✓	✓	✓		
Methylamine	–	10	×	×	×	×	✓	×		
Ornithine	6	N/A	×		×		✓			
Phenylalanine	10	10	×	×	×	✓	×	×		
Pyrazine carboxylic acid	N/A	10		✓		✓		✓		
Pyroglutamic acid	10	N/A	✓		✓		✓			
Sarcosine	–	N/A	✓		✓		✓			
Succinic acid	8	10	×	×	×	✓	×	×		
Taurine	–	10	✓	×	✓	✓	×	×		
Threonine	9	N/A	×		✓		×			
Valine	10	10	×	×	✓	✓	✓	×		

*Auto = Automated BAYESIL method. Q. Man = Quantified manual method. "–" = no confidence score provided by BAYESIL (assumed to be 0). N/A = metabolites that were not identified by that method and therefore a confidence score could not be given.

Table 3 shows a longer list of significant metabolites than the result for HC vs W0 (Table 2). The metabolites boxed in red were identified as TB treatment drugs and/or their derivatives/metabolites (based upon pure compound ^{1}H NMR spectra). They were not identified by BAYESIL in any of the samples (indicated by N/A in the confidence score column) as the BAYESIL spectral library does not contain drugs/medications. For the univariate statistical rules of |logFC| >1 and t-test FDR p-value ≤ 0.05, both the BAYESIL method and 'quantified manual' method identified almost the same number of metabolites each – 7 metabolites each for |logFC| >1, and 17 (BAYESIL) and 16 ('quantified manual') metabolites for t-test FDR p-value ≤ 0.05. The statistical rule for the PLS-DA analysis identified 17 metabolites for BAYESIL and 7 metabolites for the 'quantified manual' method. Thus, when assessing a treated state (W2) vs an untreated diseased state (W0), the

BAYESIL method identified more metabolites. For this group comparison (W0 vs W2), 24% of all BAYESIL identities were important in differentiating between the groups, compared to 57% from the 'quantified manual' method.

Superficially, the BAYESIL method identified more metabolites than the 'quantified manual' method for both comparisons in our study. However, it is important to look at the details of the results and highlight some important points – as discussed below.

During final checks of all data, it was noted that the number of cases in the quantified data set did not match the number of cases in the BAYESIL data set. The reason for this was that two of the samples were identified as being of poor NMR spectral quality (due to baseline drift and peak broadening) based upon manual qualitative inspection, and could not be quantified accurately – and so were excluded from further analysis. This highlights an important point of consideration – the quality of the raw ^{1}H NMR spectral data that are input into BAYESIL is not checked by BAYESIL. Thus, some form of quality check – for example, a manual qualitative check or a computer-coded algorithm – is needed to assess the spectral quality prior to inputting into BAYESIL.

4.1.5 Discussion

The main issue that was addressed in this study is that in-house processing of ^{1}H NMR metabolomics data can be very time-consuming and vulnerable to human fatigue and error. Since automated software tools have been developed to combat this problem (such as BAYESIL and MetaboAnalyst), our focus was to test the robustness of these tools and determine if automation of the processing of untargeted ^{1}H NMR metabolomics data has the capability of providing a true and reliable metabolic profile of treated and untreated TB patient urine samples, and in so doing, identify the pitfalls of these automated tools.

To achieve this, two methods of ^{1}H NMR metabolomics data processing were compared and are discussed below. These methods were the in-house manual method of the Centre for Human Metabolomics at our university, and an automated shotgun method using open-source BAYESIL. Statistical analysis for both methods was carried out using MetaboAnalyst and two general comparisons were assessed: healthy controls (HC) versus week 0 (W0), and week 0 (W0) versus week 2 (W2).

4.1.5.1 MetaboAnalyst results

From the MetaboAnalyst results in Table 1, it is clear that all of the groups obtained better (or the same) separation when subjected to PLS-DA analysis when compared to the PCA models, as indicated by the grading scale. This is to be expected as PLS-DA is a supervised statistical test, which means that it takes class labels into account to aid in group separation (Ruiz-Perez *et al*.,

2020). Comparing each method in Table 1, group by group, shows that both methods performed similarly with the HC vs W0 group. Furthermore, only the 'quantified manual' method showed better stability, with the more complex sample group (W0 vs W2) where metabolites of both disease and medication were present, with an R^2 value of 0.83, a Q^2 value of 0.80, and an accuracy of 95%, meeting all the validation criteria given in Guidepost #6. The lower values of BAYESIL for this group (R^2 = 0.67, Q^2 = 0.54, accuracy = 86%), meeting one of 3 validation criteria, indicate that BAYESIL had some difficulty with the complex sample matrix. The most likely reason is that BAYESIL does not have an algorithm designed specifically for urine, causing some issues with the coverage of the spectrum by the fit produced by BAYESIL software. Overall, the highest accuracy for both groups was achieved by the 'quantified manual' method, with 97% for the HC vs W0 group and 95% for the W0 vs W2 group.

4.1.5.2 HC vs W0 group comparison

Table 2 includes a list of metabolites that were identified as important when differentiating between the HC and W0 groups. Some were identified by only one method, others were identified by both methods but considered significant in only one method, and others were identified as significant in both methods. The confidence score provided in columns 2 and 3 of Table 2 is a score provided by BAYESIL to indicate how confident the software is in the identification of a specific metabolite, where 0 is the lowest and 10 is the highest confidence. The lowest confidence scores generated by BAYESIL were for 3-hydroxyisovaleric acid and histidine, with scores of 6 and 7, respectively. Methylmalonic acid, which was not identified in the 'quantified manual' method at all, had no confidence score provided by BAYESIL, which was assumed to imply no confidence in its identification. For the purposes of comparison, all 'quantified manual' method identities were given a confidence score of 10, since extensive measures were taken to ensure the accuracy of these identities, including the use of two spectral libraries of pure compounds and the use of 2D ^{1}H NMR methods (JRES and COSY). Since applications of these methods include biomarker identification for disease diagnosis, it is important to have complete confidence in the identification of metabolites. Disease diagnosis, specifically of a serious disease such as TB, leaves little room for uncertainty in metabolite identification. The 'quantified manual' method clearly holds an important advantage here.

An interesting observation was the identification of isopropanol by the BAYESIL method, which was not identified by the 'quantified manual' method. Isopropanol is found in rubbing alcohol and cleaning products and its ingestion is usually associated with toxic effects, as described in a review by Slaughter *et al.* (2014). For this reason, the presence of isopropanol in these samples is considered highly unlikely and its identification by BAYESIL could possibly be a mistake.

The metabolite labelled ‘N-AcetylX’ refers to the N-acetyl moiety of a metabolite that could not be identified as it was not in the available libraries used in the ‘quantified manual’ method; however, it could be identified as an acetylated compound since it presented with a singlet in the region of the ^{1}H NMR spectrum known to be dominated by N-acetyl moieties (1.95–2.15 ppm). This shows an advantage over the manual method – the ability to use human intuition to predict where certain metabolites will lie in the spectrum. Software systems such as BAYESIL use fixed algorithms and so lack the intuition of human analysts.

4.1.5.3 W0 vs W2 group comparison

This group comparison is considered the most complex due to the presence of disease metabolites from TB infection, as well as numerous exogenous medication metabolites derived from the TB treatment. The TB medications that were identified as significant are indicated by a red outline in Table 3. The only technique that identified these metabolites was the ‘quantified manual’ method. The BAYESIL method was unable to identify any medication metabolites, which is explained by the limited library of the software – an absence of specialised metabolites such as those found in TB disease or treatment. This is a limitation of the automated BAYESIL method, as highlighted in a study by Maulidiani *et al*. (2017).

This also brings to attention an important advantage of the ‘quantified manual’ method, which is the ability and opportunity to expand the available spectral library to aid in the identification of specialised metabolites, such as TB disease and treatment metabolites. In this study, a TB medication library was compiled after analysing the pure compounds of these medications and their metabolites/derivatives on the NMR spectrometer. This made the identification of these metabolites possible. The identification of these medications as important by statistical analysis is expected, since these compounds are the separating factor between the week 0 and week 2 groups, since week 0 samples are untreated TB samples. Thus, a suggestion from us to the creators of the BAYESIL software is to open their databases to allow the public to upload their own pure compound ^{1}H NMR spectral data – pending approval by BAYESIL, of course. A database by the public, for the public.

4.1.5.4 General comparison of methods

Although BAYESIL was able to identify a greater number of metabolites (90 overall) than the manual method, most of them were not identified as significant (16% significant for HC vs W0 and 24% significant for W0 vs W2), whereas the alternative method identified fewer compounds overall (22 and 28 metabolites for the HC vs W0 and W0 vs W2 comparisons, respectively), but more of them were significant – 36% and 57%, respectively. (See Tables S2 and S3 in the supplementary information for the lists of metabolites that were not found to be significant.) Although the

BAYESIL method produced more information, which was expected from a high-throughput technique, most of this information was not relevant to this study. Moreover, the identification of fewer metabolites by the 'quantified manual' method was to be expected as it performs pre-selection of metabolites based on the same selection criteria. One may then simply relax the initial selection criteria to expand the list of bins and so the list of metabolites evaluated by the manual method.

These findings point to a significant strength of the 'quantified manual' method, which is its ability to produce more accurate and more relevant results and identify metabolites no matter the complexity of the samples, using the advantage of user expertise and input/intuition. The fact that BAYESIL does not have an algorithm for urine is a disadvantage of this platform. Since there is currently no alternative, well-established automated method for untargeted ^{1}H NMR urine metabolomics data processing, BAYESIL remains the best option for these studies. This identifies a relatively large gap in automated metabolomics data processing. Since urine is one of the most used biofluids in metabolomics research – it confers many advantages, including easy non-invasive sample collection, little to no pre-treatment requirements and lower protein content (Zhang *et al.,* 2012) – one would expect better automated options for this biofluid. Numerous metabolomics research groups are undoubtedly attempting to address this matter. Recently, one such research group from Denmark (Khakimov *et al*., 2020) has presented a new peak-picking algorithm called signature mapping (SigMa), designed specifically for urine samples. A beta version of SigMa became available at the end of 2020, but rigorous testing is needed before it can be considered as well-established as BAYESIL; hence, our current choice is to focus on BAYESIL instead of SigMa.

During data analysis, some discrepancies were noticed between the concentrations of metabolites obtained by the BAYESIL method and by the 'quantified manual' method. Other problems that were experienced with the BAYESIL method include the inability to identify some metabolites due to its limited library, issues with the fit produced by BAYESIL and there being no algorithm specifically designed for use on urine. These limitations to the BAYESIL system are similar to those reported by Maulidiani and colleagues (2017), which they attributed to poor baseline correction by BAYESIL, as well as issues with peak shifting. The study by Maulidiani *et al*. (2017) found that despite these issues observed when using BAYESIL, it is still suitable for the automation of metabolite quantification of ^{1}H NMR data and holds the advantage over the BATMAN system (an alternative semi-automated NMR data processing platform) of being more user friendly and not requiring coding experience to use it.

In contrast, the 'quantified manual' method was able to identify and quantify metabolites more accurately but requires significant intervention by a skilled NMR analyst. This method is also greatly time consuming and can take up to a few days to provide a complete metabolite profile. It

is therefore clear that both methods have advantages and disadvantages, highlighting the importance of finding a balance between high-throughput processing and user expertise. Another pitfall that was identified in this study is that there is a lack of standardisation in data processing and statistical analysis, which leads to many inconsistencies in the results obtained. For this study, the available pipeline was used for data processing, which determined the statistical tests and rules to be used. This may, however, differ from what another analyst would or can do, which may produce a different result. Furthermore, a limitation of the MetaboAnalyst platform itself is that it has very strict requirements for the format of the data to be uploaded, which may make it challenging for a novice user who has limited experience in data handling, statistics, and bioinformatics.

This paper provides a functional guide on how to process untargeted ^{1}H NMR metabolomics data for novice users in the future and was compiled by a fellow novice. Throughout the research that led to this paper, a few areas were identified where mistakes were easily made by a novice. All the advice and tips that were picked up along the way are included in this paper as guideposts, which we hope will provide useful guidance to those who lack confidence during the many hours of data processing.

4.1.6 Conclusions

When comparing the automated BAYESIL and the 'quantified manual' methods, we found that the latter method generates a metabolite profile with higher specificity and greater confidence. However, the biggest disadvantage of the manual method is that it is time-consuming, labour-intensive and requires an analyst with experience of ^{1}H NMR metabolomics data. In contrast, the BAYESIL method is high-throughput and quick, is easier to use and requires no trained expertise. It does, however, perform with less specificity and identifies metabolites with a lower confidence than the other method. For these reasons, we conclude that the automated BAYESIL method is currently better suited for shotgun metabolic profiling to obtain a bird's-eye view of metabolism, whereas the in-house 'quantified manual' method, with its higher specificity and confidence, is well-suited for more detailed research studies. It may also be useful to combine both methods, to allow the BAYESIL system to supplement the 'quantified manual' method, as it may identify metabolites missed by a human expert. Currently, there is a large gap in the automation of ^{1}H NMR metabolomics data processing, since there is no established online tool specifically designed for the widely used and information-rich biofluid – urine. Clearly, the automation of ^{1}H NMR metabolomics data processing has potential but is still lacking in some areas and retains a measure of error associated with it. Moreover, we hope that the guideposts that we have provided here will help future novice users of ^{1}H NMR metabolomics data to understand better and navigate the data processing needed to achieve the maximum out of their data.

4.1.7 Declarations

Author Conflict of Interest Statement

Funding: This work is based on the research supported wholly/in part by the National Research Foundation of South Africa (Grant number 130913).

Conflict of Interest: All authors state that they have no conflict of interest to declare.

Author Contribution Statement

SM conceived and designed the study. JvdW conducted all laboratory experiments and analysis. JvdW processed the data, with guidance from MvR as the bioinformatician. SM and MvR provided guidance as supervisors. JvdW wrote the manuscript. All authors read and approved the manuscript.

Compliance with Ethical Standards

All procedures performed in this study involving human samples were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Ethical clearance was granted by the Health Research Ethics Committee (HREC) of the NWU under the following number: NWU-00127-11-A1-02.

Data Availability Statement

The complete ^{1}H NMR data set from this metabolomics project will be made publicly available later, once this project has been completed (all the data published). Immediate sub-sets of data from the study reported here can be made available upon special request: nmr.nwu@gmail.com.

Software or Database Availability Statement

BAYESIL is available online (www.bayesil.ca).

MetaboAnalyst is available online (www.metaboanalyst.ca).

4.1.8 References

Emwas, A.H., Saccenti, E., Gao, X., McKay, R.T., Martins dos Santos, V.A.P., Roy, R. & Wishart, D.S. 2018. Recommended strategies for spectral processing and post-processing of 1D 1H-NMR data of biofluids with a particular focus on urine. *Metabolomics*, 14(3):31. https://doi.org/10.1007/s11306-018-1321-4

Gowda, G.A.N and Raftery, D. 2019. Overview of NMR spectroscopy-based metabolomics: opportunities and challenges. In: Gowda, G.A.N. & Raftery, D., eds. *NMR-Based Metabolomics*. New York: Springer. pp 3-14.

Khakimov, B., Mobaraki, N., Trimigno, A., Aru, V. & Engelsen, S.B. 2020. Signature mapping (SigMa): An efficient approach for processing complex human urine 1H NMR metabolomics data. *Analytica Chimica Acta*, 1108: 142-151. DOI: 10.1016/j.aca.2020.02.025

Maulidiani, Rudiyanto, Mediani, A., Khatib, A., Ismail, A., Hamid, M., Lajis, N.H., Shaari, K. & Abas, F. 2017. Application of BATMAN and BAYESIL for quantitative ^{1}H-NMR based metabolomics of urine: discriminant analysis of lean, obese, and obese-diabetic rats. *Metabolomics*, 13(11). DOI: 10.1007/s11306-017-1273-0

Ravanbakhsh, S., Liu, P., Bjordahl, T.C., Mandal, R., Grant, J.R., Wilson, M., Eisner, R., Sinelnikov, I., Hu, X., Luchinat, C., Greiner, R. & Wishart, D.S. 2015. Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS ONE*, 10(5), e0124219. DOI: 10.1371/journal.pone.0124219

Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K, & Narasimhan, G. 2020. So you think you can PLS-DA? *BMC Bioinformatics*, 21(Suppl 1):2. https://doi.org/10.1186/s12859-019-3310-7.

Selegato, D.M., Pilon, A.C. and Neto, F.C. 2019. Plant metabolomics using NMR spectroscopy. In: Gowda, G.A.N. & Raftery, D., eds. NMR-Based Metabolomics. New York: Springer. pp 345-362.

Slaughter, R.J., Mason, R.W., Beasley, D.M.G., Vale, J.A. & Schep, L.J. 2014. Isopropanol poisoning. *Clinical Toxicology*, 52 (5): 470-478. DOI: 10.3109/15563650.2014.914527

van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K. & van der Werf, M.J. 2006. Centering, scaling and transformations: improving the biological information content of metabolomics data. BMC Genomics, 7:142. DOI: 10.1186/1471-2164-7-142.

Zhang, A., Sun, H., Wu, X. & Wang, X. 2012. Urine metabolomics. *Clinica Chimica Acta*, 414 (2012):65-69. http://dx.doi.org/10.1016/j.cca.2012.08.016.

Zheng, C., Zhang, S., Ragg, S., Raftery, D. & Vitek, O. 2011. Identification and quantification of metabolites in 1H-NMR spectra by Bayesian model selection. *Bioinformatics*, 27(12):1637-1644. doi:10.1093/bioinformatics/btr118

CHAPTER 5 DISCUSSION

A discussion of the results of this study is given in the manuscript in Chapter 4. In this chapter, the objectives of this study (listed in Chapter 1.3) will be discussed, and focus will be placed on how each objective was met and how the overall aim was achieved. Insight(s) will also be offered where applicable.

5.1 Objective 1: The manual method

5.1.1 Objective 1.1

Objective 1.1 involved manually preparing a spectral binned data matrix of the raw ^{1}H-NMR output.

This was successfully completed by combining all the raw ^{1}H-NMR outputs from all samples analysed by the ^{1}H-NMR spectrometer, after the necessary spectral processing steps were completed. These spectral processing steps were carried out using the software program Bruker Topspin (V3.5) and involved calibrating the spectra to TSP at 0.00 ppm, phasing to adjust the symmetry of the peaks and baseline correction. All regions of the spectra that were considered uninformative after qualitative assessment of the spectra (such as the beginning and end parts of the spectra where no discernible peaks were seen) were removed from the data matrix. The remaining regions were divided into bins with a spectral width of 0.02 ppm.

Since spectra obtained from ^{1}H-NMR analyses contain thousands of peaks, the corresponding data will have just as many variables. This results in an overwhelming amount of data that is difficult to process into useful information. Binning is therefore applied to divide these vast spectra of thousands of peaks into approximately 500 pieces/segments of equal-sized (equidistant) spectral regions called bins. These bins represent the spectra, and each bin may correspond to one or more spectral peak, or none if it is a noise region. Binning reduces the dimensionality of the data to make it easier to work with and reduces the effect of peak shifts caused by small variations in pH, temperature, and other exogenous effects (Cobas, 2011). The main drawback of binning, however, is that some information is lost, since peaks corresponding to more than one metabolite may be included in the same bin (Cobas, 2011).

After binning, the data were normalized to the methylene moiety of creatinine found as a singlet at the chemical shift of 4.05 ppm. This yielded the spectral binned data matrix used for further analysis in the manual method.

5.1.2 Objective 1.2

Objective 1.2, as listed in Chapter 1.3 of this dissertation, is as follows: “Subject this data matrix to statistical analysis via MetaboAnalyst”.

Before this objective could be met, the data matrix had to undergo some data clean-up and meet a few requirements. The data clean-up steps involved the removal of the water region in the spectra and noise filtering. The water region of the spectra was identified to be between 4.69 – 4.89 ppm, upon qualitative inspection. This region is easy to identify, due to the water suppression program used during spectral collection pushing the water signal below the baseline. It is important to suppress this water signal during analysis because without suppression, the high abundance of water in biofluids, such as urine, will over-shadow the signals from all the metabolites present in less abundance and produce smaller signals.

All analytical instruments, including the NMR spectrometer, have ‘noise’ regions in their outputs. This ‘noise’ is usually the result of electrical interference and/or other environmental/technical aspects and is of no practical relevance. For this reason, this ‘noise’ is filtered out of the spectra to ensure that these regions are not incorrectly identified as statistically significant. Noise filtering involved the calculation of a noise threshold, and the replacement of all values that fell below this threshold with zero.

Unless instructed otherwise, MetaboAnalyst automatically treats missing and zero values by first removing variables with more than 50% zero or missing values, then replacing the remaining by one fifth the minimum, and removing variables showing too little variation across experimental groups based on the inter quantile range.

The data matrix could then be uploaded onto MetaboAnalyst in comma separated value (.csv) format. For a document to be read by the MetaboAnalyst software, it must be contained in a delimited text file where a comma separates the values. It is therefore important to check that the binned data matrix has point (.) decimals and not comma (,) decimals, otherwise the binned data will be incorrectly converted. This is the first important guidepost that was noted.

Another important tip is to assess the information output by MetaboAnalyst after the data integrity check, to make sure that the uploaded data was read correctly. Important things to double check include: the number of rows and columns, missing values, and group sizes. All requirements provided by MetaboAnalyst for the data were met, allowing us to continue with the statistical analysis successfully. Therefore, objective 1.2 was successfully completed.

5.1.3 Objective 1.3

Objective 1.3 was the following: “Identify ^{1}H-NMR spectral regions (bins) that differentiate between experimental groups within this data matrix and assign metabolite names to them”.

Once the statistical analysis was complete, a list of discriminating/important bins was obtained. This list was produced by combining the results obtained from the volcano plots, t-tests and PLSDA tests. These statistical tests, as well as their selection criteria, were carefully selected after consultation with a biostatistician since there are several different settings and aspects that could influence the statistical result.

For univariate analysis, the data were only transformed and not scaled. Scaling is excluded at this point to avoid negative means commonly associated with scaled data since this is unconventional in a clinical setting. Transformation, on the other hand, is essential to reduce the skew of data that metabolomics is well-known for. The univariate tests and their selection criteria used in our study included fold change analysis with a threshold of 1, an independent/dependent t-test with the assumption of unequal group variance and false discovery rate (FDR) adjusted p values, and a volcano plot with an FDR p-value cut-off of 0.05 and assumed unequal group variance. As explained in Chapter 3.6, unequal group variance is assumed to avoid the assumption of homoscedasticity by using a variation on the t-test that accommodates heteroscedasticity. The validity of findings is improved by controlling the rate at which false discoveries are made, which is made possible by the option to select an ‘FDR-adjusted p value’ on MetaboAnalyst. This is also explained in more detail in Chapter 3.6.

For multivariate tests, the data are transformed as well as scaled. Scaling in the multivariate setting ensures that greater importance is not placed on metabolites present in higher abundance. Multivariate analysis included PCA and PLS-DA models. The purpose of these models is to summarise the variation in the data and visualize how variability relates to the groups of interest. The difference between these two models is that PLS-DA is a supervised model, meaning that it knows which groups are present in the data. This tends to overly inflate the power of a PLS-DA model (overfit). To assess the severity of the overfit, PLS-DA models must always be validated. As discussed in Chapter 3.6, the most basic and lenient validation procedure is called LOOCV and involves leaving a sample out repeatedly and rebuilding the model every time. If the model performance is unstable, it is unlikely that the model will accurately predict the group membership of a new dataset. Since LOOCV is not very robust, it is complemented by the performance metric, prediction accuracy. The proportion of variability explained by the group structure (the alternative performance metric) is not considered appropriate since only a small number of metabolites may be affected.

Since this was a metabolomics study and the focus was on metabolites, these bins had to be converted into a more useful form of information i.e., a list of metabolite names. These metabolites' names were obtained by using a combination of two semi-automated software systems, with built-in pure compound spectral libraries: Chenomx profiler (V8.2) and Bruker Amix (V3.9.14). Metabolite identities were confirmed with 2D correlation spectroscopy (COSY) and 2D ^{1}H-^{1}H J-resolved (JRES) NMR spectroscopy. Although these software systems were the main source of information for the annotation of metabolite identities, knowledge obtained from previous NMR training was also used to predict the chemical shift and positioning of some metabolites based on their functional groups, see Fig 5-1 below. Objective 1.3 was therefore successfully completed.

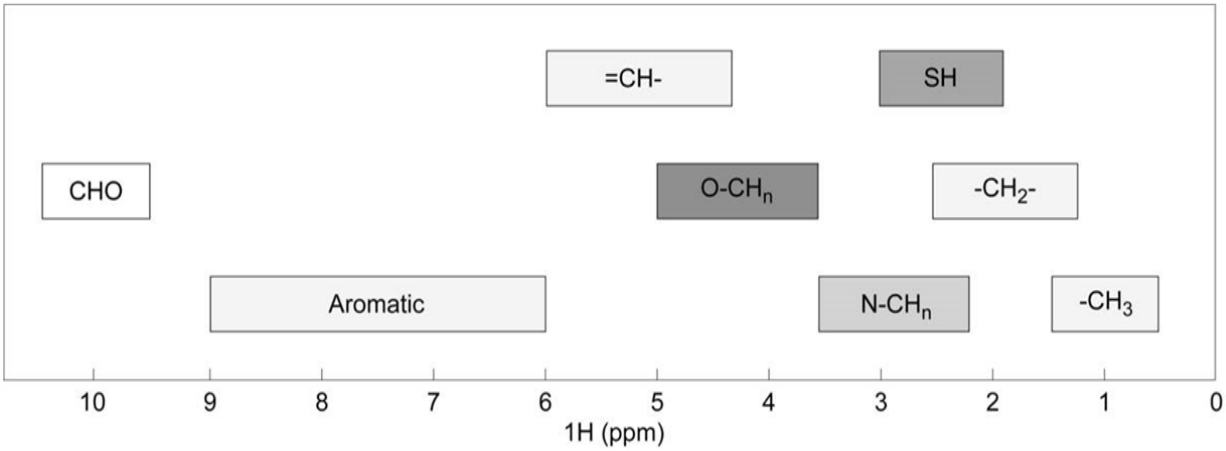


Figure 5-1: Regions of a ^{1}H-NMR spectrum where different functional groups are known to be present, allowing the prediction of where certain metabolites will lie. A ^{1}H-NMR spectrometer detects protons that are attached to a carbon atom (-CH$_n$ groups). The chemical environment around these groups determines the positions of the -CH$_n$ groups of different metabolites. More electronegative environments (such as a nearby oxygen atom – CHO) will move more downstream (closer to 10 ppm).

5.1.4 Objective 1.4

As a reminder, objective 1.4 states: "Calculate absolute concentrations (µM) of these differentiating metabolites."

It is not enough to only know which metabolites are present, the concentrations of these metabolites are also important. This is especially true for this study, since the BAYESIL system this manual method is being compared to produces a full metabolite profile of samples, including the absolute concentrations of identified metabolites. To make accurate and fair comparisons, the outputs of these two methods had to be in the same format, hence the calculation of the absolute concentrations of the identified important metabolites.

All metabolites were made relative to creatinine, as described in Chapter 3.6. We could also quantify creatinine in absolute terms by using the creatinine peak at 4.05 ppm. Firstly, the integral of this creatinine peak was divided by the number of protons in creatinine. The same is done for the TSP peak where its integral is divided by 9. These two integrals were then divided by one another and multiplied by the concentration of TSP (1mM) to obtain the absolute concentration of creatinine (mM). By multiplying the relative concentrations of the other metabolites (mmol/mol creatinine) with the absolute concentration of creatinine, we obtained the absolute concentrations (expressed in µM) of all the important metabolites, thereby completing objective 1.4 successfully.

5.1.5 Objective 1.5

Objective 1.5 of this study was to subject the quantified metabolites that potentially differentiate between groups to further statistical analysis to obtain a comparable metabolite profile.

Now that the output of the manual method is in the same format as the BAYESIL output, the metabolite concentration data matrix was subjected to a final round of statistical analysis via MetaboAnalyst, identical to the previous round. The purpose of this was to use univariate and multivariate statistical tests, with their various selection criteria, to obtain the final list of important metabolites that will be used in the comparison with the BAYESIL method. The important metabolite list was a combination of metabolites identified as potential predictors of group membership by the t-test, volcano plot, fold change analysis and PLS-DA.

This last round of statistical analysis resulted in a final list of important metabolites, along with their absolute concentrations, which formed the metabolite profile for these samples using the manual method. Therefore, objective 1.5 was successfully met.

5.2 Objective 2: The automated BAYESIL method

5.2.1 Objective 2.1

Objective 2.1 involved the upload of the raw ^{1}H-NMR spectral data to the BAYESIL website and the combination of the metabolite profile data into a single data matrix, carried out in parallel to the manual method.

This automated method did not require the same manual spectral processing or data clean-up steps because the BAYESIL software is programmed to carry these out automatically. Therefore, the raw ^{1}H-NMR spectral data in the form of FID files were directly uploaded onto the BAYESIL website as individually zipped folders per sample. Since BAYESIL is governed by software and algorithms, it lacks the human quality of intuition – which means it does not have the ability to look at the spectra and conduct a quality check, i.e., a human is able to look at a spectrum and decide

qualitatively whether the spectrum is of good quality. Usually, this is determined by looking at the water region (around 5 ppm) and confirming that the water signal was sufficiently suppressed below the baseline. The beginning of the spectrum is also scrutinized to ensure that the TSP signal was accurately calibrated to exactly 0 ppm and that the resolution is good by measuring the width of the TSP peak. If the width of the TSP peak is <1 Hz, the spectrum is determined to have a good resolution. Herein highlights that the BAYESIL method should not be used as a fully automated tool (yet) and that a trained NMR analyst should inspect all spectra first before uploading the raw NMR data to BAYESIL.

After selecting the relevant parameters and submitting each sample for analysis, a processing time of approximately ten minutes per sample file produced an exportable Excel® document as an output with the list of identified metabolites and their concentrations. Although the processing time is 10 minutes per sample, in reality the BAYESIL analysis took a lot longer. This was mostly because the BAYESIL system queues all uploaded samples and runs them one at a time. When conducting my analysis, another user simultaneously used the website, causing my samples to fall behind theirs in the queue. On average, my samples were 8th in the queue every time, meaning that I had to wait 70 minutes before my sample was analysed. BAYESIL has the potential to produce quick results, but the more popular it gets, the more people will start using it, the longer it will take for each person to obtain their results.

Another important point to raise regarding the BAYESIL system is that almost every sample had a problem with the coverage of the spectrum by the fit produced by BAYESIL. This ranged from 0% - 102% not covered by the fit. This was most likely caused by the absence of a urine algorithm, which led to inaccuracies with the metabolite profile produced by the automated BAYESIL method. This coverage problem also contributed to uncertainty with some of the metabolite identities, such as methylmalonic acid, sarcosine, and taurine, which BAYESIL indicated as having zero confidence in identification.

The data contained in the individual Excel® sheets were combined into a single concentration data matrix for all samples to simplify further data processing. Objective 2.1 was, therefore, successful.

5.2.2 Objective 2.2

This objective states the following: Upload results of BAYESIL to the MetaboAnalyst website and perform the same pre-processing and statistical analysis as performed on the manually generated dataset.

For both the manual and automated methods, all statistical analyses were carried out using MetaboAnalyst. The same statistical tests were applied with identical validation criteria to avoid any unnecessary variation that would impact our ability to compare these methods. Subjecting the BAYESIL concentration data matrix to statistical analysis via MetaboAnalyst yielded a list of metabolites and their concentrations that were identified as being statistically significant. This list differed greatly from the list obtained by the manual method and lacked the TB medication metabolites. This was expected because we knew that BAYESIL had a limited library and may have trouble with the identification of some metabolites. BAYESIL is, however, a high-throughput method and was able to produce a longer list of metabolites. However, these metabolites were yet to be evaluated for their predictive ability and so most of which were not important to us. In contrast, the manual method removes insignificant bins during the first round and produces a shorter list of (more relevant) metabolites, more likely to discriminate between groups. Finally, several metabolite identities produced by BAYESIL carried at least some uncertainty, which is not ideal for application in the clinical setting.

Objective 2.2 was, therefore, successfully completed.

5.3 Objective 3

The last objective for this study was objective 3 and addresses the comparison of the outputs of the in-house and automated processing methods as well as the identification of any differences and what they imply, as well as pitfalls for novice users.

Since both metabolite profiles were obtained in an identical way with the same statistical tests and selection criteria, they are comparable with one another in such a way that would be fair and unbiased. After comparison of these methods, some important advantages and pitfalls were noticed for each method and are provided in the article resulting from this study (Chapter 4).

The most important findings were that the automated BAYESIL method struggled to identify some metabolites due to its limited library, such as those arising from TB disease or the associated medication. BAYESIL also had some issues with the coverage of the spectra and identified metabolites with relatively low confidence and lower specificity. These are very important pitfalls of this method. Advantages of this method included being high-throughput, easy to use, and less time-consuming.

The manual method on the other hand is very time-consuming, labour intensive and requires specialised training and experience. Important advantages include its ability to identify metabolites resulting from the ingestion of TB medication, higher-specificity, and higher confidence in identification of metabolites. The manual method also has the advantage of involving a human,

which allows the use of intuition, gained from experience, to aid in the identification of some metabolites.

The two methods involved in this study could therefore successfully be compared, highlighting some important advantages and pitfalls to both methods. This successfully allowed the completion of objective 3.

5.4 References

Cobas, C. 30 January 2011. *Alignment of NMR spectra – Part II: Binning/bucketing* [Blog post]. http://nmr-analysis.blogspot.com/2011/01/alignment-of-nmr-spectra-part-ii.html Date of access: 14 Mar. 2022

CHAPTER 6 CONCLUSIONS AND FUTURE PROSPECTS

6.1 Conclusion

The successful completion of all objectives set out in Chapter 1.3 allows me to conclude that the aim of this study, which was to determine if automation of the processing of untargeted ^{1}H-NMR metabolomics data has the capability to obtain a metabolic profile of treated and untreated TB patient urine samples was successfully achieved.

Each of the two methods investigated in this study has advantages and disadvantages, some with a greater impact than others. The automated method BAYESIL is high-throughput, identifies more but potentially less predictive metabolites, is easy to use, requires no training, and remains the best currently established automated method for untargeted ^{1}H-NMR metabolomics data processing. Some disadvantages to this method include that it has no algorithm designed for urine, it struggles with complex sample matrices, it has some uncertainty in identification, has a limited library and most importantly – it lacks the human quality of intuition.

On the other hand, the manual method identified fewer metabolites as only those showing some predictive ability are quantified, is time-consuming and labour intensive, and requires an analyst that has received prior ^{1}H-NMR training. The advantages of this method include its ability to use human intuition to aid in the identification of metabolites, there is a possibility to expand the library at any time, more of the metabolites that are identified are significant and the results from this method are generally more accurate and more specific.

From the advantages and pitfalls identified for each method, mentioned here and discussed in more detail in Chapter 4 of this dissertation, it is concluded that the BAYESIL method is currently recommended for shotgun metabolic profiling where not too much detail is desired. If one wishes to perform more detailed research on metabolism, the manual method remains the best suited for this purpose. It may also be useful to combine these methods, so that they may supplement one another's shortcomings.

Although we achieved our aim, the question we set out to investigate does not have a straightforward answer. The automation of processes such as these is already becoming more common and advancing rapidly. Currently, the available automated methods still have quite a way to go before they are fully capable of obtaining a comprehensive metabolite profile of such complex samples. For all we know, automated systems might never reach that point as they will always lack what humans have – intuition, emotion, and thought. It is fascinating and exciting to watch automated metabolomics unfold, but how far will it go?

6.2 Avenues for further research

This study has identified several prospects for future research, such as:

1. During this research, it was noticed that the concentrations of some metabolites obtained by BAYESIL differed from the concentrations obtained by the manual method. This opens an avenue for further investigation into BAYESIL and how to adapt its algorithm to be more accurate in its quantification processes for urine samples.
2. Since statistical analyses across different metabolomics studies are not standardised, there are many alterations that can be made. It would be interesting to investigate the impact on the final result given different steps and options. This could be explored further with a closer look into the statistics and the various available tests. Hopefully this can lead to a more standardised and more optimised statistical analysis process or at least a better understanding of the consequence(s).
3. Collaboration between BAYESIL and the public could possibly lead to the extension of its limited library. For example, including metabolites arising from diseases such as TB and medication metabolites that are of interest.
4. Repeating this study and comparing the manual method with other software systems available for the (semi)-automation of ^{1}H-NMR metabolomics data, such as BQuant, BATMAN, and MetaboHunter. Further investigation into these systems might reveal a better alternative.

ANNEXURE A: HREC APPROVAL LETTER

Private Bag X1290, Potchefstroom
South Africa 2520

Tel: 086 016 9698
Web: http://www.nwu.ac.za/

North-West University Health Research Ethics Committee (NWU-HREC)

Tel: 018 299-1206
Email: Ethics-HRECApply@nwu.ac.za (for human studies)

7 December 2021

ETHICS APPROVAL LETTER OF STUDY

Based on approval by the North-West University Health Research Ethics Committee (NWU-HREC) on 07/12/2021, the NWU-HREC hereby approves your study as indicated below. This implies that the NWU-HREC grants its permission that, provided the general conditions specified below are met and pending any other authorisation that may be necessary, the study may be initiated, using the ethics number below.

Study title: Automated processing of untargeted 1H-NMR metabolomics data of urine from treated TB patients
Principal Investigator/Study Supervisor/Researcher: Dr SW Mason
Student: J van der Westhuizen - 28710665

Ethics number:

N	W	U	-	0	0	1	2	7	-	1	1	-	A	1	-	0	2
Institution				Study Number						Year			Status		Sub-study		

Status: S = Submission; R = Re-Submission; P = Provisional Authorisation; A = Authorisation

Application Type: Sub-study
Commencement date: 07/12/2021
Expiry date: 28/02/2023

Risk: Minimal

Approval of the study is provided for a year, after which continuation of the study is dependent on receipt and review of an annual monitoring report and the concomitant issuing of a letter of continuation. A monitoring report is due at the end of February annually until completion of the study.

General conditions:

While this ethics approval is subject to all declarations, undertakings and agreements incorporated and signed in the application form, the following general terms and conditions will apply:

- *The principal investigator/study supervisor/researcher must report in the prescribed format to the NWU-HREC:*
  - *Annually on the monitoring of the study, whereby a letter of continuation will be provided annually, and upon completion of the study; and*
  - *without any delay in case of any adverse event or incident (or any matter that interrupts sound ethical principles) during the course of the study.*
- *The approval applies strictly to the proposal as stipulated in the application form. Should any amendments to the proposal be deemed necessary during the course of the study, the principal investigator/study supervisor/researcher must apply for approval of these amendments at the NWU-HREC, prior to implementation. Should there be any deviations from the study proposal without the necessary approval of such amendments, the ethics approval is immediately and automatically forfeited.*
- *Annually a number of studies may be randomly selected for active monitoring.*
- *The date of approval indicates the first date that the study may be started.*
- *In the interest of ethical responsibility, the NWU-HREC reserves the right to:*
  - *request access to any information or data at any time during the course or after completion of the study;*
  - *to ask further questions, seek additional information, require further modification or monitor the conduct of your research or the informed consent process;*

- *withdraw or postpone approval if:*
  - *any unethical principles or practices of the study are revealed or suspected;*
  - *it becomes apparent that any relevant information was withheld from the NWU-HREC or that information has been false or misrepresented;*
  - *submission of the annual monitoring report, the required amendments, or reporting of adverse events or incidents was not done in a timely manner and accurately; and/or*
  - *new institutional rules, national legislation or international conventions deem it necessary.*
- *NWU-HREC can be contacted for further information via Ethics-HRECApply@nwu.ac.za or 018 299 1206*

Special conditions of the research approval due to the COVID-19 pandemic:

Please note: Due to the nature of the study i.e. (laboratory work involving the molecular analysis of previously collected biological samples), this study will be able to proceed during the current alert level, following receipt of the approval letter. No additional COVID-19 restrictions have been placed on the study except that the researcher must ensure that before proceeding with the study that all research team members have reviewed the North-West University COVID-19 Occupational Health and Safety Standard Operating Procedure.

The NWU-HREC would like to remain at your service and wishes you well with your study. Please do not hesitate to contact the NWU-HREC for any further enquiries or requests for assistance.

Yours sincerely,

Digitally signed by
Prof Petra Bester
Date: 2021.12.07
10:55:50 +02'00'

Chairperson NWU-HREC

Current details (25239522) G:\My Drive\9. Research and Postgraduate Education\9.1.5.4 Templates\9.1.5.4.2_NWU-HREC_EAL.docm
20 August 2019
File Reference: 9.1.5.4.2

ANNEXURE B: ADVANCED TRAINING RESULTS

Urine

Day	Analyst	Repeat		Valine	Leucine	3-Hydroxybutyric acid	3-Hydroxyisovaleric acid	Lactic acid	Alanine	2-Hydroxyisobutyric acid	Acetic acid	Succinic acid	Citric acid	Methylamine	Dimethylamine	Creatine	Creatinine	Choline	Carnitine	Glycine	Tyrosine	Formic acid
	1 D		1	128,341	89,8379	219,941	75,5496	410,264	367,98	118,374	413,052	211,937	1232,05	61,8405	107,297	3019,27	6342,88	176,348	195,931	1311,24	1559,85	121,628038
	1 D		2	129,896	90,5345	218,709	77,8986	412,073	369,742	119,078	414,963	211,931	1232,85	61,6336	107,173	3012,07	6410,73	175,725	195,752	1311,79	1556,97	121,864553
	1 D		3	125,152	87,2254	214,962	74,7174	398,17	359,094	114,569	401,669	206,056	1194,47	60,6934	103,874	2930,97	6197,3	170,493	190,677	1276,62	1515,83	118,221269
	1 D		4	126,454	89,3844	218,379	75,2493	408,313	365,452	117,439	413,336	210,611	1223,36	61,8415	106,463	3000	6357,92	174,412	195,016	1306,09	1553,47	120,868715
	1 D		5	127,639	89,5213	217,97	76,7884	406,023	367,273	117,828	407,525	208,356	1213,02	60,6759	105,43	2973,01	6314,7	172,972	193,469	1295,11	1537,28	120,129184
	2 D		1	127,924	90,0538	264,476	77,9262	411,097	371,618	117,653	415,616	212,471	1222,05	62,7942	105,706	3014,26	6274,81	172,544	195,633	1311,43	1553,32	123,694877
	2 D		2	127,801	89,6268	242,123	75,1614	407,404	365,082	115,801	414,291	210,72	1224,43	61,3333	105,748	2998,96	6272,25	174,065	193,734	1299,19	1543,61	123,035154
	2 D		3	128,701	89,6471	274,182	79,1112	414,787	373,164	118,855	417,086	212,209	1221,72	62,7414	105,579	3015,19	6242,65	173,593	195,743	1311,78	1551,44	124,766119
	2 D		4	127,021	89,2732	249,931	77,0115	407,939	365,909	115,834	410,916	209,597	1211,54	61,9233	104,416	2978,87	6226,05	171,743	193,035	1296,63	1537,07	120,888997
	2 D		5	127,227	89,9726	260,233	77,7233	410,546	369,777	117,52	416,177	211,555	1223,88	61,9617	105,642	3005,14	6282,65	173,866	194,864	1305,17	1545,78	123,825608
	3 D		1	130,367	90,934	224,99	76,9487	417,263	377,363	119,394	426,256	216,665	1249,59	63,2181	107,597	3082,06	6438,22	176,781	200,083	1337,95	1586,85	126,069226
	3 D		2	127,506	88,8232	217,943	76,7357	404,829	366,226	115,333	410,304	208,991	1194,22	61,067	103,719	2947,09	6203,45	169,416	191,461	1283,88	1517,71	120,087588
	3 D		3	125,738	87,6377	217,375	75,843	406,865	367,899	115,554	409,051	208,631	1197,95	61,1907	103,666	2955,75	6161,47	170,509	191,442	1288,42	1519,65	120,82128
	3 D		4	117,796	81,676	202,366	70,0719	375,374	338,609	106,705	382,286	194,853	1116,46	57,1514	96,4823	2743,88	5822,36	158,758	178,333	1197,19	1417,4	112,686284
	3 D		5	126,842	89,6543	217,559	76,7123	408,061	367,394	115,993	415,423	211,944	1218,04	60,878	105,684	2996,1	6338,86	173,474	193,496	1300,89	1540,47	123,637638
	4 D		1	127,66	89,7044	221,296	75,9673	409,951	368,816	114,374	419,112	213,535	1222,61	62,5005	105,879	3016,83	6351,73	172,917	195,462	1310,95	1552,03	125,831237
	4 D		2	137,978	95,6533	236,661	83,4702	441,907	396,92	123,518	448,976	230,062	1311,25	67,1136	112,987	3246,77	6762,74	185,128	211,677	1405,8	1666,96	133,919355
	4 D		3	127,101	89,3502	219,524	75,9693	409,516	367,791	114,805	419,263	212,581	1216,28	62,2426	105,855	3017,17	6289,27	173,416	195,494	1311,15	1550,53	126,541985
	4 D		4	129,932	89,9003	220,736	76,5706	411,296	369,156	115,728	418,153	214,581	1229,09	61,9876	106,332	3032,85	6359,96	174,199	196,442	1316,63	1558,07	126,437257
	4 D		5	127,252	89,4665	218,599	75,0807	409,173	366,456	114,26	418,046	213,72	1222,29	61,3485	106,149	3013,75	6363,5	173,612	194,925	1308,25	1547,72	124,637275
				127,496	89,3007	217,992	76,0406	406,969	365,908	117,457	410,109	209,778	1219,15	61,337	106,047	2987,06	6324,7	173,99	194,169	1300,17	1544,68	120,542352
1		mean		1,80782	1,24231	1,84708	1,28737	5,40921	4,10781	1,72809	5,48948	2,54295	15,9559	0,60152	1,42345	35,9654	79,3182	2,34609	2,18015	14,7713	18,3405	1,46556945
		STD		1,41794	1,39116	0,84731	1,69301	1,32915	1,12264	1,47125	1,33854	1,21221	1,30877	0,98068	1,34227	1,20404	1,2541	1,3484	1,12281	1,1361	1,18733	1,21581288
		CV%		127,735	89,7147	258,189	77,3867	410,354	369,11	117,132	414,818	211,31	1220,73	62,1508	105,418	3002,49	6259,68	173,162	194,602	1304,84	1546,24	123,242151
2		mean		0,65992	0,31171	12,5079	1,45522	2,94849	3,52268	1,30829	2,40462	1,17255	5,2637	0,61615	0,56374	14,8104	24,1628	0,98614	1,18772	6,91046	6,48945	1,45327571
		STD		0,51663	0,34744	4,84447	1,88045	0,71852	0,95437	1,11693	0,57968	0,55489	0,43119	0,99138	0,53476	0,49327	0,38601	0,56949	0,61033	0,5296	0,41969	1,17920346
		CV%		125,65	87,7451	216,047	75,2623	402,478	363,498	114,596	408,664	208,217	1195,25	60,7011	103,43	2944,98	6192,87	169,788	190,963	1281,67	1516,42	120,660403
3		mean		4,71171	3,59936	8,28837	2,93235	15,8839	14,609	4,70878	16,2315	8,1338	49,2125	2,19888	4,20887	124,463	234,462	6,79739	7,89895	51,7784	61,9578	5,05123395
		STD		3,74987	4,10207	3,83638	3,89617	3,94651	4,019	4,10904	3,97184	3,90641	4,11733	3,62248	4,0693	4,22628	3,78599	4,00347	4,13637	4,03993	4,0858	4,18632278
		CV%		129,985	90,8149	223,363	77,4116	416,369	373,828	116,537	424,71	216,896	1240,3	63,0386	107,44	3065,47	6425,44	175,854	198,8	1330,56	1575,06	127,473422
4		mean		4,61192	2,71307	7,50726	3,4283	14,2991	12,9517	3,945	13,5761	7,39417	39,9183	2,31787	3,107	101,621	190,964	5,20425	7,21933	42,1731	51,5102	3,68211078
		STD		3,54805	2,98748	3,36101	4,42867	3,43424	3,46461	3,38519	3,19656	3,40909	3,21843	3,6769	2,89183	3,31501	2,972	2,95941	3,63145	3,16959	3,27036	2,88853216
		CV%		2,30812	2,20703	3,22229	2,97458	2,3571	2,39016	2,5206	2,27166	2,27065	2,26893	2,31786	2,20954	2,30965	2,09952	2,22019	2,37524	2,21881	2,2408	2,36746782
		Total CV%																				

Urine CV				
Average	2,38	Min	2,10	Max	3,22

Serum

Day	Analyst	Repeat	Valine	Lactate	Alanine	2-Hydroxyisobutyrate	Acetate	N-acetylglutamine	Acetoacetate	Citrate	Choline	Creatine	Creatinine	Glucose	Tyrosine	Nicotinamide	Tryptophan
1	D	1	321,722	54,7851	776,267	89,1884	188,695	34226,1	393,09	1105,41	89,2523	185,994	253,767	1229,76	277,861	868,521	60,42265377
1	D	2	325,318	55,0645	783,44	79,138	191,754	33268,9	377,854	1116,77	90,512	186,703	255,154	1251,13	280,225	877,962	61,10123599
1	D	3	323,452	54,8346	783,391	81,3201	188,526	32999	393,843	1113,94	89,8863	186,23	252,481	1243,3	279,185	872,127	60,8583761
1	D	4	324,12	55,5099	782,769	82,776	192,963	34328,6	372,868	1117,75	91,3381	191,11	270,98	1403,57	281,9	875,162	60,99527032
1	D	5	329,244	55,1701	789,436	84,01	193,725	33886,4	397,89	1129,63	91,7088	188,751	258,662	1266,03	283,971	890,246	62,08180525
2	D	1	316,036	53,8142	758,404	82,1123	185,542	32801	374,347	1085,71	88,469	180,272	250,171	1214,71	272,281	848,968	59,71899049
2	D	2	314,497	53,7936	757,937	79,8798	187,045	32370,6	373,676	1082,2	88,2828	179,236	250,534	1209,15	271,693	847,156	59,1516858
2	D	3	314,51	53,4164	754,409	79,1464	187,42	32345,7	369,807	1082,01	88,1361	178,916	248,306	1210,27	271,717	847,567	58,65478633
2	D	4	316,371	55,1885	760,741	79,5513	185,969	32194,1	373,153	1084,81	88,3474	178,645	248,314	1213,58	271,93	848,396	59,11920077
2	D	5	316,181	54,1338	759,038	81,0041	186,457	32825,6	373,077	1085,3	88,5692	179,308	249,661	1215,21	272,715	850,468	59,54501315
3	D	1	319,466	54,8221	769,743	79,5762	188,466	33193	369,291	1098,06	89,9754	180,051	254,224	1228,66	275,679	855,889	59,61438158
3	D	2	321,454	54,6864	770,128	78,9582	190,732	32757,1	372,626	1102,81	90,6072	178,833	256,44	1233,82	277,428	863,061	59,73642989
3	D	3	322,73	54,7483	776,109	77,8526	191,052	33063,6	369,793	1107,91	91,0746	179,711	256,346	1239,66	278,642	867,15	60,38403975
3	D	4	321,802	55,2792	774,885	78,4005	190,641	33233,9	364,315	1107,23	90,8484	181,337	255,602	1236,91	278,114	864,221	59,87555257
3	D	5	324,25	54,7317	778,265	76,8841	191,344	33240,9	364,502	1114,89	91,4945	181,115	257,715	1247,11	280,118	872,615	59,80728216
4	D	1	325,539	55,5215	780,304	76,6641	192,468	33321,7	363,548	1118,06	92,8745	179,799	262,421	1252,35	281,04	876,802	59,69309695
4	D	2	322,841	55,3951	774,327	76,6981	191,737	33263	356,682	1109,69	92,0359	177,838	258,664	1243,16	278,32	869,608	59,56079217
4	D	3	334,865	57,0523	802,695	86,8222	197,508	35669,8	381,347	1148,31	95,2655	183,364	271,448	1275,96	289,518	900,085	61,42857516
4	D	4	326,631	55,8169	784,035	77,9667	192,455	33781,6	365,003	1121,79	93,2867	179,57	262,624	1257,09	281,784	879,932	60,71197401
4	D	5	325,551	56,0744	781,53	78,0412	194,445	33626	362,692	1117,45	92,8322	179,049	263,278	1253,33	280,846	877,51	60,50341306
1		mean	324,771	55,0729	783,06	83,2865	191,132	33741,8	387,109	1116,7	90,5395	187,758	258,209	1278,76	280,629	876,804	61,09186829
		STD	2,81818	0,2916	4,66719	3,76477	2,40784	586,347	11,0203	8,70862	1,01108	2,167	7,50287	70,9971	2,38202	8,29444	0,610688297
		CV%	0,86774	0,52947	0,59602	4,52026	1,25978	1,73775	2,84683	0,77985	1,11673	1,15415	2,90574	5,55204	0,84882	0,94599	0,999622886
2		mean	315,519	54,0693	758,106	80,3388	186,487	32507,4	372,812	1084	88,3609	179,275	249,397	1212,58	272,067	848,511	59,23793531
		STD	0,93457	0,67532	2,32349	1,20838	0,76542	287,41	1,7551	1,76836	0,16728	0,61701	1,03961	2,71604	0,43202	1,30156	0,414520955
		CV%	0,2962	1,24898	0,30649	1,5041	0,41044	0,88414	0,47077	0,16313	0,18931	0,34417	0,41685	0,22399	0,15879	0,15339	0,699755912
3		mean	321,94	54,8536	773,826	78,3343	190,447	33097,7	368,105	1106,18	90,8	180,21	256,066	1237,24	277,996	864,587	59,88353719
		STD	1,75595	0,24293	3,75435	1,03332	1,14195	203,302	3,60684	6,26968	0,56504	1,03198	1,279	6,86835	1,6301	6,10603	0,29600114
		CV%	0,54543	0,44288	0,48517	1,31911	0,59961	0,61425	0,97984	0,56679	0,62229	0,57265	0,49948	0,55514	0,58638	0,70624	0,494294683
4		mean	327,085	55,972	784,578	79,2384	193,723	33932,4	365,854	1123,06	93,259	179,924	263,687	1256,38	282,302	880,787	60,37957027
		STD	4,56884	0,65908	10,7372	4,29085	2,34359	994,506	9,22235	14,7856	1,20961	2,06723	4,70023	12,0852	4,23981	11,4533	0,769396475
		CV%	1,39684	1,17752	1,36853	5,41511	1,20977	2,93084	2,52077	1,31655	1,29705	1,14895	1,7825	0,96191	1,50187	1,30035	1,274266233
	Total CV%		0,77655	0,84971	0,68905	3,18965	0,8699	1,54174	1,70455	0,70658	0,80635	0,80498	1,40114	1,82327	0,77396	0,77649	0,866984928

Serum CV					
Average	1,17	Min	0,69	Max	3,19

ANNEXURE C: SUPPLEMENTARY INFORMATION FOR 'A NOVICE'S GUIDE TO PROCESSING UNTARGETED ^{1}H NMR METABOLOMICS DATA'

Buffer recipe

Table S1. Recipe used to make 100 mL of NMR buffer (adapted from the BAYESIL website)

Ingredients	Mass to be added (g)	Concentration (in buffer)	Final concentration (in each sample)
KH_2PO_4 MW = 136.09 g/mol	23.816	1.75 M	150 mM
2-chloropyrimidine-5-carboxylic acid MW = 158.54 g/mol	0.0926	5.84 mM	500 µM
TSP MW = 172.27 g/mol	0.202	11.7 mM	1 mM

Table S2. Full MetaboAnalyst results, including those for the manual binned data

PCA			***PLS-DA***					**No. of important compounds**		
Method	**Group**	**Separation**	**Separation**	**Stability**		**Performance**		**PLS-DA**	**Volcano**	**t-test**
				R^2	**Q^2**	**P value**	**Accuracy (%)**			
BAYESIL	HC vs W0	1	2	0.75	0.67	< 0.0005	92	12	7	8
Manual (bins)	HC vs W0	1	3	0.87	0.75	< 0.0005	94	-	-	-
Quantified manual	HC vs W0	2	2	0.77	0.72	< 0.0005	97	8	1	5
BAYESIL	W0 vs W2	1	2	0.67	0.54	< 0.0005	86	17	7	17
Manual (bins)	W0 vs W2	1	2	0.80	0.65	0.006	90	-	-	-
Quantified manual	W0 vs W2	2	2	0.83	0.80	< 0.0005	95	7	7	16

Statistical figures

Quantified manual data:

(a) (b)

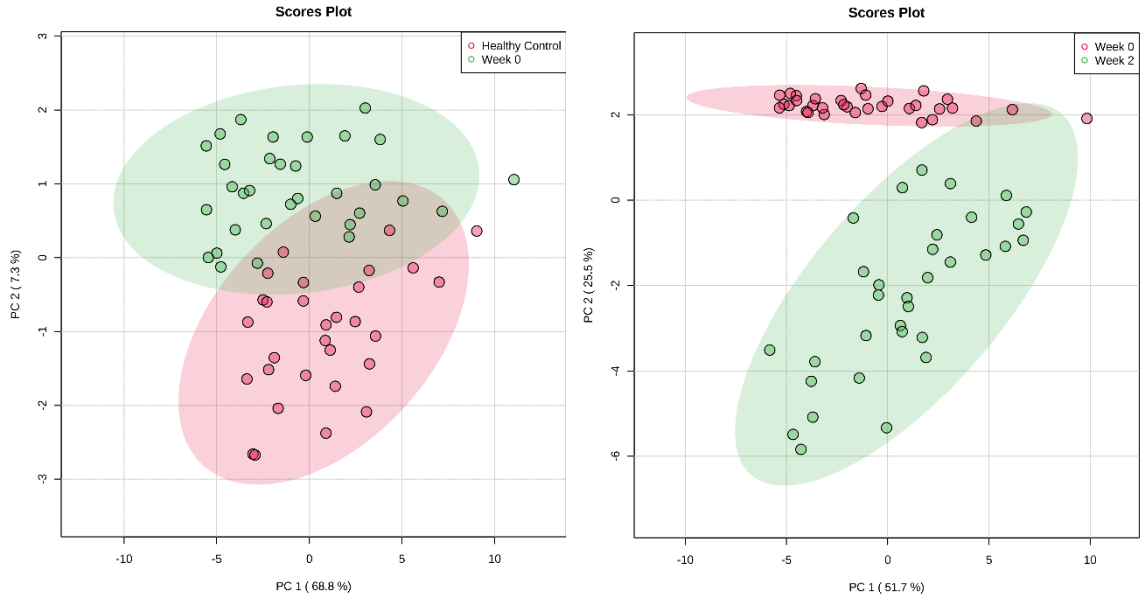


Figure S1. (a) The PCA result for the HC vs W0 comparison of quantified manual data. (b) The PCA result for the W0 vs W2 groups of quantified manual data

(a) (b)

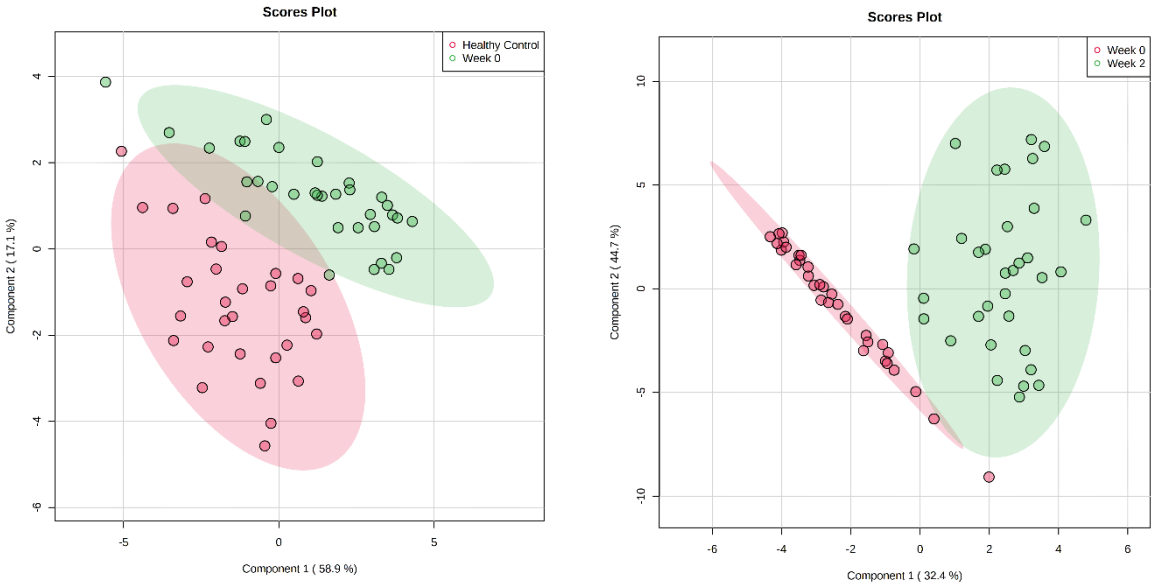


Figure S2. (a) The PLS-DA result for the HC vs W0 groups of the quantified manual data. (b) The PLS-DA result for the W0 vs W2 groups of the quantified manual data

(a) (b)

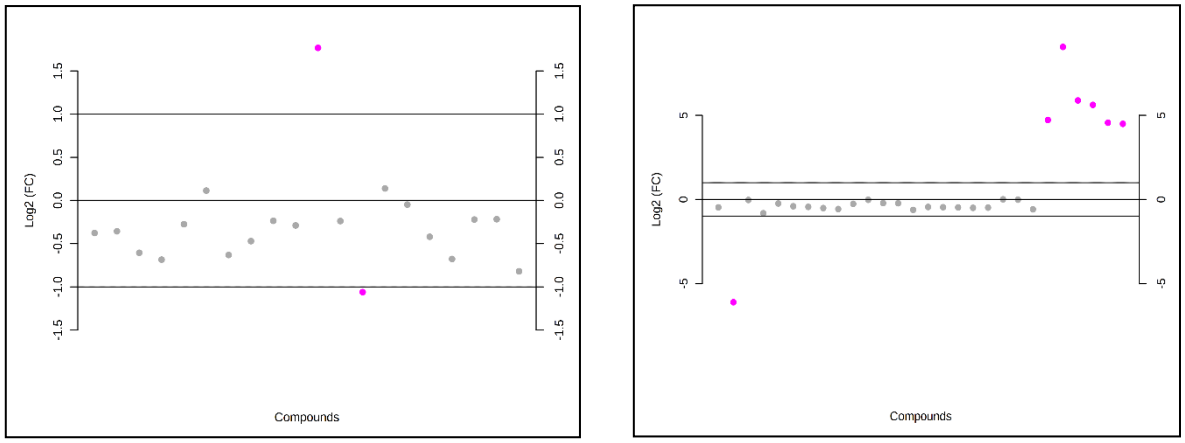


Figure S3. (a) The fold change analysis of the HC vs W0 groups of the quantified manual data. (b) The fold change analysis of the W0 vs W2 groups of the quantified manual data

(a) (b)

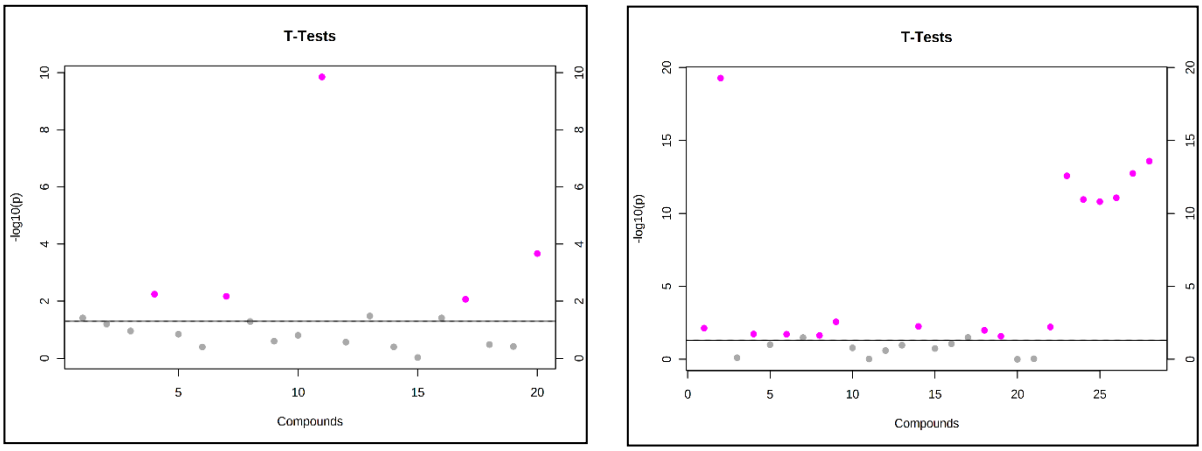


Figure S4. (a) The unpaired t-test result from the HC vs W0 groups of the quantified manual data. (b) The paired t-test result of the W0 vs W2 groups of the quantified manual data

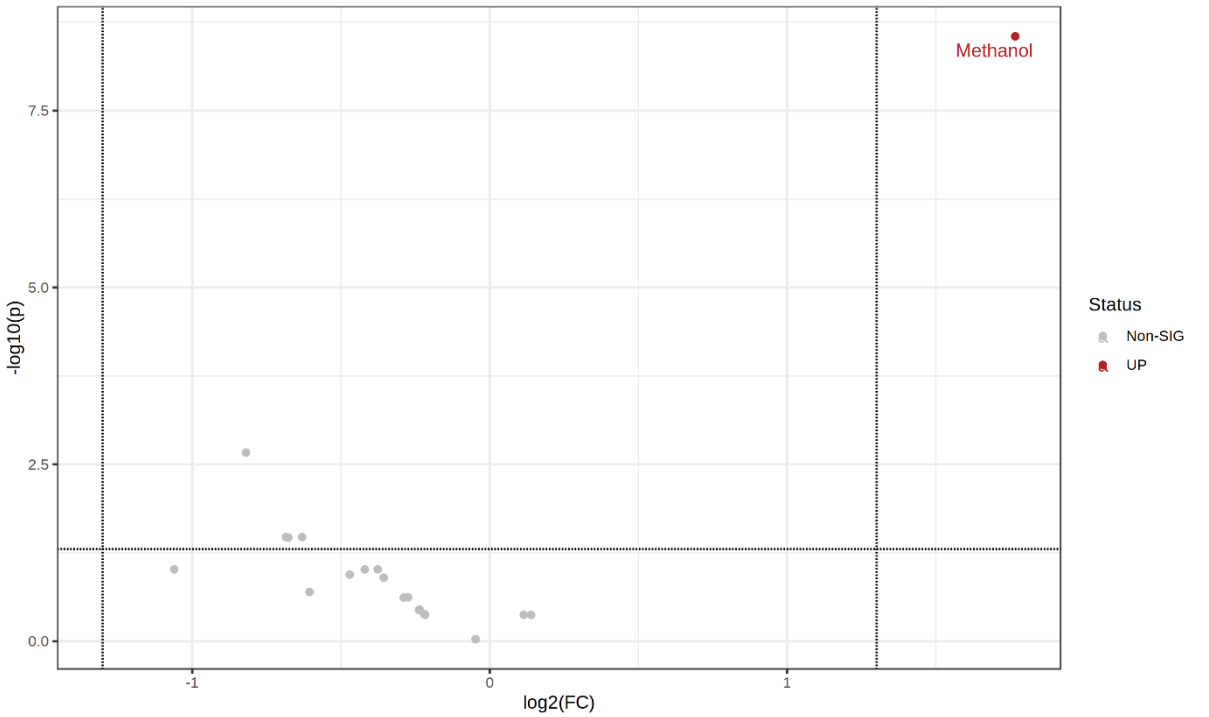


Figure S5. The volcano plot result from the HC vs W0 groups of the quantified manual data, showing only methanol as significant

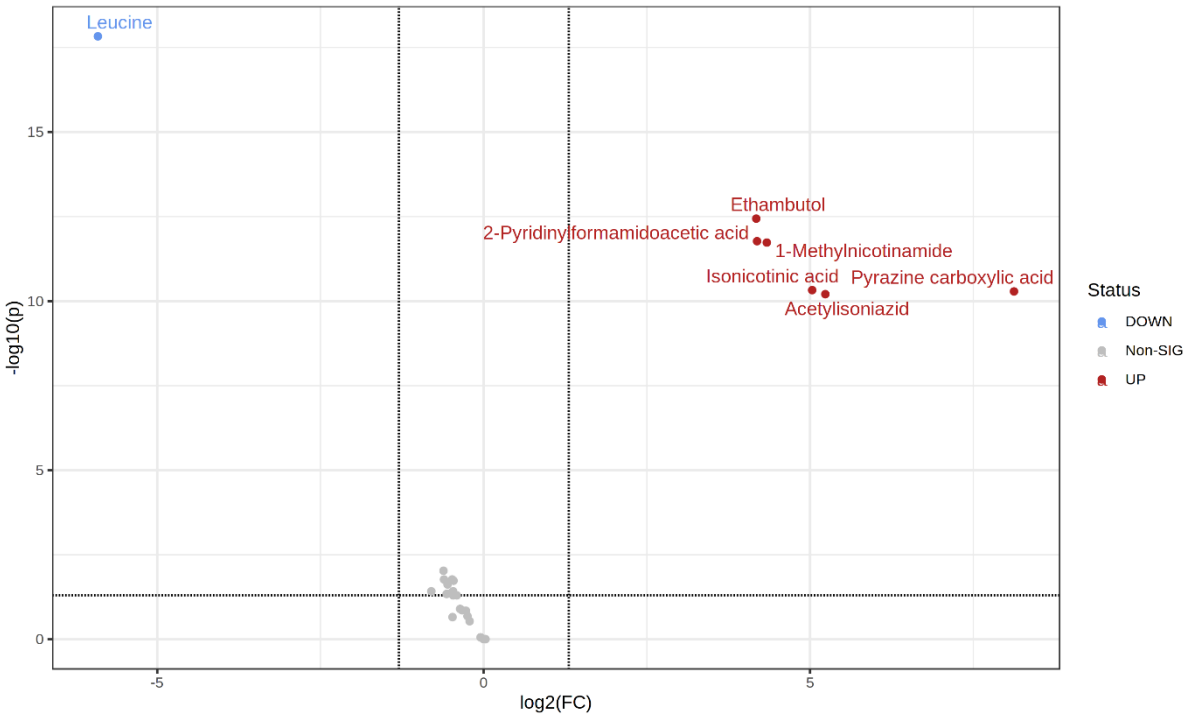


Figure S6. The volcano plot from the W0 vs W2 groups of the quantified manual data, showing mainly TB medication metabolites as significant

BAYESIL data:

(a) (b)

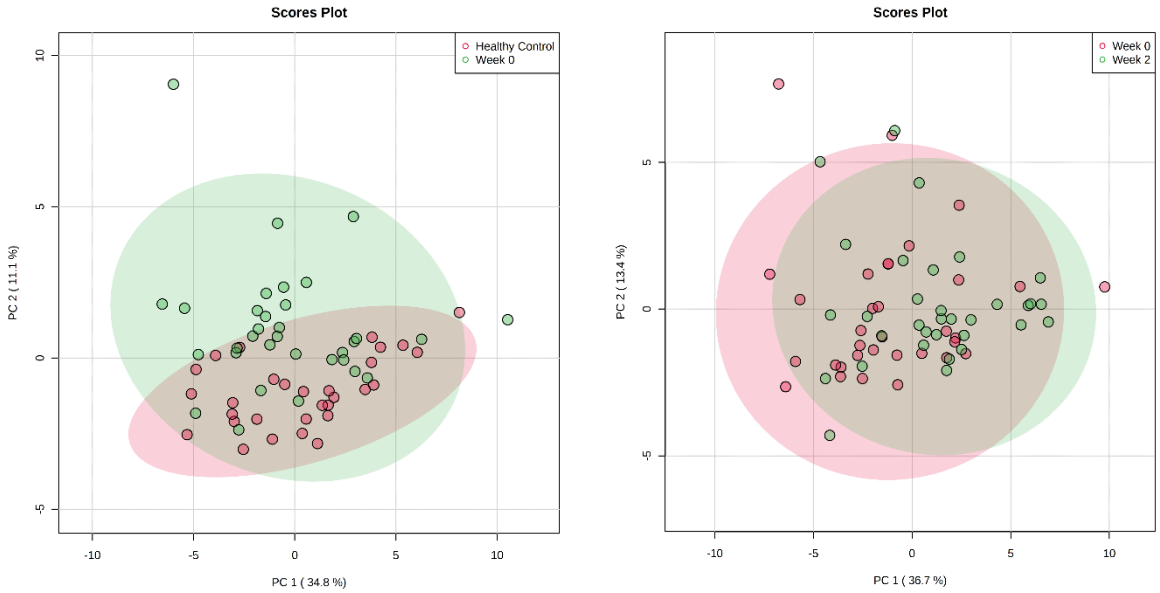


Figure S7. (a) The PCA result from the HC vs W0 groups of the BAYESIL data. (b) The W0 vs W2 PCA result for the BAYESIL data

(a) (b)

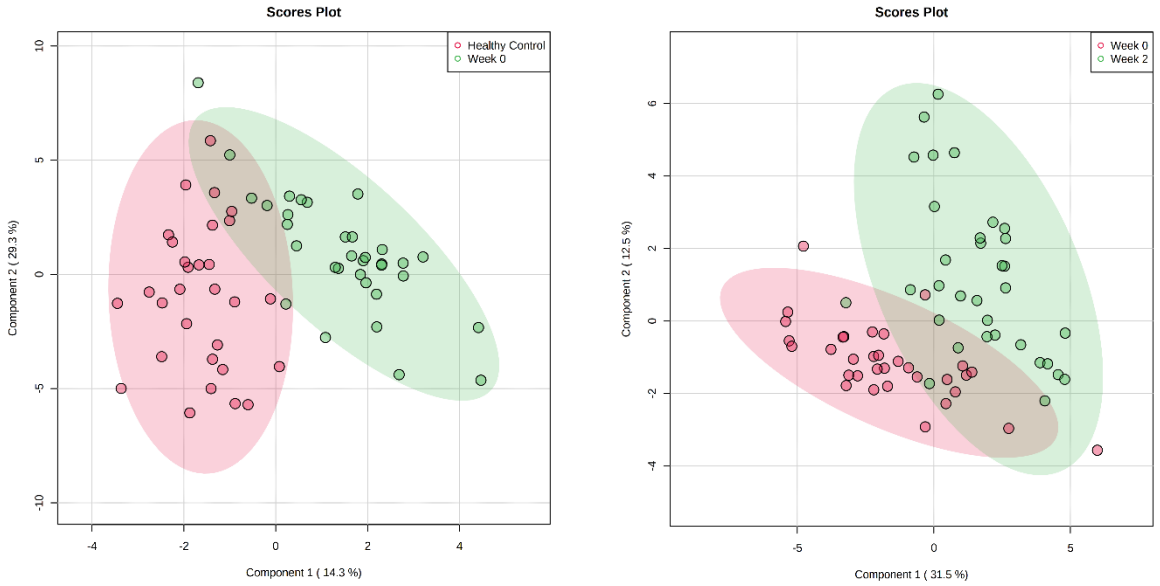


Figure S8. (a) The PLS-DA result of the HC vs W0 group comparison of the BAYESIL data. (b) The result after subjecting the W0 vs W2 groups of the BAYESIL data to PLS-DA analysis

(a) (b)

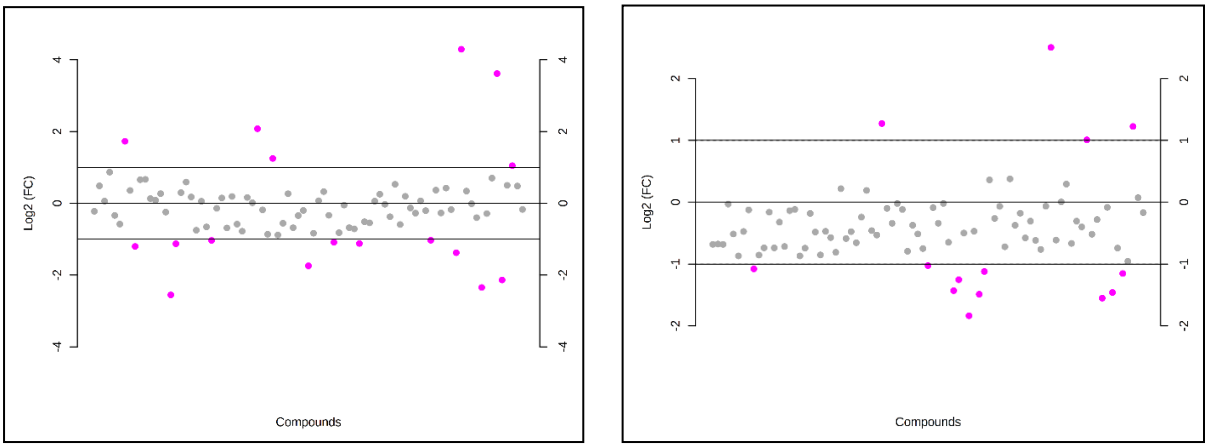


Figure S9. (a) The result after subjecting the HC vs W0 groups of the BAYESIL data to fold change analysis. (b) The fold change result for the W0 vs W2 groups of the BAYESIL data

(a) (b)

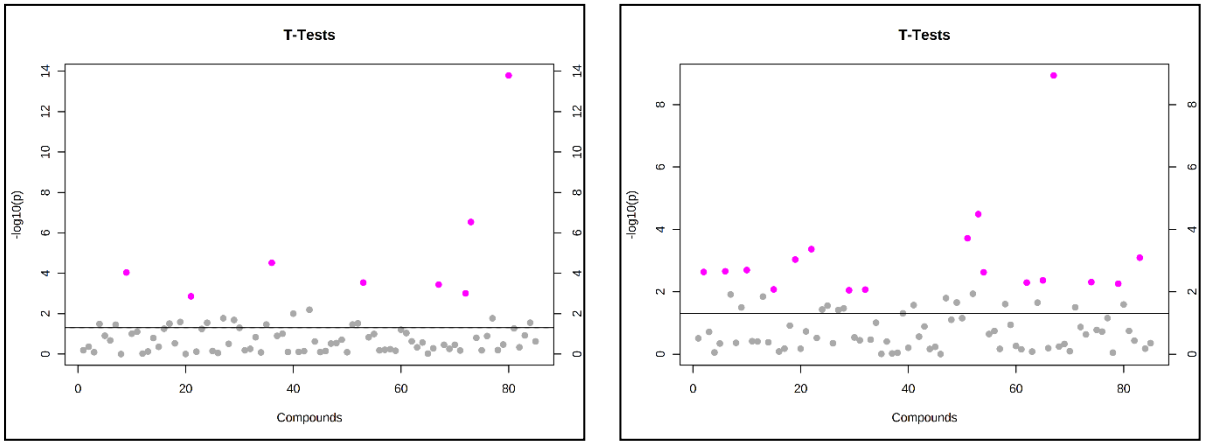


Figure S10. (a) The result of the unpaired t-test conducted on the HC vs W0 groups of the BAYESIL data. (b) The paired t-test result from the W0 vs W2 groups of the BAYESIL data.

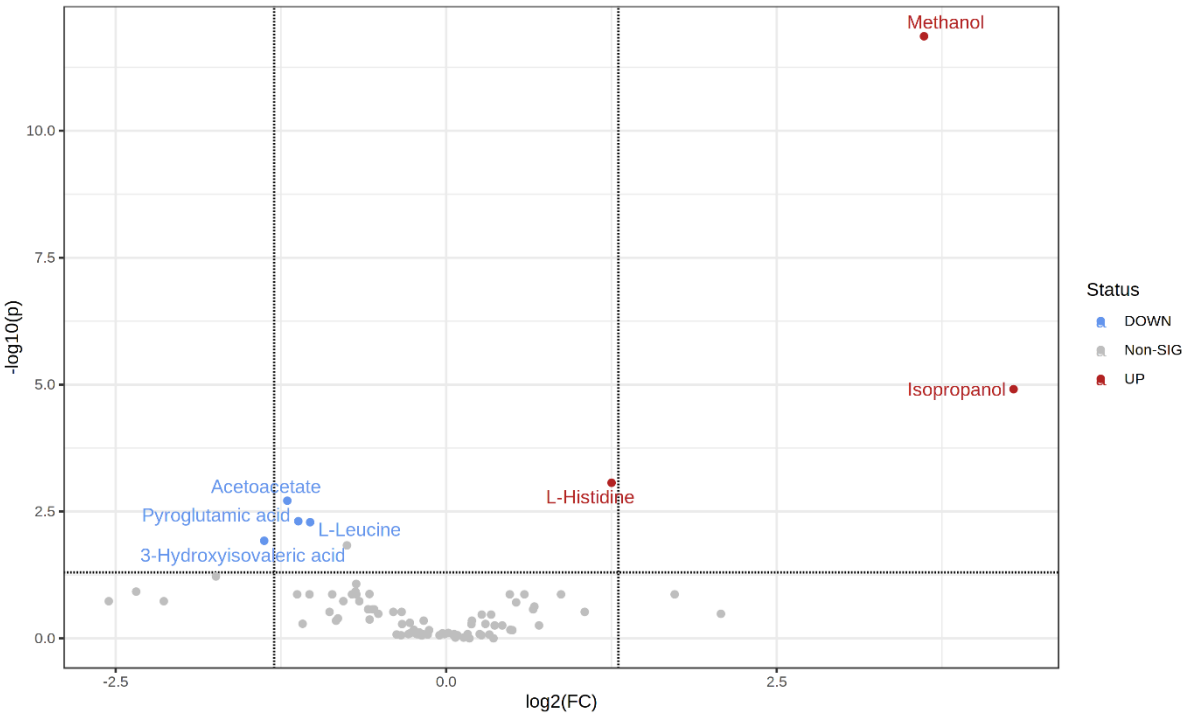


Figure S12. The volcano plot for the HC vs W0 groups of the BAYESIL data, showing the significant metabolites in blue and red

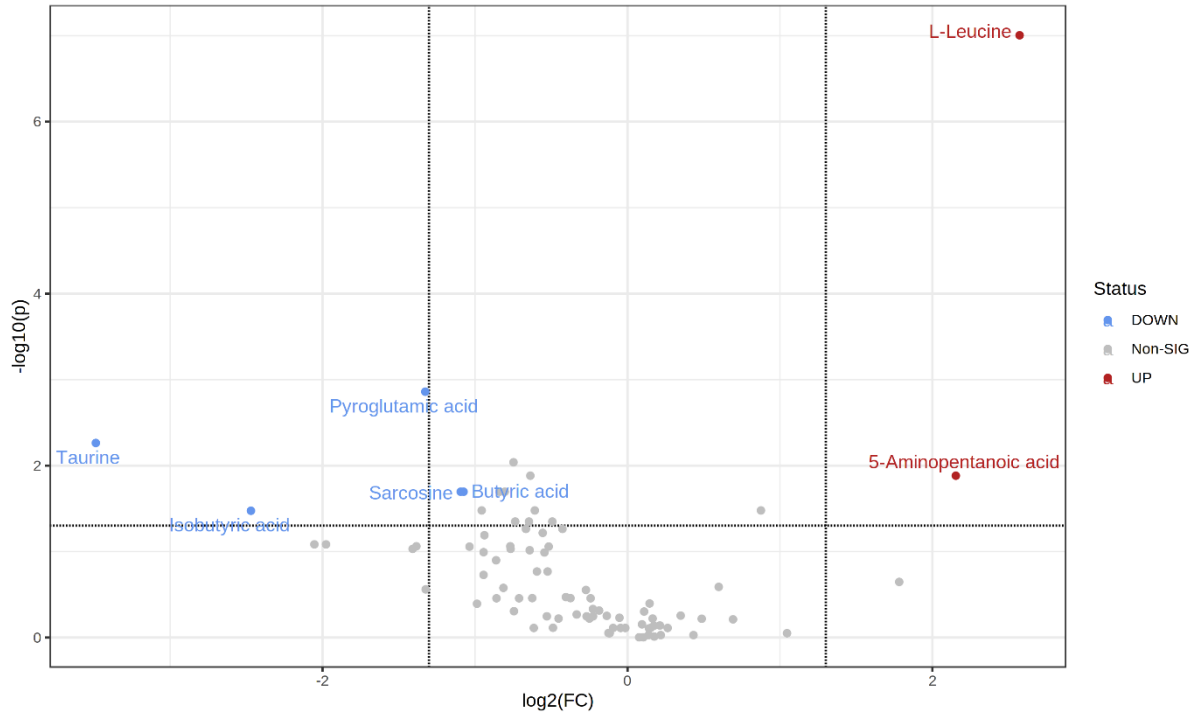


Figure S11. The volcano plot result for the W0 vs W2 groups of the BAYESIL data, indicating significant metabolites in red and blue

Identified metabolites

Table S3. The list of metabolites that were identified by each method, but were not found to be significant for the HC vs W0 groups

			Rule for significance					
	Confidence in ID		*|logFC| ≥ 1*		*t-test FDR p-value ≤ 0.05*		*PLSDA VIP ≥ 1 for LV 1 or 2*	
Compound	**Auto***	**Q. Man**	**Auto***	**Q. Man**	**Auto***	**Q. Man**	**Auto***	**Q. Man**
Trimethylamine	–	10	×	×	×	×	×	×
Citric acid	10	N/A	×		×		×	
Valine	10	10	×	×	×	×	×	×
Isoleucine	7	10	×	×	×	×	×	×
3-Hydroxybutyric acid	10	10	×	×	×	×	×	×
Methyl guanidine	N/A	10		×		×		×
Methylamine	–	10	×	×	×	×	×	×
Taurine	–	10	×	×	×	×	×	×
Sucrose	N/A	10		×		×		×
3-Hydroxyphenylacetic acid	N/A	10		×		×		×
Histamine	N/A	10		×		×		×
Hypoxanthine	10	10	×	×	×	×	×	×
1-Methylhistidine	5	N/A	×		×		×	
2-Hydroxybutyrate	9	N/A	×		×		×	
2-Oxoisovalerate	–	N/A	×		×		×	
p-Hydroxyphenyl acetic acid	–	N/A	×		×		×	
3-Hydroxyisobutyrate	9	N/A	×		×		×	
Butyric acid	–	N/A	×		×		×	
Acetic acid	10	N/A	×		×		×	
Ascorbic acid	–	N/A	×		×		×	
Carnitine	9	N/A	×		×		×	
Creatine	7	N/A	×		×		×	
Dimethylamine	10	N/A	×		×		×	
Dimethylglycine	–	N/A	×		×		×	
Choline	10	N/A	×		×		×	
Ethanol	9	N/A	×		×		×	
Glycine	10	N/A	×		×		×	
Guanidoacetic acid	–	N/A	×		×		×	
Glycerol	9	N/A	×		×		×	
Fumaric acid	–	N/A	×		×		×	
Galactose	–	N/A	×		×		×	

Glutamic acid	10	N/A	×	×	×
Ethanolamine	–	N/A	×	×	×
Tyrosine	10	N/A	×	×	×
Alanine	10	N/A	×	×	×
Threonine	9	N/A	×	×	×
Asparagine	3	N/A	×	×	×
Mannose	3	N/A	×	×	×
Fucose	–	N/A	×	×	×
Lysine	9	N/A	×	×	×
Alpha-Lactose	–	N/A	×	×	×
Serine	4	N/A	×	×	×
Aspartic acid	9	N/A	×	×	×
Cystine	3	N/A	×	×	×
2-Oxoglutaric acid	–	N/A	×	×	×
Phenylacetic acid	–	N/A	×	×	×
Myoinositol	9	N/A	×	×	×
Ornithine		N/A	×	×	×
Oxalacetic acid	–	N/A	×	×	×
Pyruvic acid	8	N/A	×	×	×
Sorbitol	–	N/A	×	×	×
Sarcosine	–	N/A	×	×	×
Xanthine	7	N/A	×	×	×
Urea	8	N/A	×	×	×
Uracil	–	N/A	×	×	×
2-Hydroxyisovaleric acid	10	N/A	×	×	×
3-Methylhistidine	–	N/A	×	×	×
Arginine	6	N/A	×	×	×
Creatinine	10	N/A	×	×	×
Cysteine	3	N/A	×	×	×
Gluconic acid	–	N/A	×	×	×
Glutamine	10	N/A	×	×	×
Fructose	7	N/A	×	×	×
Malonic acid	7	N/A	×	×	×
Methionine	10	N/A	×	×	×
Hippuric acid	–	N/A	×	×	×
Isovalerate	–	N/A	×	×	×
4-Hydroxyphenyllactic acid	–	N/A	×	×	×
Tryptophan	5	N/A	×	×	×
Phosphorylcholine	–	N/A	×	×	×
Acetone	10	N/A	×	×	×
Caffeine	–	N/A	×	×	×
Isobutyric acid	10	N/A	×	×	×
Propylene glycol	10	N/A	×	×	×
1,5-Anhydrosorbitol	9	N/A	×	×	×
5-Aminopentanoic acid	–	N/A	×	×	×
Dimethyl sulfone	10	N/A	×	×	×
p-Cresol sulfate	–	N/A	×	×	×
EDTA	10	N/A	×	×	×

Key:	
×	Identified by method, but not significant according to stats
	Not identified by method
–	No confidence score given (assumed to be 0)
Auto*	Automated BAYESIL method
Q. Man	Quantified manual method

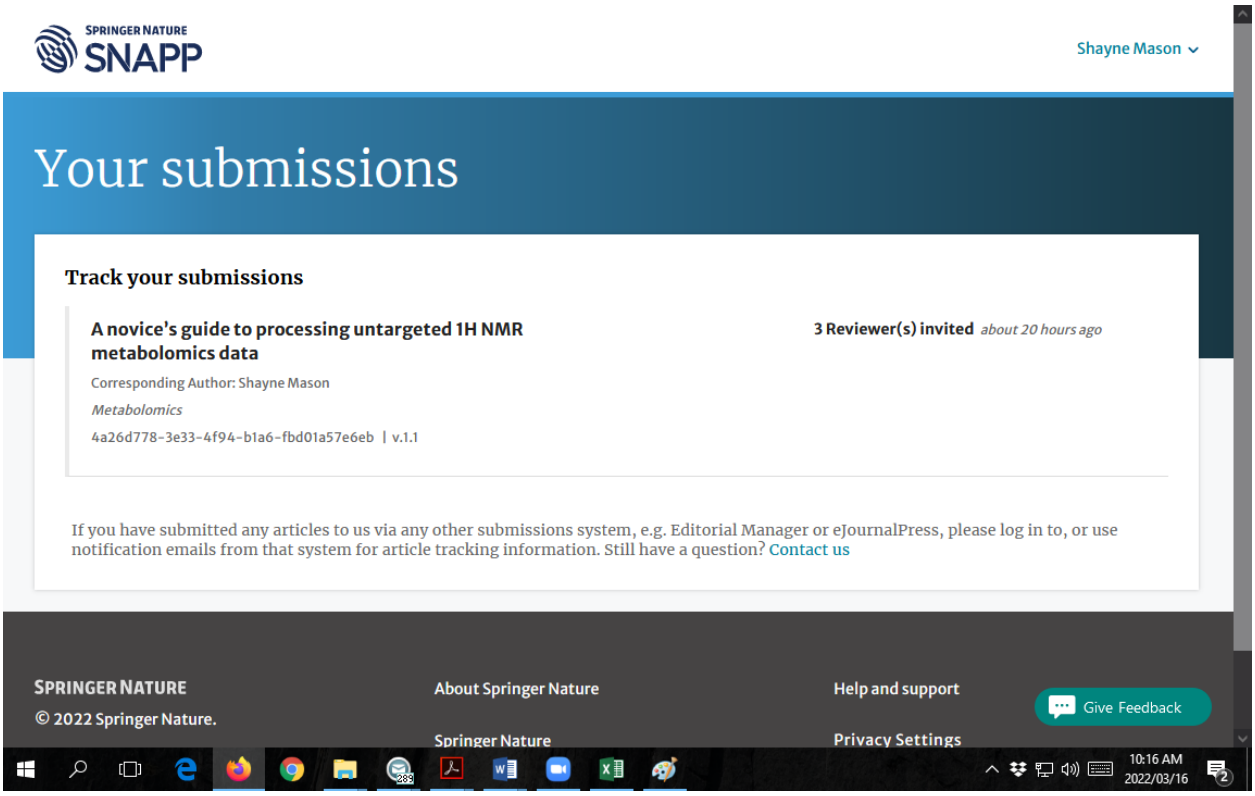
Table S4. The list of metabolites that were identified by both methods for the W0 vs W2 groups, but were not found to be significant

			Rule for significance					
			|logFC| ≥ 1		t-test FDR p-value ≤ 0.05		PLSDA VIP ≥ 1 for LV 1 or 2	
	Confidence in ID							
Compound	**Auto***	**Q. Man**	**Auto***	**Q. Man**	**Auto***	**Q. Man**	**Auto***	**Q. Man**
Isoleucine	7	10	×	×	×	×	×	×
3-Hydroxyisovaleric acid	6	10	×	×	×	×	×	×
N-AcetylX	N/A	10		×		×		×
Methyl guanidine	N/A	10		×		×		×
Trimethylamine	–	10	×	×	×	×	×	×
Methanol	10	10	×	×	×	×	×	×
Sucrose	N/A	10		×		×		×
Anhydro glucose	N/A	10		×		×		×
Glucose	10	10	×	×	×	×	×	×
Histamine	N/A	10		×		×		×
1-Methylhistidine	5	N/A	×		×		×	
2-Oxoisovalerate	–	N/A	×		×		×	
p-Hydroxyphenyl acetic acid	–	N/A	×		×		×	
3-Hydroxyisobutyrate	9	N/A	×		×		×	
Acetic acid	10	N/A	×		×		×	
Betaine	10	N/A	×		×		×	
Ascorbic acid	–	N/A	×		×		×	
Carnitine	9	N/A	×		×		×	
Creatine	7	N/A	×		×		×	
Dimethylglycine	–	N/A	×		×		×	
Choline	10	N/A	×		×		×	
Ethanol	9	N/A	×		×		×	
Guanidoacetic acid	–	N/A	×		×		×	
Glycerol	9	N/A	×		×		×	
Fumaric acid	–	N/A	×		×		×	

Galactose	–	N/A	×	×	×
Glutamic acid	10	N/A	×	×	×
Ethanolamine	–	N/A	×	×	×
Tyrosine	10	N/A	×	×	×
Proline	10	N/A	×	×	×
Asparagine	3	N/A	×	×	×
Mannose	3	N/A	×	×	×
Fucose	–	N/A	×	×	×
Histidine	7	N/A	×	×	×
Lysine	9	N/A	×	×	×
Alpha-Lactose	–	N/A	×	×	×
Serine	4	N/A	×	×	×
Aspartic acid	9	N/A	×	×	×
Cystine	3	N/A	×	×	×
Methylmalonic acid	–	N/A	×	×	×
2-Oxoglutaric acid	–	N/A	×	×	×
Phenylacetic acid	–	N/A	×	×	×
Myoinositol	9	N/A	×	×	×
Oxalacetic acid	–	N/A	×	×	×
Pyruvic acid	8	N/A	×	×	×
Sorbitol	–	N/A	×	×	×
Xanthine	7	N/A	×	×	×
Urea	8	N/A	×	×	×
Uracil	–	N/A	×	×	×
3-Methylhistidine	–	N/A	×	×	×
Arginine	6	N/A	×	×	×
Cysteine	3	N/A	×	×	×
Gluconic acid	–	N/A	×	×	×
Fructose	7	N/A	×	×	×
Malonic acid	7	N/A	×	×	×
Methionine	10	N/A	×	×	×
Hippuric acid	–	N/A	×	×	×
Isovalerate	–	N/A	×	×	×
4-Hydroxyphenyllactic acid	–	N/A	×	×	×
Isopropanol	10	N/A	×	×	×
Tryptophan	5	N/A	×	×	×
Phosphorylcholine	–	N/A	×	×	×
Acetone	10	N/A	×	×	×
Caffeine	–	N/A	×	×	×
Propylene glycol	10	N/A	×	×	×
1,5-Anhydrosorbitol	9	N/A	×	×	×
Dimethyl sulfone	10	N/A	×	×	×
p-Cresol sulfate	–	N/A	×	×	×
EDTA	10	N/A	×	×	×

Key:	
×	Identified by method, but not significant according to stats
	Not identified by method
–	No confidence score given (assumed to be 0)
Auto*	Automated BAYESIL method
Q. Man	Quantified manual method

ANNEXURE D: A SCREENSHOT FROM THE ONLINE SUBMISSION TRACKING SYSTEM



This screenshot shows the latest update (16 March 2022) from the online submission tracking system, regarding the manuscript included as Chapter 4 of this dissertation, indicating that 3 Reviewers have been invited (i.e., it us under peer-review).

ANNEXURE E: CERTIFICATE FROM THE 27TH SASBMB CONFERENCE AWARDED FOR 2ND PLACE IN THE POSTER PRESENTATION CATEGORY

POSTER PRESENTATION AWARD

THIS CERTIFIES THAT

JESSICA VAN DER WESTHUIZEN

has been awarded the 2nd prize for her Poster Presentation entitled:

A novice's guide towards processing untargeted ^{1}H-NMR metabolomics data

AT THE SASBMB CONFERENCE

23 – 26 January 2022

2022 SASBMB BIOCHEMISTRY MATTERS

Wolf-Dieter Schubert
Conference Chairperson

Paul Abidemi Kappo
SASBMB President

ANNEXURE F: GUIDELINES FOR SUBMISSION TO THE JOURNAL *METABOLOMICS*

Instructions for Authors

Types of Papers

Original Articles

The journal publishes Original Research articles that contextualize their hypothesis and research to a broader global issue, and/or present a novel technological approach to the study of metabolomics. All research paper submissions are supported by sound experimental design, methodology, and data interpretation. Submissions of Original Research articles should be 2,500-4,000 words in length (or a maximum of 7 printed pages), should have a maximum of 6 figures/tables (combined), and data should be made accessible via public metabolomics data repositories. Supplementary information is welcome.

Reviews

There are several types of reviews that Metabolomics aims to publish:

- Critical Reviews which review an area in a critical fashion. These reviews aim to critique objectively an area and in doing so synthesis knowledge giving the reader a feel for the critical aspects in terms of what has been currently achieved and what is needed to move a particular field forward. These are typically 7-15 pp. printed text.
- Tutorial Reviews are used to guide a reader through a particular area, be it analytical in nature, computational, or of a biological/medical focus. These should contain a teaching element aimed at earlier career scientists These are typically 7-15 pp. printed text.
- Mini-reviews are highly focused and report on a current hot topic of importance within metabolomics. These should aim to review and critique a few key papers in the field. These are a maximum of 4 pp. printed text.
- Opinion Pieces like mini-reviews are focused on a particular area and should review a few key papers in the field. As these are opinions, if there is potential contention then authors of papers that are criticized may be asked for their views first, and these along with the reviews comments may be published. These are a maximum of 4 printed pages.

Short Communications

• This should follow the same style as a full paper (please refer to section above).

• The Abstract contains less that 100 words.

• This is followed by a brief Introduction section.

• Material and Methods should allow work to be reproduced and extensive material should be provided in supporting Supplementary Information.

• The Results and Discussion must be combined.

• References contain no more than 20 references.

• These are a maximum of 4 pages of printed text, with a combined maximum of 2 Figures and/or Tables.

Software/Databases

These highlight recent developments in software and computer programs along with database developments. Submissions should follow the original article format with structured abstracts and within the abstract and paper details of where to access the software and databases. Articles of this type can also be produced in brief formats as short communications.

General Format for all Types of Papers

Format

1. Abstract

Original Papers require a structured abstract of less than 250 words arranged under the following headings:

• Introduction: Describe the topic's mechanisms, history, and/or how it relates to a problem.

• Objectives: Explain the purpose. What does the study try to demonstrate?

• Methods: Briefly describe the experimental design.

• Results: Report your findings.

• Conclusion: Analyze your results and link it back to the purpose.

Reviews also require an abstract. Please adhere to the following structure:

• Background

• Aim of Review

• Key Scientific Concepts of Review

An abstract is not required of Commentaries or Letters to the Editor.

Each heading should be composed of 1-2 sentences. All abstracts must not include equations, tables, reference citations, or graphics.

2. Introduction: briefly introduces the area and provides a road map of what is to come.

3. Headed Sections: a series of these are then used to bring various aspects together

4. Overall Conclusions: should be made to bring the review to a conclusion.

5. Figures and Tables: should be used to break up the text.

A well-used figure or table can be used to synthesis knowledge, depict time lines etc. The data presented in these formats should stand-alone and the author should avoid repeating these within the body of the text.

Manuscript Preparation

Ethical Statements and Journal Specific Requirements:

• Conflicts of Interest Statement

• Author Contributions

• Ethical Statements

• Data Availability

• Software Availability

Author Conflict of Interest Statement

Funding: This study was funded by X (grant number X). (optional - could be left out in case no funding was received)

Conflict of Interest: Author1 declares that he/she has no conflict of interest.

Author2 declares that he/she has no conflict of interest. ...

Author Contribution Statement

Authors must provide a short description of the contributions made by each listed author (please use initials). This will be published in a separate section in front of the Acknowledgments. For example: AM and DB conceived and designed research. AM and BB conducted experiments. GR contributed new reagents or analytical tools. AM and GR analyzed data. AM wrote the manuscript. All authors read and approved the manuscript.

Authorship Policy

Authorship should incorporate and should be restricted to those who have contributed substantially to the work in one or more of the following categories:

• Conceived of or designed study

• Performed research

• Analyzed data

• Contributed new methods or models

• Wrote the paper

Compliance with Ethical Standards

Please included a Compliance with Ethical Standards section before the References as shown in the samples below.

- In case animals were involved:

  "All applicable international, national, and/or institutional guidelines for the care and use of animals were followed."

- And/or in case humans were involved:

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

- Informed consent: "Informed consent was obtained from all individual participants included in the study.
- If articles do not contain studies with human participants or animals by any of the authors, please select one of the following statements:

  "This article does not contain any studies with human and/or animal participants performed by any of the authors."

Data Availability Statement

"The metabolomics and metadata reported in this paper are available via [insert repository and URL] study identifier [insert study identifier/project ID etc]"

Software or Database Availability Statement

"The software developed in this study is available via [inset URL]"

and/or

"The database reported in this study is accessible via [inset URL]"

Data Availability Statement

The Metabolomics strongly encourages open data! If you have not already done so, please make your research data available – for example on the following repositories:

Metabolights https://www.ebi.ac.uk/metabolights/

Metabolnote http://metabolonote.kazusa.or.jp/Main_Page

Metabolomics Workbench http://www.metabolomicsworkbench.org/

Metaspace http://metaspace2020.eu/#/about

MassIVE https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp

MetabolomeXchange http://www.metabolomexchange.org/site/

Global Natural Product Social molecular networking http://gnps.ucsd.edu/

Offprints

25 offprints of each article will be provided free of charge. Additional offprints can be ordered by means of an offprint order form supplied with the proofs.

Manuscript Submission

Manuscript Submission

Submission of a manuscript implies: that the work described has not been published before; that it is not under consideration for publication anywhere else; that its publication has been approved by all co-authors, if any, as well as by the responsible authorities – tacitly or explicitly – at the institute where the work has been carried out. The publisher will not be held legally responsible should there be any claims for compensation.

Permissions

Authors wishing to include figures, tables, or text passages that have already been published elsewhere are required to obtain permission from the copyright owner(s) for both the print and online format and to include evidence that such

permission has been granted when submitting their papers. Any material received without such evidence will be assumed to originate from the authors.

Online Submission

Please follow the hyperlink “Submit manuscript” and upload all of your manuscript files following the instructions given on the screen.

Source Files

Please ensure you provide all relevant editable source files at every submission and revision. Failing to submit a complete set of editable source files will result in your article not being considered for review. For your manuscript text please always submit in common word processing formats such as .docx or LaTeX.

Plagiarism Check with CrossCheck

Plagiarism prevention with CrossCheck

Springer is a participant of CrossCheck, a multi-publisher plagiarism detection initiative to screen published and submitted content for originality. CrossCheck consists of two products: a database of scholarly publications (CrossCheck) and a web-based tool (iThenticate) to check an authored work against that database.

This journal uses the plagiarism tool to detect instances of overlapping and similar text in submitted manuscripts and your manuscript may be screened upon submission for plagiarism against previously published works.

Title Page

Title Page

Please make sure your title page contains the following information.

Title

The title should be concise and informative.

Author information

- The name(s) of the author(s)
- The affiliation(s) of the author(s), i.e. institution, (department), city, (state), country
- A clear indication and an active e-mail address of the corresponding author
- If available, the 16-digit ORCID of the author(s)

If address information is provided with the affiliation(s) it will also be published.

For authors that are (temporarily) unaffiliated we will only capture their city and country of residence, not their e-mail address unless specifically requested.

Abstract

Please provide an abstract of 150 to 250 words. The abstract should not contain any undefined abbreviations or unspecified references.

For life science journals only (when applicable)

- Trial registration number and date of registration for prospectively registered trials
- Trial registration number and date of registration, followed by “retrospectively registered”, for retrospectively registered trials

Keywords

Please provide 4 to 6 keywords which can be used for indexing purposes.

Statements and Declarations

The following statements should be included under the heading "Statements and Declarations" for inclusion in the published paper. Please note that submissions that do not include relevant declarations will be returned as incomplete.

- **Competing Interests:** Authors are required to disclose financial or non-financial interests that are directly or indirectly related to the work submitted for publication. Please refer to “Competing Interests and Funding” below for more information on how to complete this section.

Please see the relevant sections in the submission guidelines for further information as well as various examples of wording. Please revise/customize the sample statements according to your own needs.

Text

Text Formatting

Manuscripts should be submitted in Word.

- Use a normal, plain font (e.g., 10-point Times Roman) for text.
- Use italics for emphasis.
- Use the automatic page numbering function to number the pages.
- Do not use field functions.
- Use tab stops or other commands for indents, not the space bar.
- Use the table function, not spreadsheets, to make tables.
- Use the equation editor or MathType for equations.
- Save your file in docx format (Word 2007 or higher) or doc format (older Word versions).

Manuscripts with mathematical content can also be submitted in LaTeX. We recommend using Springer Nature's LaTeX template.

Headings

Please use the decimal system of headings with no more than three levels.

Abbreviations

Abbreviations should be defined at first mention and used consistently thereafter.

Footnotes

Footnotes can be used to give additional information, which may include the citation of a reference included in the reference list. They should not consist solely of a reference citation, and they should never include the bibliographic details of a reference. They should also not contain any figures or tables.

Footnotes to the text are numbered consecutively; those to tables should be indicated by superscript lower-case letters (or asterisks for significance values

and other statistical data). Footnotes to the title or the authors of the article are not given reference symbols.

Always use footnotes instead of endnotes.

Acknowledgments

Acknowledgments of people, grants, funds, etc. should be placed in a separate section on the title page. The names of funding organizations should be written in full.

Scientific style

- Please always use internationally accepted signs and symbols for units (SI units).
- Nomenclature: Insofar as possible, authors should use systematic names similar to those used by Chemical Abstract Service or IUPAC.
- Genus and species names should be in italics.
- Generic names of drugs and pesticides are preferred; if trade names are used, the generic name should be given at first mention.
- Please use the standard mathematical notation for formulae, symbols, etc.:Italic for single letters that denote mathematical constants, variables, and unknown quantities Roman/upright for numerals, operators, and punctuation, and commonly defined functions or abbreviations, e.g., cos, det, e or exp, lim, log, max, min, sin, tan, d (for derivative) Bold for vectors, tensors, and matrices.

References

Citation

Cite references in the text by name and year in parentheses. Some examples:

- Negotiation research spans many disciplines (Thompson, 1990).
- This result was later contradicted by Becker and Seligman (1996).
- This effect has been widely studied (Abbott, 1991; Barakat et al., 1995; Kelso & Smith, 1998; Medvec et al., 1999).

Authors are encouraged to follow official APA version 7 guidelines on the number of authors included in reference list entries (i.e., include all authors up to 20; for larger groups, give the first 19 names followed by an ellipsis and the final

author's name). However, if authors shorten the author group by using et al., this will be retained.

Reference list

The list of references should only include works that are cited in the text and that have been published or accepted for publication. Personal communications and unpublished works should only be mentioned in the text.

Reference list entries should be alphabetized by the last names of the first author of each work.

Journal names and book titles should be *italicized*.

If available, please always include DOIs as full DOI links in your reference list (e.g. "https://doi.org/abc").

- Journal article Grady, J. S., Her, M., Moreno, G., Perez, C., & Yelinek, J. (2019). Emotions in storybooks: A comparison of storybooks that represent ethnic and racial groups in the United States. *Psychology of Popular Media Culture, 8*(3), 207–217. https://doi.org/10.1037/ppm0000185
- Article by DOI Hong, I., Knox, S., Pryor, L., Mroz, T. M., Graham, J., Shields, M. F., & Reistetter, T. A. (2020). Is referral to home health rehabilitation following inpatient rehabilitation facility associated with 90-day hospital readmission for adult patients with stroke? *American Journal of Physical Medicine & Rehabilitation*. Advance online publication. https://doi.org/10.1097/PHM.0000000000001435
- Book Sapolsky, R. M. (2017). *Behave: The biology of humans at our best and worst*. Penguin Books.
- Book chapter Dillard, J. P. (2020). Currents in the study of persuasion. In M. B. Oliver, A. A. Raney, & J. Bryant (Eds.), *Media effects: Advances in theory and research* (4th ed., pp. 115–129). Routledge.
- Online document Fagan, J. (2019, March 25). *Nursing clinical brain*. OER Commons. Retrieved January 7, 2020, from https://www.oercommons.org/authoring/53029-nursing-clinical-brain/view

EndNote-Style for Metabolomics

- Please find an EndNote citation style that had been customised for the journal at the Link below:

Link (Download zip, 2 kB)

Tables

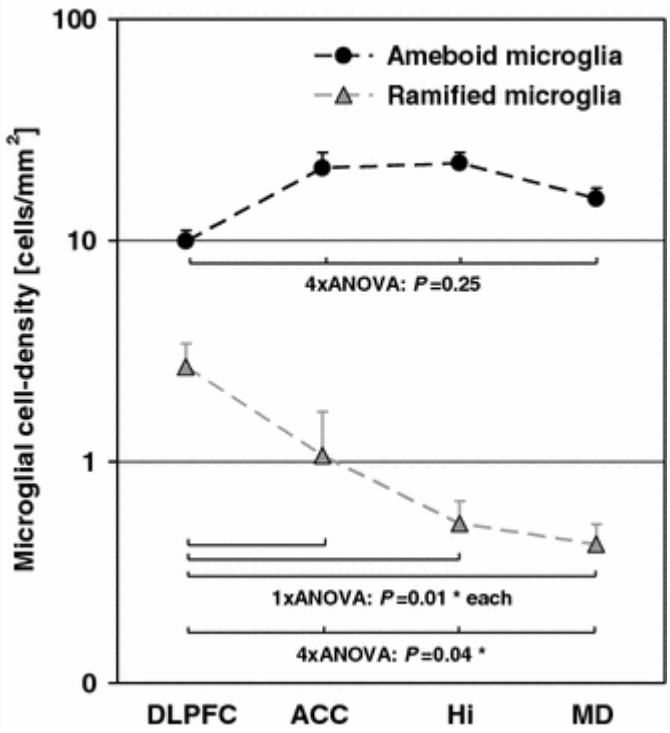
- All tables are to be numbered using Arabic numerals.
- Tables should always be cited in text in consecutive numerical order.
- For each table, please supply a table caption (title) explaining the components of the table.
- Identify any previously published material by giving the original source in the form of a reference at the end of the table caption.
- Footnotes to tables should be indicated by superscript lower-case letters (or asterisks for significance values and other statistical data) and included beneath the table body.

Artwork and Illustrations Guidelines

Electronic Figure Submission

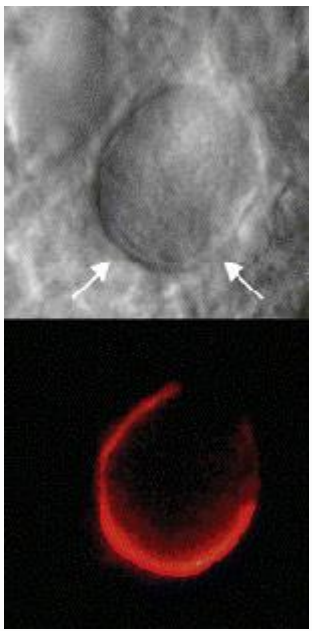
- Supply all figures electronically.
- Indicate what graphics program was used to create the artwork.
- For vector graphics, the preferred format is EPS; for halftones, please use TIFF format. MSOffice files are also acceptable.
- Vector graphics containing fonts must have the fonts embedded in the files.
- Name your figure files with "Fig" and the figure number, e.g., Fig1.eps.

Line Art



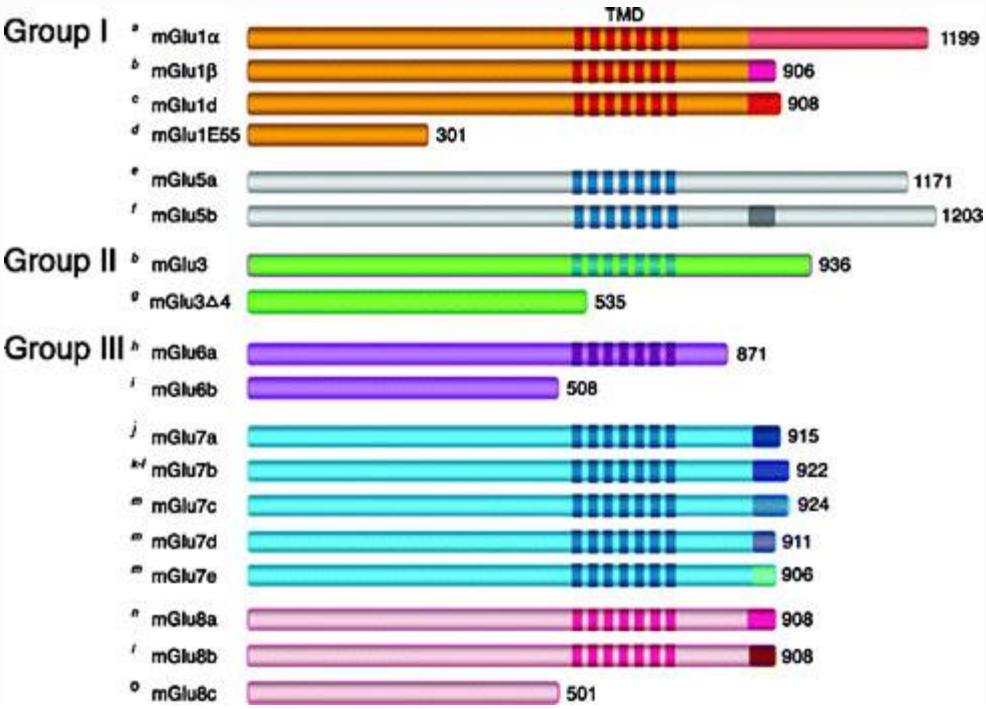
- Definition: Black and white graphic with no shading.
- Do not use faint lines and/or lettering and check that all lines and lettering within the figures are legible at final size.
- All lines should be at least 0.1 mm (0.3 pt) wide.
- Scanned line drawings and line drawings in bitmap format should have a minimum resolution of 1200 dpi.
- Vector graphics containing fonts must have the fonts embedded in the files.

Halftone Art



- Definition: Photographs, drawings, or paintings with fine shading, etc.
- If any magnification is used in the photographs, indicate this by using scale bars within the figures themselves.
- Halftones should have a minimum resolution of 300 dpi.

Combination Art



- Definition: a combination of halftone and line art, e.g., halftones containing line drawing, extensive lettering, color diagrams, etc.
- Combination artwork should have a minimum resolution of 600 dpi.

Color Art

- Color art is free of charge for online publication.
- If black and white will be shown in the print version, make sure that the main information will still be visible. Many colors are not distinguishable from one another when converted to black and white. A simple way to check this is to make a xerographic copy to see if the necessary distinctions between the different colors are still apparent.
- If the figures will be printed in black and white, do not refer to color in the captions.
- Color illustrations should be submitted as RGB (8 bits per channel).

Figure Lettering

- To add lettering, it is best to use Helvetica or Arial (sans serif fonts).
- Keep lettering consistently sized throughout your final-sized artwork, usually about 2–3 mm (8–12 pt).
- Variance of type size within an illustration should be minimal, e.g., do not use 8-pt type on an axis and 20-pt type for the axis label.
- Avoid effects such as shading, outline letters, etc.
- Do not include titles or captions within your illustrations.

Figure Numbering

- All figures are to be numbered using Arabic numerals.
- Figures should always be cited in text in consecutive numerical order.
- Figure parts should be denoted by lowercase letters (a, b, c, etc.).
- If an appendix appears in your article and it contains one or more figures, continue the consecutive numbering of the main text. Do not number the appendix figures,"A1, A2, A3, etc." Figures in online appendices [Supplementary Information (SI)] should, however, be numbered separately.

Figure Captions

- Each figure should have a concise caption describing accurately what the figure depicts. Include the captions in the text file of the manuscript, not in the figure file.
- Figure captions begin with the term Fig. in bold type, followed by the figure number, also in bold type.

- No punctuation is to be included after the number, nor is any punctuation to be placed at the end of the caption.
- Identify all elements found in the figure in the figure caption; and use boxes, circles, etc., as coordinate points in graphs.
- Identify previously published material by giving the original source in the form of a reference citation at the end of the figure caption.

Figure Placement and Size

- Figures should be submitted separately from the text, if possible.
- When preparing your figures, size figures to fit in the column width.
- For large-sized journals the figures should be 84 mm (for double-column text areas), or 174 mm (for single-column text areas) wide and not higher than 234 mm.
- For small-sized journals, the figures should be 119 mm wide and not higher than 195 mm.

Permissions

If you include figures that have already been published elsewhere, you must obtain permission from the copyright owner(s) for both the print and online format. Please be aware that some publishers do not grant electronic rights for free and that Springer will not be able to refund any costs that may have occurred to receive these permissions. In such cases, material from other sources should be used.

Accessibility

In order to give people of all abilities and disabilities access to the content of your figures, please make sure that

- All figures have descriptive captions (blind users could then use a text-to-speech software or a text-to-Braille hardware)
- Patterns are used instead of or in addition to colors for conveying information (colorblind users would then be able to distinguish the visual elements)
- Any figure lettering has a contrast ratio of at least 4.5:1

Supplementary Information (SI)

Springer accepts electronic multimedia files (animations, movies, audio, etc.) and other supplementary files to be published online along with an article or a book chapter. This feature can add dimension to the author's article, as certain information cannot be printed or is more convenient in electronic form.

Before submitting research datasets as Supplementary Information, authors should read the journal's Research data policy. We encourage research data to be archived in data repositories wherever possible.

Submission

- Supply all supplementary material in standard file formats.
- Please include in each file the following information: article title, journal name, author names; affiliation and e-mail address of the corresponding author.
- To accommodate user downloads, please keep in mind that larger-sized files may require very long download times and that some users may experience other problems during downloading.
- High resolution (streamable quality) videos can be submitted up to a maximum of 25GB; low resolution videos should not be larger than 5GB.

Audio, Video, and Animations

- Aspect ratio: 16:9 or 4:3
- Maximum file size: 25 GB for high resolution files; 5 GB for low resolution files
- Minimum video duration: 1 sec
- Supported file formats: avi, wmv, mp4, mov, m2p, mp2, mpg, mpeg, flv, mxf, mts, m4v, 3gp

Text and Presentations

- Submit your material in PDF format; .doc or .ppt files are not suitable for long-term viability.
- A collection of figures may also be combined in a PDF file.

Spreadsheets

- Spreadsheets should be submitted as .csv or .xlsx files (MS Excel).

Specialized Formats

- Specialized format such as .pdb (chemical), .wrl (VRML), .nb (Mathematica notebook), and .tex can also be supplied.

Collecting Multiple Files

- It is possible to collect multiple files in a .zip or .gz file.

Numbering

- If supplying any supplementary material, the text must make specific mention of the material as a citation, similar to that of figures and tables.
- Refer to the supplementary files as “Online Resource”, e.g., "... as shown in the animation (Online Resource 3)", “... additional data are given in Online Resource 4”.
- Name the files consecutively, e.g. “ESM_3.mpg”, “ESM_4.pdf”.

Captions

- For each supplementary material, please supply a concise caption describing the content of the file.

Processing of supplementary files

- Supplementary Information (SI) will be published as received from the author without any conversion, editing, or reformatting.

Accessibility

In order to give people of all abilities and disabilities access to the content of your supplementary files, please make sure that

- The manuscript contains a descriptive caption for each supplementary material
- Video files do not contain anything that flashes more than three times per second (so that users prone to seizures caused by such effects are not put at risk)

Ethical Responsibilities of Authors

This journal is committed to upholding the integrity of the scientific record. As a member of the Committee on Publication Ethics (COPE) the journal will follow the COPE guidelines on how to deal with potential acts of misconduct.

Authors should refrain from misrepresenting research results which could damage the trust in the journal, the professionalism of scientific authorship, and ultimately the entire scientific endeavour. Maintaining integrity of the research and its presentation is helped by following the rules of good scientific practice, which include*:

- The manuscript should not be submitted to more than one journal for simultaneous consideration.
- The submitted work should be original and should not have been published elsewhere in any form or language (partially or in full), unless the new work concerns an expansion of previous work. (Please provide transparency on the re-use of material to avoid the concerns about text-recycling ('self-plagiarism').
- A single study should not be split up into several parts to increase the quantity of submissions and submitted to various journals or to one journal over time (i.e. 'salami-slicing/publishing').
- Concurrent or secondary publication is sometimes justifiable, provided certain conditions are met. Examples include: translations or a manuscript that is intended for a different group of readers.
- Results should be presented clearly, honestly, and without fabrication, falsification or inappropriate data manipulation (including image based manipulation). Authors should adhere to discipline-specific rules for acquiring, selecting and processing data.
- No data, text, or theories by others are presented as if they were the author's own ('plagiarism'). Proper acknowledgements to other works must be given (this includes material that is closely copied (near verbatim), summarized and/or paraphrased), quotation marks (to indicate words taken from another source) are used for verbatim copying of material, and permissions secured for material that is copyrighted.

Important note: the journal may use software to screen for plagiarism.

- Authors should make sure they have permissions for the use of software, questionnaires/(web) surveys and scales in their studies (if appropriate).

- Research articles and non-research articles (e.g. Opinion, Review, and Commentary articles) must cite appropriate and relevant literature in support of the claims made. Excessive and inappropriate self-citation or coordinated efforts among several authors to collectively self-cite is strongly discouraged.
- Authors should avoid untrue statements about an entity (who can be an individual person or a company) or descriptions of their behavior or actions that could potentially be seen as personal attacks or allegations about that person.
- Research that may be misapplied to pose a threat to public health or national security should be clearly identified in the manuscript (e.g. dual use of research). Examples include creation of harmful consequences of biological agents or toxins, disruption of immunity of vaccines, unusual hazards in the use of chemicals, weaponization of research/technology (amongst others).
- Authors are strongly advised to ensure the author group, the Corresponding Author, and the order of authors are all correct at submission. Adding and/or deleting authors during the revision stages is generally not permitted, but in some cases may be warranted. Reasons for changes in authorship should be explained in detail. Please note that changes to authorship cannot be made after acceptance of a manuscript.

*All of the above are guidelines and authors need to make sure to respect third parties rights such as copyright and/or moral rights.

Upon request authors should be prepared to send relevant documentation or data in order to verify the validity of the results presented. This could be in the form of raw data, samples, records, etc. Sensitive information in the form of confidential or proprietary data is excluded.

If there is suspicion of misbehavior or alleged fraud the Journal and/or Publisher will carry out an investigation following COPE guidelines. If, after investigation, there are valid concerns, the author(s) concerned will be contacted under their given e-mail address and given an opportunity to address the issue. Depending on the situation, this may result in the Journal's and/or Publisher's implementation of the following measures, including, but not limited to:

- If the manuscript is still under consideration, it may be rejected and returned to the author.
- If the article has already been published online, depending on the nature and severity of the infraction:

  - an erratum/correction may be placed with the article

  - an expression of concern may be placed with the article

  - or in severe cases retraction of the article may occur.

The reason will be given in the published erratum/correction, expression of concern or retraction note. Please note that retraction means that the article is **maintained on the platform**, watermarked "retracted" and the explanation for the retraction is provided in a note linked to the watermarked article.

- The author's institution may be informed
- A notice of suspected transgression of ethical standards in the peer review system may be included as part of the author's and article's bibliographic record.

Fundamental errors

Authors have an obligation to correct mistakes once they discover a significant error or inaccuracy in their published article. The author(s) is/are requested to contact the journal and explain in what sense the error is impacting the article. A decision on how to correct the literature will depend on the nature of the error. This may be a correction or retraction. The retraction note should provide transparency which parts of the article are impacted by the error.

Suggesting / excluding reviewers

Authors are welcome to suggest suitable reviewers and/or request the exclusion of certain individuals when they submit their manuscripts. When suggesting reviewers, authors should make sure they are totally independent and not connected to the work in any way. It is strongly recommended to suggest a mix of reviewers from different countries and different institutions. When suggesting reviewers, the Corresponding Author must provide an institutional email address for each suggested reviewer, or, if this is not possible to include other means of verifying the identity such as a link to a personal homepage, a link to the publication record or a researcher or author ID in the submission letter.

Please note that the Journal may not use the suggestions, but suggestions are appreciated and may help facilitate the peer review process.

Authorship principles

These guidelines describe authorship principles and good authorship practices to which prospective authors should adhere to.

Authorship clarified

The Journal and Publisher assume all authors agreed with the content and that all gave explicit consent to submit and that they obtained consent from the responsible authorities at the institute/organization where the work has been carried out, **before** the work is submitted.

The Publisher does not prescribe the kinds of contributions that warrant authorship. It is recommended that authors adhere to the guidelines for authorship that are applicable in their specific research field. In absence of specific guidelines it is recommended to adhere to the following guidelines*:

All authors whose names appear on the submission

1) made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work;

2) drafted the work or revised it critically for important intellectual content;

3) approved the version to be published; and

4) agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

* Based on/adapted from:

ICMJE, Defining the Role of Authors and Contributors,

Transparency in authors' contributions and responsibilities to promote integrity in scientific publication, McNutt at all, PNAS February 27, 2018

Disclosures and declarations

All authors are requested to include information regarding sources of funding, financial or non-financial interests, study-specific approval by the appropriate ethics committee for research involving humans and/or animals, informed consent if the research involved human participants, and a statement on welfare of animals if the research involved animals (as appropriate).

The decision whether such information should be included is not only dependent on the scope of the journal, but also the scope of the article. Work submitted for publication may have implications for public health or general welfare and in those cases it is the responsibility of all authors to include the appropriate disclosures and declarations.

Data transparency

All authors are requested to make sure that all data and materials as well as software application or custom code support their published claims and comply with field standards. Please note that journals may have individual policies on (sharing) research data in concordance with disciplinary norms and expectations.

Role of the Corresponding Author

One author is assigned as Corresponding Author and acts on behalf of all co-authors and ensures that questions related to the accuracy or integrity of any part of the work are appropriately addressed.

The Corresponding Author is responsible for the following requirements:

- ensuring that all listed authors have approved the manuscript before submission, including the names and order of authors;
- managing all communication between the Journal and all co-authors, before and after publication;*
- providing transparency on re-use of material and mention any un-published material (for example manuscripts in press) included in the manuscript in a cover letter to the Editor;

- making sure disclosures, declarations and transparency on data statements from all authors are included in the manuscript as appropriate (see above).

* The requirement of managing all communication between the journal and all co-authors during submission and proofing may be delegated to a Contact or Submitting Author. In this case please make sure the Corresponding Author is clearly indicated in the manuscript.

Author contributions

In absence of specific instructions and in research fields where it is possible to describe discrete efforts, the Publisher recommends authors to include contribution statements in the work that specifies the contribution of every author in order to promote transparency. These contributions should be listed at the separate title page.

Examples of such statement(s) are shown below:

• Free text:

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [full name], [full name] and [full name]. The first draft of the manuscript was written by [full name] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Example: CRediT taxonomy:

• Conceptualization: [full name], …; Methodology: [full name], …; Formal analysis and investigation: [full name], …; Writing - original draft preparation: [full name], …]; Writing - review and editing: [full name], …; Funding acquisition: [full name], …; Resources: [full name], …; Supervision: [full name],….

For **review articles** where discrete statements are less applicable a statement should be included who had the idea for the article, who performed the literature search and data analysis, and who drafted and/or critically revised the work.

For articles that are based primarily on the **student's dissertation or thesis**, it is recommended that the student is usually listed as principal author:

A Graduate Student's Guide to Determining Authorship Credit and Authorship Order, APA Science Student Council 2006

Affiliation

The primary affiliation for each author should be the institution where the majority of their work was done. If an author has subsequently moved, the current address may additionally be stated. Addresses will not be updated or changed after publication of the article.

Changes to authorship

Authors are strongly advised to ensure the correct author group, the Corresponding Author, and the order of authors at submission. Changes of authorship by adding or deleting authors, and/or changes in Corresponding Author, and/or changes in the sequence of authors are **not** accepted **after acceptance** of a manuscript.

- **Please note that author names will be published exactly as they appear on the accepted submission!**

Please make sure that the names of all authors are present and correctly spelled, and that addresses and affiliations are current.

Adding and/or deleting authors at revision stage are generally not permitted, but in some cases it may be warranted. Reasons for these changes in authorship should be explained. Approval of the change during revision is at the discretion of the Editor-in-Chief. Please note that journals may have individual policies on adding and/or deleting authors during revision stage.

Author identification

Authors are recommended to use their ORCID ID when submitting an article for consideration or acquire an ORCID ID via the submission process.

Deceased or incapacitated authors

For cases in which a co-author dies or is incapacitated during the writing, submission, or peer-review process, and the co-authors feel it is appropriate to include the author, co-authors should obtain approval from a (legal) representative which could be a direct relative.

Authorship issues or disputes

In the case of an authorship dispute during peer review or after acceptance and publication, the Journal will not be in a position to investigate or adjudicate. Authors will be asked to resolve the dispute themselves. If they are unable the Journal reserves the right to withdraw a manuscript from the editorial process or in case of a published paper raise the issue with the authors' institution(s) and abide by its guidelines.

Confidentiality

Authors should treat all communication with the Journal as confidential which includes correspondence with direct representatives from the Journal such as Editors-in-Chief and/or Handling Editors and reviewers' reports unless explicit consent has been received to share information.

Compliance with Ethical Standards

To ensure objectivity and transparency in research and to ensure that accepted principles of ethical and professional conduct have been followed, authors should include information regarding sources of funding, potential conflicts of interest (financial or non-financial), informed consent if the research involved human participants, and a statement on welfare of animals if the research involved animals.

Authors should include the following statements (if applicable) in a separate section entitled "Compliance with Ethical Standards" when submitting a paper:

- Disclosure of potential conflicts of interest
- Research involving Human Participants and/or Animals
- Informed consent

Please note that standards could vary slightly per journal dependent on their peer review policies (i.e. single or double blind peer review) as well as per journal

subject discipline. Before submitting your article check the instructions following this section carefully.

The corresponding author should be prepared to collect documentation of compliance with ethical standards and send if requested during peer review or after publication.

The Editors reserve the right to reject manuscripts that do not comply with the above-mentioned guidelines. The author will be held responsible for false statements or failure to fulfill the above-mentioned guidelines.

Competing Interests

Authors are requested to disclose interests that are directly or indirectly related to the work submitted for publication. Interests within the last 3 years of beginning the work (conducting the research and preparing the work for submission) should be reported. Interests outside the 3-year time frame must be disclosed if they could reasonably be perceived as influencing the submitted work. Disclosure of interests provides a complete and transparent process and helps readers form their own judgments of potential bias. This is not meant to imply that a financial relationship with an organization that sponsored the research or compensation received for consultancy work is inappropriate.

Editorial Board Members and Editors are required to declare any competing interests and may be excluded from the peer review process if a competing interest exists. In addition, they should exclude themselves from handling manuscripts in cases where there is a competing interest. This may include – but is not limited to – having previously published with one or more of the authors, and sharing the same institution as one or more of the authors. Where an Editor or Editorial Board Member is on the author list they must declare this in the competing interests section on the submitted manuscript. If they are an author or have any other competing interest regarding a specific manuscript, another Editor or member of the Editorial Board will be assigned to assume responsibility for overseeing peer review. These submissions are subject to the exact same review process as any other manuscript. Editorial Board Members are welcome to submit papers to the journal. These submissions are not given any priority over other manuscripts, and Editorial Board Member status has no bearing on editorial consideration.

Interests that should be considered and disclosed but are not limited to the following:

Funding: Research grants from funding agencies (please give the research funder and the grant number) and/or research support (including salaries, equipment, supplies, reimbursement for attending symposia, and other expenses) by organizations that may gain or lose financially through publication of this manuscript.

Employment: Recent (while engaged in the research project), present or anticipated employment by any organization that may gain or lose financially through publication of this manuscript. This includes multiple affiliations (if applicable).

Financial interests: Stocks or shares in companies (including holdings of spouse and/or children) that may gain or lose financially through publication of this manuscript; consultation fees or other forms of remuneration from organizations that may gain or lose financially; patents or patent applications whose value may be affected by publication of this manuscript.

It is difficult to specify a threshold at which a financial interest becomes significant, any such figure is necessarily arbitrary, so one possible practical guideline is the following: "Any undeclared financial interest that could embarrass the author were it to become publicly known after the work was published."

Non-financial interests: In addition, authors are requested to disclose interests that go beyond financial interests that could impart bias on the work submitted for publication such as professional interests, personal relationships or personal beliefs (amongst others). Examples include, but are not limited to: position on editorial board, advisory board or board of directors or other type of management relationships; writing and/or consulting for educational purposes; expert witness; mentoring relations; and so forth.

Primary research articles require a disclosure statement. Review articles present an expert synthesis of evidence and may be treated as an authoritative work on a subject. Review articles therefore require a disclosure statement.Other article types such as editorials, book reviews, comments (amongst others) may, dependent on their content, require a disclosure statement. If you are unclear

whether your article type requires a disclosure statement, please contact the Editor-in-Chief.

Please note that, in addition to the above requirements, funding information (given that funding is a potential competing interest (as mentioned above)) needs to be disclosed upon submission of the manuscript in the peer review system. This information will automatically be added to the Record of Cross-Mark, however it is **not added** to the manuscript itself. Under 'summary of requirements' (see below) funding information should be included in the '**Declarations**' section.

Summary of requirements

The above should be summarized in a statement and placed in a 'Declarations' section before the reference list under a heading of 'Funding' and/or 'Competing interests'. Other declarations include Ethics approval, Consent, Data, Material and/or Code availability and Authors' contribution statements.

Please see the various examples of wording below and revise/customize the sample statements according to your own needs.

When all authors have the same (or no) conflicts and/or funding it is sufficient to use one blanket statement.

Examples of statements to be used when funding has been received:

- Partial financial support was received from [...]
- The research leading to these results received funding from [...] under Grant Agreement No[...].
- This study was funded by [...]
- This work was supported by [...] (Grant numbers [...] and [...]

Examples of statements to be used when there is no funding:

- The authors did not receive support from any organization for the submitted work.
- No funding was received to assist with the preparation of this manuscript.
- No funding was received for conducting this study.

- No funds, grants, or other support was received.

Examples of statements to be used when there are interests to declare:

- **Financial interests:** Author A has received research support from Company A. Author B has received a speaker honorarium from Company Wand owns stock in Company X. Author C is consultant to company Y.

  Non-financial interests: Author C is an unpaid member of committee Z.

- **Financial interests:** The authors declare they have no financial interests.

  Non-financial interests: Author A is on the board of directors of Y and receives no compensation as member of the board of directors.

- **Financial interests:** Author A received a speaking fee from Y for Z. Author B receives a salary from association X. X where s/he is the Executive Director.

  Non-financial interests: none.

- **Financial interests:** Author A and B declare they have no financial interests. Author C has received speaker and consultant honoraria from Company M and Company N. Dr. C has received speaker honorarium and research funding from Company M and Company O. Author D has received travel support from Company O.

  Non-financial interests: Author D has served on advisory boards for Company M, Company N and Company O.

Examples of statements to be used when authors have nothing to declare:

- The authors have no relevant financial or non-financial interests to disclose.
- The authors have no competing interests to declare that are relevant to the content of this article.
- All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
- The authors have no financial or proprietary interests in any material discussed in this article.

Authors are responsible for correctness of the statements provided in the manuscript. See also Authorship Principles. The Editor-in-Chief reserves the right to reject submissions that do not meet the guidelines described in this section.

Research involving human participants, their data or biological material

Ethics approval

When reporting a study that involved human participants, their data or biological material, authors should include a statement that confirms that the study was approved (or granted exemption) by the appropriate institutional and/or national research ethics committee (including the name of the ethics committee) and certify that the study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. If doubt exists whether the research was conducted in accordance with the 1964 Helsinki Declaration or comparable standards, the authors must explain the reasons for their approach, and demonstrate that an independent ethics committee or institutional review board explicitly approved the doubtful aspects of the study. If a study was granted exemption from requiring ethics approval, this should also be detailed in the manuscript (including the reasons for the exemption).

Retrospective ethics approval

If a study has not been granted ethics committee approval prior to commencing, retrospective ethics approval usually cannot be obtained and it may not be possible to consider the manuscript for peer review. The decision on whether to proceed to peer review in such cases is at the Editor's discretion.

Ethics approval for retrospective studies

Although retrospective studies are conducted on already available data or biological material (for which formal consent may not be needed or is difficult to obtain) ethics approval may be required dependent on the law and the national ethical guidelines of a country. Authors should check with their institution to make sure they are complying with the specific requirements of their country.

Ethics approval for case studies

Case reports require ethics approval. Most institutions will have specific policies on this subject. Authors should check with their institution to make sure they

are complying with the specific requirements of their institution and seek ethics approval where needed. Authors should be aware to secure informed consent from the individual (or parent or guardian if the participant is a minor or incapable) See also section on **Informed Consent**.

Cell lines

If human cells are used, authors must declare in the manuscript: what cell lines were used by describing the source of the cell line, including when and from where it was obtained, whether the cell line has recently been authenticated and by what method. If cells were bought from a life science company the following need to be given in the manuscript: name of company (that provided the cells), cell type, number of cell line, and batch of cells.

It is recommended that authors check the NCBI database for misidentification and contamination of human cell lines. This step will alert authors to possible problems with the cell line and may save considerable time and effort.

Further information is available from the International Cell Line Authentication Committee (ICLAC).

Authors should include a statement that confirms that an institutional or independent ethics committee (including the name of the ethics committee) approved the study and that informed consent was obtained from the donor or next of kin.

Research Resource Identifiers (RRID)

Research Resource Identifiers (RRID) are persistent unique identifiers (effectively similar to a DOI) for research resources. This journal encourages authors to adopt RRIDs when reporting key biological resources (antibodies, cell lines, model organisms and tools) in their manuscripts.

Examples:

Organism: *Filip1*$^{tm1a(KOMP)Wtsi}$ **RRID:MMRRC_055641-UCD**

Cell Line: RST307 cell line **RRID:CVCL_C321**

Antibody: Luciferase antibody DSHB Cat# LUC-3, **RRID:AB_2722109**

Plasmid: mRuby3 plasmid **RRID:Addgene_104005**

Software: ImageJ Version 1.2.4 **RRID:SCR_003070**

RRIDs are provided by the Resource Identification Portal. Many commonly used research resources already have designated RRIDs. The portal also provides authors links so that they can quickly register a new resource and obtain an RRID.

Clinical Trial Registration

The World Health Organization (WHO) definition of a clinical trial is "any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes". The WHO defines health interventions as “A health intervention is an act performed for, with or on behalf of a person or population whose purpose is to assess, improve, maintain, promote or modify health, functioning or health conditions” and a health-related outcome is generally defined as a change in the health of a person or population as a result of an intervention.

To ensure the integrity of the reporting of patient-centered trials, authors must register prospective clinical trials (phase II to IV trials) in suitable publicly available repositories. For example www.clinicaltrials.gov or any of the primary registries that participate in the WHO International Clinical Trials Registry Platform.

The trial registration number (TRN) and date of registration should be included as the last line of the manuscript abstract.

For clinical trials that have not been registered prospectively, authors are encouraged to register retrospectively to ensure the complete publication of all results. The trial registration number (TRN), date of registration and the words 'retrospectively registered’ should be included as the last line of the manuscript abstract.

Standards of reporting

Springer Nature advocates complete and transparent reporting of biomedical and biological research and research with biological applications. Authors are

recommended to adhere to the minimum reporting guidelines hosted by the EQUATOR Network when preparing their manuscript.

Exact requirements may vary depending on the journal; please refer to the journal's Instructions for Authors.

Checklists are available for a number of study designs, including:

Randomised trials (CONSORT) and Study protocols (SPIRIT)

Observational studies (STROBE)

Systematic reviews and meta-analyses (PRISMA) and protocols (Prisma-P)

Diagnostic/prognostic studies (STARD) and (TRIPOD)

Case reports (CARE)

Clinical practice guidelines (AGREE) and (RIGHT)

Qualitative research (SRQR) and (COREQ)

Animal pre-clinical studies (ARRIVE)

Quality improvement studies (SQUIRE)

Economic evaluations (CHEERS)

Summary of requirements

The above should be summarized in a statement and placed in a 'Declarations' section before the reference list under a heading of 'Ethics approval'.

Please see the various examples of wording below and revise/customize the sample statements according to your own needs.

Examples of statements to be used when ethics approval has been obtained:

• All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was approved by the Bioethics Committee of the Medical University of A (No. ...).

• This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of University B (Date.../No. ...).

• Approval was obtained from the ethics committee of University C. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

• The questionnaire and methodology for this study was approved by the Human Research Ethics committee of the University of D (Ethics approval number: ...).

Examples of statements to be used for a retrospective study:

• Ethical approval was waived by the local Ethics Committee of University A in view of the retrospective nature of the study and all the procedures being performed were part of the routine care.

• This research study was conducted retrospectively from data obtained for clinical purposes. We consulted extensively with the IRB of XYZ who determined that our study did not need ethical approval. An IRB official waiver of ethical approval was granted from the IRB of XYZ.

• This retrospective chart review study involving human participants was in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The Human Investigation Committee (IRB) of University B approved this study.

Examples of statements to be used when no ethical approval is required/exemption granted:

• This is an observational study. The XYZ Research Ethics Committee has confirmed that no ethical approval is required.

• The data reproduced from Article X utilized human tissue that was procured via our Biobank AB, which provides de-identified samples. This study was reviewed and deemed exempt by our XYZ Institutional Review Board. The BioBank protocols are in accordance with the ethical standards of our institution and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Authors are responsible for correctness of the statements provided in the manuscript. See also Authorship Principles. The Editor-in-Chief reserves the right to reject submissions that do not meet the guidelines described in this section.

Informed consent

All individuals have individual rights that are not to be infringed. Individual participants in studies have, for example, the right to decide what happens to the (identifiable) personal data gathered, to what they have said during a study or an interview, as well as to any photograph that was taken. This is especially true concerning images of vulnerable people (e.g. minors, patients, refugees, etc) or the use of images in sensitive contexts. In many instances authors will need to secure written consent before including images.

Identifying details (names, dates of birth, identity numbers, biometrical characteristics (such as facial features, fingerprint, writing style, voice pattern, DNA or other distinguishing characteristic) and other information) of the participants that were studied should not be published in written descriptions, photographs, and genetic profiles unless the information is essential for scholarly purposes and the participant (or parent/guardian if the participant is a minor or incapable or legal representative) gave written informed consent for publication. Complete anonymity is difficult to achieve in some cases. Detailed descriptions of individual participants, whether of their whole bodies or of body sections, may lead to disclosure of their identity. Under certain circumstances consent is not required as long as information is anonymized and the submission does not include images that may identify the person.

Informed consent for publication should be obtained if there is any doubt. For example, masking the eye region in photographs of participants is inadequate

protection of anonymity. If identifying characteristics are altered to protect anonymity, such as in genetic profiles, authors should provide assurance that alterations do not distort meaning.

Exceptions where it is not necessary to obtain consent:

• Images such as x rays, laparoscopic images, ultrasound images, brain scans, pathology slides unless there is a concern about identifying information in which case, authors should ensure that consent is obtained.

• Reuse of images: If images are being reused from prior publications, the Publisher will assume that the prior publication obtained the relevant information regarding consent. Authors should provide the appropriate attribution for republished images.

Consent and already available data and/or biologic material

Regardless of whether material is collected from living or dead patients, they (family or guardian if the deceased has not made a pre-mortem decision) must have given prior written consent. The aspect of confidentiality as well as any wishes from the deceased should be respected.

Data protection, confidentiality and privacy

When biological material is donated for or data is generated as part of a research project authors should ensure, as part of the informed consent procedure, that the participants are made aware what kind of (personal) data will be processed, how it will be used and for what purpose. In case of data acquired via a biobank/biorepository, it is possible they apply a broad consent which allows research participants to consent to a broad range of uses of their data and samples which is regarded by research ethics committees as specific enough to be considered "informed". However, authors should always check the specific biobank/biorepository policies or any other type of data provider policies (in case of non-bio research) to be sure that this is the case.

Consent to Participate

For all research involving human subjects, freely-given, informed consent to participate in the study must be obtained from participants (or their parent or

legal guardian in the case of children under 16) and a statement to this effect should appear in the manuscript. In the case of articles describing human transplantation studies, authors must include a statement declaring that no organs/tissues were obtained from prisoners and must also name the institution(s)/clinic(s)/department(s) via which organs/tissues were obtained. For manuscripts reporting studies involving vulnerable groups where there is the potential for coercion or where consent may not have been fully informed, extra care will be taken by the editor and may be referred to the Springer Nature Research Integrity Group.

Consent to Publish

Individuals may consent to participate in a study, but object to having their data published in a journal article. Authors should make sure to also seek consent from individuals to publish their data prior to submitting their paper to a journal. This is in particular applicable to case studies.

Summary of requirements

The above should be summarized in a statement and placed in a 'Declarations' section before the reference list under a heading of 'Consent to participate' and/or 'Consent to publish'. Other declarations include Funding, Competing interests, Ethics approval, Consent, Data and/or Code availability and Authors' contribution statements.

Please see the various examples of wording below and revise/customize the sample statements according to your own needs.

Sample statements for **"Consent to participate"**:

Informed consent was obtained from all individual participants included in the study.

Informed consent was obtained from legal guardians.

Written informed consent was obtained from the parents.

Verbal informed consent was obtained prior to the interview.

Sample statements for **"Consent to publish"**:

The authors affirm that human research participants provided informed consent for publication of the images in Figure(s) 1a, 1b and 1c.

The participant has consented to the submission of the case report to the journal.

Patients signed informed consent regarding publishing their data and photographs.

Sample statements if identifying information about participants is available in the article:

Additional informed consent was obtained from all individual participants for whom identifying information is included in this article.

Authors are responsible for correctness of the statements provided in the manuscript. See also Authorship Principles. The Editor-in-Chief reserves the right to reject submissions that do not meet the guidelines described in this section.

Images will be removed from publication if authors have not obtained informed consent or the paper may be removed and replaced with a notice explaining the reason for removal.

Research Data Policy and Data Availability Statements

This journal operates a type 2 research data policy (life sciences). A submission to the journal implies that materials described in the manuscript, including all relevant raw data, will be freely available to any researcher wishing to use them for non-commercial purposes, without breaching participant confidentiality.

The journal strongly encourages that all datasets on which the conclusions of the paper rely should be available to readers. We encourage authors to ensure that their datasets are either deposited in publicly available repositories (where available and appropriate) or presented in the main manuscript or additional supporting files whenever possible. Please see Springer Nature's information on recommended repositories.

List of Repositories

Research Data Policy

General repositories - for all types of research data - such as figshare and Dryad may be used where appropriate.

Datasets that are assigned digital object identifiers (DOIs) by a data repository may be cited in the reference list. Data citations should include the minimum information recommended by DataCite: authors, title, publisher (repository name), identifier.

DataCite

Where a widely established research community expectation for data archiving in public repositories exists, submission to a community-endorsed, public repository is mandatory. Persistent identifiers (such as DOIs and accession numbers) for relevant datasets must be provided in the paper.

If the journal that you're submitting to uses double-blind peer review and you are providing reviewers with access to your data (for example via a repository link, supplementary information or data on request), it is strongly suggested that the authorship in the data is also blinded. There are data repositories that can assist with this and/or will create a link to mask the authorship of your data.

For the following types of data set, submission to a community-endorsed, public repository is mandatory:

Mandatory deposition	Suitable repositories
Protein sequences	Uniprot
DNA and RNA sequences	Genbank DNA DataBank of Japan (DDBJ) EMBL Nucleotide Sequence Database (ENA)
DNA and RNA sequencing data	NCBI Trace Archive NCBI Sequence Read Archive (SRA)

Genetic polymorphisms	dbSNP dbVar European Variation Archive (EVA)
Linked genotype and phenotype data	dbGAP The European Genome-phenome Archive (EGA)
Macromolecular structure	Worldwide Protein Data Bank (wwPDB) Biological Magnetic Resonance Data Bank (BMRB) Electron Microscopy Data Bank (EMDB)
Microarray data (must be MIAME compliant)	Gene Expression Omnibus (GEO) ArrayExpress
Crystallographic data for small molecules	Cambridge Structural Database

For more information:

Research Data Policy Frequently Asked Questions

Data availability

The journal encourages authors to provide a statement of Data availability in their article. Data availability statements should include information on where data supporting the results reported in the article can be found, including, where applicable, hyperlinks to publicly archived datasets analysed or generated during the study. Data availability statements can also indicate whether data are available on request from the authors and where no data are available, if appropriate.

Data Availability statements can take one of the following forms (or a combination of more than one if required for multiple datasets):

- 1. The datasets generated during and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS]

- 2. The datasets generated during and/or analysed during the current study are not publicly available due [REASON WHY DATA ARE NOT PUBLIC] but are available from the corresponding author on reasonable request.
- 3. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.
- 4. Data sharing not applicable to this article as no datasets were generated or analysed during the current study.
- 5. All data generated or analysed during this study are included in this published article [and its supplementary information files].

More examples of template data availability statements, which include examples of openly available and restricted access datasets, are available:

Data availability statements

Authors who need help understanding our data sharing policies, help finding a suitable data repository, or help organising and sharing research data can access our Author Support portal for additional guidance.

Please note:

Metabolomics data should be submitted following the MSI guidelines.

Preferred Metabolomics data repositories are MassIVE and MetaboLights

After acceptance

Upon acceptance, your article will be exported to Production to undergo typesetting. Once typesetting is complete, you will receive a link asking you to confirm your affiliation, choose the publishing model for your article as well as arrange rights and payment of any associated publication cost.

Once you have completed this, your article will be processed and you will receive the proofs.

Article publishing agreement

Depending on the ownership of the journal and its policies, you will either grant the Publisher an exclusive licence to publish the article or will be asked to transfer copyright of the article to the Publisher.

Offprints

Offprints can be ordered by the corresponding author.

Color illustrations

Online publication of color illustrations is free of charge. For color in the print version, authors will be expected to make a contribution towards the extra costs.

Proof reading

The purpose of the proof is to check for typesetting or conversion errors and the completeness and accuracy of the text, tables and figures. Substantial changes in content, e.g., new results, corrected values, title and authorship, are not allowed without the approval of the Editor.

After online publication, further changes can only be made in the form of an Erratum, which will be hyperlinked to the article.

Online First

The article will be published online after receipt of the corrected proofs. This is the official first publication citable with the DOI. After release of the printed version, the paper can also be cited by issue and page numbers.

Open Choice

Open Choice allows you to publish open access in more than 1850 Springer Nature journals, making your research more visible and accessible immediately on publication.

Article processing charges (APCs) vary by journal – view the full list

Benefits:

- Increased researcher engagement: Open Choice enables access by anyone with an internet connection, immediately on publication.

- Higher visibility and impact: In Springer hybrid journals, OA articles are accessed 4 times more often on average, and cited 1.7 more times on average*.
- Easy compliance with funder and institutional mandates: Many funders require open access publishing, and some take compliance into account when assessing future grant applications.

It is easy to find funding to support open access – please see our funding and support pages for more information.

*) Within the first three years of publication. Springer Nature hybrid journal OA impact analysis, 2018.

English Language Editing

For editors and reviewers to accurately assess the work presented in your manuscript you need to ensure the English language is of sufficient quality to be understood. If you need help with writing in English you should consider:

- Getting a fast, free online grammar check.
- Asking a colleague who is proficient in English to review your manuscript for clarity.
- Visiting the English language tutorial which covers the common mistakes when writing in English.
- Using a professional language editing service where editors will improve the English to ensure that your meaning is clear and identify problems that require your review. Two such services are provided by our affiliates Nature Research Editing Service and American Journal Experts. Springer authors are entitled to a 10% discount on their first submission to either of these services, simply follow the links below.

Free online grammar check

English language tutorial

Nature Research Editing Service

American Journal Experts

Please note that the use of a language editing service is not a requirement for publication in this journal and does not imply or guarantee that the article will be selected for peer review or accepted.

If your manuscript is accepted it will be checked by our copyeditors for spelling and formal style before publication.

Open access publishing

To find out more about publishing your work Open Access in *Metabolomics*, including information on fees, funding and licenses, visit our Open access publishing page.