Metabolomics Data Processing and Quality Control

Platform and Instrumentation

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

Sample Preparation

Serum samples were thawed on ice, vortex-mixed, and subjected to protein precipitation with methanol under vigorous agitation followed by centrifugation. 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.

Chromatographic Methods

Four analytical methods were applied to maximize metabolite coverage. Reverse-phase liquid chromatography with positive ion mode (early) was used for hydrophilic compounds, while reverse-phase chromatography with positive ion mode (late) targeted hydrophobic compounds. Reverse-phase chromatography with negative ion mode was applied for acidic and anionic metabolites, and hydrophilic interaction chromatography (HILIC) with negative ion mode was used for polar metabolites. Mass scan range covered approximately 70–1000 m/z.

Compound Identification

Metabolites were identified by comparison to an authenticated in-house reference library based on retention index matching, accurate mass matching within instrument tolerance, and forward and reverse MS/MS spectral similarity scoring.

Batch Normalization

Raw peak areas were quantified using area-under-the-curve integration. For samples spanning multiple analytical batches, run-day block correction was applied by registering batch medians to 1.00 and normalizing each data point proportionately. The delivered dataset reflects these batch-normalized values. Missing values were imputed using the per-metabolite minimum observed value.

Outlier Detection

Multivariate outlier detection was performed using principal component analysis (PCA). Features with >20% missing values were temporarily excluded; remaining missing values were imputed using a random forest–based method solely for this quality assessment step. Abundance values were log2-transformed, mean-centered, and scaled to unit variance prior to PCA. Mahalanobis distance was calculated in principal component space, and samples exceeding the 99th percentile of the chi-square distribution were classified as outliers and removed. A total of 8 samples were excluded from the final metabolomics analytic dataset (999 processed → 991 final).

Sheet Name	Description
Chemical Annotation	Annotation information for all metabolites in the dataset, including pathway classification, database IDs, and chemical identifiers
Method	Documentation of the analytical methods and platforms used for metabolite detection
Sample Meta Data (999 Samples)	Metadata for all samples in the dataset, including sample identifiers, matrix type, and analytical platform assignments
Batch Norm Data	Batch-normalized peak area data representing relative abundance of each metabolite, where raw peak areas are normalized within each analytical batch to account for instrument variability across batches. Each column represents a metabolite identified by CHEM_ID, and each
outliers	Samples or metabolites flagged as statistical outliers during Metabolon's quality control process

Chemical Annotation	
Column	**Description**
MASS	Mass-to-charge ratio (m/z) used for mass spectrometry identification
LIB_ID	Spectral library entry ID; when multiple datasets are merged, appears as LIB_ID_1, LIB_ID_2, etc.; blank if not used in that dataset
CHRO_LIB_ENTRY_ID	Chromatography library entry ID corresponding to retention time and mass spectra information
TYPE	Compound identification type; typically Named (known compound) or Unknown/X- (unknown but reproducibly detected)
INCHIKEY	Standardized chemical structure string (International Chemical Identifier Key) for cross-database matching
SMILES	Text-based chemical structure representation (Simplified Molecular Input Line Entry System)
CHEMICAL_NAME	Systematic chemical name (IUPAC or other standard nomenclature)
PLOT_NAME	Shortened display name used in figures and plots, typically an abbreviated version of BIOCHEMICAL
CHEMSPIDER	Compound ID from the ChemSpider database
PUBCHEM	Compound ID (CID) from the PubChem database

Sample Meta Data (999 Samples)	
Column	**Description**
PARENT_SAMPLE_NAME	Metabolon's internal sample identifier assigned to each sample
CLIENT_IDENTIFIER	Sample identifier provided by the client/researcher, used to match back to original study metadata
NEG	Indicates whether the sample was analyzed on the LC/MS Negative ion mode platform
POLAR	Indicates whether the sample was analyzed on the LC/MS Polar platform
POS EARLY	Indicates whether the sample was analyzed on the LC/MS Positive Early ion mode platform
POS LATE	Indicates whether the sample was analyzed on the LC/MS Positive Late ion mode platform
SAMPLE_AMOUNT	The amount of sample submitted for analysis
SAMPLE_AMOUNT_UNITS	The unit of measurement for the sample amount (e.g., mg, µL, µg)
CLIENT_MATRIX	The sample matrix type as provided by the client (e.g., plasma, urine, serum, tissue)