Supporting Information

Representing the Metabolome with High Fidelity: Range and Response as Quality Control Factors in LC-MS Based Global Profiling

Caroline J. Sands[1,2]*, María Gómez-Romero[1,2], Gonçalo Correia[1,2], Elena Chekmeneva[1,2], Stephane Camuzeaux[1,2], Chioma Izzi-Engbeaya[3], Waljit S. Dhillo[3], Zoltan Takats[1,2], Matthew R. Lewis[1,2]*

[1] National Phenome Centre, Department of Metabolism, Digestion and Reproduction, Imperial College London, Hammersmith Campus, London W12 0NN, United Kingdom

[2] Section of Bioanalytical Chemistry, Department of Metabolism, Digestion and Reproduction, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

[3] Section of Endocrinology and Investigative Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, Hammersmith Campus, London W12 0HS, United Kingdom

Corresponding Authors

*E-mail: Caroline.Sands01@imperial.ac.uk (C.J.S.).
*E-mail: Matthew.Lewis@imperial.ac.uk (M.R.L.).

Table of Contents

Figure S1 – Overview of key LC-MS metabolomics workflow steps

Overview of the key stages in the metabolomics workflow for the transformation of LC-MS untargeted profiling measurements into interpretable data, including definitions, common terminology and currently available open-source software.

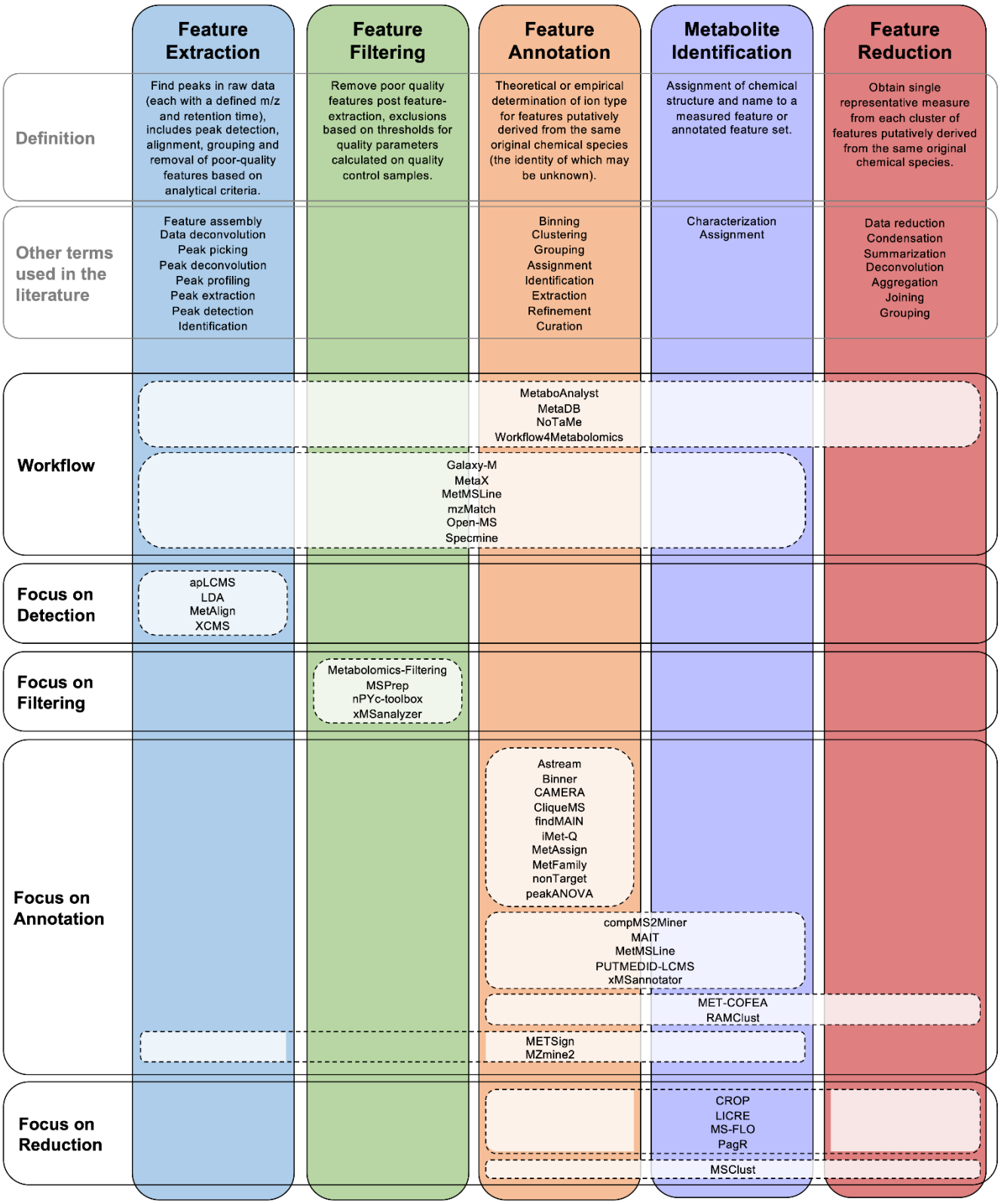


Figure S2 – Advanced filtering applied to a range of full profiling feature sets

Assessment of the impact of applying rudimentary response and range-based filtering methods to full profiling feature sets. Feature sets selected to cover two biofluids (serum and urine), three chromatographic methods (small molecule SmMol RPC, Lipid RPC, HILIC) and two ion modes (positive and negative). In each Venn diagram[1], the numbers of features not meeting each filtering strategy are given according to the following inclusion criteria: RSD, RSD in pooled QC samples ≤ 30; Correlation, Pearson correlation coefficient between dQC series and dilution factor ≥ 0.7; Response, greater than 80% of study samples within an intensity range where fold-change error ≤ 20%; Range, greater than 80% of study samples within a range covered by the dQC series samples. For full experimental details see Methods S1 and S3.

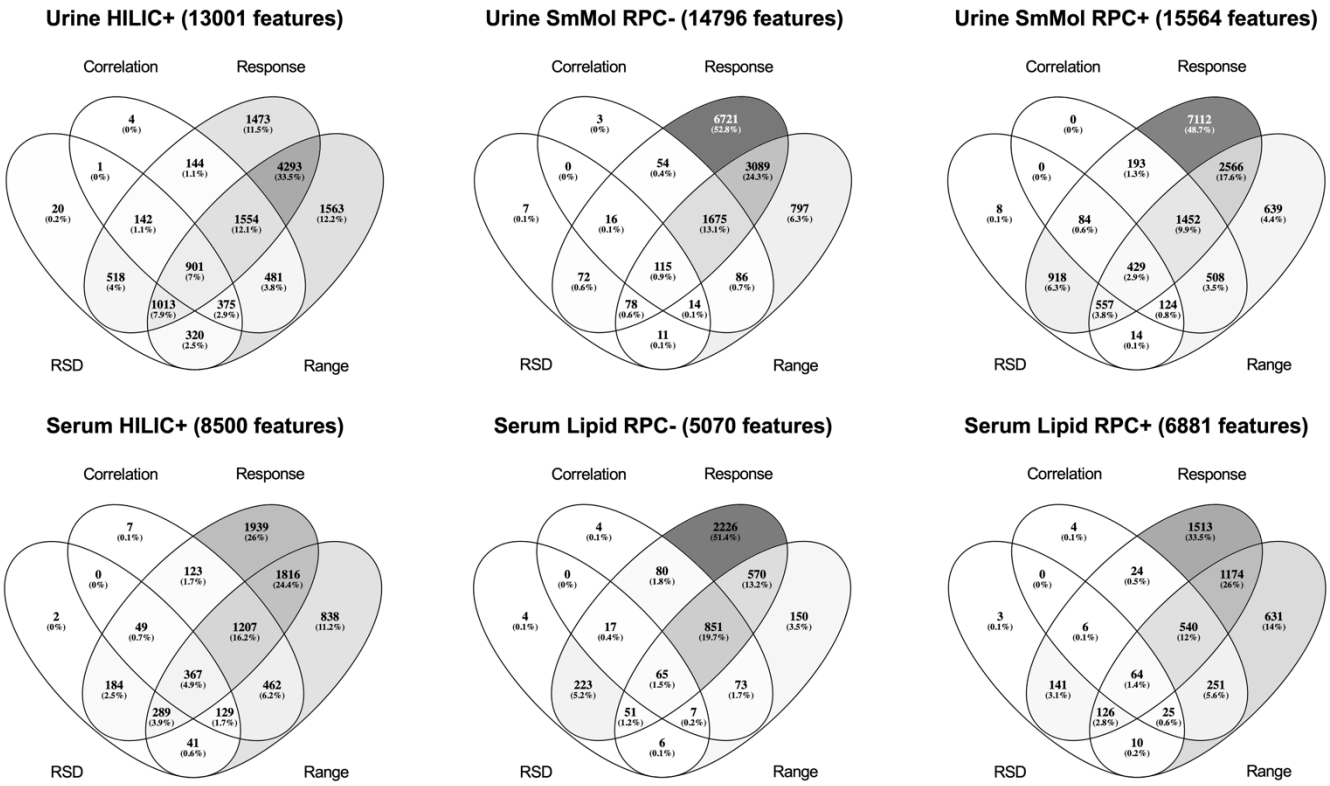
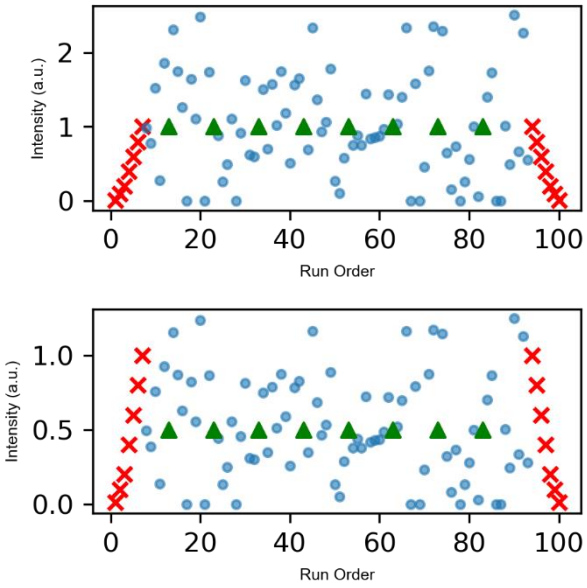


Figure S3 – Dilution series design

Dilution of study and pooled QC samples (blue circles and green triangles respectively) allows more complete dQC series (red crosses) range coverage in metabolic profiling studies without need for sample concentration. When undiluted (top panel), approximately half of the study samples are above the upper boundary of the dQC series, hindering assessment of their quality with respect to the high end of linear dynamic range. Adding a dilution step to all study samples and the pooled QC (bottom panel) captures more of the dataset within the boundaries of the dilution series, allowing better assessment of feature response across the range.



Methods S1 – Experimental details of the serum feature sets

The serum datasets used as examples in this paper have been previously published[2]. Fifteen healthy men were recruited to investigate the effects of kisspeptin on glucose-stimulated insulin secretion and appetite. Serum samples (N = 112) were taken pre-infusion (T = -15 minutes) and at steady state (T = 45 minutes). Sample handling (sorting, formatting, preparation), UPLC-MS and data pre-processing was performed as previously described[2]. For QC assessment and data pre-processing, a QC sample was initially prepared by pooling equal parts of each study sample, and a dilution series was created from the pooled QC sample (10x 100%, 5x 80%, 3x 60%, 3x 40%, 5x 20%, 10x 1%). Samples were subjected to RPC tailored for complex lipid separation, while HILIC was used to separate small polar metabolites. Aliquots (50 µL) were taken from each study sample and pooled QC and diluted 1:1 v/v with ultrapure water and protein was removed by addition of organic solvent (diluted sample/isopropanol in 1:4 v/v ratio for lipid RPC profiling and diluted sample/acetonitrile in 1:3 v/v ratio for HILIC profiling). Mixtures of method specific chemichal standards were added (at dilution stage for HILIC and protein precipitation stage for RPC) in order to monitor data quality during acquisition (see Izzi-Engbeaya *et al.* supplementary information (Metabolite profiling) for full details[2]). All analyses were performed on Acquity UPLC instruments, coupled to Xevo G2-S TOF mass spectrometers (Waters Corp., Manchester, UK) via a Z-spray electrospray ionization (ESI) source. The lipid RPC profiling was conducted in both positive and negative ion modes (generating the serum lipid RPC+ and lipid RPC- feature sets, respectively), while the HILIC assay was performed in positive ion mode only (generating the serum HILIC+ feature set). For QC assessment and data pre-processing, the pooled QC sample was acquired every 10 study samples throughout the analysis and a set of dilution series samples were acquired immediately prior to and after the study sample analysis. Feature extraction and retention time alignment were performed in Progenesis QI (Waters Corp., Milford, MA) and data pre-processing for the elimination of potential run-order effects was performed using the nPYc-Toolbox[3].

Methods S2 – Data pre-processing, modelling and lipid assignment details for Figure 2

Data was acquired as per Methods S1. For each feature set (lipid RPC+, lipid RPC-, HILIC+) features were filtered using the nPYc-Toolbox[3] according to the following inclusion criteria: RSD in pooled QC $\leq$ 20, dQC series Pearson correlation to dilution factor $\geq$ 0.8, RSD in study samples $\geq$ 1.1* RSD in pooled QC. As previously detailed[2], for each final dataset, linear mixed effect (LME) models were generated using the lmer4 R package[4] for each feature according to the formula: model <− Feature ~ Time*Class + (1|SubjectID) + (1|Challenge), including fixed effects for the interaction between class (kisspeptin or vehicle alone) and time (T = -15 minutes and T = 45 minutes), and random affects for participant and challenge (owing to the presence of multiple challenges per participant). Statistical significance was determined by local FDR correction[5] of the appropriate LME model estimates (local FDR-corrected value < 0.05). Where possible, chemical identity of significant features was assigned by matching accurate mass and tandem mass spectrometry (MS/MS) fragmentation measurements to reference spectra using LIPID MAPS online tools (for lipid species)[6] or, where available, to authentic chemical standards. Figure 2 shows a Manhattan style plot of the 5200 features measured by serum lipid RPC+. Of these, 392 significantly changed over time with kisspeptin administration (colored red for increasing and blue for decreasing). See Izzi-Engbeaya *et al* supplementary information (Table S4) for full details[2].

Methods S3 – Experimental details of the urine feature sets

The urine feature sets used as examples in this paper were generated as part of an ongoing study where data has been acquired for 126 urine samples. Sample handling (sorting, formatting, preparation), UPLC-MS and data pre-processing was performed as previously described[7], with an additional sample dilution step as detailed below. For QC assessment and data pre-processing, a QC sample was initially prepared by pooling equal parts of each study sample, and a dilution series was created from the pooled QC sample (10x 100%, 5x 80%, 3x 60%, 3x 40%, 5x 20%, 10x 1%). Samples were subjected to RPC tailored for small molecule separation, while HILIC was used to separate small polar metabolites. Initially, aliquots (75 µL for RPC and 25 µL for HILIC) were taken from each study sample and pooled QC and diluted 1:1 v/v with ultrapure water. Subsequent stages follow that previously published[7]. In brief, samples were diluted 1:1 v/v with ultrapure water and with assay specific chemical standards for monitoring data quality during acquisition (see Lewis *et al.* for full details[7]). In order to better match the initial solvent conditions, acetonitrile was added to diluted samples for HILIC analysis (diluted sample/acetonitrile in 1:3 v/v ratio). All analyses were performed on Acquity UPLC instruments, coupled to Xevo G2-S TOF mass spectrometers (Waters Corp., Manchester, UK) via a Z-spray electrospray ionization (ESI) source. The small molecule (SmMol) RPC profiling was conducted in both positive and negative ion modes (generating the urine SmMol RPC+ and SmMol RPC- feature sets, respectively), while the HILIC assay was performed in positive ion mode only (generating the urine HILIC+ feature set). For QC assessment and data pre-processing, the pooled QC sample was acquired every 10 study samples throughout the analysis and a set of dilution series samples were acquired immediately prior to and after the study sample analysis. Raw data was converted to the mzML open source format and signals below an absolute intensity threshold of 100 counts were removed using the MSConvert tool in ProteoWizard[8]. Feature extraction was performed by XCMS[9] and data pre-processing for the elimination of potential run-order effects was performed using the nPYc-Toolbox[3].

Table S1 – Feature reduction strategies and open-source software

Strategies for reducing features putatively derived from the same compound into a single representative measurement. Software tools in this table are restricted to those which perform (or can perform) feature reduction (either by combination or selection). Note, most strategies highlight the utility of reduction for statistical analysis, but to retain all features for future reference. Abbreviations: *m/z,* mass-to-charge ratio; RT, retention time.

Approach	Strategy	Tool	Details
Combine	Sum	MS-FLO[10]	Feature set inspected for putative related molecular and adduct ions based on user defined parameters (including expected adduct types and *m/z* and RT tolerances). Feature pairs meeting these criteria and with peak height correlation of $R^2 \geq 0.8$ across all samples, are automatically joined by summing their intensity values. Features meeting these criteria but with $R^2 < 0.8$, or multiple features meeting criteria with $R^2 \geq 0.8$ are flagged for manual review.
	Mean	PagR[11]	Presents results of four peak aggregation (feature reduction) strategies: three ways of combining and one way of selecting features (see below). All methods resulted in a significant increase in predictive power compared to the non-reduced dataset.
	Principal component analysis decomposition		
	Non-negative matrix factorisation reduction		
	Weighted mean	RAMClust[12]	Unsupervised method using RT and correlation between features across all samples (including MS/MS if available) to group features into spectra. Outputs include a dataset where grouped features are condensed into spectral intensities using a weighted mean function (where more abundant signals contribute more to the spectral intensity).
Select	Largest mean peak area	CROP[13]	Features grouped based on Pearson's pairwise correlations and RT, with each resulting group represented in the final dataset by the feature with the highest mean peak area.
		PagR[11]	Presents results of four peak aggregation (feature reduction) strategies: three ways of combining (see above) and one way of selecting features. All methods resulted in a significant increase in predictive power compared to the non-reduced dataset.
	Largest median peak area	LICRE[14]	(Lipid) features grouped based on correlation, for each final node (set of highly correlated features) the feature with highest median measurement is retained in the final dataset.
		NoTaMe[15]	Features grouped using a novel undirected graph approach based on Pearson's pairwise correlations and RT. Each resulting group represented in the final dataset by the feature with the highest median peak area.

	(De)protonated ion	MET-COFEA[16]	Used (de)protonated ions for quantitation in exemplar datasets, illustrating success of sample class separation and ease of interpretation in reduced dataset.
		MetaboAnalyst[17]	Optional filter (using the FormatPeakList function, for more details see online tutorial documentation[18]) post annotation (using CAMERA[19]) to remove all adducts except for the (de)protonated ions.
	Highest degree of connection (having the most relationships to other features) Highest intensity	MetaDB[20]	Pre-processing workflow includes optional output of a relative intensity measure for chemical compounds rather than features, representative feature selected based on abundance and cluster membership (uses MSClust[21]).
		MS-CleanR[22]	Post feature clustering (using MS-DIAL[23]) the user can select the number of features to keep between three selection strategies (highest connectivity, highest intensity, or both).
	Highest intensity	Workflow4Metabolomics[24]	Options for feature reduction using the Analytic Correlation Filtration (ACorF) tool. After grouping (using CAMERA[19]) the user can choose between one of these four strategies to select a representative feature from each group.
	Highest mass		
	Highest mass^2 average intensity		
	Highest mass among the top highest average intensities		

Table S2 – Feature filtering strategies and open-source software

Post extraction feature filtering options. Throughout the text sample definitions have been unified: SS: study sample, comprising the biological/experimental sample set; QC: quality control, comprising repeated injections of a representative sample (e.g., a pool of SS); dQC series: diluted QC series, comprising the QC sample diluted to a number of different relative concentrations. Other abbreviations: CV, coefficient of variation; IQR, interquartile range; RSD, relative standard deviation.

Approach	Strategy	Tool	Details
Biological variance	Feature must exhibit sufficient variance, or more variance in SS than in QC samples	MetaboAnalyst[17]	Optional filter to exclude low-variance features. This filtering is based on either IQR, CV, or standard deviation, and empirical rules are applied (less than 250 variables: 5% will be filtered; between 250 - 500 variables: 10% will be filtered; between 500 - 1000 variables: 25% will be filtered; over 1000 variables: 40% will be filtered). NOTE, no filtering is only an option for datasets with less than 5000 features, otherwise some filtering must be applied.
		Metabolomics-Filtering[25]	By manual pre-definition of a subset of high and low-quality peaks, and visualisation of the corresponding distributions of intra-class correlation coefficients (ICC, or

			proportion of between-subject variation to total variation, where repeated measures of a pooled QC sample are considered a 'pseudo-subject') an appropriate (data specific) threshold can be selected for feature filtering. ICC simultaneously considers both technical and biological variability, thus a large ICC indicates that much of the total variation is biological (regardless of the magnitude of the CV).
		NoTaMe[15]	Flags or excludes features with D-ratio < 0.4 (where D-ratio = standard deviation in QC/standard deviation in SS[26]).
		nPYc-Toolbox[3]	Excludes features where QC RSD * threshold > SS RSD. Default threshold 1.1.
		Specmine[27] for more details see online documentation[28]	Optional filter to exclude low-variance features. Filtering is based on either IQR, RSD, standard deviation or median absolute deviation, and features excluded either by "percent" variables in the dataset or "threshold" absolute values. Percent or threshold values defined by the user or determined automatically if required.
		Workflow4Metabolomics[29] for more details see online tutorial documentation[30]	Optional filter (using quality metric computation and generic filter) to flag and exclude features where QC CV/SS CV < threshold%. User defined threshold. Flexible, where QC samples not included, features can be filtered based on overall standard deviation or CV values.
Intensity	Feature must be present at sufficient intensity	MetaboAnalyst[17]	Optional filter to exclude low-value features. Filtering is based on either sample means or medians, and empirical rules are applied (less than 250 variables: 5% will be filtered; between 250 - 500 variables: 10% will be filtered; between 500 - 1000 variables: 25% will be filtered; over 1000 variables: 40% will be filtered). NOTE, no filtering is only an option for datasets with less than 5000 features, otherwise some filtering must be applied.
		mzMatch[31, 32] for more details see online tutorial documentation [33]	Optional filter to exclude features not meeting threshold intensity. User defined threshold.
		Specmine[27] for more details see online documentation[28]	Optional filter to exclude low-value features. Filtering is based on either sample means or medians, and features excluded either by "percent" variables in the dataset or "threshold" absolute values. Percent or threshold values defined by the user or determined automatically if required.
		Workflow4Metabolomics[29] for more details see online tutorial documentation [30]	Optional filter (using quality metric computation and generic filter) to flag and exclude features not meeting threshold intensity (e.g., in mean intensity across samples). User defined threshold.

Linear response	Feature must respond in expected way to dilution of replicate samples (dilution series)	mzMatch[31, 32] for more details see online tutorial documentation[33]	Excludes features with Pearson's correlation of dilution factor order to binary logarithm of the peak intensities in dQC series samples < -0.85.
		nPYc-Toolbox[3]	Excludes features with Pearson's correlation of dilution factor to intensity in dQC series samples ≥ threshold. Default threshold 0.7.
		Workflow4Metabolomics[29] for more details see online tutorial documentation[30]	Optional filter (using quality metric computation and generic filter) to flag and exclude features with correlation of dilution factor to intensity in dQC series samples < threshold%. User defined threshold.
Non-biological source	Feature must not be present in procedural blank samples	Galaxy-M[34]	Excludes features that appear to be as strong in the blanks as in the biological spectra. User defined thresholds.
		Metabolomics-Filtering[25]	By manual pre-definition of a subset of high and low-quality peaks, and visualisation of mean-difference plot between feature abundances in blank and SS an appropriate (data specific) threshold can be selected for feature filtering.
		mzMatch[31, 32] for more details see online tutorial documentation[33]	Optional filter to exclude features where signal intensity in blanks is greater than or equal to that in SS.
		nPYc-Toolbox[3]	Optional filter to exclude features where average intensity is greater than that seen in procedural blank injections *threshold. Default threshold 1.1.
		Workflow4Metabolomics[29] for more details see online tutorial documentation[30]	Optional filter (using quality metric computation and generic filter) to flag and exclude features where signal intensity in blanks exceeds a certain threshold (or is greater than that in SS). User defined threshold.
Precision	Feature must be present with less than a certain CV/RSD in replicate samples CV/RSD = σ/μ, where σ is the standard deviation and μ the mean intensity across sample replicates	MetaboAnalyst[17]	Optional filter to exclude features which show low repeatability, i.e., RSD in QC samples > x% (suggested threshold 20% for LC-MS data). NOTE, no filtering is only an option for datasets with less than 5000 features, otherwise some filtering must be applied.
		MetaDB[20]	Calculates QC RSD (suggests threshold 0.2).
		MetMSLine[35]	Excludes features with QC RSD > 0.3.
		MSPrep[36]	Three technical replicates per sample required. User specified threshold for CV. Only features found in at least two replicates are retained. If CV < threshold, average of replicates is used; if CV > threshold and found in 2/3 replicates, observation is left blank; if CV > threshold and found in 3/3 replicates, median of replicates is used. Dataset subsequently filtered by prevalence (see above).
		mzMatch[31, 32] for more details see online tutorial documentation [33]	Optional filter to exclude features irreproducible in biological and/or technical replicates. User defined RSD threshold (0.3 used in tutorial).

		NoTaMe[15]	Flag or exclude features with QC RSD > 0.2.
		nPYc-Toolbox[3]	Excludes features with QC RSD > threshold%. Default threshold 30%.
		Workflow4Metabolomics[29] for more details see online tutorial documentation[30]	Optional filter (using quality metric computation and generic filter) to flag and exclude features with QC RSD < threshold%. User defined threshold.
		xMSanalyzer[37]	If analytical replicates acquired, uses QC CV or percent intensity difference (PID = absolute intensity difference/mean intensity*100) between analytical replicates to define the best quality features.
Prevalence	Feature must be present in at least a certain number or percentage of samples	Galaxy-M[34]	Excludes features not present in x-out-of-n study samples in total or in any sample class. User defined threshold for x.
		MetaboAnalyst[17]	Optional filter to exclude low-prevalence features, by exclusion of features with > threshold% missing values (default threshold 50%). Also options for missing value imputation. NOTE, no filtering is only an option for datasets with less than 5000 features, otherwise some filtering must be applied.
		Metabolomics-Filtering[25]	By manual pre-definition of a subset of high and low-quality peaks, and visualisation of the corresponding distributions of percent missing values an appropriate (data specific) threshold can be selected for feature filtering.
		MetaX[38]	Excludes features not present in > 50% QC samples and > 20% SS. Post filtering options for missing value imputation.
		MSPrep[36]	Excludes features not present in > threshold% samples. Threshold set by user (80% used in example). Post filtering options for missing value imputation.
		mzMatch[31, 32] for more details see online tutorial documentation[33]	Optional filter to exclude features not present in x samples. User defined threshold.
		NoTaMe[15]	Flags features not present in > 70% QC samples. NOTE, features are excluded from analysis but retained in case useful for future metabolite identification.

References

1. Oliveros, J. C. *VENNY. An interactive tool for comparing lists with Venn Diagrams.*, 2007.
2. Izzi-Engbeaya, C.; Comninos, A. N.; Clarke, S. A.; Jomard, A.; Yang, L.; Jones, S.; Abbara, A.; Narayanaswamy, S.; Eng, P. C.; Papadopoulou, D.; Prague, J. K.; Bech, P.; Godsland, I. F.; Bassett, P.; Sands, C.; Camuzeaux, S.; Gomez-Romero, M.; Pearce, J. T. M.; Lewis, M. R.; Holmes, E.; Nicholson, J. K.; Tan, T.; Ratnasabapathy, R.; Hu, M.; Carrat, G.; Piemonti, L.; Bugliani, M.; Marchetti, P.; Johnson, P. R.; Hughes, S. J.; James Shapiro, A. M.; Rutter, G. A.; Dhillo, W. S., The effects of kisspeptin on beta-cell function, serum metabolites and appetite in humans. *Diabetes Obes Metab* **2018,** *20* (12), 2800-2810.
3. Sands, C. J.; Wolfer, A. M.; Correia, G. D. S.; Sadawi, N.; Ahmed, A.; Jimenez, B.; Lewis, M. R.; Glen, R. C.; Nicholson, J. K.; Pearce, J. T. M., The nPYc-Toolbox, a Python module for the pre-processing, quality-control and analysis of metabolic profiling datasets. *Bioinformatics* **2019,** *35* (24), 5359-5360.
4. Bates, D. M., M.; Bolker, B.; Walker, S., Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **2015,** *67* (1), 48.
5. Efron, B. T., B.; Narasimhan, B.; Strimmer, K. Computation of local false discovery rates, https://cran.r-project.org/web/packages/locfdr/.
6. Fahy, E.; Sud, M.; Cotter, D.; Subramaniam, S., LIPID MAPS online tools for lipid research. *Nucleic Acids Res* **2007,** *35* (Web Server issue), W606-12.
7. Lewis, M. R.; Pearce, J. T.; Spagou, K.; Green, M.; Dona, A. C.; Yuen, A. H.; David, M.; Berry, D. J.; Chappell, K.; Horneffer-van der Sluis, V.; Shaw, R.; Lovestone, S.; Elliott, P.; Shockcor, J.; Lindon, J. C.; Cloarec, O.; Takats, Z.; Holmes, E.; Nicholson, J. K., Development and Application of Ultra-Performance Liquid Chromatography-TOF MS for Precision Large Scale Urinary Metabolic Phenotyping. *Anal Chem* **2016,** *88* (18), 9004-13.
8. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **2012,** *30* (10), 918-20.
9. Smith, C. A.; Want, E. J.; 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.
10. DeFelice, B. C.; Mehta, S. S.; Samra, S.; Cajka, T.; Wancewicz, B.; Fahrmann, J. F.; Fiehn, O., Mass Spectral Feature List Optimizer (MS-FLO): A Tool To Minimize False Positive Peak Reports in Untargeted Liquid Chromatography-Mass Spectroscopy (LC-MS) Data Processing. *Anal Chem* **2017,** *89* (6), 3250-3255.
11. Fernandez-Albert, F.; Llorach, R.; Andres-Lacueva, C.; Perera-Lluna, A., Peak aggregation as an innovative strategy for improving the predictive power of LC-MS metabolomic profiles. *Anal Chem* **2014,** *86* (5), 2320-5.
12. Broeckling, C. D.; Afsar, F. A.; Neumann, S.; Ben-Hur, A.; Prenni, J. E., RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data. *Anal Chem* **2014,** *86* (14), 6812-7.
13. Kouril, S.; de Sousa, J.; Vaclavik, J.; Friedecky, D.; Adam, T., CROP: correlation-based reduction of feature multiplicities in untargeted metabolomic data. *Bioinformatics* **2020,** *36* (9), 2941-2942.

14. Wong, G.; Chan, J.; Kingwell, B. A.; Leckie, C.; Meikle, P. J., LICRE: unsupervised feature correlation reduction for lipidomics. *Bioinformatics* **2014,** *30* (19), 2832-3.
15. Klavus, A.; Kokla, M.; Noerman, S.; Koistinen, V. M.; Tuomainen, M.; Zarei, I.; Meuronen, T.; Hakkinen, M. R.; Rummukainen, S.; Farizah Babu, A.; Sallinen, T.; Karkkainen, O.; Paananen, J.; Broadhurst, D.; Brunius, C.; Hanhineva, K., "notame": Workflow for Non-Targeted LC-MS Metabolic Profiling. *Metabolites* **2020,** *10* (4).
16. Zhang, W.; Chang, J.; Lei, Z.; Huhman, D.; Sumner, L. W.; Zhao, P. X., MET-COFEA: a liquid chromatography/mass spectrometry data processing platform for metabolite compound feature extraction and annotation. *Anal Chem* **2014,** *86* (13), 6245-53.
17. Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D. S.; Xia, J., MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res* **2018,** *46* (W1), W486-W494.
18. https://www.metaboanalyst.ca/docs/RTutorial.xhtml.
19. Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T. R.; 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.
20. Franceschi, P.; Mylonas, R.; Shahaf, N.; Scholz, M.; Arapitsas, P.; Masuero, D.; Weingart, G.; Carlin, S.; Vrhovsek, U.; Mattivi, F.; Wehrens, R., MetaDB a Data Processing Workflow in Untargeted MS-Based Metabolomics Experiments. *Front Bioeng Biotechnol* **2014,** *2*, 72.
21. Tikunov, Y. M.; Laptenok, S.; Hall, R. D.; Bovy, A.; de Vos, R. C., MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. *Metabolomics* **2012,** *8* (4), 714-718.
22. Fraisier-Vannier, O.; Chervin, J.; Cabanac, G.; Puech, V.; Fournier, S.; Durand, V.; Amiel, A.; Andre, O.; Benamar, O. A.; Dumas, B.; Tsugawa, H.; Marti, G., MS-CleanR: A Feature-Filtering Workflow for Untargeted LC-MS Based Metabolomics. *Anal Chem* **2020,** *92* (14), 9971-9981.
23. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M., MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* **2015,** *12* (6), 523-6.
24. Monnerie, S.; Petera, M.; Lyan, B.; Gaudreau, P.; Comte, B.; Pujos-Guillot, E., Analytic Correlation Filtration: A New Tool to Reduce Analytical Complexity of Metabolomic Datasets. *Metabolites* **2019,** *9* (11).
25. Schiffman, C.; Petrick, L.; Perttula, K.; Yano, Y.; Carlsson, H.; Whitehead, T.; Metayer, C.; Hayes, J.; Rappaport, S.; Dudoit, S., Filtering procedures for untargeted LC-MS metabolomics data. *BMC Bioinformatics* **2019,** *20* (1), 334.
26. Broadhurst, D.; Goodacre, R.; Reinke, S. N.; Kuligowski, J.; Wilson, I. D.; Lewis, M. R.; Dunn, W. B., Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* **2018,** *14* (6), 72.
27. Costa, C.; Maraschin, M.; Rocha, M., An R package for the integrated analysis of metabolomics and spectral data. *Comput Methods Programs Biomed* **2016,** *129*, 117-24.
28. https://rdrr.io/cran/specmine/man/flat_pattern_filter.html.
29. Guitton, Y.; Tremblay-Franco, M.; Le Corguille, G.; Martin, J. F.; Petera, M.; Roger-Mele, P.; Delabriere, A.; Goulitquer, S.; Monsoor, M.; Duperier, C.; Canlet, C.; Servien, R.; Tardivel, P.; Caron, C.; Giacomoni, F.; Thevenot, E. A., Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. *Int J Biochem Cell Biol* **2017,** *93*, 89-101.
30. https://workflow4metabolomics.org/sites/workflow4metabolomics.org/files/files/w4e-2016-data_processing.pdf.

31. Jankevics, A.; Merlo, M. E.; de Vries, M.; Vonk, R. J.; Takano, E.; Breitling, R., Separating the wheat from the chaff: a prioritisation pipeline for the analysis of metabolomics datasets. *Metabolomics* **2012,** *8* (Suppl 1), 29-36.
32. Scheltema, R. A.; Jankevics, A.; Jansen, R. C.; Swertz, M. A.; Breitling, R., PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. *Anal Chem* **2011,** *83* (7), 2786-93.
33. http://mzmatch.sourceforge.net/tutorial.mzmatch.r.advanced.php.
34. Davidson, R. L.; Weber, R. J.; Liu, H.; Sharma-Oates, A.; Viant, M. R., Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. *Gigascience* **2016,** *5*, 10.
35. Edmands, W. M.; Barupal, D. K.; Scalbert, A., MetMSLine: an automated and fully integrated pipeline for rapid processing of high-resolution LC-MS metabolomic datasets. *Bioinformatics* **2015,** *31* (5), 788-90.
36. Hughes, G.; Cruickshank-Quinn, C.; Reisdorph, R.; Lutz, S.; Petrache, I.; Reisdorph, N.; Bowler, R.; Kechris, K., MSPrep--summarization, normalization and diagnostics for processing of mass spectrometry-based metabolomic data. *Bioinformatics* **2014,** *30* (1), 133-4.
37. Uppal, K.; Walker, D. I.; Jones, D. P., xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data. *Anal Chem* **2017,** *89* (2), 1063-1067.
38. Wen, B.; Mei, Z.; Zeng, C.; Liu, S., metaX: a flexible and comprehensive software for processing metabolomics data. *BMC Bioinformatics* **2017,** *18* (1), 183.