Data Processing and Quality Control Documentation

Project Information

Project Title: Trans-omics Integration of Multi-omics Studies for Osteoporosis

PI: Dr. Hong-Wen Deng
Grant support: U19AG055373
Version: v1.0
Latest updated date: 02/23/2026

1. Overview

This document describes the data generation, preprocessing, and quality control procedures applied to serum metabolomics, lipidomics, and short-chain fatty acid (SCFA) datasets generated from male participants in the Louisiana Osteoporosis Study (LOS) and/or the Trans-omics Integration of Multi-omics Studies for Male Osteoporosis (MOMO).

All datasets are de-identified. No personally identifiable information is included. The data are prepared for controlled-access distribution through the National Institute on Aging Research Biobank. Only serum samples are included in this submission.

2. Cohort and Biospecimen Information

Cohort description: Male non-fasting serum samples from Louisiana Osteoporosis study (LOS) and/or the Trans-omics Integration of Multi-omics Studies for Male Osteoporosis (MOMO)

Biospecimen type: Serum

Storage condition prior to analysis: −80 degrees Celsius

All shared datasets contain harmonized study identifiers only. The identifier crosswalk file is maintained separately under controlled access.

3. Vendor Analytical Platform and Data Generation

Serum metabolomics and lipidomics profiling were performed using standardized ultra-high performance liquid chromatography–tandem mass spectrometry workflows. SCFA quantification was performed using mass spectrometry–based targeted methods.

3.1. Sample Handling

Upon receipt, non-fasting serum samples were inventoried and stored at −80°C. All samples were tracked using a laboratory information management system to ensure chain-of-custody and sample integrity.

3.2. Global Metabolomics

Untargeted metabolomic profiling was performed using a Waters ACQUITY UPLC system coupled to a Thermo Scientific Q-Exactive high-resolution/accurate mass spectrometer equipped with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operating at 35,000 mass resolution.

3.2.1. Sample Preparation

Serum samples were thawed on ice and vortex-mixed. Proteins were precipitated by the addition of methanol under vigorous agitation, followed by centrifugation to remove insoluble material. Recovery standards were added prior to extraction to monitor extraction efficiency and instrument performance. The resulting supernatant was divided into multiple fractions for analysis under complementary chromatographic and ionization conditions. Extracts were dried under nitrogen and reconstituted in solvents compatible with each analytical method.

3.2.2. Chromatography and Mass Spectrometry

Samples were analyzed using UPLC coupled to a high-resolution mass spectrometer equipped with a heated electrospray ionization source and Orbitrap mass analyzer.

Four analytical methods were applied to maximize metabolite coverage:

- Reverse-phase chromatography with positive ion mode optimized for hydrophilic compounds
- Reverse-phase chromatography with positive ion mode optimized for hydrophobic compounds
- Reverse-phase chromatography with negative ion mode
- Hydrophilic interaction chromatography with negative ion mode
- Mass scan range typically covered approximately 70–1000 m/z.

3.2.3. Compound Identification

Metabolites were identified by comparison to an authenticated reference library based on:

- Retention index matching
- Accurate mass matching (within instrument tolerance)
- Forward and reverse MS/MS spectral similarity scoring

3.2.4. Data Processing and Normalization

Peaks were quantified using area-under-the-curve integration. For studies spanning multiple days, run-day block correction was performed by registering medians to equal 1.00 and normalizing each data point proportionately. The dataset provided for data sharing corresponds to the batch-normalized values.

3.3. Complex Lipid Profiling

Lipidomic profiling was performed using a methyl-tert-butyl ether (MTBE)–based extraction protocol followed by targeted liquid chromatography–tandem mass spectrometry (LC–MS/MS). Deuterated internal standards spanning major lipid classes were added prior to extraction to enable absolute quantification and monitor analytical performance.

3.3.1. Extraction

Serum samples were mixed with methanol to precipitate proteins, followed by addition of MTBE to induce phase separation. After incubation and centrifugation, the organic phase containing lipids was collected. Extracts were dried under nitrogen and reconstituted in an appropriate solvent system for LC–MS/MS analysis.

3.3.2. LC–MS/MS Analysis

Lipid species were analyzed using a Shimadzu liquid chromatography system coupled to a SCIEX 5500 QTRAP mass spectrometer operating in multiple reaction monitoring (MRM) mode. Chromatographic separation was achieved using reversed-phase liquid chromatography optimized for lipid class separation. Analyses were conducted in both positive and negative electrospray ionization modes to enable detection of diverse lipid subclasses, including phospholipids, sphingolipids, glycerolipids, and sterol lipids. MRM transitions were selected to target class-specific and species-specific precursor-to-product ion pairs.

3.3.3. Quantification

Absolute concentrations were calculated using analyte-to-internal-standard peak area ratios multiplied by the known concentration of corresponding internal standards.

Background subtraction was performed using process blanks. Quality control samples were interspersed throughout analytical batches to monitor assay performance. Run-day normalization was applied when appropriate to reduce inter-batch variability.

3.4. Short-Chain Fatty Acid Quantification

Short-chain fatty acids (SCFAs; C2–C6) were quantified in serum using a targeted LC-MS/MS assay (Metabolon Method TAM148).

3.4.1. Sample Preparation

Samples were spiked with stable isotope-labeled internal standards and subjected to protein precipitation using an organic solvent. After centrifugation, an aliquot of the supernatant was derivatized prior to analysis. Derivatized extracts were injected onto an Agilent 1290 / SCIEX QTRAP 5500 LC-MS/MS system equipped with a C18 reverse-phase UHPLC column. The mass spectrometer was operated in negative electrospray ionization (ESI) and quantified using multiple reaction monitoring (MRM) transitions (analyte product ions measured against the corresponding internal standards).

3.4.2. Quantification

Quantitation was performed using **weighted linear least-squares regression** based on fortified calibration standards prepared concurrently with study samples. Internal standards were used to correct for extraction efficiency and instrument variability. Concentrations were reported in **ng/mL**. LC-MS/MS raw data were collected using SCIEX Analyst 1.7.3 and processed using SCIEX OS-MQ v3.1.6.

4. Analytical Sample Summary

Table 1 summarizes the total number of serum samples processed, the final analytic sample size, and the number of samples excluded as multivariate outliers for each omics platform. Prior to data delivery, all submitted serum samples underwent vendor-level quality assessment. Samples that did not meet predefined analytical acceptance criteria (e.g., insufficient volume, compromised integrity, or failure during extraction or instrumental analysis) were not advanced to final data generation and were therefore not included in the delivered dataset. Such samples are not reflected in the counts summarized in Table 1. For global metabolomics, 999 serum samples were processed and included in the delivered dataset. Following multivariate outlier detection, 8 samples were excluded, resulting in a final analytic sample size of 991. For serum lipidomics, 997 samples were processed. Final analytic sample sizes varied slightly by data layer after quality control. Specifically, 991 samples remained for lipid class concentrations (6 outliers removed), 985 for lipid species concentrations (12 outliers removed), and 987 for fatty acid concentrations (10 outliers removed). For serum short-chain fatty acid (SCFA) quantification, 1000 samples were processed. After multivariate outlier filtering, 994 samples remained for downstream analysis (6 outliers removed). Outlier detection was conducted independently within each analytical platform using the procedures described in

Section 5.4.

Table 1 Sample Summary Across Analytical Platforms

Platform	Subcategory	Total Samples Received	Final Analytic Sample	Outliers Removed
Serum Metabolomics	—	999	991	8
Serum Lipidomics	Lipid Class Concentrations	997	991	6
Serum Lipidomics	Lipid Species Concentrations	997	985	12
Serum Lipidomics	Fatty Acid Concentrations	997	987	10
Serum SCFA	Serum-only	1000	994	6

5. Sample-Level Multivariate Outlier Detection

To identify samples exhibiting abnormal multivariate profiles that may reflect technical artifacts or compromised sample integrity, a principal component–based outlier detection procedure was applied independently within each analytical platform. Because principal component analysis requires a complete data matrix, features with more than 20 percent missing values were temporarily excluded for this quality assessment step only. Remaining missing values were temporarily imputed using a random forest–based method to enable stable covariance estimation. This temporary imputation was performed solely for multivariate quality assessment. No imputed values were retained in the final analytic datasets used for downstream statistical analyses. Prior to PCA, abundance values were log2-transformed (log2[X + 1]), mean-centered, and scaled to unit variance. Mahalanobis distance was calculated for each sample in principal component space. Samples exceeding the 99th percentile of the chi-square distribution were classified as multivariate outliers and removed. Outlier filtering was performed once per dataset.

6. Sample Identifier Harmonization

Sample MM130113 was excluded because it did not meet qualification criteria and could not be mapped to a harmonized study identifier. Identifier mapping is maintained in a controlled-access file.

7. Compliance Statement

All data are de-identified. The study was conducted under Tulane Institutional Review Board approval. Data are shared under controlled-access terms for research use only. These data are not intended for clinical decision-making.