A workflow for UPLC-MS non-targeted metabolomic profiling in large human population-based studies

Andrea Ganna*[1,2], Tove Fall*[1], Woojoo Lee[3], Corey D. Broeckling[4], Jitender Kumar[1], Sara Hägg[1,2], Patrik K.E. Magnusson[2], Jessica Prenni[4,5], Lars Lind[6], Yudi Pawitan[2], Erik Ingelsson[1]

* denotes equal contribution

1. Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden.
2. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.
3. Department of Statistics, Inha University, Incheon, Korea.
4. Proteomics and Metabolomics Facility, Colorado State University, Fort Collins, Colorado, USA.
5. Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, CO, USA.
6. Department of Medical Sciences, Uppsala University, Uppsala, Sweden.

Address for correspondence:

Erik Ingelsson, MD, PhD, FAHA

Department of Medical Sciences, Molecular Epidemiology, Uppsala University Hospital, SE-751 85 Uppsala, Sweden

Phone: +46-70-7569422; Fax: +46-18-515570; E-mail: erik.ingelsson@medsci.uu.se

ABSTRACT

Metabolomic profiling is an emerging technique in life sciences. Human studies using these techniques have been performed in a small number of individuals or have been targeted at a restricted number of metabolites. In this article, we propose a data analysis workflow to perform non-targeted metabolomic profiling in large human population-based studies using ultra performance liquid chromatography-mass spectrometry (UPLC-MS). We describe challenges and propose solutions for quality control, statistical analysis and annotation of metabolic features. Using the data analysis workflow, we detected more than 8,000 metabolic features in serum samples from 2,489 fasting individuals. As an illustrative example, we performed a non-targeted metabolome-wide association analysis of high-sensitive C-reactive protein (hsCRP) and detected 407 metabolic features corresponding to 90 unique metabolites that could be replicated in an external population. Our results reveal unexpected biological associations, such as metabolites identified as monoacylphosphorylcholines (LysoPC) being negatively associated with hsCRP. R code and fragmentation spectra for all metabolites are made publically available. In conclusion, the results presented here illustrate the viability and potential of non-targeted metabolomic profiling in large population-based studies.

INTRODUCTION

Metabolomic profiling, or metabolomics, can be described as a holistic approach to the study of low-weight molecules (<1,500 Da) called metabolites. These chemical entities, which are the intermediates or end products of metabolism, serve as direct signatures of biochemical activities and play an important role in many common diseases, such as type 2 diabetes and cardiovascular diseases [1-5].

Improvements in instrumental technologies and advances in bioinformatics tools have provided the possibility to perform metabolomics on large prospective epidemiological studies with thousands of individuals and hundreds of phenotypes. Combining this rapid progress with improved understanding of the genetic determinants of metabolic changes [6,7], this 'omics' technology has already generated new hypotheses for therapeutic interventions and biomarkers discovery [8].

Two different types of analytical approaches are currently used in metabolomics studies: targeted and non-targeted. Most published studies in the human metabolomics field have used a targeted approach [9]. This approach relies on the measurements of a specific subset of metabolites, typically focusing on pathways of interest. At present, more than 40,000 endogenous and exogenous metabolites have been identified in humans [10]. Given that most of the human studies use a targeted approach investigating up to 200 metabolites, a large number of disease-related metabolites are likely to have been missed. The 'non-targeted' approach has the advantage to simultaneously measure as many metabolites as possible from a biological sample. Ultra performance liquid chromatography (UPLC) and gas

chromatography (GC) coupled with mass-spectrometry (MS) have been widely used in non-targeted metabolomics efforts [11].

There are two different approaches to non-targeted metabolomics.

In the first approach, metabolite identification is performed on the entire metabolic profile before the statistical analysis, which is then carried out only on the annotated metabolites or on both the annotated metabolites and those that could not be annotated with the current library of standards. The second approach, which is used in the workflow of the present study, detects as many metabolic features (metabolic fragments obtained during ionization process or an intact [molecular] ion of the original metabolite) as possible and treats each of them as separate variables in the analysis. A metabolic spectrum is reconstructed for each metabolic feature of interest and used for annotation through public databases or private libraries. Moreover, the features that cannot be annotated are reported and the spectra can be shared with the scientific community. This non-targeted metabolic feature-based approach has not been fully explored in large population studies due to complexity in alignment, analysis and annotation of the data generated from MS.

There are few previous papers describing the analytical procedures for large-scale non-targeted metabolic profiling in serum or plasma [12,13] and in tissues [14], especially focusing on sample preparation, liquid chromatography approaches and data pre-processing. However, less attention has been paid in describing the bioinformatics and statistical analysis of data from large population-based studies.

In this article, we aim to describe our data analysis workflow for non-targeted metabolomics analysis in large epidemiological studies, from raw data to reporting

statistical associations and identification of metabolites. The data analysis workflow is illustrated by an example in which we have investigated the metabolites associated with high-sensitive C-reactive protein (hsCRP) levels in two population-based studies of 2,489 individuals. The R code used to process and annotate non-targeted metabolomics data is made available at https://github.com/andgan/metabolomics_pipeline.

RESULTS

Three fundamental quantities: retention time, mass charge ratio and intensity

In mass spectrometry, a three-dimensional signal comprising retention time, mass charge ratio (m/z) and intensity is generated. The retention time is the time of elution for any metabolite in liquid chromatography. The m/z is the mass over charge of the metabolite and reflects the molecular mass, adducts, and in-source fragments of a compound. While the m/z and retention time are the two fundamental components to identify a particular metabolite, the intensity describes the abundance of this metabolite.

All signals contained in visual output from the mass spectrometer are shown in the form of a chromatogram; see **Figure 1** for an example. Each peak in the chromatogram represents a metabolic feature and the peak area is proportional to the relative abundance of that feature. A metabolite feature can be either a molecular fragment obtained during the ionization process or an intact (molecular) ion of the original metabolite. Multiple metabolite features representing a single metabolite are often detected. This phenomenon is due to the occurrence of in-source fragmentation, adduct and multimer formation, naturally occurring isotopes, and multiple charge states.

Workflow description

We have schematically outlined the analytical workflow in **Figure 2**. The workflow can be divided in four modules:

1) In the first module, peaks from each chromatogram generated by the mass spectrometer are detected, aligned and grouped across samples. Each group of peaks with unique mass-to-charge ratio (m/z) and retention time is called a ‘feature’.

2) In the second module, the intensities of features are log-transformed and normalized to take into account factors of unwanted variation. Quality control is performed to exclude samples with unusual total feature intensity and features with poor replicability.

3) In the third module, univariate statistical analysis is performed to identify features associated with the outcome of interest. False discovery rate (FDR) is controlled to select significant features without being too conservative. These features are then independently tested in a separately processed study to minimize the number of false-positives.

4) Finally, using an indiscriminate data acquisition workflow coupled with correlational grouping [15], both indiscriminant (id) MS and tandem MS (idMS/MS) spectra (if available) are generated and used to identify significant features through private library matching or annotate them through public databases. Four identification and annotation steps with different levels of confidence are described.

Module 1: feature detection

Peak Detection

As can be seen in **Figure 1**, a chromatogram contains several thousands peaks with large intensity differences across a spectrum of m/z and retention times. To obtain a true non-targeted metabolite profile, it is important to quantify as many true peaks as possible. To perform this task we use XCMS [16], a bioinformatics package implemented in R and widely used by the metabolomics community. The algorithm that performs the peak detection is called *centWave* [17] and it is implemented in the *xcmsSet* function. Two instrument-dependent parameters have a key role in the algorithm performances: (1) *ppm*, indicating the mass spectrometer accuracy; and (2) *peakwidth*, indicating the chromatographic peak-width range. The *ppm* parameter should be set to a generous multiple of the mass accuracy of the mass spectrometer (e.g. 25 ppm if the mass accuracy is 2-3 ppm using a multiple of 10). The *peakwidth* parameter is used to give an approximate estimate of the peak width range. Visualization tools like AMDIS (http://chemdata.nist.gov/mass-spc/amdis/) or instrument-specific software can be used define this parameter. It is important to note that the *peakwidth* parameter should not be interpreted as a stringent threshold since it allows peak detection in a slightly larger range than specified. To decide the values of remaining parameters in the *xcmsSet* function, we suggest an approach based on iterative testing of different settings as discussed at the end of this module, or to evaluate the algorithm performances by looking at the plot obtained, for one representative chromatogram, with the *findPeaks* function (see R code at https://github.com/andgan/metabolomics_pipeline). In **Figure 3**, examples of well and badly detected peaks are illustrated. Importantly, when a large number of samples

are processed, the *xcmsSet* function can be parallelized using the *nslave* parameter, which can considerably speed up the peak detection step.

Peaks Alignment

Chromatographic shifts over time represent a common characteristic of chromatography-coupled mass spectrometry. Without a proper retention time alignment, peaks representing the same compound would not be correctly grouped across different samples because of differences in retention time. This issue is critical in large studies, where the mass spectrometer might run for months and some level of analytical variation is unavoidable. The *retcorr* function from the XCMS package realigns the samples by correcting the retention time shifting. We suggest using the *obiwarp* algorithm [18] (implemented in the *retcorr* function), as it is more stable for large number of samples. This algorithm uses the sample with the largest number of peaks as reference for alignment. It is therefore important to check that the abundance of peaks in the 'centring sample' is not due to abnormal laboratory conditions. Since the amount of correction in retention time for each peak can be obtained from the *retcorr* function, samples with an abnormal total retention time correction can be easily identified and removed. Similarly, visualization of retention time corrections across samples can be informative of batch effects or laboratory issues.

Peak Grouping

In this step, the peaks that have been detected and aligned are now grouped across samples. Each group, typically called 'metabolic feature' or simply 'feature', has unique retention time and m/z. The *group* function from the XCMS package performs this task. Three parameters play a key role: (1) *bw*, the retention time deviation to be

allowed for grouping; (2) *mzwid*, m/z width to determine the peak grouping across sample; (3) *minfrac*, minimum fraction of samples in each group needed to call it a valid feature. Simulation approaches based on iterative testing of different settings (see discussion at the end of this module) and the plot obtained from the *group* function (see R code in **Supplementary Material**) is useful to determine the right values for these three parameters. With a large number of samples, we suggest to keep the *minfrac* parameter relatively low (for example at 0.03) to allow rare and exogenous (e.g. cotinine, a metabolite of nicotine) metabolites to be included in the analyses. However, the disadvantage of using a too low *minfrac* parameter is an increased risk of detecting false-positive features or noise. In **Figure 4**, we report the graphical output of the *group* function and discuss some examples of well and badly grouped peaks.

Filling missing features

The majority of features are not detected in all samples. Indeed, most of the feature intensities are missing when arriving at this step. This might be due to a true lack of signals for certain samples (for example, cotinine should not be detectable in non-smokers) or, most likely, because some peaks are missed by the peak detection algorithm due to the inherent uncertainty when the intensity is close to the signal-to-noise cutoff. To overcome this problem, the *fillpeak* function from the XCMS package uses the raw data, following retention time correction, to fill the intensity values for each of the missing features. However, even after the *fillpeak* function it is possible that some features are not detected in certain samples and a zero value is assigned to these.

Finally, the *groupval* function extracts the *m* by *n* matrix of intensities, with *m* number of features and *n* number of samples. It should be noted that the *n* samples include replicates, and both MS and idMS/MS data (when idMS/MS data were collected, i.e. Waters MS^E [19]); hence, *n* is four times the number of samples originally collected from study participants. Each feature is uniquely tagged by a mass and retention time, approximated to the third decimal place for accurate mass MS systems.

How to detect best parameter configuration in XCMS

The parameters used in XCMS to detect, align and group peaks can drastically change the number and quality of the identified features. Brodsky and colleagues have shown how the parameter tuning affects the inter-replicate correlation and which parameters are likely to have the largest influence [20]. The authors of XCMS have suggested parameter values for different UPLC/MS instruments; both in a published paper [21] and in the online version of XCMS [22]. Our advice is to randomly select a small number of samples (n=20-40) with duplicate or even triplicate injections and to try several parameter configurations within a reasonable interval around the suggested values. The parameter configuration that maximizes the intra-replicates correlation is likely to be the best choice. If replicates are not available, manual inspection of the plots generated by the peak detection and grouping algorithms can help to indicate whether the selected configuration is appropriate (**Figure 3 and 4**).

Module 2: quality control

Log_2 transformation, sample outlier exclusion, normalization

Our data acquisition workflow involves the simultaneous collection of low and high collision energy in alternating scans [15,19]. Thus, for every sample two corresponding data files are generated. In the previous module, we jointly processed idMS (low collision energy) and idMS/MS (high collision energy) data. This was done to ensure common features names in both datasets. In the next steps (module 2-3), only the idMS data are used, while the idMS/MS data is used in the last module for annotation of significant features.

First, feature intensities are transformed to the log_2 base scale to approximate normal distribution while keeping the interpretation of the coefficients straightforward; e.g. 1-unit change is equivalent to a doubling on the original scale. To identify potential outliers, the total intensity (sum of the intensities) is calculated for each sample. A very low total intensity might indicate sample degradation or technical errors and these samples should be excluded.

Second, feature intensities are normalized to allow comparability across samples. Normalization is usually performed using statistical models to derive an optimal scaling factor based on data distribution (e.g. quantile or median normalization) or by inclusion of internal or external standards. The latter approach has larger efficiency, but adds complexity to sample preparation [23]. Among methods for statistical normalization, quantile normalization has been shown to outperform linear normalization [20,24]. When sources of unwanted variability, such as batch effect, seasonal effect or storage time affect the comparison between samples, ANOVA-type

normalization methods can be advantageous [25] [26]. A simple approach can be applied by obtaining the residuals from a linear model for association between each feature and the factors of unwanted variability. These factors can be identified by studying the association between several technical variables and the first principal components. This and other methods are commonly used to adjust for batch effects in microarray data and are discussed in depth by Leek and colleagues [27]

In addition to visual investigation of the data before and after normalization, correlation between replicates can be used as criteria to compare and select the optimal normalization approach.

Finally, feature intensities are averaged between duplicates to reduce the inherent instrumental variability. Those features with poor correlation between replicates are excluded.

Module 3: analysis

Univariate statistical analysis

Similar to genome-wide association studies (GWAS) in the field of genomics, we propose a non-targeted metabolome-wide association study (MWAS) design, using metabolites as independent variables (instead of genetic variants as in a GWAS). We suggest performing a univariate analysis (e.g. linear regression for continuous outcomes, logistic regression for dichotomous outcomes or Cox regression for time-to-event outcomes) for each feature. These models are typically adjusted for age, sex and analysis-specific covariates; but for specific outcomes, additional biological covariates can be included depending on the research question.

Given the large number of statistical tests performed, correction for multiple testing needs to be considered. Two widely used methods to address the issue of multiple testing are Bonferroni correction and false discovery rate (FDR). Bonferroni's method controls the probability of making any false positives among all tests, which is reasonable in GWAS, where few discoveries are expected from the large number of statistical tests. However, in non-targeted metabolomics studies, especially on metabolism-related outcomes, a large number of discoveries are expected due to the high degree of correlation between features, and strong biological links between circulating metabolites and development of these outcomes. In such situations, Bonferroni's method is much too conservative and therefore the FDR method should be considered. This method is more appropriate as it gives better power to detect a larger number of biologically significant findings while controlling the expected proportion of false positives. We used simulations to investigate the behaviour of these two multiple-testing correction methods in a setting with highly clustered data,

similar to what is observed in metabolomics data (**Figure 5**). Each set of highly clustered simulated data constitute a block, and blocks are independent. We assumed an exchangeable correlation structure within the block. Therefore, the test statistics within the block are strongly correlated, but not correlated with the test statistics in other blocks. Two scenarios were studied: (1) 40% true signals, all with a moderate effect size; and (2) 1% true signals with large effect size. We simulated 1,000 individuals (500 cases and 500 controls) and 10,000 features clustered in 500 high correlated blocks. The correlation of features in the block was 0.9. The two-sample t-statistics were used, and the P-values were obtained by permuting group labels. The standard estimate of FDR as a function of the P-values was computed [28].

The Bonferroni method selected only a small fraction of the true signals (**Figure 5**; panel A), even after allowing a higher P-value threshold. On the other hand, the FDR method captured a much larger number of features, by allowing a small fraction of false discoveries. The effects of block independence can be seen from the perspective of the false discovery proportion (FDP), which is the random proportion of false discoveries among the rejected nulls. Note that true FDPs varied from realization to realization. When compared with the true FDP, the FDR estimate was nearly unbiased and maintained low variability for small FDR (**Figure 5**; panel C). When we simulated a scenario with few, highly significant features, the two methods performed similarly (**Figure 5**; panel B) with an increased variability of FDR estimates (**Figure 5**; panel D), partially due to the correlated structure of the data. This observation has been described and discussed previously [28].

The FDR threshold to use for inclusion in the replication phase is quite arbitrary and study-dependent. When there is a large number of expected significant features (e.g.

when investigating metabolic traits), a conservative FDR (1% to 5%) may be used to facilitate the replication and annotation procedures (by having fewer features to follow-up in the downstream workflow), while if there are fewer expected findings (such as for non-metabolic traits) the FDR threshold may be set higher (e.g. up to 20%). This should also be balanced against the proportion of expected false positive findings that you are willing to accept in that specific setting.

Replication in an external population

Even if the discovery and replication studies have been analysed in the same laboratory and under the same experimental conditions, the metabolic features detected might be different because of study-specific sampling, storage and handling or due to different bioinformatics data processing (e.g. the *minfrac* parameter depends on the number of samples as more features are detected when larger number of samples are jointly processed). In order to determine whether two features represent the same compound, both m/z and retention time need to be matched. The m/z match can be done within a certain confidence interval, depending on the accuracy of the mass spectrometer (e.g. $\pm$0.02 m/z differences). The retention time matching is more challenging and depends on the retention time correction applied during peak alignment. In general, a larger number of features that are significantly replicated among all those that are matched indicate a better quality of the matching strategy.

Among the features taken forward to validation and that can be matched in the replication sample, the number of promising features that can be replicated is controlled by the FDR, but also by the underlying effect sizes and the validation sample sizes.

Module 4: identification and annotation

Generation of idMS and idMS/MS spectra for significant features

Tandem mass spectrometry (MS/MS) using precursor ion selection is a well-established tool to elucidate metabolite structure. In traditional non-targeted metabolomics analysis, global MS analysis is followed by targeted MS/MS experiments to confirm the putative identification of significant features. However, this additional step requires additional analytical work, samples, instrumental time and data processing. Recent developments in mass spectrometer technologies have allowed for the acquisition of both MS and MS/MS simultaneously in the same experiment, by alternating low and high collision energy scans. Using this approach, which is a unique feature of Waters systems, ions are fragmented in an indiscriminate manner (i.e. no precursor ion selection) [19]. The challenge of this acquisition approach is the correct assignment of precursor-product ion relationships. Recently, we have shown that by taking advantage of inherent variability within the data, correlational relationships can be used to make these assignments and reconstruct both idMS and idMS/MS spectra for each feature [15]. Feature identification can be obtain by in-house spectral libraries and feature annotation can then be facilitated by spectral matching against publically available databases (e.g. METLIN [29], MassBank [30]). R code for generating idMS and idMS/MS spectra is provided in the https://github.com/andgan/metabolomics_pipeline.

Identification and annotation

Results of an MWAS are represented as clusters of highly correlated features. Each cluster includes features with very similar retention time, but different masses, due to fragments and derivative ions of the same metabolites. There are several approaches to the identification or annotation of metabolic features, each reflecting different

levels of confidence. In **Figure 6**, we report the workflow that we use to identify or annotate the features and the level of confidence obtained within each step. Some steps can only be performed given the availability of idMS/MS spectra, which are specific of our platform. For a detailed discussion of the levels of confidence, we refer to the definition by the Metabolomics Standard Initiative (MSI) [31].

The first approach has the highest confidence (level 1 according to MSI) and it is based on the direct matching of idMS and/or idMS/MS generated spectra to an in-house spectral library of authentic standards collected under same experimental conditions. We define the matching at this confidence level as 'identification', in contrast to 'annotation', which is used as a lower level of confidence.

Thus, a level 1 identification is based on matching accurate mass, fragmentation pattern, and retention time. The main limitation with this approach is the lack of authentic standards and the additional analytical and instrumental effort needed to create a spectral library.

The following three approaches correspond to the MSI confidence level 2 and are therefore referred to as approach 2a, 2b and 2c, where 2a is the most confident. They are all based on spectrum and/or m/z similarities, but not retention time similarity, and their annotation relies on information available in public databases. The three levels can be summarized as follows: level 2a corresponds to matching both accurate mass and fragmentation pattern to library spectra in a public database; level 2b corresponds to a match based on only fragmentation pattern (precursor ion cannot be assigned or is undetected); level 2c corresponds to matching of only accurate mass to a public database or literature. Both levels 2a and 2c require assignment of a molecular

precursor ion in the spectrum. This process can be challenging as often the M+H ion of the metabolite is either not present or is in low abundance compared to the base peak. Several programs, such as CAMERA [32] and PeakML/mzMatch [33] can help by detecting highly correlated clustered features (based on co-elution in a single sample) and annotating isotopes and predictable adducts and in-source fragments. For example, m/z differences between spectral peaks can inform about the potential adducts present in the spectra and allow back-calculation of the molecular weight; typical adducts observed in positive mode are $[M+H]^+$, $[M+Na^{]+}$, $[M+K]^+$, $[2M+H]^+$, $[2M+Na]^+$, and $[2M+K]^+$. Less common adducts might also be observed and complicate the back-calculation of the molecular weight [34].

Level 2a requires the ability to search public spectral libraries based on both accurate precursor mass and fragmentation pattern. This approach can be implemented in the MS/MS spectrum match option in METLIN [29] and the MS/MS search option in HMDB [10]. Level 2b involves spectral searching only (i.e. precursor ion is not easily assigned). This approach can be implemented in the spectrum search option in MassBank [30], multiple fragments option in METLIN, or in-source spectral search in the NIST search program (http://chemdata.nist.gov/mass-spc/ms-search/). This strategy usually provides a larger number of "matched" metabolites per feature, and hence subsequent screening based on visual inspection of the top findings is needed to determine the most reliable match. Level 2c requires only matching based on accurate mass of the assigned precursor in the idMS spectrum. This approach is utilized when spectral library data is not available for the putative metabolite. Caution should be used when assigning level 2c annotation as multiple metabolites (e.g. multiple

compounds with the same empirical molecular formula but different structure) will share the same molecular weight.

The next approach, which corresponds to MSI confidence level 3, uses a combination of spectral data, accurate mass, and retention time to assign the metabolite to a chemical class (without knowing the exact origin of the metabolite). Finally, if all the other approaches have failed in the annotation of the metabolite or metabolite class, the metabolite is annotated as “unknown” (level 4 according to MSI).

It is our view that spectral data should be provided for all identified and annotated features, included those annotated as “unknown”. This ensures data transparency and enables prospective annotation. For a thoughtful discussion regarding identification and annotation of non-targeted metabolomic features, we suggest the article from Dunn and colleagues [34], and a recently published protocol on metabolite characterization using METLIN [35].

Example of application: high-sensitive c-reactive protein

We applied the workflow described above to serum specimens from two population-based studies including 2,489 fasting individuals (1,519 from TwinGene and 970 from PIVUS); see the **Materials and Methods** section for a detailed description of the study participants. The section is organized as follow. First, we describe the processing of raw metabolomics data (module 1) and the quality control (module 2). A similar approach was applied both to TwinGene and PIVUS; however, details are reported only for results from PIVUS to exemplify the workflow. Second, we describe a MWAS on levels of hsCRP levels in TwinGene with replication of the significant findings in PIVUS (module 3). Third, we show how to annotate the validated features using both publically available databases and a private spectral library (module 4).

Iterative testing of different settings to determine the best XCMS parameters

Inter-replicate correlation was evaluated under different parameter configurations. Thirty individuals were randomly chosen (120 files; including duplicate injections, and both idMS and idMS/MS data) and run through the detection, alignment, grouping and filling peaks steps for 2,161 combinations of different levels for five parameters (*snthresh, mzdiff, minfrac, mzwid, bw*). In **Figure 7**, the average correlation between duplicates is reported separately for each parameter suggesting that a configuration with relatively higher *snthresh* and *minfrac*, and lower *mzwid* and *bw* improved the intra-replicates correlation. The *mzdiff* parameter did not influence the intra-replicates correlation. Similar to our observations, Brodsky and colleagues [20] also reported that an increase in the *minfrac* and decrease in the *mzwid* parameters were associated with higher between-replicates correlation.

To maximize the number of detected features while maintaining a high inter-replicate correlation, a *minfrac* parameter (minimum fraction of sample in each group for calling it as a valid feature) slightly lower than the optimal value was employed.

The following parameters were used in the final analysis for peak detection (xcmsSet *function*): *method="centWave", ppm=25, peakwidth=c(2:15), snthresh=8, mzCenterFun="wMean", integrate=2, mzdiff=0.05 prefilter=c(1,5);* peak alignment (*rector* function): *method="obiwarp",plottype="deviation";* peak grouping (*group* function): *bw=2, minfrac=0.03, max=100, mzwid=0.01* and peak filling (*fillPeaks.chrom* function). All the other parameters were set to default.

Module 1: Processing of metabolomics data in PIVUS

Using these parameters in XCMS, 8,185 features were detected by processing 3,880 files (including replicates, and both idMS and idMS/MS data) from PIVUS. In total, 31,755,397 peaks (approximately 8,206 per data file) were detected and aligned across files. Data files that could not be properly aligned resulted in a high retention time correction (difference between retention time before and after alignment). In **Supplementary Figure 1,** red dots indicate files with abnormal retention time correction that needed to be excluded from further analysis. After exclusion, the average retention time correction in absolute values was 0.37 seconds (max: 2.35 seconds). Among the 8,185 aligned features, 86% of them was detected in less than half of the samples with a signal/noise level greater than 8. The *fillpeak* algorithm identifies the intensities for the missing data files by using the original raw data.

Module 2: Quality control in PIVUS

In the next step, idMS/MS data was excluded and quality control was performed on 1,898 files (including replicates). Two files were removed because of the unusually low total intensity (sum of all the features intensities).

We calculated the correlation between replicates and the median of the coefficients of variation for each feature using different normalization approaches (**Supplementary Table 1**). The ANOVA-type approach outperformed normalization methods based on single parameter scaling. To determine which factors of unwanted variability should be included in the normalization, we studied the association between several technical variables and the scores obtained form the first two principal components. We observed that season of sample collection (P-value: $2.8x10^{-48}$), amount of retention time correction (P-value: $4.1x10^{-48}$) and storage time (P-value: $2.5x10^{-8}$) had the strongest association with the first principal component, while date of analysis (P-value: $7.0x10^{-104}$) had the strongest association with the second. The ANOVA-type approach simple regresses the feature intensity on the aforementioned variables and uses residuals to re-scale the data.

Distribution of log2-transformed feature intensities (**Figure 8**, panel A and B) and scores form the first two principal components (**Figure 8**, panel C and D) before and after normalization suggest an increased comparability between data files.

Finally, duplicates were averaged and the average intensity values used in the downstream analysis.

Module 3: MWAS of hsCRP levels in TwinGene and replication in PIVUS

The age, sex and hsCRP distributions in TwinGene and PIVUS are reported in **Supplementary Table 2**. Participants with hsCRP higher than 20 mg/L (N=88 in

TwinGene; N=21 in PIVUS) were excluded since they are likely to have acute infections or chronic inflammatory diseases, while our objective was to assess associations with hsCRP levels within the lower-moderately elevated range to study the metabolome in low-grade asymptomatic chronic inflammation.

In the remaining 1,431 participants of TwinGene, statistical analysis was performed by fitting 11,056 multivariable linear regressions (one for each feature detected in TwinGene) including age and sex as covariates and using log-transformed hsCRP as outcome. The Q-Q plot from the analysis in TwinGene (**Supplementary Figure 2)** highlights a major deviation from the null distribution, compatible with a scenario of a large number of highly correlated significant findings.

Among the 998 features with a FDR lower than 1%, 526 could be matched in PIVUS. Requirements for feature matching across studies included m/z difference of ± 0.02 and retention time differences of ± 3 seconds. The relatively low proportion of matched features could be explained by the inherent biological variability between the two studies and specifically by the larger size of TwinGene, which allows the detection of rare exogenous metabolites. The appropriate retention time window was determined by selecting the value that maximized the ratio between number of features replicated in PIVUS with a P-value < 0.05 and the total number of matched features (a high ratio indicates a good matching strategy). **Supplementary Figure 3** illustrates the behavior of these quantities over different time windows.

In a replication effort, univariate analysis (similarly to what was done in TwinGene) was performed in participants from PIVUS (N=949). Out of the 526 features with

FDR <1% in TwinGene that also could be detected in PIVUS, 435 features had a P-value < 0.05. Out of these 435 features, 407 were taken forward for identification and annotation. The remaining 28 features had a retention time of 32±2 seconds representing metabolites that are not retained on the chromatographic column. These compounds are highly polar, and co-elution presents challenges in quantitation, due to ionization suppression; and annotation, due to high complexity.

Module 4: Identification and annotation of features significantly associated with hsCRP in both TwinGene and PIVUS

Both idMS and idMS/MS spectra were generated for the remaining 407 features. Those with highly similar spectra, strong correlation and similar retention time were deemed to be from the same metabolite. Using this manual clustering approach, the 407 features were clustered in 90 groups, each representing a unique metabolite. Features belonging from chemically-linked metabolites (e.g. phosphocholines) are highly correlated, but retention times are different; hence, such features are not grouped together. Of the 90 metabolites, fifteen had missing idMS and/or idMS/MS spectra, indicating either that a single feature was detected for that compound, or that the feature represents a false positive result.

Each of the remaining 75 spectra was taken forward to the annotation step. Spectral matching was performed against our private library using the spectrum matching function implemented in the NIST mass spectral search program (http://chemdata.nist.gov/mass-spc/ms-search/). Using this first approach, three metabolites were identified with confidence level 1 according to the MSI definition.

These metabolites were 1-oleoyl-2-hydroxy-sn-glycero-3-phosphocholine (P-value: $2.1x10^{-11}$), hippuric acid (P-value: $1.5x10^{-6}$), and 18:1 fatty acid (P-value: $1.6x10^{-6}$).

For the remaining 72 candidate metabolites, we attempted the approach 2a (see detailed description above) based on inferred molecular weight and spectral matching against publically available databases. To determine the molecular weight of the compound, we inspected the idMS spectra and looked for expected patterns of m/z differences between peaks. Typically, a m/z difference of 21.982 between two peaks indicates that one peak is $[M + H]^+$and the other is $[M + Na]^+$. Similarly, isotope peaks will be observed from the contribution of naturally occurring carbon. We annotated a plausible molecular weight in 36 metabolites and performed a spectral comparison in METLIN as described by Zhu and colleagues [35]. For five metabolites, we had good matches and we assigned these as being of level 2a confidence. These were gamma-glutamyl-leucine (P-value: 1.0×10^{-10}), phenylalanylphenylalanine (P-value: $1.2x10^{-7}$), 3,4,5-trimethoxycinnamic acid (P-value: $1.1x10^{-8}$), 3-indolepropionic acid (P-value: $3.5x10^{-8}$) and 11β-hydroxyandrost-4-ene-3,17-dione (P-value: $3.1x10^{-5}$).

Annotation of the remaining 67 metabolites was attempted using the NIST public spectral library in-source fragmentation search and the METLIN *multiple fragment* search; neither of which require the annotation of the molecular weight of the metabolite (approach 2b). One metabolite was annotated as a prostaglandin, with several typical peaks, although the specific type of prostaglandin could not be determined using spectra matching with the NIST public library. Metabolites identified or annotated with these approaches (approach 1, 2a or 2b) are reported in the **Table 1**.

In the approach 2c, we attempted to annotate the remaining metabolites by matching the plausible molecular weights with METLIN and HMDB. Using this approach, we defined 27 additional metabolites as being lipids, with a specified molecular formula, while the exact structure could not be determined. The large majority of these lipids were phosphatidylcholines (PC) and the remaining were phosphatidylethanolamines (PE) or phosphoserines. Two were glycerolipids, two sphingolipids and one a fatty acid ester.

Among the remaining non-annotated metabolites, two had a peak in their corresponding idMSMS spectra at m/z 184.074, corresponding to a phosphocholine fragment, which indicates that these are phosphocholine-containing compounds (annotation level 3).

In summary, among the 90 metabolites significantly associated with hsCRP, 36 could be identified or annotated at MSI confidence level 1 or 2. Of these, 9 were identified with high confidence by matching the idMS/MS spectra with the private library or public databases. The strongest metabolite-hsCRP association that we could identify was 1-oleoyl-2-hydroxy-sn-glycero-3-phosphocholine. The number of features per metabolite, strength and direction of association in each study and metabolite class for all 90 metabolites are reported in **Supplementary Table 3**. All PCs with m/z<570.3 (LysoPC) had an inverse relationship with hsCRP. Spectra for all the 90 metabolites are made publically available to the scientific community (**Supplementary Material**).

DISCUSSION

The role of metabolomics in biomedical research

Metabolomic profiling is an important emerging technique in life sciences and biological research that can play a central role in systems biology [36]. The understanding of genetic and environmental determinants of diseases can be enhanced by the knowledge of their biochemical signature. Moreover, metabolomics can be a powerful tool to refine the pathways of association between diseases and genetic variants, or to generate new hypotheses regarding the underlying biological processes [37].

The growing expectation of the utility of metabolomics in medical and pharmacological research has been justified by several important population-based studies that have focused on genotype-metabolite [6,35] or metabolite-phenotype associations [1,38]. These studies have highlighted the heritability of metabolic traits, discovered new metabolite-genotype associations and suggested new biomarkers for common diseases. However, none of these studies have implemented a non-targeted metabolic feature-based approach.

There are several reasons why a non-targeted metabolic feature-based metabolomics approach has not been commonly performed in large human population studies. First, the processing and statistical analysis of data from mass spectrometry-based metabolomics in larger study samples is cumbersome, and there is a lack of methods descriptions and protocols of how to do this. Notable exceptions are the paper by Dunn and colleagues [12] and the recent tutorial by our group [13], both of which however are mainly focused on laboratory procedures. Second, annotation of findings

from non-targeted metabolomics experiments have been complicated by the lack of comprehensive spectral databases. Only in the past few years, these databases have gathered detailed information on a large number of metabolites and collected MS/MS spectra to facilitate annotation. Third, metabolomics results are highly dependent on sample collection, storage and analysis [39], which pose challenges for the replication of findings across studies.

Nevertheless, there are a few good examples of non-targeted metabolomics studies that have been performed in smaller clinical populations to identify markers of cardiovascular disease [40], in rats to investigate the chemical basis of neuropathic pain [41] and in glioblastoma cells [42]. The main advantage of the non-targeted over the targeted approach is the larger number of detected compounds, which translates into a greater potential for novel biomarker discovery, the ability to build annotation-independent metabolic scores for prediction, unbiased biological discoveries and the possibility to share uncharacterized metabolites with the scientific community, enhancing open-source, collaborative research.

Strengths and limitations of our approach

We believe our study have several important strengths and novel aspects. First, most of the articles about metabolomics methods are technical, presuming a strong chemical background and do not focus on large human populations and, are thus less suitable for the epidemiological community. Here we present a pipeline to process metabolomics data using language and concepts that are more familiar to researchers in population-based and clinical research. Second, the processing of the data using

XCMS software is extremely important and it is often underappreciated by other reports. We believe that a clear illustration of the XCMS parameters is of great importance. Third, statistical analyses have typically been described for small sample sizes and mainly rely on multivariate analysis (e.g. PCA). Here, we propose an analytical framework similar to those used in other areas within the molecular epidemiology field (e.g. genetic epidemiology) and that rely on univariate analysis. Finally, metabolic features annotation is of great importance. Researchers that use metabolomics data from a commercial supplier (which is the common situation for most epidemiological researchers) might not be aware of this aspect since they do not perform the annotation. Here we provide a clear illustration of the annotation process, which is important for researchers active in this field.

We also acknowledge several limitations. Although the analytical workflow we describe is based on the experience gathered from a specific type of data and platform, general considerations are valid for most metabolomics studies in large human populations, and our workflow can be easily implemented on data coming from other UPLC/MS platforms. Moreover, the workflow that we present is directly transferrable to any set of single channel data (e.g. GC/MS) with the caveat that settings should be adjusted to reflect different instruments (e.g. peak width and mass accuracy).

We further recognize that the idMS/MS spectra annotation based on data collected already at the first-pass analysis is a unique feature of our platform and it is not available in most UPLC/MS systems, though the feature grouping methods could be applied to low collision energy in-source MS spectra, which can be highly informative. However, most of the steps of feature annotation described in our paper

can be extended to experimentally generated MS/MS spectra, similarly to what has been described by Zhu and colleagues [35]. In this paper, we did not describe laboratory procedures, since this topic has been covered by previous protocols [12,14,15]. Different experimental workflows (e.g. use of internal standards) might require different data analysis procedures that are not covered in this study. In general, a rigorous experimental protocol will simplify the data quality control. Other software such as MetAlign [43] and MZMine 2 [44] can also be used to process and align mass-spectrometry data. We focused on XCMS since it is widely used in the scientific community and it has an online implementation that can be used without previous knowledge of the R language.

Conclusion

Non-targeted metabolomics enables investigation of a large number of biological and clinical questions in different areas. In this paper, we have presented a data analysis workflow for non-targeted metabolomics in biological specimens from large studies of human populations. We described challenges and proposed solutions for quality control, statistical analysis and annotation of metabolic features. Moreover, we demonstrate how this approach could be applied to real data from two epidemiological studies. Finally, we performed an MWAS of hsCRP to illustrate the potential of the non-targeted metabolomics.

METHODS

Ethic Statement

All participants of both studies gave a written consent and the Ethics Committees of Karolinska Institutet and Uppsala University approved the study.

Study description

TwinGene

The Swedish Twin Registry is a population-based national register including over 170,000 Swedish twins born from 1886 to 2000 [45]. TwinGene is a longitudinal sub-study within the Swedish Twin Register that was initiated to examine associations between genetic factors and cardiovascular diseases in Swedish twins. In TwinGene, we performed a case-cohort design by selecting all the incident cases of coronary heart diseases, type 2 diabetes, ischemic strokes and dementia up to 31st December 2009 and a sub-cohort (controls) of 1,796 individuals. Since it has been previously shown to improve the study efficiency [46], the subcohort was stratified on median age and sex, and for each of the four strata we randomly selected a number of participants proportional to the corresponding number of cases in the strata. The following analysis is conducted only on the individuals from the sub-cohort since this is representative of the original population. Of them, 1,519 passed the quality control and had hsCRP measured with a high-sensitivity method by Synchron LX systems (Beckman Coulter).

PIVUS

Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) (http://www.medsci.uu.se/pivus/pivus.htm) is a community-based study where all

men and women at age 70 living in Uppsala, Sweden were invited to participate in 2001 [47]. Of the 2,025 subjects invited, 1,016 subjects participated. Blood samples were available for 972 participants selected for metabolomic profiling. Of them, 970 passed the quality control. High-sensitive C-reactive protein (hsCRP) was measured with an ultra-sensitive particle enhanced immunoturbidimetric assay (Orion Diagnostica, Espoo, Finland).

Laboratory procedures and mass spectrometry

Metabolomics profiling was performed at the Proteomics and Metabolomics Facility, Colorado State University. Each sample was injected in non-consecutive duplicates in a randomized manner and analysed using ultra-performance liquid chromatography system (Waters Acquity UPLC system). Data was collected in positive ion mode. Scans were collected alternatively in MS mode at collision energy of 6 V and in idMS/MS mode using higher collision energy (15–30 V). idMS/MS (also called MS^E) allows for unbiased view of MS/MS fragmentation without additional experiments [15], as discussed further in module 4. Further information regarding the laboratory procedures is available in the **Supplementary Material**.

ACKNOLEDGMENTS

We thank Dr. Alexandra Jauhiainen for helpful insight and comments.

REFERENCES

1. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, et al. (2011) Metabolite profiles and the risk of developing diabetes. Nat Med 17: 448-453.
2. Sabatine MS, Liu E, Morrow DA, Heller E, McCarroll R, et al. (2005) Metabolomic identification of novel biomarkers of myocardial ischemia. Circulation 112: 3868-3875.
3. Shah SH, Bain JR, Muehlbauer MJ, Stevens RD, Crosslin DR, et al. (2010) Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events. Circ Cardiovasc Genet 3: 207-214.
4. Shah SH, Sun JL, Stevens RD, Bain JR, Muehlbauer MJ, et al. (2012) Baseline metabolomic profiles predict cardiovascular events in patients at risk for coronary artery disease. Am Heart J 163: 844-850 e841.
5. Floegel A, Stefan N, Yu Z, Muhlenbruch K, Drogan D, et al. (2013) Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. Diabetes 62: 639-648.
6. Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, et al. (2012) Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat Genet 44: 269-276.
7. Gieger C, Geistlinger L, Altmaier E, Hrabe de Angelis M, Kronenberg F, et al. (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet 4: e1000282.
8. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, et al. (2011) Human metabolic individuality in biomedical and pharmaceutical research. Nature 477: 54-60.
9. Dudley E, Yousef M, Wang Y, Griffiths WJ (2010) Targeted metabolomics and mass spectrometry. Adv Protein Chem Struct Biol 80: 45-83.
10. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, et al. (2013) HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res 41: D801-807.
11. Buscher JM, Czernik D, Ewald JC, Sauer U, Zamboni N (2009) Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. Anal Chem 81: 2135-2143.
12. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, et al. (2011) Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc 6: 1060-1083.
13. Broeckling CD, Heuberger AL, Prenni JE (2013) Large scale non-targeted metabolomic profiling of serum by ultra performance liquid chromatography-mass spectrometry (UPLC-MS). J Vis Exp: e50242.
14. Want EJ, Masson P, Michopoulos F, Wilson ID, Theodoridis G, et al. (2013) Global metabolic profiling of animal and human tissues via UPLC-MS. Nat Protoc 8: 17-32.
15. Broeckling C, Heuberger A, Prince J, Ingelsson E, Prenni J (2013) Assigning precursor–product ion relationships in indiscriminant MS/MS data from non-targeted metabolite profiling studies. Metabolomics 9: 33-43.
16. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78: 779-787.

17. Tautenhahn R, Bottcher C, Neumann S (2008) Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics 9: 504.
18. Prince JT, Marcotte EM (2006) Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. Anal Chem 78: 6140-6152.
19. Plumb RS, Johnson KA, Rainville P, Smith BW, Wilson ID, et al. (2006) UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. Rapid Commun Mass Spectrom 20: 1989-1994.
20. Brodsky L, Moussaieff A, Shahaf N, Aharoni A, Rogachev I (2010) Evaluation of peak picking quality in LC-MS metabolomics data. Anal Chem 82: 9177-9187.
21. Patti GJ, Tautenhahn R, Siuzdak G (2012) Meta-analysis of untargeted metabolomic data from multiple profiling experiments. Nat Protoc 7: 508-516.
22. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G (2012) XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem 84: 5035-5039.
23. Sysi-Aho M, Katajamaa M, Yetukuri L, Oresic M (2007) Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics 8: 93.
24. Lee J, Park J, Lim MS, Seong SJ, Seo JJ, et al. (2012) Quantile normalization approach for liquid chromatography-mass spectrometry-based metabolomic data from healthy human volunteers. Anal Sci 28: 801-805.
25. Kerr MK, Churchill GA (2001) Experimental design for gene expression microarrays. Biostatistics 2: 183-201.
26. Kerr MK, Churchill GA (2001) Statistical design and the analysis of gene expression microarray data. Genet Res 77: 123-128.
27. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11: 733-739.
28. Pawitan Y, Calza S, Ploner A (2006) Estimation of false discovery proportion under general dependence. Bioinformatics 22: 3025-3031.
29. Sana TR, Roark JC, Li X, Waddell K, Fischer SM (2008) Molecular formula and METLIN Personal Metabolite Database matching applied to the identification of compounds generated by LC/TOF-MS. J Biomol Tech 19: 258-266.
30. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, et al. (2010) MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45: 703-714.
31. Sumner L, Amberg A, Barrett D, Beale M, Beger R, et al. (2007) Proposed minimum reporting standards for chemical analysis. Metabolomics 3: 211-221.
32. Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S (2012) CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem 84: 283-289.
33. Scheltema RA, Jankevics A, Jansen RC, Swertz MA, Breitling R (2011) PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. Anal Chem 83: 2786-2793.
34. Dunn W, Erban A, Weber RM, Creek D, Brown M, et al. (2013) Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics 9: 44-66.

35. Zhu ZJ, Schultz AW, Wang J, Johnson CH, Yannone SM, et al. (2013) Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database. Nat Protoc 8: 451-460.
36. Patti GJ, Yanes O, Siuzdak G (2012) Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol 13: 263-269.
37. Suhre K, Gieger C (2012) Genetic variation in metabolic phenotypes: study designs and applications. Nat Rev Genet 13: 759-769.
38. Cheng S, Rhee EP, Larson MG, Lewis GD, McCabe EL, et al. (2012) Metabolite profiling identifies pathways associated with metabolic risk in humans. Circulation 125: 2222-2231.
39. Yin P, Peter A, Franken H, Zhao X, Neukamm SS, et al. (2013) Preanalytical aspects and sample quality assessment in metabolomics studies of human blood. Clin Chem 59: 833-845.
40. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, et al. (2011) Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature 472: 57-63.
41. Patti GJ, Yanes O, Shriver LP, Courade JP, Tautenhahn R, et al. (2012) Metabolomics implicates altered sphingolipids in chronic pain of neuropathic origin. Nat Chem Biol 8: 232-234.
42. Dang L, White DW, Gross S, Bennett BD, Bittinger MA, et al. (2009) Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature 462: 739-744.
43. Lommen A (2009) MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem 81: 3079-3086.
44. Pluskal T, Castillo S, Villar-Briones A, Oresic M (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11: 395.
45. Lichtenstein P, De Faire U, Floderus B, Svartengren M, Svedberg P, et al. (2002) The Swedish Twin Registry: a unique resource for clinical, epidemiological and genetic studies. J Intern Med 252: 184-205.
46. Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P, et al. (2012) Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. Am J Epidemiol 175: 715-724.
47. Lind L, Fors N, Hall J, Marttala K, Stenborg A (2005) A comparison of three different methods to evaluate endothelium-dependent vasodilation in the elderly: the Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) study. Arterioscler Thromb Vasc Biol 25: 2368-2375.

FIGURE LEGENDS

Figure 1. Example of chromatogram. Each peak represents a compound (e.g. metabolite) with unique m/z and retention time, and the peak area is proportional to the amount of the compound.

Figure 2. Workflow for doing metabolomics in large human populations from raw data to reporting statistical associations and identification of metabolites.

Figure 3. Examples of well and badly detected peaks obtained from the *findPeaks* function. **Panel A:** well detected peak; the Gaussian curve is correctly fitted to the intensity points. **Panel B:** a fictitious peak is detected in a highly noisy area (low signal-to-noise threshold); increasing the signal-to-noise threshold would avoid misidentifying this peak. **Panel C:** a single peak is split in two peaks; increasing the *peakwidth* parameter allows identifying only one peak. **Panel D:** a large peak is identified in an unrealistically large time window; reducing the *peakwidth* parameters allows avoiding this error.

Figure 4. Examples of well and badly grouped peaks obtained from the *group* function. **Panel A:** well grouped peaks; most of the peaks for a specific m/z (500.90-500.91) have been grouped in a similar retention time window. **Panel B:** peaks are spread across all the retention times and do not seem to cluster in a specific time window, however the algorithm detected a group around 200 seconds retention time. Increasing the *minfrac* parameter should avoid detecting groups with only few peaks. **Panel C:** a well-detected group is split in two groups; increasing the *bw* parameter allows to increase the retention time window for the group detection. **Panel D:** for

this specific m/z (159.05-159.06), only one group is identified; however two other groups seem to be present at higher retention times; decreasing the *minfrac* parameter allows to detect groups with fewer peaks.

Figure 5. Simulations to investigate the behaviour of two multiple testing correction methods in a metabolomics-like scenario. **Panel A** and **B:** Simulation to investigate the proportion of significant findings identified by varying the inclusion cut-off with two multiple testing correction methods: false discovery rate (FDR; black dashed lines) and Bonferroni (red dashed lines). Their averages are denoted by black solid and red solid line, respectively. The *x-axis* is the Bonferroni-corrected P-value or the FDR, depending on the method. In **panel A**, we simulate a scenario with 40 % true signals, all with a moderate effect size. In **panel B**, the true signals are 1 % of all signals and have large effect sizes. **Panel C** and **D:** We use the same simulation settings as in **panel A** and **B**, respectively, to investigate the bias and variability of FDR estimates compared to the true false discovery proportion. The average estimates of the 25 simulations (one each line) is overlapping with the identity line, indicating no bias. In **panel D**, we observe larger variability, which is however modest for small FDR.

Figure 6. Decision tree for identification and annotation of features from metabolomics data.

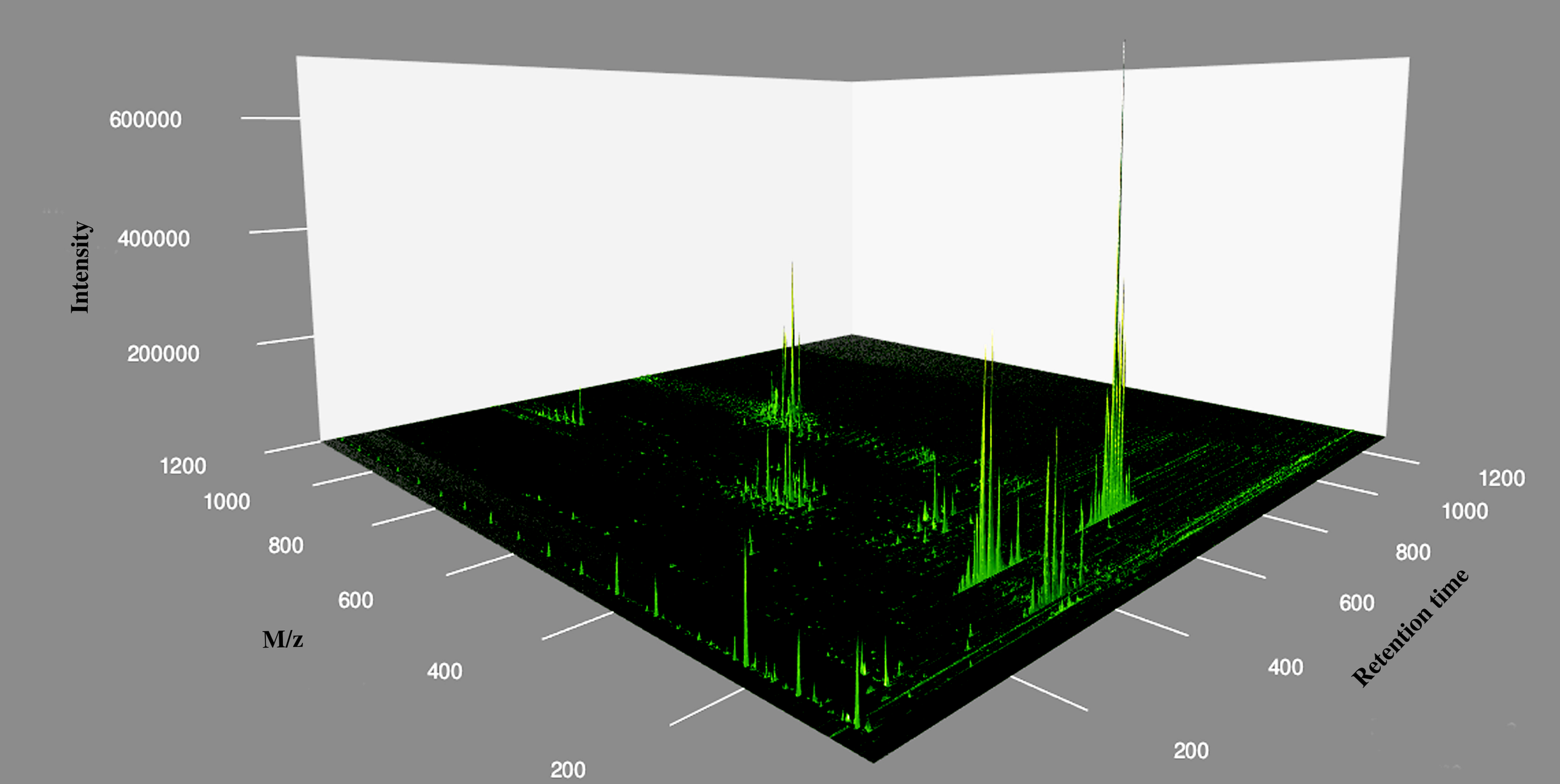
Figure 7. Results of iterative testing of different parameters for detection, alignment, grouping and filling steps on 30 random individuals (120 files) from PIVUS. We ran all 2,161 possible combinations of values within the reported ranges for five

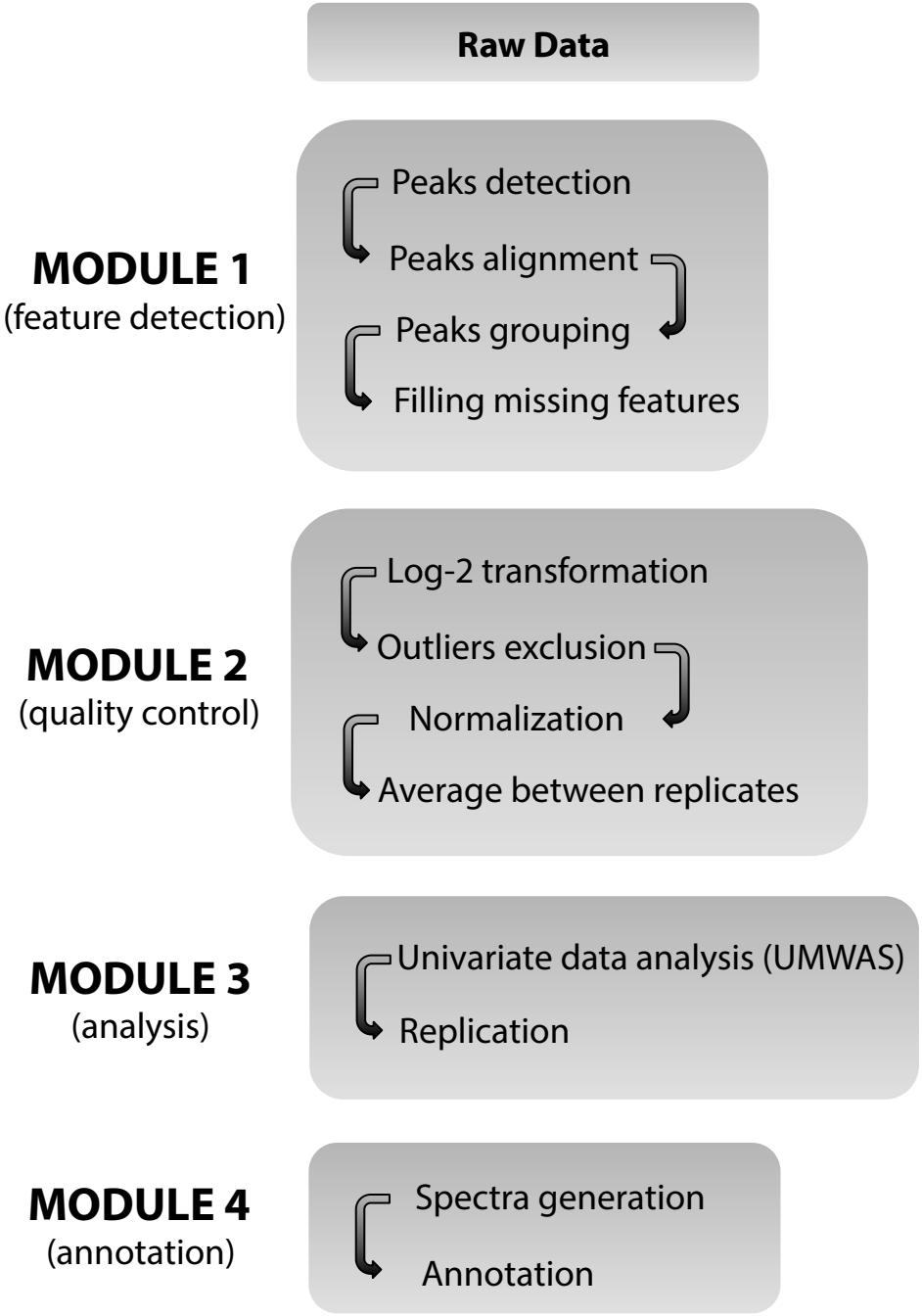
parameters (*sntresh, mzdiff, minfrac, mzwid, bw*). The reported correlations were the average of correlations for each parameter, across values of the other parameters. The dot in each panel indicates the value that has been used in the full analysis. The last panel indicates the observed correlation for number of features detected across all possible parameter combinations.

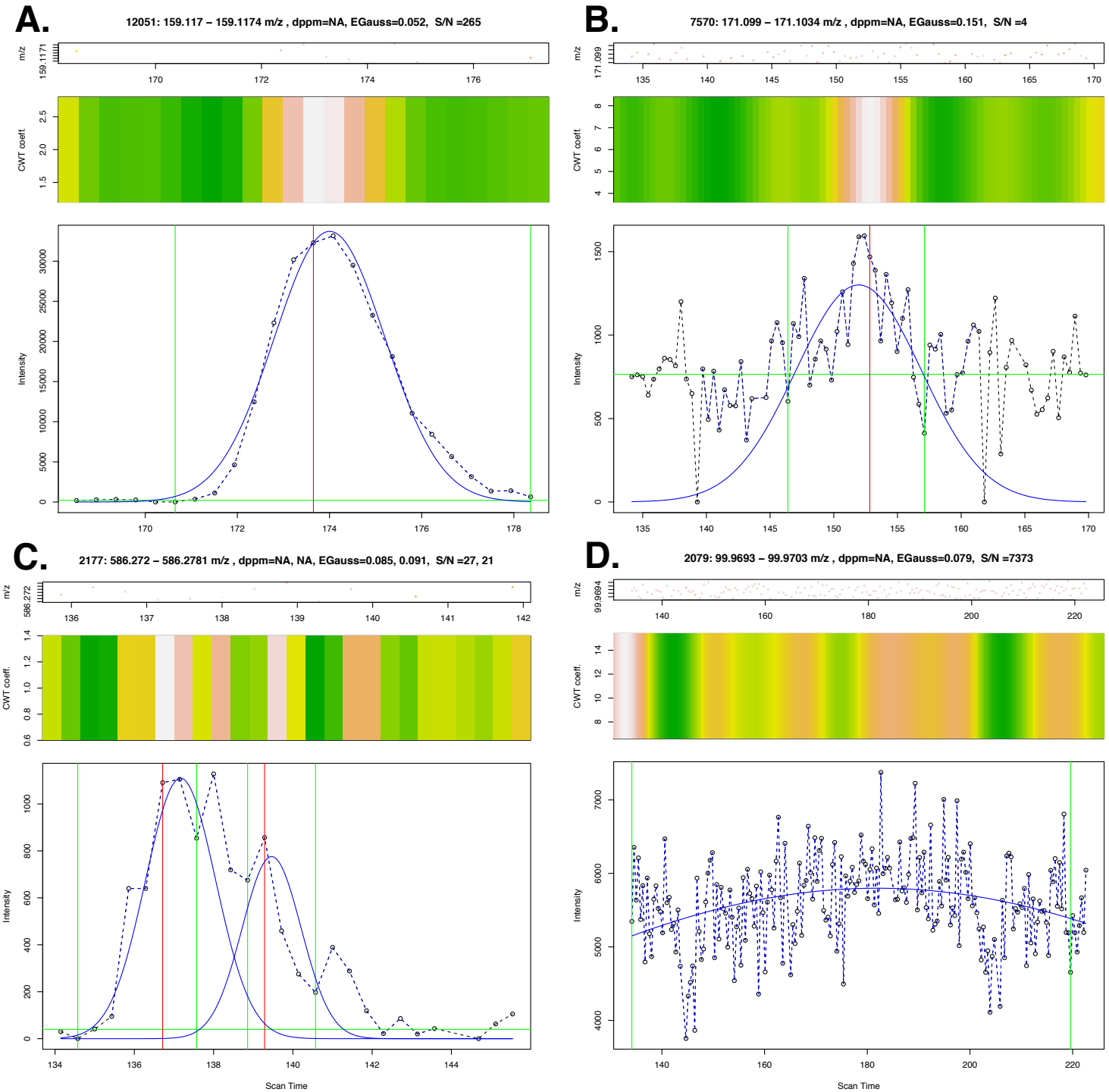
Figure 8. Normalization and removal of unwanted variability. Feature distribution of 100 random samples from PIVUS before (**panel A)** and after (**panel B**) ANOVA-type normalization. First two principal components in PIVUS before (**panel A)** and after (**panel B**) adjustment for season and storage time.

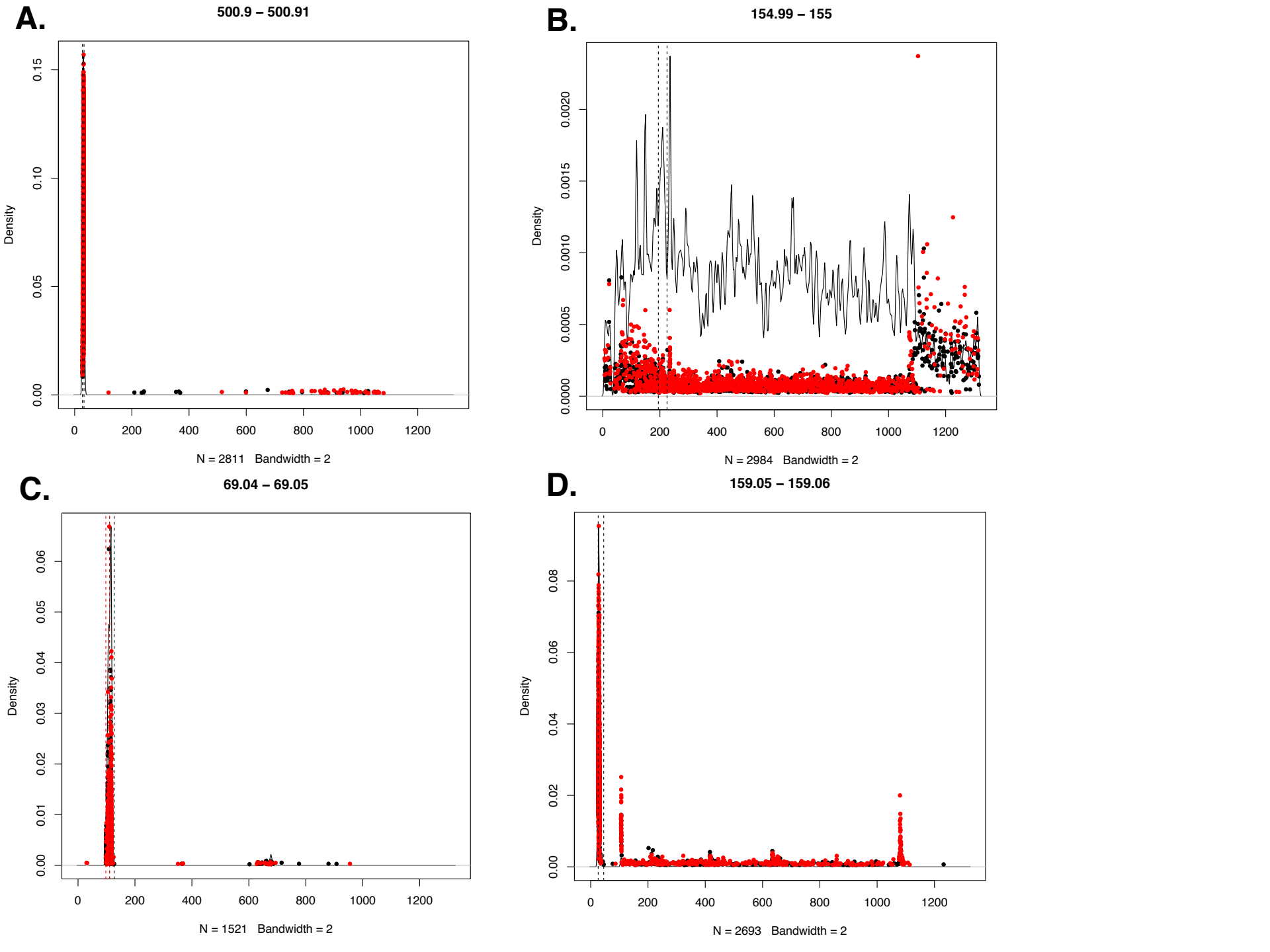
Table 1. Metabolites associated with hsCRP and annotated through spectra matching with a private compound library or public databases (approach 1, 2a or 2b)

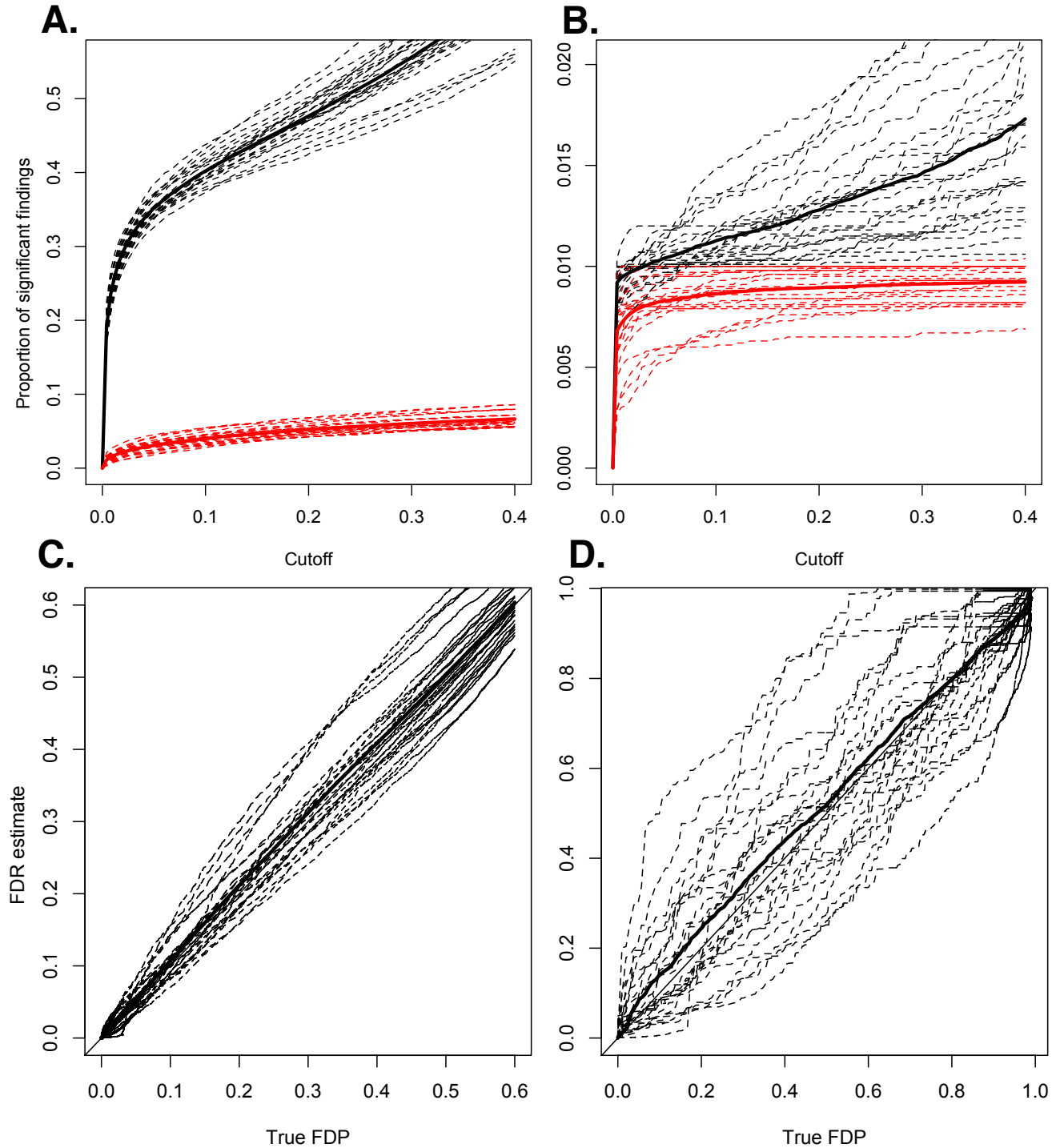
Feature ID in TwinGene	Molecular Weight	N. of Significant Features From the Same Metabolite	P-value TwinGene	P-value Pivus	Direction of Association Combined	P-value Combined	Identification and Annotation approach	Compound Subclass	Metabolite Name
M524.362T382.723	521.348	39	6.0E-04	1.5E-09	-	2.1E-11	1	Glycerophosphocholines	1-oleoyl-2-hydroxy-sn-glycero-3-phosphocholine
M221.014T112.295	179.058	6	1.2E-05	1.9E-02	-	1.5E-06	1	Alpha Amino Acids and Derivatives	Hippuric Acid
M565.520T452.724	282.256	3	2.1E-04	2.3E-03	+	1.6E-06	1	Unsaturated Fatty Acids	18:1 fatty acid
M261.144T109.900	260.137	3	1.7E-05	2.0E-07	+	1.0E-10	2a	Alpha Amino Acids and Derivatives	Gamma-glutamyl-Leucine
M296.130T119.570	312.147	1	6.9E-04	6.8E-06	+	1.2E-07	2a	Peptides	Phenylalanylphenylalanine
M190.087T152.558	189.079	3	1.1E-04	8.5E-05	-	3.5E-08	2a	Indolyl Carboxylic Acids and Derivatives	3-Indolepropionic acid
M239.092T142.656	238.084	13	1.6E-05	1.5E-04	-	1.1E-08	2a	Hydroxycinnamic Acid Derivatives	3,4,5-Trimethoxycinnamic acid
M303.196T165.056	302.188	5	5.3E-04	1.9E-02	-	3.1E-05	2a	Androgens and Derivatives	11β-Hydroxy-4-androstene-3,17-dione
M317.212T196.185	-	9	9.8E-06	5.1E-06	-	2.5E-10	2b	Prostaglandins and related compounds	

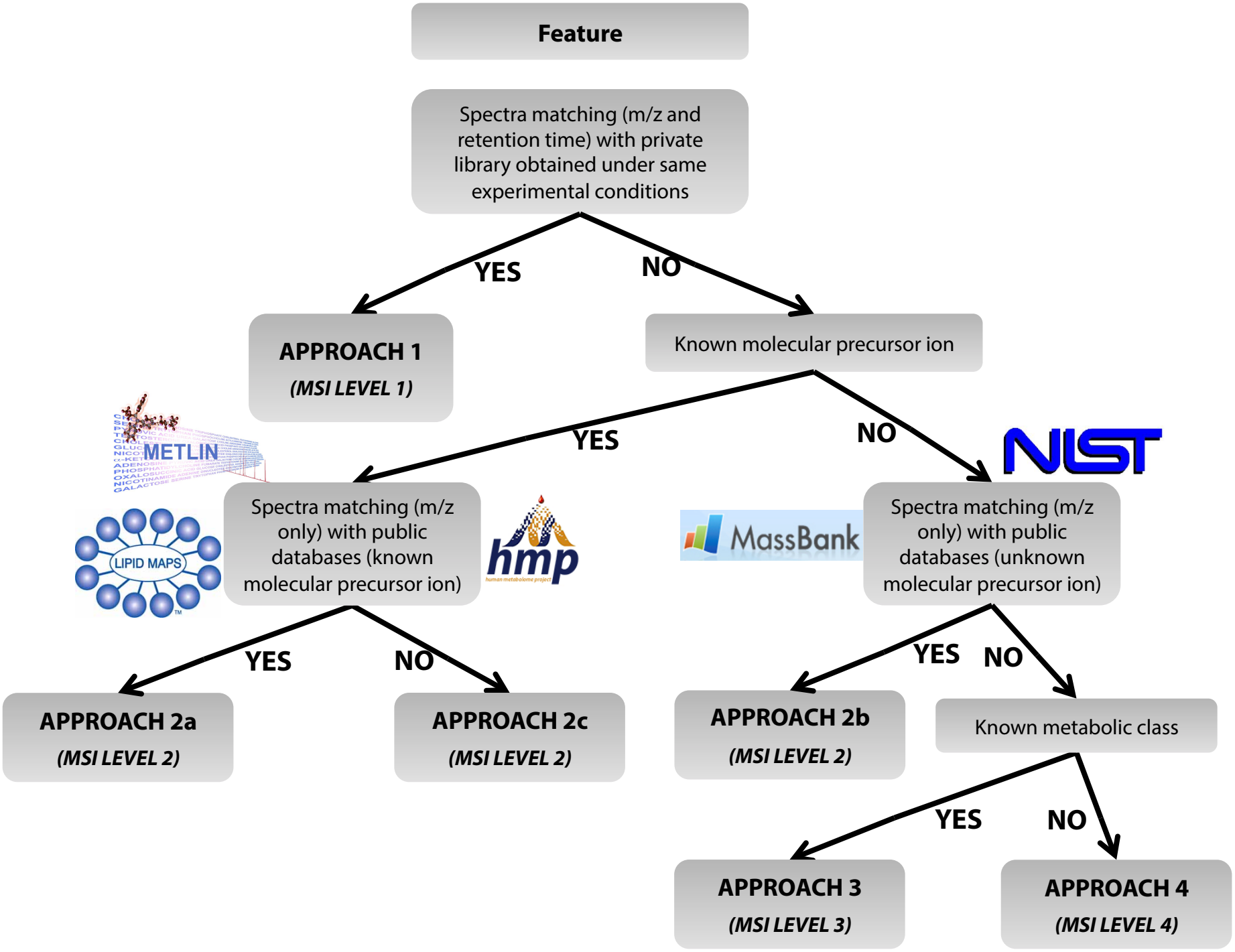

Intensity
600000
400000
200000
M/z
1200
1000
800
600
400
200
Retention time
1200
1000
800
600
400
200

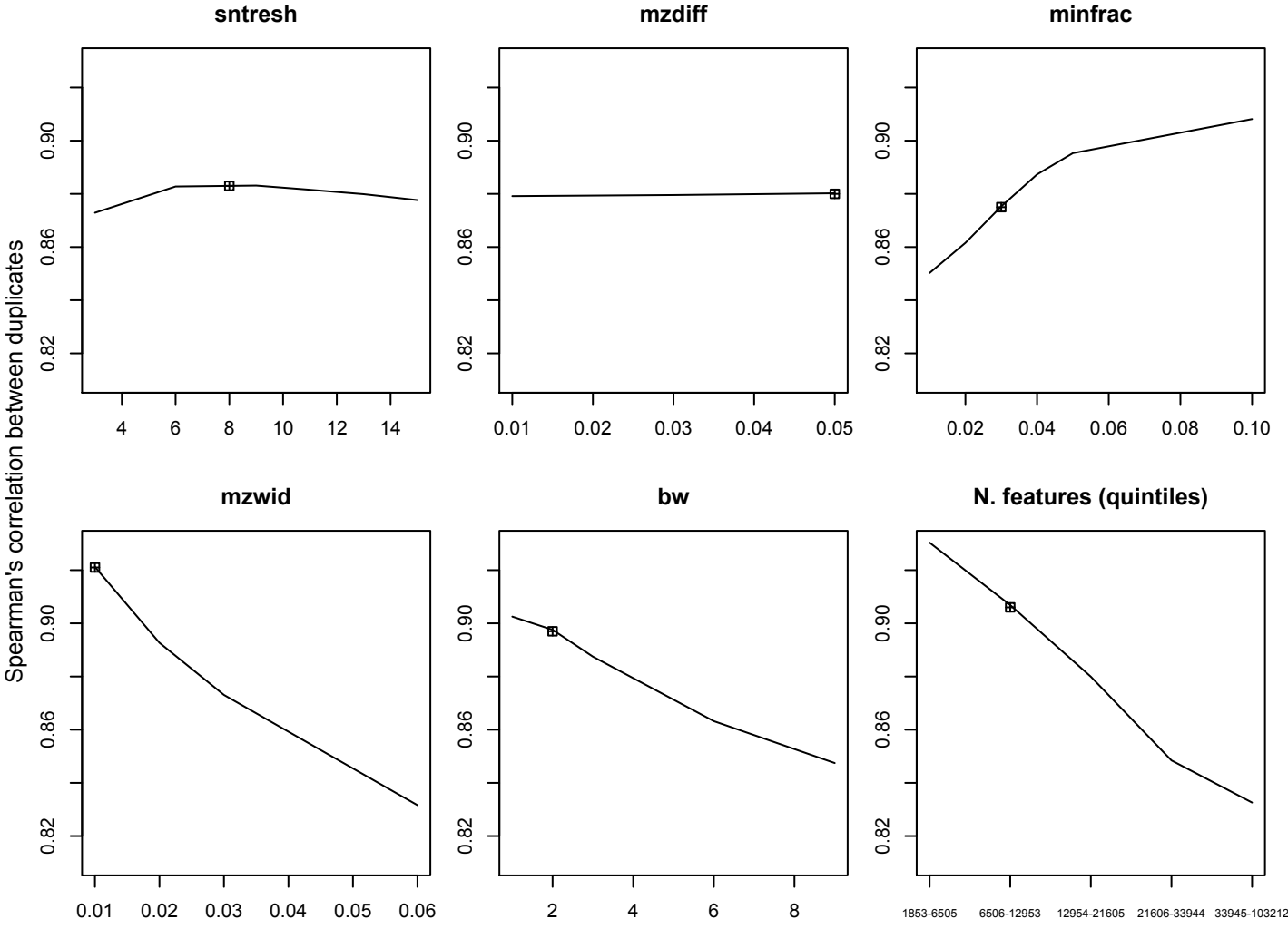

Raw Data
MODULE 1
(feature detection)
Peaks detection
Peaks alignment
Peaks grouping
Filling missing features
MODULE 2
(quality control)
Log-2 transformation
Outliers exclusion
Normalization
Average between replicates
MODULE 3
(analysis)
Univariate data analysis (UMWAS)
Replication
MODULE 4
(annotation)
Spectra generation
Annotation

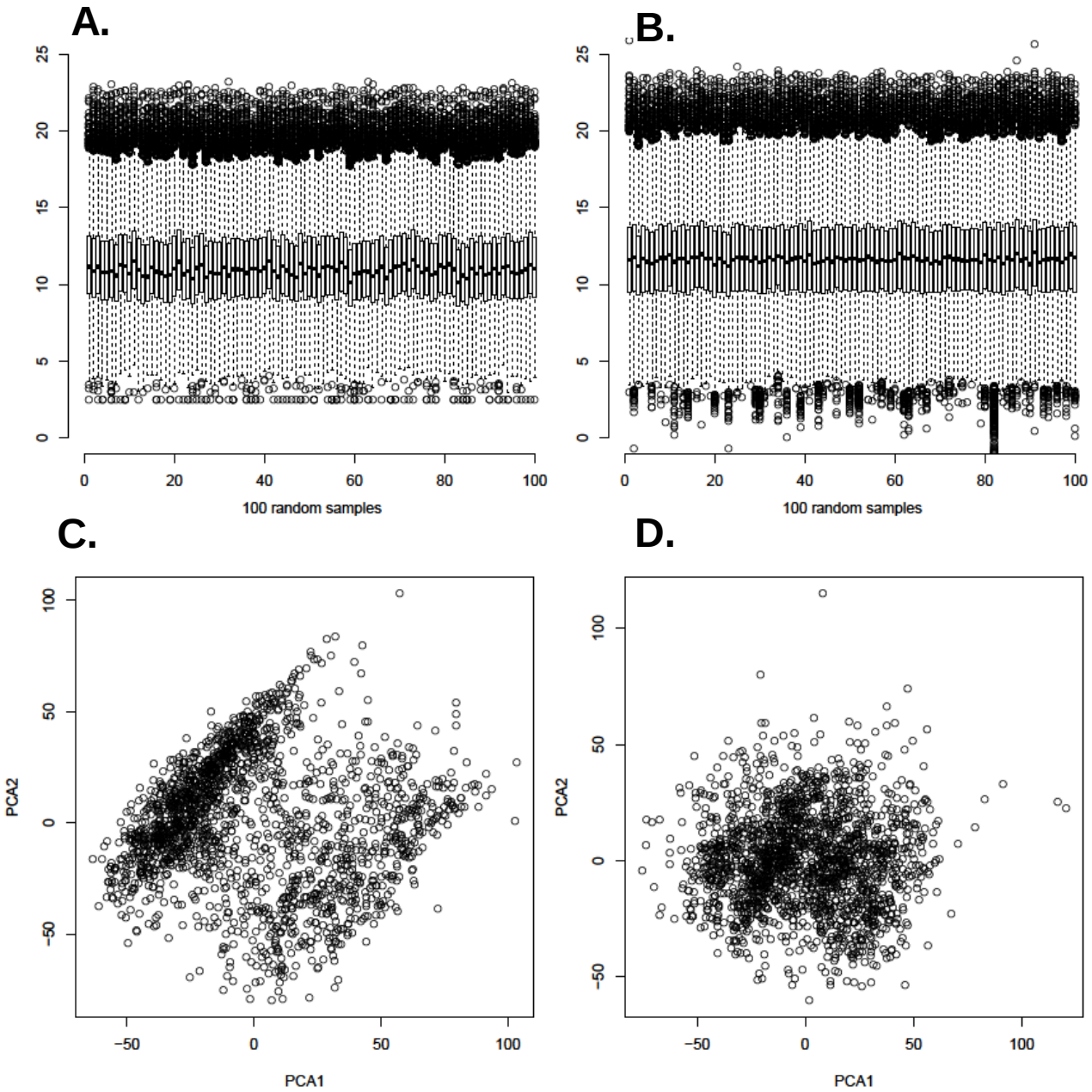

A.
12051: 159.117 – 159.1174 m/z , dppm=NA, EGauss=0.052, S/N =265
m/z
159.1171
CWT coeff.
Intensity
B.
7570: 171.099 – 171.1034 m/z , dppm=NA, EGauss=0.151, S/N =4
m/z
171.099
CWT coeff.
Intensity
C.
2177: 586.272 – 586.2781 m/z , dppm=NA, NA, EGauss=0.085, 0.091, S/N =27, 21
m/z
586.272
CWT coeff.
Intensity
Scan Time
D.
2079: 99.9693 – 99.9703 m/z , dppm=NA, EGauss=0.079, S/N =7373
m/z
99.9694
CWT coeff.
Intensity
Scan Time


A.
500.9 – 500.91
Density
0.00
0.05
0.10
0.15
0
200
400
600
800
1000
1200
N = 2811 Bandwidth = 2
B.
154.99 – 155
Density
0.0000
0.0005
0.0010
0.0015
0.0020
0
200
400
600
800
1000
1200
N = 2984 Bandwidth = 2
C.
69.04 – 69.05
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
200
400
600
800
1000
1200
N = 1521 Bandwidth = 2
D.
159.05 – 159.06
Density
0.00
0.02
0.04
0.06
0.08
0
200
400
600
800
1000
1200
N = 2693 Bandwidth = 2


A.
B.
C.
D.
Proportion of significant findings
Cutoff
Cutoff
FDR estimate
True FDP
True FDP
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.000
0.005
0.010
0.015
0.020
0.8
1.0


Feature
Spectra matching (m/z and retention time) with private library obtained under same experimental conditions
YES
NO
APPROACH 1
(MSI LEVEL 1)
Known molecular precursor ion
YES
NO
METLIN
NIST
Spectra matching (m/z only) with public databases (known molecular precursor ion)
LIPID MAPS
hmp
MassBank
Spectra matching (m/z only) with public databases (unknown molecular precursor ion)
YES
NO
YES
NO
APPROACH 2a
(MSI LEVEL 2)
APPROACH 2c
(MSI LEVEL 2)
APPROACH 2b
(MSI LEVEL 2)
Known metabolic class
YES
NO
APPROACH 3
(MSI LEVEL 3)
APPROACH 4
(MSI LEVEL 4)


sntresh
mzdiff
minfrac
mzwid
bw
N. features (quintiles)
Spearman's correlation between duplicates
0.82
0.86
0.90
4
6
8
10
12
14
0.01
0.02
0.03
0.04
0.05
0.06
0.10
2
1853-6505
6506-12953
12954-21605
21606-33944
33945-103212


A.
B.
C.
D.
100 random samples
100 random samples
PCA1
PCA2
PCA1
PCA2