Nucleic Acids Research, 2026, **54**, D734–D742
https://doi.org/10.1093/nar/gkaf1190
Advance access publication date: 24 November 2025
Database issue

GMrepo v3: a curated human gut microbiome database with expanded disease coverage and enhanced cross-dataset biomarker analysis

Can Liu [1,†], Xiaohan Wang [1,†], Zhilin Zhang [1,†], Wenboxin Wang [1], Tianci Wang [1], Yuchen Zhao [1], Mingyu Wang [1], Wei-Hua Chen [1,2,*]

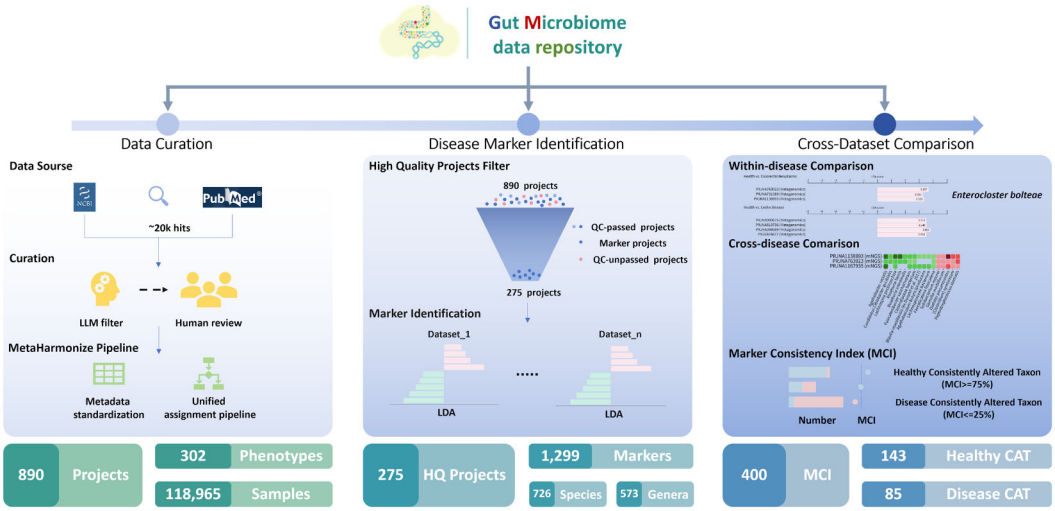
[1]Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-Imaging, Center for Artificial Intelligence Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
[2]School of Biological Science, Jining Medical University, Rizhao 272111, China

*To whom correspondence should be addressed. Email: weihuachen@hust.edu.cn
†The first three authors should be regarded as Joint First Authors.

Abstract

GMrepo (Gut Microbiome Data Repository) is a curated and consistently annotated database of human gut metagenomes, designed to improve data reusability and enable cross-project and cross-disease comparisons. In this latest release, GMrepo v3 has been expanded to 890 projects and 118 965 runs/samples, including 87 048 16S rRNA and 31 917 metagenomic datasets. The number of annotated diseases has increased from 133 to 302, allowing more comprehensive disease-related microbiome analyses. We systematically identified microbial markers between phenotype pairs (e.g. healthy versus diseased) at the project level and compared them across datasets to detect reproducible signatures. As of this release, GMrepo v3 includes 1299 marker taxa (726 species and 573 genera) associated with 167 phenotype pairs, derived from 275 carefully curated projects. To assess marker stability, we developed the Marker Consistency Index (MCI), which summarizes the prevalence and directional consistency of markers across studies. Among 400 markers showing altered abundances in ≥10 projects, 143 were consistently enriched in healthy controls (MCI > 75%), while 85 were enriched in diseases (MCI < 25%). A marker-centric interface enables users to explore marker behavior across diseases. The GMrepo v3 database is freely accessible at https://gmrepo.humangut.info.

Graphical abstract



Received: September 9, 2025. **Revised:** October 12, 2025. **Accepted:** October 13, 2025

Introduction

The human gut microbiota plays a fundamental role in host physiology [1–8], influencing immune function [9, 10], metabolism [11, 12], energy balance [13, 14], and disease susceptibility [15–17] across the lifespan. Its composition and function are shaped by a range of host and environmental factors, including age [18], sex [19], body mass index (BMI) [20], geography [21], diet [4], antibiotic use [22], and host genetics [23]. As a result, understanding the human gut microbiome has become a central focus in biomedical research [24, 25], driving numerous large-scale metagenomic studies based on 16S rRNA gene amplicon and whole-genome shotgun (WGS) sequencing.

Although raw sequencing data are commonly deposited in public repositories such as the European Nucleotide Archive (ENA) [26] and the NCBI Sequence Read Archive (SRA) [27], and curated platforms such as MGnify [28], gcMeta [29], and Qiita [30] provide processed outputs, key limitations remain. In particular, inconsistent or incomplete phenotype and metadata annotations hinder data reusability and large-scale comparative analyses. Our previous study [31, 32] revealed that nearly 30% of publicly available gut microbiome samples lack essential metadata such as age, sex, or BMI, even after extensive manual curation. While curatedMetagenomicData [33] has addressed part of this issue by offering taxonomic and functional profiles in an R package, its utility is limited for users lacking programming expertise, and it does not support large-scale cross-project comparisons of disease-associated microbiota.

To address these limitations, we developed GMrepo (Gut Microbiome Data Repository), a curated, phenotype-centered database of human gut microbiome datasets. GMrepo v1 [31] introduced a unified pipeline for metadata curation and taxonomic annotation, with samples organized by phenotype to facilitate targeted search and comparison. GMrepo v2 [32] expanded the database to 71 642 samples, introduced project-level disease marker identification, and enabled cross-project marker comparisons, including a marker-centric view to assess specificity and reproducibility of microbial signatures.

Here, we present GMrepo v3, a major update that substantially expands the scope of the resource. This version includes 890 newly curated human gut microbiome projects, comprising 118 965 samples (both 16S and WGS), and covering 301 diseases. Compared to GMrepo v2 (353 projects, 71 642 samples), v3 represents a significant enhancement in data volume and phenotypic diversity. In GMrepo v3, the functionality for microbial marker discovery has also been strengthened. In GMrepo v3, we developed the Marker Consistency Index (MCI) to quantify the prevalence and directional stability of microbial markers across phenotype comparisons. Specifically, MCI measures the proportion of comparisons in which a species is enriched in health, providing a straightforward metric to identify markers consistently associated with health or disease. This is particularly useful for interpreting markers in the context of multimorbidity, where a species may appear in multiple disease comparisons with varying patterns. A total of 1299 disease-associated taxa (573 genera and 726 species) were identified across 167 phenotype pairs from 275 rigorously selected projects.

Materials and methods

Data collection and metadata curation

To assemble a comprehensive resource for human gut microbiome studies, we first queried the NCBI BioProject [34] (https://www.ncbi.nlm.nih.gov/bioproject) and PubMed (https://pubmed.ncbi.nlm.nih.gov/) databases using multiple keyword combinations, including "human gut microbiota," "human metagenome," "feces metagenome," "gut metagenome," and the taxonomic identifier txid408170 for Homo sapiens. The systematic search and screening were conducted between September 2024 and December 2024, capturing all publicly available human gut microbiome projects up to the end of 2024. This resulted in the retrieval of ∼20 000 candidate projects. For each project, textual metadata, including project name, title, and description, were extracted and evaluated to determine whether the study included human gut (especially fecal) microbiome samples.

To improve the efficiency and scalability of this screening process, we applied a large language model (ERNIE-3.5-128K) via the Baidu Qianfan platform (Fig. 1). The model was prompted to determine whether each project was likely to contain human gut-related samples based on the textual metadata. The prompt was formulated to enforce binary (yes/no) decisions in natural language, helping reduce ambiguity. Projects predicted as "relevant" by the model were then manually reviewed by curators to confirm eligibility, correct potential misclassifications, and extract phenotype and metadata information. ERNIE-3.5-128k was used solely as an initial triage tool to highlight potentially relevant projects. The outputs were not retained; all final inclusion decisions were made through manual review by curators. Among roughly 20 000 candidate projects, ERNIE-3.5-128k classified ∼30% as "yes" (potentially human gut–related) and 70% as "no." Manual curation of the "yes" subset showed that ∼10% of those indeed met the inclusion criteria for GMrepo v3. Projects predicted as "no" were rarely relevant, confirming that the model's negative-class precision was very high. To ensure unambiguous binary outputs, we used a concise instruction equivalent to: "Based on the project title and abstract, decide whether this study contains human gut (including fecal) samples. Answer only "yes" or "no," without explanation". The actual prompt was implemented in Chinese to minimize ambiguity and enforce strictly one-character responses. All projects flagged "yes" underwent full manual verification, and only curator-confirmed studies were retained.

For all included projects, we retrieved the associated raw sequencing data from the NCBI SRA (https://www.ncbi.nlm.nih.gov/sra) and the ENA (https://www.ebi.ac.uk/ena), using in-house Python (v3.10.0) and R (v4.3.0) scripts. Two rounds of manual metadata curation were then performed to ensure data quality. In the first round, metadata related to sequencing technology (e.g. platform, data type—16S rRNA amplicon or WGS sequencing, and read counts) and host information (e.g. phenotype, age, sex, BMI, country, diet, and antibiotic usage) were extracted and manually annotated. Disease terms were standardized using the controlled vocabulary from the Medical Subject Headings (MeSH) database (https://meshb.nlm.nih.gov/) to ensure consistency across projects. In the second round, independent curators reviewed and verified the annotations to ensure accuracy and consistency across projects.

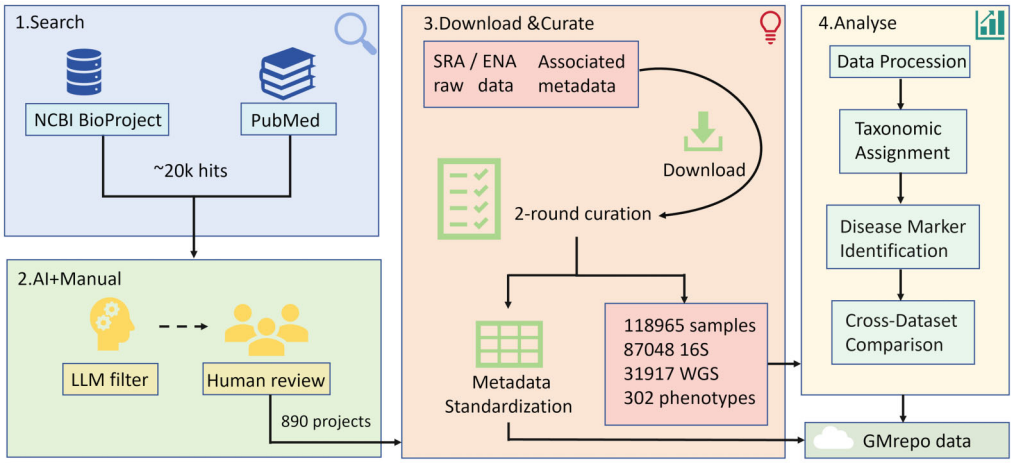


Figure 1. Overview of the GMrepo v3 database workflow.

Taxonomic assignment and relative abundance calculation

The newly collected raw sequencing data were processed following the same general workflow as in the previous version of GMrepo [31], with updated software versions and configurations to ensure consistency and compatibility. Briefly, we first evaluated the overall quality of the raw sequencing data using FastQC (version 0.11.8, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low-quality bases and sequencing adapters were trimmed using Trimmomatic [35] with default parameters. For 16S rRNA sequencing data, we used QIIME2 [36] (version 2024.10.1) for read processing and taxonomic assignment. Amplicon sequence variants (ASVs) were inferred with the DADA2 plugin using dataset-specific parameters (`-p-trim-left`, `-p-trunc-len`). Taxonomic classification was performed with the pre-trained SILVA 138 (99% sequence identity) Naïve Bayes classifier [37] (`silva-138-99-nb-classifier.qza`). Relative abundances were calculated at both genus and species levels, normalized to 100% within each level. For whole-genome metagenomic shotgun sequencing data, we used MetaPhlAn4 [38] (version 4.1.0) with default settings to assign taxonomy to microbial reads. Relative abundances at both genus and species levels were computed and normalized to sum to 100% within each level.

To ensure data quality and consistency across datasets, a two-step quality control procedure was applied (Supplementary Fig. 1). In the first step, raw sequences were trimmed using Trimmomatic to remove adapter sequences and low-quality regions (average quality <20). Samples with fewer than 20 000 reads after trimming, or with post-trimming read counts <50% of the original, were flagged as "QC failed" (QC status = 0) in GMrepo. The second step involved evaluating the taxonomic composition after classification. Samples with only a single microbial species identified were also marked as "QC failed." This criterion reflects the well-established high diversity of the human gut microbiota: genuine stool samples invariably contain complex communities, and a profile dominated by a single species is overwhelmingly likely to result from contamination, sequencing artifacts, or sample mislabeling rather than true biology. Following this process, 68 723 runs passed quality control, while 39 453 did not.

Marker identification and cross-dataset comparison

Projects were eligible for marker analysis if they contained at least two phenotype groups with ≥10 quality-controlled runs per group. For each eligible project, LEfSe (Linear discriminant analysis Effect Size) [39] analysis was performed at both species and genus levels using the R package *microbiomeMarker* (v1.13.2). The analysis was conducted within each project separately, comparing two phenotype groups at a time, using default parameters. Significant taxa were recorded as project-specific microbial markers. To evaluate cross-dataset reproducibility, we calculated the MCI, defined as the proportion of comparisons in which a taxon was enriched in healthy samples relative to its total occurrences across all eligible comparisons; a low MCI indicates consistent disease enrichment. High-MCI taxa identified across multiple diseases may represent core microbiome components involved in shared pathophysiology, whereas low- or intermediate-MCI taxa may reflect context-dependent dysbiosis or environmental influences.

Web interface and data access

To support the expanded dataset and ensure long-term sustainability, the GMrepo v3 web interface was fully reimplemented using a modern front-end/back-end architecture. The new system is built with Vue 3 for the front-end and Django for the back-end, improving modularity, performance, and maintainability. While the underlying infrastructure has been updated, the overall layout and user experience remain consistent with GMrepo v2 to minimize disruption. New GMrepo v3 data are accessible through the new platform at https://gmrepo.humangut.info. For compatibility and reference, the GMrepo v2 interface remains available at https://gmrepo2022.humangut.info.

Results

In total, GMrepo v3 includes 118 965 curated samples from 890 projects, comprising 87 048 16S rRNA amplicon and 31 917 WGS metagenomic sequencing samples. These samples span 302 distinct human phenotypes (301 diseases and healthy controls), representing a substantial increase from the 133 phenotypes covered in the previous version and providing

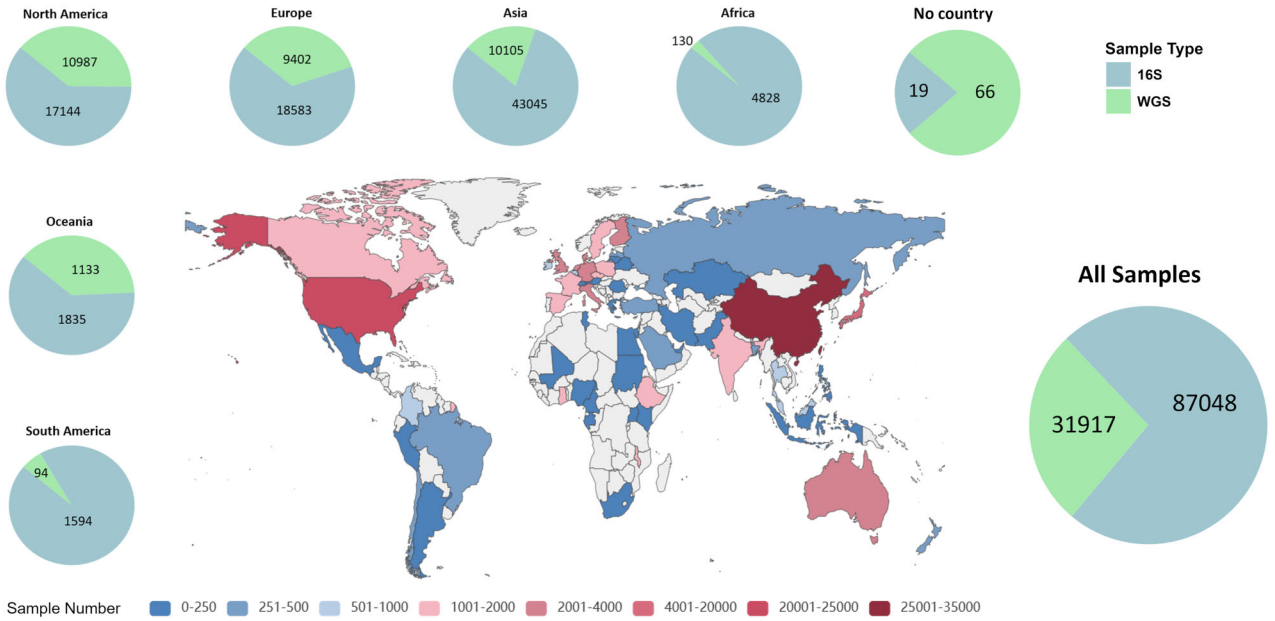


Figure 2. Geographic distribution and sequencing type composition in GMrepo v3. The central world map is colored by the number of gut microbiome samples per country. Surrounding pie charts show the proportion of 16S rRNA and WGS sequencing across six continents, samples with unreported origin, and the entire dataset. Most samples originate from Asia, Europe, and North America. While 16S data dominate overall, WGS samples are relatively more prevalent in North America, Europe, Asia, and Oceania than in South America and Africa.

a broad foundation for cross-regional and cross-technology comparative analyses.

As shown in Fig. 2, samples are globally distributed, with the largest contributions from China and the United States. While 16S rRNA sequencing remains predominant across all continents, the proportion of WGS samples is comparatively higher in North America, Europe, Oceania, and Asia than in South America and Africa, reflecting regional differences in sequencing strategies.

Adopting the disease vocabulary used in the Medical Subject Headings (MeSH) database for our projects, we found that digestive system diseases and neoplasms are the most frequently represented categories, each associated with over 100 projects (Fig. 3a). At the disease level, colorectal neoplasms, Crohn's disease, COVID-19, and Parkinson's disease are among the most commonly studied (Fig. 3b). In terms of demographic metadata, middle-aged adults comprise the largest proportion of samples across age groups, and ~30% of samples are annotated with gender information (Fig. 3c; see Supplementary Fig. 2 for age distribution including all samples). The broad phenotypic coverage, together with available demographic annotations, provides a valuable foundation for investigating microbiome–disease associations across diverse conditions and population subgroups.

Disease marker identification and cross-dataset comparison

GMrepo v3 continues to support disease marker identification and cross-dataset comparison as core functionalities. Building on the Linear discriminant analysis Effect Size (LEfSe)-based analytical framework established in GMrepo v2 [31], the current release significantly expands both the scale and phenotype coverage of marker analysis.

Of the 890 projects in GMrepo v3, 664 contain at least one sequencing run passing quality control. Among these, 275 projects met stringent inclusion criteria for marker identification, each comprising at least two phenotype groups with ≥10 quality-controlled samples (Fig. 4a). This represents a more than three-fold increase over GMrepo v2 (83 projects), enabling more comprehensive and statistically robust phenotype comparisons.

Within each qualifying project, LEfSe was applied independently to identify phenotype-associated microbial taxa. In total, GMrepo v3 reports 1299 marker taxa (726 species and 573 genera) across 167 phenotype pairs, providing a broad and systematically annotated resource for microbiome-based disease research.

To assess the consistency of microbial markers across studies, GMrepo offers cross-project comparison views. For instance, CRC markers identified from ten metagenomic projects consistently show enrichment of *Fusobacterium nucleatum* and *Gemella morbillorum* in CRC samples (Fig. 4b), which is in agreement with previous studies [40–42]. Interactive visualizations for each phenotype pair are available on the website, providing an intuitive interface for further exploration. Additionally, a marker-centric view enables users to evaluate whether a specific taxon is disease-specific or shared across conditions. For example, *Enterocloster bolteae* has been identified as enriched in multiple phenotype comparisons, including CRC, Crohn's disease, and COVID-19, corroborating findings from prior research [43–45] (Fig. 4c). These tools allow for in-depth examination of microbial signatures, which may have diagnostic or mechanistic implications.

To provide a global view of consistently reported microbial signatures, GMrepo v3 summarizes the top 50 most frequently identified species-level markers across 167 phenotype comparisons (Fig. 5). For each species, both the frequency of occurrence and the direction of enrichment—toward health or disease—are recorded. We introduce an MCI, defined as the proportion of comparisons in which the enrichment direc-

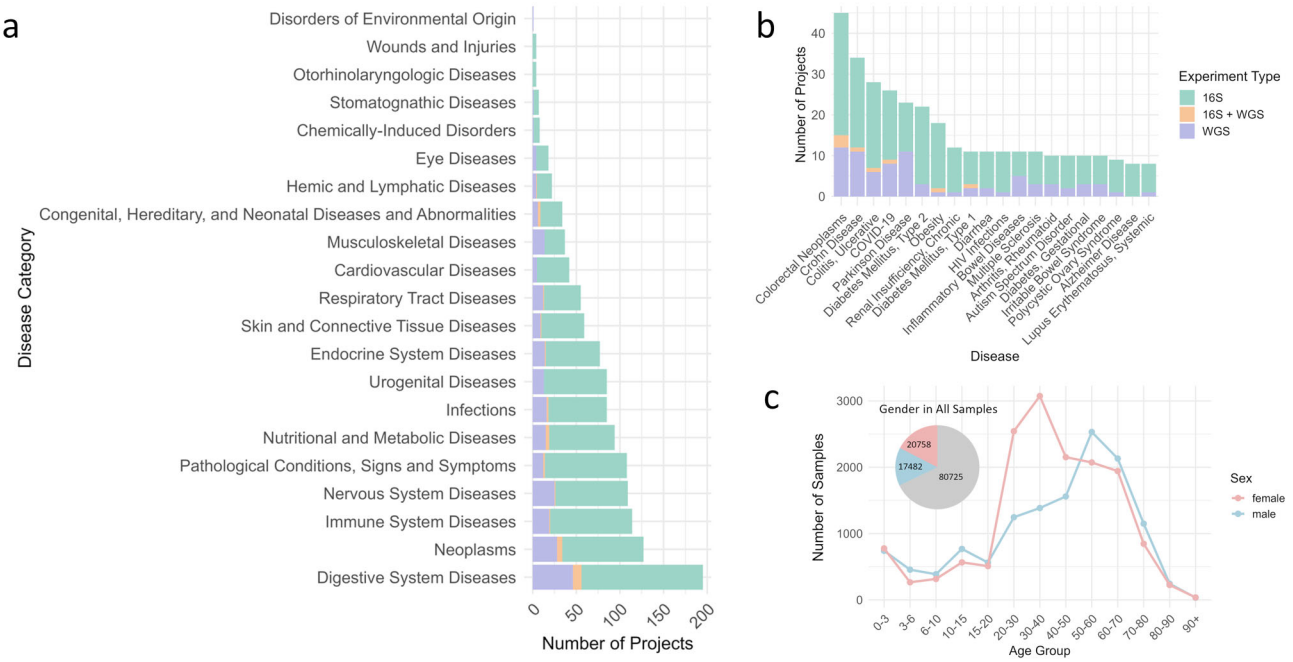


Figure 3. Phenotypic and demographic composition of GMrepo v3. (**a**) Number of projects by disease category, with digestive system diseases and neoplasms being the most represented (each with over 100 projects). (**b**) Top 20 specific diseases ranked by project count, including colorectal neoplasms, Crohn's disease, COVID-19, and Parkinson's disease. Some disease labels correspond to MeSH subtypes of inflammatory bowel diseases (IBD). For clarity, Crohn's disease and "Colitis, Ulcerative" (commonly referred to as ulcerative colitis) are the two primary IBD subtypes. We retained the original project annotations so that users may analyze these subtypes separately or merge them into IBD for downstream analyses. (**c**) Age and gender distribution of samples: the line chart shows the number of male and female samples across age groups, with the highest counts in middle-aged adults; the pie chart indicates that ∼30% of samples have gender information.

tion is associated with health (range: 0%–100%). Species with MCI > 75% are considered consistently enriched in health, and those with MCI < 25% consistently enriched in disease. Such species may represent ideal candidates for intervention, whereas those with intermediate MCI values should be interpreted with caution due to possible context- or disease-specific effects. The MCI is available on the GMrepo website for each species identified as a marker in ≥10 projects, enabling rapid assessment of both prevalence and stability of associations across conditions.

For example, *Faecalibacterium prausnitzii* was frequently identified as a health-enriched species across multiple comparisons, aligning with its established role in maintaining gut homeostasis and exerting anti-inflammatory [46, 47] and anti-tumor [7] effects. In contrast, *Hungatella hathewayi* was predominantly disease-enriched across multiple comparisons. It has been implicated in CRC [48–50] and multiple sclerosis [51, 52].

Overall, this aggregate analysis highlights microbial taxa with stable health- or disease-related patterns across diverse conditions and provides a data-driven foundation for prioritizing candidate species in microbiome-based diagnostics and therapeutics.

Discussion

GMrepo v3 enhances the utility of the database with 118 965 samples from 890 projects covering 302 phenotypes. It introduces 1299 microbial markers across 167 phenotype pairs, enabling deeper insights into disease-specific microbial signatures. The MCI provides a quantitative measure of marker stability and directionality across comparisons. Species with high or low MCI values represent consistently health- or disease-associated markers, while intermediate values indicate context-dependent effects. This framework is particularly relevant for studying multimorbidity, as it highlights markers that reliably appear across multiple diseases. The redesigned web platform, built on Vue 3 and Django, offers improved usability and scalability, broadening the database's research applications.

Compared to GMrepo v2, v3 substantially increases sample size, phenotypic diversity, and the robustness of cross-dataset marker analyses. These improvements allow researchers to prioritize microbial taxa with stable patterns across diseases, facilitating more reliable microbiome-based biomarker discovery and translational applications. Beyond increasing scale, the database enables systematic cross-project comparisons, allowing identification of core microbial taxa that exhibit consistent associations with specific phenotypes. The integration of MCI allows users to distinguish reproducible signals from context-dependent or cohort-specific effects, which is critical for robust biomarker discovery and mechanistic hypothesis generation.

High-MCI taxa identified across multiple diseases may represent core microbiome components involved in shared pathophysiology, whereas low-MCI taxa indicate consistent enrichment in disease, and intermediate MCI values reflect context- or disease-specific effects. For example, *Faecalibacterium prausnitzii* consistently appears as a health-associated marker across multiple conditions, supporting its known anti-inflammatory and gut homeostasis functions, whereas *Hungatella hathewayi* is enriched in disease phenotypes such as CRC and multiple sclerosis, illustrating phenotype-specific pathogenic associations. Such patterns provide insight into

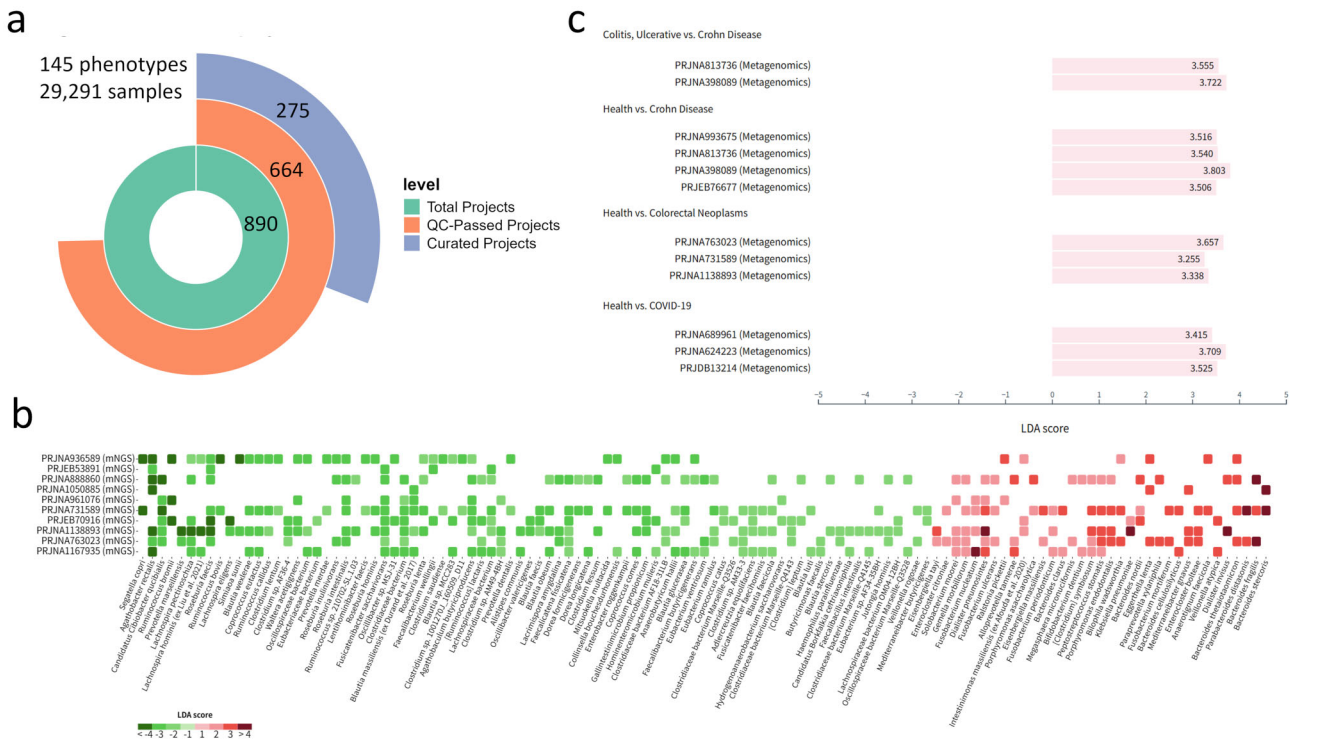


Figure 4. Disease marker identification and cross-dataset comparisons in GMrepo v3. (**a**) Overview of project selection for marker analysis. The innermost ring represents all projects ($n = 890$); the middle ring shows the subset with at least one sequencing run that passed quality control ($n = 664$); the outermost ring denotes the 275 projects selected for marker identification, each containing at least two phenotype groups with ≥10 QC-passed runs. (**b**) Cross-dataset comparison of microbial markers associated with colorectal cancer (CRC) across ten metagenomic datasets. Taxa with LDA scores <−2 are enriched in healthy samples, while those with LDA scores >2 are enriched in disease samples. Green and red indicate health- and disease-enriched taxa, respectively; color intensity reflects the magnitude of enrichment. An interactive version is available at https://gmrepo.humangut.info/phenotypes/comparisons/D006262/D015179. (**c**) Cross-disease comparison of Enterocloster bolteae enrichment patterns across multiple projects. Interactive version is available at https://gmrepo.humangut.info/taxon/208479.

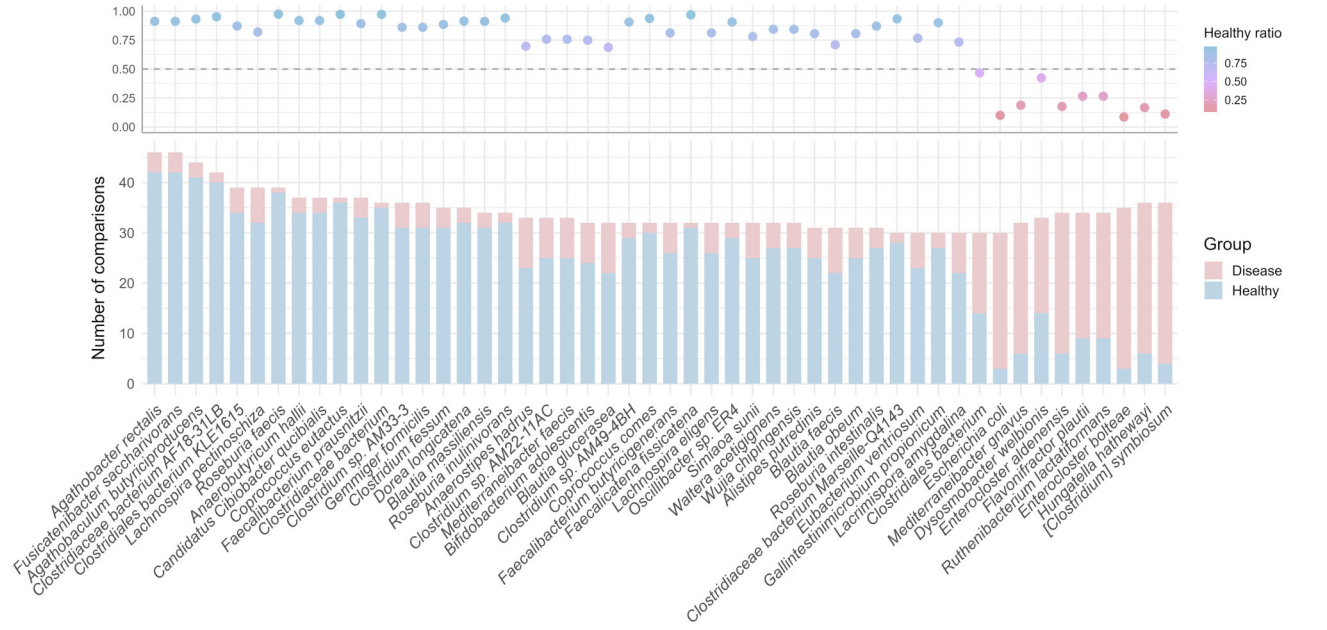


Figure 5. Top 50 species-level microbial markers across 167 phenotype comparisons in GMrepo v3. Bottom panel: total occurrences of each species as health- or disease-enriched markers. Top panel: Marker Consistency Index (MCI; 0%–100%), the proportion of health-associated occurrences; MCI > 75% indicates consistent health enrichment, and MCI < 25% indicates consistent disease enrichment. MCI is provided on the GMrepo website for species identified in ≥10 projects.

microbial roles in disease mechanisms and potential intervention targets.

The global and multi-technology coverage of GMrepo v3, including 16S rRNA and WGS samples from diverse geographic regions, further enables analyses of regional and methodological variation in microbial signatures. However, integrating 890 projects inevitably introduces heterogeneity arising from upstream experimental differences, such as DNA extraction kits and sequencing platforms. These batch effects may influence both marker identification and MCI scores, potentially confounding cross-study comparisons. We therefore recommend that users carefully consider these factors when interpreting MCI values and cross-dataset markers, and we provide metadata fields to facilitate stratified or sensitivity analyses. These capabilities allow researchers to assess reproducibility, identify shared and context-specific markers, and explore interactions between microbial composition, host phenotype, and environmental factors. In addition, GMrepo v3 continues to employ LEfSe as the primary approach for phenotype-associated marker identification. We chose LEfSe to ensure methodological continuity with GMrepo v2, facilitating comparability across database versions. LEfSe integrates nonparametric testing with effect size estimation, providing interpretable results and widely used visualization formats. Nevertheless, we acknowledge that relying on a single method has limitations, as different algorithms may yield method-specific marker sets. In future updates, we plan to incorporate additional compositionality-aware approaches (e.g. ANCOM-BC, ALDEx2) and provide consensus marker lists to further strengthen robustness and reproducibility.

Despite the expanded global coverage, the current dataset remains dominated by samples from China and the United States, reflecting the uneven availability of publicly released gut microbiome projects worldwide. This regional imbalance may limit the generalizability of certain marker associations or MCI values to underrepresented populations. We therefore caution users when extrapolating findings to global populations and will prioritize the integration of newly released datasets from Africa, South America, and other undersampled regions in future GMrepo updates.

Interactive visualization tools, including project- and marker-centric views, facilitate exploration of cross-disease and cross-cohort patterns. For instance, CRC markers consistently enriched for *Fusobacterium nucleatum* and *Gemella morbillorum* across multiple metagenomic datasets exemplify reproducible disease associations. *Enterocloster bolteae* shows enrichment across CRC, Crohn's disease, and COVID-19, highlighting taxa with pleiotropic disease associations. By linking taxa prevalence, enrichment direction, and MCI values, GMrepo v3 provides an intuitive and quantitative platform for prioritizing species for experimental validation or therapeutic targeting.

Future updates of GMrepo will expand microbial profiling, incorporating viral abundance data and metabolic pathway annotations to provide a more comprehensive view of the gut microbiome, and will also include an online tool enabling users to compare their own data with public datasets through interactive visualization, as well as the integration of additional repositories such as the Genome Sequence Archive (GSA) to broaden data coverage and population diversity. Transformer-based models will be developed to predict phenotypes and gene functions, enhancing the potential for personalized medicine. Parallel efforts will focus on developing and refining disease-specific predictive models using machine learning approaches, building on our previous work on microbiome-based disease classification and optimal model construction [53, 54].

In summary, GMrepo v3 offers a powerful, scalable resource for microbiome research, with plans for further expansions to enhance its diagnostic and therapeutic applications.

Acknowledgements

Author contributions: Can Liu (Data curation [lead], Formal analysis [lead], Software [lead], Visualization [lead], Writing—original draft [lead]), Xiaohan Wang (Data curation [supporting], Formal analysis [supporting]), Zhilin Zhang (Data curation [supporting], Formal analysis [supporting]), Wenboxin Wang (Data curation [supporting]), Tianci Wang (Data curation [supporting]), Yuchen Zhao (Data curation [supporting]), and Mingyu Wang (Validation [supporting]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This research is supported by National Key Research and Development Program of China [2024YFA0918500 to W.H.C.], Double First-Class Construction Funds for Provincial Departments Affiliated Universities in Hubei Province [5001170159 to W.H.C.], and Yunnan Expert Workstation [202505AF350080]. Funding to pay the Open Access publication charges for this article was provided by National Key Research and Development Program of China [2024YFA0918500 to W.H.C.], Double First-Class Construction Funds for Provincial Departments Affiliated Universities in Hubei Province [5001170159 to W.H.C.], and Yunnan Expert Workstation [202505AF350080].

Data availability

GMrepo v3 data are accessible through the new platform at https://gmrepo.humangut.info. Users can download curated sample metadata, taxonomic profiles, and microbial marker information for research purposes in help page. For compatibility and reference, the GMrepo v2 interface remains available at https://gmrepo2022.humangut.info. In addition, all code used for data processing and analysis is available upon request.

References

1. Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. *Nat Rev Microbiol* 2021;19:55–71. https://doi.org/10.1038/s41579-020-0433-9
2. Rinninella E, Raoul P, Cintoni M *et al.* What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. *Microorganisms* 2019;7:14. https://doi.org/10.3390/microorganisms7010014

3. Almeida A, Mitchell AL, Boland M *et al.* A new genomic blueprint of the human gut microbiota. *Nature* 2019;568:499–504. https://doi.org/10.1038/s41586-019-0965-1
4. Ross FC, Patangia D, Grimaud G *et al.* The interplay between diet and the gut microbiome: implications for health and disease. *Nat Rev Microbiol* 2024;22:671–86. https://doi.org/10.1038/s41579-024-01068-4
5. Luis AS, Hansson GC. Intestinal mucus and their glycans: a habitat for thriving microbiota. *Cell Host Microbe* 2023;31:1087–100. https://doi.org/10.1016/j.chom.2023.05.026
6. Hou K, Wu ZX, Chen XY *et al.* Microbiota in health and diseases. *Signal Transduct Target Ther* 2022;7:135. https://doi.org/10.1038/s41392-022-00974-4
7. Shi Z, Li M, Zhang C *et al.* Butyrate-producing *Faecalibacterium prausnitzii* suppresses natural killer/T-cell lymphoma by dampening the JAK–STAT pathway. *Gut* 2025;74:557–70. https://doi.org/10.1136/gutjnl-2024-333530
8. Zhu J, Jiang MZ, Chen X *et al.* Systematic pairwise co-cultures uncover predominant negative interactions among human gut bacteria. *Microbiome* 2025;13:161. https://doi.org/10.1186/s40168-025-02156-0
9. Yoo J, Groer M, Dutra S *et al.* Gut microbiota and immune system interactions. *Microorganisms* 2020;8:1587. https://doi.org/10.3390/microorganisms8101587
10. Yang D, Jacobson A, Meerschaert KA *et al.* Nociceptor neurons direct goblet cells via a CGRP–RAMP1 axis to drive mucus production and gut barrier protection. *Cell* 2022;185:4190–205. https://doi.org/10.1016/j.cell.2022.09.024
11. de Vos WM, Tilg H, Van Hul M *et al.* Gut microbiome and health: mechanistic insights. *Gut* 2022;71:1020–32. https://doi.org/10.1136/gutjnl-2021-326789
12. Jyoti, Dey P. Mechanisms and implications of the gut microbial modulation of intestinal metabolic processes. *NPJ Metab Health Dis* 2025;3:24. https://doi.org/10.1038/s44324-025-00066-1
13. Asnicar F, Berry SE, Valdes AM *et al.* Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat Med* 2021;27:321–32. https://doi.org/10.1038/s41591-020-01183-8
14. Guo D, Deng Y, Yang Q *et al.* Symbiotic probiotic communities with multiple targets successfully combat obesity in high-fat-diet-fed mice. *Gut Microbes* 2024;16:2420771. https://doi.org/10.1080/19490976.2024.2420771
15. De Luca R, Nardone S, Arrè V *et al.* Gastrointestinal microbiota and inflammasomes interplay in health and disease: a gut feeling. *Gut* 2025. https://doi.org/10.1136/gutjnl-2025-334938
16. Shen Y, Fan N, Ma SX *et al.* Gut microbiota dysbiosis: pathogenesis, diseases, prevention, and therapy. *MedComm* 2025;6:e70168. https://doi.org/10.1002/mco2.70168
17. Jiang P, Wu S, Luo Q *et al.* Metagenomic analysis of common intestinal diseases reveals relationships among microbial signatures and powers multidisease diagnostic models. *mSystems* 2021;6:e00112–21. https://doi.org/10.1128/mSystems.00112-21
18. Fu J, Qiu W, Zheng H *et al.* Ageing trajectory of the gut microbiota is associated with metabolic diseases in a chronological age-dependent manner. *Gut* 2023;72:1431–3. https://doi.org/10.1136/gutjnl-2022-328034
19. Zhang X, Zhong H, Li Y *et al.* Sex- and age-related trajectories of the adult human gut microbiota shared across populations of different ethnicities. *Nat Aging* 2021;1:87–100. https://doi.org/10.1038/s43587-020-00014-2
20. Watanabe K, Wilmanski T, Diener C *et al.* Multiomic signatures of body mass index identify heterogeneous health phenotypes and responses to a lifestyle intervention. *Nat Med* 2023;29:996–1008. https://doi.org/10.1038/s41591-023-02248-0
21. Maghini DG, Oduaran OH, Olubayo LAI *et al.* Expanding the human gut microbiome atlas of Africa. *Nature* 2025;638:718–28. https://doi.org/10.1038/s41586-024-08485-8
22. Fishbein SRS, Mahmud B, Dantas G. Antibiotic perturbations to the gut microbiome. *Nat Rev Microbiol* 2023;21:772–88. https://doi.org/10.1038/s41579-023-00933-y
23. Zhernakova DV, Wang D, Liu L *et al.* Host genetic regulation of human gut microbial structural variation. *Nature* 2024;625:813–21. https://doi.org/10.1038/s41586-023-06893-w
24. Robles Alonso V, Guarner F. Linking the gut microbiota to human health. *Br J Nutr* 2013;109:S21–6. https://doi.org/10.1017/S0007114512005235
25. Chibani CM, Mahnert A, Borrel G *et al.* A catalogue of 1,167 genomes from the human gut archaeome. *Nat Microbiol* 2022;7:48–61. https://doi.org/10.1038/s41564-021-01020-9
26. O'Cathail C, Ahamed A, Burgin J *et al.* The European Nucleotide Archive in 2024. *Nucleic Acids Res* 2025;53:D49–55. https://doi.org/10.1093/nar/gkae975
27. Sayers EW, Bolton EE, Brister JR *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2022;50:D20–6. https://doi.org/10.1093/nar/gkab1112
28. Richardson L, Allen B, Baldi G *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res* 2023;51:D753–9. https://doi.org/10.1093/nar/gkac1080
29. Shi W, Qi H, Sun Q *et al.* gcMeta: a global catalogue of metagenomics platform to support the archiving, standardization and analysis of microbiome data. *Nucleic Acids Res* 2019;47:D637–48. https://doi.org/10.1093/nar/gky1008
30. Gonzalez A, Navas-Molina JA, Kosciolek T *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nat Methods* 2018;15:796–8. https://doi.org/10.1038/s41592-018-0141-9
31. Dai D, Zhu J, Sun C *et al.* GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Res* 2022;50:D777–84. https://doi.org/10.1093/nar/gkab1019
32. Wu S, Sun C, Li Y *et al.* GMrepo: a database of curated and consistently annotated human gut metagenomes. *Nucleic Acids Res* 2020;48:D545–53. https://doi.org/10.1093/nar/gkz764
33. Pasolli E, Schiffer L, Manghi P *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* 2017;14:1023–4. https://doi.org/10.1038/nmeth.4468
34. Barrett T, Clark K, Gevorgyan R *et al.* BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res* 2012;40:D57–63. https://doi.org/10.1093/nar/gkr1163
35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–20. https://doi.org/10.1093/bioinformatics/btu170
36. Bolyen E, Rideout JR, Dillon MR *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019;37:852–7. https://doi.org/10.1038/s41587-019-0209-9
37. Quast C, Pruesse E, Yilmaz P *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;41:D590–6. https://doi.org/10.1093/nar/gks1219
38. Blanco-Míguez A, Beghini F, Cumbo F *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* 2023;41:1633–44. https://doi.org/10.1038/s41587-023-01688-w
39. Segata N, Izard J, Waldron L *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol* 2011;12:R60. https://doi.org/10.1186/gb-2011-12-6-r60
40. Queen J, Cing Z, Minsky H *et al. Fusobacterium nucleatum* is enriched in invasive biofilms in colorectal cancer. *npj Biofilms Microbiomes* 2025;11:81. https://doi.org/10.1038/s41522-025-00717-7
41. Zhang Y, Luo J, Chen K *et al.* Cross-cohort analysis identifies shared gut microbial signatures and validates microbial risk scores for colorectal cancer. *J Transl Med* 2025;23:676. https://doi.org/10.1186/s12967-025-06676-z
42. Zhou Z, Niu Y, Ma Y *et al.* Lipid accumulation inhibition strategies alleviate *Fusobacterium nucleatum*-infected colorectal cancer. *Microbiome* 2025;13:181. https://doi.org/10.1186/s40168-025-02133-7

43. Yang Y, Du L, Shi D *et al*. Dysbiosis of human gut microbiome in young-onset colorectal cancer. *Nat Commun* 2021;12:6757. https://doi.org/10.1038/s41467-021-27112-y
44. Shen M, Gao S, Zhu R *et al*. Multimodal metagenomic analysis reveals microbial InDels as superior biomarkers for pediatric Crohn's disease. *J Crohns Colitis* 2025;19:jjaf039. https://doi.org/10.1093/ecco-jcc/jjaf039
45. Liu Q, Su Q, Zhang F *et al*. Multi-kingdom gut microbiota analyses define COVID-19 severity and post-acute COVID-19 syndrome. *Nat Commun* 2022;13:6806. https://doi.org/10.1038/s41467-022-34535-8
46. Lopez-Siles M, Duncan SH, Garcia-Gil LJ *et al*. *Faecalibacterium prausnitzii*: from microbiology to diagnostics and prognostics. *ISME J* 2017;11:841–52. https://doi.org/10.1038/ismej.2016.176
47. Quévrain E, Maubert MA, Michon C *et al*. Identification of an anti-inflammatory protein from *Faecalibacterium prausnitzii*, a commensal bacterium deficient in Crohn's disease. *Gut* 2016;65:415–25. https://doi.org/10.1136/gutjnl-2014-307649
48. Xia X, Wu WKK, Wong SH *et al*. Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer. *Microbiome* 2020;8:108. https://doi.org/10.1186/s40168-020-00847-4
49. Huang Z, Wang C, Huang Q *et al*. Hungatella hathewayi impairs the sensitivity of colorectal cancer cells to 5-FU through decreasing CDX2 expression. *Human Cell* 2023;36:2055–65. https://doi.org/10.1007/s13577-023-00938-y
50. Zhang J, He Y, Xia L *et al*. Expansion of colorectal cancer biomarkers based on gut bacteria and viruses. *Cancers* 2022;14:4662. https://doi.org/10.3390/cancers14194662
51. Consortium iSMS. Gut microbiome of multiple sclerosis patients and paired household healthy controls reveal associations with disease risk and course. *Cell* 2022;185:3467–86.e16.
52. Ouyang Q, Yu H, Xu L *et al*. Relationship between gut microbiota and multiple sclerosis: a scientometric visual analysis from 2010 to 2023. *Front Immunol* 2024;15:1451742. https://doi.org/10.3389/fimmu.2024.1451742
53. Li P, Li M, Chen WH. Best practices for developing microbiome-based disease diagnostic classifiers through machine learning. *Gut Microbes* 2025;17:2489074. https://doi.org/10.1080/19490976.2025.2489074
54. Li M, Liu J, Zhu J *et al*. Performance of gut microbiome as an independent diagnostic tool for 20 diseases: cross-cohort validation of machine-learning classifiers. *Gut Microbes* 2023;15:2205386. https://doi.org/10.1080/19490976.2023.2205386

Received: September 9, 2025. **Revised:** October 12, 2025. **Accepted:** October 13, 2025