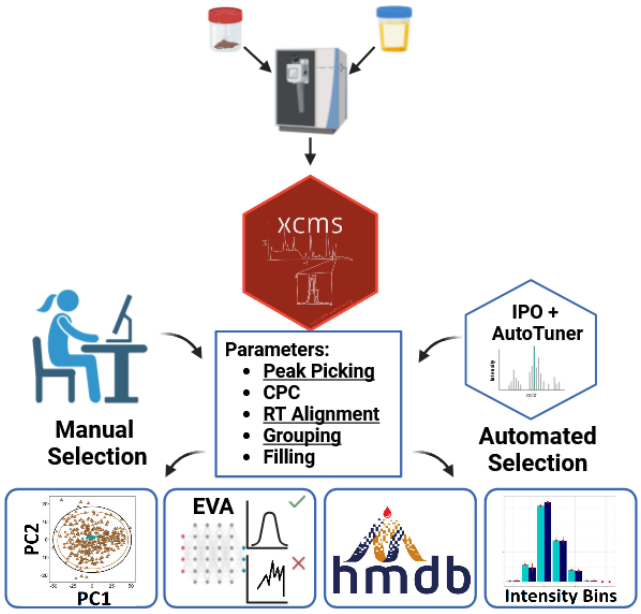
Towards automated preprocessing of untargeted LC-MS-based metabolomics feature lists from human biofluids

Amy Hughes[1], Pablo Vangeenderhuysen[2], Marilyn De Graeve[2,3], Beata Pomian[2], Tim S Nawrot[4,5], Jeroen Raes[6,7], Simon J S Cameron[1], Lynn Vanhaecke[1,2*]

1. Institute for Global Food Security, School of Biological Sciences, Queen's University Belfast, BT9 5DL, Northern Ireland. **2.** Laboratory of Integrative Metabolomics (LIMET), Ghent University, 9820 Merelbeke, Belgium. **3.** Institute for Biomedicine, Eurac Research, 39100 Bolzano, Italy. **4.** Centre for Environmental Sciences, Hasselt University, Diepenbeek, 3590 Diepenbeek, Belgium. **5.** School of Public Health, Occupational and Environmental Medicine, Leuven University, 3000 Leuven, Belgium. **6.** Laboratory of Molecular Bacteriology, Rega Institute, Katholieke Universiteit Leuven, 3000 Leuven, Belgium. **7.** Centre for Microbiology, Vlaams Instituut voor Biotechnologie (VIB), 3001 Leuven, Belgium.

ABSTRACT: Maximizing the extraction of true, high-quality, non-redundant features from biofluids analyzed via LC-MS systems is challenging. Here, the R packages IPO and AutoTuner were used to optimize XCMS parameter settings for the retrieval of metabolite or lipid features in both ionisation modes from either faecal or urine samples from two cohorts (n = 621). The feature lists obtained were compared with those where the parameter values were selected manually. Three categories were used to compare feature lists: 1) feature quality through removing false positives, 2) tentative metabolite identification using the Human Metabolome Database (HMDB) and 3) feature utility such as analyzing the proportion of features within intensity threshold bins. Furthermore, a PCA-based approach to feature filtering using QC samples and variable loadings was also explored under this category. Overall, more features were observed after automated selection of parameter values for all datasets (1.3- 3.7-fold), which propagated through comparative exercises. For example, a greater number of features (on average 51 *vs.* 45 %) had a coefficient of variation (CV) < 30 %. Additionally, there was a significant increase (7.6 – 10.4 %) in the number of faecal metabolites that could be tentatively annotated, and more features were present in higher intensity threshold bins. Considering the overlap across all three categories, a greater number of features were also retained. Automated approaches that guide selection of optimal parameter values for preprocessing are important to decrease the time invested for this step, whilst taking advantage of the wealth of data that LC-MS systems provide.



Metabolomic analyses of biofluids can provide useful insights into the functioning of an organism's metabolic pathways, helping to monitor physiological responses and phenotypes associated with health and disease states (1, 2). Ultra-high performance liquid chromatography coupled to high resolution mass spectrometry (UHPLC-HRMS) is regarded as one of the most superior LC-MS technologies, owing to its efficient separation and high sensitivity detection of metabolites, facilitating broad coverage of the metabolome (3, 4).

Bioinformatic tools must be applied to the large datasets obtained from LC-MS systems upon metabolomic analyses. The aim is to extract a list of true, non-redundant features in tabular format containing the metabolic variable information *m/z* and retention time, and for each sample, a value for the peak area or maximum peak intensity, from the thousands of spectral peaks which are contained within the LC-MS raw datafiles (5). Noise from contaminants, unwanted signal spikes, false positive peaks, and redundant information in the form of

adducts, fragments and isotopes must be accounted for to maximize the extraction of high-quality features (1, 3). In turn, these features need to have a strong signal to aid downstream statistical analyses and biological interpretation (5, 6).

The steps involved in data preprocessing include peak picking, retention time alignment, grouping and gap filling (7, 8). Different software exists ranging from open source such as MZmine 3 (9), XCMS (10) and MS-DIAL (11), to commercial ones such as Compound Discoverer and MassLynx (7). Each software uses different algorithms for certain stages such as peak picking (12). Thus, dissimilar feature lists can be obtained. The factors implicated in the choice of application include cost, coding skills and algorithm transparency (7). Software where the latter applies would improve reproducibility of the preprocessing steps for untargeted metabolomic studies.

For this study, XCMS was chosen as a free, widely used, open-source R package (8, 10), that has performed well in detecting true positive features relative to alternatives (13), and scored approximately 70 % when considering the FAIR research principles governing the use of software that is easy to find, accessible, interoperable and reusable/modifiable (14).

Choosing data appropriate parameter settings for preprocessing is important to maximize the value of the raw metabolomic data and optimize the extraction of useful features (15). Whilst time-consuming, this can be done manually through e.g., visualizing the extracted ion chromatogram (EIC) of internal standards; however, it requires a high level of expertise. Alternatively, the process can be automated using software packages such as AutoTuner or IPO (16, 17).

In this study, both packages were used to automate optimization of XCMS parameters. Whilst the most computationally intensive, IPO is more established and has been employed in multiple studies including e.g. characterizing metabolite alterations associated with bariatric surgery (18). Few attempts have been made to compare LC-MS feature lists obtained from manual *vs.* automated selection of XCMS parameters (19, 20). One study used Kruskal-Wallis tests to compare the features in plasma and liver that were significantly different between groups of piglets that were either newborn, neonates, or infants (19). Whilst the other used machine learning approaches such as random forest to compare how well the features obtained from fingerprints could classify between males and females or time points (20). Thus, both studies have made the manual *vs.* automated XCMS parameter selection comparison through the ability of the features obtained to discriminate between different groups.

To the best of our knowledge, no approach has been reported that compares preprocessing of both metabolomic and lipidomic features in an unsupervised manner, i.e. in the absence of a defined endpoint. Two different complex human biofluids (n = 621 in total) collected from two Flemish prospective cohorts were utilized; faecal samples obtained from the longitudinal Flemish Gut Flora Project (FGFP) (21) and urine samples from children of the Environmental Influence on Early Ageing (ENVIRonAGE) cohort (22). The aim was to compare feature lists through methods assigned to one of three categories 1) feature quality (removal of false positives), 2) tentative metabolite identification using the HMDB and 3) feature utility/filtering through e.g. comparing how features were distributed across intensity threshold bins and evaluating the usefulness of a novel PCA-based approach to feature selection/filtering. The number of features that overlapped across methods and categories was also evaluated for each dataset and compared between workflows.

Methodology

Reagents The analytical standards and reagents used are documented by Vangeenderhuysen et al (23).

Samples A total of 292 faecal samples were collected from the FGFP participants, as detailed by Falony et al (21). Urine samples from 329 children (aged 4-10 years) from the ENVIRonAGE cohort were obtained as described by Janssen et al (22). Upon collection of samples, cold chain monitoring was implemented for transport (21). Upon arrival, urine samples were frozen (-80 °C), whilst faecal samples were freeze-dried, ground, and sieved, then stored at -80 °C. Quality control (QC) samples were formed by pooling the biological samples together (n= 58 for faecal and n = 66 for urine). Two QCs were injected at regular intervals after 10 proceeding biological randomized samples to monitor the stability of the LC-MS system and provide a measure of repeatability and signal correction within and between each batch (5, 24).

Sample Extraction The protocol for dual faecal metabolome and lipidome extraction is described in detail by Vangeenderhuysen et al (23). Briefly, for the faecal metabolome, 100 mg of faeces was added to 2 mL of ultra-pure water (UPW), followed by 12.5 µl of a 100 ng/µl mixture of 6 internal standards (ISTDS: l-tyrosine-d2, indole 3-acetic acid-d5, dopamine-d4, deoxycholic acid-d4, l-phenylalanine-d2 and alanine-d3). 0.5 mL of a 75:25 v/v mixture of ice-cold methanol and UPW was then added. Thereafter, the solution was vortexed (1 min), rotated (10 mins) and centrifuged (10 min at 10,000 x g). The supernatant was collected and passed through a polyamide filter (25 mm diameter, 0.45 µm pore size, Macherey-Nagel, Germany). 500 µl of this diluted extract was transferred to a glass LC vial. To extract the faecal lipidome, the residual fraction of the stool sample following faecal metabolome extraction was subjected to the protocol outlined in Supplementary Text 2. Urine sample extraction is detailed by De Paepe et al (25). In short, 300 µL of urine was added to a 1.5 mL Eppendorf tube, together with 30 µL of 6 ISTDs. Next, samples were centrifuged for 8 minutes at 1000 x g at 4°C. 100 µL of supernatant was collected and diluted (1:10) with UPW, and then transferred to a glass LC vial.

Instrumentation The UHPLC used for the faecal metabolome was the Vanquish Horizon with an Acquity high strength silica (HSS) T3 column (detailed by De Paepe et al (25)) coupled to the Orbitrap Exploris 120 MS (Thermo Fisher Scientific, San José, CA, USA) (23). MS detection was preceded by heated electrospray ionisation (HESI) in polarity switching mode and the following instrument parameters were used: sheath gas flow rate (55 arbitrary units, au), auxiliary gas flow rate (25 au) and sweep gas flow rate (3 au), heater and capillary temperature (both 300 °C), S-lens RF level 50 V, spray voltage of 2.9 kV for both ionisations modes, m/z scan range of 53 to 800 Da, maximum injection time of 70 ms, with automatic gain control target of $1e^6$ ions and mass resolution of 120,000 full width at half-maximum (FWHM, 1 Hz) (23). Van Meulebroek et al (26) details the instrumentation used for faecal lipidome separation and detection, also present in Supplementary Text 2. The UHPLC for urine metabolites was the Vanquish Flex (HSS T3 column), coupled to the Orbitrap Q-Exactive MS (Thermo Fisher Scientific, San José, CA, USA) (25). The MS detection was preceded by HESI in polarity switching mode. Some instrument parameters remained the same as the faecal metabolome; those that differed were a sheath gas flow of 50 au, a heater and capillary temperature

Table 1: Preprocessing parameters used for each workflow*.

Setting	Faecal Metabolites + Lipids	Faecal Metabolites	Faecal Lipids	Urinary Metabolites
Work-flow	1	2	2	2
PEAK PICKING SETTINGS				
ppm	6	5	5	5
peak width (min, max)	5, 45	3, 50	20, 65	4, 55
prefilter	3, 1000	3, 100	3, 100	3, 100
mzdiff	0.0500	-0.0065	0.0450	0.0050
fitgauss	TRUE	FALSE	FALSE	FALSE
RETENTION TIME ALIGNMENT				
binSize	1 (Default)	0.79	0.79	0.53
gap Init	0.30 (Default)	0.67	0.42	0.81
gap Extend	2.40 (Default)	2.34	2.50	2.00
GROUPING				
bw	30.00 (Default)	0.88	0.88	0.88
mzwid	0.050	0.010	0.010	0.015

**Supplementary Table 2A contains definitions of each parameter.*

of 350 °C and 250 °C, a spray voltage of ± 4.0 kV, and mass resolution of 140 000 FWHM (25). The sample injection volume for both faeces and urine was 10 µl.

Data Conversion and Preprocessing Raw Thermo Fisher files were converted into mzML formats using default settings of Proteowizard MS convert prior to preprocessing (27). Centroiding was applied to each data file, in each ionisation mode. R Studio (version 4.2.2.) was used for data analysis and plots. The R package XCMS (version 3.2.0) was used for preprocessing (10).

After peak picking using the "centWave" algorithm (8, 28), low quality, false positive peaks were filtered with the comprehensive peak characterization (CPC) algorithm (version 0.1) (29) and any incorrectly split peaks were detected and merged ((30). Retention time alignment was performed using the "obiwarp" method (31), to ensure the same peak *m/z* across different samples had the same retention time (8). Subsequently, peaks that clustered closely around a retention time were grouped into a feature (8). Gap filling was then used to account for missing peaks that were not initially detected or aligned correctly (7, 8, 10). The feature lists consisting of *m/z*-retention time pairs, with corresponding peak area/height information for each sample were then annotated by CAMERA (version 1.54), where possible isotopes, fragments and adducts were identified (32).

Manual Selection of XCMS Parameters Manual selection was conducted by a scientist in the field with expert knowledge. Initially, scientific articles which utilized the same or very similar UHPLC-HRMS instruments and had used XCMS for preprocessing were reviewed for their parameters e.g., (33) , in conjunction with tutorials on XCMS parameter selection. Next, the selected parameters were fine-tuned based on manual inspection of EICs of known compounds in biological faecal samples, QCs, and analytical standards (Supplementary Table 1) from a subset of samples collected from the Rombouts et al study (34).

Automated Selection of XCMS Parameters For each sample type (faeces/urine), each ionisation mode (+/-) and each category (metabolites/lipids), IPO was run in R using all QC samples (version 1.24). This took between 6 – 17 hours depending upon the dataset. AutoTuner was run using 15 randomly selected QC samples (version 1.3). Whilst more interactive, AutoTuner took on average < 1 hour per dataset. Both outputs were considered and some parameter values such as ppm were further refined in line with the high resolution of the equipment used. The definitions of the parameters altered, and the settings used to run IPO/Autotuner are shown in Supplementary Table 2A and B respectively.

Comparative Analyses For each dataset comparison either all features were used, or those with a CV < 30 % in the QC samples. Peak height was the chosen metric.

Here within, workflow 1 refers to the manual selection of parameters, whilst workflow 2 refers to the automation. Preprocessing parameters were kept the same for both ionisation modes, particularly as the outcomes from IPO/Autotuner were remarkably similar for both (Table 1). Workflow 1 settings were used for both metabolite and lipid preprocessing. Noise (15,000) and S/N threshold (10) were kept the same for both workflows. IPO/Autotuner gave low noise (~5000) and too high ppm values. The use of a ppm that is close to the mass accuracy of the mass spectrometer is usually advised, thus 5 ppm was chosen for workflow 2. When grouping peaks into features, the minimum sample fraction was set to 0.2 (20 %) for both workflows.

PCA The normalization, scaling and transformations applied to data before principal component analysis (PCA) were determined through a combination of visualizing how well QCs clustered, assessing the number of outliers, the % of variation explained by PC1 and 2 and the number of metabolites with normal distributions as determined by Shapiro-Wilk tests. In addition, a *k*-nearest neighbors' approach (*k*=3) was used to calculate an accuracy score based on the confusion matrix from classifying samples as either QC or biological sample. From each cross-validation (n=7), the average score was calculated (35) . The same combination was used for each workflow for fair comparison. Normalization methods tested included total ion current (TIC) normalization, locally estimated scatter plot smoother (LOESS) and QC normalization (24, 36). Transformations assessed included logarithmic, inverse hyperbolic sine and Yeo-Johnson, whilst scaling methods included range, auto and pareto.

EVA EVA (PyCharm version 2.5.0 in Anaconda 3 environment (37)), is a deep learning neural network approach which has been trained on over 25,000 manually recognized EIC peaks from data of various sample types, LC-MS configurations and spectra acquisition rates, was utilized to estimate the number of true and false positive peaks from each workflow (37) based on good or poor quality EIC shape respectively. A feature was deemed true if it had at least one peak with a good EIC shape. A two proportion Z-test was used to compare workflow 1 *vs.* 2, considering the proportion of true or false metabolic features for each category (metabolites/lipids) and ionisation mode (+/-).

HMDB Tentative annotation (level 4 as proposed by Schymanski et al (38)), was carried out using the different adduct *m/z's* for each entity/accession in the HMDB. Ppm values were calculated for all features of each workflow with

each adduct in separate ionisation modes. To mitigate unlikely adduct matches, we prioritized $[M + H]^+$ and $[M - H]^-$ adduct matching over others if the ppm < 5, given their higher likelihood of formation during electrospray ionisation (7). For features with multiple HMDB accession matches, the lowest ppm was used to annotate, with prioritization given to $[M + H]^+$ and $[M - H]^-$ adducts. In cases where multiple matches shared the same molecular formula, ppm difference, and adduct type, annotations could not be further refined due to the absence of fragmentation data and the lack of universal retention times in HMDB. As these annotations were intended only for comparing numbers across workflows and not for biological interpretation, the first HMDB accession within the group of matches sharing the same molecular formula was selected. This approach ensured consistency and avoided over-inflating annotation counts. To elucidate proportion of metabolites annotated per workflow, a < 1 ppm for metabolites and < 10 ppm for lipids cut-off was utilised. Proportion Z-tests were used for comparisons.

ICC To estimate the biological variation in the features that would remain after CV filtering and to compare these between workflows, the intra-class correlation coefficient (ICC) was calculated using the method outlined in Schifmann et al (39). Only the features left after CV filtering were used, because there was too much multi-collinearity using some of the full feature lists. Since ICC considers both biological and technical sources of variation, a feature with a high ICC is indicative of one that may have higher biological variation compared to one with a low ICC (39).

Signal Strength Intensity threshold bins were created for each category (metabolites/lipids) and ionisation mode (+/-). QC samples were used to determine the number of features per bin, where the average number of features across all QCs for each intensity bin were calculated. Proportion Z-tests were used to compare workflow 1 *vs.* 2 at each threshold.

Feature Selection PCA was used to test an alternative feature filtering approach to using CV < 30 %. A combination of QC normalization, log transformation and pareto scaling was applied to the QC data, as it caused the QC samples to cluster best. The number of PCs to retain was determined through permutation testing (n=1000), keeping only the PCs which were significant ($p < 0.05$). For each feature, the sum of loadings across those PCs was then calculated. The threshold was determined arbitrarily at half the maximum summed loading value. Cut-offs were unique for each category/ionisation mode and workflow. It was assumed that features with a small loading would show low variation across QCs and not drive any separation; thus, the CV of features below this threshold was also checked for congruency.

All statistical tests (except permutation testing) were adjusted for multiple comparison via the Benjamini-Hochberg procedure (false discovery rate, FDR), with an FDR q value < 0.01 deemed significant.

Based on these approaches, we assigned methods to compare workflows to one of three categories:

- ***Feature Quality*** (EVA and CV calculations checking false positives)
- ***Metabolite Identification*** (tentative based on HMDB annotations).
- ***Feature Utility / Filtering for Downstream Analyses*** (Intensity thresholds, ICC calculations, PCA filtering method).

The comparisons between workflows were initially evaluated using each method singularly and then examined via the overlap of features across categories that matched criteria. For example, examining the overlap between features that were recognized as true by EVA, had a CV < 30 %, could be tentatively annotated by HMDB at the respective ppm, were in the top 3 or 4 intensity bins with signal exceeding 2 million, and were retained after filtering using PCA loadings. To include the ICC results in the evaluations, the overlap was also examined using only features remaining after CV filtering. UpSet plots were used to visualize the overlaps for each dataset and workflow (40, 41).

Results and Discussion

Impact of parameter selection on feature numbers More features were present in workflow 2 (fold change ranging from 1.3 for metabolites to 3.7 for lipids) (Table 2). Decreasing peak width and increasing band width were found to be the most influential parameters driving the difference in feature number between workflow 1 *vs.* 2 (discussed in Supplementary Text 1). Similarly, the use of IPO increased the number of features observed compared to manually chosen XCMS settings in both liver and plasma tissue in positive and negative mode respectively, whilst AutoTuner led to the biggest increase in features detected in human fingerprints (19, 20).

Overlap between feature lists The *m/z* values for each workflow, category and ionisation mode, were rounded to 3 decimals prior to assessing common features among workflows. Urinary metabolites had greater overlap, with 69 and 65 % commonality in positive and negative mode, respectively (Supplementary Fig. 1). Faecal lipids showed the least overlap (22 % for both ion modes, Supplementary Fig. 1). The scale of the continuous wavelet transform applied to the ROI during peak picking is influenced by peak width (28), which could lead to retention time deviations. The latter affects peak grouping and the feature *m/z* values that those groups become; hence, the observed dissimilarity and less overlap for those datasets (10).

Considering variability across the QCs as technical replicates, on average 45 and 51 % remained after removing features with CV ≥ 30 % for workflow 1 and 2, respectively (Table 2). A similar % of overlap between workflows was also observed when only using these CV-filtered features (Supplementary Fig. 2).

Contrary to this, Alboniga et al showed that alternative methods of selecting XCMS parameters had minimal effect on the number of features left after CV filtering, although they used a lower CV threshold of 20 % (19). Furthermore, whilst features are expected to exert low variation across technical replicates like pooled sample QCs, they should also exhibit high variability across biological samples (39), which can be assessed through ICC. The distribution and density of the ICC values per workflow were compared for each dataset using violin plots (Supplementary Fig. 3). Notably, for lipids in negative mode, workflow 1 had a larger number of features with a lower ICC, whilst workflow 2 demonstrated a greater number with larger ICC values. Only faecal metabolites in positive mode and lipids in negative mode had significantly different proportions of features with an ICC > 0.5 (Supplementary Table 3).

Table 2: Number and percentage of features for each workflow before and after CV filtering.

Dataset	Workflow 1	CV < 30 %	Workflow 2	CV < 30 %
Positively ionized faecal metabolites (FGFP)	8732	2682 (31 %)	15343	6324 (41%)
Negatively ionized faecal metabolites (FGFP)	4445	1884 (42 %)	7610	4779 (63 %)
Positively ionized faecal lipids (FGFP)	2351	1361 (58 %)	7983	4693 (59 %)
Negatively ionized faecal lipids (FGFP)	1470	795 (54 %)	5435	2680 (49 %)
Positively ionized urinary metabolites (ENVIRonAGE)	1157	427 (37 %)	1543	640 (41 %)
Negatively ionized urinary metabolites (ENVIRonAGE)	1442	670 (46 %)	2103	1072 (51 %)

Multivariate analysis of sample clustering PCA score plots were used to compare workflows based on the number of outliers outside of Hotelling's T^2 ellipse, and on how well QCs clustered together as an indicator of the LC-MS system stability (5). Normalization, transformation, and scaling reduced the number of outliers and improved clustering among QCs for all workflows, categories (metabolites and lipids) and ionisation modes. Whilst workflow 2 had on average 2.9 % greater variation explained by PC1 compared to workflow 1 (Fig. 1 and Supplementary Fig. 4), all plots indicated that irrespective of the method of parameter selection, the UHPLC-HRMS data acquisition and XCMS preprocessing ensured high quality data, and the success of data pretreatment strategies in reducing any unwanted variation (42).

FEATURE QUALITY: Comparing true and negative metabolic features based on EIC EVA, a convolutional neural based network approach trained to recognise true and false metabolic features based on EICs, was chosen to efficiently evaluate whether manual or automated selection of XCMS parameters could affect peak quality (37). Whilst different selection methods for the preprocessing parameters affected feature number, it did not affect the percentage of true metabolic features for faecal and urinary metabolites in positive mode and faecal lipids in negative mode when comparing workflow 1 *vs.* 2. The same was observed by Guo et al (37). Here however, individual proportion tests showed a significant ($q < 0.01$) difference for metabolites in negative mode (both faecal and urinary) and faecal lipids in positive mode (Supplementary Fig. 5). If considering the total number of features from these datasets, which are both estimated as being true as determined by EVA with a CV < 30 %, workflow 2 is better (Supplementary Table 4). Whilst EVA was used here to evaluate feature quality because it circumvented the need to train machine learning models (43) and was quick to implement for 12 datasets, it only evaluates whether a feature is true/false based on EICs from one representative sample, not across all samples. Depending upon time constraints, number and size of datasets and hardware limitations, future research could explore how other feature quality metrics focused on peak shape that take into consideration multiple samples per feature compare to EVA (44). Currently, there is no standard definition of feature quality and measurements range from considering EIC shape like EVA to checking for retention time misalignment (45). To enhance reproducibility, a standardised definition of feature quality with software to assess this efficiently in large datasets is also warranted.

METABOLITE INDENTIFICATION: Assessing the number of tentative HMDB annotations per workflow There were 32 and 15 positively and negatively ionized HMDB adducts respectively, used to tentatively annotate features from each workflow (annotation level 4). The $[M+H]^+$ adduct usage ranged from 20 – 45 %, whilst the $[M-H]^-$ ranged from 35 – 70 %. Interestingly, more faecal metabolites could be identified at 1 ppm compared to urine, which is consistent with studies where more faecal metabolites could be identified relative to urine and plasma (46, 47). Considering significant proportion ($p < 0.05$) test results, a higher % of features could be annotated by HMDB based on adduct matching for faecal metabolites for workflow 2 (both ionisation modes), whilst workflow 1 had a higher % of features when comparing positively ionized lipids (Supplementary Table 5A). This was confirmed when using the features that remained after CV filtering. Considering unique HMDB features only, a significant difference was also observed (Supplementary Table 5B). Although, when considering number of features rather than proportions, workflow 2 had a higher number of features tentatively annotated by HMDB for all datasets and thus more unique compounds.

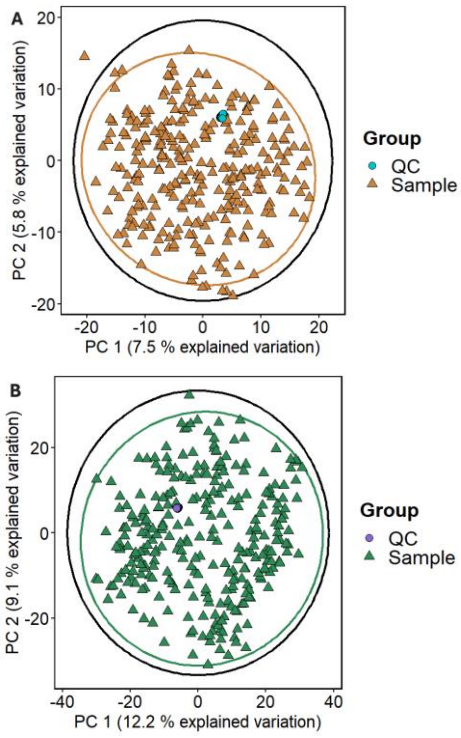


Figure 1: Exemplar PCA plots to compare QC clustering and number of outliers. Faecal metabolites in negative mode for workflow 1 (A) and 2 (B) following QC normalization, Yeo-Johnson transformation and pareto scaling.

It is acknowledged that using solely MS1 data for annotation is limiting and thus, retention time and MS2 data would be needed to accurately determine whether the automated parameter selection method increased metabolite identification.

FEATURE UTLITY: Comparing strength of feature signals Features with the highest captured intensity signal were present in the lipids category in negative mode, whereas faecal metabolites in this mode exhibited the lowest intensity signals (Supplementary Fig. 6). There was no significant difference between workflows when comparing the proportion of features within each intensity threshold bin using urinary metabolites (Supplementary Fig.6). Faecal metabolites showed significant differences when comparing proportions of features within lower to median intensity thresholds. Instead, faecal lipids, particularly in the positive ionization mode, showed significant differences when considering the upper intensities (Fig. 2). These observations held when using only features with CV < 30 % (Supplementary Fig. 6). If considering number of features (rather than proportion or percentage of total) present in higher intensity threshold bins, which could be useful for downstream analyses, then workflow 2 would be the best to proceed with.

FEATURE FILTERING: Using PCA loadings as a potential approach to filtering features Filtering features obtained from preprocessing software such as XCMS is important to maximize the retention of true positive features and remove noise (39). Approaches used include calculating CVs for features using QC samples and removing those exceeding a certain threshold, typically 25-30 % for untargeted LC-MS data, as used here, or only keeping features present in at least 80 % of biological samples (39). However, without considering data adaptive thresholds for these filtering methods or data-centric approaches, inevitably some true positive features would be lost. PCA is typically used for feature extraction, where a number of PCs are retained rather than the original high dimensional dataset, which differs from feature selection methods used to select a subset of original features (48). The use of loadings from PCA to rank variables of interest has been explored as a possible unsupervised feature selection method, where after determining a threshold for the loadings, the original features within this limit are retained for downstream analyses (49, 50). Whilst there is no clear consensus how best to rank variables using loadings, previous attempts have used a weighted technique through e.g., multiplying the loading by the proportion of variance explained by the PC and then adding these values across a pre-selected number of PCs (49, 50).

Here, QCs were used to assess the ability of PCA for unsupervised feature selection. Workflow 2 across all modes and categories had greater variation explained in PC1 compared to workflow 1 and fewer PCs were significant from permutation testing (proportion of variance explained by them ranged from 65 to 79 %). The loadings were then added across these PCs. Notably, a weighted approach was not used here, particularly as the sum of the squared loadings equals the proportion of variance explained, thus, it avoided altering what the original loading represented.

On average, this approach retained 52 % of features for both workflows, performing better for faecal features compared to urine. Nevertheless, excluding faecal lipids in the negative mode, workflow 2 for each mode and category had a higher number of filtered features that also had a CV < 30 % (Supplementary Table 6), on average 70.6 *vs.* 62.0 %. One would expect a higher % to have a low CV for both workflows if low loadings might reflect features with low variation across QCs. Different combinations of transformations and scaling methods were also tested; however, all performed similarly, albeit the features that were retained were different. Van den Berge et al (50) also found that different data pretreatment methods affected the rank of the metabolite based on the cumulative contribution of the weighted loading values across the first three PCs. Nevertheless, the PCA based approach outlined here highlights a data centric approach to filtering, which could be useful in conjunction with established methods (e.g. CV calculation).

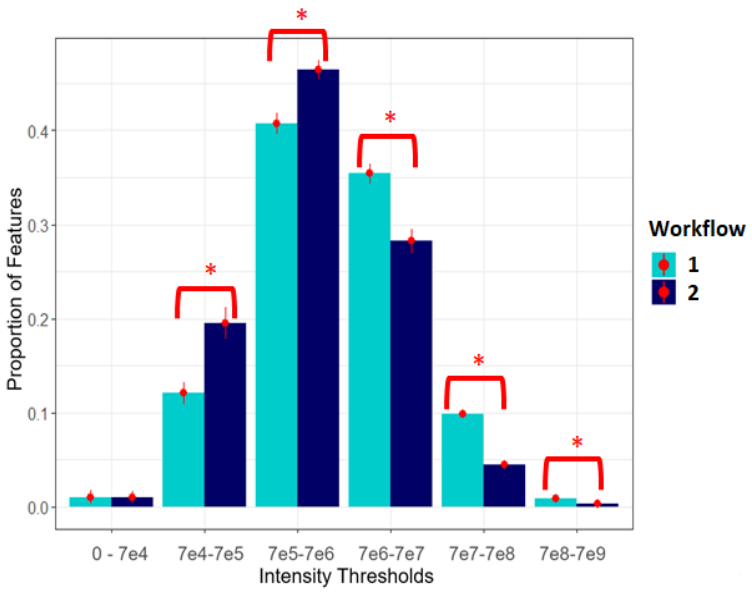


Figure 2: Bar plot showing the mean proportion of features +/- SD in intensity threshold bins using all features for faecal lipids in positive mode. *represents FDR q value < 0.01 for comparing proportions between workflows.

Overlap between categories

For each dataset (metabolites/lipids and positive/negative ion mode), when comparing workflows, the number of features unique to the 2-, 3-, 4-, and 5-set overlap were examined. When using all features, the ranking of the overlaps and the method used was the same for each workflow for 9/12 datasets (Supplementary Fig. S7). For example, considering a 2-set overlap for urine metabolites in positive mode, the EVA-Intensity overlap had the most features (213 and 219 features in workflow 1 and 2 respectively). This pattern held for the other 3-, 4- and 5- set overlaps even though the feature counts varied between workflows. The dataset that had the most features across all categories/methods (5-set overlap) was faecal lipids in positive mode for both workflows (Fig. 3). As workflow 2 had more features, a greater number was still found consistently in the 5-set overlaps across all datasets. Similar observations were also made when using the features that remained after CV filtering. The ICC distributions across intersections were also comparable for these datasets and workflows, except for faecal lipids in negative mode when comparing workflow 1 vs 2 (Supplementary Fig. S8). Finally, it was expected that methods within categories such as feature quality (EVA true and CV < 30 %) would have had strong overlaps. However, using all features, the EVA-CV was only the largest 2-set overlap for urinary metabolites in negative mode (13 % and 19 % of total features for workflow 1 and 2 respectively).

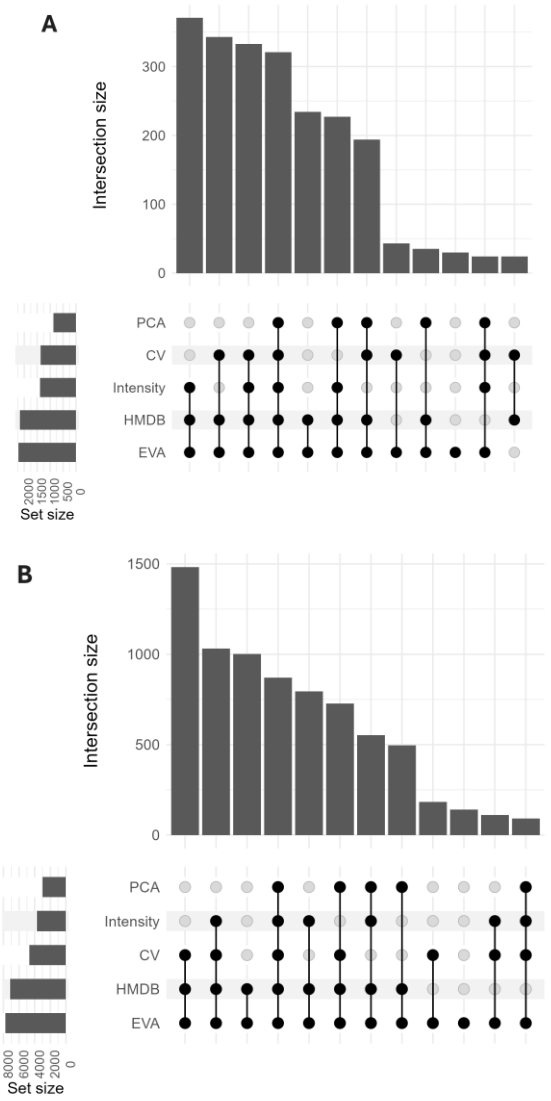


Figure 3: UpSet plots showing the overlap between different methods/categories for faecal lipids in positive mode using all features for workflow 1 (A) and 2 (B).

Similarly, features with low loadings retained through PCA filtering (likely true positives), were expected to have large overlaps with the EVA and/or CV category. Here, the EVA-CV-PCA 3-set overlap was the largest for only faecal metabolites in positive and negative mode (6 and 10 % respectively, both workflow 1) and urinary metabolites in negative mode (13 and 11 % for workflow 1 and 2). Future research understanding the discrepancy between feature quality metrics assessing the same criteria (e.g. true positives) will be important for method refinement or development.

Conclusions

A crucial part of any metabolomics study is to obtain a high-quality feature list that will be useful and informative for downstream statistical analyses. Therefore, an important step is not only selecting which preprocessing software to employ, but also the parameter values governing the algorithms used. Manual selection can be more time-intensive, compared to automated selection based on packages such as IPO and AutoTuner for XCMS software. However, some of the parameter values of the automatic approaches are likely to need finetuning based on knowledge of equipment used. This paper highlights some approaches to compare feature lists in an untargeted approach, where parameter values were either selected manually or where more > 75 % were chosen automatically. Considering the actual number of features, not proportions, the feature lists obtained from automated selection of parameter values performed better for each comparison than manual selection and a greater number of features were also found across all categories/methods of feature evaluation as identified using UpSet plots. It is noteworthy that features were present in at least 40 % of samples and therefore it remains unclear how the approaches used here to compare parameter selection methods would have affected the removal of potentially useful, but low abundant metabolites. Future exploration is warranted, given that some important metabolites may only appear in a small subset of samples due to dietary habits or gut microbiome composition. Nonetheless, this work underscores the importance of exploring these automatic parameter optimization packages to optimize preprocessing and capitalize on the plethora of data obtained by LC-MS from human biofluids.

ASSOCIATED CONTENT

Supplementary Material file containing tables and figures (Venn diagrams, violin plots, PCA, bar plots and UpSet diagrams).

AUTHOR INFORMATION

Corresponding Author

* Email: Lynn.Vanhaecke@UGent.be

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by a Department for the Economy PhD studentship (Northern Ireland) and the Interuniversity Special Research Fund (iBOF) from Flanders (Grant number BOFIBO2021001102). The FGFP was funded in part with the support of the Flemish government (IWT130359) and the Research Fund Flanders (FWO) Odysseus program (G.0924.09). The ENVIRonAGE was supported by the European Union research council “project ENVIRONAGE” (ERC-2012-StG 310,890) and Flemish Scientific Fund (G073315N/G048420N).

References

1. Alves S, Paris A, Rathahao-Paris E. Mass spectrometry-based metabolomics for an in-depth questioning of human health. Adv Clin Chem. 2020;99:147-91.
2. Wörheide MA, Krumsiek J, Kastenmüller G, Arnold M. Multi-omics integration in biomedical research - A metabolomics-centric review. Anal Chim Acta. 2021;1141:144-62.
3. Riquelme G, Bortolotto EE, Dombald M, Monge ME. Model-driven data curation pipeline for LC-MS-based untargeted metabolomics. Metabolomics. 2023;19(3):15.
4. Zhang XW, Li QH, Xu ZD, Dou JJ. Mass spectrometry-based metabolomics in health and medical science: a systematic review. RSC Adv. 2020;10(6):3092-104.
5. Tugizimana F, Steenkamp PA, Piater LA, Dubery IA. A Conversation on Data Mining Strategies in LC-MS Untargeted Metabolomics: Pre-Processing and Pre-Treatment Steps. Metabolites. 2016;6(4).
6. Bauermeister A, Mannochio-Russo H, Costa-Lotufo LV, Jarmusch AK, Dorrestein PC. Mass spectrometry-based metabolomics in microbiome investigations. Nat Rev Microbiol. 2022;20(3):143-60.
7. Spicer R, Salek RM, Moreno P, Cañueto D, Steinbeck C. Navigating freely-available software tools for metabolomics analysis. Metabolomics. 2017;13(9):106.
8. Domingo-Almenara X, Siuzdak G. Metabolomics Data Processing Using XCMS. Methods Mol Biol. 2020;2104:11-24.

9. Schmid R, Heuckeroth S, Korf A, Smirnov A, Myers O, Dyrlund TS, et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. Nat Biotechnol. 2023;41(4):447-9.
10. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem. 2006;78(3):779-87.
11. 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. 2015;12(6):523-6.
12. Guo J, Huan T. Mechanistic Understanding of the Discrepancies between Common Peak Picking Algorithms in Liquid Chromatography-Mass Spectrometry-Based Metabolomics. Anal Chem. 2023;95(14):5894-902.
13. Hu Y, Cai B, Huan T. Enhancing Metabolome Coverage in Data-Dependent LC-MS/MS Analysis through an Integrated Feature Extraction Strategy. Anal Chem. 2019;91(22):14433-41.
14. Du X, Dastmalchi F, Ye H, Garrett TJ, Diller MA, Liu M, et al. Evaluating LC-HRMS metabolomics data processing software using FAIR principles for research software. Metabolomics. 2023;19(2):11.
15. Hemmer S, Manier SK, Fischmann S, Westphal F, Wagmann L, Meyer MR. Comparison of Three Untargeted Data Processing Workflows for Evaluating LC-HRMS Metabolomics Data. Metabolites. 2020;10(9).
16. Libiseller G, Dvorzak M, Kleb U, Gander E, Eisenberg T, Madeo F, et al. IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics. 2015;16:118.
17. McLean C, Kujawinski EB. AutoTuner: High Fidelity and Robust Parameter Selection for Metabolomics Data Processing. Anal Chem. 2020;92(8):5724-32.
18. Narath SH, Mautner SI, Svehlikova E, Schultes B, Pieber TR, Sinner FM, et al. An Untargeted Metabolomics Approach to Characterize Short-Term and Long-Term Metabolic Changes after Bariatric Surgery. PLoS One. 2016;11(9):e0161425.
19. Albóniga OE, González O, Alonso RM, Xu Y, Goodacre R. Optimization of XCMS parameters for LC-MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. Metabolomics. 2020;16(1):14.
20. Lassen J, Nielsen KL, Johannsen M, Villesen P. Assessment of XCMS Optimization Methods with Machine-Learning Performance. Anal Chem. 2021;93(40):13459-66.
21. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, et al. Population-level analysis of gut microbiome variation. Science. 2016;352(6285):560-4.
22. Janssen BG, Madhloum N, Gyselaers W, Bijnens E, Clemente DB, Cox B, et al. Cohort Profile: The ENVIRonmental influence ON early AGEing (ENVIRONAGE): a birth cohort study. Int J Epidemiol. 2017;46(5):1386-7m.
23. Vangeenderhuysen P, Van Arnhem J, Pomian B, De Graeve M, De Commer L, Falony G, et al. Dual UHPLC-HRMS Metabolomics and Lipidomics and Automated Data Processing Workflow for Comprehensive High-Throughput Gut Phenotyping. Anal Chem. 2023;95(22):8461-8.
24. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc. 2011;6(7):1060-83.
25. De Paepe E, Van Meulebroek L, Rombouts C, Huysman S, Verplanken K, Lapauw B, et al. A validated multi-matrix platform for metabolomic fingerprinting of human urine, feces and plasma using ultra-high performance liquid-chromatography coupled to hybrid orbitrap high-resolution mass spectrometry. Anal Chim Acta. 2018;1033:108-18.
26. Van Meulebroek L, De Paepe E, Vercruysse V, Pomian B, Bos S, Lapauw B, et al. Holistic Lipidomics of the Human Gut Phenotype Using Validated Ultra-High-Performance Liquid Chromatography Coupled to Hybrid Orbitrap Mass Spectrometry. Anal Chem. 2017;89(22):12502-10.
27. Adusumilli R, Mallick P. Data Conversion with ProteoWizard msConvert. Methods Mol Biol. 2017;1550:339-68.
28. Tautenhahn R, Böttcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics. 2008;9:504.
29. Pirttilä K, Balgoma D, Rainer J, Pettersson C, Hedeland M, Brunius C. Comprehensive Peak Characterization (CPC) in Untargeted LC-MS Analysis. Metabolites. 2022;12(2).
30. Johannes Rainer MG-A. refineChromPeaks-merge [Available from: https://rdrr.io/bioc/xcms/man/refineChromPeaks-merge.html.
31. Prince JT, Marcotte EM. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. Anal Chem. 2006;78(17):6140-52.
32. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem. 2012;84(1):283-9.
33. Hsu JY, Hsu JF, Chen YR, Shih CL, Hsu YS, Chen YJ, et al. Urinary exposure marker discovery for toxicants using ultra-high pressure liquid chromatography coupled with Orbitrap high resolution mass spectrometry and three untargeted metabolomics approaches. Anal Chim Acta. 2016;939:73-83.
34. Rombouts C, Van Meulebroek L, De Spiegeleer M, Goethals S, Van Hecke T, De Smet S, et al. Untargeted Metabolomics Reveals Elevated L-Carnitine Metabolism in Pig and Rat Colon Tissue Following Red Versus White Meat Intake. Mol Nutr Food Res. 2021;65(7):e2000463.
35. De Graeve M, Van de Walle E, Van Hecke T, De Smet S, Vanhaecke L, Hemeryck LY. Exploration and optimization of extraction, analysis and data normalization strategies for mass spectrometry-based DNA adductome mapping and modeling. Anal Chim Acta. 2023;1274:341578.
36. Jacob E.Wulff MWM. A Comparison of Various Normalization Methods for LC/MS Metabolomics Data Advances in Bioscience and Biotechnology 2018;9.
37. Guo J, Shen S, Xing S, Chen Y, Chen F, Porter EM, et al. EVA: Evaluation of Metabolic Feature Fidelity Using a Deep Learning Model Trained With Over 25000 Extracted Ion Chromatograms. Anal Chem. 2021;93(36):12181-6.
38. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ Sci Technol. 2014;48(4):2097-8.
39. Schiffman C, Petrick L, Perttula K, Yano Y, Carlsson H, Whitehead T, et al. Filtering procedures for untargeted LC-MS metabolomics data. BMC Bioinformatics. 2019;20(1):334.
40. M K. ComplexUpset. 20202.
41. A. Lex NG, H. Strobelt, R. Vuillemot and H. Pfister. UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics.20(12):1983-92.
42. De Livera AM, Sysi-Aho M, Jacob L, Gagnon-Bartsch JA, Castillo S, Simpson JA, et al. Statistical methods for handling unwanted variation in metabolomics data. Anal Chem. 2015;87(7):3606-15.
43. Kantz ED, Tiwari S, Watrous JD, Cheng S, Jain M. Deep Neural Networks for Classification of LC-MS Spectral Peaks. Anal Chem. 2019;91(19):12407-13.
44. Kumler W, Hazelton BJ, Ingalls AE. Picky with peakpicking: assessing chromatographic peak quality with simple metrics in metabolomics. BMC Bioinformatics. 2023;24(1):404.
45. Wu CT, Wang Y, Wang Y, Ebbels T, Karaman I, Graça G, et al. Targeted realignment of LC-MS profiles by neighbor-wise compound-specific graphical time warping with misalignment detection. Bioinformatics. 2020;36(9):2862-71.
46. Zhao L, Wang C, Peng S, Zhu X, Zhang Z, Zhao Y, et al. Pivotal interplays between fecal metabolome and gut microbiome reveal functional signatures in cerebral ischemic stroke. J Transl Med. 2022;20(1):459.
47. Zhgun ES, Ilina EN. Fecal Metabolites As Non-Invasive Biomarkers of Gut Diseases. Acta Naturae. 2020;12(2):4-14.
48. Hira ZM, Gillies DF. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Adv Bioinformatics. 2015;2015:198363.
49. Seoung Bum Kim PR. Unsupervised feature selection using weighted principal components. Expert Systems with Applications. 2011;38(5):5704-10.
50. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006;7:142.