Published online 12 November 2021 *Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D777–D784***
https://doi.org/10.1093/nar/gkab1019

GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison

Die Dai[1,†], Jiaying Zhu[1,†], Chuqing Sun[1], Min Li[1], Jinxin Liu[2], Sicheng Wu[1], Kang Ning[1], Li-jie He[3,*], Xing-Ming Zhao[2,4,5,*] and Wei-Hua Chen[1,6,*]

[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, China, [2]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China, [3]Department of Oncology, The People's Hospital of Liaoning Province, People's Hospital of China Medical University 110016 Shenyang, China, [4]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Ministry of Education, China, [5]Research Institute of Intelligent Complex System, Fudan University, Shanghai 200433, China and [6]Institution of Medical Artificial Intelligence, Binzhou Medical University, Yantai 264003, China

Received July 28, 2021; Revised October 05, 2021; Editorial Decision October 11, 2021; Accepted October 13, 2021

ABSTRACT

GMrepo (data repository for Gut Microbiota) is a database of curated and consistently annotated human gut metagenomes. Its main purposes are to increase the reusability and accessibility of human gut metagenomic data, and enable cross-project and phenotype comparisons. To achieve these goals, we performed manual curation on the meta-data and organized the datasets in a phenotype-centric manner. GMrepo v2 contains 353 projects and 71,642 runs/samples, which are significantly increased from the previous version. Among these runs/samples, 45,111 and 26,531 were obtained by 16S rRNA amplicon and whole-genome metagenomics sequencing, respectively. We also increased the number of phenotypes from 92 to 133. In addition, we introduced disease-marker identification and cross-project/phenotype comparison. We first identified disease markers between two phenotypes (e.g. health versus diseases) on a per-project basis for selected projects. We then compared the identified markers for each phenotype pair across datasets to facilitate the identification of consistent microbial markers across datasets. Finally, we provided a marker-centric view to allow users to check if a marker has different trends in different diseases. So far, GMrepo includes 592 marker taxa (350 species and 242 genera) for 47 phenotype pairs, identified from 83 selected projects. GMrepo v2 is freely available at: https://gmrepo.humangut.info.

INTRODUCTION

Gut microbiota is important in maintaining normal physiology of host throughout life (1–4). Growing evidence suggests that microbiota disruption can affect the immune function (5,6), metabolism (7,8), energy production (9,10) and cause various diseases (11–18). Many factors including age (19), sex (20), body-mass-index (BMI) (21), country, environment (22), genetics (23), diet (24) and recent antibiotics usage (25) can influence composition of gut microbial communities (26). In recent years, the study of the impact of gut microbiota on human health is a rapidly moving field of research and has been widely considered as an exciting advancement in biomedicine (27–29). Increasing numbers of human gut metagenomic data (including both 16S amplicon and shotgun metagenomics sequencing data) has been rapidly generated. Raw sequencing data are often deposited into several general purpose databases, including European Nucleotide Archive (ENA) (30) (https://www.ebi.ac.uk/ena) and NCBI Sequence Read Archive (SRA) (31) (https://www.ncbi.nlm.nih.gov/sra); in addition, several other public resources, including MGnify (32), gcMeta (33) and Qiita (34) have collected processed data and organized them according to the habitats from which the samples were taken, while gutMDisorder (35), GIMICA (36) and DISBIOME (37) have linked gut micro-

*To whom correspondence should be addressed. Tel: +1 582 735 4263; Email: weihuachen@hust.edu.cn
Correspondence may also be addressed to Li-jie He. Email: 17702488896@163.com
Correspondence may also be addressed to Xing-Ming Zhao. Email: xmzhao@fudan.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

biota dysbiosis with various human diseases. These existing databases greatly promoted data reuse. However, obstacles of the reusability and accessibility of the rapidly growing human metagenomic data still remain, especially the inaccurate and/or incomplete phenotype information and/or missing metadata; for example, our previous analysis revealed that ∼30% of gut metagenomic samples did not have any of the three basic information including age, gender and BMI, even after several rounds of manual curation (38). Recently, curatedMetagenomicData, a curated metagenomic data resource became available; it provides standardized, curated human microbiome data with pre-calculated taxonomic and functional annotations (39), which will greatly facilitate data reusability and promote novel analysis of human metagenomics. However, it is available as a R package and might be difficult to use for non-R users; in addition, cross-project comparisons are not straightforward in curatedMetagenomicData such as the prevalence of a species of interests across samples of multiple diseases and whether a disease marker species is specific to that disease or shared by multiple diseases.

In 2020, we introduced GMrepo v1 (data repository for Gut Microbiota) (38) as an online database of curated and consistently annotated human gut metagenomes to facilitate the reusability and accessibility of the increasing human metagenomic data. We performed extensive meta-data curation for each collected run/sample and include all possible related meta-data, such as age, sex, country, body-mass-index (BMI) and recent antibiotics usage. Further, we consistently annotated microbial contents by assigning the sequencing reads to taxonomic units and pre-computed species/genus relative abundances using state-of-the-art toolsets. We organized the collected samples based on their associated phenotypes and added within- and cross-phenotype statistics including taxonomic abundances, prevalence and co-occurrences, to facilitate researchers to easily explore the distribution of a species/genus in diseases of interests and compare to that of healthy controls. In addition, we provided programmable access to achieve most of contents in GMrepo by representational state transfer (REST) application programming interfaces (APIs). GMrepo is equipped with powerful graphical query builders to make users search the collected samples and projects more conveniently (38).

In this study, we introduce an updated version of GMrepo. In this new version, we collected more projects, runs/samples and phenotypes. Most importantly, we added disease marker identification and cross-project/phenotype comparisons of the identified markers. The main features of GMrepo v2 include: (a) identification of disease markers between two phenotypes (e.g. health versus adenoma) on per-project basis for selected projects, (b) cross-dataset disease marker comparison to facilitate the identification of consistent microbial markers across multiple datasets of the same diseases, (c) cross-disease marker comparison to allow users to check if microbial markers are unique to a specific disease or shared by multiple diseases, and if they have different trends in different diseases.

DATA GENERATION

Collection of sequencing reads and manual curation of meta-data

To obtain more human gut metagenomic data, we searched recently updated projects in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject) and publications in the NCBI PubMed database (https://pubmed.ncbi.nlm.nih.gov/) using 'human gut microbiota' as the keyword. Projects with clearly defined phenotype information and public raw sequencing data were collected for further analysis. The raw sequencing reads were downloaded from NCBI SRA (Sequence Read Archive, https://www.ncbi.nlm.nih.gov/sra) (31) and EBI ENA (European Nucleotide Archive, https://www.ebi.ac.uk/ena) (30) database as reported in GMrepo v1 (38). Related meta-data were also downloaded using in-house PERL (version 5.30.0) and R (version 4.0.4) scripts. We then performed two rounds of manual curation on the meta-data. For the first round, meta-data were extracted and manually examined, including technical meta-data such as the sequencing platform, type of sequences obtained (i.e. 16S rRNA amplicon or whole-genome metagenomic) and number of sequences, and the human host related meta-data such as phenotypes (health or diseases), BMI, age, sex, diet, country and antibiotic usage of the associated samples/runs. For the second round, different curators from the first round reviewed the collected meta-data and made necessary corrections. GMrepo v2 now contains 71 642 runs/samples from 353 projects. Among these runs/samples, 45 111 and 26 531 were obtained by 16S rRNA amplicon (16S for short) and whole-genome metagenomics (mNGS for short) sequencing, respectively. This collection represents a significant increase compared to our previous version, which contained 58 903 runs/samples 253 projects. In addition, the number of phenotypes (i.e. health and diseases) has also been increased from 92 to 133. The newly added diseases include COVID-19 (four projects; https://gmrepo.humangut.info/phenotypes/D000086382) and many others.

Taxonomic assignment and relative abundances calculation

The newly collected raw sequencing data were processed as reported in the previous version (38). In short, we first evaluated the overall quality of the downloaded data using FastQC (version 0.11.8, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), followed by removing low-quality bases and sequencing vectors using Trimmomatic with default parameters (40). Then we annotated microbial contents and calculated relative abundances. For 16S sequences, QIIME2 (41) was used to analyze the obtained clean data and assign taxonomic classification information to the reads; ASV (Amplicon Sequence Variant) instead of OTU (Operational Taxonomic Units) results were used, as the former can provide more precise measurement of sequence variation and be able to easily compare sequences between different studies (tractability and reproducibility) (41). DADA2 (version 1.18.0) (42) pipeline implemented in QIIME2 was used to filter the sequencing reads and construct ASVs table without any

sequence clustering. Deblur (43) was used for denoising and chimera removal. The taxonomy classification database used was Greengenes version 13.8. Relative abundances were then calculated for each sample at both species and genus levels, with totaling abundances of 100% at both levels respectively. For whole-genome metagenomic sequences, MetaPhlAn2 (44) was applied with default parameters for the taxonomic assignments to the sequencing reads. We calculated the relative abundances at species and genus levels, with totaling abundances of 100% at both levels respectively.

DISEASE MARKER IDENTIFICATION AND CROSS-DATASET COMPARISON

One of the main features that distinguish GMrepo from other metagenomic databases is the cross-dataset comparison. To bring this to the next level, we introduced disease-marker identification for selected high-quality projects and provided tools to facilitate cross-project/phenotype comparisons of the identified markers.

Identification of disease markers between two phenotypes

To better understand the relationships between gut microbiota dysbiosis and human diseases, we performed in-depth analysis to identify differential bacteria and disease markers between two phenotypes (e.g., health versus adenoma) on per-project basis for selected projects, especially those with high-quality data. To get reliable results of the analysis, we selected projects that met all the criteria: (i) the project must include at least two phenotypes, very often between a disease phenotype and healthy controls. For those having different disease stages, we also compared the samples between different stages; (ii) numbers of samples must be >20. Furthermore, manually curation was performed for selected projects in order to: (a) select usable runs with clearly defined phenotype(s), (b) merge multiple runs if they correspond to the same sample, (c) calculate taxon abundances on per-sample basis instead of per-run basis, (d) group samples according to their corresponding phenotypes and (e) identify marker taxa between a pair of phenotypes of interests, e.g., Health versus colorectal cancer (CRC). Here, a 'marker taxon' refers to a species or genus whose relative abundances showed significant differences between phenotypes. In GMrepo, marker taxa were identified using LEfSe (Linear discriminant analysis Effect Size) analysis (45) implemented in the 'microbiomeMarker' package (version 0.0.1.9000) of R (version 4.0.4). Linear discriminant analysis (LDA) scores were used to describe the extents of the differences, with larger values indicating more significant differences. A LDA cutoff of 2 was used as the cutoff for marker taxa. Markers were identified on a per-dataset basis in order to control for project-specific confounding factors such as DNA extraction methods and sequencing platforms. For 16s rDNA sequencing datasets, genus level markers were identified; for whole-genome metagenomic datasets (also known as metagenomic next-generation sequencing datasets, mNGS for short), both species and genus level markers were identified.

Since our marker identification was on per-project basis, for each project we added bar plots to visualize the marker taxa between two phenotypes. Shown in Figure 1 are the marker species identified between Health and Adenoma for BioProject PRJEB6070. For mNGS datasets like PRJEB6070, genus level markers are also available; researchers can choose to show markers at either species or genus level, or both levels together (see Figure 1 for more details; see also https://gmrepo.humangut.info/data/project/PRJEB6070/D006262/D000236). BioProject PRJEB6070 contained samples of three phenotypes, namely health, adenoma and CRC. Thus, in addition to markers between heath and adenoma, we also identified marker taxa for health versus CRC, and adenoma versus CRC respectively. See https://gmrepo.humangut.info/data/project/PRJEB6070 for more details. Note the newly added marker identification results, the phenotype pairs and their corresponding runs and groups can be found in the 'in-depth analysis' section of the webpage.

So far, GMrepo includes 592 marker taxa (350 species and 242 genera) for 47 phenotype pairs, identified from 83 selected projects; more projects will be analyzed in the future. The detailed information of these marker taxa is listed in https://gmrepo.humangut.info/taxon/markertaxa. Additional links to the NCBI taxonomy (46), ENA taxonomy and NCBI MeSH Browser were also provided for each of the marker taxa, in order to facilitate researchers to obtain more information. More external databases will be included in the future.

Cross-dataset disease marker comparison

In GMrepo, a disease could be covered by multiple datasets/projects. A recent meta-analysis on CRC-associated gut microbiome projects suggested that disease-related microbial markers were not always consistent across studies/projects (47). Thus, to facilitate cross-project comparisons of the identified markers within each project, we added a dedicated page for each phenotype pair (e.g. health versus liver cirrhosis, or adenoma versus colorectal cancer) to systematically show the consistent and non-consistent disease-associated microbial markers across datasets. Shown in Figure 2 are two typical examples. For example, Figure 2A shows the biomarkers between heath and CRC across seven datasets (See also https://gmrepo.humangut.info/phenotypes/comparisons/D006262/D015179 for details). We observed consistent disease-associated microbial markers across the seven projects; for example, known marker species of CRC including *Fusobacterium nucleatum* (48–52), *Parvimonas micra* (53,54) and *Gemella morbillorum* (55) were identified to be enriched in CRC patients in most studies. However, in the phenotype comparisons of health and 'arthritis, rheumatoid', we observed non-consistent disease-associated microbial makers (Figure 2B, see also https://gmrepo.humangut.info/phenotypes/comparisons/D006262/D001172 for details). So far, GMrepo includes 47 phenotype comparisons, see https://gmrepo.humangut.info/phenotypes/comparisons for a complete list.

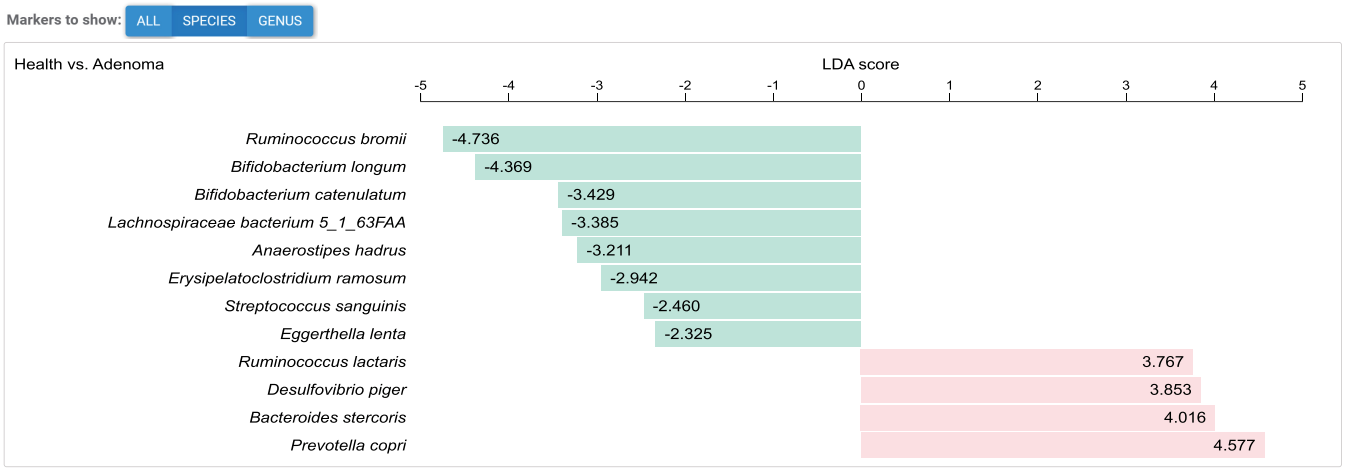

Figure 1. Disease markers identified between two phenotypes in a project. Here data from BioProject PRJEB6070 are used as an example; health and disease (adenoma) enriched species are plotted in green and pink respectively. The markers were identified using LEfSe. LDA (linear discriminant analysis) scores (X-axis) were used to show the extents of their enrichment. For whole-genome metagenomic dataset like PRJEB6070, genus level markers were also identified. Users can use the widgets (blue buttons) to choose the markers to show. For 16S rRNA datasets, only genus level markers were identified; thus, the 'Species' button will be unclickable.

Cross-disease marker comparison

We also provided a marker-centric view to allow users to check if a microbial marker is unique to a specific disease or shared by multiple diseases, and if it has different trends in different diseases. Take *F. nucleatum* as an example, it has been identified as a marker species in eight phenotype comparisons and showed consistent trends as a disease-enriched maker (Figure 3A, see also https://gmrepo.humangut.info/taxon/851). In addition to being a CRC marker, *F. nucleatum* is also associated with multiple diseases in GMrepo v2, including Cardiovascular Disease, Inflammatory Bowel Diseases, Liver Cirrhosis and COVID-19. Interestingly, although *F. nucleatum* was enriched in CRC samples in the adenoma versus CRC comparison (Figure 3A), it was not enriched in the adenoma samples as compared with the healthy controls (see also https://gmrepo.humangut.info/phenotypes/comparisons/D006262/D000236), suggesting it came at the latter stages of CRC (and maybe other diseases). These results are consistent with recent publications that *F. nucleatum* is not a marker for gut microbiota-based adenoma diagnostic models (56,57). Conversely, *Prevotella copri* was found to have inconsistent trends between phenotype pairs (Figure 3B and also https://gmrepo.humangut.info/taxon/165179) and even between projects of the same phenotype comparisons. *P. copri* was reported to be associated with gut microbial enterotypes whose abundances could be affected by diet, age and gender (58,59). The inconsistent trends may indicate either the undetected biases between disease and control groups in the related datasets, or an equilibrium state for *P. copri* in the gut should be maintained.

Future directions

In addition to continuously adding new human gut metagenomic data to GMrepo in the future, we plan to add new contents to GMrepo, including (but not limited to) functional profiles and metabolic pathway profiles for the collected samples. It is also necessary to re-analyze all data with the latest version of the tools, or use new tools that become available in the future. In addition, we plan to include genomic sequences for the identified species, especially those directly assembled from human gut metagenome datasets (60). These will further facilitate the reusability and accessibility of human gut metagenomic data and will contribute to better understanding of the relationships between gut microbiota dysbiosis and human diseases.

CONCLUSIONS

In this study, we introduced GMrepo v2, an updated version of the online database of curated, consistently annotated meta-data and human gut metagenomic data. Updates since the last version include increased numbers of projects, samples/runs and phenotypes by multiple rounds of extensive manual curation of the meta-data. One of the main features that distinguish GMrepo from other metagenomic databases is cross-dataset comparison. To bring this to the next level, we introduced disease-marker identification and performed cross project/phenotype comparisons, including: (i) identification of disease markers between two phenotypes on per-project basis for selected projects, especially those with high-quality data; (ii) cross-dataset disease marker comparison to facilitate the identification of consistent microbial markers across datasets; (iii) cross-disease marker comparison to provide a marker-centric view to allow users to check if microbial markers have different trends in different diseases. So far, GMrepo includes 592 marker taxa (350 species and 242 genera) for 47 phenotype pairs, identified from 83 selected projects; more projects will be analyzed in the future. We believe that GMrepo v2 is expected to be a highly useful and an important database for biologists and bioinformaticians studying gut microbiome. In the future, we aim to update GMrepo regularly to provide up-to-date contents and include more functionalities.

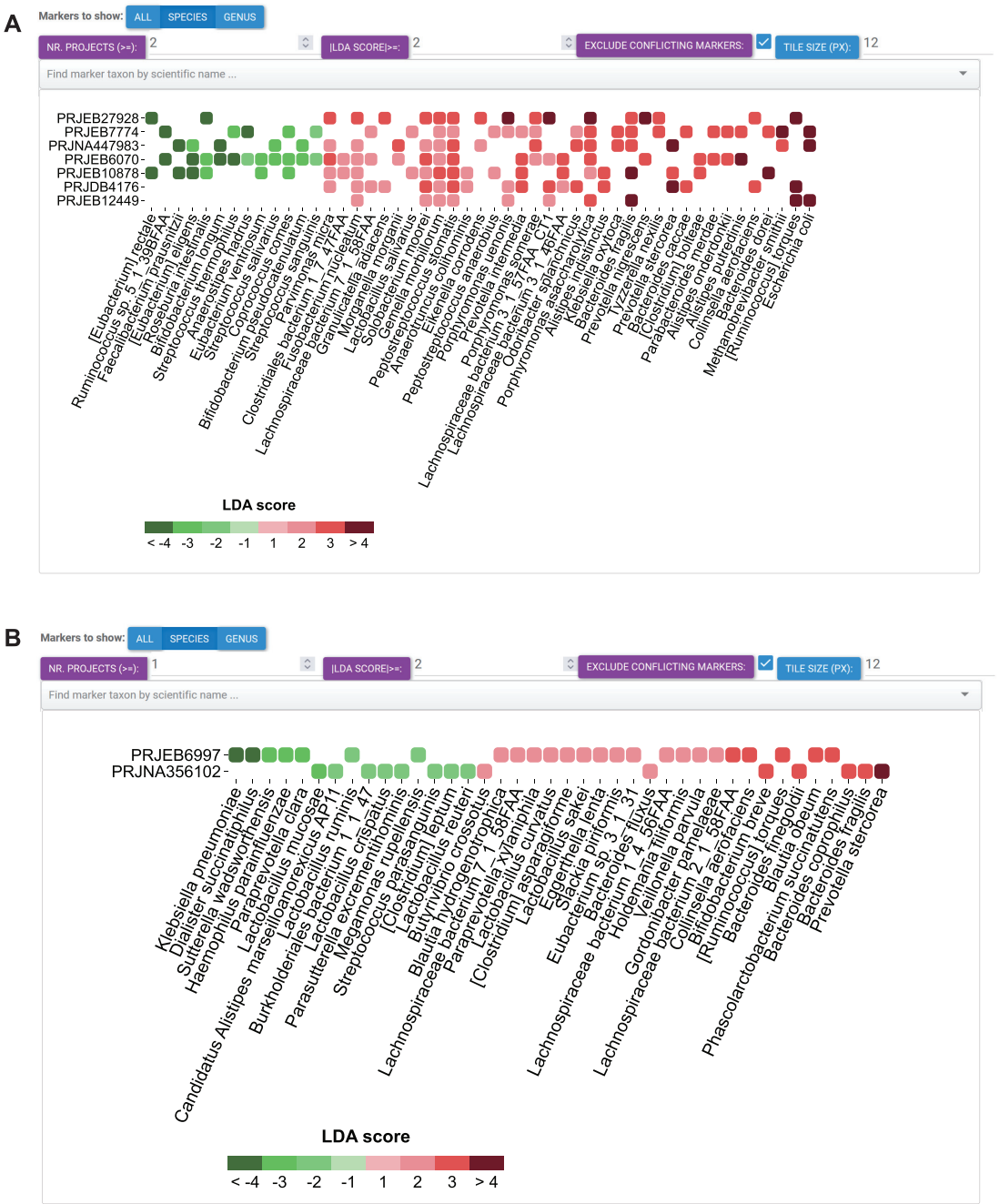

Figure 2. Cross-study comparison of microbial markers. (**A**) Comparison of marker species for colorectal cancer in seven metagenomic projects. (**B**) Comparison of marker species for 'arthritis, rheumatoid' in two projects. Marker taxa with LDA <-2 are health enriched, while those with LDA >2 are disease enriched. Health and disease enriched markers are shown in green and red respectively, with deeper color indicate increased enrichment. To facilitate users to explore the markers, a few widgets are included to allow users to 1) filter markers according to the number of projects they are identified, 2) filter markers according to the absolute LDA scores, 3) exclude markers that show inconsistent trends (e.g. those are significantly decreased in disease in one project but significantly increased in others) among projects and 4) change the size of the tiles. Users can also save the resulting visualization as SVG or PNG format. Please consult https://gmrepo.humangut.info/phenotypes/comparisons/D006262/D015179 and https://gmrepo.humangut.info/phenotypes/comparisons/D006262/D001172 for the interactive versions on our website; for the second link, please change the value of the 'NR.PROJECTS (> =):' widget on the webpage to '1' in order to show the markers.

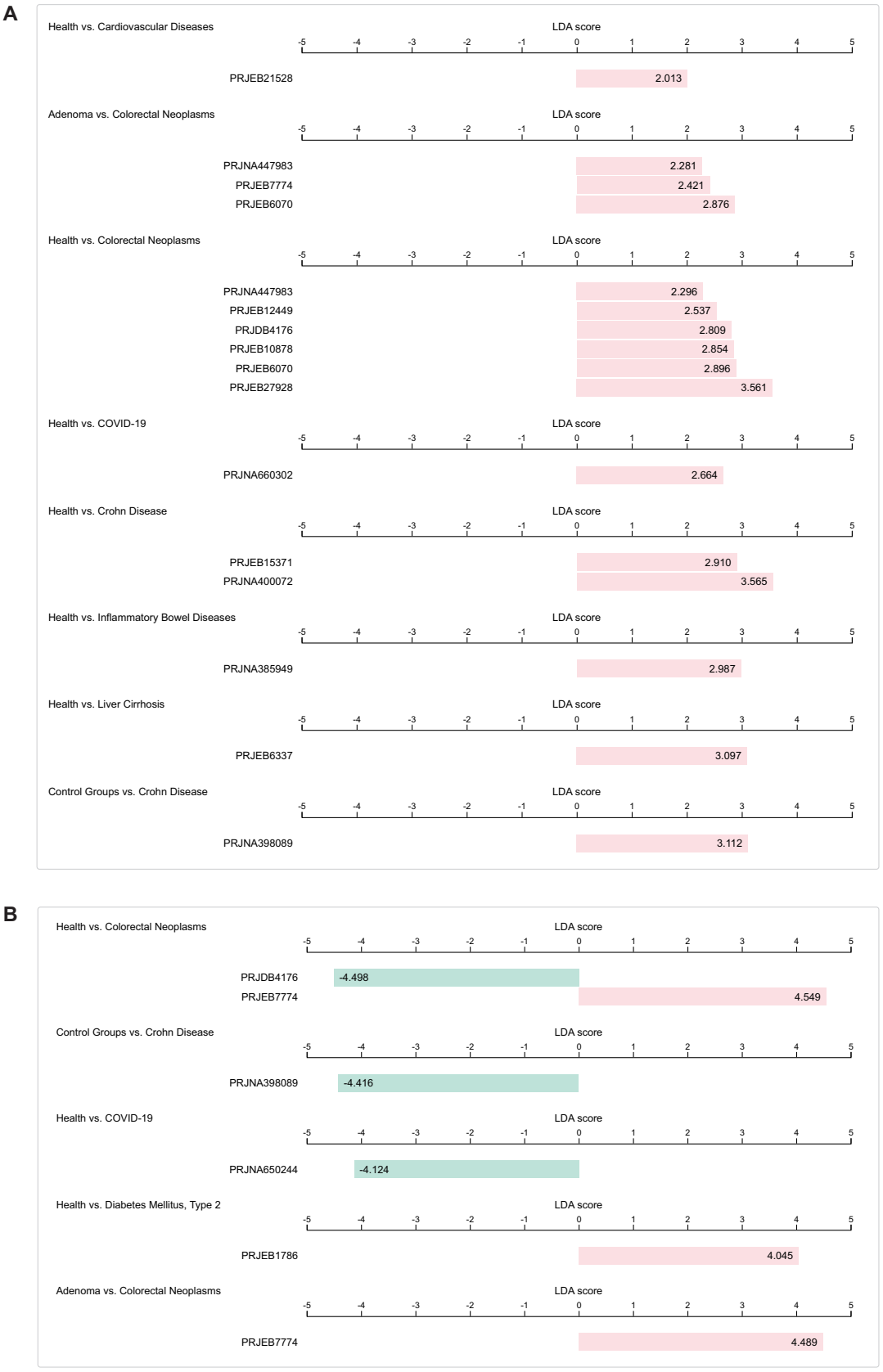

Figure 3. Cross-disease comparison of marker taxa. (**A**) Enrichment trends of *Fusobacterium nucleatum* across diseases and projects. (**B**) A marker-centric view of *Prevotella copri* across diseases and projects. Please consult https://gmrepo.humangut.info/taxon/851 and https://gmrepo.humangut.info/taxon/165179 for the online versions.

DATA AVAILABILITY

All data are freely accessible to all academic users. This work is licensed under a Creative Commons Attribution Non-Commercial 3.0 Unported License (CC BY-NC 3.0). Users can download dataset from the 'Data downloads' section of the 'Help' page. Users can also download individual datasets or combined datasets for individual project/phenotype/species via the 'Browse' page. We also provided programmable access through REST APIs. And users can obtain our datasets based on the detailed instructions on using R, Perl and Python at the 'Programmable access' section of the 'Help' page or our GitHub page: https://github.com/evolgeniusteam/GMrepoProgrammableAccess.

FUNDING

National Key Research and Development Program of China [2019YFA0905600 to W.H.C.]; National Natural Science Foundation of China [81803850 to D.D., 61932008, 61772368 to X.M.Z.]; National Key R&D Program of China [2020YFA0712403 to X.M.Z.]; Shanghai Science and Technology Innovation Fund [19511101404 to X.M.Z.]; Shanghai Municipal Science and Technology Major Project [2018SHZDZX01 to X.M.Z.]. Funding for open access charge: National Key Research and Development Program of China [2019YFA0905600 to W.H.C.].
Conflict of interest statement. None declared.

REFERENCES

1. Fan,Y. and Pedersen,O. (2021) Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.*, **19**, 55–71.
2. Rinninella,E., Raoul,P., Cintoni,M., Franceschi,F., Miggiano,G.A., Gasbarrini,A. and Mele,M.C. (2019) What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. *Microorganisms*, **7**, 14.
3. Thursby,E. and Juge,N. (2017) Introduction to the human gut microbiota. *Biochem. J.*, **474**, 1823–1836.
4. Almeida,A., Mitchell,A.L., Boland,M., Forster,S.C., Gloor,G.B., Tarkowska,A., Lawley,T.D. and Finn,R.D. (2019) A new genomic blueprint of the human gut microbiota. *Nature*, **568**, 499–504.
5. Maslowski,K.M. and Mackay,C.R. (2011) Diet, gut microbiota and immune responses. *Nat. Immunol.*, **12**, 5–9.
6. Yoo,J.Y., Groer,M., Dutra,S.V., Sarkar,A. and McSkimming,D.I. (2020) Gut microbiota and immune system interactions. *Microorganisms*, **8**, 1587.
7. Gomes,A.C., Hoffmann,C. and Mota,J.F. (2018) The human gut microbiota: metabolism and perspective in obesity. *Gut Microbes*, **9**, 308–325.
8. Yadav,M., Verma,M.K. and Chauhan,N.S. (2018) A review of metabolic potential of human gut microbiome in human nutrition. *Arch. Microbiol.*, **200**, 203–217.
9. Wang,Y., Shou,J.-W., Li,X.-Y., Zhao,Z.-X., Fu,J., He,C.-Y., Feng,R., Ma,C., Wen,B.-Y., Guo,F. *et al.* (2017) Berberine-induced bioactive metabolites of the gut microbiota improve energy metabolism. *Metabolism*, **70**, 72–84.
10. Bliss,E.S. and Whiteside,E. (2018) The gut-brain axis, the human gut microbiota and their integration in the development of obesity. *Front. Physiol.*, **9**, 900.
11. Meng,C., Bai,C., Brown,T.D., Hood,L.E. and Tian,Q. (2018) Human gut microbiota and gastrointestinal cancer. *Genomics Proteomics Bioinformatics*, **16**, 33–49.
12. Fei,N., Bruneau,A., Zhang,X., Wang,R., Wang,J., Rabot,S., Gérard,P., Zhao,L. and,Ruby Edward G. (2020) Endotoxin producers overgrowing in human gut microbiota as the causative agents for nonalcoholic fatty liver disease. *mBio*, **11**, e03263-19.
13. Jiang,P., Wu,S., Luo,Q., Zhao,X., Chen,W.-H. and Bucci,V. (2021) Metagenomic analysis of common intestinal diseases reveals relationships among microbial signatures and powers multidisease diagnostic models. *mSystems*, **6**, e00112-21.
14. Wu,S., Jiang,P., Zhao,X.M. and Chen,W.H. (2021) Treatment regimens may compromise gut-microbiome-derived signatures for liver cirrhosis. *Cell Metab.*, **33**, 455–456.
15. Ryan,F.J., Ahern,A.M., Fitzgerald,R.S., Laserna-Mendieta,E.J., Power,E.M., Clooney,A.G., O'Donoghue,K.W., McMurdie,P.J., Iwai,S., Crits-Christoph,A. *et al.* (2020) Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease. *Nat. Commun.*, **11**, 1512.
16. Lavelle,A. and Sokol,H. (2020) Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.*, **17**, 223–237.
17. Caruso,R., Lo,B.C. and Núñez,G. (2020) Host–microbiota interactions in inflammatory bowel disease. *Nat. Rev. Immunol.*, **20**, 411–426.
18. Wu,H., Tremaroli,V., Schmidt,C., Lundqvist,A., Olsson,L.M., Krämer,M., Gummesson,A., Perkins,R., Bergström,G. and Bäckhed,F. (2020) The gut microbiota in prediabetes and diabetes: a population-based cross-sectional study. *Cell Metab.*, **32**, 379–390.
19. Maffei,V.J., Kim,S., Blanchard,E.I.V., Luo,M., Jazwinski,S.M., Taylor,C.M. and Welsh,D.A. (2017) Biological aging and the human gut microbiota. *J. Gerontol. A*, **72**, 1474–1482.
20. Kim,Y.S., Unno,T., Kim,B.-Y. and Park,M.-S. (2019) Sex differences in gut microbiota. *wjmh*, **38**, 48–60.
21. Gao,X., Zhang,M., Xue,J., Huang,J., Zhuang,R., Zhou,X., Zhang,H., Fu,Q. and Hao,Y. (2018) Body mass index differences in the gut microbiota are gender specific. *Front. Microbiol.*, **9**, 1250.
22. Rothschild,D., Weissbrod,O., Barkan,E., Kurilshikov,A., Korem,T., Zeevi,D., Costea,P.I., Godneva,A., Kalka,I.N., Bar,N. *et al.* (2018) Environment dominates over host genetics in shaping human gut microbiota. *Nature*, **555**, 210–215.
23. Khachatryan,Z.A., Ktsoyan,Z.A., Manukyan,G.P., Kelly,D., Ghazaryan,K.A. and Aminov,R.I. (2008) Predominant role of host genetics in controlling the composition of gut microbiota. *PLoS One*, **3**, e3064.
24. Merra,G., Noce,A., Marrone,G., Cintoni,M., Tarsitano,M.G., Capacci,A. and De Lorenzo,A. (2021) Influence of mediterranean diet on human gut microbiota. *Nutrients*, **13**, 7.
25. Feng,J., Li,B., Jiang,X., Yang,Y., Wells,G.F., Zhang,T. and Li,X. (2018) Antibiotic resistome in a large-scale healthy human gut microbiota deciphered by metagenomic and network analyses. *Environ. Microbiol.*, **20**, 355–368.
26. Moeller,A.H. and Ochman,H. (2013) Factors that drive variation among gut microbial communities. *Gut Microbes*, **4**, 403–408.
27. Clemente,Jose C., Ursell,Luke K., Parfrey,Laura W. and Knight,R. (2012) The impact of the gut microbiota on human health: an integrative view. *Cell*, **148**, 1258–1270.
28. Robles Alonso,V. and Guarner,F. (2013) Linking the gut microbiota to human health. *Br. J. Nutr.*, **109**, S21–S26.
29. Conlon,M.A. and Bird,A.R. (2015) The impact of diet and lifestyle on gut microbiota and human health. *Nutrients*, **7**, 17–44.
30. Harrison,P.W., Ahamed,A., Aslam,R., Alako,B.T.F., Burgin,J., Buso,N., Courtot,M., Fan,J., Gupta,D., Haseeb,M. *et al.* (2021) The European Nucleotide Archive in 2020. *Nucleic Acids Res.*, **49**, D82–D85.
31. Leinonen,R., Sugawara,H., Shumway,M. and on behalf of the International Nucleotide Sequence Database, C. (2011) The Sequence Read Archive. *Nucleic Acids Res.*, **39**, D19–D21.
32. Mitchell,A.L., Almeida,A., Beracochea,M., Boland,M., Burgin,J., Cochrane,G., Crusoe,M.R., Kale,V., Potter,S.C., Richardson,L.J. *et al.* (2020) MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, **48**, D570–D578.
33. Shi,W., Qi,H., Sun,Q., Fan,G., Liu,S., Wang,J., Zhu,B., Liu,H., Zhao,F., Wang,X. *et al.* (2019) gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. *Nucleic Acids Res.*, **47**, D637–D648.
34. Gonzalez,A., Navas-Molina,J.A., Kosciolek,T., McDonald,D., Vázquez-Baeza,Y., Ackermann,G., DeReus,J., Janssen,S., Swafford,A.D., Orchanian,S.B. *et al.* (2018) Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods*, **15**, 796–798.

35. Cheng,L., Qi,C., Zhuang,H., Fu,T. and Zhang,X. (2020) gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.*, **48**, D554–D560.
36. Tang,J., Wu,X., Mou,M., Wang,C., Wang,L., Li,F., Guo,M., Yin,J., Xie,W., Wang,X. *et al.* (2021) GIMICA: host genetic and immune factors shaping human microbiota. *Nucleic Acids Res.*, **49**, D715–D722.
37. Janssens,Y., Nielandt,J., Bronselaer,A., Debunne,N., Verbeke,F., Wynendaele,E., Van Immerseel,F., Vandewynckel,Y.P., De Tre,G. and De Spiegeleer,B. (2018) Disbiome database: linking the microbiome to disease. *BMC Microbiol.*, **18**, 50.
38. Wu,S., Sun,C., Li,Y., Wang,T., Jia,L., Lai,S., Yang,Y., Luo,P., Dai,D., Yang,Y.-Q. *et al.* (2020) GMrepo: a database of curated and consistently annotated human gut metagenomes. *Nucleic Acids Res.*, **48**, D545–D553.
39. Pasolli,E., Schiffer,L., Manghi,P., Renson,A., Obenchain,V., Truong,D.T., Beghini,F., Malik,F., Ramos,M., Dowd,J.B. *et al.* (2017) Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods*, **14**, 1023–1024.
40. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
41. Bolyen,E., Rideout,J.R., Dillon,M.R., Bokulich,N.A., Abnet,C.C., Al-Ghalith,G.A., Alexander,H., Alm,E.J., Arumugam,M., Asnicar,F. *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, **37**, 852–857.
42. Callahan,B.J., McMurdie,P.J., Rosen,M.J., Han,A.W., Johnson,A.J. and Holmes,S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods*, **13**, 581–583.
43. Amir,A., McDonald,D., Navas-Molina,J.A., Kopylova,E., Morton,J.T., Zech Xu,Z., Kightley,E.P., Thompson,L.R., Hyde,E.R., Gonzalez,A. *et al.* (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *Msystems*, **2**, e00191-16.
44. Truong,D.T., Franzosa,E.A., Tickle,T.L., Scholz,M., Weingart,G., Pasolli,E., Tett,A., Huttenhower,C. and Segata,N. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, **12**, 902–903.
45. Segata,N., Izard,J., Waldron,L., Gevers,D., Miropolsky,L., Garrett,W.S. and Huttenhower,C. (2011) Metagenomic biomarker discovery and explanation. *Genome Biol.*, **12**, R60.
46. Schoch,C.L., Ciufo,S., Domrachev,M., Hotton,C.L., Kannan,S., Khovanskaya,R., Leipe,D., McVeigh,R., O'Neill,K., Robbertse,B. *et al.* (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, **2020**, baaa062.
47. Wirbel,J., Pyl,P.T., Kartal,E., Zych,K., Kashani,A., Milanese,A., Fleck,J.S., Voigt,A.Y., Palleja,A., Ponnudurai,R. *et al.* (2019) Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.*, **25**, 679–689.
48. Komiya,Y., Shimomura,Y., Higurashi,T., Sugi,Y., Arimoto,J., Umezawa,S., Uchiyama,S., Matsumoto,M. and Nakajima,A. (2019) Patients with colorectal cancer have identical strains of *Fusobacterium nucleatum* in their colorectal cancer and oral cavity. *Gut*, **68**, 1335.
49. Abed,J., Maalouf,N., Manson,A.L., Earl,A.M., Parhi,L., Emgård,J.E.M., Klutstein,M., Tayeb,S., Almogy,G., Atlan,K.A. *et al.* (2020) Colon cancer-associated *Fusobacterium nucleatum* may originate from the oral cavity and reach colon tumors via the circulatory system. *Front. Cell. Infect. Microbiol.*, **10**, 400.
50. Hashemi Goradel,N., Heidarzadeh,S., Jahangiri,S., Farhood,B., Mortezaee,K., Khanlarkhani,N. and Negahdari,B. (2019) Fusobacterium nucleatum and colorectal cancer: a mechanistic overview. *J. Cell. Physiol.*, **234**, 2337–2344.
51. Sun,C.-H., Li,B.-B., Wang,B., Zhao,J., Zhang,X.-Y., Li,T.-T., Li,W.-B., Tang,D., Qiu,M.-J., Wang,X.-C. *et al.* (2019) The role of *Fusobacterium nucleatum* in colorectal cancer: from carcinogenesis to clinical management. *Chronic Dis.Transl. Med.*, **5**, 178–187.
52. Wu,J., Li,Q. and Fu,X. (2019) Fusobacterium nucleatum contributes to the carcinogenesis of colorectal cancer by inducing inflammation and suppressing host immunity. *Transl. Oncol.*, **12**, 846–851.
53. Löwenmark,T., Löfgren-Burström,A., Zingmark,C., Eklöf,V., Dahlberg,M., Wai,S.N., Larsson,P., Ljuslinder,I., Edin,S. and Palmqvist,R. (2020) Parvimonas micra as a putative non-invasive faecal biomarker for colorectal cancer. *Sci. Rep.*, **10**, 15250.
54. Xu,J., Yang,M., Wang,D., Zhang,S., Yan,S., Zhu,Y. and Chen,W. (2020) Alteration of the abundance of *Parvimonas micra* in the gut along the adenoma–carcinoma sequence. *Oncol. Lett.*, **20**, 106.
55. Kwong,T.N.Y., Wang,X., Nakatsu,G., Chow,T.C., Tipoe,T., Dai,R.Z.W., Tsoi,K.K.K., Wong,M.C.S., Tse,G., Chan,M.T.V. *et al.* (2018) Association between bacteremia from specific microbes and subsequent diagnosis of colorectal cancer. *Gastroenterology*, **155**, 383–390.
56. Liang,J.Q., Li,T., Nakatsu,G., Chen,Y.-X., Yau,T.O., Chu,E., Wong,S., Szeto,C.H., Ng,S.C., Chan,F.K.L. *et al.* (2019) A novel faecal Lachnoclostridium marker for the non-invasive diagnosis of colorectal adenoma and cancer. *Gut*, **69**, 1248–1257.
57. Wu,Y., Jiao,N., Zhu,R., Zhang,Y., Wu,D., Wang,A.-J., Fang,S., Tao,L., Li,Y., Cheng,S. *et al.* (2021) Identification of microbial markers across populations in early detection of colorectal cancer. *Nat. Commun.*, **12**, 3063.
58. Wu,G.D., Chen,J., Hoffmann,C., Bittinger,K., Chen,Y.Y., Keilbaugh,S.A., Bewtra,M., Knights,D., Walters,W.A., Knight,R. *et al.* (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science (New York, N.Y.)*, **334**, 105–108.
59. Arumugam,M., Raes,J., Pelletier,E., Le Paslier,D., Yamada,T., Mende,D.R., Fernandes,G.R., Tap,J., Bruls,T., Batto,J.M. *et al.* (2011) Enterotypes of the human gut microbiome. *Nature*, **473**, 174–180.
60. Almeida,A., Nayfach,S., Boland,M., Strozzi,F., Beracochea,M., Shi,Z.J., Pollard,K.S., Sakharova,E., Parks,D.H., Hugenholtz,P. *et al.* (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.*, **39**, 105–114.