



OPEN

DATA DESCRIPTOR

# Compilation of longitudinal gut microbiome, serum metabolome, and clinical data in acute myeloid leukemia

Armin Rashidi<sup>1</sup>✉, Maryam Ebadi<sup>1</sup>, Tauseef Ur Rehman<sup>1</sup>, Heba Elhousseini<sup>1</sup>, Hossam Halaweish<sup>2</sup>, Thomas Kaiser<sup>2</sup>, Shernan G. Holtan<sup>1</sup>, Alexander Khoruts<sup>3</sup>, Daniel J. Weisdorf<sup>1</sup> & Christopher Staley<sup>2</sup>

Induction chemotherapy for patients with acute myeloid leukemia (AML) is a unique clinical scenario. These patients spend several weeks in the hospital, receiving multiple antibiotics, experiencing gastrointestinal mucosal damage, and suffering severe impairments in their immune system and nutrition. These factors cause major disruptions to the gut microbiota to a level rarely seen in other clinical conditions. Thus, the study of the gut microbiota in these patients can reveal novel aspects of microbiota-host relationships. When combined with the circulating metabolome, such studies could shed light on gut microbiota contribution to circulating metabolites. Collectively, gut microbiota and circulating metabolome are known to regulate host physiology. We have previously deposited amplicon sequences from 566 fecal samples from 68 AML patients. Here, we provide sample-level details and a link, using de-identified patient IDs, to additional data including serum metabolomics (260 samples from 36 patients) and clinical metadata. The detailed information provided enables comprehensive multi-omics analysis. We validate the technical quality of these data through 3 examples and demonstrate a method for integrated analysis.

## Background & Summary

The standard curative-intent chemotherapy for patients with acute myeloid leukemia (AML) is accompanied by major intestinal microbiota disruptions due mainly to high antibiotic pressure during several weeks of hospitalization<sup>1–3</sup>. The impact of dysbiosis in these patients has only partially been characterized and includes neutropenic fever<sup>4</sup>, bloodstream infection<sup>3,5</sup>, and increased mortality<sup>6</sup>. The involved mechanisms are even less clear and include intestinal domination<sup>6</sup>, gut barrier damage<sup>5</sup>, and altered microbiota-host crosstalk<sup>4</sup>. One of the reasons for our poor understanding of how dysbiosis may influence clinical outcomes in these patients is the inherently high-risk nature of mechanistic studies in this patient population. Intensive chemotherapy suppresses the bone marrow and the resulting decline in platelets and white blood cells make safe access to the intestinal tract and associated interventions challenging and often infeasible due to high bleeding and infection risk. The period of neutropenia typically starts during week 1 of chemotherapy and, in patients achieving a complete remission, ends in week 4.

With these limitations in mind, a good alternative strategy to achieve mechanistic knowledge is by analyzing longitudinal datasets of multi-omic data. Multi-omics data assess different but related aspects of pathogenesis which are often on a causal link. For example, the gut microbiota regulates and significantly contributes to the circulating metabolome<sup>7–9</sup>. In addition, longitudinal data provide insight into intra-individual patterns of change over time<sup>10</sup> which should be distinguished from inter-individual variation due to the personalized features of the microbiota and relevant host-specific factors<sup>11</sup>. Knowledge obtained from longitudinal multi-omics studies facilitates the generation of testable hypotheses for subsequent therapeutic trials<sup>12</sup>.

<sup>1</sup>Division of Hematology, Oncology, and Transplantation, Department of Medicine, University of Minnesota, Minneapolis, MN, USA. <sup>2</sup>Department of Surgery, University of Minnesota, Minneapolis, MN, USA. <sup>3</sup>Division of Gastroenterology, Hepatology, and Nutrition, Department of Medicine, University of Minnesota, Minneapolis, MN, USA. ✉e-mail: [arashidi@umn.edu](mailto:arashidi@umn.edu)

Here, we provide patient- and sample-level longitudinal gut microbiota and circulating microbiome data from AML patients hospitalized to receive induction chemotherapy at the University of Minnesota. Since the initiation of our effort to collect this biorepository, we have been depositing the raw 16S amplicon sequences from the stool samples in the form of paired-end fastq files at the NCBI's Sequence Read Archive<sup>13</sup>. In addition, we recently analyzed a subset of longitudinal serum samples obtained from the same patients for metabolomics and reported them in aggregate<sup>4</sup>. Here, we provide patient- and sample-level metabolomic data and a link between serum and stool samples for each patient. In addition, we provide granular antibiotic exposure data (facilitated by the patients' hospitalized and their closely monitored status) and other clinical metadata for each patient. Collectively, the user will have access to the longitudinal gut microbiome, serum metabolome, and clinical metadata of each patient. This unique database will enable hypothesis generation about gut microbiota and circulating metabolomic changes within and between individuals over time, their possible causal connections, and how baseline and subsequent clinical factors may influence the microbiome and metabolome. Prolonged hospitalization, heavy antibiotic exposure to prevent and treat infections, severe decline in the immune system, cytotoxic damage to the intestinal barrier, and nutritional compromise make the patient population of interest in this study unique. Therefore, we expect the well-annotated multi-omics data compiled here to lead to new discoveries in humans experiencing severe multi-faceted perturbations.

## Methods

**Participants and clinical metadata.** Sample collection and analysis was approved by the University of Minnesota Institutional Review Board (ClinicalTrials.gov Identifier: NCT03316456). All participants provided signed informed consent. Clinical metadata was obtained by reviewing the electronic medical records. In the first step, 3 of the investigators independently collected data. In the second step, a fourth investigator compared the findings between the reviewers and resolved any conflicts. All identifiable data were removed. Patient\_ID is a non-identifiable indicator that can be used to link clinical metadata to omics sample data. All dates are relative to the first day of chemotherapy, which itself will remain confidential. This protocol was initiated in 2017 and closed in 2021. This article includes all data from the study.

**Sample collection and fecal 16S rRNA gene sequencing.** Serum and stool sample collection started with hospital admission and continued twice weekly (Mon/Thu +/− 2 days) until day 28 of chemotherapy or discharge (whichever occurred first). Serum samples were collected twice weekly (Mon/Thu; preprandial) between 6–8 AM in standard red-top tubes, split in 250 μL aliquots, and stored at −80 °C within 2 hours of collection. Stool samples were collected in 95% ethanol-filled sterile tubes and stored at −80 °C. DNA from the stool samples was extracted using the DNeasy PowerSoil DNA isolation kit (QIAGEN, Hilden, Germany). qPCR was used to quantify 16S rRNA gene content in each sample. The V4 hypervariable region of the 16S rRNA gene was amplified on an Illumina MiSeq platform (2 × 300 paired-end mode)<sup>14</sup>. Adaptor trimming was done using SHI7<sup>15</sup>, and the resulting demultiplexed fastq files were used as input to DADA2<sup>16</sup> to infer exact amplicon sequence variants (ASVs) (*dada2* package v1.18.0 in R). For filtering, we used DADA2 default parameters (PHRED score threshold of 2, maximum number of expected errors of 2 for both forward and reverse reads) and truncation lengths of 220 (forward) and 150 (reverse). De-replication, de-noising, merging, and chimera removal were done using DADA2 default parameters. Taxonomic assignment was done by the naive Bayesian classifier implemented in DADA2 and the SILVA non-redundant v138.1 training set<sup>17</sup>. Clinical metadata and ASV abundances were merged into a phyloseq object in R (R Foundation for Statistical Computing, Vienna, Austria) for analysis.

**Serum metabolome profiling.** Serum samples were sent to Metabolon (Morrisville, NC) for untargeted, ultrahigh performance liquid chromatography-tandem mass spectroscopy (UPLC-MS/MS).

**Sample preparation.** Samples were prepared using the automated MicroLab STAR<sup>®</sup> system from Hamilton Company. A total of 100 μL of sample was extracted under vigorous shaking for 2 min (Glen Mills GenoGrinder 2000) with methanol 80%, containing the following recovery standards: DL-2-fluorophenylglycine, tridecanoic acid, d6-cholesterol, and DL-4-chlorophenylalanine. The resulting extract was divided into 5 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, and one for analysis by HILIC/UPLC-MS/MS with negative ion mode ESI. The remaining aliquot was reserved for backup. Samples were placed briefly on a TurboVap<sup>®</sup> (Zymark) to remove the organic solvent. The sample extracts were stored overnight under nitrogen before preparation for analysis.

**Mass spectrometry.** All methods utilized a Waters ACQUITY 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  $R = 35,000$  mass resolution. The sample extract was dried then reconstituted in solvents compatible to each of the four methods. For each sample, two aliquots of each sample were reconstituted in 50 μL of 6.5 mM ammonium bicarbonate in water (pH 8) for the negative ion analysis and another two aliquots of each were reconstituted using 50 μL 0.1% formic acid in water (pH ~3.5) for the positive ion method. Each reconstitution solvent contained a series of standards at fixed concentrations to ensure injection and chromatographic consistency. The internal standards consist of a variety of deuterium labeled or halogenated biochemicals specifically designed both to cover the entire chromatographic run and to not interfere with the detection of any endogenous biochemicals. Authentic standards of d7-glucose, d3-leucine, d8-phenylalanine, and d5-tryptophan were purchased from Cambridge Isotope Laboratories (Andover, MA). d5-hippuric acid, d5-indole acetic acid, and d9-progesterone were procured from C/D/N Isotopes, Inc. (Pointe-Claire, Quebec). Bromophenylalanine was provided by Sigma-Aldrich Co. LLC. (St. Louis, MO) and

Standard	HARAM LOD ng/mL	UMR LOD ng/mL
d7-glucose	1.0	50.0
d3-leucine	0.25	5.0
d8-phenylalanine	0.25	3.0
d5-tryptophan	0.25	25.0
d5-hippuric acid	0.25	5.0
Br-phenylalanine	0.25	3.0
d5-indole acetic acid	3.0	25.0
amitriptyline	0.5	3.0
d9-progesterone	1.0	25.0

**Table 1.** Limit of detection (LOD) for standards in a dilution series using reverse-phase chromatography.

Type	Description	Purpose
CMTRX	Pool created by taking a small aliquot from every customer samples.	Assess the effect of a non-plasma matrix on the Metabolon process and distinguish biological variability from process variability.
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.** Description of metabolon QC samples.

amitriptyline was from MP Biomedicals, LLC. (Aurora, OH). Recovery standards of DL-2-fluorophenylglycine and DL-4-chlorophenylalanine were from Aldrich Chemical Co. (Milwaukee, WI). Tridecanoic acid was purchased from Sigma-Aldrich (St. Louis, MO) and d6-cholesterol was from Cambridge Isotope Laboratories (Andover, MA). Standards for the HILIC dilution series of alpha-ketoglutarate, ATP, malic acid, NADH, and oxaloacetic acid were purchased from Sigma-Aldrich Co. LLC. (St. Louis, MO) while succinic acid, pyruvic acid and NAD<sup>+</sup> were purchased from MP Biomedicals, LLC. (Santa Ana, CA). Limit of detection (LOD) for standards analyzed in a dilution series using reverse phase chromatography is available in Table 1.

One aliquot was analyzed using acidic positive ion conditions (LC pos), chromatographically optimized for more hydrophilic compounds. In this method, the extract was gradient eluted from a C18 column (Waters UPLC BEH C18-2.1 × 100 mm, 1.7 μm) using water and methanol, containing 0.05% perfluoropentanoic acid (PFPA) and 0.1% formic acid (FA) at pH = 2.5. Elution was performed at 0.35 mL min<sup>-1</sup> in a linear gradient from 5% to 80% of methanol containing 0.1% FA and 0.05% PFPA over 3.35 min. A second 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 aforementioned C18 column using methanol 50%, acetonitrile 50%, water, 0.05% PFPA, and 0.01% FA at pH = 2.5 and was operated at an overall higher organic content. Elution was performed at 0.60 mL/min in a linear gradient from 40% to 99.5% over 1 min, hold 2.4 min at 99.5% of methanol 50%, acetonitrile 50%, 0.05% PFPA, and 0.01% FA. A third aliquot was analyzed using basic negative ion-optimized conditions with a separate dedicated C18 column (LC neg). The basic extracts were gradient eluted from the column using methanol 95% and water 5%, with 6.5 mM ammonium bicarbonate at pH 8. Elution was performed at 0.35 mL min<sup>-1</sup> with a linear gradient from 0.5% to 70% of methanol 95%, water 5% with 6.5 mM ammonium bicarbonate over 4 min, followed by a rapid gradient to 99% in 0.5 min. The sample injection volume was 5 μL and a 2 × needle loop overfill was used. Separations utilized separate acid and base-dedicated 2.1 mm × 100 mm Waters BEH C18 1.7 μm columns held at 40 °C. The fourth aliquot was analyzed via negative ionization following elution from an HILIC column (LC HILIC) (Waters UPLC BEH Amide 2.1 × 150 mm, 1.7 μm, held at 40 °C) using a gradient consisting of water (15%), methanol (5%), and acetonitrile (80%) with 10 mM ammonium formate, pH 10.16. Elution flow rate was 0.5 mL/min with a linear gradient from 5% to 50% in 3.5 min, followed by a linear gradient from 50% to 95% in 2 min, of water (50%), acetonitrile (50%) with 10 mM ammonium formate, pH 10.6. The MS analysis alternated between MS and data-dependent MSn scans using dynamic exclusion. The scan range varied slightly between methods but covered 70–1000 *m/z*.

**Quality assurance and quality control (QA/QC).** 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 served as a technical replicate throughout the platform run; extracted water samples served as process blanks; and a cocktail of QC standards (carefully chosen not to interfere with the measurement of endogenous compounds) spiked into every analyzed sample allowed instrument performance monitoring and aided chromatographic alignment. Tables 2–4 describe QC samples and standards. 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 (median RSD = 3–4%). Instruments are calibrated at least weekly in the utilized polarity using thermo and mass accuracy is monitored at the batch level for the internal standards. A batch fails QC if any of the internal standards are more than 5 ppm away from the theoretical mass. Overall process variability was determined by calculating the median RSD for all endogenous metabolites (i.e.,

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

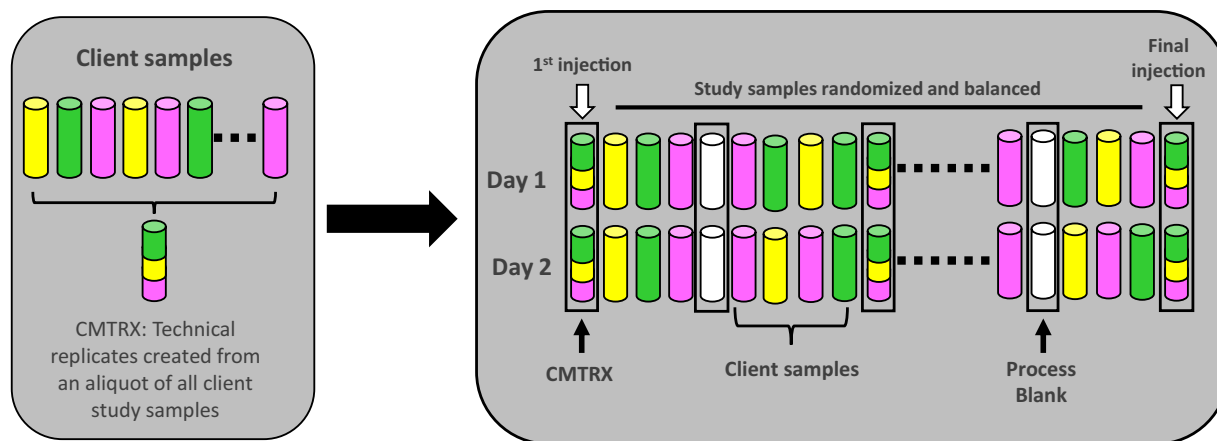
**Table 3.** Metabolon QC standards.

Condition	Internal standards
LC neg	d7-glucose
	d3-methionine
	d3-leucine
	d8-phenylalanine
	d5-tryptophan
	Br-phenylalanine
	d15-octanoic acid
	d19-decanoic acid
	d27-tetradecanoic acid
	d35-octadecanoic acid
	d2-eicosanoic acid
LC HILIC	d35-octadecanoic acid
	d5-indole acetic acid
	Br-phenylalanine
	d5-tryptophan
	d4-tyrosine
	d3-serine
	d3-aspartic acid
	d7-ornithine
	d4-lysine
	LC pos
d3-methionine	
d3-leucine	
d8-phenylalanine	
d5-tryptophan	
Br-phenylalanine	
d4-tyrosine	
d5-indole acetic acid	
d5-hippuric acid	
amitriptyline	
d9-progesterone	
d4-diethylphthalate	

**Table 4.** Quality control internal standards.

non-instrument standards) present in 100% of the pooled matrix (MTRX) samples, which are technical replicates created from a large pool of extensively characterized human plasma. The median RSD for MTRX samples was 9–10%. Five MTRX samples and three process blank samples were processed per every batch of 30 samples. Experimental samples were randomized across the platform run with QC samples spaced evenly among the injections, as outlined in Fig. 1. All studies include the analysis of a technical replicate of a sample pooled from the experimental samples. This pool was analyzed 16 times over the course of the analysis of the experimental samples in the present study.

**Compound identification and quantification.** Raw data were extracted, peak-identified, and QC processed using Metabolon pipelines. Compounds were identified by comparison to library entries of purified standards or recurrent unknown entities<sup>18,19</sup>. Briefly, Metabolon maintains a library based on authenticated standards (analyzed using the same methodology as the experimental samples) 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 (typically within a 5 second window), accurate mass match to the library  $\pm 10$  ppm (typically well within a 5 ppm window), 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



**Fig. 1** Preparation of client-specific technical replicates. A small aliquot of each sample (colored cylinders) is pooled to create a CMTRX technical replicate sample (multi-colored cylinder), which is then injected periodically throughout the platform run. Variability among consistently detected biochemicals can be used to calculate an estimate of overall process and platform variability.

to distinguish and differentiate biochemicals. 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. Peaks were quantified using area-under-the-curve. 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 1 and normalizing each data point proportionately. Metabolites were assigned to pathways based on three publicly available key chemical information resources: PubChem, HMDB and KEGG pathway database.

### Data Records

Raw sequences from the stool samples were deposited at the NCBI's Sequence Read Archive<sup>13</sup>. The following data were compiled and deposited in Figshare<sup>20</sup>:

MetabolitesIDTable: A table containing all detected serum metabolites

- Metabolite\_ID: metabolite identifier
- Metabolite: scientific name of the metabolite. Compounds for which no authentic standards were available for confirmation are marked with an "\*" after the compound name to designate these as Metabolomics Standards Initiative level 2/3<sup>21</sup>. All other metabolites were identified at level 1. The Metabolomics Standards Initiative has provided a consensus classification and notation for the level of confidence in metabolite identification. Level 1 indicates the highest level of confidence in the identity of the compound where at least 2 orthogonal properties of an authentic chemical standard are compared to experimental data acquired in the same laboratory with the same analytical methods. More than 3300 commercially available purified standard compounds have been acquired by Metabolon and registered for analysis on all platforms for determination of their analytical characteristics. Levels 2 and 3 indicate reasonable confidence in metabolite identification despite lack of an authentic standard. Specifically, level 2 (putative identification) reveals probable structure using fragmentation data from literature and/or libraries and databases, while level 3 (tentative structural identification) includes a unique match with data searched through literature and/or libraries and databases. Most such identifications are based on the experimental signature having the same characteristics as the compound class. For example, sphingomyelins all have a conserved fragmentation spectrum and so have a highly diagnostic pattern in order to permit the identification of the experimental signature as a sphingomyelin.
- SuperPathway: superpathway containing the metabolite
- SubPathway: subpathway containing the metabolite
- HMDB: The Human Metabolome Database identifier for the metabolite<sup>22</sup>
- KEGG: The Kyoto Encyclopedia of Genes and Genomes identifier for the metabolite<sup>23</sup>
- PUBCHEM: PubChem identifier for the metabolite<sup>24</sup>

AntibioticsTable: A table containing antibacterial antibiotic exposure data between hospital admission and day 28 of chemotherapy or discharge (whichever occurred first)

- Patient\_ID
- AntibacterialABx: antibiotic name
- StartDayRelativeToD1Chemo: initiation date of the antibiotic relative to day 1 of chemotherapy
- StopDayRelativeToD1Chemo: end date of the antibiotic relative to day 1 of chemotherapy

- Route: route of administration

SerumMetabolitesRawTable: A table containing metabolite levels in serum samples (260 samples from 36 patients)

- Patient\_ID
- SampleDayRelativeToD1Chemo: serum sample collection date relative to day 1 of chemotherapy
- Columns C and later: raw metabolite levels, with column name corresponding to Metabolite\_ID in the MetabolitesIDTable file

SerumMetabolitesNormalizedTable: A table containing batch normalized (by median) metabolite levels in SerumMetabolitesRawTable

StoolSamplesTable: A table containing stool sample data (566 samples from 68 patients)

- Accession\_Number: SRR accession number for each sample
- Patient\_ID
- SampleDayRelativeToD1Chemo: stool sample collection date relative to day 1 of chemotherapy
- Quality: quality control data. A minimum threshold of 1000 copies/mL of 16S rRNA gene quantified by qPCR is considered adequate by the University of Minnesota Genomics Center. Smaller values should be treated with caution.

PatientMetadata: A table containing patient metadata (68 patients)

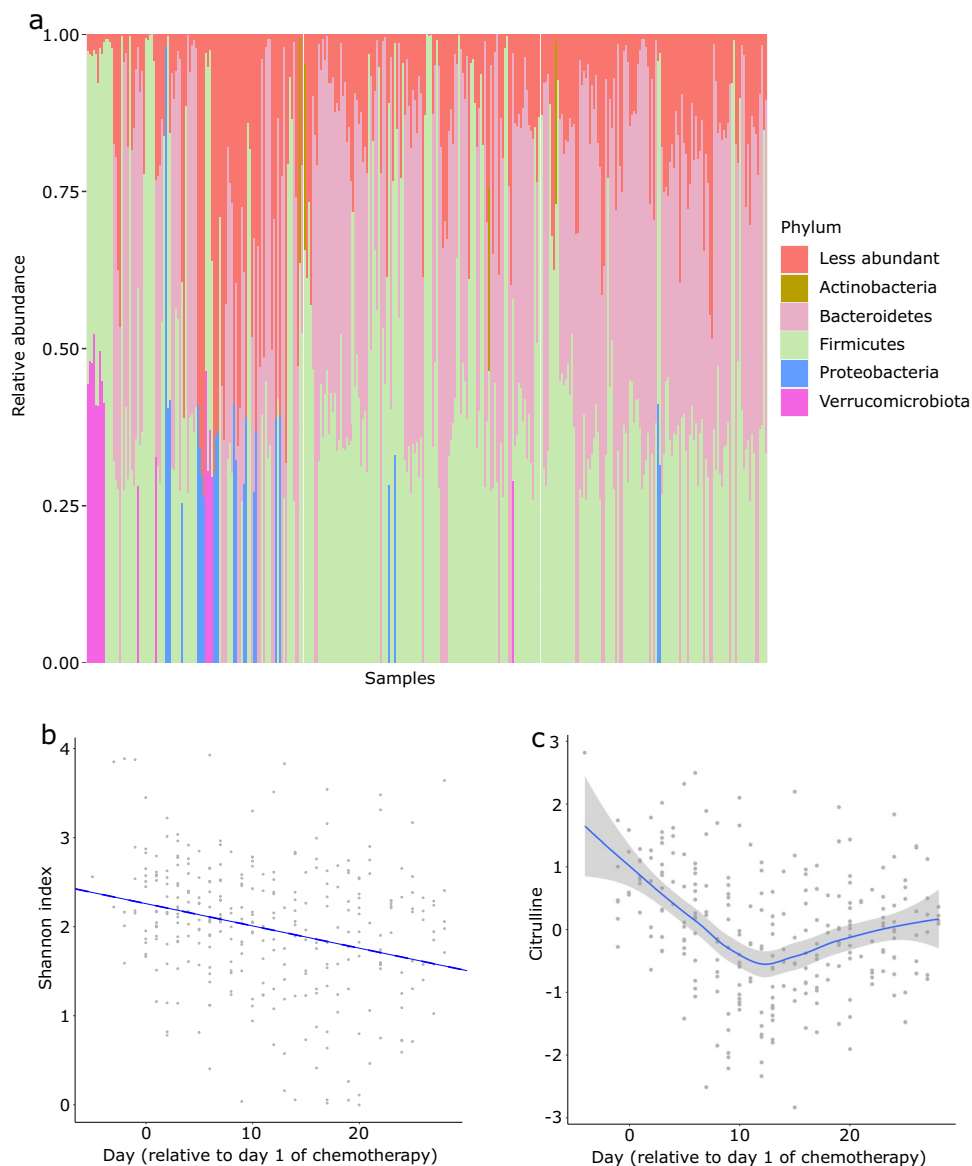
- Patient\_ID
- Age
- Sex
- Disease\_Phase: disease phase
- Chemotherapy: induction chemotherapy regimen.
- FirstNFDayRelativeToD1Chemo: day of the first neutropenic fever relative to day 1 of chemotherapy
- DiarrheaStartDayRelativeToD1Chemo: beginning of diarrhea relative to day 1 of chemotherapy
- DiarrheaEndDayRelativeToD1Chemo: end of diarrhea relative to day 1 of chemotherapy
- FirstCDiffDayRelativeToD1Chemo: day of the first positive test for *Clostridioides difficile* infection relative to day 1 of chemotherapy

## Technical Validation

We have previously published on both the gut microbiome and serum metabolome of these patients as interim analyses<sup>2,4,5,25</sup>. The material provided in this article contains all data from the entire study. For technical validation, we use the subset of patients from whom we analyzed both stool and serum samples. We perform 3 analyses to support the technical quality of data. In the first analysis, we evaluate the microbiota database. In the second analysis, we evaluate the metabolomic database. In the third analysis, we integrate the two databases.

**Alpha diversity in the gut microbiota.** The distribution of the 5 most abundant phyla among the samples is shown in Fig. 2a, with Firmicutes and Bacteroidetes being the two most abundant phyla. Previous studies have shown a decrease in alpha diversity during induction chemotherapy in AML patients<sup>3,26,27</sup>. We evaluated whether our dataset captures this pattern. After removing samples with < 1000 copies/mL of 16S rRNA gene or < 5000 reads, we used scaling with ranked subsampling (SRS package)<sup>28</sup> with normalization to the lowest sequencing depth (5,021 reads) to adjust for sample depth variability. We aggregated ASVs at the genus level. Using the Shannon index<sup>29</sup> to estimate alpha diversity (package *vegan*), a decline in diversity over time was apparent (Fig. 2b). To quantify this decline while accounting for the longitudinal nature of data (*i.e.*, multiple timepoints per patient), we built a linear mixed effect model (*lme4* package in R) in the form of Shannon index ~ (1|patient ID) + day, where patient ID was considered a random effect and day was the sample collection day relative to day 1 of chemotherapy. After controlling for patient ID, there was a significant decline in alpha diversity over time (regression line in Fig. 1a). The regression coefficient for “day” was  $-0.025$ , with a 95% confidence interval of  $-0.034$  to  $-0.017$ , indicating a negative slope ( $p < 0.01$  from 200 bootstraps using *bootMer*).

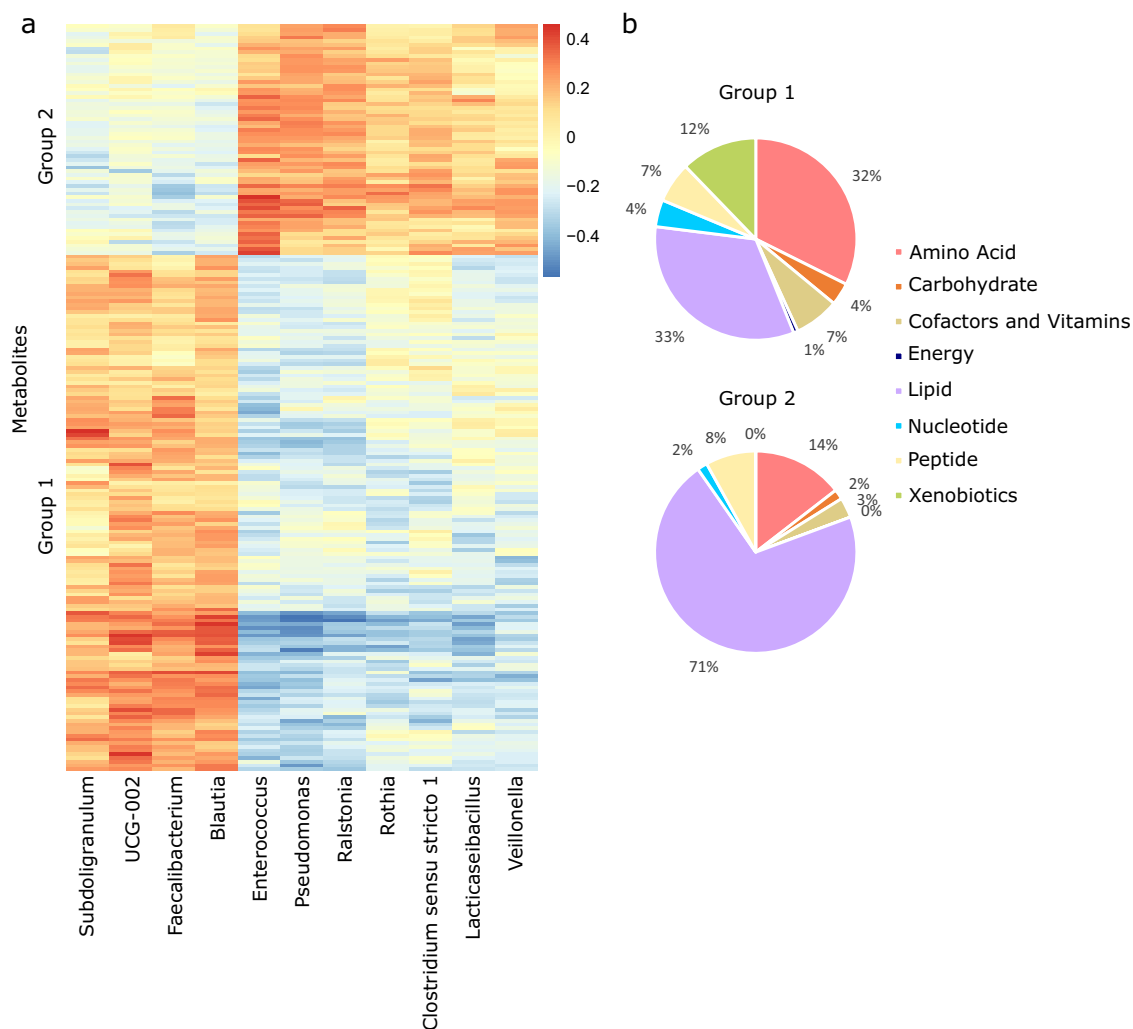
**Citrulline dynamics in the serum metabolome.** Citrulline is an amino acid produced exclusively by intestinal epithelial cells<sup>30</sup>, with circulating levels indicating total functioning enterocyte mass. Citrulline has been established as a biomarker for intestinal epithelial health<sup>30–32</sup>, with lower levels indicating intestinal epithelial cell loss. Between 10–14 days after starting mucotoxic chemotherapy, citrulline levels reach a nadir, followed by gradual recovery<sup>31,33</sup>. We evaluated whether our metabolomic dataset captures this pattern. Preprocessing of the metabolomic database included removing metabolites present in fewer than 50% of the samples and zero imputation using the half minimum of the observed values for each remaining metabolite. Rank-based inverse normal transformation was used for normalization. Using a loess smoother, we found a marked decline in citrulline until about day 12, after which citrulline slowly rose towards baseline (Fig. 2c). Week 2 of chemotherapy corresponds with maximal cytotoxic damage to the intestinal epithelium, thus supporting the validity of our database.



**Fig. 2** Gut microbiota taxonomic distribution, alpha diversity, and serum citrulline dynamics. **(a)** Distribution of the 5 most abundant phyla among samples. **(b)** Shannon index on the gut microbiota over time. The regression line was derived from a mixed effect model with patient ID as a random effect. **(c)** Serum citrulline levels over time. Citrulline levels are after rank-based inverse normal transformation.

**Integration of multi-omics data.** A unique feature of the present study is the multi-omics nature of data in patients with AML. Many microbial metabolites are normally found in the blood<sup>7,8</sup>. Examples include microbial derived metabolites of dietary tyrosine and tryptophan, with important effects on host physiology. We identified 139 same-patient pairs of serum and stool samples, with the stool sample collected within 24 hours prior to the serum sample. Using these pairs, we evaluated whether biologically meaningful connections can be found between the gut microbiota and next-day serum metabolites. We chose a 24 hour interval between samples in each pair to minimize intervening events and to account for the short half-life of many circulating metabolites. Preprocessing of the microbiota database included removing samples with < 1000 copies/mL of 16S rRNA gene or < 5000 reads, removing ASVs present in < 10% of the samples, and removing genera with a mean relative abundance < 0.005. This process yielded 33 genera for further analysis. Taxon abundances were centered log-ratio transformed to account for data compositionality<sup>34</sup>. Preprocessing of the metabolomic database was similar to the previous technical validation using citrulline dynamics and yielded 632 metabolites. The 33 genera were used as predictors of the 632 metabolites in sparse canonical correlation analysis (sCCA, *PMA* package, function *CCA*)<sup>35</sup>.

sCCA is a method to integrate multi-omics datasets with the ability to select more biologically relevant sets of features. sCCA identifies strongly associated metabolite-microbe groups by finding linear combinations of variables from each dataset maximally correlated with each other while simultaneously thresholding variable specific weights to induce sparsity and performing variable selection. This procedure applies L1-penalized matrix decomposition of the cross-product matrix similar to a LASSO regression problem<sup>36</sup>, thus variables are



**Fig. 3** Integrated multi-omics. **(a)** Heatmap correlogram showing Pearson correlation coefficient between each final gut microbiota genus and each final serum metabolite remaining in the final results of sparse canonical correlation analysis. UCG-002 is a genus in the *Oscillospiraceae* family. **(b)** Distribution of metabolites in groups 1 and 2 in panel (a) in different metabolic pathways.

selected based on their importance to the overall microbe-metabolite covariance. As a result, taxa and metabolites with non-zero loading coefficients are those driving the overall correlation between the two datasets. Hyperparameter tuning was done through 50 permutations (*CCA.permute* function) and the best set of penalty values for each dataset were used to fit the final model. An overall correlation coefficient between the two datasets was also estimated using the correlation coefficient in the first canonical variable. The 99% confidence interval for the correlation coefficient was estimated from 1000 bootstraps. Pairwise Pearson correlation coefficients were calculated between each selected genus and metabolite and the correlogram was visualized by a heatmap (*pheatmap* package).

Eleven genera and 201 metabolites drove the association between the gut microbiome and serum metabolome. The overall correlation coefficient between the two datasets was 0.79 (99% confidence interval: 0.71–0.83,  $p < 0.001$ ). The heatmap in Fig. 3a visualizes univariate correlations between these taxa and metabolites, and MicrobiomeMetabolomeHeatmap deposited in Figshare<sup>20</sup> details them in tabular format. There was a clear separation of genera into two groups based on their metabolite associations. The first group contained obligate anaerobic commensal genera in the Clostridia class (*Faecalibacterium*, *Subdoligranulum*, *Blautia*, and an *Oscillospiraceae* genus UCG-002) with a plethora of beneficial effects such as butyrate production and anti-inflammatory properties<sup>37–39</sup>. The second group contained genera with frequently pathogenic species in patients with cancer including *Enterococcus*, *Pseudomonas*, *Rothia*, and *Veillonella*. These 2 groups showed stark differences in the metabolic pathways of their positively associated metabolites (Fig. 3b). Metabolites associated with the first group were enriched in amino acid and xenobiotic pathways, while the second group metabolites were enriched in the lipid pathway. Among metabolites in group 1 were known microbial metabolites of dietary tryptophan (indoleacetate<sup>40</sup>) and tyrosine (p-cresol sulfate<sup>41</sup>) as well as butyrate/isobutyrate.



## Usage Notes

We present repositories for a longitudinal dataset of gut microbiota and serum metabolomics from hospitalized patients with AML receiving induction chemotherapy at the University of Minnesota. These data, combined with the curated clinical metadata presented here, provide a unique opportunity for hypothesis generation. As an example, we illustrated how the microbiome and metabolome datasets can be integrated to identify novel associations for further testing in future studies. To our knowledge, this is the first public, patient- and sample-level, multi-omics database offering the interested user access to raw amplicon sequences, metabolomic data, and detailed clinical metadata in patients with AML receiving induction chemotherapy. In our multi-omics example, and by providing access to the code, we have provided a step-by-step tutorial on how sCCA may be used to integrate microbiome and metabolomics data while accounting for data compositionality and sparsity.

One limitation of this work is the lack of curated dietary data as an important determinant of both gut microbiota<sup>42</sup> and serum metabolome<sup>9,43</sup>. In addition, although antibacterial antibiotic prophylaxis in this patient population utilizes fluoroquinolones in most centers<sup>44</sup>, it is not universal. Therefore, specific patterns of microbiota change may not be generalizable worldwide. Similarly, microbiome-metabolome associations found here are likely not fully generalizable to healthy individuals because of the multitude of insults to the intestinal barrier and gut microbiota in patients with AML. Finally, species-level inferences cannot be reliably made from short amplicon data<sup>45</sup>, a limitation that can be overcome by shotgun sequencing.

## Code availability

The custom R code is available in Figshare<sup>20</sup>.

Received: 14 February 2022; Accepted: 27 July 2022;

Published online: 02 August 2022

## References

- Rashidi, A. *et al.* Gut dysbiosis during antileukemia chemotherapy versus allogeneic hematopoietic cell transplantation. *Cancer* **126**, 1434–1447 (2020).
- Rashidi, A. *et al.* Dysbiosis patterns during re-induction/salvage versus induction chemotherapy for acute leukemia. *Sci. Rep.* **9**, 6083 (2019).
- Galloway-Peña, J. R. *et al.* Gut Microbiome Signatures are Predictive of Infectious Risk Following Induction Therapy for Acute Myeloid Leukemia. *Clin. Infect. Dis.* <https://doi.org/10.1093/cid/ciz777> (2019).
- Rashidi, A. *et al.* Altered microbiota-host metabolic cross talk preceding neutropenic fever in patients with acute leukemia. *Blood Adv* **5**, 3937–3950 (2021).
- Rashidi, A. *et al.* Specific gut microbiota changes heralding bloodstream infection and neutropenic fever during intensive chemotherapy. *Leukemia* **34**, 312–316 (2020).
- Messina, J. A. *et al.* Enterococcus Intestinal Domination is Associated with Increased Mortality in the Acute Leukemia Chemotherapy Population. *Clin. Infect. Dis.* <https://doi.org/10.1093/cid/ciab1043> (2021).
- Wikoff, W. R. *et al.* Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc. Natl. Acad. Sci. USA* **106**, 3698–3703 (2009).
- Visconti, A. *et al.* Interplay between the human gut microbiome and host metabolism. *Nat. Commun.* **10**, 4505 (2019).
- Bar, N. *et al.* A reference map of potential determinants for the human serum metabolome. *Nature* **588**, 135–140 (2020).
- Gerber, G. K. The dynamic microbiome. *FEBS Lett.* **588**, 4131–4139 (2014).
- Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448–454 (2020).
- Kim, D. *et al.* Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* **5**, 52 (2017).
- NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRP141394> (2021).
- Gohl, D. M. *et al.* Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* **34**, 942–949 (2016).
- Al-Ghalith, G. A., Hillmann, B., Ang, K., Shields-Cutler, R. & Knights, D. SHI7 Is a Self-Learning Pipeline for Multipurpose Short-Read DNA Quality Control. *mSystems* **3**, (2018).
- Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**, 581–583 (2016).
- Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).
- DeHaven, C. D., Evans, A. M., Dai, H. & Lawton, K. A. Organization of GC/MS and LC/MS metabolomics data into chemical libraries. *J. Cheminform.* **2**, 9 (2010).
- DeHaven, C. D., Evans, A. M., Dai, H. & Lawton, K. A. Software techniques for enabling high-throughput analysis of metabolomic datasets. *Metabolomics* **10**, 167–192 (2012).
- Rashidi, A. *et al.* Compilation of longitudinal gut microbiome, serum metabolome, and clinical data in acute myeloid leukemia. *figshare* <https://doi.org/10.6084/m9.figshare.c.5835884.v1> (2022).
- Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211–221 (2007).
- Wishart, D. S. *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res. Database* (2007).
- Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
- Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).
- Rashidi, A. *et al.* Gut microbiota response to antibiotics is personalized and depends on baseline microbiota. *Microbiome* **9**, 211 (2021).
- Galloway-Peña, J., Brumlow, C. & Shelburne, S. Impact of the Microbiota on Bacterial Infections during Cancer Treatment. *Trends Microbiol.* **25**, 992–1004 (2017).
- Galloway-Peña, J. R. *et al.* Characterization of oral and gut microbiome temporal variability in hospitalized cancer patients. *Genome Med.* **9**, 21 (2017).
- Heidrich, V., Karlovsky, P. & Beule, L. 'SRS'R Package and 'q2-srs' QIIME 2 Plugin: Normalization of Microbiome Data Using Scaling with Ranked Subsampling (SRS). *NATO Adv. Sci. Inst. Ser. E Appl. Sci.* (2021).
- Shannon, C. E. & Weaver, W. The mathematical theory of communication. *Illinois press, Urbana, I 11*, 117 (1949).
- Demacker, P. N. M. *et al.* Plasma citrulline measurement using UPLC tandem mass-spectrometry to determine small intestinal enterocyte pathology. *Journal of Chromatography B* **877**, 387–392 (2009).
- Hueso, T. *et al.* Impact and consequences of intensive chemotherapy on intestinal barrier and microbiota in acute myeloid leukemia: the role of mucosal strengthening. *Gut Microbes* **12**, 1800897 (2020).

32. van Vliet, M. J. *et al.* Citrulline as a marker for chemotherapy induced mucosal barrier injury in pediatric patients. *Pediatr. Blood Cancer* **53**, 1188–1194 (2009).
33. Blijlevens, N. M. A., Lutgens, L. C. H. W., Schattenberg, A. V. M. B. & Donnelly, J. P. Citrulline: a potentially simple quantitative marker of intestinal epithelial damage following myeloablative therapy. *Bone Marrow Transplant.* **34**, 193–196 (2004).
34. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. & Pawłowsky-Glahn, V. Logratio Analysis and Compositional Distance. *Math. Geol.* **32**, 271–275 (2000).
35. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534 (2009).
36. Cao, D.-S., Liu, S., Zeng, W.-B. & Liang, Y.-Z. Sparse canonical correlation analysis applied to -omics studies for integrative analysis and biomarker discovery. *J. Chemom.* **29**, 371–378 (2015).
37. Duncan, S. H., Barcenilla, A., Stewart, C. S., Pryde, S. E. & Flint, H. J. Acetate utilization and butyryl coenzyme A (CoA):acetate-CoA transferase in butyrate-producing bacteria from the human large intestine. *Appl. Environ. Microbiol.* **68**, 5186–5190 (2002).
38. Holmström, K., Collins, M. D., Møller, T., Falsen, E. & Lawson, P. A. *Subdoligranulum variabile* gen. nov., sp. nov. from human feces. *Anaerobe* **10**, 197–203 (2004).
39. Jenq, R. R. *et al.* Intestinal Blautia Is Associated with Reduced Death from Graft-versus-Host Disease. *Biol. Blood Marrow Transplant.* **21**, 1373–1383 (2015).
40. Agus, A., Planchais, J. & Sokol, H. Gut Microbiota Regulation of Tryptophan Metabolism in Health and Disease. *Cell Host Microbe* **23**, 716–724 (2018).
41. Gryp, T., Vanholder, R., Vanechoutte, M. & Glorieux, G. p-Cresyl Sulfate. *Toxins* **9**, (2017).
42. Johnson, A. J. *et al.* Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host Microbe* **25**, 789–802. e5 (2019).
43. Esko, T. *et al.* Metabolomic profiles as reliable biomarkers of dietary composition. *Am. J. Clin. Nutr.* **105**, 547–554 (2017).
44. Bucaneve, G. *et al.* Levofloxacin to prevent bacterial infection in patients with cancer and neutropenia. *N. Engl. J. Med.* **353**, 977–987 (2005).
45. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029 (2019).

### Acknowledgements

Sequence data from UMN samples were analyzed using the resources of the Minnesota Supercomputing Institute. Serum samples were analyzed for metabolomics by Metabolon, Inc. We thank Dr. Sivapriya Ramamoorthy from Metabolon for her assistance with Figure 1. This work was supported by the National Institutes of Health's National Center for Advancing Translational Sciences grants KL2TR002492 and UL1TR002494. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health's National Center for Advancing Translational Sciences.

### Author contributions

A.R. and C.S. conceived the study and performed the analyses. A.R., M.E., T.U.R., and H.E. collected metadata. T.K., H.H., and C.S. extracted and sequenced DNA. A.R. wrote the manuscript. M.E., D.J.W., S.G.H., A.K., and C.S. provided critical feedback on the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to A.R.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022