Meta2DB: Curated Shotgun Metagenomic Feature Sets and Metadata for Health State Prediction

Car Reen Kok[1], Nisha J. Mulakken[1], James B. Thissen[1], Jose Manuel Martí[1], Ryan Lee[1,2], Jacob B. Trainer[1], Andre R. Goncalves[1], Hiranmayi Ranganathan[1], Aram Avila-Herrera[1], Crystal J. Jaing[1], Nicholas A. Be[1]

[1]Lawrence Livermore National Laboratory, Livermore, CA
[2]University of California, Merced, Merced, CA

Corresponding author: Nicholas A. Be, Ph.D.
Biosciences and Biotechnology Division
Lawrence Livermore National Laboratory
Livermore, CA 94550
Tel: 925-423-1612
Email: be1@llnl.gov

ABSTRACT

Meta2DB is a curated metagenomic and metadata database that provides structurally consistent microbiome taxonomy feature count tables for 13,897 samples across 84 studies, 23 disease states, and 34 geographical locations. All samples were uniformly processed using a streamlined metagenomic classification pipeline that employs a reference database indexed to contain all sequences across all kingdoms of life that were present in the NCBI Nucleotide (nt) database retrieved on Jan 04, 2023. This pipeline leverages high-performance computing (HPC) resources at Lawrence Livermore National Laboratory and was used to process 50TB of publicly available raw metagenomic sequence data. Extensive metadata curation was carried out through a combination of manual curation and automated parsing, producing a consistent inter-study metadata table specifically structured to facilitate training of ML models for prediction of human health.

INTRODUCTION

The human microbiome represents a target with tremendous potential for diagnosing and predicting human health. Microbiome features have been indicated as prognostic factors for a range of disease states, including cancer, gut health disruption, neurodegenerative disease, environmental exposure, and mental health. Development of models that leverage microbiome features for prediction of human health states could facilitate tools with clinically actionable utility and identify microbiome-centric variables for therapeutic intervention.

Machine learning (ML), and specifically deep learning (DL), approaches have demonstrated potential in exploring and classifying microbiomes in the context of human health [1, 2]. Convolutional neural networks (CNNs) have demonstrated the capacity to predict various modes of disease outcomes such as disease severity in ulcerative colitis, cirrhosis and inflammatory bowel disease using microbiome-derived data [3-5].

Microbiome data employed for predictive modeling are often derived from metagenomic sequence data generated from the specimens of interest. Such profiling can involve amplicon sequencing, where a specific gene or region is amplified and assigned to microbial taxonomy, most often the bacterial 16S rRNA sequence. Alternatively, shotgun metagenomic sequencing can be performed, whereby all gDNA within an extracted specimen are sequenced and the resultant sequence reads are aligned to references via a metagenomic classification software platform. The output from such analyses is a feature table, where a taxonomic identifier for a given taxonomic rank (e.g., genus, species) is assigned a numeric value (number of classified

55 sequences), which can then be transformed as desired and used as input for the ML
56 platform of choice.
57 The high-dimensional nature of microbiome data derived from many features (*p*)
58 alongside small sample sizes (*n*) often result in large *p* and small *n* datasets. This
59 presents a critical challenge in ML modeling, resulting in overfitting and limiting
60 generalizability to new or unseen data. One strategy for sample size expansion is to
61 integrate publicly available datasets across different studies. While health-relevant
62 microbiome datasets are becoming increasingly available, dataset integration for use in
63 a cohesive model training approach is difficult as 1) microbiome feature tables are
64 generated by distinct metagenomic classification platforms that impart systematic
65 biases, 2) independent studies implement distinct quality and sequence filtering steps,
66 and 3) metadata are inconsistently and sparsely represented.
67 One avenue for addressing this need is to perform a meta-acquisition of the raw
68 sequencing data corresponding to many parallel studies, and to process all such
69 datasets via one consistent metagenomic classification pipeline. The Human
70 Microbiome Compendium [6] and the curatedMetagenomicData data package [7]
71 implements such an approach providing an excellent resource for consistently
72 structured microbiome profiles across multiple studies using 16S rRNA amplicon
73 sequencing and shotgun metagenomic sequencing respectively. The Human
74 Microbiome Compendium pipeline incorporates the DADA2 algorithm [8] to produce
75 Amplicon Sequence Variants for taxonomic identification. The profiles contained in the
76 curatedMetagenomicData data package are produced via MetaPhlAn3 [9], which
77 employs a marker library reference for assigning shotgun metagenomic data, leveraging

78 the benefits of increased taxonomic resolution of shotgun sequence. The marker library
79 facilitates a computationally efficient workflow, though it is limited to the taxonomic
80 entities for which markers are identified and indexed.

81 Our approach (**Figure 1**) employs a reference database indexed to contain all
82 sequences present in NCBI Nucleotide (nt) from all kingdoms of life [10] (Database
83 available here: https://benlangmead.github.io/aws-indexes/centrifuge). This approach is
84 inherently more comprehensive and high resolution, but computationally intensive. To
85 provide a complementary resource for microbiome-centric model training, we leveraged
86 high-performance computing (HPC) resources at Lawrence Livermore National
87 Laboratory to perform metagenomic classification on 50TB of publicly available raw
88 metagenomic sequence data. We generated structurally consistent microbiome
89 taxonomy feature count tables for 13,897 samples, spanning across 84 studies [11-94].
90 We further performed extensive metadata curation of all assessed studies through a
91 combination of manual curation and automated parsing, producing a consistent inter-
92 study metadata table specifically structured to facilitate training of ML models for
93 prediction of human health. These curated datasets are available here as supplemental
94 material and will soon be hosted on a website (Meta2DB) whereby a user interface
95 supported by relational databases is available for data exploration.
96

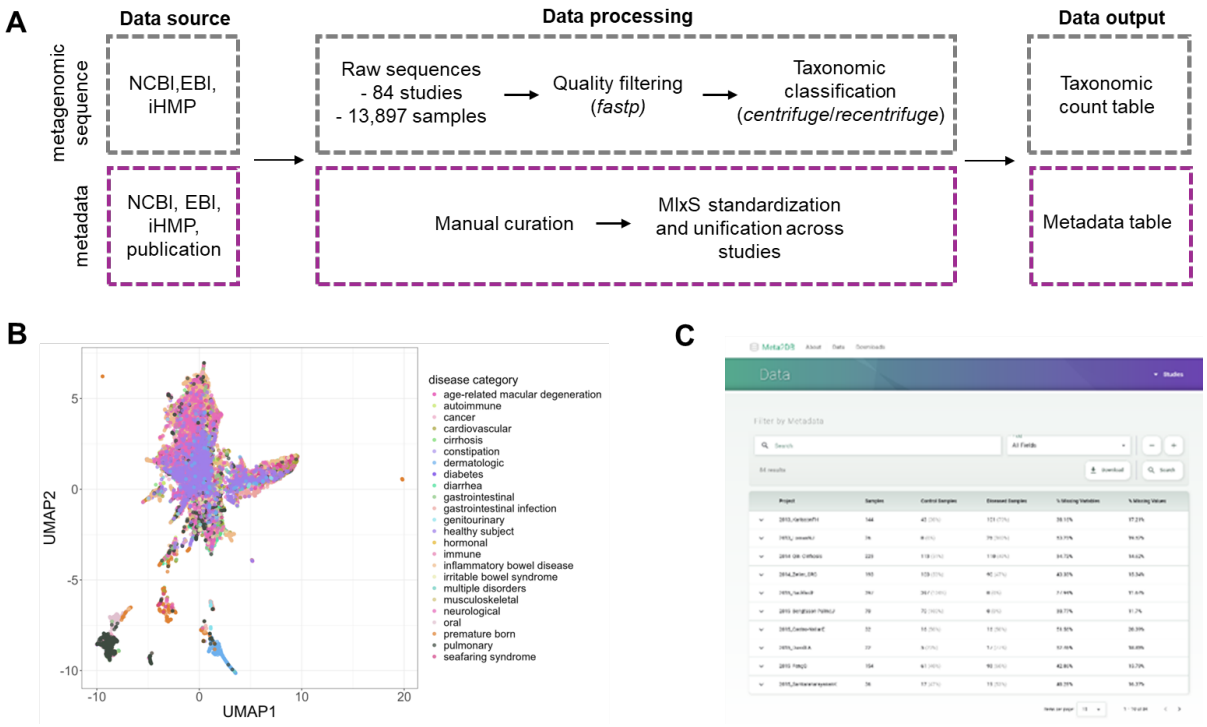

Figure 1. Meta2DB pipeline for the curation of shotgun metagenomic data. A) Data intake, processing and integration of publicly available shotgun metagenomic sequences and associated metadata. B) UMAP plot of 13,897 samples based on computed Euclidean distances. Samples are colored according to different health and disease states. C) Meta2DB web interface for data download and exploration.

METHODS

Selection of studies for meta-analysis

Literature studies were selected for data intake and curation if they met each of the following requirements: 1) biospecimens were processed via whole metagenome short-read sequencing using an Illumina platform, 2) raw sequence data and study metadata were publicly accessible, 3) the study was performed in the context of one or more defined human disease states, and 4) each assessed biospecimen could be assigned one of two binary labels corresponding to “control” or “diseased.”

The data collection process includes leveraging an existing resource of curated microbiome data and conducting a strategic search in PubMed. The curatedMetagenomicData package is a pre-existing, well-curated collection of studies, many of which met the characteristics described above [95]. A subset of studies from this curated package was selected for inclusion. To supplement this list of studies, a literature search was carried out in PubMed to retrieve relevant articles. The following search terms were used in combination; “shotgun”, “metagenomics”, “disease”, and studies were manually curated to ensure that they met the requirements listed above.

These parameters were selected for consistency with the aim of facilitating downstream analyses; it is however acknowledged that future analyses could benefit from incorporation of additional data modalities such as amplicon sequences and long-read sequencing, which will be the subject of future iterations.

Metadata curation

Metadata was extracted from multiple potential sources depending on the individual study of interest, including from the publication and supplementary data or

134 sequence repository. Data sources also included large databases, like the Inflammatory
135 Bowel Disease database of the iHMP, which contains thousands of metagenomics
136 samples from patients with either Crohn's Disease (CD) or ulcerative colitis (UC), as
137 well as control samples. The Genomic Standards Consortium's Minimum Information
138 about any Sequence (MIxS) metadata standards were used to group and standardize
139 patient health, treatment, and demographic metadata across all studies. Controlled
140 vocabulary for defined disorders and health states were used to create consistent
141 formatting and naming. In addition to metadata common to all studies, metadata specific
142 to each disorder was also preserved. Metadata around sample collection and type of
143 assay performed were tracked to account for possible effects on outcomes. To
144 distinguish different types of missing data, "not applicable" was used for fields not
145 relevant to a sample, "not available" was used when the data is relevant and missing,
146 and "none" was used when was used when a field that is a treatment is not applied to
147 the patient or when a health status is not observed for the patient.

148 *Raw sequence acquisition*

149 Raw sequence data (fastq files) were downloaded from public repositories
150 including the NCBI Sequence Read Archive (SRA) [96] and the European Nucleotide
151 Archive (ENA) [97]. The Integrative Human Microbiome Project (iHMP) sequences were
152 obtained from the HMP Data Coordination Center [62].

153 *Sequence pre-processing*

154 Raw sequence data were pre-processed via fastp [98]. These parameters
155 include using a qualified quality phred value of Q15, an unqualified base percentage of
156 40%, and a minimum read length of 15. Individual post-filtering of samples was

157 performed with the following read count thresholds: 5M paired-end reads for fecal
158 samples and 1M paired-end reads for skin, oral, and nasal samples. Samples not
159 meeting these thresholds were removed from downstream analysis. Reads were
160 subsequently aligned to the GRCh38 human reference genome using minimap2 [99] for
161 the removal of human reads.

162 *Metagenomic classification*

163 Resultant reads were processed for metagenomic classification via Centrifuge
164 [100], applying taxonomic labels via Lowest Common Ancestor (LCA) classification
165 strategy. A Minimum Hit Length (MHL) of 15 or 25 was employed for initial Centrifuge
166 processing. Post-processing was performed via Recentrifuge [101] to propagate and
167 normalize Centrifuge assignment scores and perform additional filtering. For purposes
168 of downstream analysis, taxonomic assignments were filtered at a MHL threshold of 40
169 and feature count tables were produced.

170 *UMAP visualization*

171 UMAP dimensionality reduction was applied to all 13,897 samples using the
172 umap (version 0.2.10) package in R (version 4.2.2) with default parameters
173 (n_neighbors=15, n_components=2, n_epochs=200). The local neighborhood structure
174 was constructed using the UMAP algorithm on pairwise Euclidean distances.

175

176 **DATA AVAILABILITY**

177 The processed feature tables and metadata are available here as supplemental
178 material. The Meta2DB website will be available soon and will host all processed
179 feature tables, metadata tables along with example notebooks and scripts for data

intake and analyses. The indexed version of the centrifuge database based on the entirety of NCBI nt is available at https://benlangmead.github.io/aws-indexes/centrifuge for processing of datasets not included in Meta2DB.

COMPETING INTERESTS

Not applicable

FUNDING

This study received internal financial support from an award at Lawrence Livermore National Laboratory.

ACKNOWLEDGEMENTS

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL Disclaimer: This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors

expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

REFERENCES

1. Namkung J. Machine learning methods for microbiome studies. Journal of Microbiology. 2020;58(3):206-16; doi: 10.1007/s12275-020-0066-8.
2. Wang Y, Bhattacharya T, Jiang Y, Qin X, Wang Y, Liu Y, et al. A novel deep learning method for predictive modeling of microbiome data. Briefings in Bioinformatics. 2020;22(3); doi: 10.1093/bib/bbaa073.
3. Fung DLX, Li X, Leung CK, Hu P. A self-knowledge distillation-driven CNN-LSTM model for predicting disease outcomes using longitudinal microbiome data. Bioinformatics Advances. 2023;3(1); doi: 10.1093/bioadv/vbad059.
4. Lo C, Marculescu R. MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks. BMC Bioinformatics. 2019;20(12):314; doi: 10.1186/s12859-019-2833-2.
5. Sharma D, Paterson AD, Xu W. TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction. Bioinformatics. 2020;36(17):4544-50; doi: 10.1093/bioinformatics/btaa542.
6. Abdill RJ, Graham SP, Rubinetti V, Albert FW, Greene CS, Davis S, et al. Integration of 168,000 samples reveals global patterns of the human gut microbiome. bioRxiv. 2023:2023.10.11.560955; doi: 10.1101/2023.10.11.560955.
7. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nature Methods. 2017;14(11):1023-4; doi: 10.1038/nmeth.4468.
8. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods. 2016;13(7):581-3; doi: 10.1038/nmeth.3869.
9. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife. 2021;10:e65088; doi: 10.7554/eLife.65088.
10. Martí JM, Kok CR, Thissen JB, Mulakken NJ, Avila-Herrera A, Jaing CJ, et al. Addressing the dynamic nature of reference data: a new nt database for robust metagenomic classification. bioRxiv. 2024:2024.06.12.598617; doi: 10.1101/2024.06.12.598617.
11. Asnicar F, Berry SE, Valdes AM, Nguyen LH, Piccinno G, Drew DA, et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat Med. 2021;27(2):321-32; doi: 10.1038/s41591-020-01183-8.

12. Asnicar F, Manara S, Zolfo M, Truong DT, Scholz M, Armanini F, et al. Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. mSystems. 2017;2(1); doi: 10.1128/mSystems.00164-16.
13. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe. 2015;17(5):690-703; doi: 10.1016/j.chom.2015.04.004.
14. Bai X, Narayanan A, Skagerberg M, Ceña-Diez R, Giske CG, Strålin K, et al. Characterization of the Upper Respiratory Bacterial Microbiome in Critically Ill COVID-19 Patients. Biomedicines. 2022;10(5):982.
15. Bedarf JR, Hildebrand F, Coelho LP, Sunagawa S, Bahram M, Goeser F, et al. Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients. Genome Med. 2017;9(1):39; doi: 10.1186/s13073-017-0428-y.
16. Bengtsson-Palme J, Angelin M, Huss M, Kjellqvist S, Kristiansson E, Palmgren H, et al. The Human Gut Microbiome as a Transporter of Antibiotic Resistance Genes between Continents. Antimicrob Agents Chemother. 2015;59(10):6551-60; doi: 10.1128/aac.00933-15.
17. Boktor JC, Sharon G, Verhagen Metman LA, Hall DA, Engen PA, Zreloff Z, et al. Integrated Multi-Cohort Analysis of the Parkinson's Disease Gut Metagenome. Mov Disord. 2023;38(3):399-409; doi: 10.1002/mds.29300.
18. Bommana S, Richards G, Kama M, Kodimerla R, Jijakli K, Read TD, et al. Metagenomic Shotgun Sequencing of Endocervical, Vaginal, and Rectal Samples among Fijian Women with and without Chlamydia trachomatis Reveals Disparate Microbial Populations and Function across Anatomic Sites: a Pilot Study. Microbiol Spectr. 2022;10(3):e0010522; doi: 10.1128/spectrum.00105-22.
19. Boolchandani M, Blake KS, Tilley DH, Cabada MM, Schwartz DJ, Patel S, et al. Impact of international travel and diarrhea on gut microbiome and resistome dynamics. Nature Communications. 2022;13(1):7485; doi: 10.1038/s41467-022-34862-w.
20. Brooks B, Olm MR, Firek BA, Baker R, Thomas BC, Morowitz MJ, et al. Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. Nat Commun. 2017;8(1):1814; doi: 10.1038/s41467-017-02018-w.
21. Byrd AL, Deming C, Cassidy SKB, Harrison OJ, Ng WI, Conlan S, et al. Staphylococcus aureus and Staphylococcus epidermidis strain diversity underlying pediatric atopic dermatitis. Sci Transl Med. 2017;9(397); doi: 10.1126/scitranslmed.aal4651.
22. Castro-Nallar E, Bendall ML, Pérez-Losada M, Sabuncyan S, Severance EG, Dickerson FB, et al. Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. PeerJ. 2015;3:e1140; doi: 10.7717/peerj.1140.
23. Chang H-W, Yan D, Singh R, Bui A, Lee K, Truong A, et al. Multiomic Analysis of the Gut Microbiome in Psoriasis Reveals Distinct Host–Microbe Associations. JID Innovations. 2022;2(3):100115; doi: 10.1016/j.xjidi.2022.100115.

24. Chng KR, Tay AS, Li C, Ng AH, Wang J, Suri BK, et al. Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. Nat Microbiol. 2016;1(9):16106; doi: 10.1038/nmicrobiol.2016.106.
25. David LA, Weil A, Ryan ET, Calderwood SB, Harris JB, Chowdhury F, et al. Gut microbial succession follows acute secretory diarrhea in humans. mBio. 2015;6(3):e00381-15; doi: 10.1128/mBio.00381-15.
26. Deshpande NP, Riordan SM, Castaño-Rodríguez N, Wilkins MR, Kaakoush NO. Signatures within the esophageal microbiome are associated with host genetics, age, and disease. Microbiome. 2018;6(1):227; doi: 10.1186/s40168-018-0611-4.
27. Espinoza JL, Harkins DM, Torralba M, Gomez A, Highlander SK, Jones MB, et al. Supragingival Plaque Microbiome Ecology and Functional Potential in the Context of Health and Disease. mBio. 2018;9(6); doi: 10.1128/mBio.01631-18.
28. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun. 2015;6:6528; doi: 10.1038/ncomms7528.
29. Ferreiro AL, Choi J, Ryou J, Newcomer EP, Thompson R, Bollinger RM, et al. Gut microbiome composition may be an indicator of preclinical Alzheimer's disease. Science Translational Medicine. 2023;15(700):eabo2984; doi: doi:10.1126/scitranslmed.abo2984.
30. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol. 2019;4(2):293-305; doi: 10.1038/s41564-018-0306-4.
31. Ghensi P, Manghi P, Zolfo M, Armanini F, Pasolli E, Bolzan M, et al. Strong oral plaque microbiome signatures for dental implant diseases identified by strain-resolution metagenomics. NPJ Biofilms Microbiomes. 2020;6(1):47; doi: 10.1038/s41522-020-00155-7.
32. Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews MC, Karpinets TV, et al. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. Science. 2018;359(6371):97-103; doi: 10.1126/science.aan4236.
33. Guillén Y, Noguera-Julian M, Rivera J, Casadellà M, Zevin AS, Rocafort M, et al. Low nadir CD4+ T-cell counts predict gut dysbiosis in HIV-1 infection. Mucosal Immunol. 2019;12(1):232-46; doi: 10.1038/s41385-018-0083-7.
34. Guo C, Che X, Briese T, Ranjan A, Allicock O, Yates RA, et al. Deficient butyrate-producing capacity in the gut microbiome is associated with bacterial network disturbances and fatigue symptoms in ME/CFS. Cell Host Microbe. 2023;31(2):288-304.e8; doi: 10.1016/j.chom.2023.01.004.
35. Gupta A, Dhakan DB, Maji A, Saxena R, P KV, Mahajan S, et al. Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. mSystems. 2019;4(6); doi: 10.1128/mSystems.00438-19.
36. Hall AB, Yassour M, Sauk J, Garner A, Jiang X, Arthur T, et al. A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. Genome Med. 2017;9(1):103; doi: 10.1186/s13073-017-0490-5.

37. Hannigan GD, Duhaime MB, Ruffin MTt, Koumpouras CC, Schloss PD. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. mBio. 2018;9(6); doi: 10.1128/mBio.02248-18.
38. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016;2:16180; doi: 10.1038/nmicrobiol.2016.180.
39. Hu Y, Feng Y, Wu J, Liu F, Zhang Z, Hao Y, et al. The Gut Microbiome Signatures Discriminate Healthy From Pulmonary Tuberculosis Patients. Front Cell Infect Microbiol. 2019;9:90; doi: 10.3389/fcimb.2019.00090.
40. Jie Z, Xia H, Zhong SL, Feng Q, Li S, Liang S, et al. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun. 2017;8(1):845; doi: 10.1038/s41467-017-00900-1.
41. Jo S, Kang W, Hwang YS, Lee SH, Park KW, Kim MS, et al. Oral and gut dysbiosis leads to functional alterations in Parkinson's disease. NPJ Parkinsons Dis. 2022;8(1):87; doi: 10.1038/s41531-022-00351-6.
42. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013;498(7452):99-103; doi: 10.1038/nature12198.
43. Khachatryan L, Xiang Y, Ivanov A, Glaab E, Graham G, Granata I, et al. Results and lessons learned from the sbv IMPROVER metagenomics diagnostics for inflammatory bowel disease challenge. Sci Rep. 2023;13(1):6303; doi: 10.1038/s41598-023-33050-0.
44. Kieser S, Sarker SA, Sakwinska O, Foata F, Sultana S, Khan Z, et al. Bangladeshi children with acute diarrhoea show faecal microbiomes with increased Streptococcus abundance, irrespective of diarrhoea aetiology. Environ Microbiol. 2018;20(6):2256-69; doi: 10.1111/1462-2920.14274.
45. Laske C, Müller S, Preische O, Ruschil V, Munk MHJ, Honold I, et al. Signature of Alzheimer's Disease in Intestinal Microbiome: Results From the AlzBiom Study. Front Neurosci. 2022;16:792996; doi: 10.3389/fnins.2022.792996.
46. Li J, Zhao F, Wang Y, Chen J, Tao J, Tian G, et al. Gut microbiota dysbiosis contributes to the development of hypertension. Microbiome. 2017;5(1):14; doi: 10.1186/s40168-016-0222-x.
47. Lin B, Wang M, Gao R, Ye Z, Yu Y, He W, et al. Characteristics of Gut Microbiota in Patients with GH-Secreting Pituitary Adenoma. Microbiol Spectr. 2022;10(1):e0042521; doi: 10.1128/spectrum.00425-21.
48. Liu Q, Mak JWY, Su Q, Yeoh YK, Lui GC, Ng SSS, et al. Gut microbiota dynamics in a prospective cohort of patients with post-acute COVID-19 syndrome. Gut. 2022;71(3):544-52; doi: 10.1136/gutjnl-2021-325989.
49. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ, Quick J, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. Jama. 2013;309(14):1502-10; doi: 10.1001/jama.2013.3231.
50. Loomba R, Seguritan V, Li W, Long T, Klitgord N, Bhatt A, et al. Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metab. 2017;25(5):1054-62.e5; doi: 10.1016/j.cmet.2017.04.001.

51. Mancabelli L, Milani C, Lugli GA, Turroni F, Mangifesta M, Viappiani A, et al. Unveiling the gut microbiota composition and functionality associated with constipation through metagenomic analyses. Sci Rep. 2017;7(1):9879; doi: 10.1038/s41598-017-10663-w.
52. Matson V, Fessler J, Bao R, Chongsuwat T, Zha Y, Alegre ML, et al. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. Science. 2018;359(6371):104-8; doi: 10.1126/science.aao3290.
53. Maya-Lucas O, Murugesan S, Nirmalkar K, Alcaraz LD, Hoyo-Vadillo C, Pizano-Zárate ML, et al. The gut microbiome of Mexican children affected by obesity. Anaerobe. 2019;55:11-23; doi: 10.1016/j.anaerobe.2018.10.009.
54. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems. 2018;3(3):10.1128/msystems.00031-18; doi: doi:10.1128/msystems.00031-18.
55. Nagata N, Nishijima S, Kojima Y, Hisada Y, Imbe K, Miyoshi-Akiyama T, et al. Metagenomic Identification of Microbial Signatures Predicting Pancreatic Cancer From a Multinational Study. Gastroenterology. 2022;163(1):222-38; doi: 10.1053/j.gastro.2022.03.054.
56. Nagy-Szakal D, Williams BL, Mishra N, Che X, Lee B, Bateman L, et al. Fecal metagenomic profiles in subgroups of patients with myalgic encephalomyelitis/chronic fatigue syndrome. Microbiome. 2017;5(1):44; doi: 10.1186/s40168-017-0261-y.
57. Oh S, Yap GC, Hong PY, Huang CH, Aw MM, Shek LP, et al. Immune-modulatory genomic properties differentiate gut microbiota of infants with and without eczema. PLoS One. 2017;12(10):e0184955; doi: 10.1371/journal.pone.0184955.
58. Olm MR, Brown CT, Brooks B, Firek B, Baker R, Burstein D, et al. Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. Genome Res. 2017;27(4):601-12; doi: 10.1101/gr.213256.116.
59. Pienkowska K, Pust MM, Gessner M, Gaedcke S, Thavarasa A, Rosenboom I, et al. The Cystic Fibrosis Upper and Lower Airway Metagenome. Microbiol Spectr. 2023;11(2):e0363322; doi: 10.1128/spectrum.03633-22.
60. Piper HG, Fan D, Coughlin LA, Ho EX, McDaniel MM, Channabasappa N, et al. Severe Gut Microbiota Dysbiosis Is Associated With Poor Growth in Patients With Short Bowel Syndrome. JPEN J Parenter Enteral Nutr. 2017;41(7):1202-12; doi: 10.1177/0148607116658762.
61. Polster SP, Sharma A, Tanes C, Tang AT, Mericko P, Cao Y, et al. Permissive microbiome characterizes human subjects with a neurovascular disease cavernous angioma. Nat Commun. 2020;11(1):2659; doi: 10.1038/s41467-020-16436-w.
62. Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, Zhou W, et al. The Integrative Human Microbiome Project. Nature. 2019;569(7758):641-8; doi: 10.1038/s41586-019-1238-8.

63. Pust MM, Wiehlmann L, Davenport C, Rudolf I, Dittrich AM, Tümmler B. The human respiratory tract microbial community structures in healthy and cystic fibrosis infants. NPJ Biofilms Microbiomes. 2020;6(1):61; doi: 10.1038/s41522-020-00171-7.
64. Qi X, Yun C, Sun L, Xia J, Wu Q, Wang Y, et al. Gut microbiota-bile acid-interleukin-22 axis orchestrates polycystic ovary syndrome. Nat Med. 2019;25(8):1225-33; doi: 10.1038/s41591-019-0509-0.
65. Qian Y, Yang X, Xu S, Huang P, Li B, Du J, et al. Gut metagenomics-derived genes as potential biomarkers of Parkinson's disease. Brain. 2020;143(8):2474-89; doi: 10.1093/brain/awaa201.
66. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014;513(7516):59-64; doi: 10.1038/nature13568.
67. Rosa BA, Supali T, Gankpala L, Djuardi Y, Sartono E, Zhou Y, et al. Differential human gut microbiome assemblages during soil-transmitted helminth infections in Indonesia and Liberia. Microbiome. 2018;6(1):33; doi: 10.1186/s40168-018-0416-5.
68. Rubel MA, Abbas A, Taylor LJ, Connell A, Tanes C, Bittinger K, et al. Lifestyle and the presence of helminths is associated with gut microbiome composition in Cameroonians. Genome Biol. 2020;21(1):122; doi: 10.1186/s13059-020-02020-4.
69. Sankaranarayanan K, Ozga AT, Warinner C, Tito RY, Obregon-Tito AJ, Xu J, et al. Gut Microbiome Diversity among Cheyenne and Arapaho Individuals from Western Oklahoma. Curr Biol. 2015;25(24):3161-9; doi: 10.1016/j.cub.2015.10.060.
70. Shi B, Chang M, Martin J, Mitreva M, Lux R, Klokkevold P, et al. Dynamic changes in the subgingival microbiome and their potential for diagnosis and prognosis of periodontitis. mBio. 2015;6(1):e01926-14; doi: 10.1128/mBio.01926-14.
71. Sun H, Guo Y, Wang H, Yin A, Hu J, Yuan T, et al. Gut commensal Parabacteroides distasonis alleviates inflammatory arthritis. Gut. 2023;72(9):1664-77; doi: 10.1136/gutjnl-2022-327756.
72. Sun Z, Zhang M, Li M, Bhaskar Y, Zhao J, Ji Y, et al. Interactions between Human Gut Microbiome Dynamics and Sub-Optimal Health Symptoms during Seafaring Expeditions. Microbiol Spectr. 2022;10(1):e0092521; doi: 10.1128/spectrum.00925-21.
73. Tay ASL, Li C, Nandi T, Chng KR, Andiappan AK, Mettu VS, et al. Atopic dermatitis microbiomes stratify into ecologic dermotypes enabling microbial virulence and disease severity. J Allergy Clin Immunol. 2021;147(4):1329-40; doi: 10.1016/j.jaci.2020.09.031.
74. Ventura RE, Iizumi T, Battaglia T, Liu M, Perez-Perez GI, Herbert J, et al. Gut microbiome of treatment-naïve MS patients of different ethnicities early in disease course. Sci Rep. 2019;9(1):16396; doi: 10.1038/s41598-019-52894-z.
75. Vincent C, Miller MA, Edens TJ, Mehrotra S, Dewar K, Manges AR. Bloom and bust: intestinal microbiota dynamics in response to hospital exposures and

469 Clostridium difficile colonization or infection. Microbiome. 2016;4:12; doi:
470 10.1186/s40168-016-0156-3.
471 76. Wallen ZD, Demirkan A, Twa G, Cohen G, Dean MN, Standaert DG, et al.
472 Metagenomics of Parkinson's disease implicates the gut microbiome in multiple
473 disease mechanisms. Nature Communications. 2022;13(1):6958; doi:
474 10.1038/s41467-022-34667-x.
475 77. Wang Z, Liang L, Liu L, Wang Z, Wang Y, Yu Z, et al. Changes in the Gut
476 Microbiome Associated with Intussusception in Patients with Peutz-Jeghers
477 Syndrome. Microbiol Spectr. 2023;11(2):e0281922; doi:
478 10.1128/spectrum.02819-22.
479 78. Wen C, Zheng Z, Shao T, Liu L, Xie Z, Le Chatelier E, et al. Quantitative
480 metagenomics reveals unique gut microbiome biomarkers in ankylosing
481 spondylitis. Genome Biol. 2017;18(1):142; doi: 10.1186/s13059-017-1271-6.
482 79. Weng YJ, Gan HY, Li X, Huang Y, Li ZC, Deng HM, et al. Correlation of diet,
483 microbiota and metabolite networks in inflammatory bowel disease. J Dig Dis.
484 2019;20(9):447-59; doi: 10.1111/1751-2980.12795.
485 80. Xiong R, Gunter C, Fleming E, Vernon SD, Bateman L, Unutmaz D, et al. Multi-
486 'omics of gut microbiome-host interactions in short- and long-term myalgic
487 encephalomyelitis/chronic fatigue syndrome patients. Cell Host Microbe.
488 2023;31(2):273-87.e5; doi: 10.1016/j.chom.2023.01.001.
489 81. Xiong Z, Peng K, Song S, Zhu Y, Gu J, Huang C, et al. Cerebral
490 Intraparenchymal Hemorrhage Changes Patients' Gut Bacteria Composition and
491 Function. Front Cell Infect Microbiol. 2022;12:829491; doi:
492 10.3389/fcimb.2022.829491.
493 82. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al.
494 Metagenomic and metabolomic analyses reveal distinct stage-specific
495 phenotypes of the gut microbiota in colorectal cancer. Nat Med. 2019;25(6):968-
496 76; doi: 10.1038/s41591-019-0458-7.
497 83. Yan Q, Gu Y, Li X, Yang W, Jia L, Chen C, et al. Alterations of the Gut
498 Microbiome in Hypertension. Front Cell Infect Microbiol. 2017;7:381; doi:
499 10.3389/fcimb.2017.00381.
500 84. Ye Z, Zhang N, Wu C, Zhang X, Wang Q, Huang X, et al. A metagenomic study
501 of the gut microbiome in Behcet's disease. Microbiome. 2018;6(1):135; doi:
502 10.1186/s40168-018-0520-6.
503 85. Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic analysis
504 of faecal microbiome as a tool towards targeted non-invasive biomarkers for
505 colorectal cancer. Gut. 2017;66(1):70-8; doi: 10.1136/gutjnl-2015-309800.
506 86. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of
507 fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol.
508 2014;10(11):766; doi: 10.15252/msb.20145645.
509 87. Zhang F, Wan Y, Zuo T, Yeoh YK, Liu Q, Zhang L, et al. Prolonged Impairment
510 of Short-Chain Fatty Acid and L-Isoleucine Biosynthesis in Gut Microbiome in
511 Patients With COVID-19. Gastroenterology. 2022;162(2):548-61.e4; doi:
512 10.1053/j.gastro.2021.10.013.

88. Zhou T, Wu J, Zeng Y, Li J, Yan J, Meng W, et al. SARS-CoV-2 triggered oxidative stress and abnormal energy metabolism in gut microbiota. MedComm (2020). 2022;3(1):e112; doi: 10.1002/mco2.112.
89. Zhu F, Ju Y, Wang W, Wang Q, Guo R, Ma Q, et al. Metagenome-wide association of gut microbiome features for schizophrenia. Nat Commun. 2020;11(1):1612; doi: 10.1038/s41467-020-15457-9.
90. Zhu Q, Hou Q, Huang S, Ou Q, Huo D, Vázquez-Baeza Y, et al. Compositional and genetic alterations in Graves' disease gut microbiome reveal specific diagnostic biomarkers. Isme j. 2021;15(11):3399-411; doi: 10.1038/s41396-021-01016-7.
91. Zinkernagel MS, Zysset-Burri DC, Keller I, Berger LE, Leichtle AB, Largiadèr CR, et al. Association of the Intestinal Microbiome with the Development of Neovascular Age-Related Macular Degeneration. Sci Rep. 2017;7:40826; doi: 10.1038/srep40826.
92. Zuo T, Zhang F, Lui GCY, Yeoh YK, Li AYL, Zhan H, et al. Alterations in Gut Microbiota of Patients With COVID-19 During Time of Hospitalization. Gastroenterology. 2020;159(3):944-55.e8; doi: 10.1053/j.gastro.2020.05.048.
93. Zuo W, Wang B, Bai X, Luan Y, Fan Y, Michail S, et al. 16S rRNA and metagenomic shotgun sequencing data revealed consistent patterns of gut microbiome signature in pediatric ulcerative colitis. Sci Rep. 2022;12(1):6421; doi: 10.1038/s41598-022-07995-7.
94. Zysset-Burri DC, Keller I, Berger LE, Largiadèr CR, Wittwer M, Wolf S, et al. Associations of the intestinal microbiome with the complement system in neovascular age-related macular degeneration. NPJ Genom Med. 2020;5:34; doi: 10.1038/s41525-020-00141-0.
95. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nat Methods. 2017;14(11):1023-4; doi: 10.1038/nmeth.4468.
96. Leinonen R, Sugawara H, Shumway M, Collaboration obotINSD. The Sequence Read Archive. Nucleic Acids Research. 2010;39(suppl_1):D19-D21; doi: 10.1093/nar/gkq1019.
97. Yuan D, Ahamed A, Burgin J, Cummins C, Devraj R, Gueye K, et al. The European Nucleotide Archive in 2023. Nucleic Acids Research. 2023;52(D1):D92-D7; doi: 10.1093/nar/gkad1067.
98. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i90; doi: 10.1093/bioinformatics/bty560.
99. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100; doi: 10.1093/bioinformatics/bty191.
100. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26(12):1721-9; doi: 10.1101/gr.210641.116.
101. Marti JM. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. PLoS Comput Biol. 2019;15(4):e1006967; doi: 10.1371/journal.pcbi.1006967.