Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

Protocol

"NoTaMe": Workflow for non-targeted LC-MS metabolic profiling

Marietta Kokla[1,†,*], Anton Klåvus[1,†,*], Stefania Noerman[1], Ville M Koistinen[1], Marjo Tuomainen[1], Iman Zarei[1], Topi Meuronen[1], Merja R. Häkkinen[2], Soile Rummukainen[2], Ambrin Farizah Babu[1], Taisa Sallinen[1,2], Olli Kärkkäinen[2], Jussi Paananen[3], David Broadhurst[4], Carl Brunius[5], Kati Hanhineva[1,*]

[1] University of Eastern Finland, Department of Clinical Nutrition and Public Health, Kuopio
[2] University of Eastern Finland, School of Pharmacy, Faculty of Health Sciences, Kuopio
[3] University of Eastern Finland, Institute of Biomedicine
[4] Centre for Integrative Metabolomics & Computational Biology, School of Science, Edith Cowan University, Joondalup 6027, Australia; d.broadhurst@ecu.edu.au
[5] Chalmers University of Technology, Department of Biology and Biological Engineering; carl.brunius@chalmers.se

† Equal contribution
* Correspondence: marietta.kokla@uef.fi; anton.klavus@uef.fi; kati.hanhineva@uef.fi

Abstract

Metabolomics analysis generates vast arrays of data, necessitating comprehensive workflows involving expertise in analytics, biochemistry and bioinformatics, in order to provide coherent and high-quality data that enables discovery of robust and biologically significant metabolic findings. In this protocol article, we introduce NoTaMe, an analytical workflow for non-targeted metabolic profiling approaches utilizing liquid chromatography–mass spectrometry analysis. We provide an overview of lab protocols and statistical methods that we commonly practice for the analysis of nutritional metabolomics data. The paper is divided into three main sections: the first and second sections introducing the background and the study designs available for metabolomics research, and the third section describing in detail the steps of the main methods and protocols used to produce, preprocess and statistically analyze metabolomics data, and finally to identify and interpret the compounds that have emerged as interesting.

Keywords: Metabolomics, LC-MS, mass spectrometry, metabolic profiling, Computational Statistical; Unsupervised learning; Supervised learning; Pathway analysis

© 2020 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

1. Introduction

The rapid technical development of instrumentation for biomolecule analysis has opened a wider view on metabolomics analysis than ever before. Due to its very high sensitivity and the ability to concomitantly assess thousands of molecular features, liquid chromatography coupled with mass spectrometry (LC-MS) is making its way as the key analytical tool in the field of discovery-driven metabolic profiling (1–3) The LC-MS platform generates large amounts of signals – biological signals from metabolites, their adducts, fragments, isotopes and instrument noise, thereby necessitating good computational tools to process, analyze, and interpret the data (4,5). Although the data processing solutions for complex metabolomics data are accumulating with increasing speed, they continue to be the bottleneck within the analysis, especially the metabolite identification process (6–8). Starting from the acquisition of data to the identification of metabolites, the metabolic profiling workflow involves numerous steps that require expertise in analytical chemistry, biochemistry, bioinformatics and statistics – thus rendering it impossible to achieve reliably by any click-and-go online tools, but necessitates cooperation of scientists with various backgrounds and expertise.

Firstly, the production of good-quality metabolomics data requires good quality samples originating from studies with meaningful research questions, good sample preparation and know-how in operating MS instruments in order to get out the maximum performance of the sensitive measurements. The acquired data needs to undergo several preprocessing steps, starting from data collection (peak picking), where it is imperative to understand the detection threshold and signal-to-noise ratios of the measurement. This is then followed by a multi-step processing phase involving imputation, normalization, data reduction and clean-up, which determines the quality of the data that goes into the data-analysis, identification and biological interpretation of the results. All of these steps need to follow necessary quality assurance and quality control procedures for reliable outcome of the metabolomics analysis (9,10). Finally, the compounds that have emerged as interesting, potentially differential compounds in the given study setup need to be identified using a combination of automated metabolite identification algorithms and exploration of the raw LC-MS/MS spectral data.

Although the currently proposed non-targeted metabolic profiling workflow is applicable on basically any metabolomics study, it has been developed and utilized mainly on food and nutritional approaches, and therefore examples provided here on the presentation of results are from studies within that field. In fact, food and nutrition sciences encompass a versatile array of research fields, which have adopted metabolomics as one of the most important analytical tools during the past decade (9). For example, metabolic profiling allows a comprehensive analysis of the chemical composition of food and estimating the impact of industrial processing and modifications by gut microbiota.(11,12) Likewise, when assessing the actual health outcomes of certain diets or specific foods, metabolic profiling enables pointing out the areas of metabolism that are reflecting the dietary differences, and especially

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

when the data is correlated with other, traditional clinical variables, it may raise novel hypotheses on the molecular level linkage between diet and health (13–15).

Here, we present analytical workflows suitable for any non-targeted metabolic profiling study in a systematic manner (Figure 1), with a major focus on data-analysis challenges. We also present a new R package, where we have bundled many of the data-analysis tools used in our lab so that they are easy to adopt for other scientists working in the field of metabolic profiling. This includes the pre-processing steps and visualizations in Sections 3.2.2–3.2.5., statistical tests and multivariate models in Section 3.3., as well as the visualizations in Section 3.5.

2. Experimental Design

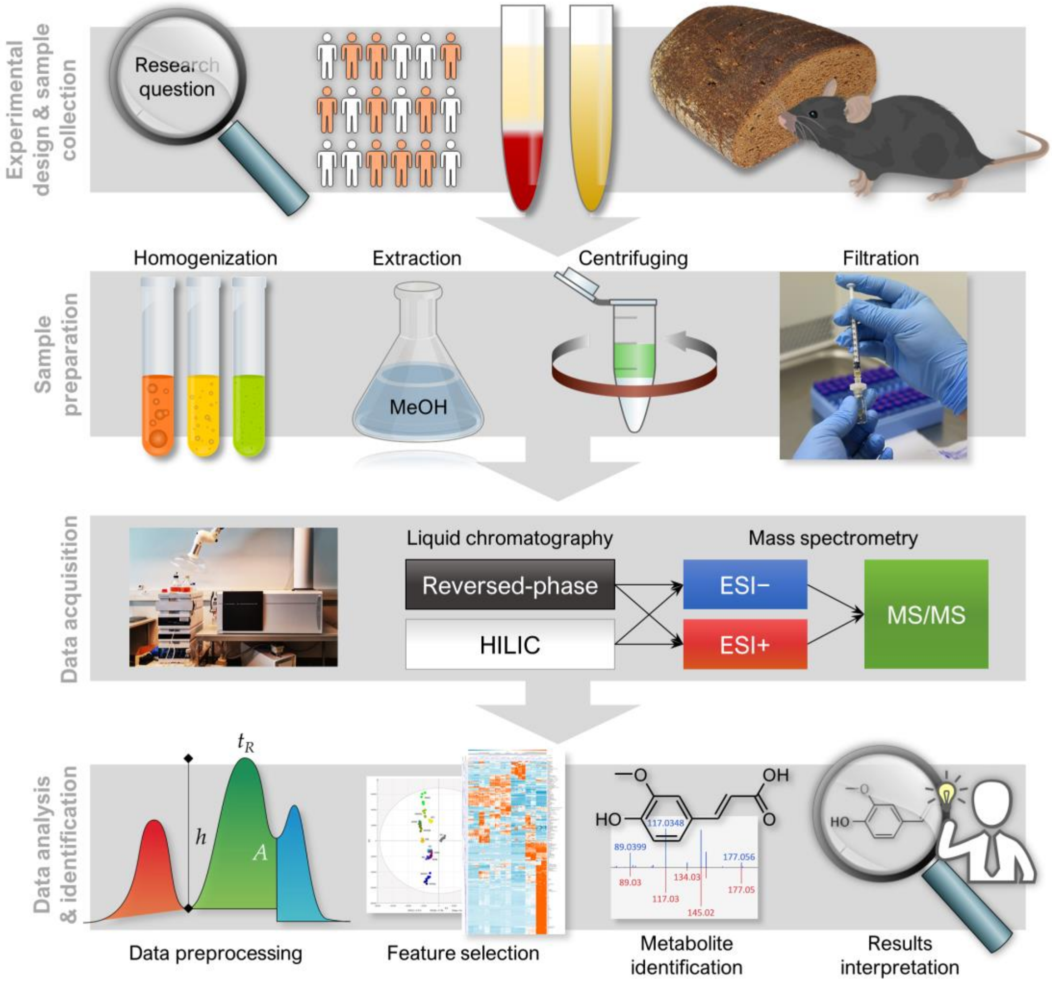

Figure 1: A general overview of NoTaMe workflow containing four important stages; 1.Experimental designs and sample collection, 2. Sample Preparation, 3. Data Acquisition, 4. Data Analysis, and Biomarker Identification Analysis

The non-targeted metabolic profiling analytical workflow presented here includes steps from sample preparation and LC-MS analysis all the way to metabolite identification (Figure 1). It is noteworthy to mention, however, that the study design and careful planning for the sampling are very important part of the study governing the quality of the results and therefore require special attention(9). Herein, we focus on metabolomics analysis performed in one batch (where the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

number of samples typically reaches 200–300 samples). However, the procedures are in general applicable for larger, multi-batch experiments, although extra procedures for quality control are in order (10,16).

2.1. Materials

Sample preparation materials:

a. 96-Well plate (Thermo Scientific, Rochester, NY, USA, Cat.No. 260252), filter plate (Agilent, Santa Clara, CA, USA, Cat.No. A5969002)
b. 96-Well cap mats (Thermo Scientific, Roskilde, Denmark, Cat.No. 276002)
c. Syringe filters (PALL Corporation, Ann Arpor, MI, USA, Cat.No. 4552T)
d. Syringe Norm-Ject® tuberculin 1 ml (Henke Sass Wolf, Tuttlingen, Germany, Cat.No 4010-200V0)
e. Wide orifice pipette tips (Thermo Scientific, Vantaa, Finland, Cat.No. 9405050)
f. Homogenizer microtubes (OMNI International, Kennesaw, GA, USA, Cat.No 19-620

LC-MS materials:

a. Reversed-phase chromatography (RP) column: Zorbax Eclipse XDB-C18, particle size 1.8 µm, 2.1 × 100 mm (Agilent Technologies, Santa Clara, CA, USA, Cat.No. 981758-902).
b. Hydrophilic interaction chromatography (HILIC) column: Acquity UPLC BEH Amide 1.7 µm, 2.1 × 100 mm (Waters Corporation, Milford, MA, USA, Cat.No. 186004801).

Reagents:

a. Acetonitrile, ACN (HiPerSolv CHROMANORM, VWR Chemicals, Fontenay-sous-Bois, France, Cat.No. 83640.320)
b. Methanol, MeOH (CHROMASOLV™ LC-MS Ultra, Riedel-de Haën™, Honeywell, Seelze, Germany, Cat.No. 14262-2L)
c. Formic acid (Optima LC/MS, Fisher Chemical, Geel, Belgium, Cat.No. A117-50)
d. Ammonium formate (CHROMASOLV™ LC-MS Ultra, Honeywell Fluka, Seelze, Germany, Cat.No. 14266-25G)
e. Ultra-pure water (Class 1, ELGA Purelab ultra Analytical, United Kingdom)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

2.2. Equipment

The current workflow is demonstrated with one suitable LC-MS instrumentation and software combination but can likewise employ any other high-accuracy LC-MS setup.

Sample preparation and LC-MS instruments:

a. Centrifuges: For 96-well plates: Heraus Megafuge 40R (ThermoFisher Scientific, Osterode, Germany), for microcentrifuge tubes: Centrifuge 5804R (Eppendorf, Hamburg, Germany)
b. Vortex: Vortex Genie 2 (Scientific Industries, Bohemia, NY, USA)
c. Homogenizer: Bead Ruptor 24 Elite with OMNI BR CRYO unit (OMNI International, Kennesaw, GA, USA)
d. Shaker: Multi Reax (Heidolph, Germany)
e. 1290 Infinity Binary UPLC system (Agilent Technologies, Waldbronn, Karlsruhe, Germany)
f. 6540 UHD accurate-mass quadrupole-time-of-flight mass spectrometer (qTOF-MS) with Jetstream ESI source (Agilent Technologies, Santa Clara, CA, USA)

Software:

a. Agilent MassHunter Acquisition B.07.00 (Agilent Technologies),

b. MS-DIAL version 3.70 (10),

c. MS-FINDER version 3.24 (50),

d. R version 3.5.0 (17)

e. Multiple Experiment Viewer (MeV) version 4.9.0 (http://mev.tm4.org/).

3. Analytical procedure and results

3.1. LC-MS analysis

3.1.1. Sample preparation

Sample preparation for the non-targeted metabolite profiling work aims to extract in a single attempt as wide range of metabolites as possible with, in general, minimal sample workup. Therefore, straightforward, simple extraction protocols are preferred. Protocol 1 is designed for extracting plasma/serum samples at a ratio of 1:5 with ACN and Protocol 2 for extracting homogenized tissue samples at a ratio of 1:6 with 80% methanol.

Protocol 1: Plasma/ Serum Samples

1. Thaw plasma/serum samples in ice water and keep them on wet ice during all the waiting periods.
2. Place the 96-well plate on wet ice for sample preparation and set the filter plate on it.
3. Add 400 µl of cold ACN to the filter plate well.
4. Vortex a plasma/serum sample 10 s at the maximum speed.
5. Add 100 µl of plasma/serum sample to the same well as ACN.
6. For the quality control (QC) samples, add 10 µl of sample to the clean microcentrifuge tube and collect aliquots of all following samples in the same tube.
7. Mix ACN and sample by pipetting four times. Use wide orifice Finn Pipette tips to avoid tip clogging.
8. Repeat steps 1-5 for all samples. Lastly, use the same procedure for the QC sample.
9. Filter the precipitated samples by centrifuging the plate for 5 min at 700 x g at 4 °C.
10. Remove the filter plate and seal the 96-well plate tightly with the 96-well cap mat to avoid sample evaporation.
11. Analyze the samples immediately or store the plate at +4 °C for a maximum of 1 day or at −20 °C until analysis.

Protocol 2: Tissue Samples

12. Weigh a maximum of 300 mg frozen tissue into 2 ml OMNI microtube with beads. Keep the samples on dry ice.
13. Add ice cold 80% methanol in a ratio of 500 µL solvent per 100 mg tissue and keep the tubes on wet ice.
14. **Optional step:** In a case of metabolite-dense sample material (*e.g.* plants), it might be necessary to use a more diluted solvent/sample ratio to avoid analytical problems, such as saturation of the detector or overloading of the column.
15. Homogenize samples with a Bead Ruptor 24 Elite homogenizer. For soft tissues, perform one homogenization cycle at the speed 6 m/s at +/− 2 °C for 30 s.
16. **Optional step:** In a case a homogenizer instrument is not available, manual tissue disruption can be performed using mortar and pestle with liquid nitrogen.
17. Extract the homogenized samples in a shaker for 5 min at RT.
18. Centrifuge samples for 10 min at 20 000 x g at +4 °C.
19. Collect the supernatants on a 96-well filter plate and centrifuge for 5 min at 700 x g at 4 °C.
20. **Optional step:** Filter the samples using solvent resistant syringes and PTFE filters into the HPLC vials.
21. Take aliquots (5–25 µl) of filtered samples and combine into one vial to be used as QC sample in the analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

22. Analyze the samples immediately or store the plate at +4 °C maximum of 1 day or −20 °C until analysis.

3.1.2 LC-MS measurement

The most commonly applied analytical technique in non-targeted metabolic profiling is mass spectrometry, often combined with liquid or gas chromatographic separation at the front end. In order to cover a wide range of polarities among the analyzable metabolites, different chromatographic methods may be utilized, *e.g.* reversed-phase chromatography (RP) and hydrophilic interaction chromatography (HILIC). MS data can then be acquired in both positive (+) and negative (−) electrospray ionization (ESI) polarities.

23. Use the following conditions for RP chromatography: Column oven temperature 50°C, flow rate 0.4 mL/min, gradient elution with water (eluent A) and methanol (eluent B) both containing 0.1% (v/v) of formic acid. Gradient profile for RP separations: 0–10 min: 2 → 100% B; 10–14.5 min: 100% B; 14.5–14.51 min: 100 → 2% B; 14.51–16.5 min: 2% B. Needle wash with 50% ACN. Set the injection volume 2 μL and sample tray at 10 °C.
24. Use the following conditions for HILIC: Column oven temperature 45°C, flow rate 0.6 mL/min, gradient elution with 50% v/v ACN in water (eluent A) and 90% v/v ACN in water (eluent B), both containing 20 mM ammonium formate (pH 3). The gradient profile for HILIC separations: 0–2.5 min: 100% B, 2.5–10 min: 100% B → 0% B; 10–10.01 min: 0% B → 100% B; 10.01–12.5 min: 100% B. Needle wash with 50% ACN. Set the injection volume at 2 μL and sample tray at 10 °C.
25. To operate at high mass accuracy (< 2 ppm), calibrate the MS daily and use the continuous mass axis calibration by monitoring two reference ions from an infusion solution throughout the analytical runs. Examples of reference ions in ESI+ mode: *m*/*z* 121.050873 and *m*/*z* 922.009798, and reference ions in ESI− mode *m*/*z* 112.985587 and *m*/*z* 966.000725.
26. Use the following conditions for Jetstream ESI source: drying gas temperature 325°C and flow 10 L/min, sheath gas temperature 350°C with a flow of 11 L/min, nebulizer pressure 45 psi, capillary voltage 3500 V, nozzle voltage 1000 V, fragmentor voltage 100 V, and skimmer 45 V. Use nitrogen as the instrument gas.
27. For data acquisition, use a 2 GHz extended dynamic range mode in both ESI + and ESI - ionization modes from m/z 50 to 1600. Collect the data in the centroid mode at an acquisition rate of 1.67 spectra/s (i.e., 600 ms/spectrum) with an abundance threshold of 150. For automatic data dependent MS/MS analyses, set the precursor isolation width to 1.3 Da. From every precursor scan cycle, 4 most abundant ions are selected for fragmentation. These ions are excluded after two product ion spectra and released again for fragmentation after a 0.25 min hold. Product ion scan time is based on precursor ion intensity, ending at 25,000 counts or after 300 ms. Use collision-induced dissociation voltage 10, 20, and 40 V in subsequent runs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

28. Generate the worklist containing analytical samples. Inject quality control samples after every 12 samples and in the beginning and end of the analysis. The injection order of samples should be randomized. If the study contains samples from multiple matrices, such as samples from different organs, it is recommended that all the samples of a matrix be injected consecutively, for example first inject all heart samples, followed by all liver samples. If there are multiple samples from the same individual, it is recommended that the samples of an individual are run consecutively. We use an in-house developed software called Wranglr to automatize the process of generating sample worklists. Wranglr automatically randomizes the sample order and adds QC and MS/MS samples. Wranglr is a web application developed with the Shiny package for R (18). Wranglr is published under an open-source software, and can be accessed at github.com/antonvsdata/wranglr.

3.2. Data collection and preprocessing

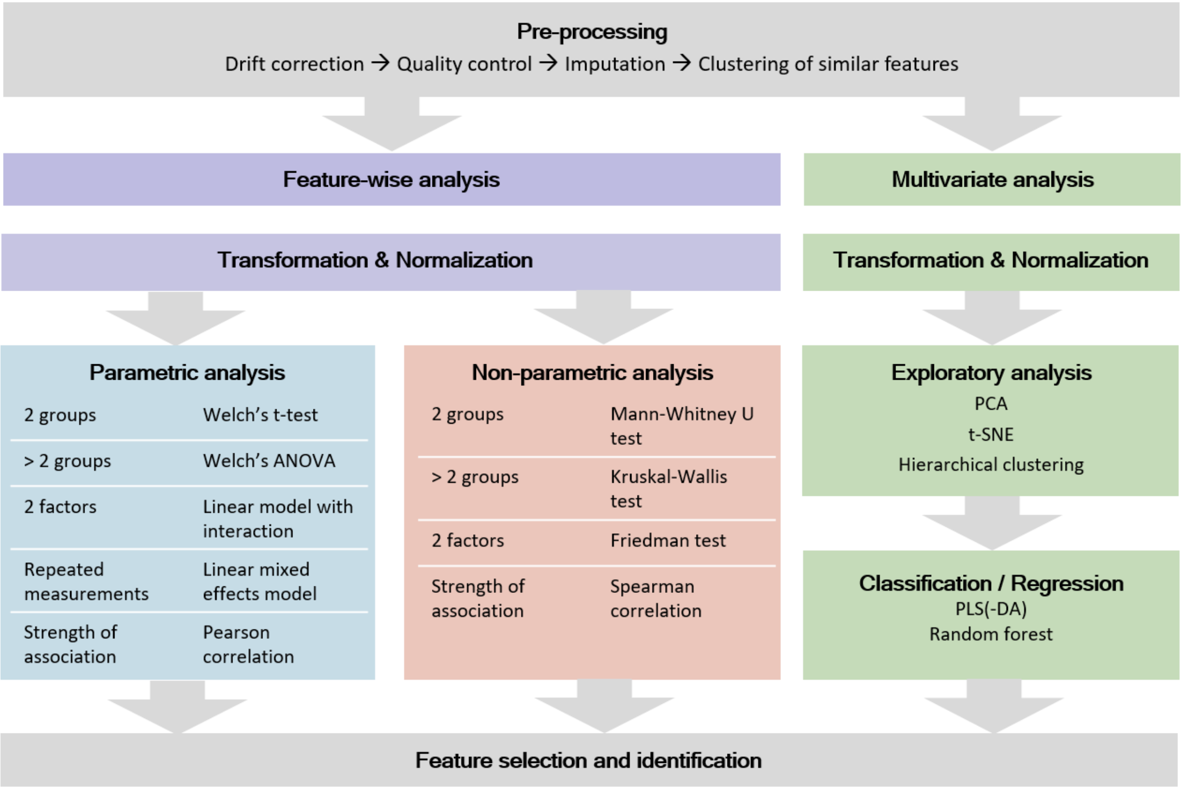

Figure 2. Workflow of the statistical analysis after the peak-picking step. The choices depend on the type of data and the research question of the study design.

The data collection (peak picking) and subsequent preprocessing of the raw data are critical steps in non-targeted metabolomics data-analysis since they determine the quality of the data for all the remaining steps (Figure 2). Various peak picking algorithms exist, utilized by vendor-specific and open-source software as well as freely available online services. In this workflow, MS-DIAL (an open-source software: http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/) (19) is used for the peak picking. After collection of the raw data, pre-processing is required to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

monitor the quality of the data, make any required transformations/corrections to the data, as well as reduce/merge the number of features originating from the same metabolite.

3.2.1. Peak picking and alignment

29. Before the peak picking, convert the raw instrumental data (i.e. *.d) to ABF format using Reifycs Abf Converter (https://www.reifycs.com/AbfConverter). Follow the vendor-specific instructions on the website.
30. For the peak picking in MS-DIAL (version 3.70), choose the following parameters:
    a. *m*/*z* tolerance according to the instrument mass accuracy (for QTOF: 0.01 Da);
    b. minimum peak height 2000 signal counts for QTOF (or at least 5 times the typical noise level of the instrument; 3000 signal counts for highly concentrated plant samples);
    c. mass slice width 0.1 Da;
    d. linear weighted moving average as the smoothing method (smoothing level 3 scans and minimum peak width 5 scans);
    e. in positive mode, select $[M + H]^+$, $[M + NH_4]^+$, $[M + Na]^+$, $[M + K]^+$, $[M + CH_3OH + H]^+$, and $[M - NH_3 + H]^+$ as the adducts and in-source fragments; in negative mode, select $[M - H]^-$, $[M - H2O - H]^-$, $[M + Cl]^-$, $[M + HCOOH - H]^-$, and $[2M - H]^-$ as the adducts and in-source fragments.
31. For the peak alignment, set the retention time tolerance at 0.05 min and MS1 tolerance at 0.015 Da. Set the detection filter (detected in at least one sample group) at 50%. Unselect the “detected in all QCs” option and select gap filling by compulsion.
32. Once the peak picking is finished, export the alignment result as peak areas into a raw data matrix as a tab-separated text file. Transform the data matrix into a datasheet in a spreadsheet software, such as Excel. Insert additional columns to each datasheet specifying the chromatography and the ionization mode before combining the datasheets into a single file. Remove columns containing peak areas from auto-MS/MS data files.

3.2.2 Drift correction and flagging low-quality features

LC-MS-based metabolomics suffers from systematic intensity drift during an LC-MS run. This means that the signal intensity of a molecular feature either decreases or increases systematically throughout the experiment. Removing this drift increases the quality of LC-MS data and allows estimating the true biological effects more accurately. Unfortunately, some molecular features show too much variation in the QC intensities even after drift correction. We use here different quality metrics defined by Broadhurst et al. (18) for measuring the quality of a molecular feature before and after drift correction. The low-quality features are flagged although they are not included for further analysis. Note that we do not recommend removing low-quality features completely, as they are sometimes

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

needed in the metabolite identification phase when searching for specific ions or fragments of known molecules.

33. Make sure that missing values are correctly represented. A peak picking software might use a numerical value (such as 0, 1 or -999) to represent missing values, while other software such as R have specific ways of representing missing values.
34. Molecular features with too low detection rate in the QC samples should be flagged. A recommended threshold is 70%, meaning that a molecular feature needs to be detected in at least 70% of the QC samples.
35. Log-transform the features prior to drift correction. Log-transformed data can conform better with the assumptions of the regression model used to model the drift. We use the natural logarithm.
36. The drift correction should then be performed again by repeating steps 37-39 for each molecular feature.
37. Model the drift function (*f_{drift}*) by fitting a smoothed cubic spline (20)onto the QC samples, where the abundance of the molecular feature is predicted by the injection order (Figure 3a). Smoothed cubic spline regression has one hyperparameter: a smoothing parameter which controls the overall curvature of the drift function. The smoothing prevents the spline from overfitting the drift function in the presence of a few deviating QC samples (see Figure 4). A suitable value for the smoothing parameter is chosen by using leave-one-out cross validation. For the R function smooth.spline (19), we recommend the smoothing parameter to be between 0.5 and 1.5.
38. Correct the abundance of each sample using the following formula (for a sample with injection order *i*):

$$x_{corrected}(i) = x_{original}(i) + mean(x_{QC}) - f_{drift}(i)$$

39. Reverse the log transformation by applying the corresponding exponential function.
40. The drift correction procedure is visualized (Figures 3 and 4) by drawing a scatter plot of the abundances against the injection order before and after drift correction. A line representing the drift function should be added to the scatter plot before correction. To save time, we usually draw drift correction plots after the statistical tests, so we only need to draw them for the interesting molecular features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

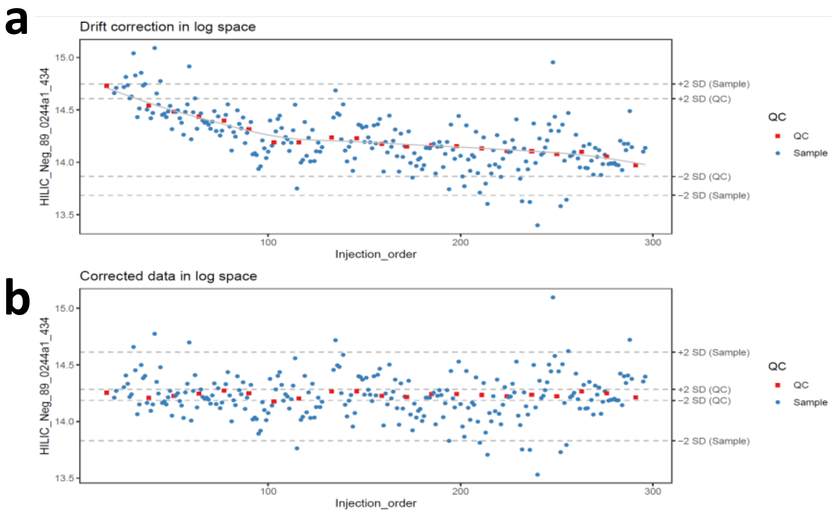

Figure 3. A molecular feature is illustrated here before (a) and after drift correction (b) by smoothing the cubic spline regression. The horizontal lines represent distances of 2 standard deviations from the mean of QC samples and biological samples. The systematic effect of the drift is corrected.

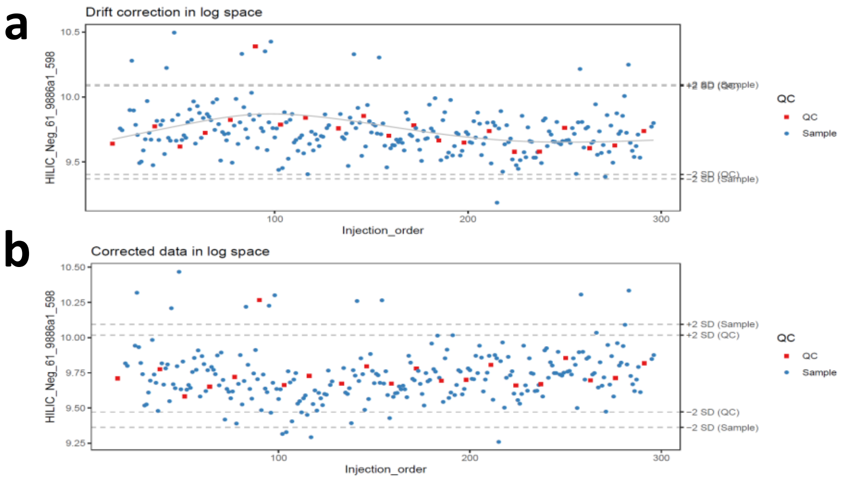

Figure 4. A molecular feature is illustrated here with a deviating QC sample before (A) and after (B) drift correction by smoothing the cubic spline regression. The horizontal lines represent distances of 2 standard deviations from the mean of QC samples and biological samples. Due to the smoothing parameter, the correction method is robust against the deviating QC sample and corrects for the global drift trend correctly.

41. **Optional step:** Compute the quality metrics after drift correction and keep only the drift-corrected values for the molecular features where the values of the quality metrics have decreased, which means that the data quality has been improved. For the other molecular features, retain the original values.
42. Flag or remove the low-quality features. As Broadhurst et al. (18) recommended, only the molecular features with RSD < 0.2 and D-ratio < 0.4 should be retained.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

3.2.3. Quality control

The raw data obtained from the peak picking software requires careful examination to estimate the need for additional preprocessing such as drift correction (see 3.2.2.). In the currently proposed workflow, the data quality is monitored at each step of the preprocessing with a set of visualizations. The example figures are based on RP positive data from a dietary intervention study (21), before and after drift correction and removal of low-quality features.

43. Draw the visualizations in steps 44-50 For the data before drift correction, after drift correction and after flagging low-quality features to monitor data quality and the effect of preprocessing.
44. Apply a linear model to each feature, where the feature levels are predicted by injection order. Then visualize the effect of drift to individual features by drawing a histogram of the p-values for the regression coefficient of injection order (Figure 5). We represent the expected uniform distribution by a horizontal line. Ideally, the p-values should roughly follow the expected uniform distribution. Unfortunately, this is rarely the case, but the closer the distribution is to uniform, the better. It is recommended to apply this procedure separately on QC samples and biological samples, which allows observing the drift patterns in both parts of the dataset.

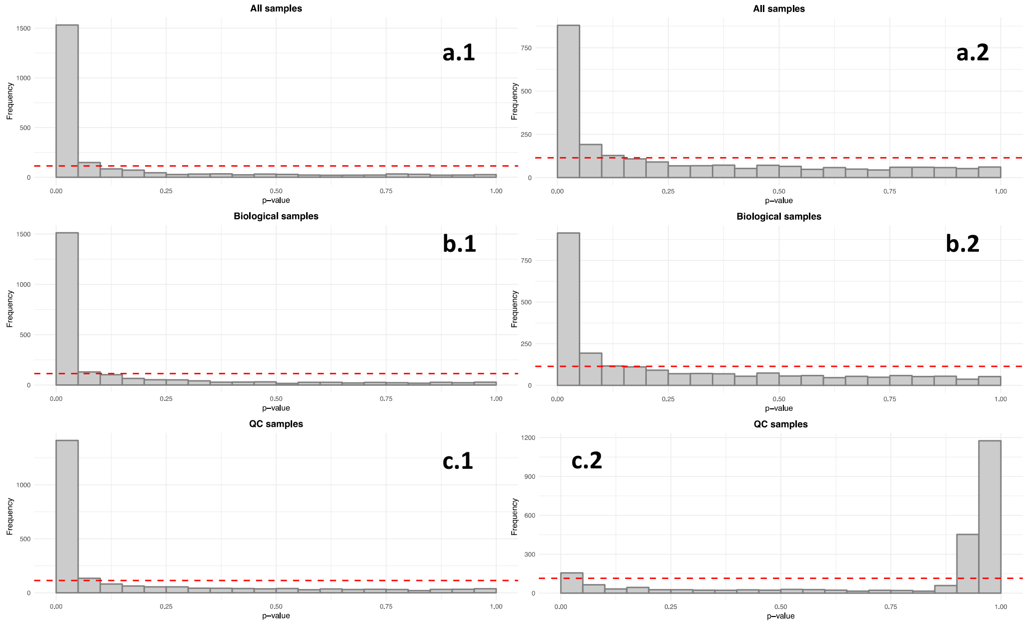

Figure 5: The six histograms illustrate p-values from linear regression models between each feature and injection order. The dashed red lines represent the uniform distribution. The histograms show the p-values from before (a.1) and after drift correction (a.2) in all the samples. The b.1 and b.2 histograms focus only in the biological samples before (b.1) and after (b.2) drift correction. Finally, the c.1 and c.2 histograms show only the p-values from the QC samples before and after drift correction. In this case, we have a strong drift in the LC-MS data because the p-values of the QCs (c.1) tend to gather close to zero. After the drift correction, (c.2), p-values for the QCs are increased.

45. Draw boxplots (Figure 6) where each individual boxplot represents the distribution of all feature levels in a sample. We have two types of boxplots: in the first type, we order the samples by the study group (a.1, a.2) (and possibly the time point). This can reveal systematic changes in the global feature levels across samples. In the second type (b.1, b.2), we order the samples by injection order, highlighting the QC samples. This allows us to observe any systematic drift across the feature levels in the samples.
46. Before the rest of the visualizations, mean center the features and divide by standard deviation.
47. Visualize the distribution of the Euclidean distances between samples using a density plot. The plot should feature two distributions, the distribution of distances between QC samples and the distances between biological samples. Ideally, the distribution of QC sample distances should be narrow and well separated from the distribution of study samples (Figure 7)

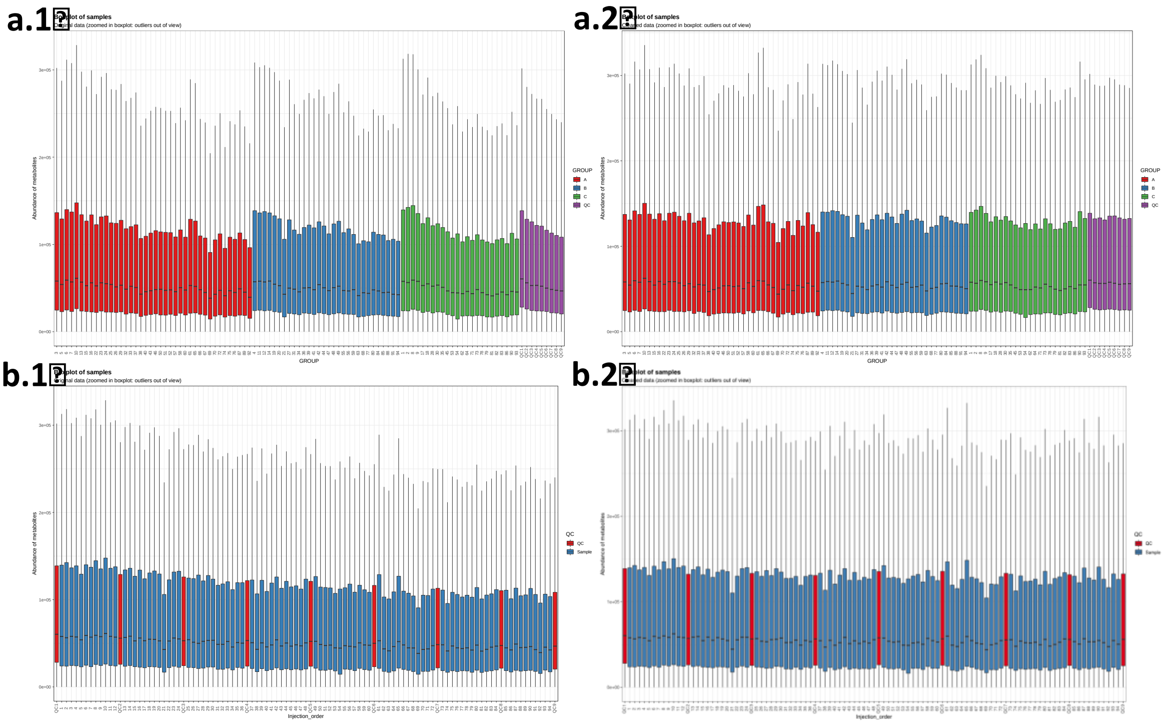

Figure 6: Boxplots of feature intensities per sample. The boxplots a.1, where the samples are ordered by study group (a.1) and b.1, where the samples are ordered by injection (b.1), show a clear systematic decrease in signal intensity during the injection sequence. After the drift correction, the drift pattern is no longer observable (in boxplots a.2 and b.2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

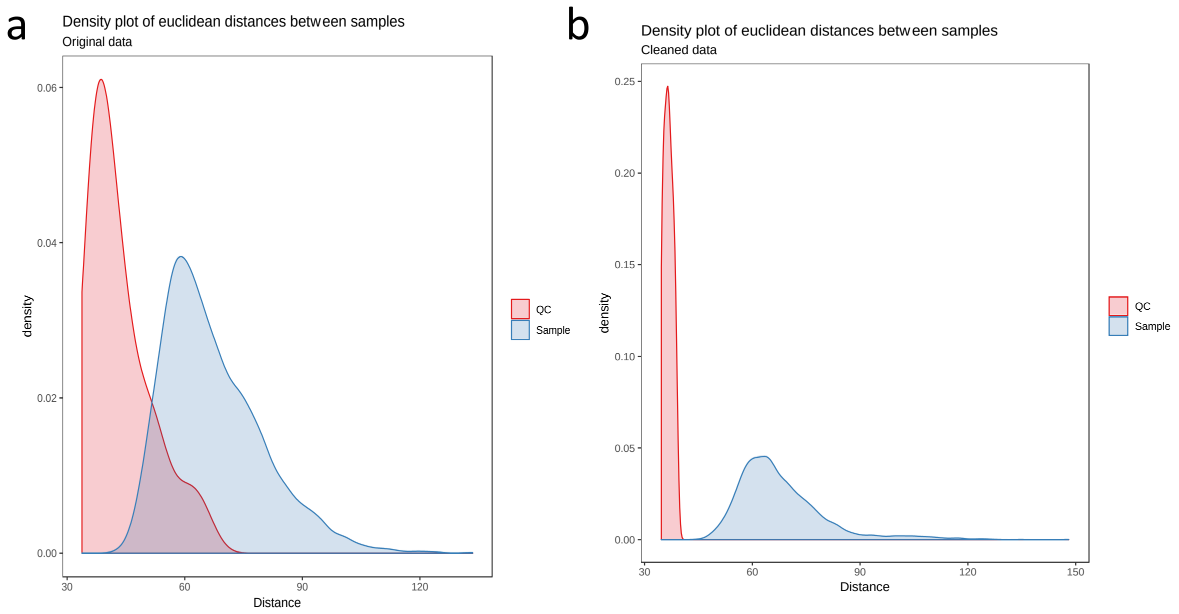

Figure 7: The density plot (a) shows a clear overlap between the distribution of QC samples and the biological samples, which indicates poor data quality. After drift correction and quality control (b), the distributions are not overlapping anymore.

Principal component analysis (PCA)(22–24) or t-distributed stochastic neighbor embedding (t-SNE)(25) can be used for observing patterns in the data by drawing scatter plots of the samples in a low-dimensional space (Figures 8, 9). PCA is a linear method, while t-SNE can also reveal non-linear patterns. For conciseness we only show t-SNE figures here.

48. Draw scatterplots of the data points using PCA and t-SNE. Samples can be highlighted by coloring the points in the scatter plot with a study factor (e.g. treatment groups or time points) to observe trends in the data. Ideally, QC samples should cluster together (Figure 8). We also draw separate plots where the samples are colored by injection order to observe drift patterns (Figure 9). If the data quality is high, there should be no visible patterns, but the color of the points should appear random (Figure 9b).
49. **Optional step**: If there is a large number of samples and the points in the t-SNE plots tend to overlap, draw a hexbin version of t-SNE scatter plots colored by injection order (Figure 10), where the plot area is divided into hexagons, and each hexagon is colored by the mean of the injection orders of the points inside that hexagon. As before, in an ideal case, there should be no visible drift patterns.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

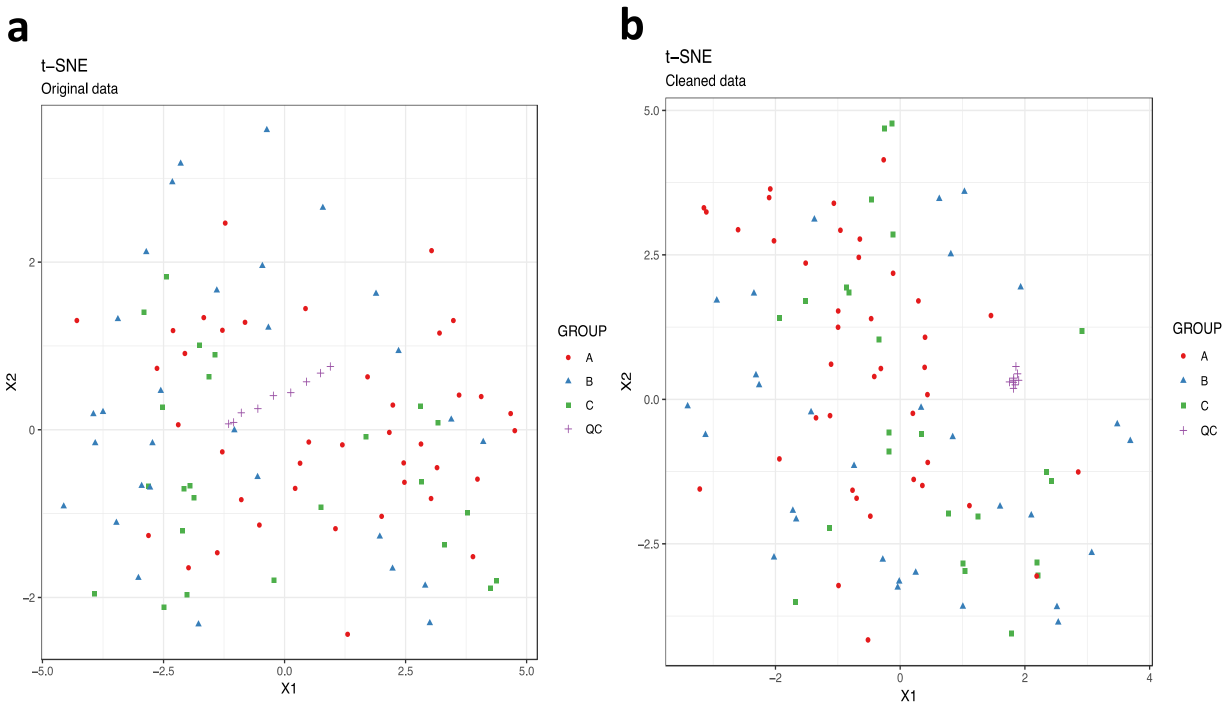

Figure 8: Investigating drift correction patterns using the t-SNE method. The QC samples are shifting systematically before drift correction (the line trend of the purple crosses symbol) (a), whereas after the drift correction (b), the line trend of the QCs is gone, and the QCs are now grouped together nicely.

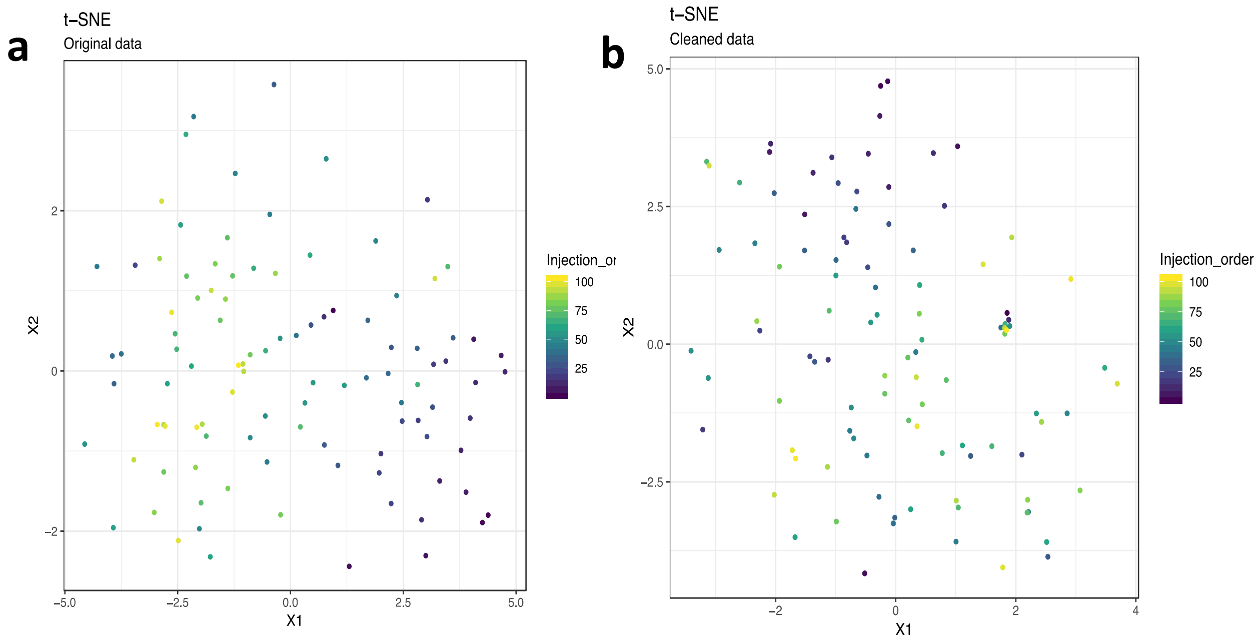

Figure 9: The drift pattern in the injection order (the color trend) using the t-SNE method is visible before drift correction (a), whereas after the drift correction, the samples are more randomly scattered (b).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

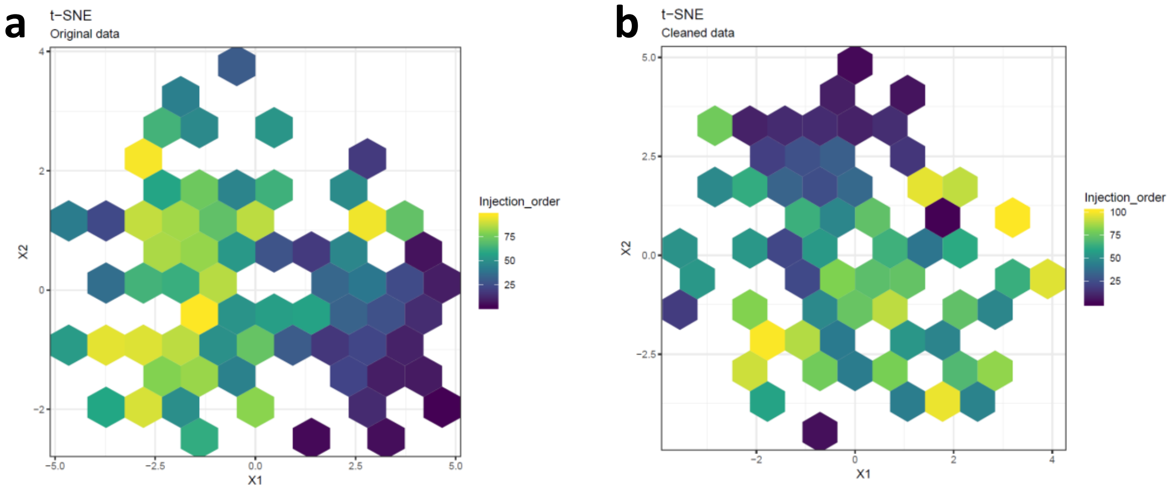

Figure 10: The hexbin plots show similar patterns as the scatterplots above: the drift pattern is more obvious before drift correction (a) than after drift correction (b). The color of each hexagon is the mean of the injection orders of the data points that fall in that hexagon.

50. Apply hierarchical clustering (26,27) to the samples and visualize the result by using a dendrogram (Figure 11a, b). The QC samples should cluster together early. We also draw a heatmap (Figure 11c, d) representing pairwise distances between samples, where samples on x and y axis are ordered by hierarchical clustering. The QC samples should have smaller inter-sample distances than other samples. Several techniques can be used for clustering. We recommend that several techniques be investigated. However, we have consistently achieved good results with hierarchical clustering using Euclidean distances and Ward's criterion for linking clusters (27).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

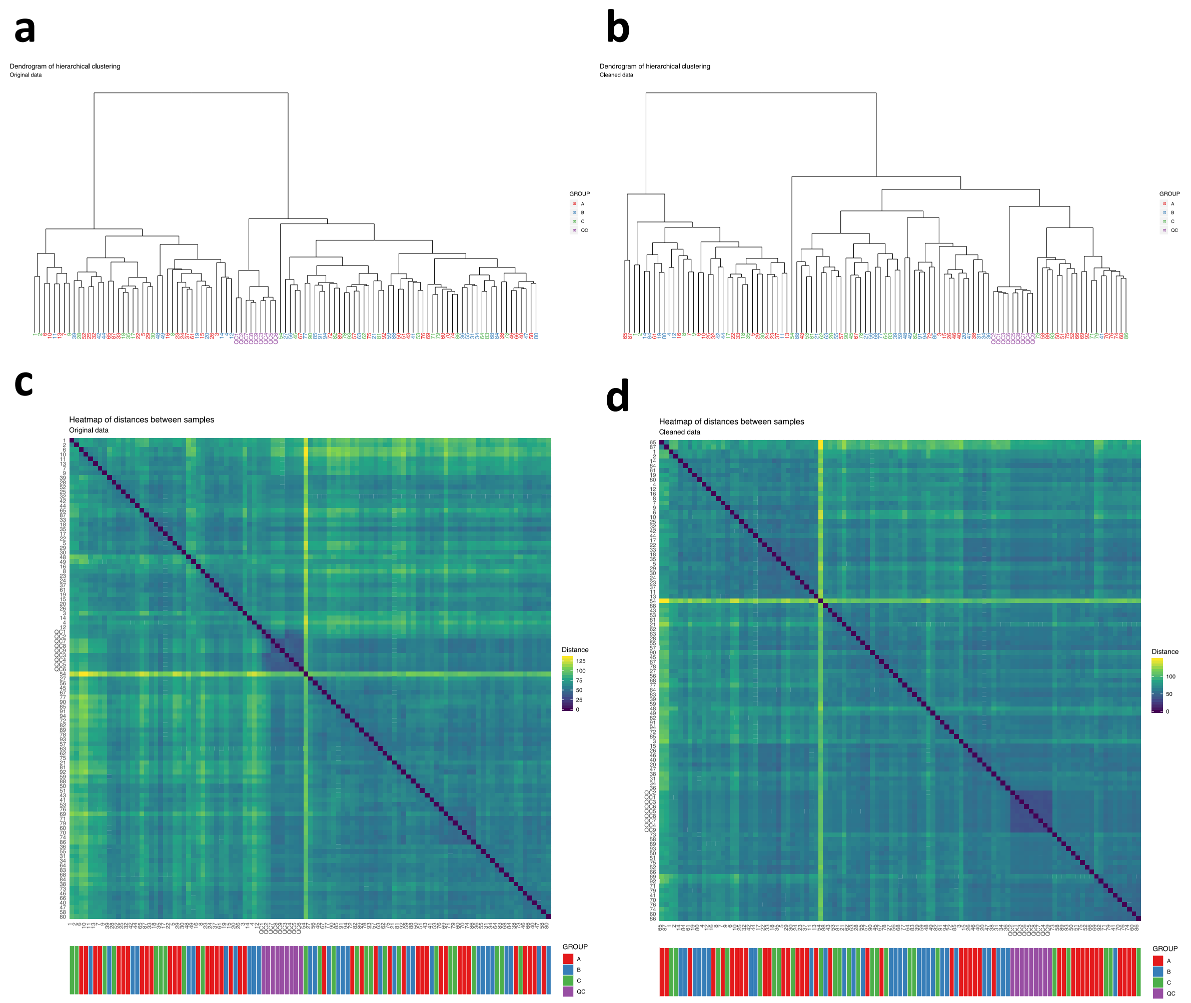

Figure 11: The hierarchical clustering algorithm clusters QC samples together even before drift correction (a) whereas, after performing drift correction (b), the QC samples cluster more clearly together. In the heatmap before drift correction (c) we cannot observe a clear QC "block" pattern (purple color code), however this pattern is clearly observed in the heatmap after drift correction (d).

3.2.4. Imputation, transformation, normalization and scaling

Missing data occur in metabolomics datasets for various reasons and it is one of the most challenging computational processes in the metabolomics data pre-processing (28). Imputation is the procedure replacing the missing data with reasonable values using a priori knowledge or information available from the existing data. In this workflow, random forest (RF) imputation is performed in order to replace the missing values with the most appropriate estimate with the missForest package (28,29), although several other procedures for imputation are available (30,31).

Normalization and transformation steps also play an important role in metabolomics analysis and they can affect the statistical analysis results (32). They are used to correct for data heteroscedasticity (e.g. when the variance of the error term depends on the independent variables in a linear regression model), and any skewed distributions that are present among the molecular features. Normalization

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

techniques can effectively convert the noise from peaks that have high intensity into systematic variation, and since heteroscedastic noise has such an influence in the data (32), it is proposed here that the data should first be transformed and then normalized. Depending on the type of statistical analysis (feature-wise or multivariate), we will proceed with different normalization and transformation approaches (33).

51. Remove the QC samples from the data. QC samples should be removed prior to imputation (and/or normalization) to ensure that imputation is based on patterns in the biological data.
52. Impute missing values using random forest imputation.
53. Transform the data either using the log transformation with the natural logarithmic (nlog) or by using the generalized logarithmic (glog) when the data are heavily skewed.(33)
54. Normalize the data by using probabilistic quotient normalization (PQN) (33,34).
55. Perform mean centering and scaling by standard deviation (autoscaling), before multivariate analysis; this is necessary with PCA and PLSDA methods, for example, but it is not always needed as in the case with RF (35).

3.2.5. Clustering molecular features originating from same metabolite

Currently used peak picking software can detect the isotopes, most common adducts and some in-source fragments, and combine those features as one entry in the data matrix. However, in LC-MS analysis, unpredictable adduct behavior and neutral loss formation typically occurs resulting in the same metabolite being redundantly represented in the data matrix, causing problems not only for the identification of the compounds but also potentially in the data-analysis step due to multiple collinearities.

We present here a novel method for clustering these features and combining them, facilitating data analysis and metabolite identification. Features originating from the same compound are assumed to be strongly correlated and have a small difference in their retention time. Thus, the algorithm initially identifies pairs of correlated features within a specified retention time window. The user specifies both the correlation threshold and the size of the retention time window. For illustration, a correlation coefficient threshold of 0.9 and a retention time window of ±1 second is used. Pearson correlation coefficient is used, as the relationship between features originating from the same compound is assumed linear.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

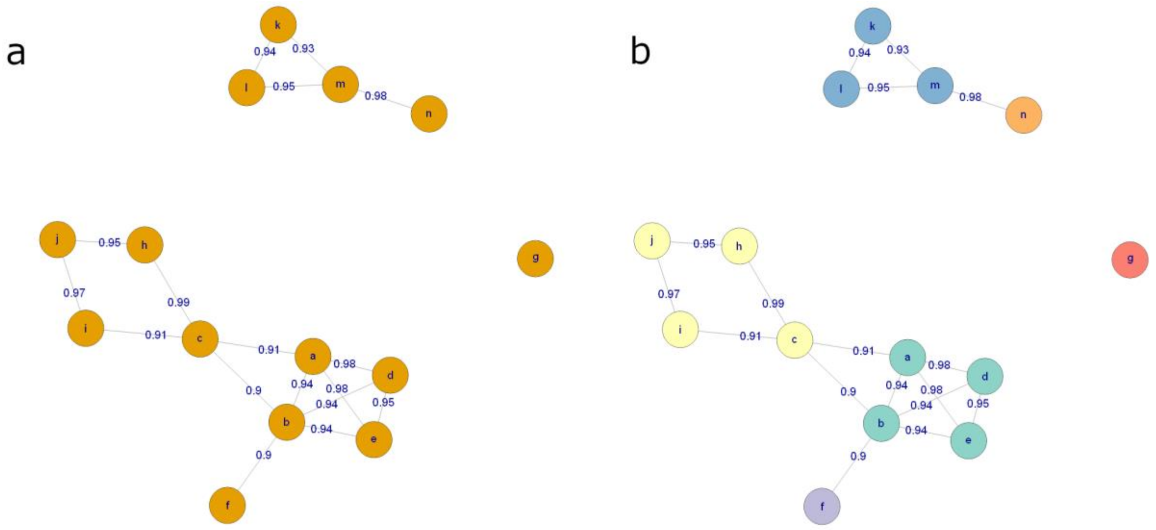

Figure 12. a) An example graph, where every node is a molecular feature and every edge represents a high correlation coefficient and a small retention time difference between the features. b) The graph after the clustering procedure. Each color corresponds to a distinct cluster of features.

Next, an undirected graph of all the connections between the features is generated, where each node represents a feature and the edge weight the corresponding Pearson correlation coefficient under the retention time constraint (Figure 12a). The graph is then decomposed into components where all the nodes are reachable from any other node. The components are then pruned by further removing nodes until all nodes have a sufficiently high degree (the number of edges of the node). This step requires a third user-defined parameter, degree threshold, defined as a percentage of the maximum possible degree. For example, in a component of five nodes, the maximum degree is 4. With a degree threshold of 0.8, each node is required to have at least $0.8 \cdot 4 = 3.2 \approx 3$ edges. If this criterion is not met, the node with the lowest degree is discarded iteratively until the criterion is met. In the case of a tie, the node with the lowest sum of edge weights is discarded. Discarded nodes are then reintroduced and can form new clusters.

After the clustering, the feature with the largest median peak area is retained for each cluster. All the features that are clustered together are recorded for future reference. Figure 12b shows the state of the graph from Figure 12a after clustering, with each final cluster colored differently.

56. Cluster the molecular features using the algorithm described above. Represent each cluster with the feature with the highest median abundance. Use these features and the clustering information for multivariate analysis and metabolite identification.

3.3. Data analysis

Once the raw data is checked for quality and analytical drift, and the features originating from same metabolites merged to reduce the data matrix entering the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

subsequent steps, the next phase in non-targeted metabolic profiling is to utilize data- analytical methods to discover the metabolites of biological significance within the taken study set-up. Preferably, combination of feature-wise and multivariate analyses can be applied (Figure 2).

57. Combine the features from the different analytical modes to a single data matrix.

3.3.1. Feature-wise analysis

In feature-wise analysis, two types of testing may be used depending on the data: parametric and non-parametric (36). The choice of the test depends on the data and the biological research questions of the study. Most typically, parametric statistical tests are used, but if the features do not satisfy the assumptions of parametric tests, they may be replaced with non-parametric alternatives. Non-parametric methods perform better when dealing with non-normal populations, unequal variances, and unequal small sample sizes.

58. For study designs with two groups and no covariates, such as case versus control studies, use a simple Welch's t-test. Welch's t-test can deal with unequal variances between the groups. For a non-parametric alternative, consider a Mann-Whitney U test.
59. For studies with multiple groups, first apply Welch's one-way analysis of variance (ANOVA) to choose interesting features based on their p-value and in order to find differences between individual groups, conduct post-hoc pairwise t-tests between the groups for the subset of important molecular features. Welch's ANOVA can deal with unequal variances between the groups. For a non-parametric alternative, consider a Kruskal-Wallis test.
60. For studies with two categorical variables, apply two-way ANOVA, which allows examining the main effect of each variable plus their interaction. If one or both variables have multiple levels, choose interesting features based on their p-values and conduct post-hoc pairwise t-tests as before. For a non-parametric alternative, consider Friedman test.
61. For studies with repeated measurements, use a linear mixed effects model with the time point, group and their interaction factors as fixed effects and the subjects as a random effect. If there are no more than two groups or time points (or if time is modeled as a continuous variable), use t-tests on the regression coefficients to assess the significance of the effects. In the case of multiple groups and/or time points, use type III F-tests for ANOVA-like tables, e.g. with the help of two R packages lme4 and lmerTest that provide all the necessary tests (37,38).
62. To test the strength of association between molecular features, or between molecular features and other variables, use Pearson correlation, or Spearman correlation as a non-parametric alternative. This can also be done post-hoc, after identification of key metabolites (39).
63. After performing the feature-wise tests for each feature, the p-values acquired from the tests need to be adjusted for multiple testing. We recommend using

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

Benjamini-Hochberg false discovery rate (FDR) approach. The FDR-adjusted p-values are sometimes referred to as q-values. (36,40,41).

3.3.2. Multivariate analysis

Multivariate analysis offers powerful tools for any metabolomics analysis. Dimensionality reduction methods like PCA or t-SNE enable us to explore the data to identify outliers and seek possible patterns like clusters of samples. Unsupervised clustering methods, such as hierarchical clustering are useful for validating findings from dimensionality reduction methods, as they allow us to observe clustering patterns in high-dimensional space.

Partial least squares (PLS), and random forest (RF) are useful supervised learning techniques for finding the most interesting molecular features (42,43). Both the PLS and RF algorithms can be used for both regression and classification. In case of classification, the PLS model is called partial least squares discriminant analysis (PLS-DA). PLS-DA contrary to PCA is a supervised dimensionality reduction method that relies on the class membership of each sample and it can be used for predictive and descriptive modeling as well as for discriminative variable selection. RF is a highly flexible model and has 3 main advantages over PLS -DA: As opposed to PLS, RF does not assume Gaussian distribution of the variables. Moreover, RF does not assume linear relationships between response and (latent) predictor variables. Finally, RF is scale invariant, which circumvents issues with scaling and transformations of the metabolomics data. On the other hand, it should be noted that PLS can produce stronger models if model assumptions are met. Both PLS-DA and RF offer statistics for evaluating the importance of individual features. PLS-DA provides variable importance in projection (VIP) values, and RF estimates the rise of error if the feature was excluded from the dataset.

64. Apply multivariate algorithm for prediction and variable selection. The multivariate analysis included in this workflow is the MUVR package in R which includes both RF and PLS-DA (42). With the RF approach, three different models are obtained as result; minimal-optimal ('min'), 'mid', and all-relevant ('max') (Figure 13). The 'max' model corresponds to maximum information content once the non-informative features have been removed. This is the model usually chosen if for example pathway analysis will be applied afterwards. The 'max' model includes the highest numbers of relevant molecular features, thought it may include some redundant features or highly correlated features. The 'min' model corresponds to the minimal-optimal set of molecular features where you're likely to find the strongest biomarker candidates. The 'mid' model corresponds to an average (geometric mean) between the 'min' and 'max' options where there is one overall "best" model while reducing the redundant signals without risking the information loss in the process. In the end, the selection of the model strictly depends on the research interest and study question, such as pathway analysis or biomarker discovery. For example, if the study aims to discover a biomarker, a min model is sufficient since we want the minimum

number of metabolites as the biomarker candidate. On the other hand, in pathway analysis, we want the maximum number of metabolites to be mapped into the metabolic pathway, so the 'max' model would be suitable for this purpose.

65. **Optional Step**: Follow this step if the MUVR package is not available (for example if other software than R is used). Evaluate performance of the multivariate model. Use cross-validation for PLS-DA and out-of-bag error estimate for RF (for more information see (43))If the model performance is satisfactory, record variable importance metric (VIP value for PLS-DA and rise in error rate for RF) for each feature.

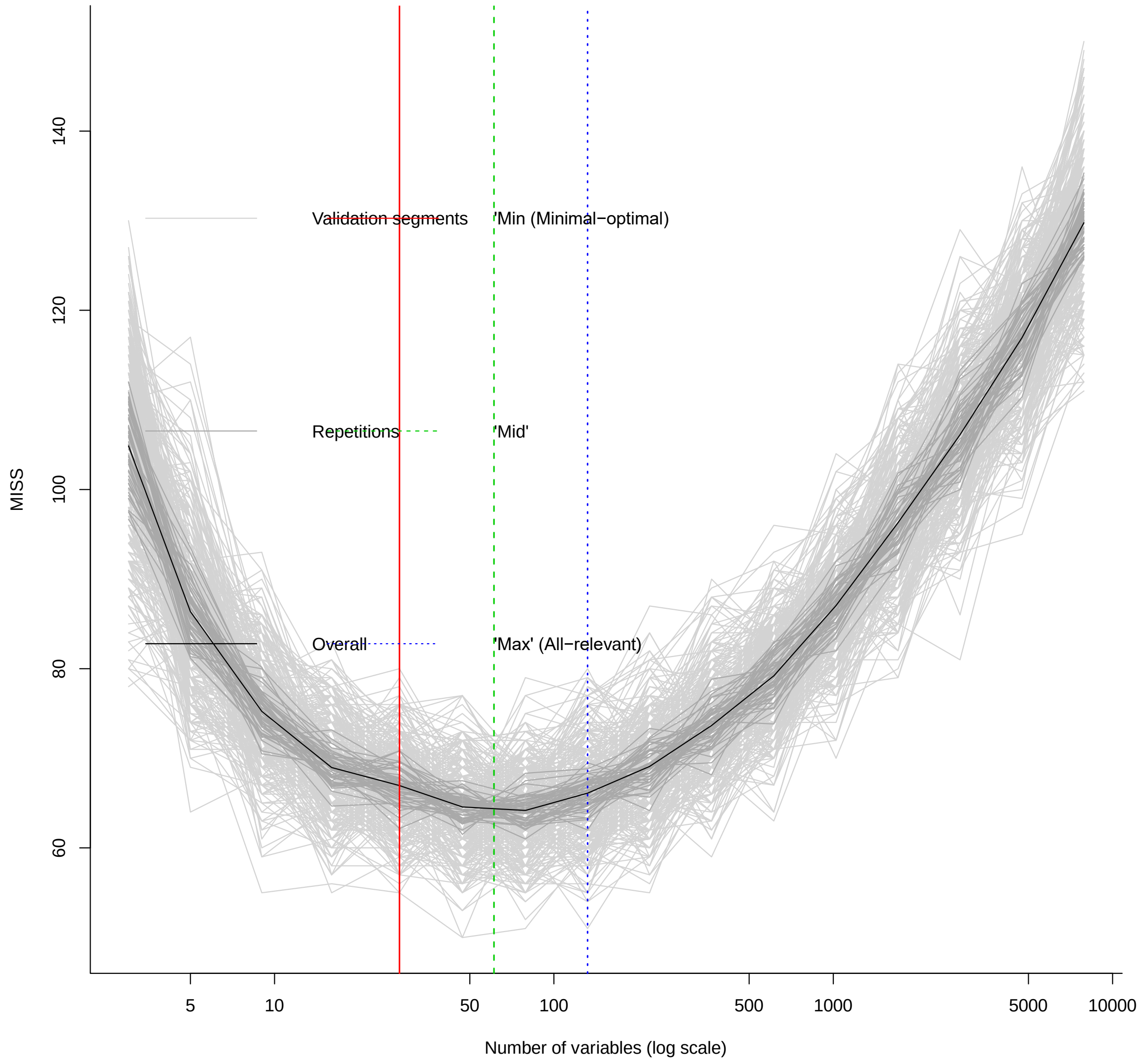

Figure 13. The MUVR validation plots for identification of the all-relevant ('max' model) and minimal-optimal ('min' model) variables. The light grey lines represent validation performance for the individual inner segments, whereas the darker grey lines represent inner segment validation curves averaged over the repetitions.

3.3.3. Ranking and filtering for variable selection

After the completion of both feature -wise and multivariate analysis, the results are combined via a ranking method in order to determine the most robust and presumably biologically relevant metabolic features to undergo identification.

66. The first step is to rank the p-values from the feature-wise analysis by giving the first number to the metabolic feature with the smallest p-value. The numbering is continued until all the metabolic features have a rank.
67. The RF variable selection method from the MUVR R package (44) provides as results three different models, as mentioned above: the min, mid and max model. In each of these different models, the metabolic features have been given a number of importance, similar to VIP rank (45). In most of the cases, we choose the max model for our analysis. However, the three different models provide different ranks, so we recommend choosing one of these models from the beginning. The second step also includes ranking the results from the RF max model with the same notion as in the feature selection case, the smallest RF rank receives the number one and the ranking continues until all metabolic features receive a number.
68. As a third step, the ranks from the same molecular features from both feature -wise and multivariate analysis are added together resulting in the FINAL rank that may be used when selecting the molecular features to undergo identification.

3.4. Visualization of results

After feature-wise and multivariate statistics, we recommend visualization of patterns of the dataset, both on a feature level and a global level as well as visualization of the p-values and effect size measures, to offer a broad view of the results.

3.4.1. Feature-wise graphs

While t-SNE figures (Figures 8 and 9) provide a solid overview of the overall patterns in the data, visualizing effects of study factors on a molecular feature level is useful when interpreting the results. The visualization type used depends on study design.

69. If the study has multiple study groups, the differences between groups can be illustrated by beeswarm boxplots separately for each group (Figure 14).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

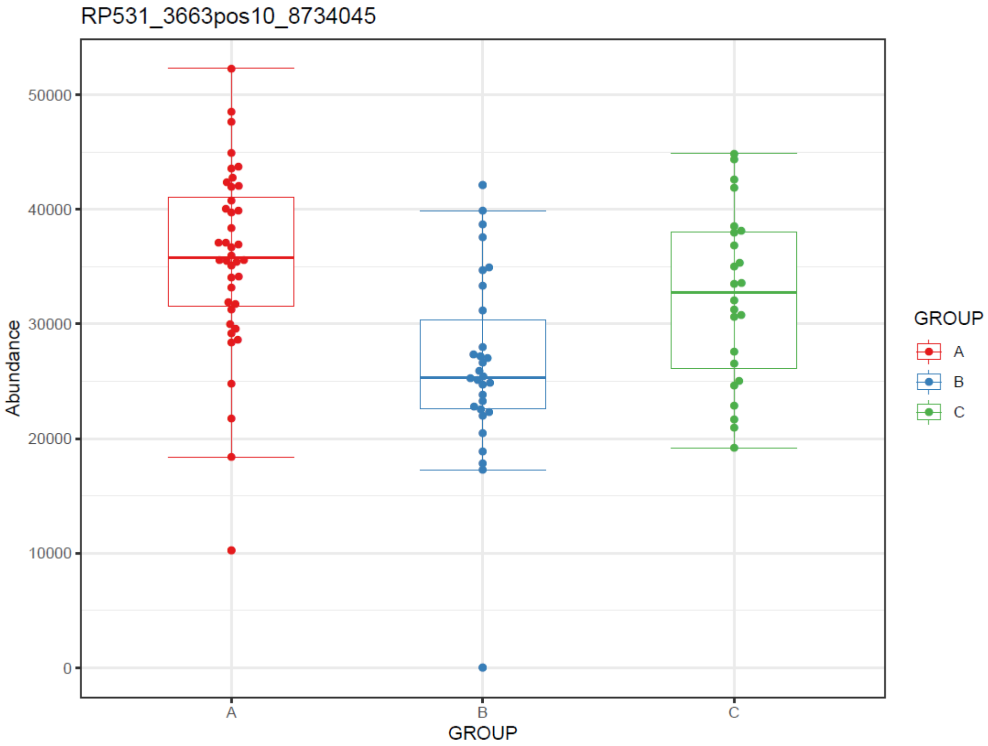

Figure 14. Three beeswarm boxplots for one molecular feature colored by study group.

70. If the study contains samples from multiple time points, the effect of time can be visualized with a line plot using one line per subject together with a thicker line representing the mean at every time point (Figure 15).

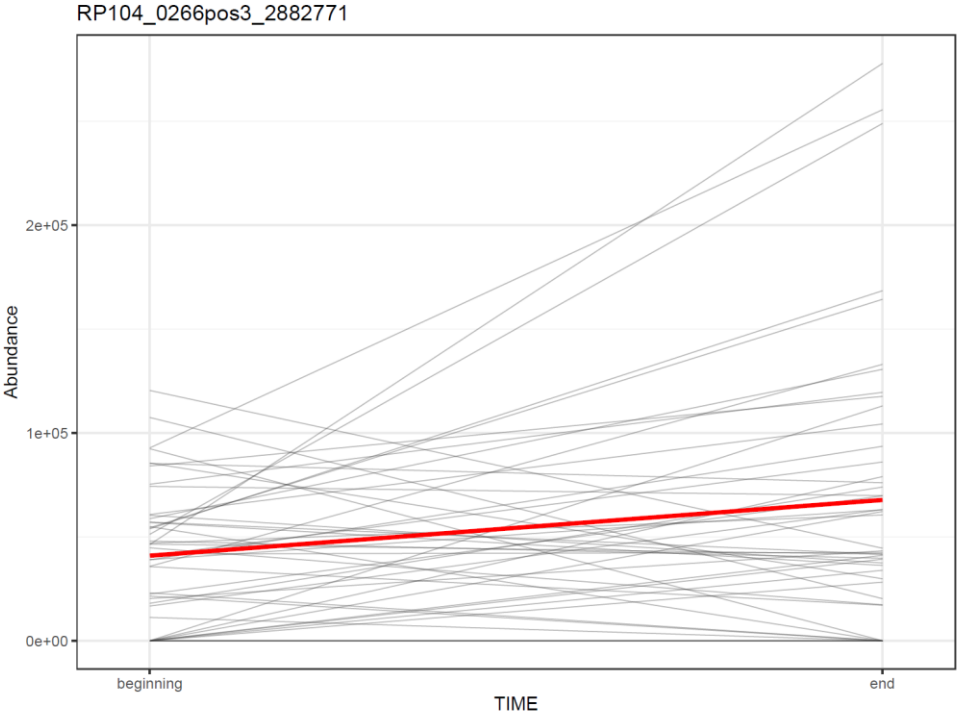

Figure 15. The change in the metabolite levels as a function of time in each subject. The thick red line represents the sample mean.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

If the study contains both multiple groups and multiple time points, consider the following visualizations:

Furthermore, for repeated measures data, plot least square means from the repeated measures model for each study group. You should also add whiskers around the points representing 95% confidence intervals, standard deviation or other measure of variability (Figure 16).

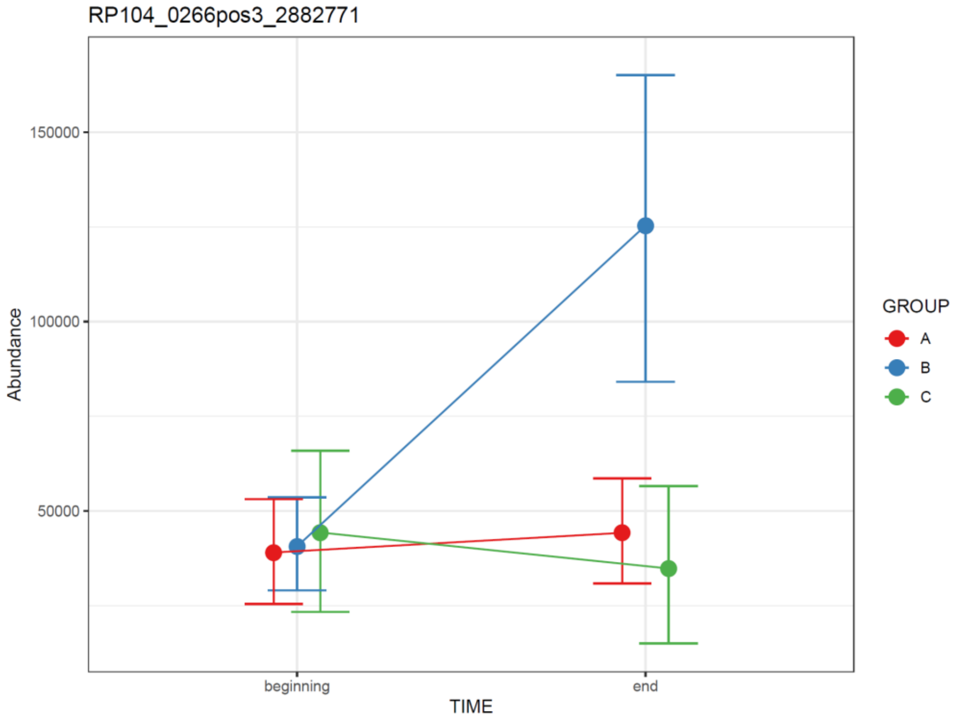

Figure 16. The change in the metabolite level as a function of time in each study group. The whiskers depict 95% confidence intervals.

71. Draw a line plot similar to the one in step 70, but color the subject lines according to group and draw separate mean lines for each group (Figure 17a). If the figure gets too cluttered, consider plotting each group separately in small multiples with a common y axis (Figure 17b).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

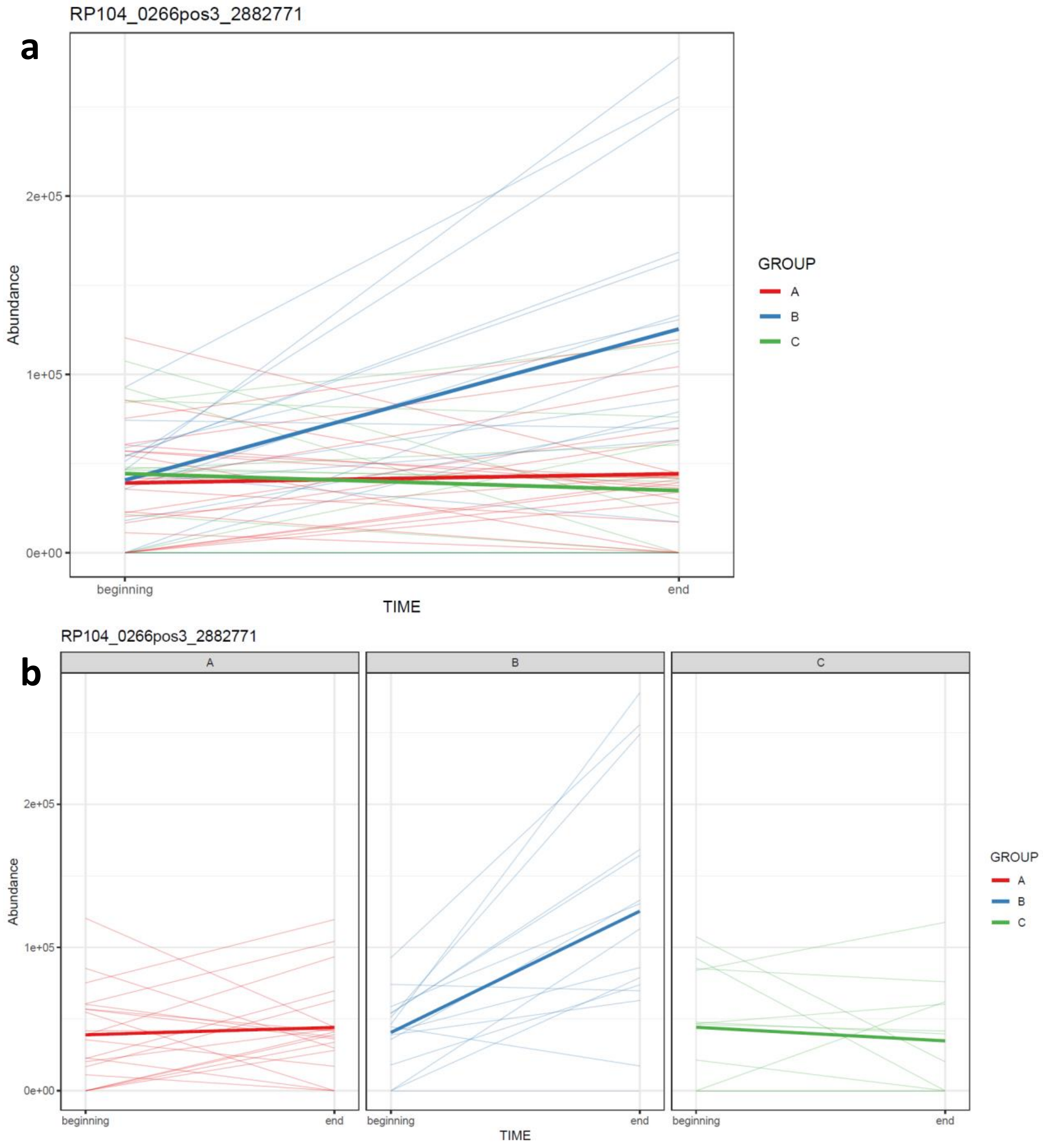

Figure 17. The change in metabolite levels between two time points in each subject, colored by group (a). Data with time series from multiple groups is easier to read when divided to small multiples (b). The bold lines represent group means. Note that the bold mean lines do not necessarily reflect an overall trend present in each subject.

3.4.2 Comprehensive visualization of results

Here, we present ways of visualizing results from both feature-wise and multivariate analysis. For illustration, we use a simple case from the RP positive mode of an intervention study, where the samples are taken from two time points, before and after an intervention. For feature-wise analysis, we used a linear model with a molecular feature as the dependent variable and the time point as the independent variable. We also calculated fold change between the two time points

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

for a scale-free measure of effect size. For multivariate analysis we fit a PLS-DA model predicting the time point from the features.

72. Visualize the patterns in the final dataset using unsupervised dimensionality reduction techniques such as PCA (23) (Figure 18) and t-SNE . If PCA reveals interesting patterns, use a PCA loadings plot to reveal which features contribute the most to the first two components that are visualized.

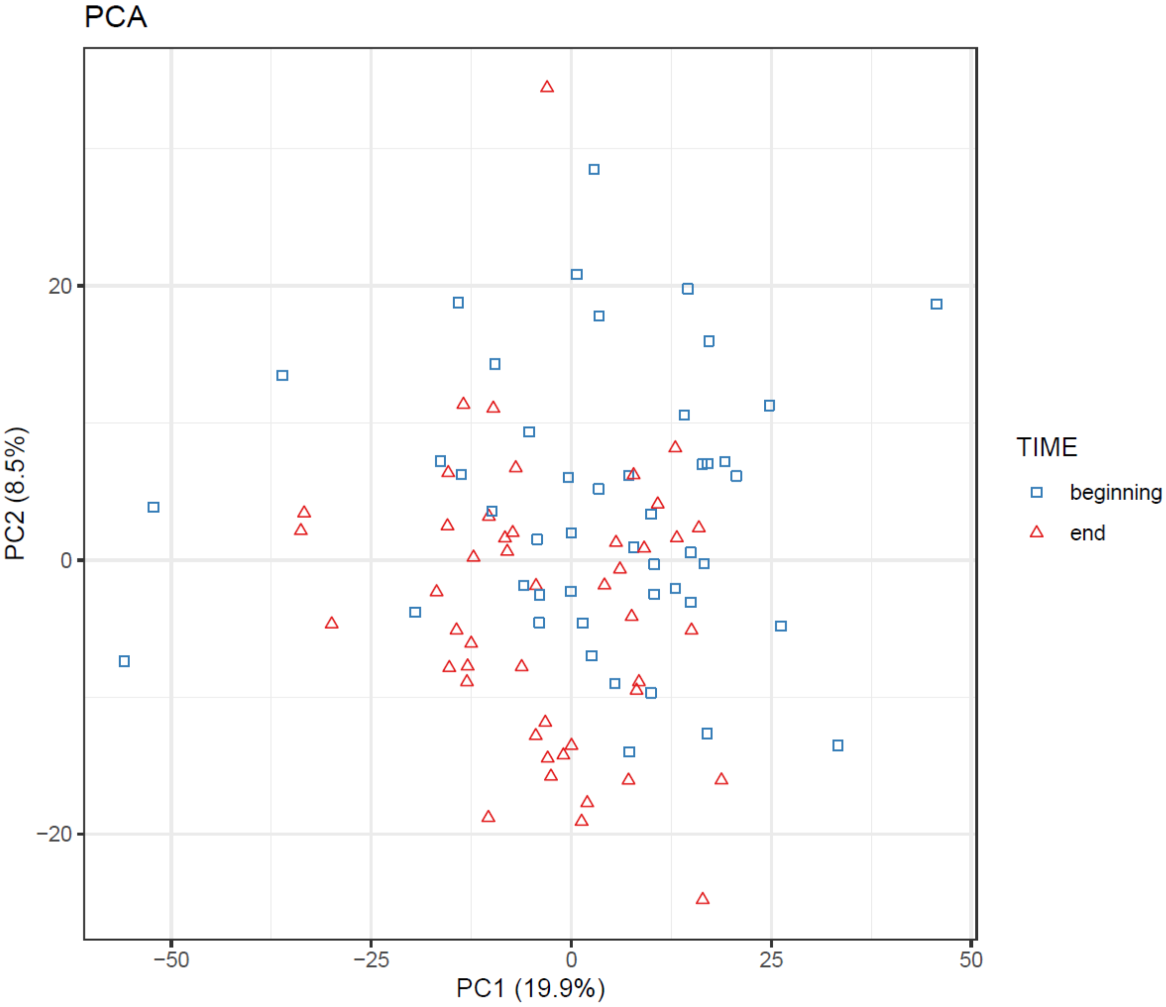

Figure 18: PCA plot of samples from an intervention study, before and after the intervention. The time points are somewhat separated, but no clear clusters or outliers are visible.

73. If PLS-DA is used, visualize the samples in the PLS component space with a scores plot. As in a PCA scatterplot, the coordinates of the points are their scores on the PLS components (see Figure 19).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

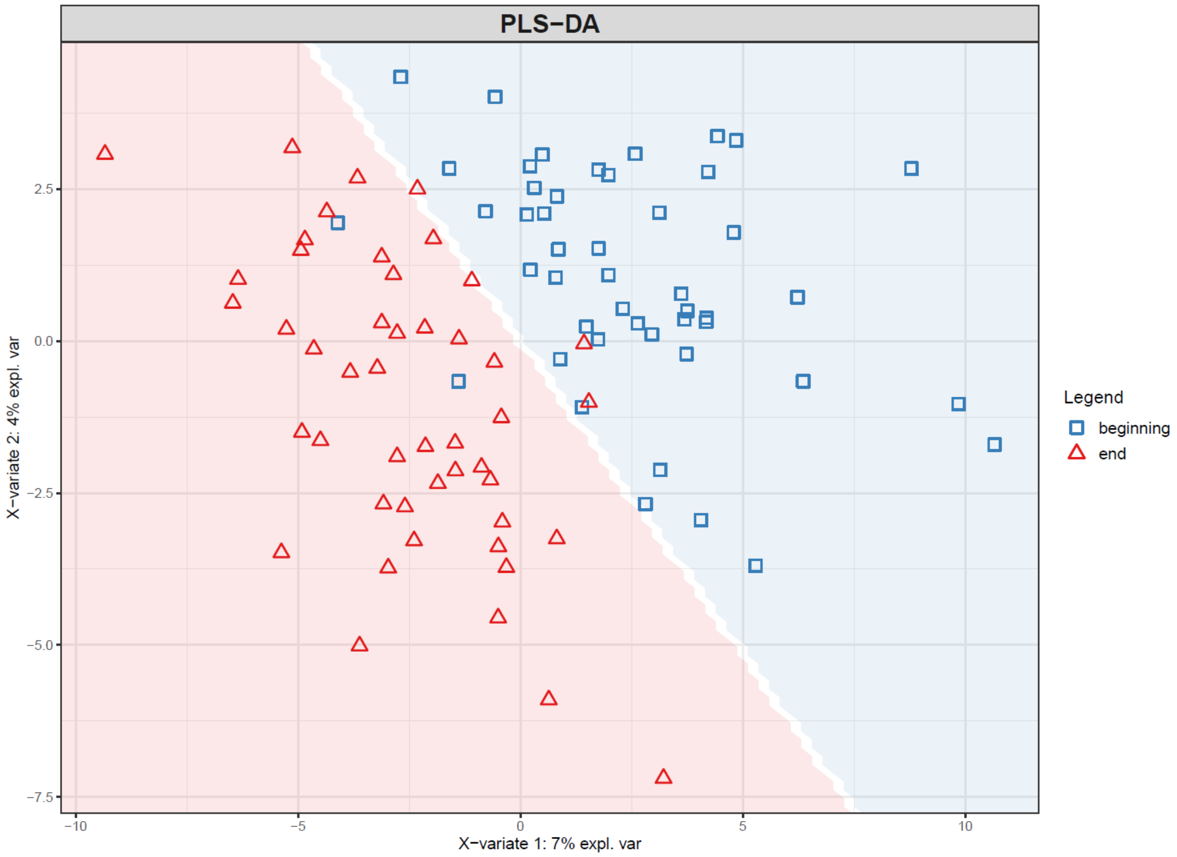

Figure 19: Scores plot of the first two components of a PLS-DA model trained to predict the time point of samples from an intervention study. The background color indicates the prediction of the model: samples in the blue area are classified to time point “beginning”, and samples in the red area to time point “end”. Note that the time points are clearly more separated as in the PCA plot in Figure 18. This is to be expected, as PLS-DA finds components that specifically separate the two time points.

74. To visualize overall changes with respect to time in studies with multiple time points, use PCA and t-SNE figures with arrows depicting change in each individual. The arrows should start at the first time point and end at the last time point for each individual. We recommend plotting each study group separately, as the plot can get crowded since the arrows occupy significantly more space than points (Figure 20).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

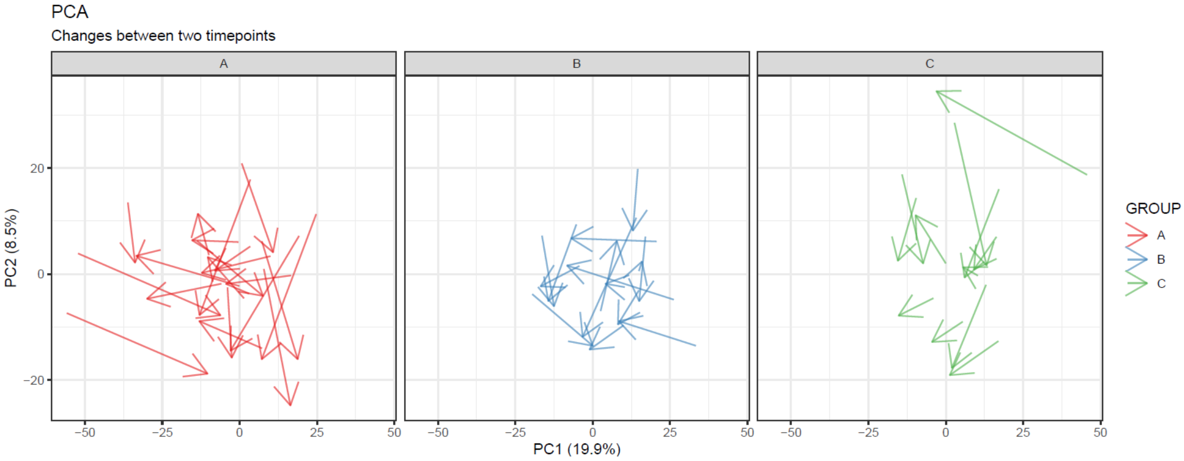

Figure 20. Changes in each subject between two time points visualized as arrows between points in a PCA plot. Samples in different groups are separated into subplots. While no group shows a systematic direction of change, we can observe that the subjects in-group A show greater overall change that subjects in the other groups.

Visualize the distribution of p-values from feature-wise analysis in a histogram. Use a line to depict the expected uniform distribution (under null hypothesis). If the distribution of the p-values deviates from the expected distribution, it can be argued that we are observing a real effect (Figure 21).

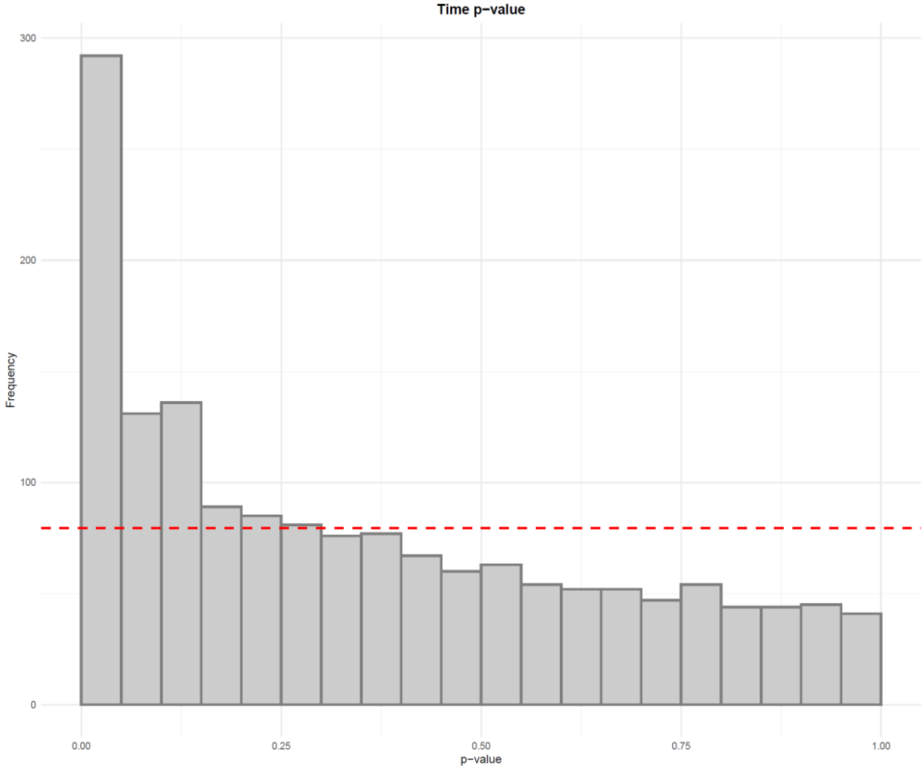

Figure 21. The distribution of p-values from linear models tests the difference of feature abundances between two time points. As the distribution clearly deviates from the uniform distribution depicted by the red line, it can be argued that there is a true difference between the two time points.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

Visualize the results of feature-wise tests in a volcano plot. Volcano plots are scatter plots with p-values on the y axis and a suitable effect size (such as fold change) on the x-axis. Add a horizontal line representing the significance threshold for FDR-adjusted q-values. To co-visualize multivariate results, the features can be colored by their relevance score in the multivariate prediction (Figure 22).

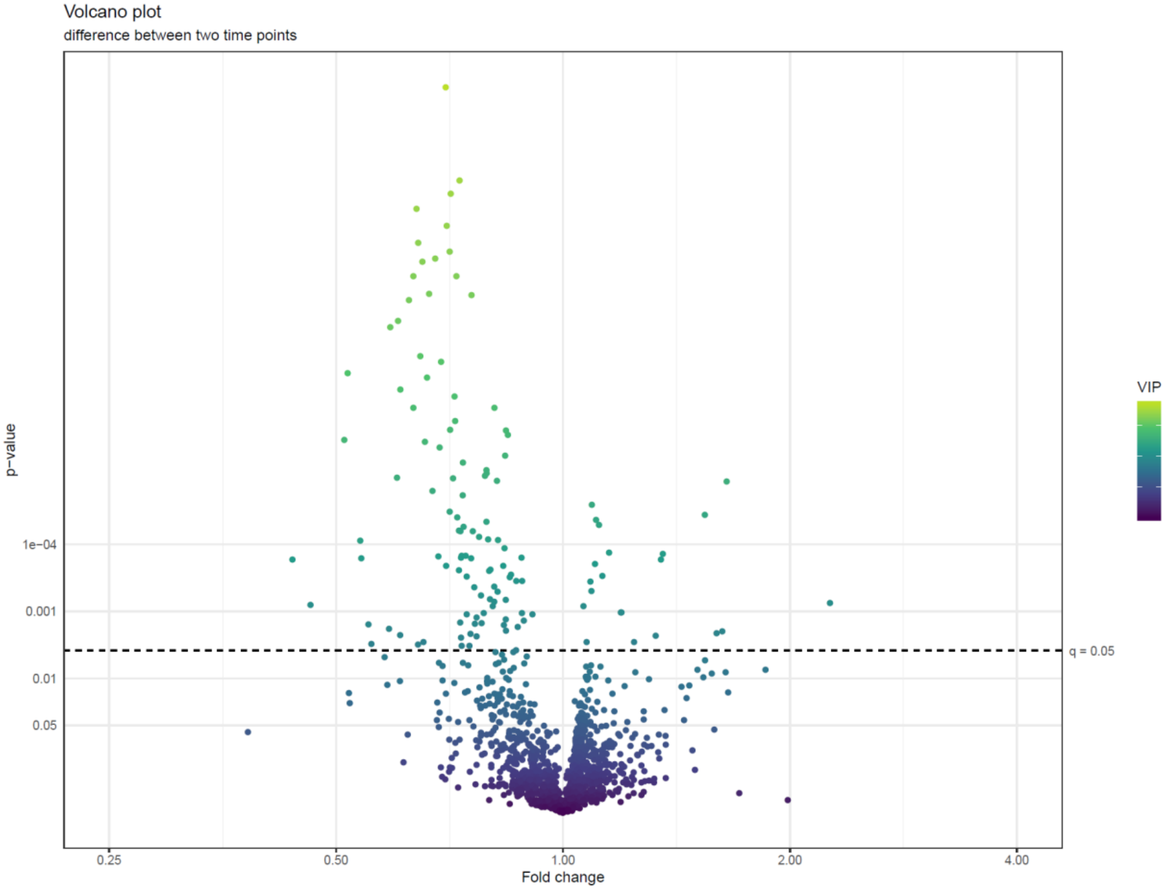

Figure 22. A volcano plot of p-values (negative log10 scale) from linear models testing the difference of feature abundances between two time points against fold changes between samples taken before and after a dietary intervention (log2 scale). The features are colored by VIP-value from a PLS-DA model trained to separate the two time points. We can observe that the features with the smallest p-values tend to have fold changes below 1, indicating that they are less abundant in the end of the intervention. Other metrics of effect size, like Cohen’s d values, can also be used in volcano plots.

Manhattan plots are commonly used in genome-wide association studies (GWAS) to study the location of the most significant single nucleotide polymorphisms on the genome. Using mass-to-charge ratio or retention time on the x-axis can use the Manhattan plots in metabolomics. In addition, in cases where direction of effect can be determined, we can multiply the y-axis by the sign of the effect to create so-called directed Manhattan plots. The Manhattan analogy is not lost, since the downward points represent the reflection of the skyline on the Hudson River. Note that Manhattan plots should always be drawn separately for each column and ionization mode, as the metabolite classes corresponding to certain m/z and retention time values depend on the column and ionization mode used.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

75. Use a Manhattan plot to connect the results of statistics to biochemical properties of the molecular features. The Manhattan plot should have either retention time or mass-to-charge ratio as the x-axis and –log10(p-value) on the y-axis. For a directed Manhattan plot, multiply –log10(p-value) by the sign of the effect. The points in the Manhattan plot can be colored by the respective VIP value from PLS-DA or another similar metric. Similarly to volcano plots, add a horizontal line to represent the significance threshold for FDR-adjusted q-values. Figures 23 and 24 show Manhattan plots with mass-to-charge ratio and retention time on the x-axis, respectively.

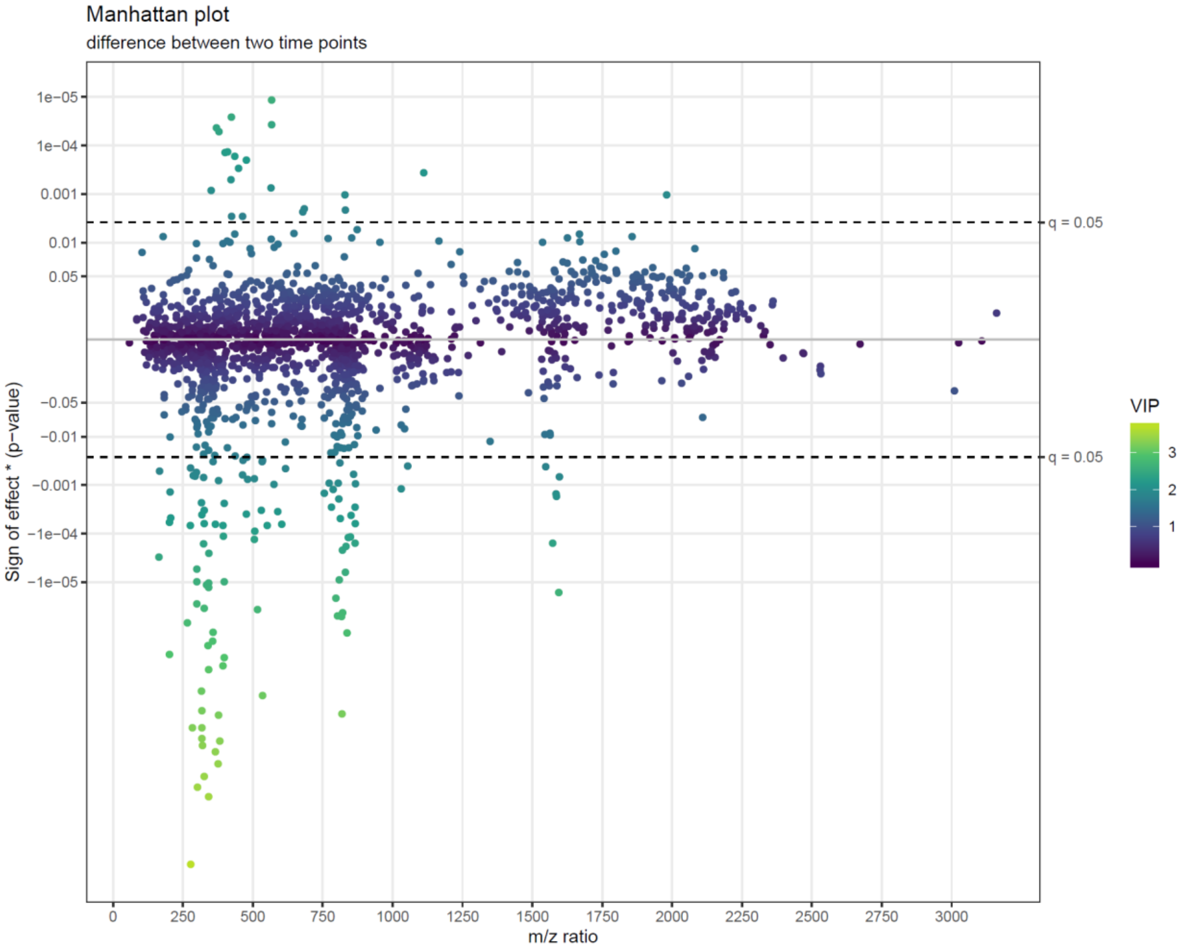

Figure 23. A directed Manhattan plot of p-values from linear models testing the difference of feature abundances between two time points with mass-to-charge ratio. The points are colored by VIP-value from a PLS-DA model trained to separate the two time points. The most interesting groups of metabolites seem to have m/z ratios around 350 and around 800. Both groups are predominantly lower in the end of the intervention.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

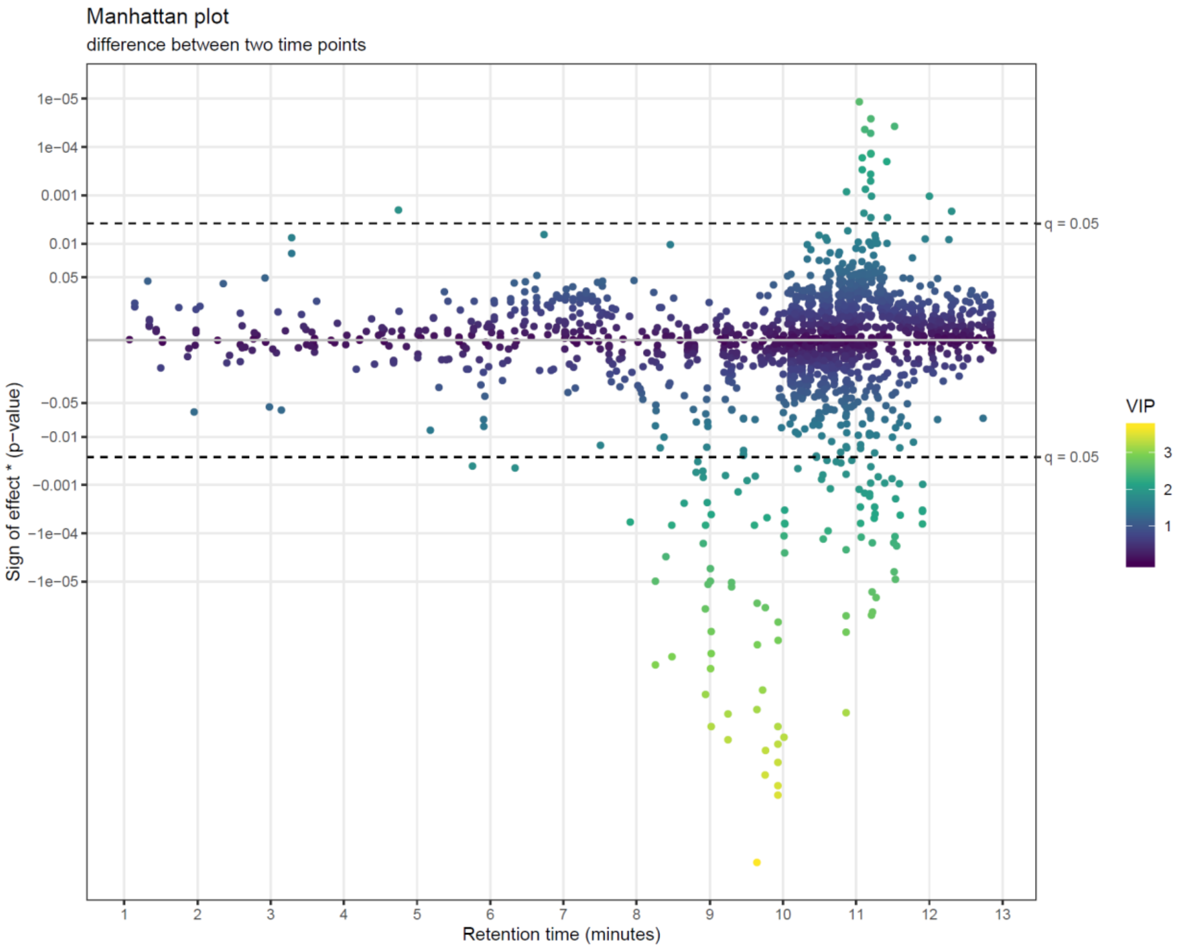

Figure 24. A directed Manhattan plot of p-values from linear models testing the difference of feature abundances between two time points with mass-to-charge ratio. The points are colored by VIP-value from a PLS-DA model trained to separate the two time points. The most interesting groups of metabolites seem to have retention times around 9–10 minutes and around 11 minutes. The first group is predominantly lower in the end of the intervention, while the metabolites in the second group have mixed associations.

76. To combine the information of both Manhattan plots, consider a scatter plot with m/z and retention time on the x- and y-axis, with the size of the point depicting p-value and the points potentially colored VIP value or other similar metric as before (Figure 25). While size is not an accurate metric in visualizations, this visualization combines mass and retention time so that the most interesting metabolite classes can be identified. As with Manhattan plots, these plots should be drawn separately for each column and ionization mode.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

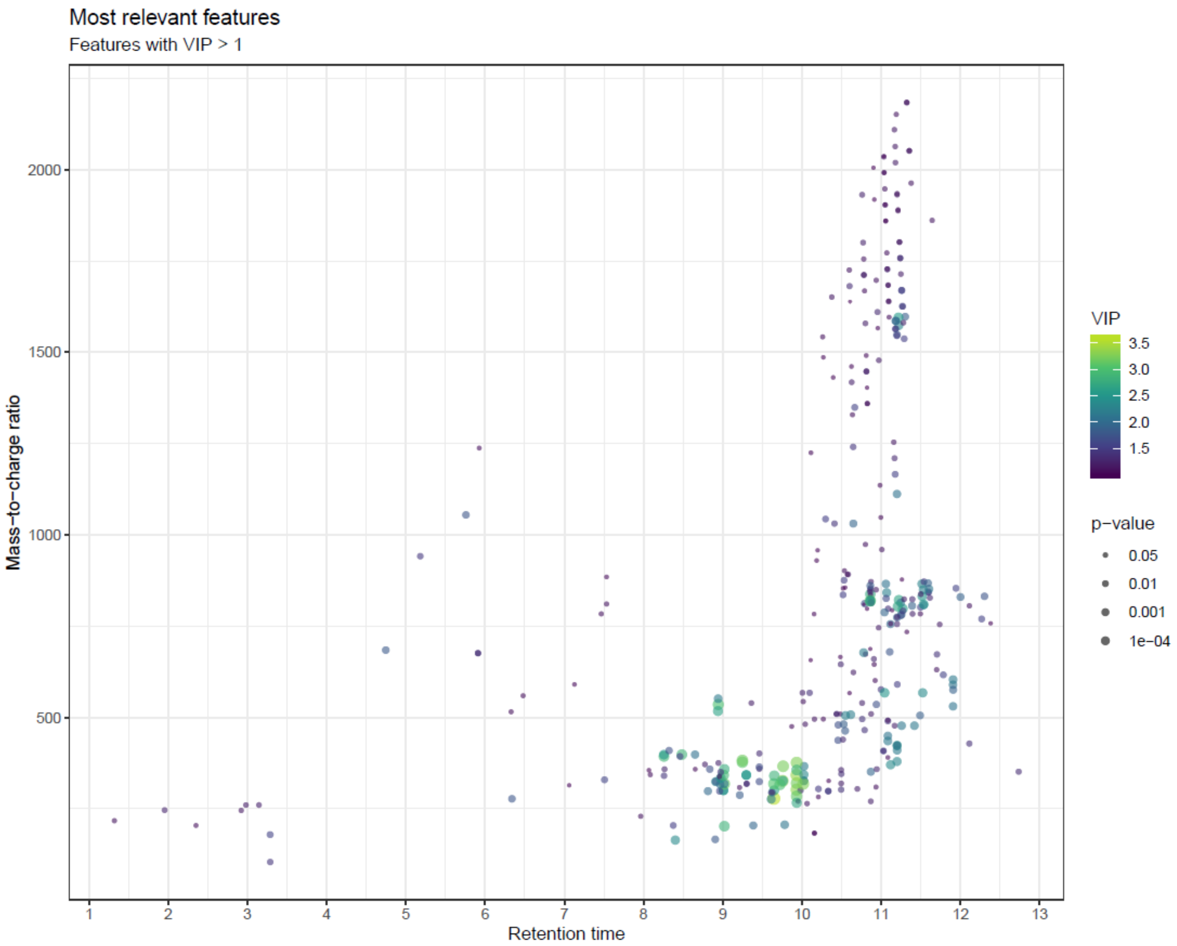

Figure 25. Scatter plot of molecular features in m/z vs retention time space, with the size of the points depicting p-values from linear models testing the difference of feature abundances between two time points with mass-to-charge ratio. The points are colored by VIP-value from a PLS-DA model trained to separate the two time points. To avoid too many overlapping points, only points with VIP value > 1 are drawn. We can observe that the most interesting group of features has retention times around 9-10 minutes and *m/z* ratios around 350.

Clustering analysis provides another effective way to view the complete data, grouping the metabolites and/or samples based on similarities in their metabolite abundance profile and providing visualization in the form of a heat map. We utilize Multiple Experiment Viewer (http://mev.tm4.org/) for *k*-means clustering and hierarchical clustering analyses, which group metabolites into separate clusters or into a hierarchy tree, respectively. The heat maps produced from the analyses can be used to assess the impact of the intervention and the number and proportion of metabolites behaving in a certain manner. An example of such heat maps is presented in Figure 26. We use the notame R package to produce heat maps of the identified metabolites and their associations with e.g. clinical markers, in which case additional information may be added to each cell, such as the statistical significance with circles, where a larger circle represents a lower p-value.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

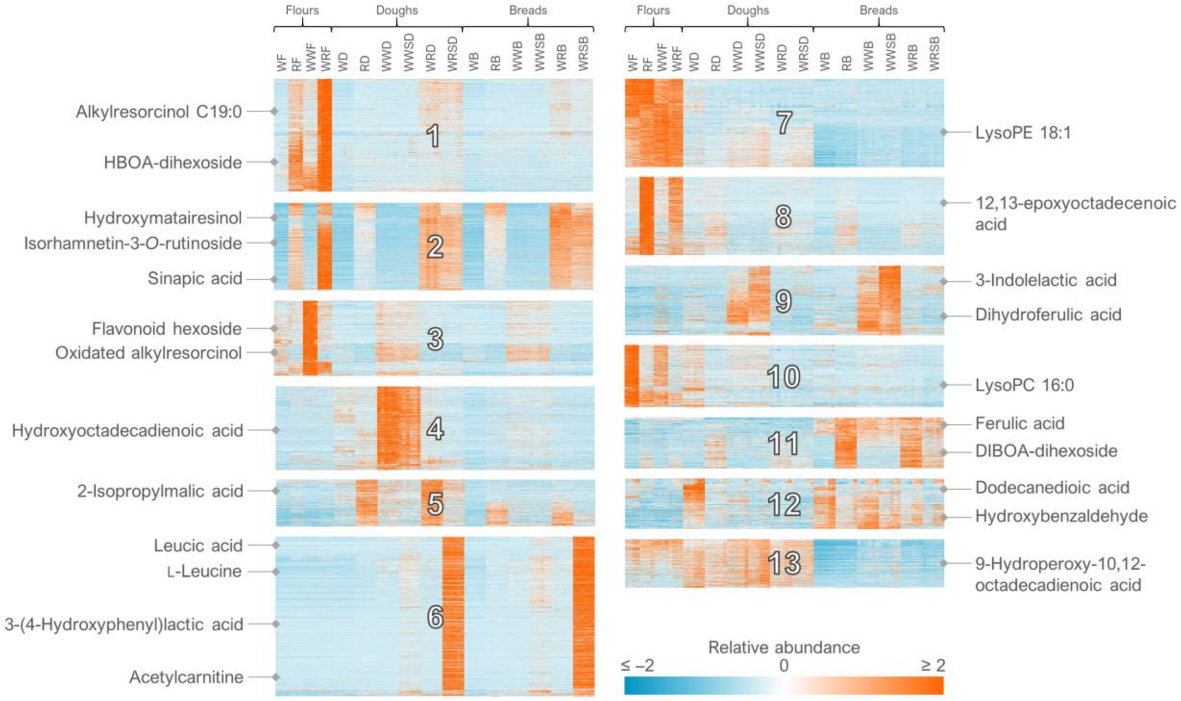

Figure 26. Heat map of all the 12 579 molecular features detected in RP negative mode from cereal samples with some of the annotated compounds highlighted. *k*-Means clustering was applied to the dataset, dividing it into distinct clusters (n = 13) based on the relative abundance of the features across samples.

77. For the clustering in Multiple Experiment Viewer, first normalize the rows (signal abundances) and select appropriate color scale limits for the normalized abundances (0 to 10% of features can be off limits). For the hierarchical clustering, choose whether to cluster only the features or samples as well; use Pearson correlation and average linkage clustering. For the *k*-means clustering, choose cluster genes, use Pearson correlation, calculate *k*-means, and choose a low number of clusters (*e.g.* 4) for the initial run. Repeat the procedure by increasing the number of clusters until no more clusters with a unique pattern emerge and choose the highest number of clusters based on this visual optimization.

3.5. Identification of metabolites

The identification and annotation of metabolites is a critical step in any metabolomics study to attribute biological meaning to the data analytical results and to enable further hypotheses to be developed for subsequent studies. In recent years, the development of new software and online tools as well as the emergence and expansion of publicly available spectral databases of metabolites have greatly facilitated the identification process (46,47). Nevertheless, metabolite identification remains perhaps the most time-consuming task where manual curation is necessary and where not all detected molecular features can be identified, leaving knowledge gaps for the interpretation of the results. Alongside with the challenges related to the instrumental differences and matching the obtained MS/MS data to databases, a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

key bottleneck restricting the level and number of identifications is the lack of reference data for the vast number of metabolites produced by living organisms, estimated up to one million for the plant kingdom (48) and more than 40 000 for humans (49). Likewise, matching the obtained MS/MS data to existing databases is not straightforward due to differences in experimental conditions used for collecting the reference data. Other limitations may be related to the poor quality (or lack) of mass spectra from metabolites with low abundance in the sample.

We utilize MS-DIAL (50) in the initial semi-automated step of metabolite identification, where the experimental characteristics (exact *m*/*z*, retention time where applicable, and MS/MS spectra in CID voltages 10, 20, and 40 V) are compared with those in databases available in NIST MSP format. These databases include MassBank (47), MoNA (51) and others available from the RIKEN Center for Sustainable Resource Science website (http://prime.psc.riken.jp/Metabolomics_Software/) combined in single files for the positive and negative ionization mode. Additionally, we have included our in-house spectral library in the MSP files. The semi-automated identification process annotates metabolites with similarity score 80% or above, after which the annotations are manually curated by assessing the similarity of the MS/MS spectra and the alternative annotations proposed by the software.

After the curation of the metabolites annotated by MS-DIAL, the remaining unknown metabolites undergo additional searches in databases that are primarily available online, including METLIN (46) for small metabolites and LIPID MAPS (52) for unknown metabolites with RP retention time in the lipid region (> 9 min). Additional attempts to characterize the unknowns are made utilizing MS-FINDER (50), which 1) calculates and scores the possible molecular formulas based on the exact mass and isotopic pattern, 2) searches for compounds corresponding to the likely molecular formulas from non-spectral chemical libraries, and 3) compares the experimental MS/MS spectrum of the unknowns with *in silico*-generated MS/MS spectra of the candidate structures.

3.5.1. Comparison with pure standard compounds (MSI level 1)

78. For the identification of metabolites (identification level 1 according to the Metabolomics Standards Initiative)(53), compare the molecular features against an in-house library (i.e. a reference standard analyzed previously with the same platform in the same chromatographic conditions). Apply the following criteria:
    a. matching *m*/*z* (within 10 ppm or according to instrument mass accuracy);
    b. similar retention time (ΔRT < 0.2–0.5 min), taking into consideration any possible near-eluting isomers;
    c. MS/MS spectra (main fragments matching within 0.02 Da in one or more CID voltage)

3.5.2. MS/MS fragmentation and database comparison (MSI levels 2-3)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

79. For the putative annotation of metabolites (ID level 2), compare the molecular features against publicly available spectral databases, including a database file (compiled in MSP format for using within MS-DIAL) and online databases. The annotation has acceptable reliability if the main fragments (excluding the molecular ion) match between the experimental and reference MS/MS spectra in only one proposed metabolite. In case several alternatives exist with similar MS/MS, the common denominator of all the alternatives (e.g. a compound class, ID level 3) is given as the annotation instead. Apply the following criteria:
    a. matching *m/z* (within 10 ppm or according to instrument mass accuracy)
    b. MS/MS spectra (main fragments matching within 0.02 Da)
80. For the putative characterization of compound class (ID level 3), use the following approaches to obtain characteristic information of the metabolite:
    a. Compare the experimental MS/MS with *in silico* generated spectra in MS-FINDER;
    b. Use the calculated molecular formula, retention time, and diagnostic MS/MS fragments to determine the compound class.

3.5.3. Pathway analysis

Once molecular features are annotated as metabolites, pathway analysis may be conducted to better understand the biological relevance of the metabolites, as well as their involvement in metabolic pathways, *e.g.* related to intervention effects of disease aetiology (1,3). We consider identification of metabolites until level 2 (putative annotation) to be essential prior to pathway analysis. Of the several pathway analysis tools that are freely available, we use predominantly MetaboAnalyst and Cytoscape. For both tools, HMDB or KEGG metabolite IDs are essential to avoid confusion from the multitude of nomenclature systems adopted different research groups.

81. Option 1: In MetaboAnalyst (54) (https://www.metaboanalyst.ca/) use Enrichment or Pathway Analysis which enables enrichment and visualization of metabolic pathways in which the metabolites could potentially be involved. For more detailed information about metabolic regulation, the Network Explorer enables inclusion of fold change data, along with gene expression data.
82. Option 2: Cytoscape (55) (https://cytoscape.org/) is a powerful stand-alone tool that is used by biomedical researchers to visualize and dynamically analyze gene/protein/metabolite interaction networks. The strength of Cytoscape is even more apparent when linked to databases, e.g. MetScape (56), which allows for visualizing and interpreting metabolomic data in the context of human metabolic networks.

The step-by-step instruction to use the software is listed in the Supplementary file. It is worth to mention that the pathway analysis may not be helpful for lipids, due to i) the limitation of non-targeted LC-MS metabolomics platform to differentiate the position of the double bonds within the lipid molecule, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

impairs the translation of lipid identity to KEGG or HMDB ID and; ii) that most pathway analysis tools would group certain lipid classes that vary greatly based on their fatty acid composition to one node, which may not be biologically meaningful. As an example, if there are five phosphatidylcholines with different acyl composition, the pathway analysis tool will group them as one node of phosphatidylcholine regardless of the acyl composition, so it may not give the whole picture of the acyl transfer. This gap hence emphasizes the need of pathway analysis tool specialized for lipid molecules.

3.6. Biological interpretation of the results

The analytical procedure described above is aimed to find out metabolites and metabolic pathways that are affected in the taken study set-up; *e.g*. differences in circulating metabolites after dietary or other intervention, or alterations caused to the phytochemical composition of a certain food due to the technological processing. Whilst the described workflow is efficient elucidating such metabolites, including non-hypothesized ones, the most critical step in terms of the value of the results is demonstration of the biological importance. Likewise, any analytical results, the findings need to be placed within the scientific context and interpreted in the light of existing biological knowledge. Optimally, the findings are validated in subsequent studies, where the most interesting/important metabolite species may be chosen for additional analysis, often encompassing development of targeted, quantitative analytical approaches, and analyzed in different study population within the same/similar biological context. As an example of such approach is the recent discovery of various trimethylated compounds related to whole grain consumption (57) and the establishment of a quantitative method within another cohort (58).

4. Conclusions

Non-targeted metabolic profiling analysis employing liquid chromatography combined with mass spectrometry has proven its usefulness in various fields of natural and medical sciences during the last couple of decades. It has greatly improved our capabilities to explore and understand a wider chemical space than ever before, in any biological sample. As introduced here, NoTaMe encompasses all the essential steps in metabolic profiling study extending from generation of data to the interpretation of the results, and is aimed to serve as general guideline for metabolomics study set-ups, as well as support the user with an in-house developed R-package for the different types of statistical analysis and visualizations useful for non-targeted metabolic profiling.

Author Contributions: MK wrote chapters of imputation and multivariate analysis**,** writing –review and editing the manuscript**. AK** wrote the chapters on quality control, drift correction, feature clustering algorithm, feature-wise analysis and visualizations, the main author of the notame R package and Wranglr web application. **SN, MT, TM** wrote the chapter on study design. **SN** and **IZ** described pathway analysis. **JP** conceived the original idea of the statistical and visualization pipeline and contributed to content of the manuscript. **VMK** wrote the chapter on the identification of metabolites and prepared illustrations, **AFB** and **TS** wrote the introduction. **SR** and **MH** wrote, reviewed and edited the section with LC-MS analysis, **KH** wrote introduction, review, editing and supervision. **OK, JP**, **DB** and **CB** writing—review, editing and supervision. All authors read, reviewed, and accepted the final manuscript

Funding: Academy of Finland, Biocenter Finland, EU Horizon 2020, Finnish Cultural Foundation, Lantmännen Research Foundation

Acknowledgments: Development of the R package and Wranglr as well as data analyses were carried out with the support of Bioinformatics Center, University of Eastern Finland, Finland. Miia Reponen is thanked for the preparation of samples and LC-MS operation.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. KH, OK, VK, AK and JP are owners of a spin-off company providing metabolomics services, Afekta Technologies Ltd.

References

{Bibliography}

1. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: Beyond biomarkers and towards mechanisms. Vol. 17, Nature Reviews Molecular Cell Biology. Nature Publishing Group; 2016. p. 451–9.

2. Manach C, Hubert J, Llorach R, Scalbert A. The complex links between dietary phytochemicals and human health deciphered by metabolomics. Vol. 53, Molecular Nutrition and Food Research. 2009. p. 1303–15.

3. Gika H, Virgiliou C, Theodoridis G, Plumb RS, Wilson ID. Untargeted LC/MS-based metabolic phenotyping (metabonomics/metabolomics): The state of the art. J Chromatogr B [Internet]. 2019 Jun 1 [cited 2019 Oct 8];1117:136–47. Available from: https://www.sciencedirect.com/science/article/abs/pii/S1570023219301370

4. Nash WJ, Dunn WB. From mass to metabolite in human untargeted metabolomics: Recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data. TrAC Trends Anal Chem. 2018 Nov;

5. Johnson CH, Ivanisevic J, Benton HP, Siuzdak G. Bioinformatics: The Next Frontier of Metabolomics. Anal Chem [Internet]. 2015 Jan 6 [cited 2019 Oct 8];87(1):147–56. Available from: https://pubs.acs.org/doi/10.1021/ac5040693

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

6. Chaleckis R, Meister I, Zhang P, Wheelock CE. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. Curr Opin Biotechnol [Internet]. 2019 Feb 1 [cited 2019 Oct 8];55:44–50. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0958166918300764

7. Misra BB, Mohapatra S. Tools and resources for metabolomics research community: A 2017–2018 update. Electrophoresis. 2019;40(2):227–46.

8. Dias DA, Jones OAH, Beale DJ, Boughton BA, Benheim D, Kouremenos KA, et al. Current and future perspectives on the structural identification of small molecules in biological systems. Vol. 6, Metabolites. MDPI AG; 2016.

9. Ulaszewska MM, Weinert CH, Trimigno A, Portmann R, Andres Lacueva C, Badertscher R, et al. Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies. Vol. 63, Molecular Nutrition and Food Research. Wiley-VCH Verlag; 2019.

10. Broadhurst D, Goodacre R, Reinke SN, Kuligowski J, Wilson ID, Lewis MR, et al. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. Metabolomics [Internet]. 2018 Jun 18 [cited 2019 Oct 8];14(6):72. Available from: http://link.springer.com/10.1007/s11306-018-1367-3

11. Koistinen VM, Hanhineva K. Microbial and endogenous metabolic conversions of rye phytochemicals. Vol. 61, Molecular Nutrition and Food Research. Wiley-VCH Verlag; 2017.

12. Koistinen VM, Hanhineva K. Mass spectrometry-based analysis of whole-grain phytochemicals. Crit Rev Food Sci Nutr. 2017 May 24;57(8):1688–709.

13. de Mello VD, Paananen J, Lindström J, Lankinen MA, Shi L, Kuusisto J, et al. Indolepropionic acid and novel lipid metabolites are associated with a lower risk of type 2 diabetes in the Finnish Diabetes Prevention Study. Sci Rep [Internet]. 2017 Dec 11 [cited 2018 Aug 24];7(1):46337. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28397877

14. Noerman S, Kärkkäinen O, Mattsson A, Paananen J, Lehtonen M, Nurmi T, et al. Metabolic Profiling of High Egg Consumption and the Associated Lower Risk of Type 2 Diabetes in Middle-Aged Finnish Men. Mol Nutr Food Res [Internet]. 2018 Dec 12 [cited 2019 Dec 12];1800605. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mnfr.201800605

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

15. Rothwell JA, Keski-Rahkonen P, Robinot N, Assi N, Casagrande C, Jenab M, et al. A Metabolomic Study of Biomarkers of Habitual Coffee Intake in Four European Countries. Mol Nutr Food Res. 2019 Nov 1;63(22).

16. Brunius C, Shi L, Landberg R. Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. Metabolomics. 2016;

17. R: The R Project for Statistical Computing [Internet]. [cited 2019 Dec 19]. Available from: https://www.r-project.org/

18. Web Application Framework for R [R package shiny version 1.4.0].

19. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods [Internet]. 2015 Jun 4 [cited 2019 Oct 8];12(6):523–6. Available from: http://www.nature.com/articles/nmeth.3393

20. Kirwan JA, Broadhurst DI, Davidson RL, Viant MR. Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. Anal Bioanal Chem. 2013 Jun;405(15):5147–57.

21. Puupponen-Pimiä R, Seppänen-Laakso T, Kankainen M, Maukonen J, Törrönen R, Kolehmainen M, et al. Effects of ellagitannin-rich berries on blood lipids, gut microbiota, and urolithin production in human subjects with symptoms of metabolic syndrome. Mol Nutr Food Res [Internet]. 2013 Dec [cited 2018 Nov 13];57(12):2258–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23934737

22. Hotelling H. Relations Between Two Sets of Variates. Biometrika [Internet]. 1936 Dec [cited 2019 Oct 8];28(3/4):321. Available from: https://www.jstor.org/stable/2333955?origin=crossref

23. Bro R, Smilde AK, Smilde AK, Hubert M, Song X, Yu R, et al. Principal component analysis. Anal Methods [Internet]. 2014 [cited 2017 Jun 1];6(9):2812. Available from: http://xlink.rsc.org/?DOI=c3ay41907j

24. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. London, Edinburgh, Dublin Philos Mag J Sci [Internet]. 1901 Nov 8 [cited 2019 Oct 8];2(11):559–72. Available from: https://www.tandfonline.com/doi/full/10.1080/14786440109462720

25. Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res [Internet]. 2008 [cited 2019 Oct 8];9(Nov):2579–605. Available from:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

http://www.jmlr.org/papers/v9/vandermaaten08a.html

26. Rokach L, Maimon O. Clustering Methods. In: Data Mining and Knowledge Discovery Handbook [Internet]. New York: Springer-Verlag; 2005 [cited 2019 Oct 8]. p. 321–52. Available from: http://link.springer.com/10.1007/0-387-25465-X_15

27. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif [Internet]. 2014 Oct 18 [cited 2019 Oct 8];31(3):274–95. Available from: http://link.springer.com/10.1007/s00357-014-9161-z

28. Kokla M, Virtanen J, Kolehmainen M, Paananen J, Hanhineva K. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. Available from: https://doi.org/10.1186/s12859-019-3110-0

29. Stekhoven DJ, Bühlmann P. Data and text mining MissForest-non-parametric missing value imputation for mixed-type data. 2012 [cited 2019 Apr 11];28(1):112–8. Available from: http://stat.ethz.ch/CRAN/.

30. Armitage EG, Godzien J, Alonso-Herranz V, López-Gonzálvez Á, Barbas C. Missing value imputation strategies for metabolomics data. Electrophoresis [Internet]. 2015 Dec 1 [cited 2017 Oct 24];36(24):3050–60. Available from: http://doi.wiley.com/10.1002/elps.201500352

31. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak [Internet]. 2016 Jul 25 [cited 2017 Nov 17];16 Suppl 3(Suppl 3):74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27454392

32. Sysi-Aho M, Katajamaa M, Yetukuri L, Orešič M. Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics [Internet]. 2007 Mar 15 [cited 2019 Oct 14];8(1):93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17362505

33. Guida R Di, Engel J, Allwood JW, Weber RJM, Jones MR, Sommer U, et al. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics [Internet]. 2016 [cited 2019 May 31];12:93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27123000

34. Kohl SM, Klein MS, Hochrein J, Oefner PJ, Spang R, Gronwald W. State-of-the art data normalization methods improve NMR-based metabolomic analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

Metabolomics [Internet]. 2012 Jun 12 [cited 2019 Oct 8];8(S1):146–60. Available from: http://link.springer.com/10.1007/s11306-011-0350-z

35. Tyralis H, Papacharalampous G, Langousis A. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. Water [Internet]. 2019 Apr 30 [cited 2019 Oct 9];11(5):910. Available from: https://www.mdpi.com/2073-4441/11/5/910

36. Vinaixa M, Samino S, Saez I, Duran J, Guinovart JJ, Yanes O. A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. Metabolites [Internet]. 2012 Oct 18 [cited 2019 May 31];2(4):775–95. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24957762

37. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models. J Stat Softw [Internet]. 2017 [cited 2019 Oct 8];82(13). Available from: http://www.jstatsoft.org/v82/i13/

38. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models using lme4. 2014 Jun 23 [cited 2019 Oct 8]; Available from: http://arxiv.org/abs/1406.5823

39. Noerman S, Kärkkäinen O, Mattsson A, Paananen J, Lehtonen M, Nurmi T, et al. Metabolic Profiling of High Egg Consumption and the Associated Lower Risk of Type 2 Diabetes in Middle-Aged Finnish Men. Mol Nutr Food Res [Internet]. 2018 Dec 12 [cited 2019 Oct 20];63(5):1800605. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mnfr.201800605

40. Claggett BL, Antonelli J, Henglin M, Watrous JD, Lehmann KA, Musso G, et al. Quantitative Comparison of Statistical Methods for Analyzing Human Metabolomics Data. 2017 Oct 10 [cited 2019 Oct 9]; Available from: http://arxiv.org/abs/1710.03443

41. Stoessel D, Stellmann J-P, Willing A, Behrens B, Rosenkranz SC, Hodecker SC, et al. Metabolomic Profiles for Primary Progressive Multiple Sclerosis Stratification and Disease Course Monitoring. Front Hum Neurosci [Internet]. 2018 [cited 2019 Oct 14];12:226. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29915533

42. Shi L, Westerhuis JA, Rosén J, Landberg R, Brunius C. Variable selection and validation in multivariate modelling. Kelso J, editor. Bioinformatics [Internet]. 2019 Mar 15 [cited 2019 Oct 8];35(6):972–80. Available from: https://academic.oup.com/bioinformatics/article/35/6/972/5085367

43. Breiman L, Leo. Random Forests. Mach Learn [Internet]. 2001 [cited 2019 Oct 14];45(1):5–32. Available from: http://link.springer.com/10.1023/A:1010933404324

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

44. Carl Brunius / MUVR · GitLab [Internet]. [cited 2019 Oct 9]. Available from: https://gitlab.com/CarlBrunius/MUVR

45. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in Partial Least Squares Regression. Chemom Intell Lab Syst [Internet]. 2012 Aug [cited 2018 Sep 19];118:62–9. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0169743912001542

46. Smith CA, Maille GO, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN. Ther Drug Monit [Internet]. 2005 Dec [cited 2019 Oct 8];27(6):747–51. Available from: https://insights.ovid.com/crossref?an=00007691-200512000-00016

47. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom [Internet]. 2010 Jul 7 [cited 2019 Oct 8];45(7):703–14. Available from: http://doi.wiley.com/10.1002/jms.1777

48. Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, et al. KNApSAcK Family Databases: Integrated Metabolite–Plant Species Databases for Multifaceted Plant Research. Plant Cell Physiol [Internet]. 2012 Feb 1 [cited 2019 Oct 8];53(2):e1–e1. Available from: https://academic.oup.com/pcp/article-lookup/doi/10.1093/pcp/pcr165

49. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, et al. HMDB 3.0—The Human Metabolome Database in 2013. Nucleic Acids Res [Internet]. 2012 Nov 17 [cited 2019 Oct 8];41(D1):D801–7. Available from: http://academic.oup.com/nar/article/41/D1/D801/1055560/HMDB-30The-Human-Metabolome-Database-in-2013

50. Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, et al. Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. Anal Chem [Internet]. 2016 Aug 16 [cited 2019 Oct 8];88(16):7946–58. Available from: https://pubs.acs.org/doi/10.1021/acs.analchem.6b00770

51. MassBank of North America [Internet]. [cited 2019 Oct 8]. Available from: https://mona.fiehnlab.ucdavis.edu/

52. Fahy E, Sud M, Cotter D, Subramaniam S. LIPID MAPS online tools for lipid research. Nucleic Acids Res [Internet]. 2007 May 8 [cited 2019 Oct 8];35(Web Server):W606–12. Available from: https://academic.oup.com/nar/article-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2020 doi:10.20944/preprints202002.0019.v1

lookup/doi/10.1093/nar/gkm324

53. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis. Metabolomics [Internet]. 2007 Sep 19 [cited 2019 Oct 8];3(3):211–21. Available from: http://link.springer.com/10.1007/s11306-007-0082-2

54. Chong J, Yamamoto M, Xia J. MetaboAnalystR 2.0: From Raw Spectra to Biological Insights. Metabolites [Internet]. 2019 Mar 22 [cited 2019 Oct 8];9(3):57. Available from: https://www.mdpi.com/2218-1989/9/3/57

55. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498–504.

56. Gao J, Tarcea VG, Karnovsky A, Mirel BR, Weymouth TE, Beecher CW, et al. Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics [Internet]. 2010 Apr 1 [cited 2019 Oct 8];26(7):971–3. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq048

57. Kärkkäinen O, Lankinen MA, Vitale M, Jokkala J, Leppänen J, Koistinen V, et al. Diets rich in whole grains increase betainized compounds associated with glucose metabolism. Am J Clin Nutr. 2018;108(5):971–9.

58. Tuomainen M, Kärkkäinen O, Leppänen J, Auriola S, Lehtonen M, Savolainen MJ, et al. Quantitative assessment of betainized compounds and associations with dietary and metabolic biomarkers in the randomized study of the healthy Nordic diet (SYSDIET). Am J Clin Nutr. 2019 Nov 1;110(5):1108–18.

© 2020 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).