Large-scale non-targeted metabolomic profiling in three human population-based studies

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

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.

Corresponding author:

Tove Fall, PhD, VMD

Department of Medical Sciences, Molecular Epidemiology, Uppsala University

Box 1115, SE-751 41 Uppsala, Sweden

Phone: +46-70-221 58 59; E-mail: tove.fall@medsci.uu.se

Abstract

Non-targeted metabolomic profiling is used to simultaneously assess a large part of the metabolome in a biological sample. Here, we describe the dataset of a large UPLC-Q-TOFMS-based metabolomic profiling effort using plasma and serum samples from participants in three Swedish population-based studies of middle-aged and older human subjects: TwinGene, ULSAM and PIVUS, including more than 3,600 participants in total. At present, more than 200 metabolites have been manually annotated in the three cohorts using an in-house library of standards and publically available spectral databases. Data available at the Metabolights repository include individual raw unprocessed data, processed data, basic demographic variables and spectra of annotated metabolites. Additional phenotypical and genetic data is available upon request to cohort steering committees.

BACKGROUND & SUMMARY

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 disease.[1-4]

A non-targeted metabolomics approach, as opposite to targeted approaches,[5] can be used 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 the preferred technologies to perform non-targeted metabolomics with high sensitivity, allowing the detection of a large number of metabolites.[6]

Recent improvements in instrumental technologies and advances in bioinformatics tools have provided the possibility to perform non-targeted metabolomics on large prospective epidemiological studies with thousands of individuals and hundreds of phenotypes measured.[7]

However, partially due to the high cost and complexity in data processing, only few epidemiological studies have undergone large-scale metabolomic profiling.[3,8] Moreover, since the majority of these analyses have been performed by private companies, both raw and processed data have not been made publically available.

We conducted one of the largest UPLC-MS-based metabolomic profiling efforts to date using plasma and serum samples from participants from three Swedish population-based studies: TwinGene, ULSAM and PIVUS, including more than 3,600 participants. Thus far, more than 200 metabolites have been manually annotated using an in-house spectral library of authentic standards and publically available spectral databases. Moreover, thousands of metabolic features, not yet annotated, have been identified across the three studies.

In addition to metabolomic profiling, information on lifestyle, anthropometrics and demographics, and measurements of established cardiovascular biomarkers are available in all participants. In two of the three studies, extensive measurements of subclinical cardiovascular disease are also available. Furthermore, all participants have been linked with national Swedish registries, allowing the identification of incident disease events for a maximum of up to 20 years of follow-up. This resource is used to perform metabolome-wide association studies (MWAS), to explore networks and pathways of metabolites, and to evaluate new methodologies for metabolite annotation. We have successfully used this data in a first analysis for association with incident coronary heart disease in 1,028 individuals (131 events) with validation in 1,670 individuals (282 events). We identified four lipid-related metabolites with evidence for clinical utility, as well as a causal role in coronary heart disease development.[9]

Our analytical strategy allowed estimation and control of instrumental variability, since each sample was analyzed in non-consecutive duplicates. Moreover, both MS and MS/MS chromatograms were collected simultaneously in the same experiment, by alternating low and elevated collision energy scans. This approach facilitated the annotation of metabolites by using correlational relationships across individuals to reconstruct both indiscriminate (id) MS and id MS/MS spectra for each feature.[10] To our knowledge, this approach has never previously been attempted in such large number of samples.

All the raw data, including MS and MS/MS spectra, as well as the processed data for the annotated metabolites have been made available on the Metabolights repository. Main demographic information and potential factors of batch effect are also accessible in the repository. Specific phenotypes are available upon request for researchers who meet the criteria for access to the confidential data. This Data Descriptor adds to the previous publication[9] by expanding the description of the data acquisition, feature identification and annotation to known metabolites the as well as a discussion on the methodological aspects of using metabolomic profiling in large datasets. Importantly, the included description of the data analysis pipeline can act as a blueprint for researchers in this rapidly increasing field and. Furthermore, it contains important notes for those that would like to use the data themselves.

Finally, GC-MS analysis of the same plasma and serum samples, as well as LC-MS analysis of urine from the same participants are ongoing and will be shared with the community as soon as the analyses are ready.

METHODS

The methods described in this section are expanded versions of descriptions in our previous work.[9]

Study populations and sample collection

TwinGene

The Swedish Twin Registry is a population-based national register including over 194,000 Swedish twins born from 1886 to 2008.[11] TwinGene is a longitudinal study within the Swedish Twin Register initiated to examine associations between genetic factors and cardiovascular disease in Swedish twins. Twins born before 1958 who participated in a telephone screening between 1998 and 2002 were re-contacted between April 2004 and December 2008. Health and medication data were collected from self-reported questionnaires, and a blood sampling kit was mailed to the subject who then contacted a local health care center for blood sampling and a health check-up. Contacts were allowed on Monday to Thursday mornings (and not the day before a national holiday), to ensure that the sample would reach the KI Biobank in Stockholm the following day by overnight mail. The participants were instructed to fast from 8 PM the previous night. A total volume of 50 ml of blood was drawn from each individual by venipuncture. Data on cardiovascular health, medication and death dates were collected through linkage with national registers. In total, 12,591 individuals (55% women) participated in the study.

The blood sampling was done as follows: First, a tube containing Ethylenediaminetetraacetic acid (EDTA) was filled and inverted 5 times immediately. These samples were used for DNA extraction. Second, three gel tubes were filled, inverted 5 times immediately, let standing for 30 minutes for coagulation in room temp, and centrifuged for 10-15 minutes at 3800 rpm. Finally, the serum was tapped from the gel tubes to a collection tube and placed in a transport cylinder. Tubes were sent to Karolinska University Laboratory by overnight post where they were frozen at -80° C up to eight years until metabolomic profiling.

Metabolomics was performed in a subsample of TwinGene. Specifically, we utilized a case-cohort design by selecting all the incident cases of coronary heart disease, type 2 diabetes, ischemic strokes and dementia up to 31st December 2010 and a sub-cohort (controls) of 1,643 individuals (43% women). 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. The subcohort was selected to avoid twins, including almost exclusively unrelated individuals. In total, 2,139 individuals underwent metabolomic profiling.

ULSAM

Men born between 1920 and 1924 in Uppsala, Sweden were invited to participate at age 50 (N=2,841) in this longitudinal cohort study, which was started in 1970 [12], and 81.7% (N=2,322) participated. Individuals have been reinvestigated at the ages of 60, 70, 77, 82 and 88 years. Information collected includes a medical questionnaire, blood pressure and anthropometric measurements, oral glucose tolerance test and 24-hour ambulatory blood pressure. Data on cardiovascular health, medication and death dates were achieved through linkage with national registers. At age 70, EDTA plasma, citrate plasma, serum and whole blood for DNA extraction were collected from fasting participants and stored at -70° C for up to 20 years until analysis. Additional EDTA plasma was collected during the oral glucose tolerance test in ~600 participants. EDTA plasma samples from the age 70 baseline were used for the metabolomic profiling. In total, 1,138 individuals underwent metabolomic profiling.

PIVUS

PIVUS is a community-based study where all men and women at age 70 living in Uppsala, Sweden in 2001-2005 were invited to participate[13]. The 1,016 participants (50% women) have been extensively phenotyped including measurements of endothelial function and arterial compliance, cardiac function and structure by ultrasound and magnetic resonance imaging, evaluation of atherosclerosis by ultrasound and magnetic resonance imaging, seven-day food intake records, electrocardiogram (ECG) analysis, cardiovascular autonomic function and body composition by dual-energy X-ray absorptiometry. Cardiovascular events have been validated through review of hospital records. Blood samples were drawn between 8

and 10 AM after an overnight fast. Blood taken in EDTA tubes was centrifuged, and plasma was aliquoted and frozen within one hour. Serum was aliquoted and frozen within two hours. These samples were stored at -70° C for up to 11 years before metabolomic profiling that was performed in serum samples from 968 individuals.

Ethics statement

All participants gave informed written consent and the Ethics Committees of Karolinska Institutet or Uppsala University approved the respective study protocol.

Sample preparation

Serum or plasma samples were thawed and 100 µL of serum or plasma were transferred to 400 µL methanol in 96-well format to precipitate proteins (**Figure 1a**). This 80% methanol solution was stored at -20°C overnight, and then centrifuged for 30 minutes at 3800g in 4°C to pellet precipitated protein. The supernatant was aliquoted to separate 96-well plates, sealed using a heat-seal foil, and stored at -20°C until analysis. Plate analysis order was completely randomized for each set of injections, and plates were run in batches of two plates, reflecting the autosampler capacity. Within each set of two plates, the run order was again completely randomized to prevent injection order artifacts. Duplicate injections were performed for all samples, with the second set of injections performed upon completion of the first set of injections for all samples. Each set of injections was randomized independently with respect to plate order, and injection order within a plate.

UPLC-QTOFMS data acquisition

Prior to each batch of two plates of samples, instrument maintenance (cone cleaning, mass calibration, and detector gain calibration) was performed, and a quality control (QC) standard mix was injected. Two conditioning and three QC injections were performed, with 1 µL injections of a 20% methanol solution containing 2 µg/mL each of caffeine, terfenadine, sulfadimethoxime, and reserpine. The QC standards were evaluated for retention time (+/- 0.05 minutes), signal intensity (<25% relative standard deviation), and mass accuracy (< 3 ppm). This approach is designed to prevent acquisition of low-quality data. The QC steps post-acquisition described in detail below were then used to remove poor injections/samples.

One µL of protein-precipitated serum or EDTA plasma was injected on a Waters Acquity UPLC system. Separation was performed using a Waters Acquity UPLC BEH C8 column (1.8 µM, 1.0 x 100 mm), using a gradient from solvent A (95% water, 5% methanol, 0.1% formic acid) to solvent B (95% methanol, 5% water, 0.1% formic acid). Injections were made in 100% A, which was held for 0.1 min, ramped to 40% B in 0.9 minutes, to 70% B over two minutes, and to 100% B over 8 minutes. The mobile phase was held at 100% B for 6 minutes, returned to starting conditions over 0.1 minute, and allowed to re-equilibrate for 5.9 minutes. Flow rate was constant

at 140 µL/min for the duration of the run. The column was held at 50°C, while samples were held at 10°C.

Column eluent was infused into a Waters Xevo G2 QTOF MS fitted with an electrospray source. Data was collected in positive ion mode, scanning from 50-1200 m/z at a rate of 5 scans per second. Scans were collected alternatively in MS mode at collision energy of 6 V, and in idMS/MS mode using elevated collision energy (15–30 V). IdMS/MS (also called MS^E) allows for an unbiased examination of both precursor and fragment ion mass spectra without additional experiments.[10]

Calibration was performed prior to every batch of 96 samples via infusion of sodium formate solution, with mass accuracy within 1 ppm. The capillary voltage was held at 2200V, the source temp at 150°C, and the desolvation temperature at 350°C at a nitrogen desolvation gas flow rate of 800 L/hr. The quadrupole was held at collision energy of 6 volts. Raw data files were converted to .cdf format using Waters DataBridge software for processing (**Figure 1a**).

Metabolic feature identification

We used the XCMS software[14] to perform peak identification, alignment, grouping and filling **(Figures 1b, 1c, 1d)**. This software is implemented in R. An example of the code used for the data processing is available at: https://github.com/andgan/metabolomics_pipeline.

Peak detection

We performed peak detection in each chromatogram using the *centWave* algorithm[15] implemented in the *xcmsSet* function **(Figure 1b)**. We noted that it is important to determine two instrument-dependent parameters: (1) *ppm*, indicating the mass spectrometer accuracy; and (2) *peakwidth*, indicating the chromatographic peak-width range. The former parameter was set as 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), as previously suggested [15]. The latter parameter was set by inspecting the peak-width in several chromatograms. To decide the values of remaining parameters in the *xcmsSet* function, we used an approach based on iterative testing of different settings as discussed in the **Technical validation** section. We also evaluated the quality of the

algorithm performance by looking at the plot obtained, for one representative chromatogram, with the *findPeaks* function.

The parameters used for peak detection in the three studies were the following:

- TwinGene: *method="centWave", ppm=25, peakwidth=c(2:15), snthresh=8, mzCenterFun="wMean", integrate=2, mzdiff=0.05, prefilter=c(1,5);*
- ULSAM: *method="centWave", ppm=25, peakwidth=c(2:15), snthresh=6, mzCenterFun="wMean", integrate=2, mzdiff=0.05, prefilter=c(1,5);*
- PIVUS: *method="centWave", ppm=25, peakwidth=c(2:15), snthresh=8, mzCenterFun="wMean", integrate=2, mzdiff=0.05, prefilter=c(1,5).*

Peak 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 retention time drift. We used the *retcorr* function to re-align the samples by correcting the retention time shifting. We used the *obiwarp* algorithm[16] (implemented in the *retcorr* function), as it is more stable for large number of samples than the *loess* algorithm, which is the default method in XCMS **(Figure 1c)**. This algorithm uses the sample with the largest number of peaks as reference for alignment. Visualization of the extent of correction of retention time for each peak across samples can be obtained from the *retcorr* function and can be informative of batch effects or laboratory issues. The parameters used for peak alignment in the three studies were the following:

- TwinGene, ULSAM and PIVUS: *method="obiwarp",plottype="deviation".*

Peak grouping

We grouped the aligned peaks across samples using the *group* function. Each group is called 'metabolic feature' or simply 'feature' **(Figure 1d)**. Three main parameters needed to be determined: (1) *bw*, the retention time deviation to be allowed for grouping; (2) *mzwid*, m/z width to determine the peak grouping across samples; (3) *minfrac*, minimum fraction of samples in each group needed to call it as a valid feature. Simulation approaches based on iterative testing of different settings (see the **Technical validation** section) and the plot obtained from the *group* function were

used to determine the right values for these three parameters. We kept the *minfrac* parameter relatively low (below 6%) to allow relatively rare and exogenous (e.g. cotinine, a metabolite of nicotine) metabolites to be included among the features brought forward. However, the disadvantage of using a too low *minfrac* parameter is an increased risk of detecting false-positive features or noise. The parameters used for peak grouping in the three studies were the following:

- TwinGene: *bw=2, minfrac=0.03, max=100, mzwid=0.01;*
- ULSAM: *bw=3, minfrac=0.05, max=100, mzwid=0.01;*
- PIVUS: *bw=2, minfrac=0.05, max=100, mzwid=0.01*.

Filling missing features

The fact that metabolic features are not detected in all samples in a cohort might be due to a true lack of signals for certain samples (for example, cotinine should only be detectable in 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, we used the *fillpeak* function to impute missing intensities for each metabolic feature. The *fillpeak* function uses the raw data, following retention time correction, to fill the missing intensity values. Notice that, if a feature has truly not been detected, the algorithm will assign a value close to the signal-to-noise threshold. Default parameters were used in all three studies.

Log_2 transformation, quality control, normalization

Our data acquisition workflow involved the simultaneous collection of low and high collision energy in alternating scans.[10,17] Thus, for every injection two corresponding data files were generated, rendering four files per individual sample. To perform peak identification, alignment, grouping and filling, we jointly processed MS (low collision energy) and MS/MS (high collision energy) chromatograms. This was done to ensure common features names in both datasets. However, in the following steps only the MS data were used, while the MS/MS data were used for metabolite annotation (see further below).

First, feature intensities were transformed to the $\log_2$ base scale to approximate normal distribution. Second, potential sample outliers were identified by plotting the total intensity for each sample. Samples exhibiting extreme total intensities might indicate sample degradation or technical errors, and such samples were excluded. Third, an ANOVA-type normalization was used to take into account factors of unwanted variation[18,19] **(Figure 1e)**. In the **Technical validation** section, we show that this normalization performed better than other commonly used normalization approaches. The normalization was done by fitting a linear regression for association between each feature intensity and several factors of unwanted variability. We then used the residuals from the regression as new feature intensities. The factors of unwanted variability can be identified by studying the association between several technical variables and the first principal components. In each study, we adjusted for the following factors:

- TwinGene: *retention time correction, analysis date, storage time, unknown cluster effect;*
- ULSAM: *retention time correction, analysis date, sample collection, plate effect;*
- PIVUS: *retention time correction, analysis date, storage time, season effect.*

Finally, feature intensities were averaged between technical duplicates to reduce the inherent instrumental variability. Features with poor correlation between duplicates were excluded. The correlation threshold for feature exclusion is study dependent and was chosen so that the correlation between technical duplicates was significant with P-value < 0.05 after adjusting for multiple testing, using a Bonferroni correction.

In total, we detected 9,755, 10,162 and 7,522 metabolic features in TwinGene, ULSAM and PIVUS, respectively (**Table 1**).

Metabolite annotation

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).[17] We have previously shown how to use the correlational relationships in the data to assign the correct precursor ion and reconstruct both idMS and idMS/MS spectra for each feature.[20] R code for generating idMS and idMS/MS spectra is provided at https://github.com/andgan/metabolomics_pipeline.

Once both idMS and idMS/MS spectra are reconstructed, we used several approaches for the annotation of metabolic features to metabolites, reflecting the different levels of confidence suggested by the Metabolomics Standard Initiative [MSI] [21] **(Figure 1f)**:

- The first approach (MSI level 1) had the highest confidence and was based on matching accurate mass, fragmentation pattern, and retention time with the in-house spectral library of authentic standards collected under the same experimental conditions.
- The second approach (MSI level 2) was based on spectrum and/or m/z similarities, and a reasonable retention time given the chemical properties, and their annotation relied on information available in public databases.
- The third approach (MSI level 3) used a combination of spectral data and accurate mass to assign the metabolite to a chemical class (without knowing the exact origin of the metabolite), and their annotation to a specific class relied on information available in public databases.
- Finally, when all the other approaches to annotation of the metabolite or metabolite class failed, the metabolite was annotated as “unknown” (MSI level 4).

In total, we detected 109 metabolites with the first approach, 102 with the second approach, and could assign the chemical class to 18 metabolites with the third approach. In the Metabolights repository, we reported the names of metabolites with level 1 and level 2 annotation, and the number of annotated metabolites per cohort is summarized in **Table 1**.

DATA RECORDS

Three data records, one for each study, have been deposited in Metabolights with accession numbers MTBLS93 (TwinGene), MTBLS124 (ULSAM) and MTBLS90 (PIVUS). Each data record contains five sections:

1. Study design description: This section contains general information about the study characteristics and related publications;
2. Protocols: This section contains a detailed description of the protocol used to collect and process the data;
3. Samples: This section reports the anonymized identification code for each participant, as well as main demographic information (e.g. sex and age);
4. Assay: In this section, all the assays (chromatograms) are reported and linked to the anonymized participant identification code. Each participant should have four assays: MS[repl. 1 and 2] and MS/MS[repl. 1 and 2] (here named MSn).We also report factors of unwanted variability (e.g. batch effects) that have been controlled for during normalization. Note that extended information about the assay characteristics can be downloaded in a .txt format.
5. Study files: In this section, all uploaded assays (chromatograms) can be downloaded in .cdf.gz format. The assays follow this filename format: '*analysis date*'_'*study name*'_'*increasing number01/02*'. MS files filename end with *01 and MS/MS files, collected on higher collision energy, end with *02. The section also contains a file called *m_'study name'_metabolite_profiling_mass_spectrometry.tsv* including all metabolites that have been identified in the corresponding study. The annotated metabolites are linked to different databases to retrieve the chemical formula and additional information, when available. The retention time and m/z values reported correspond to those observed in our data. It also includes the intensities for each metabolite or metabolic feature across the study participants.

Additional information not shown in Metabolights can be obtained by downloading the entire study or the study metadata. This file can then be open with ISAcreator for editing.

TECHNICAL VALIDATION

Description of technical variation

In **Table 1**, we report the main summary statistics about the number of chromatograms, peaks and metabolic features detected in each study. We detected less metabolic features in PIVUS, presumably mostly due to the smaller sample size. We further report measures describing the technical variability of our method. These measures were possible to obtain because each sample has been analysed in non-consecutive duplicates. All three studies were comparable in terms of mean feature correlation and mean coefficient of variation across features; the latter varying from 2.9% in PIVUS to 5.2% in ULSAM. We also report the coefficients of variation obtained from a mixture of four known standards analysed between every analytical batch (192 samples/batch). The mix was analyzed in triplicates, and the relative standard deviation (RSD) of the intensity of the M+H signal for each compound was calculated as the SD divided by the mean and expressed as percent. The mean (range) RSDs over 12 random controls were 2.9% (0.5-8.0) for caffeine, 3.9% (0.8-8.4) for sulfadime, 4.0% (0.7-9.7) for reserpine and 4.0% (0.5-10.2) for terfenadine.

Determination of optimal XCMS parameters

The parameters used in XCMS to detect, align and group peaks can drastically change the number and quality of the identified features. This aspect is often underappreciated in metabolomics studies and only few prior papers have touched upon this topic. Brodsky and colleagues have shown how the parameter tuning affects the inter-replicate correlation and which parameters are likely to have the largest influence.[22] The authors of XCMS have suggested parameter values for different UPLC/MS instruments; both in a published paper[23] and in the online version of XCMS.[24] However, the settings are both instrumental and study-dependent, and need to be fine-tuned independently for each study.

For each study, we randomly selected thirty participants (thus having 120 chromatograms; including replicate 1 and 2, and both MS and MS/MS) and ran several parameter configurations within a reasonable interval around the suggested values. Specifically, we tried 2,161 combinations of different levels for the following parameters: *snthresh*, *mzdiff*, *minfrac*, *mzwid* and *bw*. The parameter configuration that

maximizes the feature correlation between technical replicates is likely to be the best choice. In **Figure 2a**, the average feature correlation between technical replicates is reported separately for each parameter suggesting that a configuration with relatively higher *snthresh* and *minfrac*, and lower *mzwid* and *bw* improved the correlation. The *mzdiff* parameter did not influence the correlation. Similar to our observations, Brodsky and colleagues[22] also reported that an increase in the *minfrac* and decrease in the *mzwid* parameters were associated with higher feature correlation between technical replicates. To maximize the number of detected metabolic features while maintaining high 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.

Comparison of different normalization approaches

We determined which normalization approach worked better in our data by comparing the coefficient of variation and feature correlation between technical replicates (**Table 2**). For simplicity, we performed these simulations using PIVUS data since this was the study with the smallest sample size. The ANOVA-type approach clearly outperformed other normalization methods based on single parameter scaling. This can be explained by the ability of the method to adjust for several factors of unwanted variation, such as analysis date or storage time. After applying the best normalization approach, we plotted the correlation between technical replicates for each feature and compared it with what was expected under the null (**Figure 2b**). The null distribution was obtained by permuting each feature 200 times. Almost all features showed a stronger correlation than expected by chance. Nevertheless, some features had low correlation between technical replicates. Those features with a correlation coefficient of <0.3 had generally lower intensity than those with a correlation coefficient ≥ 0.7 (mean feature intensity 10.7 and 12.0 respectively, $\text{P-value}_{\text{diff}} < 0.0001$).

Finally, we performed a visual inspection of the metabolic feature distributions across samples in all three studies (**Figure 2c**). All three studies were well normalized, although ULSAM had higher variability than the other two studies, probably due to longer storage time (up to 20 years).

Targeted MS/MS analysis to determine the quality of the metabolite annotations

After the entire metabolomic profiling workflow and subsequent metabolite annotations were performed, we evaluated the quality and robustness of our untargeted approach by performing a targeted and confirmatory analysis for a random subset of metabolites which we had annotated with level 1 (n=2) or 2 (n=2). We confirmed all four of these metabolites based on matching accurate mass, fragmentation pattern, and retention time. The targeted mass analysis of the selected metabolites (acetylcarnitine, cortisol, arachidonic acid and docosahexaenoic acid) was performed in the sample with the highest concentration of that metabolite using the multiple reaction monitoring (MRM) mode by monitoring the transitions between the protonated molecular ion ($[M+H]^+$) including molecular ion adducts ($[M+Na]^+$, $[M+H-H_2O]^+$, $[M+K]^+$) and its product ions (fragments). Although limited to a small subset of the annotated metabolites, these results support the high quality of the workflow and annotation. Further, we have previously validated the level 2 annotation of four metabolites associated with coronary heart disease by targeted mass analysis (LysoPC 18:1, LysoPC 18:2 MG 18:2 and SM 28:1).[9]

USAGE NOTES

Metabolome-wide association studies

Non-targeted metabolomics can be performed in epidemiological studies to identify new biomarkers of disease. Similar to genome-wide association studies (GWAS) in the field of genomics, it is possible to conduct a non-targeted metabolome-wide association study (MWAS), using metabolites as independent variables (instead of genetic variants as in a GWAS) and a disease or other trait of interest as dependent variable. 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 and sex; but for specific outcomes, additional biological confounders can be included, depending on the research question. Features with low correlation between technical replicates or those observed in only few samples previous to peak filling should be excluded.

Given the large number of statistical tests performed, correction for multiple testing needs to be considered. False discovery rate (FDR) can be used to control the number of false positives among the metabolites declared significant. This strategy works better in situations where a large number of discoveries are expected, which is often the case with metabolomics data, especially when investigating cardiovascular or metabolic traits. Validation in a separate study is highly recommended due to the inherent technical and biological variability of this type of data. Given the differences in age range, blood collection methods and blood partition (serum and plasma); TwinGene, ULSAM and PIVUS provide an excellent opportunity for performing validation studies. Moreover, we have recently introduced a methodology to determine the expected proportion of findings that can be validated in an external study (called rediscovery rate) and the proportion of these findings that are expected to be false positives.[25]

Genetic information can also be integrated to suggest potential biological mechanisms and, importantly, to determine if the metabolite might be causative of the outcome of interest. This strategy, known as Mendelian randomization,[26] uses genetic variants, called instrumental variables, to disentangle the confounded causal relationships

between intermediate phenotypes (e.g. metabolites) and disease. As an illustration of the above procedures, we have recently performed a MWAS of coronary heart disease and identified four new lipid-related metabolites. In this work, we further integrated genetic information to indicate if the associations were likely to be causal.[9]

New methods to match features across studies

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 differences in bioinformatic data processing (e.g. the *minfrac* parameter depends on the number of samples as more features are detected when a 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. ±0.02 m/z differences). The retention time matching is more challenging and depends on the retention time correction applied during peak alignment. To our knowledge, there is no established strategy to integrate this alignment information to improve the matching of retention time across studies. Further research on this topic is encouraged and can be done using the reported data downloaded from the Metabolights repository.

Additional uses

A unique characteristic of our samples is the simultaneous collection of MS and MS/MS data. However, few papers have explored how to use this technology to improve metabolite annotation. Our resource represents a unique opportunity to identify new ways to use MS/MS data to improve annotation and analysis.

Given the large sample size, it is possible to use annotated metabolites to perform correlation-based network analysis. This analysis can be further integrated with biological information from for example KEGG or Recon X[27] to confirm correlations and suggest new potential biological pathways. Methods to integrate biological information with observed data correlation structures can also be developed using this data.

Sharing of individual level phenotypic data

Phenotypes from the TwinGene, ULSAM and PIVUS studies, other than age and sex, are not shared in Metabolights due to restrictions in the ethical permits. Nevertheless, data can be made available upon request for researchers who meet the criteria for access to the confidential data. Data from the TwinGene study are available from the Swedish Twin Registry steering committee (http://ki.se/en/research/the-swedish-twin-registry-1; contact: Patrik.Magnusson@ki.se). Data from the ULSAM study are available from the ULSAM steering committee (http://www2.pubcare.uu.se/ULSAM/res/proposal.htm; contact: vilmantas.giedraitis@pubcare.uu.se). Data from the PIVUS study are available from the PIVUS steering committee (http://www.medsci.uu.se/pivus/; contact: lars.lind@medsci.uu.se).

ACKNOWLEDGEMENTS

We thank Dr. Alexandra Jauhiainen for helpful insights and comments. Further, we want to extend our thanks to all participants of the TwinGene, ULSAM and PIVUS studies for the kind contribution to science. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under project b2011036.

FUNDING

This study was supported by grants from Knut och Alice Wallenberg Foundation (Wallenberg Academy Fellow), European Research Council (ERC-2013-StG; grant no. 335395), Swedish Diabetes Foundation (grant no. 2013-024), Swedish Heart-Lung Foundation (grant no. 20120197), the Family Ernfors Fund, the Swedish Government's strategic research area EXODIAB (Excellence of Diabetes Research in Sweden), and Swedish Research Council (grant no. 2012-1397). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Concept and design: Andrea Ganna, Tove Fall, Erik Ingelsson
Statistical methods: Andrea Ganna, Woojoo Lee, Sara Hägg, Jitender Kumar, Markus Stenemo, Yudi Pawitan
Mass spectrometry protocol and analysis: Corey D. Broeckling, Jessica Prenni
Annotation protocol: Samira Salihovic
Cohort design: Lars Lind, Patrik K.E. Magnusson, Erik Ingelsson
Manuscript revision: All authors

COMPETING FINANCIAL INTERESTS

All authors declared no conflict of interest.

SAMPLES, SUBJECTS, AND DATA OUTPUTS

We uploaded this information as ISA-Tab metadata format to the Metabolights repository.

REFERENCES

1 Floegel, A. *et al.* Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes* **62**, 639-648, doi:10.2337/db12-0495 (2013).

2 Stegemann, C. *et al.* Lipidomics profiling and risk of cardiovascular disease in the prospective population-based Bruneck study. *Circulation* **129**, 1821-1831, doi:10.1161/circulationaha.113.002500 (2014).

3 Wang-Sattler, R. *et al.* Novel biomarkers for pre-diabetes identified by metabolomics. *Molecular systems biology* **8**, 615, doi:10.1038/msb.2012.43 (2012).

4 Wang, T. J. *et al.* Metabolite profiles and the risk of developing diabetes. *Nature medicine* **17**, 448-453, doi:10.1038/nm.2307 (2011).

5 Dudley, E., Yousef, M., Wang, Y. & Griffiths, W. J. Targeted metabolomics and mass spectrometry. *Advances in protein chemistry and structural biology* **80**, 45-83, doi:10.1016/B978-0-12-381264-3.00002-3 (2010).

6 Buscher, J. M., Czernik, D., Ewald, J. C., Sauer, U. & Zamboni, N. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Anal Chem* **81**, 2135-2143, doi:10.1021/ac8022857 (2009).

7 Tzoulaki, I., Ebbels, T. M., Valdes, A., Elliott, P. & Ioannidis, J. P. Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. *Am J Epidemiol* **180**, 129-139, doi:10.1093/aje/kwu143 (2014).

8 Menni, C. *et al.* Metabolomic markers reveal novel pathways of ageing and early development in human populations. *Int J Epidemiol* **42**, 1111-1119, doi:10.1093/ije/dyt094 (2013).

9 Ganna, A. *et al.* Large-scale metabolomic profiling identifies novel biomarkers for incident coronary heart disease. *PLoS Genet* **10**, e1004801, doi:10.1371/journal.pgen.1004801 (2014).

10 Broeckling, C. D., Heuberger, A. L., Prince, J. A., Ingelsson, E. & Prenni, J. E. Assigning precursor-product ion relationships in indiscriminant MS/MS data from non-targeted metabolite profiling studies. *Metabolomics : Official journal of the Metabolomic Society* **9**, 33-43, doi:DOI 10.1007/s11306-012-0426-4 (2013).

11 Magnusson, P. K. *et al.* The Swedish Twin Registry: establishment of a biobank and other recent developments. *Twin Res Hum Genet* **16**, 317-329, doi:10.1017/thg.2012.104 (2013).

12 Ingelsson, E., Sundstrom, J., Arnlov, J., Zethelius, B. & Lind, L. Insulin resistance and risk of congestive heart failure. *JAMA* **294**, 334-341, doi:294/3/334 [pii] 10.1001/jama.294.3.334 (2005).

13 Lind, L., Fors, N., Hall, J., Marttala, K. & Stenborg, A. 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, doi:10.1161/01.ATV.0000184769.22061.da (2005).

14 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* **78**, 779-787, doi:10.1021/ac051437y (2006).

15 Tautenhahn, R., Bottcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9**, 504, doi:10.1186/1471-2105-9-504 (2008).

16 Prince, J. T. & Marcotte, E. M. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal Chem* **78**, 6140-6152, doi:10.1021/ac0605344 (2006).

17 Plumb, R. S. *et al.* UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid communications in mass spectrometry : RCM* **20**, 1989-1994, doi:10.1002/rcm.2550 (2006).

18 Kerr, M. K. & Churchill, G. A. Statistical design and the analysis of gene expression microarray data. *Genetical research* **77**, 123-128 (2001).

19 Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics* **11**, 733-739, doi:10.1038/nrg2825 (2010).

20 Broeckling, C., Heuberger, A., Prince, J., Ingelsson, E. & Prenni, J. Assigning precursor–product ion relationships in indiscriminant MS/MS data from non-targeted metabolite profiling studies. *Metabolomics* **9**, 33-43, doi:10.1007/s11306-012-0426-4 (2013).

21 Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics : Official journal of the Metabolomic Society* **3**, 211-221, doi:10.1007/s11306-007-0082-2 (2007).

22 Brodsky, L., Moussaieff, A., Shahaf, N., Aharoni, A. & Rogachev, I. Evaluation of peak picking quality in LC-MS metabolomics data. *Analytical chemistry* **82**, 9177-9187, doi:10.1021/ac101216e (2010).

23 Patti, G. J., Tautenhahn, R. & Siuzdak, G. Meta-analysis of untargeted metabolomic data from multiple profiling experiments. *Nature protocols* **7**, 508-516, doi:10.1038/nprot.2011.454 (2012).

24 Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical chemistry* **84**, 5035-5039, doi:10.1021/ac300698c (2012).

25 Ganna, A., Lee, D., Ingelsson, E. & Pawitan, Y. Rediscovery rate estimation for assessing the validation of significant findings in high-throughput studies. *Briefings in bioinformatics*, doi:10.1093/bib/bbu033 (2014).

26 Sheehan, N. A., Didelez, V., Burton, P. R. & Tobin, M. D. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Med* **5**, e177, doi:10.1371/journal.pmed.0050177 (2008).

27 Thiele, I. *et al.* A community-driven global reconstruction of human metabolism. *Nature biotechnology* **31**, 419-425, doi:10.1038/nbt.2488 (2013).

Table 1.Summary statistics of the main characteristics of the three studies

	Twin Gene	ULSAM	PIVUS
Number of individuals included in the final dataset	2,139	1,138	968
No. of total ion chromatograms *	8,126	4,776	3,846
Average no. of peaks detected per chromatogram	10,295	5,989	8,205
No. of metabolic features before exclusion	10,996	11,028	8,185
No. of metabolic features after exclusion **	9,753	10,160	7,520
Mean feature correlation between technical replicates ***	0.38	0.46	0.43
Mean coefficient of variation (%) between technical replicates	3.70%	5.20%	2.91%
N. of annotated metabolites	202	177	189

* Including MS, MS/MS and technical replicates
** Exclusion due to low correlation between technical replicates
*** Spearman correlation coefficient for one metabolic feature between all first and second replicate samples. The reported value is the average across all the metabolics features.

Table 2. Comparison of several normalization approaches in PIVUS

Normalization Method	Mean feature correlation between technical replicates *	Mean coefficient of variation (%) between technical replicates
Raw data	0.39	4.18
Median	0.42	4.67
Median within analysis date **	0.42	4.67
Quantile	0.40	3.79
ANOVA-type, adjusting for:		
analysis date	0.43	2.97
analysis date, season	0.43	2.89
analysis date, season, storage time	0.43	3.06
analysis date, season, storage time, amount of retention time correction	0.43	2.91

* Spearman correlation coefficient for one metabolic feature between all first and second replicate samples. The reported value is the average across all the metabolics features.
** Median normalization is performed within each specific analysis date.

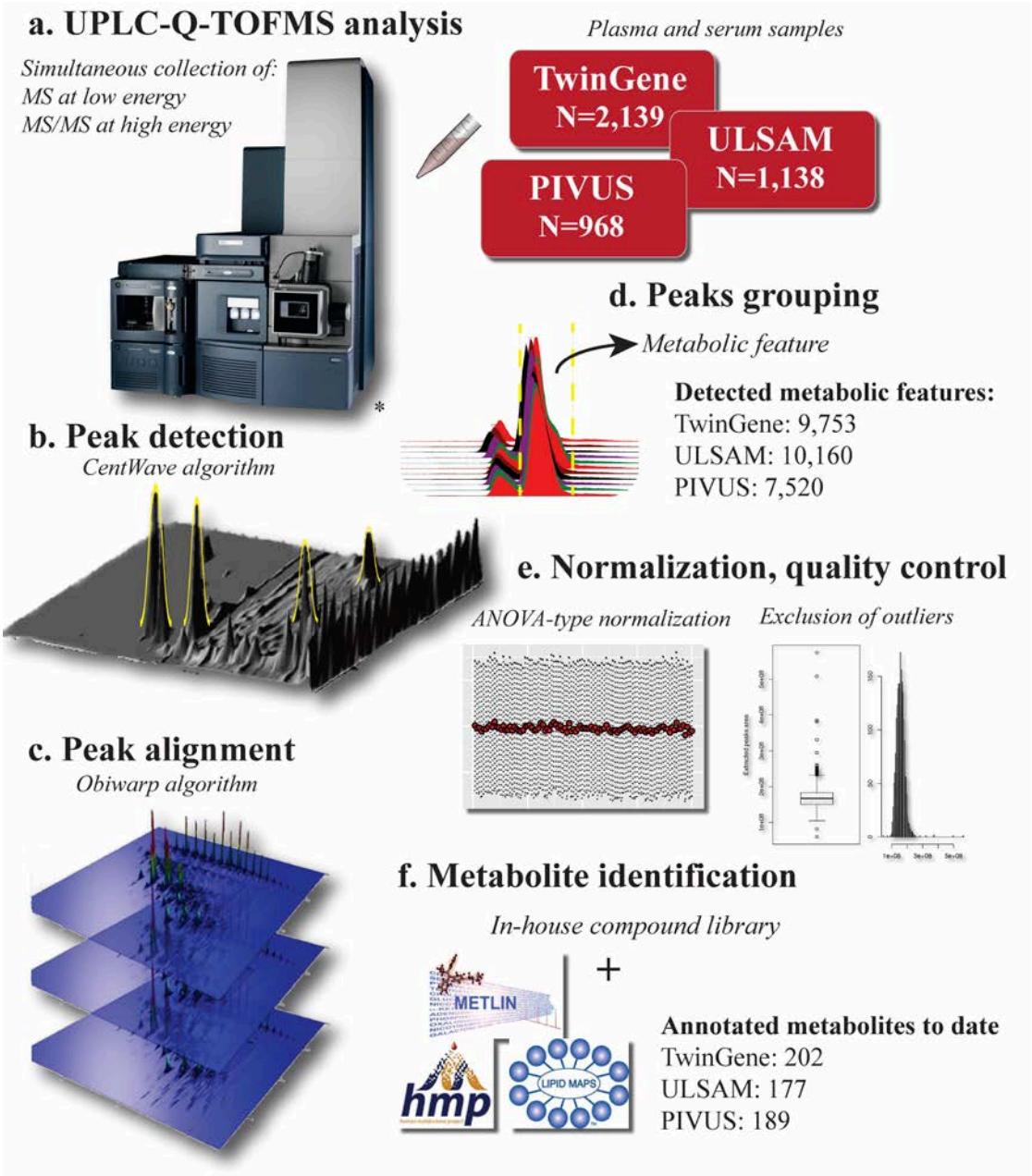

Figure 1. Schematic representation of the workflow used to obtain and process metabolomics data. **A**. Data on both MS and MS/MS were acquired from UPLC-MS. **B**. Peaks were detected for each chromatogram. **C**. Peaks were aligned across samples. **D**. Peaks were grouped across samples. Each peak group is called 'metabolic feature'. **E**. Metabolic features were log-transformed, sample outliers excluded and data were normalized using ANOVA-type normalization, which accounts for factors of unwanted variation. **F**. Metabolites were identified by matching MS/MS reconstructed spectra with the in-house compound library or using publically available databases. *Photo courtesy of Waters Corporation.

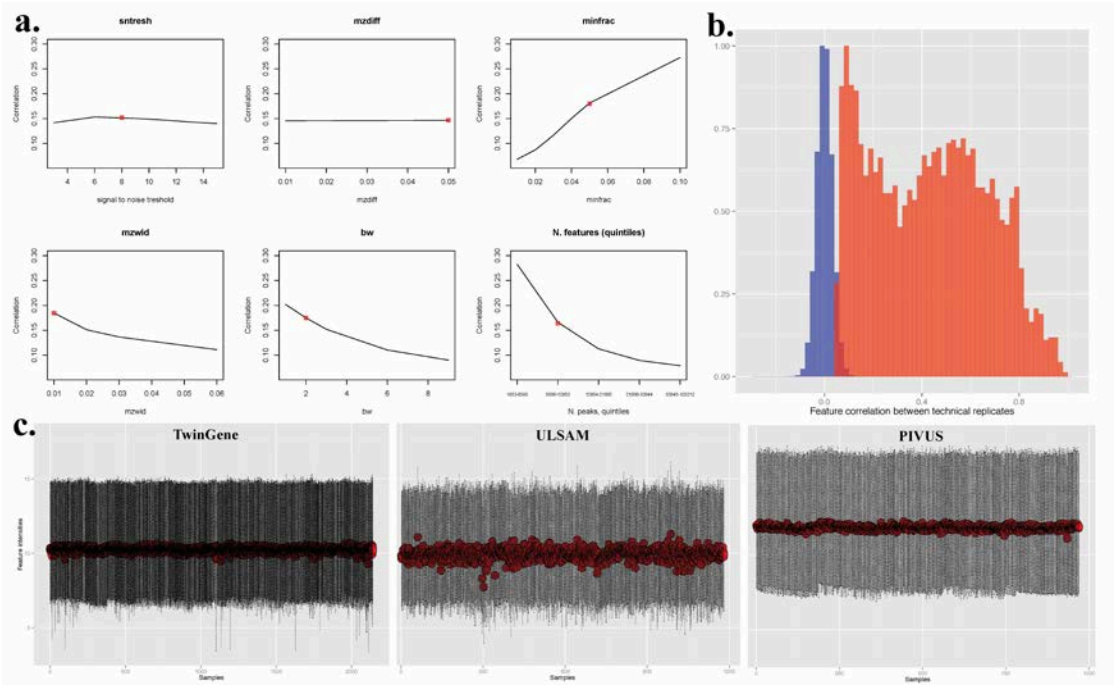

Figure 2. A. Results of iterative testing of different parameters for detection, alignment, grouping and filling steps in 30 random individuals (120 chromatograms) from PIVUS. We ran all 2,161 possible combinations of values within the reported ranges for five parameters (*sntresh*, *mzdiff*, *minfrac*, *mzwid*, *bw*). The reported correlations were the averages of the feature correlations between technical replicates for each parameter, across values of the other parameters. The dot in each panel indicates the value that has been used in the final XCMS settings. The last panel indicates the observed correlation for number of features detected across all possible parameter combinations. **B.** In blue: null distribution of the feature correlations between technical replicates obtained by permutation. In red: observed distribution of feature correlations. The observed correlations are almost always higher than those expected under the null. **C.** Feature distributions across samples in all three studies. The red dots represent the average correlations and the dotted bars represent the ranges between the 5^{th} and 95^{th} percentile of the distribution.