clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Metabolomic Profiling on 9,992 Participants using Ultra-performance Liquid Chromatography and Mass Spectrometer

Data Support Document

Greg Michelotti, Kari Wong, Vince Forgetta, Yiheng Chen, Brent Richards, Josko Ivica, Divya Joshi and Cynthia Balion

Abstract

Metabolomics is the profiling of small molecules in biological fluids, cells, and tissues. These small molecules, commonly known as metabolites are intermediate and end products of cellular metabolism processes. Thus, metabolomics can provide insight into the biological and physiological mechanisms underlying normal physiology and pathophysiological states including diseases. In this data release, we profiled metabolites using blood samples from 9,992 participants enrolled in the Comprehensive cohort of the Canadian Longitudinal Study on Aging (CLSA) at baseline. The analytical method utilized a Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operated at 35,000 mass resolution. This documentation describes the sample collection and processing, data cleaning, and quality control measures performed for metabolites. The present dataset includes data on 1,314 biochemicals, with 1,071 compounds of known identity (named biochemicals) and 243 compounds of unknown structural identity (unnamed biochemicals). Qualified researchers can access the metabolomics data release via the CLSA Data Access portal.

Contents

1.0 INTRODUCTION, PURPOSE, AND SCOPE

The metabolome is commonly described as the complete set of metabolites that are found within a specific organ, tissue, cell, or biofluid.[1] Metabolites are by-products of metabolism and can include compounds derived endogenously (e.g., sugars, amino acids, lipids, and nucleic acids) and from exogenous sources (e.g., diet, drugs, or the environment). Metabolomic analysis enables characterization of metabolites and provides insight into the alterations occurring in the genome, epigenome, proteome, and transcriptome.[2-3] Thus, metabolomics can be used to explore and understand the normal and pathophysiological processes, which can assist with disease diagnosis and prognosis, and has the potential to inform development of therapeutic interventions.[2-3] This document provides supplemental information for metabolomics data collected at baseline in the CLSA. It includes a description of how these data were obtained and the quality measures taken. Detailed procedural instructions are available in the standard operating procedure (SOP) documents available on the CLSA website.

2.0 INSTRUMENTS AND METHODOLOGY

2.1 Data Collection

2.1.1 Blood collection and processing

Participants who are part of the Comprehensive Cohort (n=30,097) and attend the Data Collection Site (DCS) are eligible for blood and urine sample collection. Of the participants in the Comprehensive cohort at baseline, 23,492 participants had provided blood and urine samples and had availability of EDTA whole blood and Buffy coat. Of these 23,492 participants, 6,268 participants had fasted for 5 or more hours, and of which, 3,000 were selected for metabolomics analyses. Additionally, another 7,000 participants were selected from the remaining 20,492 participants (**FIGURE 1**). All sample selections were made to reflect the distribution of Comprehensive cohort by DCS region (1: University of Victoria, Simon Fraser University/University of British Columbia; 2: University of Manitoba, University of Calgary; 3: McGill University, University of Ottawa; 4: University of Sherbrooke, McGill University; 5: Dalhousie University, Memorial University), age group (45-54, 55-64, 65-74, and 75-85), and sex.

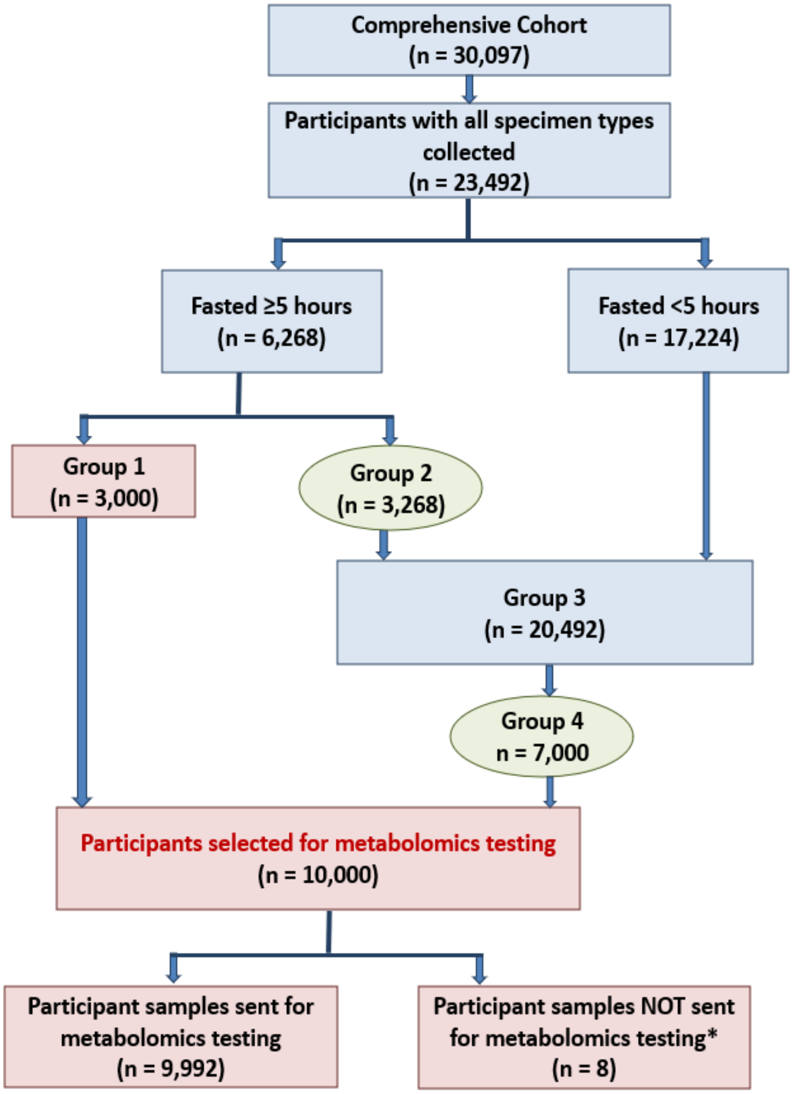


FIGURE 1. Process of selection of study participants for metabolomics analysis
*The reasons include: no samples collected, no aliquots or only aliquots with insufficient volume left

2.1.2 Sample retrieval and shipping

Aliquots from participant's EDTA plasma, stored in the Biorepository and Bioanalysis Centre (BBC) cryofreezers in liquid nitrogen vapour, were retrieved in 2 batches (September 2019 and May 2020) and shipped to the Metabolon testing laboratory in cryofreezers filled with liquid nitrogen.

2.1.3 Sample accessioning

Following receipt at the Metabolon testing facility, samples were inventoried and immediately stored at -80°C. Each sample received was accessioned into the Metabolon Laboratory Information Management System (LIMS) and was assigned by the LIMS a unique identifier that was associated with the original source identifier only. This Metabolon LIMS identifier was used to track all sample handling, tasks, and results. The samples (and all derived aliquots) were tracked by the LIMS system as well. All portions of any sample were automatically assigned their own unique identifiers by the LIMS when a new task was created; the relationship of these samples was also tracked. All samples were maintained at -80^{o}C until processed.

2.1.4 Sample preparation

Samples were prepared using the automated MicroLab STAR® system from Hamilton Company. To remove protein, dissociate small molecules bound to protein or trapped in the precipitated protein matrix, and to recover chemically diverse metabolites, proteins were precipitated with methanol under vigorous shaking for 2 min. (Glen Mills GenoGrinder 2000) followed by centrifugation. The resulting extract was divided into five fractions: two for analysis by two separate reverse-phase (RP)/UPLC-MS/MS methods with positive ion mode electrospray ionization (ESI), one for analysis by RP/UPLC-MS/MS with negative ion mode ESI, one for analysis by HILIC/UPLC-MS/MS with negative ion mode ESI, and one sample was reserved for backup. Samples were placed briefly on a TurboVap® (Zymark) to remove the organic solvent. The sample extracts were stored overnight under nitrogen before preparation for analysis.

2.1.5 Analysis

All methods utilized a Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operated at 35,000 mass resolution. The sample extract was dried then reconstituted in solvents compatible to each of the four methods. Each reconstitution solvent contained a series of standards at fixed concentrations to ensure injection and chromatographic consistency. One aliquot was analyzed using acidic positive ion conditions, chromatographically optimized for more hydrophilic compounds. In this method, the extract was gradient eluted from a C18 column (Waters UPLC BEH C18-2.1x100 mm, 1.7 µm) using water and methanol, containing 0.05% perfluoropentanoic acid (PFPA) and 0.1% formic acid (FA). Another aliquot was also analyzed using acidic positive ion conditions; however, it was chromatographically optimized for more hydrophobic compounds. In this method, the extract was gradient eluted from the same C18 column using methanol, acetonitrile, water, 0.05% PFPA and 0.01% FA and was operated at an overall higher organic content. Another aliquot was analyzed using basic negative ion optimized conditions using a separate dedicated C18 column. The basic extracts were gradient eluted from the column using methanol and water, however with 6.5 mM Ammonium Bicarbonate at pH 8. The fourth aliquot was analyzed via negative ionization following elution from a HILIC column (Waters UPLC BEH Amide 2.1x150 mm, 1.7 µm) using a gradient consisting of water and acetonitrile with 10 mM Ammonium Formate, pH 10.8. The MS analysis alternated between MS and data-dependent MS^n scans using dynamic exclusion. The scan range varied slighted between methods but covered 70-1000 m/z. Raw data files are archived and extracted as described below.

Instrument/Method:	Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer; UPLC-MS/MS
Specimen Type:	Plasma collected in 6 mL EDTA-containing BD Vacutainers
Storage Type:	Liquid nitrogen (-196°C) for long-term storage and -80°C prior to analysis
Storage Time:	Between 4 – 8 years
Tests:	Entire metabolome

2.1.6 Bioinformatics

The informatics system consisted of four major components, the Laboratory Information Management System (LIMS), the data extraction and peak-identification software, data processing tools for quality control (QC) and compound identification, and a collection of information interpretation and visualization tools for use by data analysts. The hardware and software foundations for these informatics components were the LAN backbone, and a database server running Oracle 10.2.0.1 Enterprise Edition. The purpose of the Metabolon LIMS system was to enable fully auditable laboratory automation through a secure, easy to use, and highly specialized system. The scope of the Metabolon LIMS system encompasses sample accessioning, sample preparation and instrumental analysis and reporting and advanced data analysis. All the subsequent software systems are grounded in the LIMS data structures. It has been modified to leverage and interface with the in-house information extraction and data visualization systems, as well as third party instrumentation and data analysis software.

2.1.7 Data Extraction and Compound Identification

Raw data was extracted, peak-identified and processed using Metabolon's hardware and software. These systems are built on a web-service platform utilizing Microsoft's .NET technologies, which run on high-performance application servers and fiber-channel storage arrays in clusters to provide active failover and load-balancing. Compounds were identified by comparison to library entries of purified standards or recurrent unknown entities. Metabolon maintains a library based on authenticated standards that contains the retention time/index (RI), mass to charge ratio (*m/z)*, and chromatographic data (including MS/MS spectral data) on all molecules present in the library. Furthermore, biochemical identifications are based on three criteria: retention index within a narrow RI window of the proposed identification, accurate mass match to the library +/- 10 ppm, and the MS/MS forward and reverse scores between the experimental data and authentic standards. The MS/MS scores are based on a comparison of the ions present in the experimental spectrum to the ions present in the library spectrum. While there may be similarities between these molecules based on one of these factors, the use of all three data points can be utilized to distinguish and differentiate biochemicals. More than 3,300 commercially available purified standard compounds have been acquired and registered into LIMS for analysis on all platforms for determination of their analytical characteristics. Additional mass spectral entries have been created for structurally unnamed biochemicals, which have been identified by virtue of their recurrent nature (both chromatographic and mass spectral). These compounds have the potential to be identified by future acquisition of a matching purified standard or by classical structural analysis.

2.1.8 Metabolite Quantification and Data Normalization

Peaks were quantified using area-under-the-curve. For studies where measurements were spanning multiple days, such as with the CLSA, a data normalization step was performed to correct variation resulting from instrument inter-day tuning differences. Essentially, each compound was corrected in run-day blocks by registering the medians to equal one (1.00) and normalizing each data point proportionately (termed the "block correction"; **FIGURE 2**). Metabolon performs minimum value imputation in which the minimum value detected for a given metabolite is used to replace the missing values for that same metabolite.

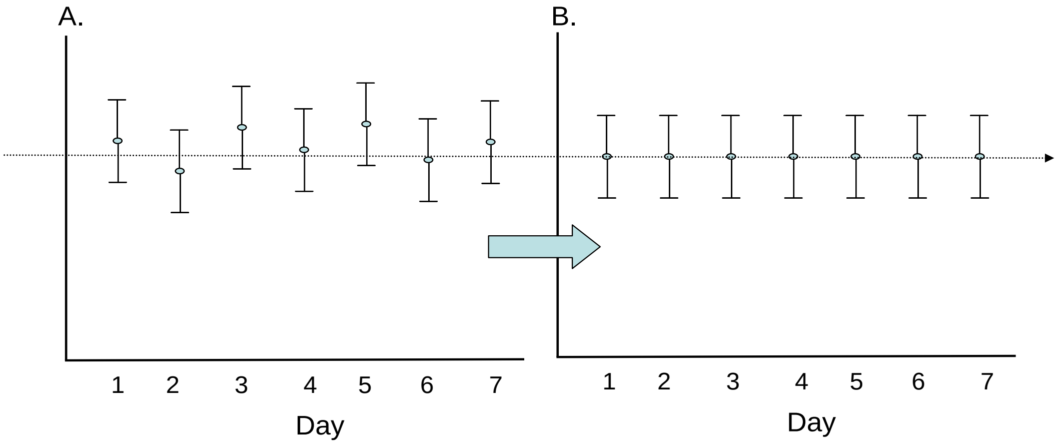


FIGURE 2. Visualization of data normalization steps for a multiday platform run

2.2 Data Cleaning

2.2.1 Participant selection

The SAS SURVEYSELECT procedure was used to select the samples. There were 30,097 Comprehensive Cohort participants at baseline of which 29,826 (99.1%) consented for blood and urine collection. From these participants, 10,000 participant sub-sample was selected for the metabolomics study, in accordance with the above-mentioned selection criteria.

2.2.2 Test results

From selected 10,000 samples, 9,992 (99.9%) participant samples were sent to the Metabolon Laboratory for testing. The present dataset comprises a total of 1,314 biochemicals, with 1,071 compounds of known identity (named biochemicals) and 243 compounds of unknown structural identity (unnamed biochemicals).

There was a small subset of samples that were potentially exposed to variable temperatures but arrived at Metabolon frozen. A principal components analysis was performed to verify that the metabolic signature of these samples was not impacted. **FIGURE 3** demonstrates that the samples potentially exposed to temperature variations (YES) did not differ from those that were not exposed to temperature variation (NO).

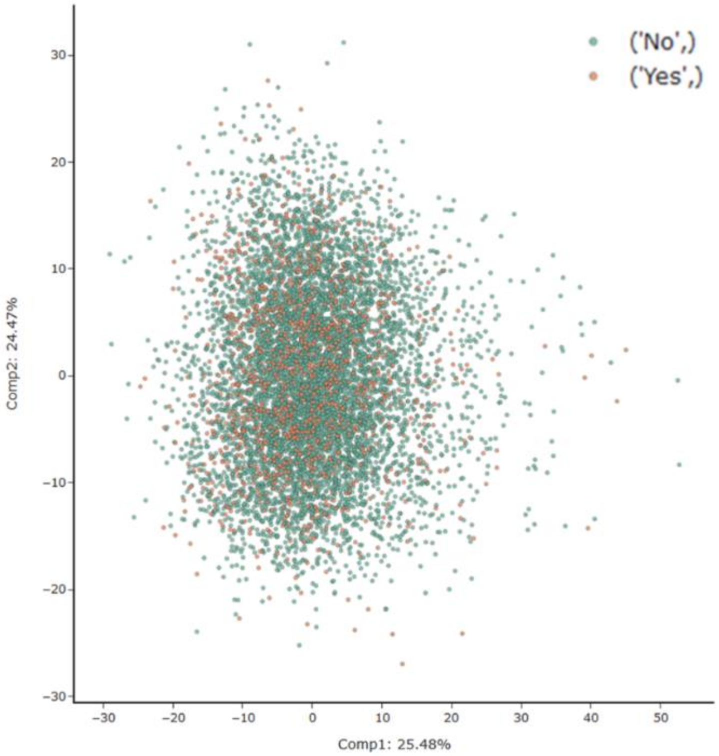


FIGURE 3. Principle Components Plot by potential temperature exposures

2.2.3 Data Merging

The data were merged using two different approaches: Batch normalization and QC Matrix normalization. Batch normalization is a run day normalization approach which involves taking the median of the experimental samples across each batch, for each metabolite. The raw values in the experimental samples were divided by the median of those samples in each instrument batch, giving each batch and thus the metabolite a median of one. The QC-normalized data are data that have been normalized based on an internal pool of EDTA plasma that was run across all projects. For each metabolite, for each batch (as determined from the "NEG", "POLAR", "POS EARLY", and "POS LATE" columns from the meta data), the peak areas for the experimental samples were divided by the median of the QC Matrix samples in that batch if at least 50% of the QC Matrix peaks areas had non-missing values in that batch. Otherwise, the experimental samples for that batch were assigned "Data_not_bridged" to all the samples in that batch. QC Matrix normalization should be used when there are differences between datasets run at different times that are necessary to be retained.

The peak area data, the batch-normalized data, and the QC-normalized data sets were merged by stacking. If a metabolite was not present in one of the individual data sets, it was assigned a value of 'Metabolite_not_called_in_this_set' (it was either absent or had data quality issues). Imputation by the minimum for the batch-normalized and QC-normalized data sets was performed after the data sets were stacked combined excluding those samples marked as 'Metabolite_called_in_this_set' or 'Data_not_bridged.' For this reason, samples indicated with either of these were left blank/missing in the merged sets with imputed values since these should not be imputed with the minimum but can be treated as missing-at-random (so can leave blank for inferential statistics). **Note:** Because of this reason, there may be columns that are blank in the imputed data sets.

2.3 Quality

Data quality assessment was performed using the Metabolon internal QC and standards, longitudinal quality control (LQC) samples, batch-based measurements, and method comparison for select metabolites. The metabolite measurements from the data in file "CLSA NORMDATAALL_COMMONMETABOLITESONLY" were used for the following analyses (except for the "Internal quality control").

2.3.1 Internal quality control

Several types of controls were analyzed in concert with the experimental samples: a pooled matrix sample generated by taking a small volume of each experimental sample (or alternatively, use of a pool of well-characterized human plasma) served as a technical replicate throughout the data set; extracted water samples served as process blanks; and a cocktail of QC standards that were carefully chosen not to interfere with the measurement of endogenous compounds were spiked into every analyzed sample, allowed instrument performance monitoring and aided chromatographic alignment. **TABLE 1** and **2** describe these QC samples and standards. Instrument variability was determined by calculating the median relative standard deviation (RSD) for the standards that were added to each sample prior to injection into the mass spectrometers. Overall process variability was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in 100% of the pooled matrix samples. Experimental samples were randomized across the platform run with QC samples spaced evenly among the injections, as outlined in **FIGURE 4.** Several recovery standards were added prior to the first step in the extraction process for QC purposes.

TABLE 1. Description of Metabolon QC Samples

Type	Description	Purpose
MTRX	Large pool of human plasma maintained by Metabolon that has been extensively characterized.	Assure that all aspects of the Metabolon process are operating within specifications.
PRCS	Aliquot of ultra-pure water	Process Blank used to assess the contribution to compound signals from the process.
SOLV	Aliquot of solvents used in extraction.	Solvent Blank used to segregate contamination sources in the extraction.

TABLE 2. Metabolon Standards

Type	Description	Purpose
RS	Recovery Standard	Assess variability and verify performance of extraction and instrumentation.
IS	Internal Standard	Assess variability and performance of instrument.

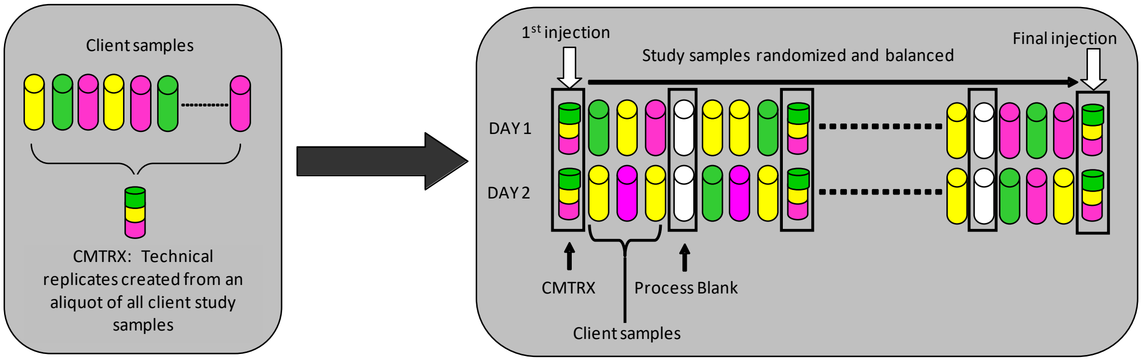

Client samples
CMTRX: Technical replicates created from an aliquot of all client study samples
1st injection
Study samples randomized and balanced
Final injection
DAY 1
DAY 2
CMTRX
Process Blank
Client samples

FIGURE 4. Use of pooled EDTA plasma technical replicates. Pooled EDTA plasma was used to create technical replicate samples (multi-colored cylinder), which was injected periodically throughout the platform run. Variability among consistently detected biochemicals can be used to calculate an estimate of overall process and platform variability.

Instrument variability was determined by calculating the median relative standard deviation (RSD) for the internal standards that were added to each sample prior to injection into the mass spectrometers. Overall process variability was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in 100% of the MTRX7 samples, which are technical replicates of pooled EDTA plasma samples. Overall process variability was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in the MTRX7 technical replicates. Values for instrument and process variability meet Metabolon's acceptance criteria as shown in **TABLE 3**.

TABLE 3. QC Results

QC Sample	***Measurement***	***Median RSD***
Internal Standards	Instrument Variability	7%
Endogenous Biochemicals	Total Process Variability	9%

A subset of samples was re-run when it was discovered that some of the QC lot of pooled EDTA plasma was adulterated by the supplier. This finding does not impact the quality of the sample data (i.e., Peak Area and Batch norm data) but does impact the QC normalized data which used these pooled EDTA plasma samples to bridge future samples. To explain further, there are two methods of data normalization to merge projects run at different times. Batch normalization (Batch-norm data) is performed by stacking the two individual sets. MTRX-normalized data (QC-norm Data) normalization is also performed. The subset of samples affected were re-run so that MTRX-normalization is possible for future studies. This approach will be important to future projects where it is important to preserve important differences that exist between the two sets (i.e., longitudinal collections).

2.3.2 Curation

A variety of curation procedures were carried out to ensure that a high-quality data set was made available for statistical analysis and data interpretation. The QC and curation processes were designed to ensure accurate and consistent identification of true chemical entities, and to remove those representing system artifacts, mis-assignments, and background noise.

Metabolon data analysts use proprietary visualization and interpretation software to confirm the consistency of peak identification among the various samples. Library matches for each compound were checked for each sample and corrected if necessary.

2.3.3 Longitudinal quality control (LQC) samples-based data quality assurance

The LQC samples were prepared by CEQAL (Vancouver, BC) with the aim to track the assay performance as well as biospecimen stability over the long term. Two types of control samples are present within each batch: (a) a plasma sample from one individual; (b) a pooled plasma from several individuals. Batch-normalized and log2-transformed metabolites measurements were used for the following analysis ($N_{metabolites}$ = 1257). Pearson's pairwise correlation coefficients were calculated by correlating the metabolite measurements of each sample to the mean of control measurements (hereby termed "reference"); where the mean is obtained separately for individual or pooled controls. As expected, all control samples (either using the individual (red points) or pooled plasma (orange points)) are highly correlated to the reference control measurement (Pearson's r > 0.9), while the CLSA participant samples showed overall lower correlation (**FIGURE 5**).

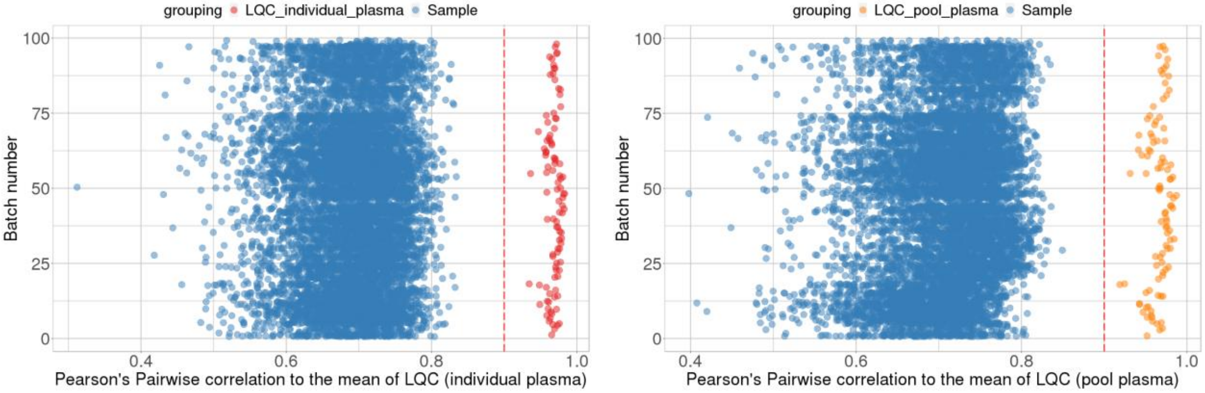


FIGURE 5. Performance of the LQC samples in the CLSA metabolomics dataset. A dot plot showing Pearson's pairwise correlation values of each sample, computed from all metabolites, using the mean of measurements in LQC samples with individual plasma (left panel) or LQC samples with pool plasma (right panel) as reference. Red dashed line: r = 0.9.

2.3.4 Batch-based data quality assurance

To assess the variability of metabolite measurements across batches, the distribution of measurements on each batch (N_{batch}=99) was plotted as density plots using batch-normalized and log2-transformed data. The measurements that are 5 standard deviations (SD) away from the mean were removed. As shown in the **FIGURE 6**, the measurements on each batch are normally distributed and the distribution from different batches are largely concordant.

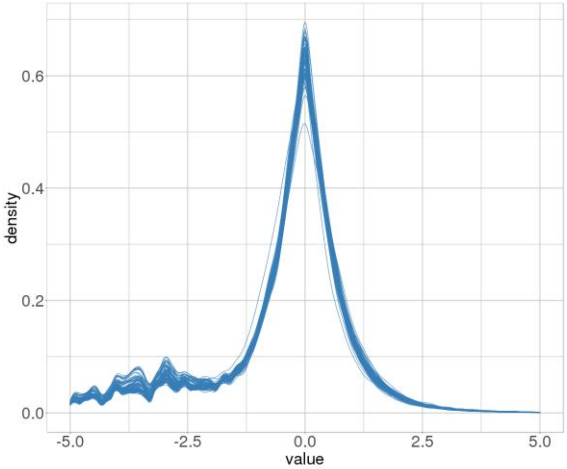


FIGURE 6. Density plot for all measurements from each batch.

2.3.5 Method comparison

The cholesterol and creatinine values from the same participants were compared to those obtained using standardized laboratory methods traceable to the respective reference measurement procedures. Batch-normalized and log2 transformed metabolomics data were used. The laboratory measurements of cholesterol and creatinine were also log2 transformed. The results show that the cholesterol and creatinine measurements from mass spectrometer measurements and laboratory tests are highly correlated (**FIGURE 7**). The correlation coefficients obtained are in accordance with those previously reported (cholesterol, Pearson's R = 0.79; creatinine, Pearson's R = 0.87).[4]

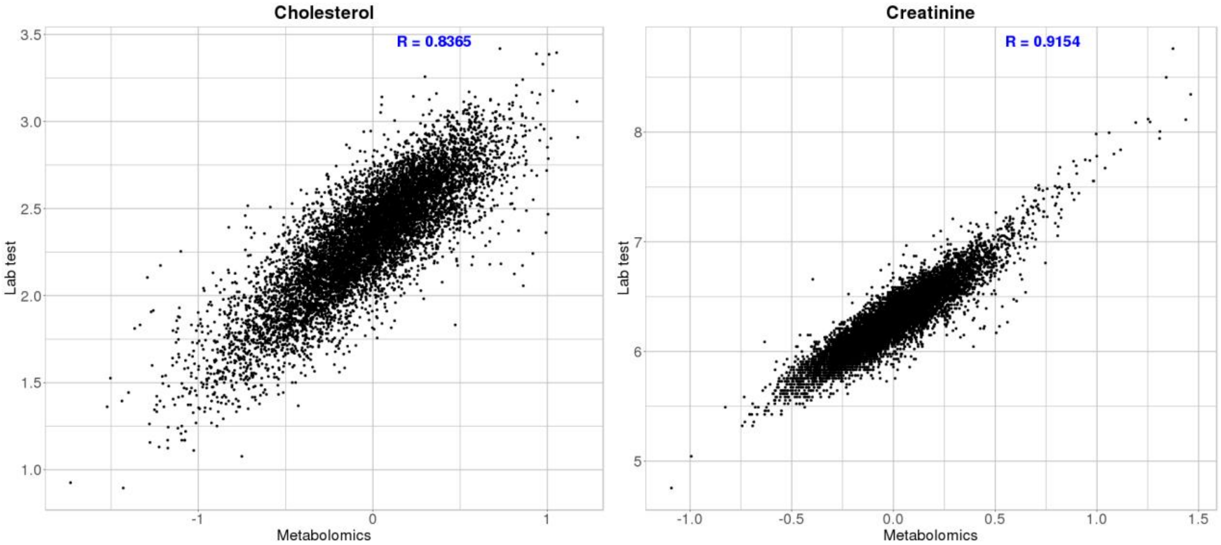


FIGURE 7. Mass-spectrometry measurements (x axis) versus log2 transformed lab tests results (y axis) for cholesterol (Pearson's R = 0.8365, P < 2.2x10^{-16}) and creatinine (Pearson's R = 0.9154, P < 2.2x10^{-16}).

2.5.6 Example of metabolomics data utilization

To further validate the utility of the metabolomics data, an exemplar genome-wide association study (GWAS) on proline using batch-normalized and log2-transformed metabolomics data (8,502 participants with European ancestry, 15,772,365 loci) was performed to validate the proline-associated single nucleotide polymorphisms (SNPs) reported previously.[6] The measurements that were >5 SD away from the mean were removed. For genotyping data from CLSA genomics baseline dataset, low-quality genetic variants were removed by retaining only those SNPs with a minor allele frequency (MAF) higher than 0.1% and imputation quality score of higher than 0.3, and individuals from European ancestry were analyzed to control for population stratification. The additive allelic effects of SNPs on proline levels were assessed using a linear mixed-model in the fastGWA software.[5] The cryptic relatedness, age, sex, hour to last meal or drink, genotyping batch and first 10 principal components of European ancestry were included as covariates in the model. The Manhattan plot **(FIGURE 8)** demonstrates that the proline GWAS largely replicated previous findings.[6]

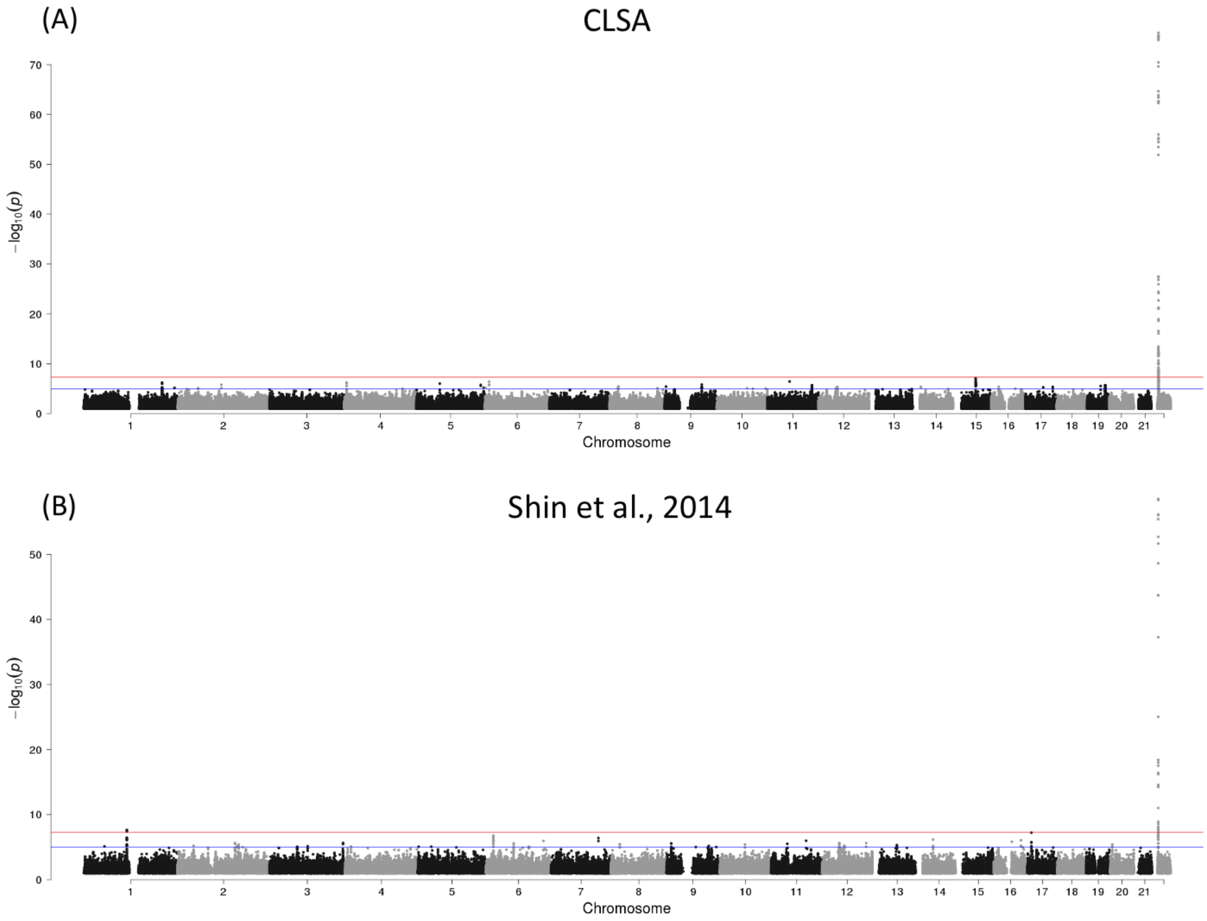


FIGURE 8. Manhattan plot of proline levels. The x-axis represents the locations of SNPs, and the y-axis represents the negative logarithm of the p-values. The red line represents the significant threshold ($P < 5x10^{-8}$) and the blue line represents the suggestive threshold ($P < 5x10^{-6}$). (A) plot using CLSA metabolomics and genomics data (B) plot using published summary statistics by Shin et al., 2014.[6]

3.0 ADDITIONAL INFORMATION AVAILABLE

The CLSA also collects additional data for tracking and quality purposes. These data are available upon request (**TABLE 4**).

TABLE 4: Meta-data

Type	Description
Collection	Venipuncture arm
Collection	Lot number of vacutainer tubes
Environmental	Room temperature during collection and processing
Environmental	Room humidity during collection and processing
Environmental	Temperature during transport and storage
Tracking	Blood collection time
Tracking	Processing time
Tracking	Storage time

3.1 Available dataset files

1. CLSA DATA_KEY_AND_EXPLANATION: Description of files and annotations
2. CLSA ANNOTATIONTABLEALL: The column IDs for the metabolites in each of the sets is "CHEM_ID" from this table. Use this table to find the name of the metabolite and other identifiers.
3. CLSA_METADATAALL: A combination of client-reported and Metabolon-generated metadata linked to each sample analyzed (e.g., age, race, BMI, client-ID, etc.).
4. CLSA PEAKAREADATAALL: This contains all the metabolites detected in all datasets. These are the unnormalized peak areas (i.e., total ion counts, integrated area-under-the-curve).
5. CLSA NORMDATAALL: This file contains batch normalized data for all metabolites detected across all datasets. If a metabolite was not present in one of the individual data sets, it is assigned a value of ‘Metabolite_not_called_in_this_set.’
6. CLSA NORMIMPDATAALL: This file contains batch normalized data that has been imputed for metabolites detected across both datasets. For each metabolite, the minimum value across all batches in the median scaled data is imputed for the missing values. Metabolites that were only detected in in one of the individual datasets are marked as ‘Metabolite_called_in_this_set’ or ‘Data_not_bridged.’
7. CLSA NORMDATAALL_COMMONMETABOLITESONLY: This file contains batch normalized data; however, only lists the metabolites that were detected in all datasets.
8. CLSA QCNORMIMPDATAALL: This is QC Normalized data (see above). Metabolites that were detected in less than 50% of the QC samples in within a dataset could be bridged and are designated “Data_not_bridged.”
9. CLSA QCNORMDATAALL_COMMONMETABOLITESONLY: This file contains QC normalized data only for metabolites that were detected in all datasets.

3.1.1 Notes on QC steps, imputation, and log transformation:

1. QC steps and imputation using minimum value were not performed for the 'CLSA NORMDATAALL' file. A pooled EDTA plasma matrix with all batches of CLSA was performed. The data from these QC samples were used to perform QC normalization to account for the inherent drift that occurs between batches using mass spectrometer technologies. QC steps and imputation using minimum value was performed for the 'CLSA QCNORMIMPDATAALL' dataset, and thus, this dataset is ready to be analyzed. The other datasets are provided for researchers who may prefer to use their own imputation and normalization processes.

2. The metabolite data available in the 'CLSA NORMDATAALL', 'CLSA NORMIMPDATAALL', and 'CLSA QCNORMIMPDATAALL' files are not log transformed.

3.2 References

1. Wishart, D. S. (2019). Metabolomics for investigating physiological and pathophysiological processes. Physiological reviews, 99(4), 1819-1875. https://doi.org/10.1152/physrev.00035.2018

2. Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nature reviews Molecular cell biology*, *17*(7), 451-459. https://doi.org/10.1038/nrm.2016.25

3. Aderemi, A. V., Ayeleso, A. O., Oyedapo, O. O., & Mukwevho, E. (2021). Metabolomics: A scoping review of its role as a tool for disease biomarker discovery in selected non-communicable diseases. *Metabolites*, *11*(7), 418. https://doi.org/10.3390/metabo11070418

4. Bar, N., Korem, T., Weissbrod, O., Zeevi, D., Rothschild, D., Leviatan, S., Kosower, N., Lotan-Pompan, M., Weinberger, A., le Roy, C. I., Menni, C., Visconti, A., Falchi, M., Spector, T. D., Adamski, J., Franks, P. W., Pedersen, O., & Segal, E. (2020). A reference map of potential determinants for the human serum metabolome. *Nature*, *588*(7836). https://doi.org/10.1038/s41586-020-2896-2

5. Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, *51*(12). https://doi.org/10.1038/s41588-019-0530-8

6. Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., Walter, K., Menni, C., Chen, L., Vasquez, L., Valdes, A. M., Hyde, C. L., Wang, V., Ziemek, D., Roberts, P., … Soranzo, N. (2014). An atlas of genetic influences on human blood metabolites. *Nature Genetics*, *46*(6). https://doi.org/10.1038/ng.2982

3.3 Supplementary Information

March 2, 2022

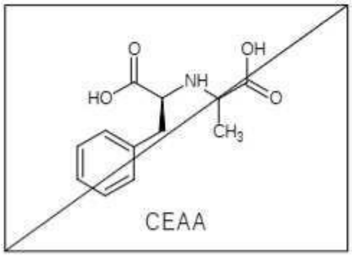
Dear Client:

We are writing to inform you that Metabolon recently identified conflicting structural data regarding carboxyethyl (CE) modified amino acids (CEAA), for example 1-carboxyethyl phenylalanine. The molecules that are being detected have been determined to be the N-lactoyl (NL) modified amino acids (NLAA) rather than CE modified amino acids. A nonissue for most studies, these metabolites represent the modification of common amino acids and ultimately turn out to be most directly related to lactate levels in the sample. **Our initial analysis indicates that your study is likely not impacted by this finding. If you have any questions about a potentially impacted study, please contact your main point of contact at Metabolon.**

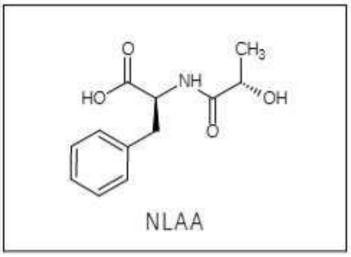
In the meantime, here is more background on this phenomenon.

What was the finding?

When lactate goes up these modified amino acids also go up, representing the lactate response rather than being a novel molecule. As a representation of lactate biology and what lactate is doing, we do not believe this will have material impact to the biological conclusions drawn from any given research. However, we are alerting all our impacted clients of this correction.



CEAA



NLAA

Before these metabolites were identified, they appeared to represent a novel class of metabolites, namely carboxyethyl modified amino acids. The desire of our scientists to understand the biology of these molecules drove them to investigate further. However, upon further investigation we identified unexpected biological signals (e.g., correlation with lactate). Additional research led us to understand that the original class of molecules was behaving unexpectedly. Further investigation into the molecules permitted us to determine that the initial synthesis of these CE modified amino acids yielded N-lactoyl amino acids as a byproduct. This was not obvious at the time because the synthetic reaction to modify amino acids with the CE moiety resulted in an unexpected and very unlikely modification and, in a rare occurrence, metabolites with the correct mass, retention time and fragmentation patterns were produced, and misidentified as the CE modified amino acids.

What compounds were affected?

A total of six (6) compounds were determined to have been affected. Below is the list of compounds whose initial identifications were determined to be incorrect and their correct identifications and Human Metabolome Database (HMDB) IDs. Note that all six (6) compounds will not necessarily be present in your data. It is likely that in your data only a subset of these were detected and reported.

Incorrect Identification	Correct Identification	Correct HMDB
1-carboxyethylvaline	N-lactoyl valine	HMDB0062181
1-carboxyethylphenylalanine	N-lactoyl phenylalanine	HMDB0062175
1-carboxyethylleucine	N-lactoyl leucine	HMDB0062176
1-carboxyethylisoleucine	N-lactoyl isoleucine	HMDB0062180
1-carboxyethyltyrosine	N-lactoyl tyrosine	HMDB0062177
1-carboxyethylhistidine	N-lactoyl histidine	N/A

Why are you notifying me?

Despite our confidence that there was a minimal impact to biological signal, our commitment to providing the best data and insights led our scientists to conduct a detailed review to understand the potential significance of these metabolites. What we do know is that these modified amino acids have not been seen to impact areas of safety or efficacy.

What action is Metabolon taking?

We initiated a Corrective and Preventive Action (CAPA) to thoroughly investigate this issue, including the root cause and its potential impact on historical studies. Further, we have implemented additional quality control measures to ensure this, or similar errors do not occur again.

Metabolon remains steadfastly committed to the highest level of quality control and reporting accuracy for all studies we perform. We will continue to execute on-going quality control assessments to maintain the highest level of accuracy and confidence in our results. Should you have any other questions regarding this update or our services, please do not hesitate to contact us at qualityinquiry@metabolon.com.

Sincerely,

Pamela J Nakhle

SignNow e-signature ID: 02479856e3...
03/02/2022 17:26:37 UTC
Pamela Nakhle (Signer)

Pamela J. Nakhle, PhD, CMQ/OE, RAC

Sr. Director, Quality Assurance & Regulatory Affairs