Enhancing untargeted LC-MS metabolomics data analysis with Bioconductor

Vilhelm Karl Mauritz Suksi

41856-403-2017, vsuksi@abo.fi

Bioscience program at Åbo Akademi, biochemistry major

BK00BM49: Master's thesis in biochemistry, 30 credits

Supervisor: Leo Lahti

Co-supervisor: Guillaume Jacquemet

Submitted in total fulfilment of the requirements of the degree of Master of Philosophy

Abstract

Insight in life science is increasingly data-driven, where new data science methodologies need to account for sparse, high-dimensional data from new experimental approaches. Untargeted LC-MS metabolomics data holds promise in teasing apart the quantities of small molecules in biological samples, the metabolome, which especially in conjunction with other omics can provide knowledge of complex, biological systems. However, due to experimental reasons and the extensive data analysis, untargeted LC-MS metabolomics data analysis meets challenges with regards to quality and reproducibility.

Analysis of omics data is well developed in Bioconductor, a centralized repository focused on high-quality open research software for life science, wherein the use of computational documents is required for documentation. This thesis aims at enhancing the quality and reproducibility of untargeted LC-MS metabolomics data analysis by leveraging Bioconductor to implement the Notame workflow, detailed in the "Metabolomics Data Processing and Data Analysis—Current Best Practices" special issue of the Metabolites journal. To highlight the reproducibility aspect, the thesis is written with a computational document system called Quarto.

Data pretreatment, feature selection and visualizations were implemented in a Bioconductor-compatible workflow using the TreeSummarizedExperiment container almost as specified by the best practices in the Notame protocol article. Data extraction and functionality related to biological context could not be implemented using the TreeSummarizedExperiment container, although these analysis steps are available using other containers in Bioconductor. This demonstration of the utility of the best practices of Notame, Bioconductor and Quarto may facilitate quality research efforts in tackling questions relating the metabolome to complex, biological systems in a reproducible fashion.

Acknowledgements

I'd like to express my sincere gratitude to professor Leo Lahti, professor Ion Petre and associate professor Guillaume Jacquemet for helping me pivot to computational biology. The patient supervision by Leo and Guillaume is well appreciated. Data was kindly provided by senior lecturer Mikael Niku. I also want to acknowledge the enthusiasm of doctoral researcher Retu Haikonen, professor Kati Hanhineva and doctoral researcher Anton Klåvus. Special thanks go out to Fanny and Vertti simply as faithful friends and for in-depth discussions about philosophy, the nature of concepts and other topics which have developed my critical faculty.

Table of contents

Abbreviations

DAWG — Data Analysis Working Group (of MSI)

FDR — Benjamin-Hochberg False Discovery Rate

GF — Germ-Free

LC-MS — Liquid Chromatography-Mass Spectrometry

LC-MS/MS — Tandem Liquid Chromatography-Mass Spectrometry

MAR — Missing At Random

MCAR — Missing Completely At Random

MNAR — Missing Not At Random

MSI — Metabolomics Standards Initiative

mQACC — Metabolomics Quality Assurance & Quality Control Consortium

m/z — Mass-to-charge ratio

PCA — Principal Components Analysis

PLS-DA — Partial Least Squares Discriminant Analysis

PQN — Probabilistic Quotient Normalization

QA — Quality Assurance

QC — Quality Control

RSD — Relative Standard Deviation

RT — Retention Time

SE — SummarizedExperiment

SPF — Specific Pathogen-Free

TSE — TreeSummarizedExperiment

t-SNE — T-Distributed Stochastic Neighbor Embedding

1 Introduction

With the development of experimental techniques suited for omics research, insight in life science increasingly leans on sophisticated data science methodologies to tease apart the complexities of biological systems (Ramos et al. 2017). From the perspective of metabolites as the continuation of the central dogma of molecular biology, metabolomics provides the closest link to phenotype and is thus of special interest, also in multi-omics research (Fiehn 2002).

Untargeted liquid chromatography-mass spectrometry (LC-MS) is the most widely used technique in metabolomics research, largely due to its broad coverage of the metabolome; the small molecules in a biological sample (Gika et al. 2019). However, due to experimental reasons and the extensive data analysis, it meets challenges with regards to quality and reproducibility (Broadhurst et al. 2018). The Notame protocol article (henceforth referred to as Notame), detailed in the "Metabolomics Data Processing and Data Analysis—Current Best Practices" special issue of the Metabolites journal, presents a contemporary analysis workflow for LC-MS metabolomics data analysis intended for new scholars entering the field (Klåvus et al. 2020). Notame and the R package of the same name specifically accommodates the needs of single-batch, untargeted LC-MS metabolomics data analysis typical of food and nutritional research. Yet, Notame is widely applicable in untargeted LC-MS metabolomics data analysis aiming to gain insight by investigating the metabolic profiles between study groups and time points.

Analysis of omics data is well developed in the R/Bioconductor ecosystem (henceforth referred to as Bioconductor), focused on high-quality open research software for life science (Ramos et al. 2017, Gentleman et al. 2004). Bioconductor can be conceptualized as consisting of data containers, software packages and a community of users and developers, who contribute functionality in a preferrably interoperable fashion. Bioconductor delivers releases consisting of a set of compatible R packages intended for compatibility only within a certain version of R, allowing for reproducible analysis (Gentleman et al. 2004). Use of computational documents is required for documentation in Bioconductor, the broader utility of which is reproducibility. Orchestration

of, for example, microbiome transcriptome data analysis with Bioconductor has been explored thoroughly (Lahti et al. 2021). Comparatively, orchestration of untargeted LC-MS metabolomics data analysis with Bioconductor is underdeveloped.

To enhance the quality and reproducibility of untargeted metabolomics data analysis, this thesis aims to implement Notame in a Bioconductor-compatible workflow. No single Bioconductor package provides the functionality needed to implement Notame. Thus, Notame is implemented using a variety of Bioconductor packages. To this end, the space of metabolomics functionality in Bioconductor is explored for implementation of Notame and showcased in an example analysis.

Promoting reproducible, open science is one of the factors identified in the Research Council of Finland's strategy contributing to the renewal, quality and societal impact of science. To highlight the reproducibility aspect of the thesis, it is written using Quarto, a computational document system which generates an output file of desired format based on text and code input (Allaire et al. 2022).

Capitalizing on the best practices of Notame and the reproducibility of Bioconductor and Quarto, can LC-MS metabolomics data analysis be of higher quality and more reproducible? Can Notame be implemented with Bioconductor using a single data container, given its philosophy of interoperability? Is Notame functionality available but lacks interoperability? Tackling the above questions, the state of Bioconductor functionality with regards to Notame is assessed. A Bioconductor-compatible workflow may promote further development of best practices in Bioconductor and quality, reproducible substantive research efforts. This may ultimately translate to knowledge of complex, biological systems and health outcomes at large.

2 Literature review

In service of enhancing LC-MS metabolomics data analysis, the literature review aims at elucidating how Notame, Bioconductor and Quarto promote quality and reproducibility. First, Notame and how it relates to research quality is addressed in an order which best explains central concepts. Notame is also contrasted with other approaches from the literature, providing context for the best practices and anticipating developments. It is important to note that a community-wide effort for best practices is being undertaken by the Metabolomics Quality Assurance & Quality Control Consortium (mQACC), although it is limited to quality assurance (QA) and quality control (QC) (Kirwan et al. 2022). The consensus practices recognized by mQACC are aligned with Notame. Best practices are not necessarily prescriptive, but serve as a guide to state-of-the-art. Accordingly, the Data Analysis Working Group (DAWG) of the Metabolomics Standards Initiative (MSI) has only acknowledged very broad best practices, such as the use of QC samples and cross-validation in supervised learning (Goodacre et al. 2020). For oversight, the metabolomics functionalities are categorized into data extraction, data pretreatment, feature selection and biological context (Figure 1). The focus on data pretreatment and feature selection, as these are catered to by the Notame R package.

Reporting standards are crucial for reproducibility irrespective of the analysis approach, and are disseminated by MSI. DAWG has set forth standards for reporting the data analysis (Goodacre et al. 2007), for which the utility of Bioconductor and Quarto will become apparent towards the end of the literature review.

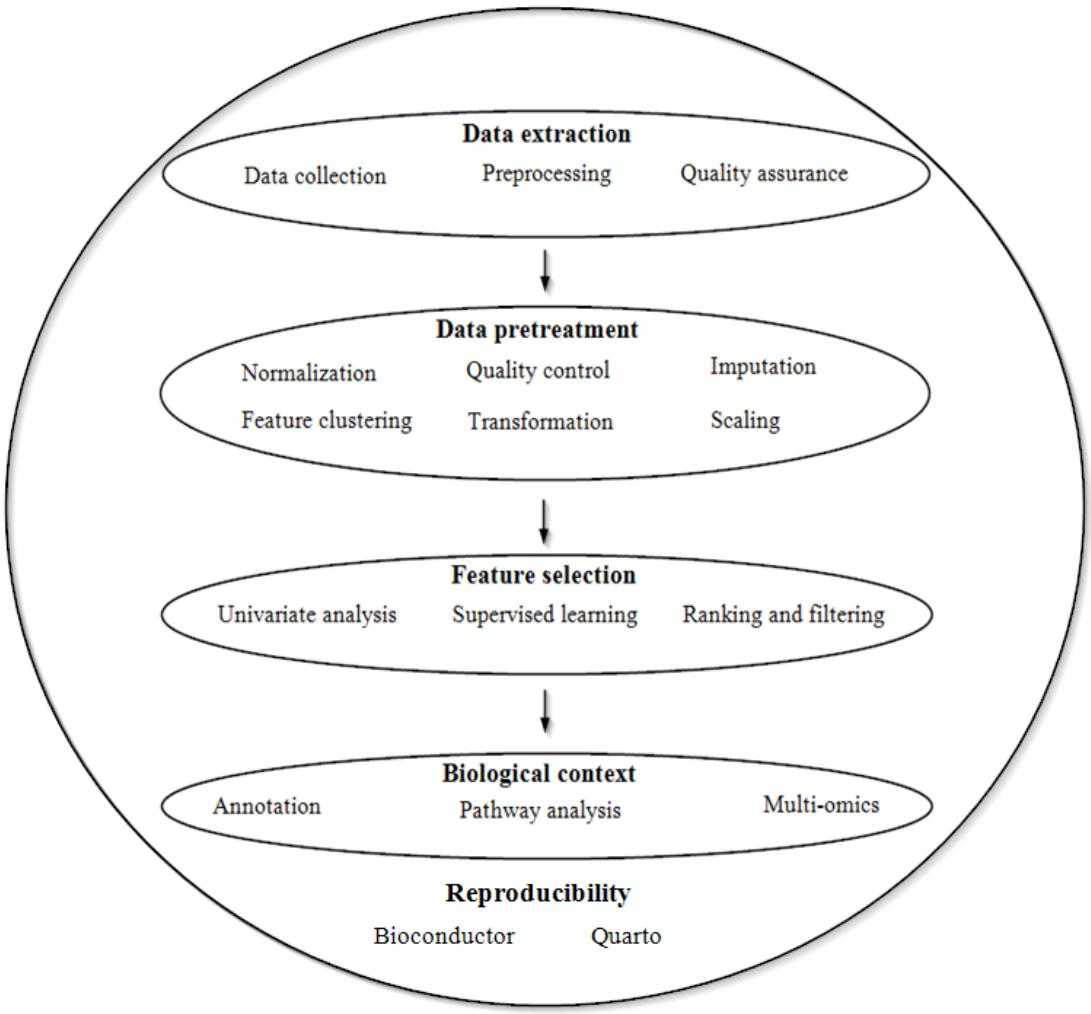

Figure 1. Overview of LC-MS metabolomics data analysis. Data extraction refers to steps leading to a quantitative dataset from data collection. Data pretreatment involves completing the dataset by way of reducing unwanted variation and data preparation dependent on downstream methods. Feature selection aims to select interesting metabolite species across study groups/time points. Functionality related to biological context elucidates the biological meaning of interesting metabolite species. Reproducibility is implied in the above steps and in research at large.

2.1 Data extraction

Data extraction results in a quantitative dataset, consisting of metabolite species characterized by their abundance across samples and metadata.

2.1.1 Data collection

Metabolomics research requires expertise in analytical chemistry, biochemistry, bioinformatics and data analysis (Klåvus et al. 2020). In metabolomics research, biological samples are typically analysed using tandem liquid chromatography-mass spectrometry (LC-MS/MS) (Gika et al. 2019). In the LC-MS step, metabolites in the liquid phase are separated in a chromatography column and ionized for detection by their mass-to-charge ratios (m/z) (McMaster 2005, p. 52). In the LC-MS/MS step, precursor ions from the first stage are fragmented for detection with improved specificity (McMaster 2005, p. 1).

The above is for simplicity; a more nuanced understanding of LC-MS/MS involves chromatography columns, ionization, fragmentation, isotopes, ion suppression, neutral loss formation and detection. Several types of chromatography columns can be used for separation of the metabolites in LC/MS. Often, two columns are used to separate metabolites with different physico-chemical characteristics (McMaster 2005, p. 23). LC-MS/MS ionization methods can be divided into soft ionization and hard ionization. For the LC-MS step, soft ionization is used, commonly electrospray ionization (McMaster 2005, p. 53). Electrospray ionization is operated in positive ion mode and negative ion mode, resulting in protonation or deprotonation of the metabolite, respectively (Ardrey 2003, p. 106). As opposed to hard ionization, the low energy of soft ionization mostly preserves the structure of the metabolite, and fragmentation of the metabolite is limited (McMaster 2005, p. 53). Metabolite species are charged due to the addition or loss of atoms and electrons, which is facilitated by volatile additives such as formic acid and ammonium formate (McMaster 2005, p. 147). Common metabolite species, for example isotopes and those promoted by additives, are characterized by known mass differences with respect to the parent metabolite. Other metabolite species, such as adducts and fragment ions, form less predictably (Schug and McNair 2003) and introduce redundancy in the dataset (Klåvus et al. 2020). These also contribute to reduced detector response, so called ion suppression, which reduces sensitivity of detection (Erngren et al. 2019). Neutral loss formation, meaning metabolite species lost as neutral molecules, also reduces the sensitivity of detection (McMaster 2005, p. 95).

Charged metabolite species are then accelerated and deflected by a magnetic field for

detection (McMaster 2005, p. 45), although the details depend on the instrumentation. The magnitude of deflection depends on the m/z of the metabolite species (McMaster 2005, p. 45). By manipulating the magnetic field, the stream of metabolite species characterized by differing m/z is nudged towards the detector (McMaster 2005, p. 45). As metabolite species hit the detector, an electric current is perturbed, which is amplified and recorded as intensity relative to m/z (Figure 2) (McMaster 2005, p. 62). The more ions arriving at the detector, the greater the perturbation and recorded intensity. In the LC-MS/MS step, a hard ionization technique like collision-induced dissociation is used to fragment metabolite species from the LC-MS step for improved specificity (McMaster 2005, p. 103).

2.1.2 Preprocessing

There are several approaches to preprocessing of the signal to quantitate the metabolite species across samples, but the two most central steps are as follows. First, the spectrum consisting of intensity values relative to m/z undergoes algorithmic peak-picking (Figure 2) to quantitate the signal from each metabolite species (Shimadzu 2023). Second, the retention time (RT), or the time a metabolite takes to pass through the chromatography column, is calculated for each metabolite species based on max peak intensities (Shimadzu 2023) or the raw data from the LC-MS spectra (Figure 2). Finally, since the RT of each metabolite varies between samples because of chromatographic conditions, contamination and other factors, RTs are aligned to match metabolite species across samples (Shimadzu 2023). These preprocessing steps are often performed using point-and-click software, such as MS-DIAL (Tsugawa et al. 2015) specified by Notame. MS-DIAL accounts for isotopes, most common adducts and some fragment ions and combine their abundances into a single entry in the quantitative dataset (Tsugawa et al. 2015). MS-DIAL also computes RTs and performs retention time alignment (Tsugawa et al. 2015). The identities of the metabolite species in the quantitative dataset are characterized by their average m/z and RT values. Quantitative analysis is concerned with quantitative datasets from the LC-MS step extracted using different chromatographic modes. RT and LC-MS/MS spectra are used in metabolite identification as per the most robust identification level specified by the Metabolomics Standards Initiative (Sumer et al. 2007).

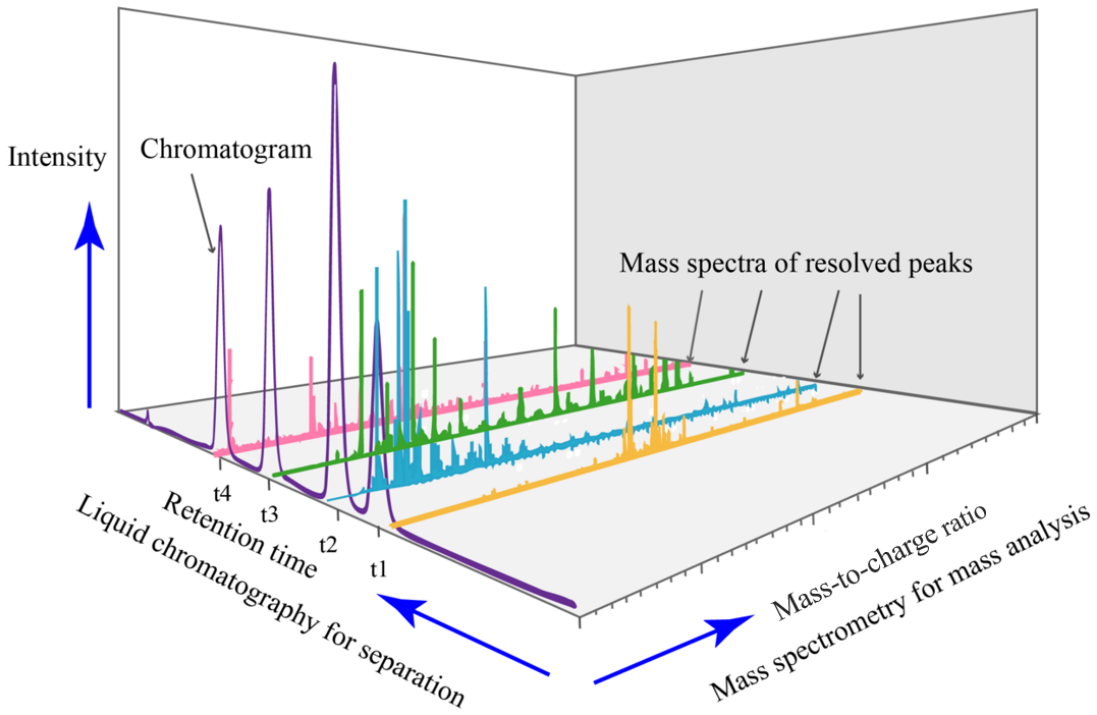

Figure 2. LC-MS data overview. The detector records chromatographically separated metabolite species' intensity relative to their m/z, which are integrated as separate peaks in the peak-picking phase of preprocessing. RT of peaks is determined from the maximum peak intensity. Finally, the same metabolite species across samples is identified with retention time alignment (not shown)

2.1.3 Quality assurance

QA and QC can be thought of as complementary, where QA refers to the steps to assure the quality of the data irrespective of whether it's effect can be reported or not, whereas QC is concerned with, well, control of the quality by flagging low-quality metabolite species (Broadhurst et al. 2018). A useful perspective on QA is that it aims to minimize unwanted experimental variation from sample preparation and data collection, instilling confidence in the results. Unwanted variation can be conceptualized as consisting of measurement error, unwanted experimental variation and unwanted biological variation. Variation is often represented as variance, the average of squared deviations from the mean for each study group/time point.

Untargeted LC-MS, in particular, suffers from scant standardization of QA (Broadhurst et al. 2018). One reason is the difficulty in using isotope-labeled internal standards for QA in untargeted approaches aiming to maximize the number and physico-chemical diversity of metabolites detected in a biological sample. As such, QA practices can't be

directly adopted from targeted approaches (Broadhurst et al. 2018). Standardized QA has not been advanced by MSI or mQACC. However, mQACC recognizes a consensus on QA for which reporting standards have been disseminated. Next, the widely cited QA practices from Broadhurst et al. (2018), in line with those recognized by mQACC and Notame, are presented from the perspective of QC samples and data collection, as these constitute important context for data pretreatment. Biological sample preparation for LC-MS metabolomics experiments is minimal and not critical for the remainder of the thesis, and is thus omitted.

QC samples have extensive utility in assuring the quality of untargeted metabolomics experiments, especially in normalization and estimating precision. To account for experimental variation in QC samples, it is critical to process the QC samples and biological samples identically. One way to prepare QC samples is to aliquot each biological sample into replicates, and randomize the order of injection. This greatly increases data collection and QC time because of multiplying the number of samples. A more time efficient alternative is to aliquot a single biological sample into > 5 replicates, which are evenly distributed in the batch. Here, the assumption that the chosen, aliquoted biological sample has a representative metabolite profile is problematic since there may be metabolites present in other biological samples which are below the detection threshold in the sample from which replicates are aliquoted. Pooled QC samples are a workable compromise and used in Notame, where aliquots from each biological sample are pooled, subaliquoted and distributed evenly in the injection sequence.

Long-term reference QC samples and standard reference materials can be used for QA across batches. Long-term reference QC samples for studies involving the same experimental procedure are simply large pooled QC samples, aliquoted and stored in liquid nitrogen. Including such samples in each batch allows for estimating the precision in and across batches. Standard reference materials, on the other hand, are suitable for QA across laboratories as they are within strict tolerances.

Another aspect of QA is assessing system suitability to reduce unwanted experimental variation. At the very beginning of an experiment, a blank excluding everything but solvents is injected. This provides an assessment of possible contaminants in the system, necessitating cleaning or column change. If the blank doesn't show any unexpected peaks, several QC samples are injected for conditioning, that is reducing absorption of

metabolites in the system, RT variability and stabilizing the detector response. A process blank is injected in the middle of the conditioning QC sample sequence to identify signals from contamination, biochemical precipitation, extraction and more. The process blank is prepared using the exact same protocol as in preparing the biological samples, but excluding the biological sample. When the signal from the conditioning QC samples has stabilized, two QC samples are injected, marking the start of the experimental run. Before the first biological sample, a standard reference or long-term reference QC sample may be run to to estimate accuracy and precision across batches, followed by two more interspersed later in the injection sequence. Biological samples are then injected, interspersed by QC samples every five samples, for example. At the end of the injection sequence, two QC samples are injected so that the regression model used to normalize for drift interpolates to the last biological samples. Then, another process blank is injected to test for cumulative carryover, that is the detection of biological signals which "leak" to the next sample, which can inform adjusting the washing between sample injections. Peaks in the processing blank satisfying specific exclusion criteria are removed from the quantitative dataset. Finally, five or so QC samples are included for LC-MS/MS data collection. The number of processing blanks, reference samples, sequential QC samples, LC-MS/MS samples and the ratio of QC samples to biological samples should be adjusted for the experiment and instrumentation at hand.

Despite QA, some unwanted variation makes it to the quantitative dataset. Some adducts and fragment ions form unpredictably and can be regarded as introducing noise (Schug and McNair 2003). Tuning experimental parameters is common practice for reducing adduct formation and fragmentation in the LC-MS step (Ardrey 2003, p. 232). However, in untargeted metabolomics, where a wide range of metabolites with diverse physico-chemical characteristics are profiled simultaneously, it is difficult to find a set of parameters that globally corrects for adduct formation and in-source fragmentation (Ardrey 2003 p. 232). Adducts and fragment ions can result in the same metabolite being redundantly represented in the dataset (Klåvus et al. 2020), although these can, to the degree that they can be identified in a reference database, be combined into a single entry before obtaining the quantitative dataset using software such as MS-DIAL (Tsugawa et al. 2015). This reduces unwanted experimental variation (Broeckling et al. 2014), multicollinearity and problems with identification (Klåvus

et al. 2020). Sources of measurement error include heteroscedastic noise (Berg et al. 2006), spectral skewing (Berg et al. 2006) and drift in the relative abundance of features as evidenced by QC samples, which can be caused by many factors in the instrumentation and non-enzymatic metabolite conversion (Broadhurst et al. 2018). An important source of unwanted biological variation is the differential dilution of samples (Dieterle et al. 2006). Another common occurrence is missing values, often because of a concentration below the limit of detection, errors in preprocessing of the data or stochastic shifts in instrument function (Wei et al. 2018).

2.2 Data pretreatment

In data pretreatment, unwanted variation is reduced with computational methods and quality is assured with QC before preparing the dataset for feature selection. There is some discussion in the literature on the use cases in which data pretreatment methods are best applied, which is dependent on the feature selection methodology. In Notame, the imputed and clustered dataset is used without any transformation, normalization or scaling for univariate analysis after QC. For supervised learning, nlog or glog transformation, probabilistic quotient normalization (PQN) and autoscaling is used, although autoscaling is not needed for scale-invariant methods such as random forest.

Though specifying data pretreatment, Notame partially omits references instilling confidence in the practices, where different combinations of pretreatment methods may yield different results. Guida et al. (2016) systematically investigated the effect of 280 permutations of normalization, imputation, transformation and scaling methods after QC, omitting feature clustering and drift correction. Conclusions were arrived at by splitting a dataset into two groups with no significant differences and then modifying the peaks of one group to reflect significant differences and non-significant differences. This modified dataset was then used to evaluate permutations of data pretreatment performed for univariate tests and partial least squares discriminant analysis (PLS-DA) using the known true positives (significant differences) and false positives (non-significant differences) in the modified dataset. The results indicate that univariate analysis is best undertaken with a non-imputed dataset normalized using PQN. For

PLS-DA, the results suggest using k-nearest neighbors imputation and glog transformation.

Then there is the order in which the pretreatment is performed. In Notame, features are first normalized for drift, followed by QC, imputation and clustering. The data is then transformed, normalized for dilution and scaled depending on the feature selection conducted. There is no discussion in the literature on the order in which the data pretreatment steps are best applied.

2.2.1 Normalization

In the LC-MS literature, normalization mostly refers to reducing feature-wise or sample-wise unwanted variation, namely from from drift and dilution. LC-MS measurements suffer from systematic intensity drift resulting in a change in abundance as a function of injection sequence, constituting systematic measurement error (Broadhurst et al. 2018). The removal of drift increases the quality of the data by reducing variation introduced by the systematic measurement error while conserving the biological variation of interest (Märtens et al. 2023). The simplest drift correction method is scaling to total feature abundance, median abundance or some other metric calculated for the dataset or study groups/time point at whole (Märtens et al. 2023). These integral normalization methods, however, assume that all metabolites experience the same pattern of variation (Märtens et al. 2023) and that there is no unwanted biological variation from dilution. Commonly, isotope-labeled internal standards included at a uniform concentration are used to correct for drift, but this isn't feasible in untargeted approahces since isotopes for thousands of metabolites would have to be inluded (Sysi-Aho et al. 2007). Computational techniques have been devised to select a subset of optimal standards to correct drift for all features (Sysi-Aho et al. 2007). Drift can also be countered by real-time correction, where the detector voltage is calibrated after each sample (Lewis et al. 2016). This is unlikely to account for all sources of drift, for example non-enzymatic metabolite conversion (Broadhurst et al. 2018).

The most widely used drift correction method in the literature is QC-based drift correction (Broadhurst et al. 2018), which is used in Notame. Pooled QC samples included at regular intervals in data collection, consisting of aliquots from each sample, are used

to remove the drift using a predictive model fitted to the features in QC samples by injection order (Broadhurst et al. 2018).

In Notame, univariate regression is used to model the relationship between the independent, explanatory variable (injection order) and the continous, dependent outcome variable (abundance). Univariate regression models the relationship by fitting a line that best fits the data, approximating the true relationship between the explanatory variable and the outcome variable. The approximation is evaluated by the residual, that is the distance between the value of the outcome variable to the modeled value of the outcome variable for a given value of the explanatory variable. Many assumptions are made in univariate regression, although these are likely to be violated in metabolomics datasets despite transformation because of the large number of features. Then again, the quality metrics used to evaluate and flag features in QC are not necessarily dependent on the assumptions as non-parametric equivalents are available.

Relaxing the linarity assumption, one can use a non-linear regression model, such as the smoothed cubic spline used in Notame. The data is nlog-transformed to better meet the assumptions of homoscedasticity and normality of residuals. The cubic spline is then used to model each feature's abundance across QC samples by connecting the observations using third order polynomials. Since QC samples are included at regular intervals interspersed with biological samples, the smoothing serves to interpolate the model to the biological samples (Broadhurst et al. 2018). The value of the smoothing parameter is optimized using leave-one-out cross validation to avoid overfitting to unwanted variation in the QC samples (Broadhurst et al. 2018). Having fit the smoothed cubic spline regression model, abundance values are corrected by adding the mean of a feature's abundance in the QC samples and subtracting the drift for the feature in each sample (Klåvus et al. 2020).

Another central source of unwanted variation is the differential dilution of samples, resulting in varying total feature abundances across samples (Dieterle et al. 2006). Normalization can reduce unwanted variation from dilution while retaining the biological variation of interest (Dieterle et al. 2006). Dilution, in this context, can refer to both experimental variation and unwanted biological variation. For example, in normalization for dilution of urine samples, each feature can be scaled to the abundance of creatinine (Waikar et al. 2010). This integral normalization method rests on the assumption that

dilution is the sole source of unwanted variation and that the excretion of creatinine can be used as a robust, constant scaling factor across samples (Dieterle et al. 2006). Another widely used method is quantile normalization. In quantile normalization each sample is given the same distribution of feature abundances (Kohl et al. 2012). The resulting feature abundances in a sample consist of the same set of values, but they are distributed differently among features (Kohl et al. 2012). Quantile normalization is problematic as a feature may end up with the same abundance across all samples, especially in the case of low- and high-abundance features (Jauhiainen et al. 2014).

The above methods do not take into account different sources of variation, which is important if it can't be reasonably assumed that the variation in total feature abundance arises predominantly from dilution as in the case of urine (Dieterle et al. 2006). In many biological samples, biological variation of interest has an undeniable effect on the total abundance of features in a sample, in which case integral normalization not only scales for differential dilution of samples, but also for the biological variation of interest (Dieterle et al. 2006).

PQN has emerged as a promising general approach, and is used in Notame. PQN posits that changes in metabolite concentrations only influence some features, whereas differential dilution influences the whole sample and that dilution is the main source of reducible unwanted variation (Dieterle et al. 2006). To accommodate biological variation and variation from dilution, PQN determines a dilution factor by calculating a distribution of quotients from the median of each feature in the biological samples and a reference spectrum (Dieterle et al. 2006). The median of these quotients is used as a dilution factor for all samples (Dieterle et al. 2006). The best reference spectrum is the median abundance of each feature across QC samples (Dieterle et al. 2006).

When normalizing for dilution between batches, an integral normalization can be performed before PQN to scale the batches to the same total abundance (Dieterle et al. 2006). PQN can also be performed on the raw data if several batches will be preprocessed as a single batch (Dieterle et al. 2006). Alternatively, identical long-term reference pooled QC samples can be used to normalize for dilution between batches (Broadhurst et al. 2018). Standard reference materials also hold promise in such batch normalization (Broadhurst et al. 2018) and even absolute quantification, although this

is a new development that suffers from having to include internal standards for each metabolite or computationally selecting a subset of standards.

2.2.2 Quality control

Including multiple pooled QC samples allows for calculation of measures of precision for each metabolite detected in the QC samples. These calculations can also be made for many batches with long-term reference QC samples (Broadhurst el al. 2018). Absolute estimates of accuracy, that is how close the measured concentration is to the true concentration, and absolute estimates of precision, that is random measurement error and experimental variation, in quantification over repeated measureument of identical samples can not be obtained using untargeted LC-MS (Broadhurst et al. 2018). This is again because of the difficulty in using internal standards for limits of detection, limits of quantification, and linearity quantifiers for each feature (Broadhurst et al. 2018). Thus three metrics, the relative standard deviation (RSD), the dispersion ratio (D-ratio) and the detection rate are widely used for QC of untargeted metabolomics data (Broadhurst et al. 2018), also in Notame.

Only the relative random measurement error in quantification of identical QC samples can be obtained as a measure of precision in untargeted LC-MS metabolomics data analysis (Broadhurst et al. 2018). If the random measurement error across identical QC samples is represented as a normal distribution, the random measurement error can be described in terms of standard deviations for each feature (Broadhurst et al. 2018). To compare the standard deviation of the features, the RSD is calculated for each feature by dividing the standard deviation with the mean of the feature (Broadhurst et al. 2018). The non-parametric median absolute deviation can also be used for such calculations, which can be scaled for comparison to RSD using the scaling factor 1.4826 (Broadhurst et al. 2018).

A measure of total experimental variation and measurement error for each feature is obtained by dividing the standard deviation of the identical QC samples with the standard deviation of the biological samples (Broadhurst et al. 2018). The result is the D-ratio, for which a non-parametric alternative is also available (Broadhurst et al. 2018).

The standard deviation of features in the identical pooled QC samples should represent only experimental variation and random measurement error, while the biological samples include experimental variation, random measurement error and biological variation (Broadhurst et al. 2018). Assuming an additive random error structure, the total variation can be simplified to experimental variation and random measurement error plus biological variation (Broadhurst et al. 2018). From this it can be concluded that a D-ratio of 0 means that there is no experimental variation or random measurement error in the feature, and only the biological variation of interest is conserved. On the other hand, a D-ratio of 1 translates to that there is only experimental variation and random measurement error, with no biological variation. The assumption that the random error structure is additive may not hold; it is commonly known that random measurement error is multiplicative as it results in heteroscedasticity (Veselkov et al. 2011), although this can be mitigated with transformation.

Finally, there is the detection rate. The detection rate is derived by dividing the observed number of each feature across QC samples with the total number of QC samples (Broadhurst et al. 2018). Features with a low detection rate are not reliable (Broadhurst et al. 2018), and probably arise because of the detection threshold. These three metrics (RSD, D-ratio, and detection rate) provide an assessment of quality that can be reported for each detected metabolite and are used to remove low quality data from the dataset prior to further data pretreatment and analysis. The exclusion criteria for the quality metrics are not set in stone. It may also be motivated to use two sets of exclusion criteria, for example to prevent removal of features with very low values in all but a few samples (Kärkkäinen et al. 2021). Notame recommends a detection rate > 0.7, RSD < 0.2 and D-ratio < 0.4.

Because of the large number of features, it is not feasible to inspect feature-wise plots at this stage of the analysis. Visual inspection is performed for each mode separately. This also serves as exploratory data analysis to form intuitions about the dataset at hand. An overview of the literature suggests that the Notame QC visualizations are quite comprehensive and somewhat original; further review is thus omitted.

2.2.3 Missing value imputation

Many downstream methods rely on a complete dataset with no missing values. Values missing due to being below an instrument's limit of detection are often referred to as missing not at random (MNAR) (Dekermanjian et al. 2022). Missing values caused by, for example, incomplete ionization because of stochastic shifts in instrument function are often referred to as missing completely at random (MCAR) (Wei et al. 2018), and are uniformly distributed in the dataset. Missingness which is dependent on other features in the dataset is called missing at random (MAR), and often arise from errors in preprocessing (Wei et al. 2018). MARs and MCARs are often indistinguishable in practice (Dekermanjian et al. 2022). Notame deals with much of the missingness by removing features that are not detected in > 70% of the samples, removing features that are at the detection threshold to a greater extent as such features also include MNARs. If any MNARs from around the detection threshold remain in the dataset, they may be imputed as if they were MCARs/MARs by the random forest method used in Notame (Wei et al. 2018). This is not optimal, as it has been found that other methods are better suited for imputing MNARs (Wei et al. 2018). Classification of missing values as MNARs and MCARs/MARs has been used to separately and more accurately impute missing values arising from different missingness mechanisms (Wei et al. 2018, Dekermanjian et al. 2022). Random forest imputation has been shown to outperform other methods in imputing MCARs/MARs (Kokla et al. 2019).

Removing features with missing values altogether is one way to deal with missing values, although this would result in a severely diminished dataset against the spirit of untargeted LC-MS metabolomics research (Wei et al. 2018). Removing samples with missing values is likely to result in an even more diminished dataset due to the the high number of features compared to samples, and reduced certainty of findings (Dekermanjian et al. 2022).

2.2.4 Clustering features originating from the same metabolite

MS-DIAL combines many isotopes, common adducts and in-source fragments into a single feature (Tsugawa et al. 2015). Still, redundant representation of the same metabolite can not be ruled out (Klåvus et al. 2020). Redundancy in the dataset can be reduced

by RT-based clustering, based on the intuition that metabolite species reflect the abundance of the parent metabolite across samples and have similar RTs (Klåvus et al. 2020). This is done separately for each mode in the dataset due to differing RTs.

In Notame, features are clustered based on correlated feature pairs within a specified RT window and correlation threshold. RT-based clustering is an advancement in relation to general clustering methods such as k-means clustering. A handful of RT-based clustering methods have been described (Broeckling et al. 2014, Joo et al. 2023), but further review is omitted as they are not as established as earlier data pretreatment steps.

The Notame method first identifies pairs of correlated features within a specified RT window, and features that don't meet a specified correlation coefficient threshold are considered groups of their own. The intuition for this is that co-eluted features with similar RTs and correlated abundances are likely to originate from the same metabolite. Notame uses Pearson's correlation for correlating the abundances, assuming that the relationship of features originating from the same metabolite is linear: the co-eluted feature abundances increase as a linear function of each other. This is not necessarily the case, especially at the lower and upper limit of quantification of the instrument (Klåvus et al. 2020).

Next, an undirected graph is generated using feature correlations as edge weights. Connected nodes are then treated as separate groups in a recursive algorithm, where features originating from the same metabolite are identified using a degree threshold. This serves to reduce the number of false annotations in a situation where a feature is connected to only one or two nodes in the group, and are unlikely to originate from the same metabolite despite co-elution. The degree threshold is defined as a percentage of the maximum possible degree, where degree is the maximum number of connections from a single node. Thus, for a group with five nodes, the degree is 4 and a degree threshold of 0.8 would mean that each node must have at least 3 edges ($0.8 \times 4 = 3.2 \approx 3$). Until this criterion is met in each group, the node with the lowest degree is discarded, or if there is a tie, the node with the lowest sum of edge weights. The result is a cluster where each node is connected to all other nodes, reflecting co-elution and abundance correlation because they originate from the same metabolite. The nodes that are initially discarded can form new clusters if the degree criterion if fulfilled. The feature with the largest median abundance is retained for each cluster.

After feature clustering, the datasets corresponding to different chromatographic modes are merged. This results in a dataset with some redundancy as many features are detected in multiple modes.

2.2.5 Transformation and scaling

Transformation and scaling completely change the feature abundance space, in contrast to normalization for drift and dilution. Multiplicative random measurement error and spectral skewing introduces heteroscedasticity and skewness in the features, respectively (Berg et al. 2006). Transformations are common for reducing heteroscedasticity and skewness to better meet the assumptions of statistical analyses (Berg et al. 2006). Moreover, in metabolomics, relations among metabolites may not always be additive, and log-transformed values can better account for multiplicative relations with linear techniques (Berg et al. 2006). A log10 transformation perfectly removes heteroscedasticity if the random measurement error increases linearly with abundance; the RSD is constant (Kvalheim et al. 1994). However, low-abundance features typically have a larger RSDs from experimental variation (Berg et al. 2006). For such non-linear increase in multiplicative random measurement error, the power transformation can be used, although this does not perfectly remove heteroscedasticity or promote conversion of multiplicative to additive relations (Berg et al. 2006). In Notame, a natural log transformation or glog transformation, in case of heavily skewed data, is used. These two options were arrived at through the comparative investigation by Guida et al. (2016), although no more discussion is available in the literature.

Another challenge in LC-MS data is that different metabolites can have very large differences in abundance. These up to 5000-fold differences in abundance do not reflect the biological importance of metabolites (Berg et al. 2006). Instead, much insight is extracted from how the intra-feature variation compares to that of other features. To make the intra-feature variation comparable across features, the variation needs to be represented on the same scale. This can be done by scaling, for example using the standard deviation of each feature as a scaling factor (Berg et al. 2006). Mean centering followed by division of abundances with the standard deviation, autoscaling, results in features having the mean at zero and a standard deviation of one, a prerequisite for

many machine learning methods based on distance measures. On the other hand, scaling results in inflation of small abundance features (Berg et al. 2006). Other scaling methods, such as pareto scaling, range scaling and vast scaling were compared in conjunction with other data pretreatment steps by Berg et al. (2006), Guida et al. (2016) and Kohl et al. (2012). Autoscaling and pareto scaling were deemed most promising. Transformation also scales the features somewhat as the difference between large and small abundance features is reduced, but this does not bring the features to the same scale (Berg et al. 2006). Notame specifies autoscaling.

2.3 Feature selection

Feature selection aims to rank features with regards to study group/time point to select a subset of features for downstream methods pertaining to biological context. DAWG has yet to put forth guidelines for reporting of how features are selected for pathway analysis, annotation and other steps which provides biological context for the findings. These steps are often labour intensive: for example, database matching represents only a putative metabolite annotation that must be confirmed by comparing the RT and/or LC-MS/MS spectra of a pure compound to that from the feature of interest (Vinaixa et al. 2012). This is time consuming and represents is the rate-limiting step (Vinaixa et al. 2012). In Notame, feature-wise and comprehensive visualization of results is used to facilitate and communicate intuitions about interesting features instead of just presenting the results in a boring table. The visualizations used inevitably shape the interpretation of the results. The literature suggests that the Notame visualizations are comprehensive, although many additional visualizations could be used to visualize the properties of interest, for example bar plots, swarm plots, violin plots and Euler diagrams. One feature-wise perspective omitted in Notame is reinspecting the raw spectra to assess the validity of the results (Grace and Hudson 2016). This could reveal falsely high-ranked features caused by scaling artifacts or spurious peak assignment (Grace and Hudson 2016). Notame includes three plots for relating RT, m/z and feature selection results which could be combined into a single cloud plot such as the one available in XCMS Online to facilitate interpretation (Patti et al. 2013).

2.3.1 Univariate analysis

Univariate, that is feature-wise, analysis can be divided into parametric and non-parametric methods. Parametric methods are based on assumptions about the population from which the samples were drawn, such that the data can be described by a mean and standard deviation. Assumptions do not burden non-parametric methods to the same degree. Some methods relax an assumption for the study design at hand, for example in repeated measures designs where the assumption of independence can be accommodated in a linear mixed effects model. Parametric and non-parametric methods differ in using the mean and median or ranks of the distribution, respectively. Parametric methods are more powerful: a non-parametric method can miss a statistically significant difference that a parametric method would recognize if assumptions are met (Vinaixa et al. 2012). With non-normal distributions, heteroscedasticity and unequal study group/time point sizes, non-parametric methods are preferred (Vinaixa et al. 2012). There are tests for assessing whether an assumption is met, although the conceptual underpinnings and interpretation are debatable. For example, normality tests aim to assess what the probability of the null hypothesis, namely that the values in the population are normally distributed, being the case on chance. With a larger population, more stringent requirements for conformity with normality are required in normality tests such that large populations seldom meet the assumption of normality, although large populations approach normality as per the central limit theorem (Vinaixa et al. 2012). For real stringent testing of assumptions, multiple testing correction may be considered, which is likely to result in the rejection of the null hypothesis in testing for normality and homoscedasticity for a large number of features. Whether or not multiple testing correction is used, the literature suggests that features not meeting some assumptions are usually not removed in metabolomics studies; instead parametric tests are often used by precedent. Parametric tests are robust against mild violations of assumptions, but what is considered mild depends on the context.

Given the large number of features in untargeted approaches and that either a parametric or non-parametric approach should solely be used, violations of assumptions of parametric tests, the reduced power of non-parametric tests or removal of features not

meeting assumptions must be accepted. Testing for normality and homoscedasticity using the Shapiro-Wilk and Levene's tests, around 40-80% of features in LC-MS datasets seem to meet the assumption of normality and homoscedasticity without transformation (Vinaixa et al. 2012), as is suggested for univariate analysis in Notame.

The probability threshold for rejection of the null hypothesis is, by convention, 5%, and can also be thought of as the probability of obtaining a false positive result. A p-value larger than 0.05 can thus be thought of as a failure to reject the null hypothesis due to lack of evidence. However, in univariate analysis for feature selection, the interpretation of statistical significance is not critical if the p-values are solely used to rank features to find a subset of interesting features, illustrating the context-dependent nature of the probability threshold.

Fold-change of a feature, a commonly used measure of effect size in metabolomics, and statistical significance do not account for sample size, so power analysis is often used to validate findings (Vinaixa et al. 2012). Post-hoc power analysis, however, is considered fraught with conceptual problems (Vinaixa et al. 2012), and a priori sample size estimation is problematic because of the highly multicollinear data (Hendriks et al. 2011). Accordingly, there is no mention of statistical power in Notame. A pilot study could be conducted to estimate the number of samples needed for an untargeted LC-MS study to make a convincing case for univariate results (Iterson et al. 2009). Then again, ethical and economical considerations mainly determine the number of samples (Vinaixa et al. 2012), especially given the ambiguity of power analysis and the labour-intensive nature of untargeted approaches.

Notame proposes the following univariate analyses, covering a range of study designs. For case versus control studies with two groups and no covariates, Welch's t-test is used as it allows for unequal variances between groups. The Mann-Whitney U-test can be used as a non-parametric alternative.

For study designs with multiple groups, Welch's ANOVA, which allows for unequal variances, is used to select interesting features based on p-values. Welch's ANOVA is quite resistant to non-normal distributions. Heteroscedasticity is a problem of similar magnitude for the non-parametric alternative, the Kruskal-Wallis test, and isn't specified by Notame. To investigate differences between multiple groups post-hoc, pairwise

Welch's t-test or Mann-Whitney U-test is used post-hoc.

In the case of two categorical study factors, two-way ANOVA is applied to examine the main effect of each factor and their interaction. If factors have multiple levels, interesting features are selected based on overall p-values and further examined using Welch's t-test. Friedman test can be used as a non-parametric alternative to two-way ANOVA, with Mann-Whitney U-test for post-hoc comparisons.

For repeated measures designs, a linear mixed effects model is used with the time point, group and interaction factors as fixed effects and subjects as random effect. To assess the significance of effects between no more than two groups or time points, t-tests are applied on the regression coefficients of the fixed effects. If multiple groups and/or time points are included, type III ANOVA is used to assess the significance of the effects, returning p-values from an F-test.

To investigate the association between features or between features and other variables, Pearson correlation or the non-parametric Spearman correlation is used.

Finally, p-values are adjusted using the Benjamin-Hochberg false discovery rate (FDR) approach for multiple testing. Univariate analysis does not suffer from sparsity typical of biomolecular data as only one feature is considered at a time. However, the large number of features considered increases the likelyhood of false positives if univariate analysis are used for hypothesis testing. Multiple testing correction is done to correct for the chance of obtaining significant results in a situation where, on average, a significance threshold of 0.05 for twenty tests would give a false positive for one test (Jafari et al. 2019).

2.3.2 Supervised learning

Complex, biological systems are multivariate by nature: they can't be described in terms of single variables, but can be approached with modelling multiple variables in the interconnected whole (Goodacre et al. 2007). However, multivariate analysis in molecular biology is complicated by "the curse of dimensionality", or sparsity: datasets have many features but relatively few samples. Adding features in a dataset results in an exponential volume increase. For example, evenly sampling 10^2=100 values

between 0 and 1 would result in a vector with a spacing of 0.01 between points. To arrive at the same spacing in a dataset with ten features, 10^20 observations would be needed. The number of features in the dataset is somewhat reduced by removal of low-quality features and feature clustering, but multivariate analyses still suffer from sparsity of the data. In practice, the curse of dimensionality is encountered when the performance of a multivariate method starts to deteriorate as the number of features increase, given constant sample number. This is referred to as overfitting. Selecting a subset of informative features (herein referred to as variable selection) and validation of the model can mitigate overfitting.

Supervised learning is widely used in molecular biology, where the data is modeled with respect to specific outcome variables, namely study group/time point. Some supervised learning methods, like Lasso regression, are well suited for sparse data since they perform embedded variable selection; variable selection is inherent to the method (Rinaudo et al. 2016). Other supervised learning methods, such as random forest and PLS-DA, often employ wrapper methods where the importance of features are assessed and selected when training the model using cross-validation or bootstrapping (Rinaudo et al. 2016). Some sources suggest that results from univariate analysis could be used for variable selection, constituting a filter method (Xia et al. 2013). However, filtering features based on results from univariate analysis doesn't reflect the intuition behind supervised learning. Univariate analysis, as noted in the previous section, is concerned with whether a difference between study groups/time points is due to chance, which doesn't translate to classification performance, which is the concern of supervised learning methods used for feature selection (Xia et al. 2013). Moreover, multicollinear features can be significant in aggregate, but aren't necessarily found so by univariate analysis (Xia et al. 2013), which may cause a filter method to eliminate informative features. Multicollinearity can be accounted for using a filter method by mutual information criterion (Lin et al. 2012) or correlation coefficient (Grissa et al. 2016).

Validation of supervised learning models is done by training on a subset of the dataset, the training set, after which the model is validated on a validation set (Xia et al. 2013). This is often done in a cross-validation scheme, where training and validation sets are repeatedly subset from the dataset so that the model doesn't overfit to a particular subset of the data (Xia et al. 2013). A further held-out test set is used to evaluate the

performance of the model on data not used in training and validating the model (Xia et al. 2013). This serves as an estimate as to how well the model generalizes to new data.

Notame specifies feature selection by supervised learning to be performed using random forest or PLS-DA, preferentially using the MUVR package, featuring unbiased variable selection (Shi et al. 2019). Selection bias is introduced when features are selected based on all or some of the data used in training the model. MUVR minimizes selection bias by repeatedly and randomly sampling the entire dataset to a training/validation set and a test set and averaging the feature ranking across such outer repetitions. For each outer repetition, a validation set is randomly sampled from the training set to select the optimum model parameters from models fit in a cross-validation scheme on the remaining training set. A proportion of low-ranking features are eliminated, and optimum model parameters and average feature ranking is obtained by repeating model fitting and validation on the reduced and anew randomly sampled training and validation set. The optimum model is then trained on a combined validation and training set, used to obtain the optimum model for feature ranking using the test set in the given outer repetition. With several outer repetitions, this method arrives at optimally ranked features and model parameters for the entire dataset. The above is for relative simplicity, as MUVR actually returns a “min”, “mid” and “max” model tailored to the analytical problem. The “min” model returns likely biomarker candidates, while the “max” model returns a set of features which encompass all the information content, used for pathway analysis. The “mid” model is a compromise between the “min” and “max” models, and is likely to perform best for classification in replicate experiments or diagnostic applications, for example. Finally, MUVR also allows for multilevel classification, which is handy for classification of both study group and time point membership.

2.3.3 Feature ranking

In Notame, feature ranking from univariate analyses and supervised learning are combined to determine the most biologically relevant features for identification. First, the ranks from supervised learning are sorted such that the most predictive feature comes first. Univariate results are sorted similarly based on the p-values. The univariate and

supervised ranks are then summed and sorted to create a combined ranking. The number of ranked features to be explored for biological meaning depends on the user.

Such combined ranking is not showcased in any likely candidate papers, and the intuition is not elaborated in Notame. Combining results from univariate analysis and supervised learning in the combined ranking constitutes a combination of hypothesis testing and the importance for classification. This also brings the assumptions of univariate analysis and variability of results in supervised learning into the picture. Other sources recommend using both univariate and multivariate analyses for feature selection to maximize the extraction of relevant information, but approaches these as separate findings (Vinaixa et al. 2012).

2.4 Biological context

The above steps in LC-MS metabolomics data analysis are used in some form in most of metabolomics research to select interesting metabolites across study groups/time points for discovery of biomarkers and the mechanistics of biochemical pathways, perhaps with increased leverage from other omics and clinical data. Univariate analyses in untargeted approaches support the above ends via selection of features for further scrutiny as in Notame, but hypothesis testing is often not an end in itself. Targeted approaches are better equipped to confirm tentative results from untargeted analyses with more specific hypotheses pertaining to differential metabolite abundance across study groups/time points because limits of detection, limits of quantification and linearity quantifiers can be addressed using standard reference materials (Broadhurst et al. 2018). As such, the quality and reproducibility of untargeted research mainly has implications for directing effort in further targeted research (Johnson et al. 2016). This somewhat lessens quality concerns in untargeted approaches, as insight gained by untargeted approaches is likely further investigated by targeted approaches before scientific consensus and application (Johnson et al. 2016).

2.4.1 Annotation

Annotation aims to identify metabolites of interest and record their physico-chemical propertires, which can prompt discussion on how the results relate to the biochemical workings of the biological system at large. MSI has set forth three levels for metabolite identification (Sumner et al. 2007). Metabolites annotated by matching m/z, LC-MS/MS spectra and RT to a reference database are considered identified. Metabolites with matching m/z and MS/MS spectra are considered putatively identified if the MS/MS spectra uniquely matches a reference metabolite. Finally, putative characterization status is given to metabolites for which only compound class can be established.

Annotation has not been fully automated although there are advances. Annotation often proceeds first with an automated step, often in point-and-click software as per Notame, followed by manual curation. Database quality is paramount in annotation. Annotation is limited by known unknowns, as only a subset of metabolites are included in databases, although some physico-chemical properties can still often be assigned (Zulfigar et al. 2023). Another issue is the relatively poor reproducibility of MS/MS spectra, affecting the two higher quality metabolite identification standards put forth by MSI (Zulfigar et al. 2023).

2.4.2 Pathway analysis

Pathway analysis puts the results in a mechanistic context, providing insight into the biochemical processes across study groups/time points. This can inspire further research into disease mechanisms, for example. Pathway analysis can also be performed in a multi-omic fashion for increased leverage. Notame recommends pathway analysis to be conducted with at least putatively annotated metabolites. Pathway analysis, in addition to preprocessing and annotation, is performed with point-and-click software in Notame.

2.4.3 Multi-omics

Although not discussed in the Notame protocol article, multi-omics expands on the promise of metabolomics in teasing apart the complexities of biological systems, for ex-

ample in complex disease. Multi-omics is promising in that the variables are molecule abundances, which can generate hypotheses for further mechanistic research and treatments.

Multi-omics makes for even sparser data. Multi-omics research often utilizes supervised or unsupervised learning. There are several strategies for integrating omics datasets for multi-omics supervised learning. The late integration strategy considers interesting features found in separate analyses of each omics dataset (Picard et al. 2021). For late integration, features could be selected from the combined ranking or multivariate ranking (Picard et al. 2021). Depending on the supervised learning methods applied, some of the other multi-omics integration strategies could in principle integrate data without pretreatment (Picard et al. 2021), but many of the above datapretreatment steps are likely to benefit almost any strategy.

Complex disease etiology can't, by definition, be reduced to simple mechanisms as in, for example, monogenic diseases. Similar clinical phenotypes may also be lumped under the same disease classification, despite possibly exhibiting different etiologies. Moreover, disease classifications are often based on spurious historical circumstance relating to social and economical factors (Scully 2004) and symptomaeology (Johansson et al. 2023). Doing research on such disease definitions may not be optimal for development of diagnostics and treatments (Johansson et al. 2023). Unsupervised learning, especially in a multi-omics context, may advance the understanding of how complex disease phenotypes arise (Johansson et al. 2023). Disease conceptualization is a central topic in philosophy of disease. Realists posit that diagnoses are natural kinds, fundamental units of inference (Watson 2023). Given the complex nature of diseases such as diabetes, a more nuanced approach is to use data from currently available experimental technologies to pragmatically reconceptualize the disease to further health outcomes. This approach views diseases as constructed groupings of unwell people (Watson et al. 2023). The question then becomes: how do we best group patients to improve their health? Ethics aside, sample clustering methods can help in answering the above question by integrating omics data, clinical data such as treatment response and other data to classify patients into disease subtypes (Johansson et al. 2023). The late integration strategy is not suited for sample clustering in multi-omics, as sample clusters would constitute separate findings across omics modalities. Instead, sample clustering

is often preceded by PCA, factor analysis, autoencoders or some other unsupervised dimensionality reduction method (Picard et al. 2021). The disease subtypes or wholly new conceptualizations of disease can then be used to generate new hypotheses for mechanistic research and treatments (Johansson et al. 2023).

2.5 Reproducibility

Reproducibility of scientific results is implicit in the science: research credibility hinges on being able to reproduce findings (Fidler et al. 2021). This requires capturing all information relevant for reproducing the results, that is provenance (Kanwal et al. 2021). The starting point for reproducing research may vary due to practical reasons. In research involving substantial computation, the notion of reproducibility is often restricted to the computations alone (Fidler et al. 2021). This is practical from the perspective of open science, as reproduction of computational results can often be done on a personal computer provided sufficient provenance.

As such, raw data is a natural starting point for reproducing of research involving computation. Indeed, "no raw data, no science" is a slogan of sorts in addressing reproducibility issues (Miyakawa et al. 2020). The editor of Molecular Brain found that upon demanding raw data to be made available, over half of submitted manuscripts are withdrawn from the publication process (Miyakawa et al. 2020). Making raw data available may limit malpractice, arising from incentives which do not always align with scientific rigor. Omission of data may be due to ethical considerations, but given public funding of research and the resolution of technical constraints, the limited availability of raw data is largely unfounded (Miyakawa et al. 2020). This was underlined by an unsuccessful attempt to find raw data with specifications suitable for this thesis. Making raw data available may support the development of best practices (Goodacre et al. 2007).

Even with sufficient provenance, analyses may not be exactly reproducible. Even using the same seed, slightly different results can be observed. Such discrepancies may result from instability of computational methods, for example embedded variable selection (Rinaudo et al. 2016).

Reproducibility supports the broader notion of replicability, denoting substantiation of single studies by triangulation from a growing literature, which may include other experimental approaches. It has been argued that the so-called replication crisis is to a large part due to overconfidence in statistical results (Amrhein et al. 2019): all models are wrong, but some are useful. Scientific generalizations need to be based on cumulative knowledge rather than on a single study, the replicability of which can be improved by quality standards (Fidler et al. 2021), perhaps best practices. As such, replicability may be thought of as research quality. One source of replication problems is the multiplicity of data analysis strategies, where different conclusions can be drawn from the same data depending on the computational methods applied (Fidler et al. 2021). This epistemic uncertainty is not addressed by uncertainty metrics like confidence intervals which apply in the context of specific methods (Hoffmann et al. 2021). In the context of supervised learning, many models may have similar performance, but yield different interpretations, for example feature importance scores. This is called the Rashomon effect (Breiman 2001). There are several approaches to addressing the variability of results, although it seems like such methods have not been adopted in metabolomics. In other disciplines methods to address the variability of results include specification curve analysis, multimodel ensembles and bayesian model averaging (Hoffmann et al. 2021).

2.5.1 Bioconductor

Bioconductor is a centralized repository for high-quality open research software, which can be conceptualized as consisting of data containers, computational methods and a community of users and developers. Interoperability is a of central focus in Bioconductor: computational methods are preferrably contributed such that they can be used flexibly according to different research needs (Gentleman et al. 2004), like Lego bricks. In Bioconductor, the interoperability continuum is apparent in package functions not relying on a rigid pipeline of functionality and the manipulation of data using data containers. Using the same data container is more interoperable than having to separate and/or create new data structures in an analysis workflow. Given the emphasis on interoperability in Bioconductor, redundancy of functionality is discouraged in Bioconductor.

Software compatibility underlies interoperability, and is largely concerned with dependencies. Dependencies may be thought of as code from other parties required to make code work correctly. Managing dependencies is not a trivial task, and may result in frustration referred to as “dependency hell”. Bioconductor delivers releases consisting of a set of compatible R packages intended for compatibility within a certain version of R (Gentleman et al. 2004). This makes analyses more provenant, as simply the R version and Bioconductor packages used can suffice with regard to reporting the software used for the analysis. There is also a quality aspect to using Bioconductor: the Bioconductor team manually tests the packages and there are requirements for documentation and testing of functionality (Gentleman et al. 2004).

When manipulating complex data, it needs to be abstracted. Abstraction, in general, is the process of reducing something to a set of essential elements. Data containers meet this description in simplifying the representation of data, while hiding its complexities and associated operations (Morgan et al. 2023). For example, instead of storing feature and sample metadata in separate tables and accessed by extensive scripting, metadata is stored in the same container instance and accessed by user-friendly operations. Bioinformatics projects typically require a data container with a count matrix, sample descriptions and annotation functionality (Morgan et al. 2023). Data containers make functionality more interoperable and easier to grasp for readers as compared to extensive scripting, promoting reproduction of the analysis. This is especially true if a single container is used for the analysis. Container converters are handy if several data containers are needed in the analysis. Data containers are also useful for comprehensive reporting of results, uncertainty and experimental details in a single instance. For example, metrics pertaining to assumptions of univariate tests can be included in the feature metadata for inspection of high-ranking features.

2.5.2 Quarto

Quarto is an open-source computational document system which generates an output file of desired format based on text and code input (Allaire et al. 2022). Provenance is improved using Quarto, as figures and results can output directly to the rendered document.

Much of the proposed minimum reporting standards from DAWG are covered by the text and code in a Quarto document, especially with informative commenting of code. The reproducibility advantages of computational document systems have been acknowledged by the scientific community, for example in Bioconductor where use of computational documents is required for documentation. IMoreover, the eLife journal accepts such so-called executable research articles. This is in stark contrast especially to results arrived at using point-and-click software, where the analysis is often not recorded and results supposedly gained as per the description of the analysis are accepted on good faith (Kanwal et al. 2017). Technologies have been developed to record the analyses performed using point-and-click software in formats which can be run from the command line, although these often present limitations as to changing of parameters, for example (Kanwal et al. 2017).

As such, it would be optimal to implement workflows fully programmatically, allowing for readers to easily reproduce the results tinker with the analysis to assess the variability of results. Using Quarto and Bioconductor packages for the analysis makes this prospect especially tractable as simply the Bioconductor version and packages can be stated. This thesis itself can be generated in a streamlined manner and is amenable to users with minimal programming experience. Moreover, providing the code in a computational document allows for flexible re-use of code, further facilitating development and dissemination of best practices.

Other technologies which increase the reproducibility of analyses include virtual environments and containerization software (Kanwal et al. 2017). Virtual environments facilitate the recording and sharing of dependencies. Containerization software additionally provides a stable machine image, which counters platform-specific variability of results (Kanwal et al. 2017).

3 Research objectives

Notame may be implementable with Bioconductor given Bioconductor's philosophy of interoperability. It is postulated that untargeted LC-MS metabolomics data analysis can be enhanced by implementing Notame in a reproducible, Bioconductor-compatible workflow. The specific aims are as follows:

1. To explore the extent and interoperability of untargeted LC-MS metabolomics data analysis functionality in Bioconductor
2. To operationalize Bioconductor and Quarto in a Bioconductor-compatible workflow spanning Notame
3. To discuss departures from Notame and other limitations in the Bioconductor-compatible workflow

Addressing the above aims, substantive research efforts may be better equipped to meet the increasing demands on reproducibility and quality, as per the Research Council of Finland's strategy contributing to the renewal, quality and societal impact of science. Moreover, Bioconductor may offer functionality accommodating varied research needs and encourage the development of best practices in Bioconductor.

4 Materials and methods

4.1 Metabolomics functionality in Bioconductor

Packages with functionality relating to LC-MS metabolomics data extraction, data pretreatment, feature selection and biological context under “metabolomics” and “mass spectrometry” in Bioconductor’s BiocViews categorization infrastructure were scrutinized. In other words, packages relating to Notame were included, excluding packages concerned with technical replicates, batch effect and more. Packages that explicitly specified focus on the closely related field of proteomics were excluded to limit the scope of the overview, although they may include relevant functionality as well. Packages utilizing solely original data containers were excluded as it quickly became apparent that they reflect non-interoperable functionality, although interoperability is hard to quantify. Deprecated packages as of Bioconductor version 3.18 were also excluded. Statistical tests, due to their general nature, are included in many packages across modalities, and were not explored systematically in the BiocViews categorization infrastructure. The same holds to a lesser extent for visualizations. Data containers in Bioconductor were also compared non-systematically.

Gaps in Bioconductor functionality were complemented with code in the spirit of not including redundant functionality in Bioconductor, and was made available in GitHub (Suksi 2024). The complementary code was mostly modified from Notame and relies on minimal non-Bioconductor dependencies. The non-Bioconductor dependencies are not concerned with computations which could affect the results. As such, the complementary code can be considered Bioconductor-compatible. Complementary code was written with interoperability and informative naming practices in mind. The Notame documentation is largely applicable.

4.2 Example analysis

The balanced data features 72 fetal brain, intestine and placentae samples from 24 fetuses collected at euthanization of six specific pathogen-free (SPF) and six germ-free

(GF) mouse dams just before expected delivery. The data was already collected and preprocessed with MS-DIAL, as per Notame, when the author received it. As such, the example analysis was restricted to data pretreatment and feature selection, the extent of the Notame R package and the functionality covered by the TreeSummarizedExperiment (TSE) data container (see Results section). Restricting to a single data container also makes sense since the Notame best practices are targeted towards new users which may lack programming proficiency. For demonstration purposes, only the first 1000 features included in the data from the HILIC chromatographic column in positive ionization mode was included to reduce execution time and allow for comprehensive visualization of QC and results. The substantive aim of the example analysis was also formulated for demonstrative utility, namely ranking features distinguishing between GF and SPF membership in intestine tissue. In accordance with the emphasis on reproducibility, the example analysis, including implementation details and instructions for rendering, was made available in GitHub (Suksi 2024). The biosigner (Rinaudo et al. 2016), MAI (Dekermanjian et al. 2022), mia (Ernst et al. 2023), phenomis (Thevenot 2023), pmp (Jankevics et al. 2023), POMA (Castellano-Escuder et al. 2021), qmtools (Joo et al. 2023), scater (McCarthy et al. 2017) and QFeatures (Gatto et al. 2023) Bioconductor packages were used in the example analysis.

Low-quality features, having a non-parametric RSD of < 0.2 and non-parametric D-ratio of < 0.4, were flagged. Features detected in less than 70% of the QC samples were removed. QC visualizations were drawn before and after drift correction, excluding flagged features, to inspect the effect of drift correction and as exploratory data analysis.

Features were temporarily nlog-transformed with an offset of one to better meet the assumptions of the smoothed cubic spline model used to normalize for drift. Unfortunately, the drift correction method almost identical to the one in Notame resulted in severely inflated values for many features in an unpredictable manner, so a similar drift correction method by Dunn et al. (2013) was adopted instead. Low-quality features were then flagged anew to inspect the results of drift correction and exclusion of low-quality features.

To complete the abundance matrix, missing values were classified as MCARs/MARs and MNARs and imputed using random forest and a linear regression-based single

imputation method, respectively. QC samples were removed before imputation to prevent bias. Imputation was first performed for quality features, followed by imputation of low-quality features. Features were then clustered to reduce the redundant representation of the same metabolite. Each cluster was aggregated by the sum of the cluster in each sample. The resulting features were named and flagged in the order they appeared in the dataset, assuming that highly correlated features are of similar quality.

Samples of all tissue types were needed for drift correction and QC despite the substantive aim of the example analysis only pertaining to intestine tissue. The imputed and clustered set was used without further normalization, transformation or scaling for univariate analysis. A non-parametric test, Mann-Whitney U-test, was used as there were only 12 intestine samples per study group. P-values were adjusted using the Benjamin-Hochberg false discovery rate (FDR) approach to correct for multiple testing where, on average, a significance threshold of 0.05 for twenty tests would give a false positive for one test.

Before supervised learning, the dataset was nlog-transformed to reduce heteroscedasticity and skewness and normalized for dilution with PQN. Random forest was then used for selecting a subset of features with respect to binary classification performance on GF/SPF membership, emphasizing a minimal, stable signature.

The supervised signature was not ranked, but tiered. Features within tiers were ranked according to fold-change to obtain supervised ranks. The univariate and supervised ranks were then combined and forced to increments of one, with any ties resolved simply by the order in which they appeared in the dataset. Features excluded from the supervised signature were ranked separately, and forced to rank after the combined ranks.

5 Results

5.1 Bioconductor is ripe with relevant functionality

5.1.1 Data containers

The MetaboSet container used in Notame is a Bioconductor base package ExpressionSet legacy container derivative. ExpressionSet is designed for array-based experiments and use with only one count matrix per instance (Gentleman et al. 2004) (Table 1). This may necessitate creation of new instances to handle transformations and other data manipulation tasks.

Table 1. Comparison of ExpressionSet and TreeSummarizedExperiment functionalities (Morgan et al. 2023, Huang et al. 2021). Basic data manipulation tasks such as subsetting were excluded from the comparison.

Functionality	ExpressionSet	TreeSummarizedExperiment
Interactive visualization	Yes	Yes
Multiple count matrices	No	Yes
On-disk option	Yes	Yes
Package compatibility	Yes	Yes
Update instance	Yes	Yes
Validity check	Yes	Yes
Hierarchical structure	No	Yes
Range representation	No	Yes
Data pairing	No	Yes
Additional metadata slots	No	Yes
Alternative feature sets	No	Yes
Low-dimensional representation	No	Yes

Although many metabolomics packages interface with ExpressionSet and are largely compatible with MetaboSet, new developments in the fast-moving field of reproducible

computation in Bioconductor increasingly lean on the modern SummarizedExperiment (SE) family of containers. SE is also based on the ExpressionSet class, but is more flexible, highly optimized and can store multiple count matrices in a single instance (Morgan et al. 2023). The TSE derivative adds to the functionality of SE by facilitating storage and manipulation of the hierarchical structure of the data as per the phyloseq package for exploring microbiome profiles (Huang et al. 2021). Moreover, since TSE is derived from SE via RangedSummarizedExperiment and SingleCellExperiment, TSE also includes functionality for representing ranges, storing data pairings as well as addition of further metadata fields, alternative feature sets and low-dimensional representations (Huang et al. 2021). TSE instances are compatible with functions using SE, but the reverse is not always true because of the additional functionality of TSE. Possible applications of the additional functionalities of TSE are as follows:

- The hierarchical structure functionality could be used to track and aggregate feature clusters.
- The range functionality could be used for subsetting by the genomic coordinates of enzymes producing specific metabolites.
- The data pairing functionality could see use in tracking the same individual in intervention studies.
- Addition of further metadata fields may be of use in separating experimental information, details on the lab, associated publications and parameters used in creating alternative feature sets.
- Storage of alternative feature sets can be relevant for storage of a subset of features selected on basis of the statistical analyses, as the assays slot doesn't accommodate assays with differing numbers of features.
- Low-dimensional representation functionality could see use in storage of results from dimensionality reduction methods, for example the principal coordinates, feature loadings and additional factor-level information from PCA.

More technical functionalities, including on-disk implementation option for large datasets, compatibility with packages designed for use with parent containers, updating of instance class and checking the validity of the object are featured in both ExpressionSet and TSE (Morgan et al. 2023, Huang et al. 2021).

Other data containers for metabolomics, chiefly a suite of data containers included in the RForMassSpectrometry initiative, show promise but do not match the interoperability of ExpressionSet and TSE. RForMassSpectrometry containers, including Spectra, QFeatures and MsExperiment, are currently almost exquisitely focused on preprocessing and are only beginning to support analysis of quantitative features. Other containers are of very specific use, limited to a single package or stage in analysis, or do not match the interoperability of ExpressionSet or TSE, as becomes apparent in later sections.

Regarding multi-omics support, MultiAssayExperiment is considered standard in Bioconductor, is inspired by SE and allows for differing numbers of samples and features in a true multi-omics fashion (Ramos et al. 2017). The MSexperiment integrative data container from the RForMassSpectrometry initiative shows promise in being based on MAE, but only one slot is available for storage of data from other modalities (Rainer 2023). Indeed, MSexperiment is not designed for multi-omics analysis per se, but for the handling various aspects of preprocessing and data pretreatment (Rainer 2023).

5.1.2 Packages

Of the 18 packages scrutinized for data pretreatment and feature selection functionality, 10 supported SE, with three packages supporting ExpressionSet (Table 2). Although difficult to quantify, the packages supporting SE also stood out with regards to interoperability.

Table 2. Rough categorization of packages providing data pretreatment and feature selection functionality, that is the span of the Notame R package, in Bioconductor. Basic data structures like data.frames, matrices and tibbles are not mentioned.

Package	Data container	Stage
MatrixQCVis	SE	Quality control
MsQuality	MsExperiment, Spectra	Quality control
MAI	SE	Imputation
NormalyzerDE	SE	Normalization
vsclust		Feature clustering
MSPrep	SE	Data pretreatment

Package	Data container	Stage
phenomis	SE, ExpressionSet	Data pretreatment
pmp	SE	Data pretreatment
POMA	SE	Univariate analysis
qmtools	SE	Data pretreatment
calm		Univariate analysis
SDAMS	SE	Univariate analysis
biosigner	SE, MAE	Supervised learning
ropls	SE, ExpressionSet	Supervised learning
INDEED		Feature selection
limma	ExpressionSet	Univariate analysis
MixOmics		Feature selection
sparsenetgls		Supervised learning
structToolbox		Feature selection
statTarget		Feature selection

Although many methods have been developed to normalize for drift, many methods, including the smoothed cubic spline used in Notame, require the optimization of a smoothing parameter for interpolating the fit to biological samples. The smoothing parameter must be validated by cross validation to minimize modeling the random measurement error in addition to the systematic measurement error, drift. The pmp package provides drift correction with cross-validation for optimization of a smoothing parameter. Moreover, drift correction in the pmp package implements the same smoothed cubic spline drift correction used in Notame. The pmp package also includes functionality for removing features affected by carryover. The pmp package supports SE. Normalization for dilution using PQN normalization can be performed using the qmtools, phenomis and pmp packages. The latter allows use of QC samples to generate a reference spectrum, which is considered more robust.

No packages can be used for quality control as per Notame, but the Notame computations meet the criteria for being considered Bioconductor-compatible set forth in materials and methods and were thus adopted in the complementary code. Promising QC

visualizations along those of Notame were available in the MatrixQCVis package, but lacked interoperability and visual appeal. All QC visualizations in Notame, except for t-SNE, rely on packages meeting the criteria for Bioconductor-compatibility and were adopted in the complementary code. Visualization using t-SNE is available in a variety of packages, including scater and qmtools, supporting SE.

A multitude of SE-supporting packages can be used for random forest imputation. The MAI (Mechanism-Aware Imputation) package features a two-step approach. First, missing values are classified as MCAR/MAR or MNAR, after which random forest imputation is applied to predict MCARs/MARs and single imputation or no-skip k-nearest neighbors is applied to predict MNARs. A small α parameter value from the model, indicating few MNARs, can be expected because of the filtering of values that are not detected in >70% of the samples disproportionally affecting features with MNARs. MAI supports SE.

Feature clustering functionality along the lines of Notame is provided by the cliqueMS and qmtools packages. However, the cliqueMS clustering method relies on parameters calculated in the xcms package workflow, and is thus not too interoperable. If configured to be as analogous to Notame as possible, the qmtools package clustering method stops at filtering by correlation coefficient where Notame starts recursively eliminating features from clusters in an undirected graph. However, the qmtools package offers many options for correlation coefficient, RT based initial grouping and elimination of features from the initial grouping. Aggregation can be performed using the QFeatures package in multiple ways, for example by retaining the sum of the cluster in each sample. The qmtools package supports SE. It is outside the scope of this thesis to systematically review and test feature clustering functionality and the utility of such functionality remains unclear.

Natural log transformation is available in, for example, the mia package and glog transformation is included in the pmp package; both support SE. Autoscaling is also available in, among others, the mia package.

Regarding univariate analysis, Welch's t-test, Mann-Whitney U test, Welch's ANOVA and two-way ANOVA are available in the POMA and phenomis packages, supporting SE. These packages also implement multiple testing correction using the false dis-

covery rate approach. Friedman test is not available in Bioconductor. The POMA package interface to the limma package caters to univariate linear modelling needs for a variety of study designs, and yields empirical Bayes moderated t-statistics and p-values. Although originally intended for microarray transcriptomics, limma is listed under metabolomics in the BiocViews categoriation infrastructure and has been used in the closely related field of proteomics. It is outside the scope of this thesis to assess the suitability of the limma package for metabolomics data analysis. Linear mixed models in limma are not as comprehensive as in the non-Bioconductor lmer or lme4 packages. Correlation tests are available in the POMA, phenomis and qmtools packages, supporting SE, although correlation between different container instances is not supported.

The MUVR package specified for feature selection using supervised learning in Notame is not included in Bioconductor. However, wrapper variable selection designed for biomarker discovery, like the "min" model in the MUVR package functionality, is available in the biosigner package (Rinaudo et al. 2016). Variable selection in the biosigner package is based on a backward procedure in which the significance of feature subsets is estimated by random permutation of the intensities in the test set. The dataset is then restricted to the significant feature subset, and the whole procedure is performed iteratively until all candidate features are found significant or until there are no features left to be tested. First, 50 boot subsets are obtained by bootstrapping. Each subset is split into a training set and a test set. The model is trained on the training set and evaluated on the test set. For each model, the features are ranked using feature importance metrics, after which the ranks over models are obtained by taking the median of the ranks across models. The largest non-significant feature subset is found by half-interval search. If the evaluation for a subset comes out in favor of the permuted test set in over 5% of all boot subsets, the subset is declared non-significant. If the feature subset is found non-significant, the next candidate feature is chosen by determining the rank closest to the mean of the latest significant feature rank -1 and the non-significant feature rank + 1. If the feature is found significant, the next candidate feature is chosen by determining the rank closest to the mean of the previously scrutinized rank - 1 and the latest non-significant feature rank + 1. The half-interval search arrives at a progressively more restricted set of features until all features are found significant, constituting

the final signature, or until there are no more features to be tested for significance. The biosigner functionality returns tiers of features instead of ranks, where features constituting the likely biomarker signature (S-tier) passed all iterations of the half interval search. Features in the other tiers were discarded in previous iterations. A disadvantage with the biosigner package is that only binary classification is supported.

The only other Bioconductor packages with wrapper variable selection using PLS-DA or random forest is ropls and MixOmics, featuring PLS-DA. Only the MixOmics package features multilevel classification. There are many packages for obtaining feature importance metrics like VIP-values for PLS-DA or out-of-bag error rate for random forest without wrapper variable selection. No packages seem to provide class-specific feature importance. Other supervised learning methods such as support vector machines are also available.

In addition to the Bioconductor functionality presented above, many functionalities are also included in low-level packages for the RForMassSpectrometry packages as interoperable functions for working with matrices and data frames. Results visualizations are also available in many packages, but lack the thoroughness and visual appeal of the Notame visualizations which were adopted in the complementary code. Notame recommends visualizing the total abundance of features across study groups in a complex heatmap using hierarchical or k-means clustering using an online tool, although such visualizations were not available in Notame or Bioconductor. Another deviation from Notame is that the pmp package drift correction functionality does not return values for the fit spline, used to visualize drift correction for high-ranking features.

Central packages for preprocessing in R/Bioconductor include the RForMassSpectrometry suite of packages and xcms, but these don’t support SE. Similarly, functionality relating to biological context is available in Bioconductor, but SE support is limited (Table 3).

Table 3. Rough categorization of packages providing preprocessing, annotation and pathway analysis functionality in Bioconductor.

Package	Data container	Stage
cosmiq	xcmsSet	Preprocessing
IPO	xcmsSet	Preprocessing
ncGTW	xcmsSet, xcmsRaw	Preprocessing
MassSpecWavelet		Preprocessing
MetaboCoreUtils		Preprocessing
MetCirc	Spectra	Preprocessing
MetaMS	N/A	Preprocessing
MSCoreUtils		Preprocessing
Msnbase	MSnExp, Spectrum	Preprocessing
msPurity	xsAnnotate, XCMSnExp, xcmsSet	Preprocessing
QFeatures	QFeatures	Preprocessing
Spectra	Spectra	Preprocessing
XCMS	XCMSnExp	Preprocessing
yamss		Preprocessing
Rdisop		Annotation
CAMERA	xsAnnotate	Annotation
clumsID	Spectrum	Annotation
compoundDb	Spectra	Annotation
MetaboAnnotation	SE, QFeatures	Annotation
Rdisop		Annotation
BioNetStat		Pathway analysis
FELLA		Pathway analysis
graphite		Pathway analysis
metapone		Pathway analysis
MWASTools		Pathway analysis
pathview		Pathway analysis
rWikiPathways		Pathway analysis
SBGNView		Pathway analysis

5.2 Example analysis demonstrates Bioconductor-compatible workflow

The number of low-quality features (342) remained unchanged after drift correction. Linear models relating each feature's abundance to injection order were fit to visualize the effect of drift correction by drawing histograms of the p-values for the regression coefficient of the models. After drift correction, the frequencies of p-values optimally follow the uniform distribution represented by a horizontal line, indicating that there is no global relationship between injection order and feature abundance (Breheny et al. 2018). The results indicate that unwanted variation from drift is reduced in that the frequency of p-values under 0.05 were reduced and more features populate the expected uniform distribution resulting from biological variance (Figure 3). The change is probably barely discernible because of the multitude of study groups and tissue types.

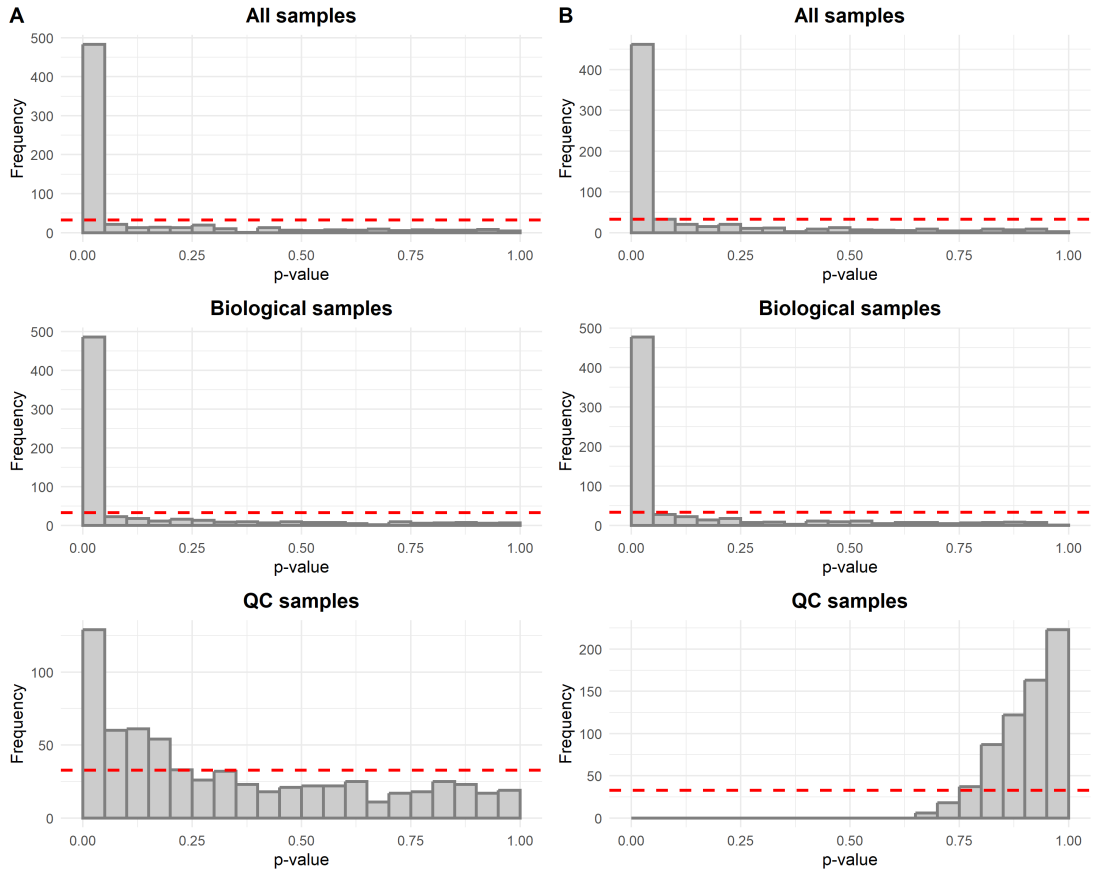

Figure 3. P-values from linear regression models relating each feature to injection order. The dashed red lines represent the expected uniform distribution. A) Before drift correction, featuring all samples, biological samples and QC samples. B) After drift correction, featuring all samples, biological samples and QC samples

No marked differences are seen across histograms of all samples and biological samples as the effect of the seven QC samples is limited in comparison to the 72 biological samples. The histograms for the QC samples alone best illustrate the effect of drift correction. Before drift correction p-values tend towards the lower end because of systemic drift. After drift correction the predictor, injection order, is globally less associated with the response, intensity, resulting in higher p-values. The uniform distribution is not populated, probably because of the majority of features violating some assumptions of linear regression.

To visualize systematic drift in global feature intensities across samples, boxplots representing the distribution of all features' abundances in each sample were drawn. In addition to being insensitive to outliers, quartiles preserve information about center and spread (Krzywinski et al. 2014). Such boxplots often show a systematic decrease or increase in signal intensity as a function of injection order, which should be reduced after drift correction. Visually speaking, there is no systematic increase or decrease in global feature intensity across the samples before or after drift correction (Figure 4). This could be because of a small amount of drift and the features experiencing drift in different directions, cancelling each other out on the global feature intensity level.

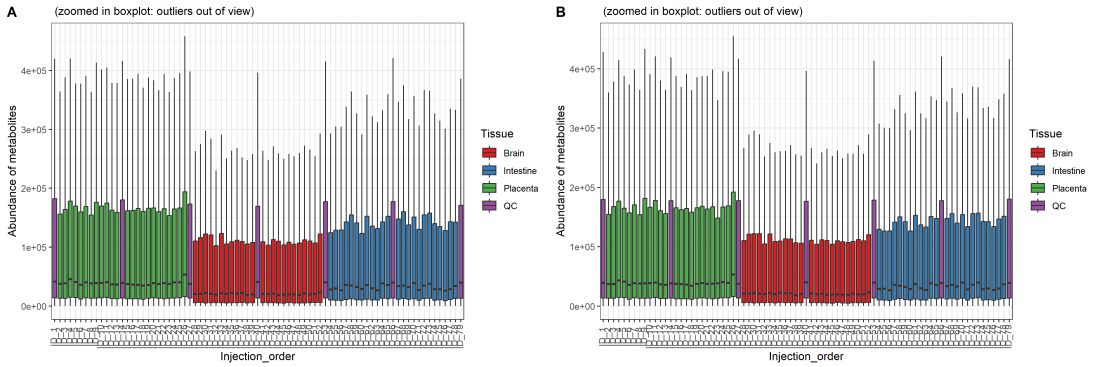

Figure 4. Boxplots representing the feature intensities in each sample in injection order, featuring the median as a black line, the interquartile range as a box and the 1.5x the interquartile range as whiskers. A) Before drift correction. B) After drift correction.

Feature variation was globally assessed using Euclidean distances between samples using density plots. Drawing such density plots before and after drift correction hopefully shows how Euclidean distances are reduced after drift correction, especially for the QC samples which optimally group independently of the biological samples as they include only random measurement error and experimental variation (Figure 5). However, drift

correction does not seem to have resulted in a reduction of Euclidean distances of features between samples.

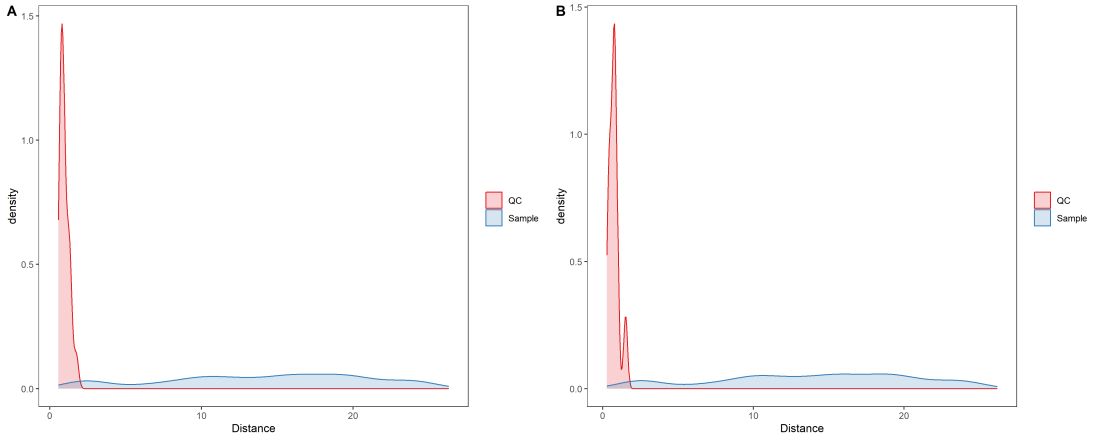

Figure 5. Density plot of Euclidean distances between samples. A) Before drift correction. B) After drift correction.

A dimensionality reduction technique, t-distributed stochastic neighbor embedding (t-SNE) was applied to visualize sample patterns in the data according to study group and QC sample membership. T-SNE can separate non-linearly separable data. Trends in the biological samples may not be apparent before or after drift correction, but the QC samples should group more tightly after drift correction. Samples can also be colored by injection order, where after drift correction, trends should dissipate. t-SNE separated the data for visualization according to tissue membership, although there is no marked difference before and after drift correction (Figure 6). The tissue groups are distinct, consisting of two somewhat distinct subgroups representing GF and SPF membership. GF/SPF membership is least well separated in intestine tissue. The QC samples group tightly.

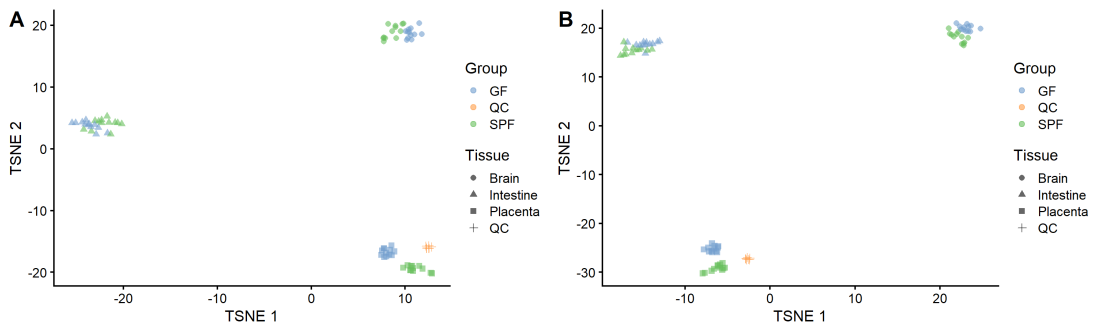

Figure 6. t-SNE plots of samples, shape by study group and color by tissue type. A) Before drift correction. B) After drift correction.

Finally, hierarchical clustering using Ward's criterion on Euclidean distances between samples was used to visualize sample clusters in a dendrogram, where the QC samples should cluster together earlier after drift correction. More distinct clusters corresponding to study groups/time points indicate higher quality after drift correction. Hierarchical clustering using Ward's criterion and PCA are complementary unsupervised learning approaches answering similar questions from different perspectives since both operate in Euclidean space (Murtagh et al. 2014). Hierarchical clustering offers higher resolution of the relationships between samples (Murtagh et al. 2014), which could reveal clusters of samples which exhibit a different metabolic response to treatment. This may guide the research and prompt further questions relating the clusters to clinical variables, for example. Herein, there are very small differences before and after drift correction concerning the clustering of the placenta and QC samples (Figure 7).

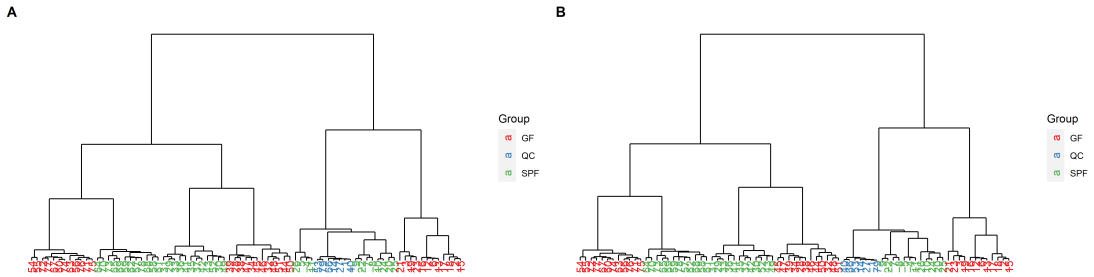

Figure 7. Dendrograms of hierarchical sample clusters using Ward's criterion on Euclidean distances between samples. A) Before drift correction. B) After drift correction.

The same clustering methodology was used for heat maps. In such a heatmap, clusters of samples with similar metabolic patterns as well as groups of discriminating metabolites that drive sample clustering can be identified (Benton et al. 2015). The Euclidean distance between samples can be expected to reduce as variation from systematic drift is reduced, resulting in more pronounced blocks of the study groups/time points and QC samples. Herein, the QC block pattern is slightly darker after drift correction (Figure 8). Tissue-wise, the effect of drift correction is negligible in the heatmap coloration. The brain and intestine blocks appear rather uniform within study groups, while the placenta block seems to consists of sub-blocks, the origin of which remains unclear.

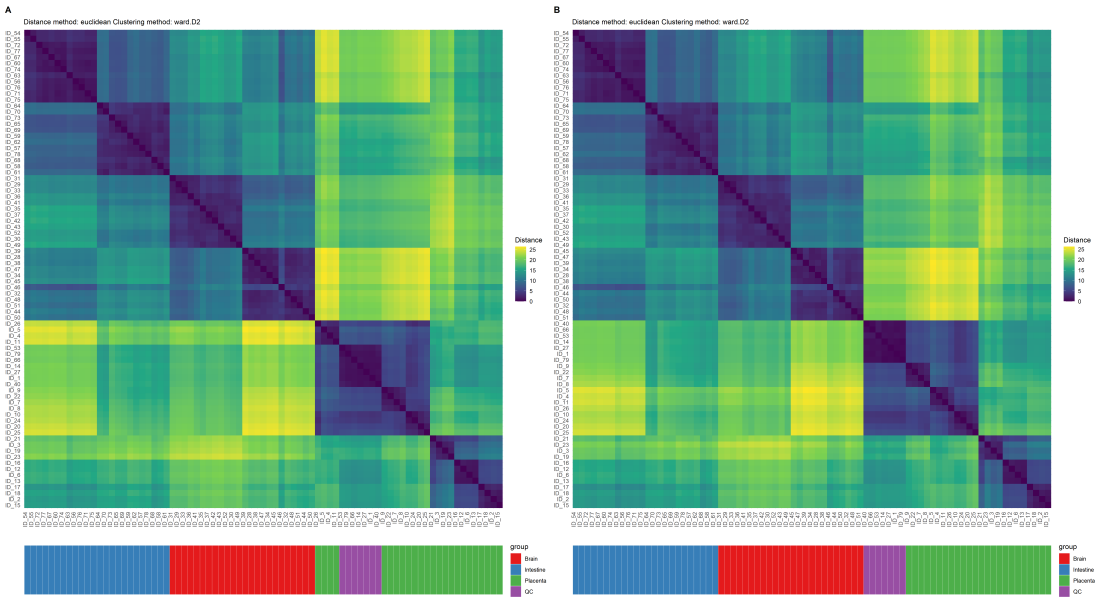

Figure 8. Heatmaps of hierarchical sample clusters using Ward's criterion on Euclidean distances between samples. A) Before drift correction. B) After drift correction.

There were a total of 312 missing values before imputation, which were all imputed using mechanism-aware imputation to complete the abundance matrix. This approach differs from Notame, in which the random forest imputation assumes that all missing values are MCARs. In mechanism-aware imputation, the alpha parameter value was five, indicating a small number of MNARs originating from the detection threshold. Thus missing values were predominantly imputed using random forest imputation, with the relatively few MNARs imputed using single imputation.

Feature clustering somewhat reduced the number of features, with 998 before and 738 features after feature clustering. The shared part of the feature clustering method differed from Notame in that instead of Pearson's correlation, Spearman's correlation was used. Moreover, the Notame feature clustering method returns the abundances of the feature with the highest median abundance in each cluster. Herein, the sum of the features in a cluster for each sample was returned.

The random forest model could classify GF and SPF intestine samples perfectly irrespective of including all features, only S-tier features or S- and A-tier features in the model. A single feature constituted the S-tier, which ranked fourth in the Mann-Whitney U test and first in the combined ranks. A total of eight features were included in the S, A, B, C and D tiers corresponding to a hierarchy of subsets found significant in the variable selection process.

Due to the methodological nature of this thesis, feature-wise visualization is limited to the most high ranking feature. In substantive research, comprehensive feature-wise inspection of a subset of high-ranking features would be undertaken. As choice of results visualizations depends on the study design, the complementary code is referred to for visualizations not applicable for the analysis performed herein, that is feature-wise plots for time-series data and PCA and PLS-DA score plots. Manhattan plots and cloud plots would be drawn separately for each mode.

The highest ranking feature suffered minimally from drift (Figure 9), although in substantive research it may be of interest to inspect the drift pattern for all features included into a single entry by clustering.

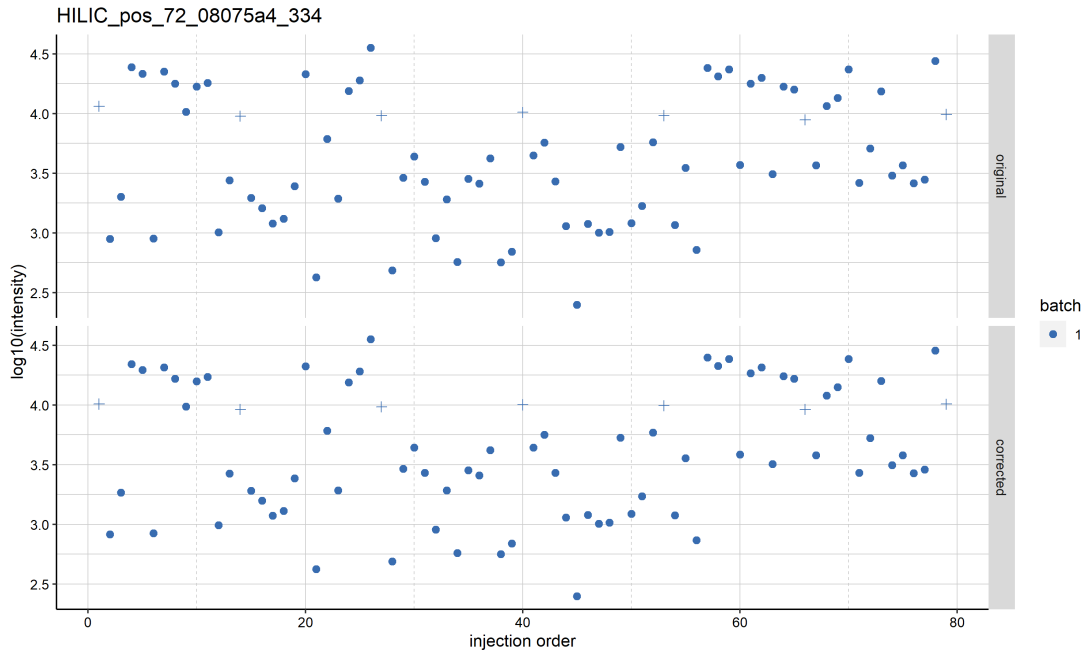

Figure 9. Drift correction for the highest-ranking feature in log10 space. A) Before drift correction. B) After drift correction.

Feature-wise plots are useful for inspecting select ranked features to compare distributions and differences in feature levels and identify outliers. The highest ranking feature in intestine tissue shows a large difference in abundance and distribution, with minimal outliers (Figure 10).

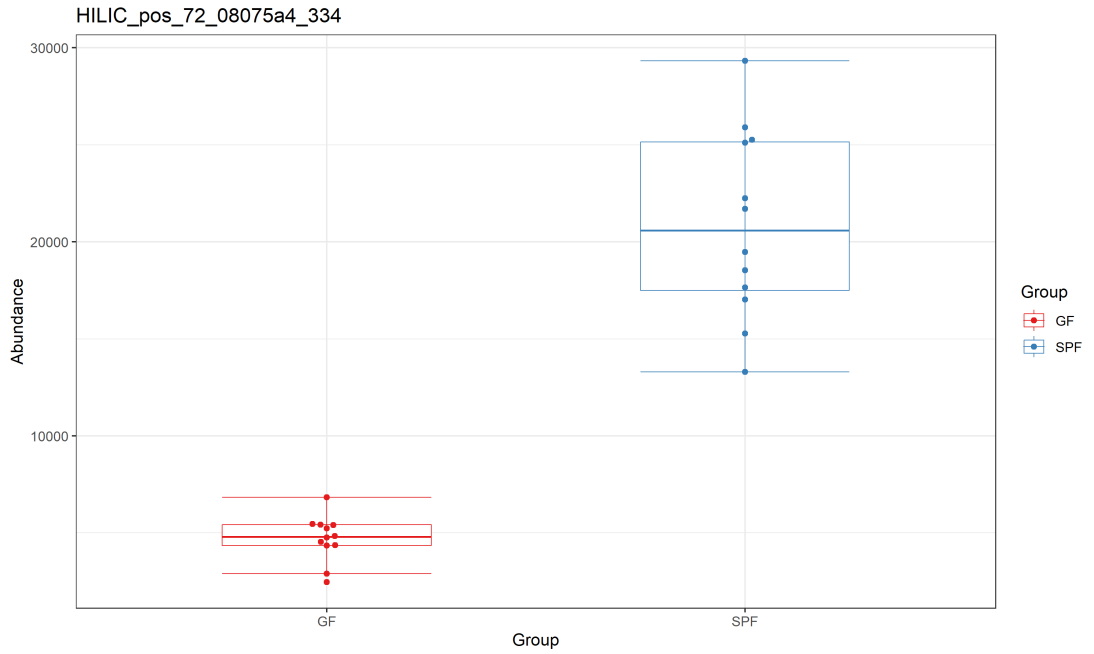

Figure 10. Beeswarm plot of the abundance of the highest ranking feature across study groups in intestine tissue. The mean is represented by a horizontal line, with hinges and whiskers representing the interquartile range and max 1.5x the interquartile range, respectively.

With regards to comprehensive visualization, dimensionality reduction using t-SNE separated the clustered dataset according to study group (Figure 11).

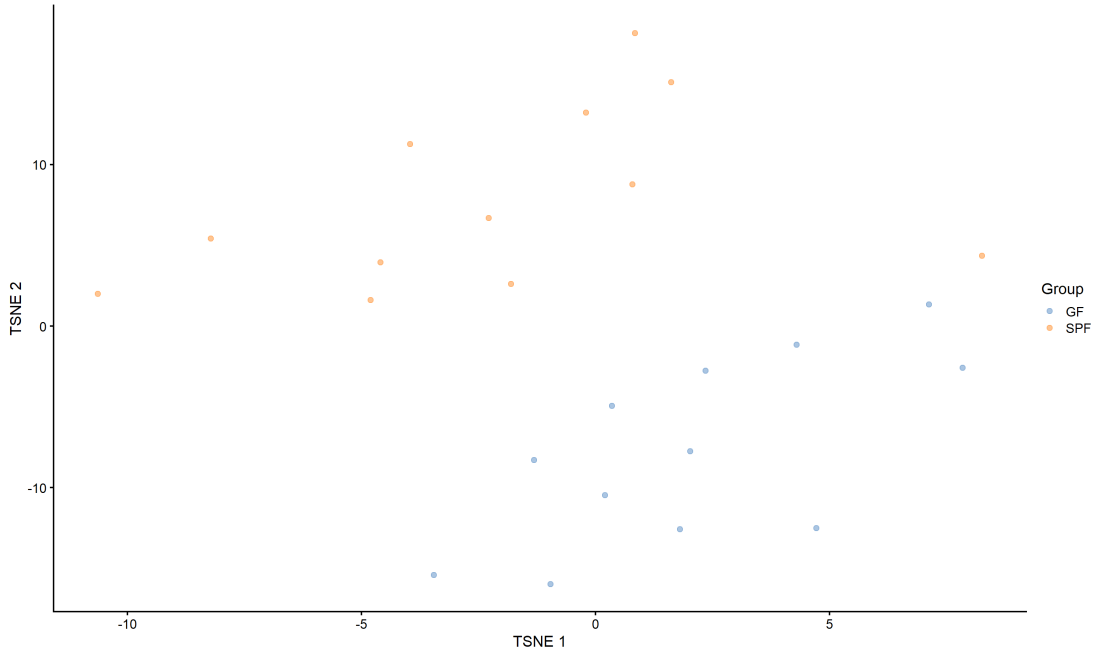

Figure 11. t-SNE plot of all samples in the dataset. Shape by tissue and color by group.

To globally assess the univariate results, a histogram depicting the distribution of p-values from Mann-Whitney U tests testing the difference in feature abundances be-

tween study groups in intestine tissue was used (Figure 12). The results indicate that there is a true difference in global feature levels between GF and SPF mice, since the distribution clearly deviates from the uniform distribution.

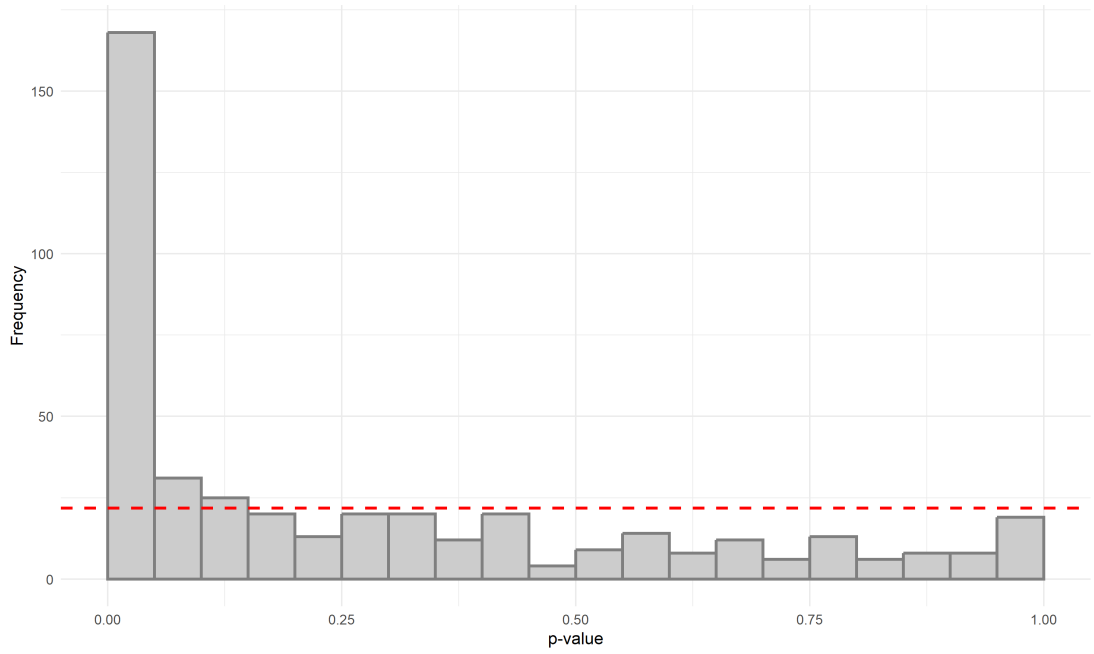

Figure 12. The distribution of p-values from Mann-Whitney U tests testing the difference in feature abundance between study groups in intestine tissue. The dashed red lines represent the uniform distribution, namely that there is no difference in feature abundance between the study groups.

To comprehensively assess the results from univariate tests and supervised learning, a volcano plot was used (Figure 13). Although Notame recommends coloring by the multivariate ranking, the features were colored by combined rank as coloring by the biosigner tiers would exclude almost all features from coloration. Eight top-ranking features included in the biosigner signature were labeled, making apparent the different perspectives that univariate and multivariate analyses have on the data. The volcano plot indicates that features more abundant in SPF intestine tissue have large fold changes in comparison to features less abundant in SPF intestine tissue, especially features with very small p-values. Small p-values seem to be somewhat disproportionately represented by features more abundant in SPF intestine tissue.

Figure 13. Volcano plot of p-values (negative log10 scale) from Mann-Whitney U tests testing the difference in feature abundances between study groups in intestine tissue against fold changes between study groups (log2 scale). Positive and negative fold changes are equidistant from the center (x = 1) of the x-axis. The features are colored by combined ranking.

Relating biochemical characteristics to molecular features was done using Manhattan plots, where features with low p-values and low ranks seem rather evenly distributed with regards to average m/z (Figure 14A). The same holds true for RT (Figure 14B), although the plot appears truncated by RT because of only including the first 1000 features in the analysis. Nonpolar metabolites elute first from HILIC columns.

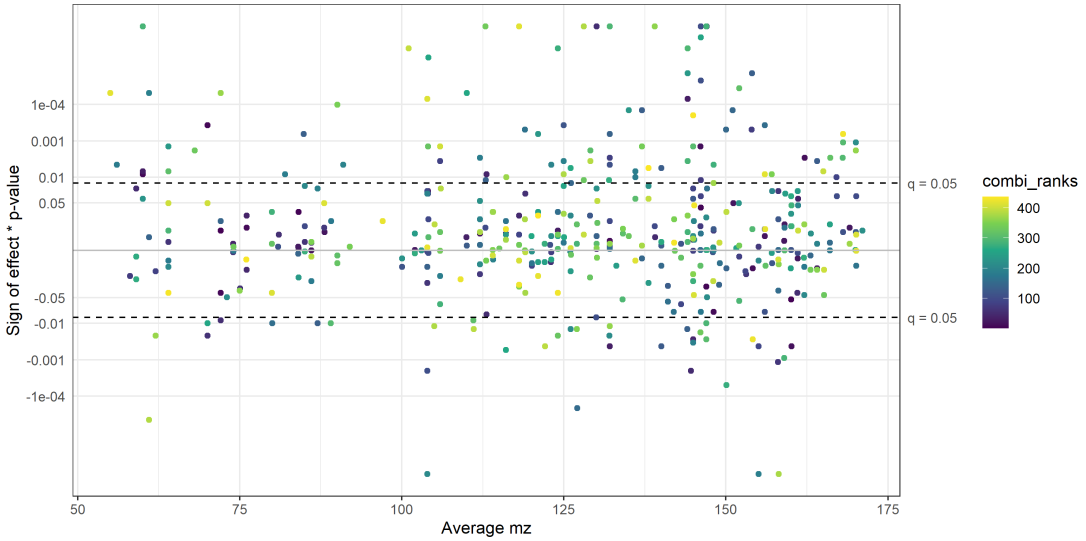

Figure 14A. A directed Manhattan plot of p-values from Mann-Whitney U tests testing the difference in feature abundances between study groups in intestine tissue with m/z of the features as x-axis. The points are colored by combined ranks.

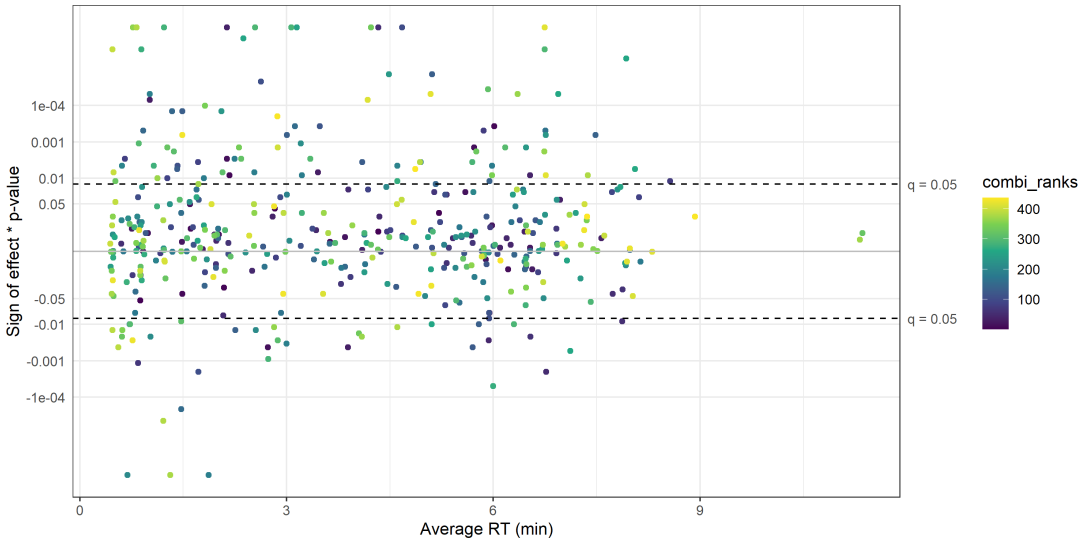

Figure 14B. A directed Manhattan plot of p-values from Mann-Whitney U tests testing the difference in feature abundances between study groups in intestine tissue with RT of the features as x-axis. The points are colored by combined ranks.

To visualize interesting features with regards to m/z, RT, p-value and combined ranks, in effect combining the data in the two Manhattan plots but without sign of effect, a cloud plot was used (Figure 15). The points are also commonly colored by fold-change (Benton et al. 2015). Again, trends are hardly discernible.

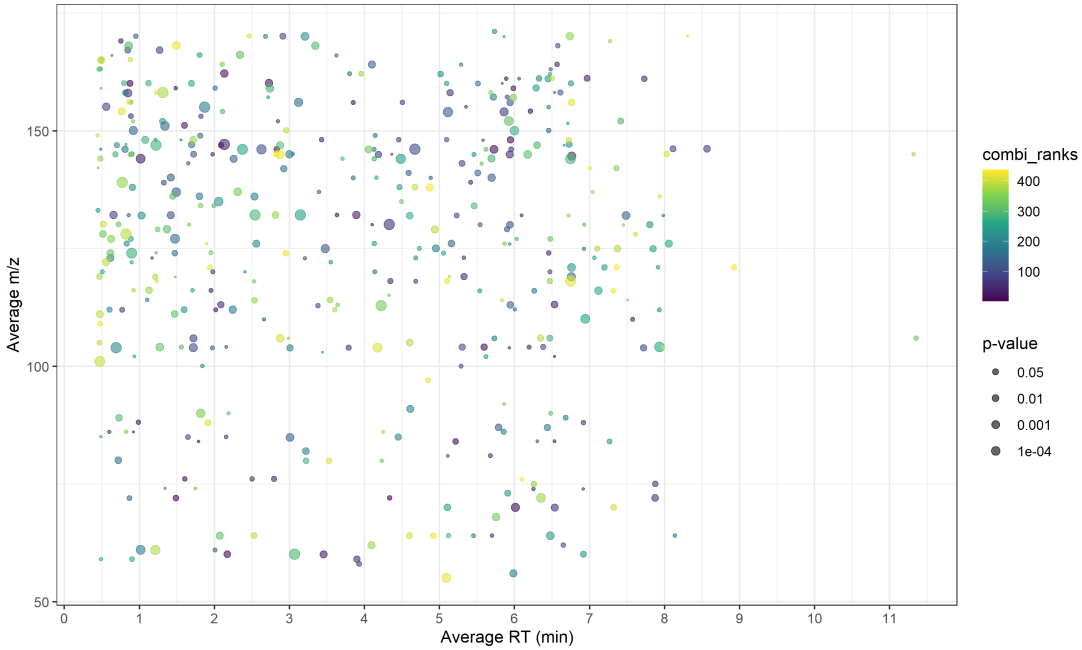

Figure 15. Cloud plot relating the m/z and RT of molecular features in intestine tissue. The size of the points reflect p-values from Mann-Whitney U test, testing the difference in feature abundances between study groups. The points are colored by combined ranking.

6 Discussion

As a relatively new technology, LC-MS research could do well with increased quality and reproducibility, which was attempted here by operationalizing Notame, Bioconductor and Quarto in a Bioconductor-compatible workflow. From a substantive perspective, this would allow science to converge more efficiently on how the metabolome relates to physiology and pathology alike. This, in turn, could inform quality targeted research, ultimately resulting in diagnostics and treatments.

Preprocessing and functionality related to biological context could not be implemented using TSE. It has come to the authors attention that a SE plug-in to MS-DIAL is being explored, which could streamline a Bioconductor-compatible workflow further by MS-DIAL outputting a SummarizedExperiment object directly. Then again, especially preprocessing packages are very developed with unique data containers, and apart from interoperability, it is not obvious what SE could bring to the table.

With the complementary code, data pretreatment could be implemented using TSE as per Notame with minimal deviations. Unfortunately, the QSRSC function in the pmp package, which is almost identical to the drift correction method used in Notame, produced severely inflated values for many features in an unpredictable manner, so a similar drift correction method by Dunn et al. (2013) in the qmtools package was adopted instead. Dunn et al. (2013) did not find a notable performance difference to the method used herein. Imputation using mechanism-aware imputation using random forest for MCARs and single imputation for MNARs is a good example of how Bioconductor can cater to evolving best practices. The third type of missing value, MAR, was not dealt with explicitly, but was rather imputed as an MCAR (Dekermanjian et al. 2023). The clustering method differed from Notame in that the feature clustering used herein ended at filtering of features by correlation in an undirected graph for each retention time window. The correlation method also differed. One can speculate usage of the non-parametric Spearman's correlation reduces the need for the recursive clustering used in the Notame method as modeling the monotonic relationship of a feature across samples may better distinguish between features originating from the same metabolite, especially if the correlation threshold is optimized. However, neither Notame nor the

qmtools package specifies a systematic approach for optimizing the parameters, so it is difficult to compare the methods in a dispassionate manner. Moreover, there is scant discussion in the literature on this approach to feature clustering. Another difference to Notame was the choice of abundance values to represent the clusters. Retaining the sum of the clusters for each sample seems reasonable and is not obviously a downgrade from Notame, where the feature with the highest median abundance is retained. The variety of options for eliminating features from the initial grouping in the qmtools package could prove useful in development and wide adoption of such feature clustering methods.

Feature selection functionality, as per Notame, is somewhat lacking in Bioconductor, especially with regards to linear mixed models and supervised learning. Repeated measures designs of various randomizations are prevalent in metabolomics studies to assess an intervention, necessitating flexible linear mixed models. Flexible linear mixed models or generalized linear mixed models could prove a useful contribution to Bioconductor, especially using SE or TSE. Biosigner was chosen for supervised learning in the example analysis for its emphasis on a restricted, stable set of features for biomarker discovery (Rinaudo et al. 2016). A restricted set of features for pathway analysis, as per the "mid" model in the MUVR package (Shi et al. 2019), could prove a useful contribution to Bioconductor, especially with multilevel classification, minimally biased variable selection, class-specific variable importance and SE-support.

The lack of some functionality in Bioconductor may reflect the lack of a need, putting into question the rationale of a Bioconductor-compatible workflow. From this perspective, a Bioconductor-compatible workflow may seem like an exercise in futility. This is supported in that the Bioconductor-compatible workflow using TSE was not very streamlined due to different coding practices in packages, especially compared to the Notame R package. The TSE lineage-specific functionalities were also not of obvious utility herein, except for addition of alternative feature sets. If Bioconductor released guidelines for coding practices, perhaps separately for modalities, packages might be written to leverage the philosophy of interoperability and functionalities of data containers. Moreover, the availability of relevant functionality supporting TSE in Bioconductor is promising for development of best practices. One must also not overlook Bioconductor's provenance, quality control and documentation.

Regarding provenance and reproducibility, using Quarto covers the reporting standards set forth by DAWG with ease. The utility of Quarto would be emphasized in a fully programmatic workflow including preprocessing and functionality related to biological context. However, the feasibility of tracking provenance so meticulously in a research context is arguable, as there is a substantial amount of work involved in preparing the documents, especially if the research article itself is rendered using Quarto. Some functionality is also missing in Quarto, such as alternative text for figures. Assessing the variability of results in support of replicability may also not be tractable due to the considerable, overnight execution times on a personal computer. Although the research objectives were largely met, this exploration, operationalization and discussion of using Notame, Bioconductor and Quarto points to issues with feasibility.

7 Conclusion

In the spirit of quality and reproducible science, Notame, Bioconductor and Quarto were operationalized in a Bioconductor-compatible workflow spanning data pretreatment and feature selection, the extent of TSE functionality. Feature selection presents limitations of the Bioconductor-compatible workflow. Notame reflects a consensus in the literature, but LC-MS metabolomics data analysis is developing quickly, especially with regards to feature selection. Given Bioconductor's philosophy of interoperability, Bioconductor could accommodate varied research needs and the further development of best practices. However, the rationale of a Bioconductor-compatible workflow is not evident. The purported interoperability of Bioconductor packages proved lacklustre, so the Bioconductor-compatible workflow is not as streamlined as expected. The provenance, quality control and documentation of Bioconductor are of more obvious utility, promoting reproducibility. Reproducibility is also promoted by using Quarto, although research culture would have to shift considerably to normalize sharing of data and tracking of provenance so meticulously. The same holds for assessing the variability of results in support of replicability.

As such, it is not straightforward to assess the significance of this thesis in a dispassionate manner, although it is safe to say that it promotes reproducible, open science as per the Research Council of Finland's strategy contributing to the renewal, quality and societal impact of science. Discounting the issues raised above, this demonstration of the utility of Notame, Bioconductor and Quarto may facilitate quality substantive research efforts in tackling question relating the metabolome to complex, biological systems in a reproducible fashion. This could ultimately translate to health outcomes at large. In accordance with the emphasis on reproducibility, the thesis along with instructions for rendering was made available in GitHub (Suksi 2024).

8 Främjande av oriktad LC-MS metabolomikdataanalys med Bioconductor

Utveckling av högeffektiva experimentella tekniker har lagt grunden för forskning i komplexa, biologiska system med hjälp av sofistikerade dataanalyser. Analys av metabolomet, det vill säga de småmolekyler som påträffas i ett biologiskt prov, är speciellt lovande eftersom metabolomet utgör den nivå i molekylärbiologins centrala dogma som mest direkt påverkar fenotypen. Därför är metabolomik centralt för såväl diagnos som vård av sjukdomar med komplex etiologi för vilka orsaksförhållanden inte kan reduceras till exempelvis en enda gen. Metabolomikdata kan även inkluderas i multi-omikanalyser där data från flera olika molekylärbiologiska nivåer används, såsom en TV-apparat med bättre resolution.

Oriktad vätskekromatografi-masspektrometri (eng. untargeted liquid chromatography-mass spectrometry, LC-MS) är en central experimentell teknik i metabolomik, där man strävar efter relativ kvantifiering av alla metaboliter i ett prov. I LC-MS separeras först metaboliterna i ett prov enligt fysisk-kemiska egenskaper i en kromatografikolumn, vartefter metaboliterna joniseras för kvantifiering i en masspektrometer. Jonerna kan även fragmenteras för att åtskilja metaboliter som eluerats ur kromatografikolumnen samtidigt. För att förbättra elueringen av metaboliter med varierande fysisk-kemiska egenskaper genomförs datainsamlingen ofta med flera kromatografikolumner, vartefter rådatan bearbetas för att erhålla data med alla metaboliter som kunde kvantifieras i provet. Datan är vid detta lag inte ännu redo för statistisk analys. Variablerna kan inte antas representera individuella metaboliter, och det förekommer drev i kvantifieringen samt saknande värden. Dessutom måste variablerna genomgå kvalitetskontroll. Före datan är färdig för statistisk analys utförs ännu någon kombination av normalisering, transformation eller skalning beroende på de statistiska metoder som används. Såsom bearbetningen och förbehandlingen, så ock är den statistiska analysen samt den algoritmiska placeringen av resultaten i biologisk kontext extensiv.

Den extensiva dataanalysen utgör en utmaning för reproducerbarhet: den involverade mjukvaran måste rapporteras och fungera för att resultaten kan bestyrkas av tredje

parter. Att utföra dataanalysen med ett programmeringsspråk som R är fördelaktigt, eftersom analysen kan då i princip köras om mycket lätt. Å andra sidan används R vanligtvis med en mélange av olika så kallade paket, varav det finns flera versioner och grader av uppdatering och kvalitetskontroll. Detta komplicerar både utförandet av dataanalysen samt reproducering eftersom paketen ofta inte är kompatibla med varandra, eller kan ge olika resultat beroende på version. Bioconductor, ett slags centraliserat arkiv för R-paket, kommer årligen ut med uppsättningar av interkompatibla paket som genomgår kvalitetskontroll. Dessutom är det lätt att rapportera dataanalysen om man använt Bioconductor-paket eftersom versionerna av paketen inte behöver specificeras; istället rapporterar man helt enkelt Bioconductor-versionen och de paket man använt. Reproducerbarhet kan främjas ytterligare med hjälp av beräkningsdokument, där kod, programmatiska resultat såsom figurer och text blandat framställs i ett önskat format, t.ex. som en pdf-fil. Detta försäkrar att de resultat man presenterar säkert härstammar från de rapporterade dataanalysstegen. Dessutom kan en tredje part manipulera dataanalysens parametrar i beräkningsdokumentet och framställa resultaten för att forma en bild över hur robust forskningen är.

En ytterligare utmaning i anslutning till den extensiva dataanalysen är bristen på brett tillämpade goda praxis, vilket leder till forskningsresultat av varierande kvalitet. Rapporteringsstandarder för oriktad LC-MS dataanalys har rekommenderats av Metabolomics Standards Initiative. För bästa praxis har ingen instans gett rekommendationer, men en bra början finns i Notame R-paketet, associerat med en vetenskaplig artikel från en serie artiklar som uttryckligen fokuserar på bästa praxis för oriktad LC-MS dataanalys.

I detta pro gradu-projekt använder jag mig av Bioconductor, beräkningsdokumentmjukvara vid namn Quarto och bästa praxis från Notame för att demonstrera en exempeldataanalys som förespråkar för reproducerbarhet och ökad kvalité av oriktad LC-MS dataanalys. Detta är i linje med öppen vetenskap som är en av de strategiska faktorer Finlands Akademi identifierat för att främja vetenskapens förnyelse, kvalitet och samhälleliga genomslag. Dessutom kan projektet utgöra en startpunkt för fortsatt utveckling av goda praxis i oriktad LC-MS dataanalys med hänsyn till reproducerbarhet.

Det var möjligt att implementera alla programmatiska steg, dvs. Notame R-paketet, i Bioconductor med minimala avvikelser. Dessutom kunde allt implementeras med

en enda databehållare, SummarizedExperiment. Notame R-paketet omfattar inte alla steg i oriktad LC-MS dataanalys: bearbetning av rådata och algoritmisk placering i biologisk kontext rekommenderas göras med hjälp av "peka-och-klicka"-mjukvara i Notame-artikeln, trots att även dessa steg finns tillgängliga i Bioconductor.

Oriktad LC-MS används ofta för utforskande forskning som inte betyngs av fasta hypoteser. Detta innebär att hypotestestning i oriktad LC-MS inte handlar om att modellera biologiska faktum, utan att generera hypoteser för bekräftande forskning. Således är dataanalysens kvalitet inte nödvändigtvis lika kritisk som i bekräftande forskning, men påverkar i förlängningen bekräftande forskning. I och med projektet är oriktad LC-MS bättre anpassat för att generera hypoteser av hög kvalitet på ett reproducerbart sätt. Detta kunde bidra till realisering av de förhoppningar vetenskapssamfundet har för maskininlärningsbaserad metabolomik- och multi-omikforskning, speciellt gällande diagnostik och vård av komplexa sjukdomar såsom Alzheimers, diabetes och cancer.

9 References

Allaire et al. (2022). Quarto, version 1.2. https://github.com/quarto-dev/quarto-cli (accessed in January 2024).

Amrhein et al. (2019). Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. The American Statistician, 73: 262-270.

Ardrey (2003). Liquid Chromatography – Mass Spectrometry: An Introduction. John Wiley & Sons.

Berg et al. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics, 7: 142-157.

Benton et al. (2015). An Interactive Cluster Heat Map to Visualize and Explore Multidimensional Metabolomic Data. Metabolomics, 11: 1029-1034.

Breheny et al. (2018). p-Value Histograms: Inference and Diagnostics. High Throughput, 7: article 23.

Breiman (2001). Statistical modeling: The Two Cultures. Statistical Science, 16: 199-215.

Broadhurst et al. (2018). Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. Metabolomics, 14: article 72.

Broeckling et al. (2014). RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. Analytical Chemistry, 86: 6812-6817.

Castellano-Escuder et al. (2021). POMAShiny: A user-friendly web-based workflow for metabolomics and proteomics data analysis. PLOS Computational Biology, 17: 1-15.

Dekermanjian et al. (2022). Mechanism-aware imputation: a two-step approach in handling missing values in metabolomics. BMC Bioinformatics, 23: article 179.

Dieterle et al. (2006). Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. Analytical Chemistry, 78: 4281-4290.

Dunn et al. (2013). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nature Protocols, 6: 1060-1083.

Erngren et al. (2019). Adduct formation in electrospray ionisation-mass spectrometry with hydrophilic interaction liquid chromatography is strongly affected by the inorganic ion concentration of the samples. Journal of Chromatography A, 1600: 174-182.

Ernst et al. (2023). mia: Microbiome analysis, version 1.10.0. https://bioconductor.org/packages/mia (accessed in January 2024)

Fidler et al. (2021). Reproducibility of Scientific Results. The Stanford Encyclopedia of Philosophy (Summer 2021 Edition).

Fiehn (2002). Metabolomics – the link between genotypes and phenotypes. Plant Molecular Biology, 48: 155-171.

Gatto et al. (2023). QFeatures: Quantitative features for mass spectrometry data. https://bioconductor.org/packages/QFeatures (accessed in January 2024)

Gentleman et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome biology 5: article 80.

Gika et al. (2019) Untargeted LC/MS-based metabolic phenotyping (metabonomics/metabolomics): The state of the art. Journal of Chromatography B, 1117: 136-147.

Goodacre et al. (2007). Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics, 3: 241-241.

Goodacre et al. (2020). Data Analysis Standards in metabolomics. https://github.com/MSI-Metabolomics-Standards-Initiative/CIMR/blob/master/cimr-DA%20standardsVer2.pdf (accessed in January 2024).

Grace and Hudson (2016). Processing and Visualization of Metabolomics Data Using R. Intech.

Grissa et al. (2016). Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data. Frontiers in Molecular Biosciences, 3: article 30.

Guida et al. (2016). Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics, 12: article 93.

Hendriks et al. (2011). Data-processing strategies for metabolomics studies. Trends in Analytical Chemistry 30: article 10.

Hoffmann et al. (2021). The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. Royal Society Open Science, 8: article 201925.

Huang et al. (2021). TreeSummarizedExperiment: a S4 class for data with hierarchical structure. F1000 Research, 9: article 1246.

Iterson et al. (2009). Relative power and sample size analysis on gene expression profiling data. BMC Genomics, 10: article 439.

Jafari et al. (2019). Why, When and How to Adjust Your P Values? Cell Journal, 20: article 4.

Jankevics et al. (2023). pmp: Peak Matrix Processing and signal batch correction for metabolomics datasets. https://bioconductor.org/packages/pmp (accessed in January 2024)

Jauhiainen et al. (2014). Normalization of metabolomics data with applications to correlation maps. Bioinformatics, 30: 2155-2161.

Johansson et al. (2023). Precision medicine in complex diseases—Molecular subgrouping for improved prediction and treatment stratification. Journal of Internal Medicine, 294: article 4.

Johnson et al. (2016). Metabolomics: beyond biomarkers and towards mechanisms. Nature Reviews Molecular Cell Biology, 17: 451-459.

Joo et al. (2023). qmtools: Quantitative Metabolomics Data Processing Tools, version 1.6.0. https://bioconductor.org/packages/qmtools (accessed in January 2024).

Kanwal et al. (2017). Investigating reproducibility and tracking provenance – A genomic workflow case study. BMC Bioinformatics, 18: article 337.

Kirwan et al. (2022). Quality assurance and quality control reporting in untargeted metabolic phenotyping: mQACC recommendations for analytical quality management. Metabolomics, 18: article 70.

Klåvus et al. (2020). "Notame": Workflow for Non-Targeted LC–MS Metabolic Profiling. Metabolites, 10: article 135.

Kohl et al. (2012). State-of-the art data normalization methods improve NMR-based metabolomic analysis. Metabolomics, 8: 146-160.

Kokla et al. (2019). Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: A comparative study. BMC Bioinformatics, 20: article 492.

Krzywinski et al. (2014). Visualizing samples with box plots. Nature Methods, 11: 119-120.

Kvalheim et al. (1994). Preprocessing of Analytical Profiles in the Presence of Homoscedastic or Heteroscedastic Noise. Analytical Chemistry, 66: 43-51.

Kärkkäinen et al. (2021). Changes in the metabolic profile of human male postmortem frontal cortex and cerebrospinal fluid samples associated with heavy alcohol use. Addiction Biology, 26: e13035.

Lahti et al. (2021). Orchestrating Microbiome Analysis with R and Bioconductor. https://microbiome.github.io/oma/ (accessed in October 2023).

Lewis et al. (2016). Development and Application of Ultra-Performance Liquid Chromatography-TOF MS for Precision Large Scale Urinary Metabolic Phenotyping. Analytical Chemistry, 88: 9004-9013.

Lin et al. (2012). A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. Journal of Chromatography B, 910: 149-155.

McCarthy et al. (2017). Scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R. Bioinformatics, 33: 1179-1186.

McMaster (2005). LC/MS: A Practical User's Guide. John Wiley & Sons.

Miyakawa et al. (2020). No raw data, no science: another possible source of the reproducibility crisis. Molecular Brain, 13: article 24.

Morgan et al. (2023). SummarizedExperiment: SummarizedExperiment container, version 1.32.0. https://bioconductor.org/packages/SummarizedExperiment (accessed in January 2024).

Murtagh et al. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification, 31: 274-295.

Märtens et al. (2023). Instrumental Drift in Untargeted Metabolomics: Optimizing Data Quality with Intrastudy QC Samples. Metabolites, 13: article 5.

Patti et al. (2013). A View from Above: Cloud Plots to Visualize Global Metabolomic Data. Analytical Chemistry, 85: 798-804.

Pessa-Morikawa et al. (2022). Maternal microbiota-derived metabolic profile in fetal murine intestine, brain and placenta. BMC Microbiology, 22: article 46.

Picard et al. (2021). Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology, 19: 3735-3746.

Rainer (2023). MsFeatures: Functionality for Mass Spectrometry Features, version 1.10.0. https://bioconductor.org/packages/MsFeatures (accessed in October 2023).

Ramos et al. (2017). Software for the Integration of Multiomics Experiments in Bioconductor. Cancer Research, 77: 39-42.

Rinaudo et al. (2016). biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data. Frontiers in Molecular Biosciences, 3: article 26.

Schug and McNair (2003). Adduct formation in electrospray ionization mass spectrometry: II. Benzoic acid derivatives. Journal of Chromatography A, 985: 531-539.

Scully (2004). What is a disease? EMBO Reports, 5: 650-653.

Shi et al. (2019). Variable selection and validation in multivariate modelling. Bioinformatics, 35: 972-980.

Shimadzu. Basics of Liquid Chromatograph-Mass Spectrometry. https://www.shimadzu.com/an/servi
support/technical-support/liquid-chromatograph-mass-spectrometry/index.html
(accessed in January 2024).

Suksi (2024). Enhancing untargeted LC-MS metabolomics data analysis with Bioconductor. https://github.com/vsuksi/Enhancing-untargeted-LC-MS-metabolomics-data-analysis-with-Bioconductor (accessed in January 2024).

Sumner et al. (2007). Proposed minimum reporting standards for chemical analysis. PMC Metabolomics, 3: article 3.

Sysi-Aho et al. (2007). Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics, 8: article 93.

Thenovot (2023). phenomis: Postprocessing and univariate analysis of omics data. https://bioconductor.org/packages/phenomis (accessed in January 2024)

Tsugawa et al. (2015). MS-DIAL: data independent MS/MS deconvolution for comprehensive metabolome analysis. Nature Methods, 12: 523-526.

Veselkov et al. (2011). Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery. Analytical Chemistry, 83: 5864-5872.

Vinaixa et al. (2012). A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. Metabolites, 2: 775-795.

Watson (2023). On the Philosophy of Unsupervised Learning. Philosophy & Technology, 36: article 28.

Waikar et al. (2010). Normalization of urinary biomarkers to creatinine during changes in glomerular filtration rate. Kidney International, 78: 486-494.

Wei et al. (2018). Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. Scientific Reports, 8: article 663.

Xia et al. (2013). Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics, 9: 280-299.

Zulfigar et al. (2023). MAW: the reproducible Metabolome Annotation Workflow for untargeted tandem mass spectrometry. Journal of Cheminformatics, 15: article 32.