D808–D816 *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 28 October 2021*
https://doi.org/10.1093/nar/gkab973

mBodyMap: a curated database for microbes across human body and their associations with health and diseases

Hanbo Jin[1,†], Guoru Hu[1,†], Chuqing Sun[1,†], Yiqian Duan[2], Zhenmo Zhang[1], Zhi Liu[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 430074, China, [2]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China, [3]Department of Biotechnology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, 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 August 14, 2021; Revised September 29, 2021; Editorial Decision October 04, 2021; Accepted October 05, 2021

ABSTRACT

mBodyMap is a curated database for microbes across the human body and their associations with health and diseases. Its primary aim is to promote the reusability of human-associated metagenomic data and assist with the identification of disease-associated microbes by consistently annotating the microbial contents of collected samples using state-of-the-art toolsets and manually curating the meta-data of corresponding human hosts. mBodyMap organizes collected samples based on their association with human diseases and body sites to enable cross-dataset integration and comparison. To help users find microbes of interest and visualize and compare their distributions and abundances/prevalence within different body sites and various diseases, the mBodyMap database is equipped with an intuitive interface and extensive graphical representations of the collected data. So far, it contains a total of 63 148 runs, including 14 401 metagenomes and 48 747 amplicons related to health and 56 human diseases, from within 22 human body sites across 136 projects. Also available in the database are pre-computed abundances and prevalence of 6247 species (belonging to 1645 genera) stratified by body sites and diseases. mBodyMap can be accessed at: https://mbodymap.microbiome.cloud.

INTRODUCTION

Microbes inhabit almost all human body parts and play critical roles in human health and disease (1–4). The human microbiota is located primarily in the gut, where the numbers and diversity from the stomach to the colon multiply continuously (3,5). However, other anatomical parts, including the lungs, skin, vagina, eyes, placenta, ears, mouth and nasal compartments also harbor microbiomes (6,7). Microbiome's composition varies depending on the anatomy (e.g. between the intestine and the lungs), between individuals and even over time (4,8–10); it can be altered by dietary changes (including the use of probiotics, the use of antibiotics and other drugs (11–14), age (15) or diseases) and other factors and is also dynamic (16–19). For instance, the human skin microbiome is highly personalized, depending on multiple factors, such as body site, age, gender and lifestyle elements (20–23). In addition to individual microbes (e.g. known pathogenic bacteria), changes in the composition of microbes (i.e. dysbiosis) are increasingly observed in many diseases, like colorectal cancer (CRC), type 2 diabetes (T2D), and inflammatory bowel disease (IBD) (24). Therefore, the importance of maintaining a healthy microbiota has garnered attention over the years, although the exact definition of 'healthy microbiota' remains to be provided (25,26). This increased attention has seen

*To whom correspondence should be addressed. Tel: +86 158 2735 4263; Fax: +86 27 8779 2072; Email: weihuachen@hust.edu.cn
Correspondence may also be addressed to Xing-Ming Zhao. Email: xmzhao@fudan.edu.cn
Correspondence may also be addressed to Zhi Liu. Email: zhiliu@hust.edu.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

probiotics, prebiotics, and synbiotics developed and used to intervene in microbial dysbiosis and/or restore 'healthy microbiota' in cases of numerous diseases (13,27–38).

Public databases, such as HMDAD (the Human Microbe-Disease Association Database) (39), Disbiome (22) and MicroPhenoDB (40) that store associations between human diseases and microbes across body sites have been established. Table 1 summarized their main features. Briefly, HMDAD and Disbiome collect text-mining-based microbe-disease associations from peer-reviewed publications and determine the strength of these associations based on the credibility of the data sources. MicroPhenoDB harvests microbe-disease relationships from the HMDAD and Disbiome databases and other open resources and is, therefore, the largest database with associations between microbes and diseases so far. However, despite the valued contributions to microbe-disease associations that these databases provide, they tend to focus primarily on individual pathogenic microbes through text-mining and largely overlook vital contributions from the microbial community as a whole on health and diseases. A comprehensive collection of curated and consistently annotated metagenomic datasets to link human-related microbes within different sites of the whole body to health and diseases, therefore, remains unavailable.

In that regard, we developed mBodyMap, a curated database for microbes across 22 human body sites and their relationships with health and diseases. Overall, we collected 63 148 metagenomic samples/runs from both 16S rRNA and metagenomic next-generation sequencing (mNGS) across 136 projects. The core mBodyMap features include: (i) manually curated healthy and diseased information for each collected run/sample and all possible related meta-data, such as age, sex, country and body-mass-index (BMI); (ii) consistently annotated microbial contents, including taxonomic assignments of sequencing reads and precomputed species/genus relative abundances using state-of-the-art toolsets; (iii) collected samples organized based on their associated health control and diseases, sample harvesting body sites and statistics, including species-prevalence and abundances; (iv) equipment with an intuitive graphical representation of the distributions and abundances/prevalence of microbes across the human body that enables users to browse the distribution of microbes across the human body and compare microbes' distribution among various diseases and health intuitively.

DATABASE CONSTRUCTION

Data collection of sequencing reads and manual curation of associated meta-data

To identify human-related metagenomic datasets, we systematically searched public databases, including the NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject/) and EBI ENA (41) (European Nucleotide Archive, https://www.ebi.ac.uk/ena) and manually examined related project information to determine the accuracy of datasets as human-associated metagenomic datasets.

Next, we downloaded the raw sequencing data from EBI ENA (41) and NCBI SRA (42) (Sequence Read Archive, https://www.ncbi.nlm.nih.gov/sra) using enaBrowserTools (https://github.com/enasequence/enaBrowserTools) and SRA-Tools (https://github.com/ncbi/sra-tools) facilitated by Aspera (a high-speed data transfer tool). For each run and sample, we also downloaded relevant meta-data, including technical metadata, such as the sequencing platform, number of reads, and read length, and biological metadata, such as the body site from which the samples were taken, as well as the age, gender, country, body mass index (BMI), and disease(s) of the human host. We manually curated the meta-data twice: round one consisted of manually inspecting the extracted meta-information with the help of in-house R or Perl scripts to find all meta-data of interest; if necessary, the related publication(s), supplementary materials, and even the corresponding authors were consulted. During the second round of manual curation, different curators from the first round reviewed the collected meta-data and made necessary corrections.

We stratified samples according to their associated human health or disease and body sites from which the samples were harvested. The body sites in question are as follows: ear, nose, oral, trachea, esophagus, upper respiratory tract, lung, stomach, uterus, cervix, fallopian tube, ovary, vagina, urethra, skin, blood, peritoneal fluid, large intestine and small intestine.

Processing of raw sequencing reads

We processed the downloaded raw sequencing reads in FASTQ format using FastQC (v0.11.8, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to evaluate the quality and Trimmomatic (43) to remove low-quality bases and sequencing vectors. Sequences shorter than two-thirds of the original read length were excluded from subsequent analyses.

For 16S sequences, we used single-ended sequencing reads directly in ensuing analyses but merged pair-ended reads using Casper (44) v0.8.2 at default parameters before subsequently analyzing them. Metagenomic sequences comprising single-ended and pair-ended sequencing reads were all underwent subsequent analyses directly.

We referred to the resulting sequences as 'clean data' and used them for further scrutiny. We also used Seqtk (https://github.com/lh3/seqtk) to convert FASTQ sequences to FASTA formats at default parameters if necessary.

Taxonomic assignment of processed sequencing reads and the calculation of relative abundances

For 16S sequences, we used MAPseq v1.2 (45) to analyze the clean data and assign taxonomic classification information to the reads. As indicated by the creators of MAPseq, we applied a combined score of 0.4 at the genus level to assign the taxonomic classification. For each sample/run, the relative abundances at the genus and species levels were subsequently calculated, with total abundance values of 100%.

For metagenomic sequences, we used MetaPhlAn2 (46) at default parameters to assign taxonomic classification data to the sequencing reads and calculate relative abundances at species and genus levels.

Table 1. Key features of mBodyMap and comparison with similar databases on microbe-human disease associations

Database	Key features	Data source	Data size		Reference
			# Disease	# Microbe	
mBodyMap	• Comprehensive collection of metagenomic data and analysis using state-of-the-art tools • Careful curation of human-related meta-data such as diseases and health • Disease-centric organization of pre-calculated microbial abundance data across the body sites and diseases • Intuitive graphical interface and extensive visualization of the microbial profiles	Metagenomics data	56	6247	This study
HMDAD	• Text mining in large quantity of publications followed by manual curation • Construction of a microbe-based human disease network	Text-mining	39	292	(39)
Disbiome	• Collection and presentation of published microbiota-disease information in a standardized way • Assessment for each study's reporting quality using a standardized questionnaire	Text-mining	372	1622	(22)
MicroPhenoDB	• Provision of non-redundant associations between microbes and human disease phenotypes across human body and relationships between unique clade-specific core genes and microbes • Development of a refined score model to prioritize the associations based on evidential metrics	Text-mining, HMDAD and Disbiome	542	1781	(40)

Quality controls for samples/runs

We conducted sample/run level quality control to guarantee the quality of our data: first, we excluded amplicon samples/runs with <5000 reads from subsequent analyses and marked them as 'failed QC (QC status = 0)' in mBodyMap and then ensured samples/runs contained only a single taxon, i.e. we also marked a species or a genus accounting for more than or equal to 99.99% of total abundance as 'failed QC (QC status = 0)'.

Database construction and web development

We loaded all data into the MySQL v5.7.25 (https://www.mysql.com/) database and coded the frontend (the webpages) of the website using HTML and JavaScript and the backend using Python v3.7.7 (https://www.python.org/) with a Flask v1.1.2 (https://flask.palletsprojects.com/) framework to support queries to the MySQL database. We bridged the front- and back- ends using the Vue.js v 2.6.12 (https://cn.vuejs.org/) framework and visualized the front-end with plotly.js v1.58.4 (https://github.com/plotly/plotly.js/). We also used several other open-source JavaScript libraries, including Element UI v2.15.1 (https://element.eleme.io/) and BootstrapVue v2.21.2 (https://code.z01.com/bootstrap-vue/). The website is hosted on an Apache v2.4.29 (https://www.apache.org/) server.

DATABASE OVERVIEW AND FUNCTIONALITY

Overview of mBodyMap

So far, mBodyMap contains 63 148 runs, including 14 401 metagenomic and 48 747 amplicon runs relating to health and 56 human diseases, linked to 22 human body sites across 136 projects (Figure 1A). Of the total, we considered 61 913 runs 'valid runs' based on our quality controls and subsequent analysis processes.

Through multiple rounds of manual curation, we assigned clear healthy or disease information to almost all collected samples, subsequently describing and organizing these information using the MeSH system (Medical Subject Headings, a controlled and hierarchically organized vocabulary produced by the National Library of Medicine). We identified health- and 56 diseases-related information from the microbiome data. Table 2 contains health and the top 10 diseases included in mBodyMap; they are ranked by the number of samples/runs they are linked-to in our database.

We also strived to collect as voluminous meta-data as possible for the microbiome datasets; however, our ardent efforts yielded only three most basic host details: the age, sex, and BMI of a very small proportion (3.97%) of the samples (Figure 1B). 22.61% of the samples contained none of the basic meta-data, while the rest contained only one or two (64.23% and 9.19%, respectively) (Figure 1C). These results are consistent with our previous discovery in gut microbiome datasets (47). They indicate the difficulties in reusing metagenomic information and call for detailing guidelines of meta-information or metagenomic samples.

We identified 6247 species belonging to 1645 genera from the 61 913 valid runs in our database, with 3710 of the species belonging to 1075 genera identified in more than one sample each (with a median relative abundance higher than 0.01% within one or more health/diseases); these results match our previous finding from gut metagenome analysis that about ∼50% of microbes are specific to individuals (47). While the prevalence of most species is low, our results demonstrate that a small number of runs contained massive amounts of taxa under abundances limitation, expanding the recognized microbiota species in various parts of the human body. We believe that further analyses of samples will

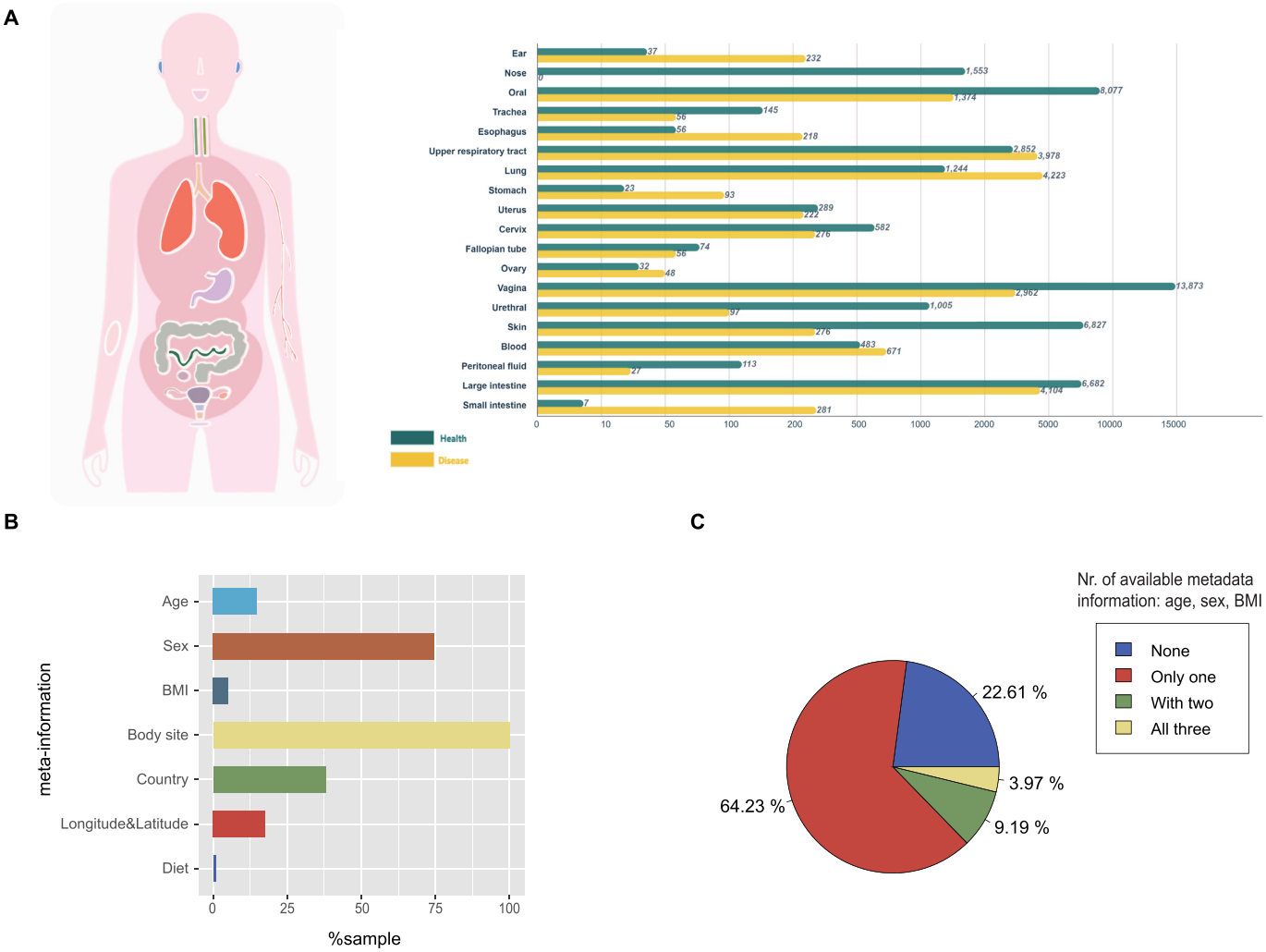

Figure 1. Overview of data in mBodyMap. (**A**) The left panel contains an interactive body map indicating clickable body sites for which metagenomic data are available; the right panel contains the number of samples for each body site, stratified by health (dark green) and diseases (yellow). (**B**) A barplot summarizing the meta-data we have collected for samples. The Y-axis represents meta-information, and the X-axis denotes the proportion of the samples comprising this meta-information. (**C**) The integrity of the metadata assessed based on age, sex and BMI.

Table 2. Statistics of health and the top 10 diseases included in mBodyMap

Health/disease	No. of associated sites	No. of processed runs	No. of valid runs	No. of associated species	No. of associated genera
Health	21	42 816	36 852	6070	1623
Respiratory tract infections	3	2357	2274	3525	1103
Cystic fibrosis	1	2129	1656	4569	1353
Pouchitis	2	1858	889	3621	1190
Bacterial vaginosis	1	1541	1538	3775	1227
Chronic obstructive pulmonary disease	3	1174	1084	4122	1292
Premature birth	2	1137	1110	3040	952
Necrotizing enterocolitis	1	1094	659	1037	323
Asthma	1	870	850	3654	1196
Crohn disease	2	714	398	1189	423
Endometrial neoplasms	8	660	604	2835	999

No. of associated sites: the number of body sites from which the sample with this health/disease was harvested.
No. of processed runs: the number of all runs with processed sequence data; all the runs are processed eventually.
No. of valid runs: the number of runs whose data passed our quality control procedure, with the corresponding species/genus relative abundances available in our database.
No. of associated species: the number of species associated with processed and valid runs.
No. of associated genera: the number of genera associated with processed and valid runs.

yield an increase in the total number of species/strains in various parts of the human body.

Web usage

mBodyMap provides a user-friendly and interactive portal for browsing and querying metagenomic data and related information. To help researchers find body site-health/disease associations, mBodyMap provides users with two search options: one requires a click on the directives depicting body sites of interest on the picture of a human body on the front page to view associated health or diseases, and the other demands choosing between health/diseases on the 'Health&Disease' page to view related body sites. For each body site-health/disease pair, we provide information about related projects and samples/runs and the associated species/genera and their relative abundances and prevalence in related samples. For example, to see related details on Chronic Obstructive Pulmonary Disease of the lung, users can select Chronic Obstructive Pulmonary Disease on the 'Health&Disease' page and then the lung as the body site, for which the query result will show that there are 4026 associated microbial species assigned to 1270 genera. Of these, we identified only 274 species (∼6.81% of the total) assigned to 86 genera (∼6.77% of the total) in more than one sample, with a median relative abundance higher than 0.01%. See https://mbodymap.microbiome.cloud/#/health&diseases/Lung/Chronic%20Obstructive%20Pulmonary%20Disease/D029424 for more details. Users may then select a species, such as *Streptococcus mitis*, to access further information, including its distribution and abundances in healthy and diseased samples; for more details, see https://mbodymap.microbiome.cloud/#/taxon/species/Lung/D029424/28037. The 'Taxa' page that includes 'Species' and 'Genera' pages is available to users who can browse through a microbe of interest to view the body sites it inhabits and the health or diseases it is associated with.

The 'Data' page provides the manually curated metadata of metagenomic projects and samples/runs for users to download. Additional links to NCBI BioProject, NCBI SRA, and NCBI MeSH Browsers for each of the projects, runs, and health/diseases are available to help researchers download data and acquire more material. Furthermore, for each microbial taxon (i.e. species and genus), we have included links to the corresponding pages (if available) in public databases, such as NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) (48), GMrepo (a comprehensive gut microbiome database stratified by human phenotypes) (47), and MVP (a microbe-phage association database) (49). We intend to create more links to external databases as we continue to improve the site.

Species relative abundance and prevalence within and across diseases and body sites

With the availability of pre-calculated relative abundances for all valid runs in mBodyMap, users can visualize the prevalence of microbes of interest in different diseases; for comparisons, the distributions of the microbes in healthy individuals are also provided. Figure 2A presents the distribution of *Haemophilus parainfluenzae*: a barplot is used to depict its prevalence in health and ten diseases associated with the upper respiratory tract (see also https://mbodymap.microbiome.cloud/#/taxon/species/Upper%20respiratory%20tract/729; by default, diseases with >10 valid runs are included in this barplot). Additionally, we visualized its relative abundances across selected body sites in healthy controls and other diseases and compared the outcomes in a box plot (by default, diseases with >10 valid runs are included in this box plot; Figure 2B). To better illustrate the proportions of samples under different relative abundance thresholds for a species/genera across each body site, we created a line plot whose Y-axis represents the percentage of runs per all valid runs within certain ranges of relative abundances and whose X-axis denotes the threshold of relative abundances. The line plot displays the distribution of the relative abundances of selected taxons across selected body sites (Figure 2C).

With mBodyMap, users can also explore the distribution of microbes of interest across body sites. Figure 3 shows a graphical representation of the human body used to show the abundances and prevalence of *Streptococcus mitis* across human body sites. We used different colors to represent various relative abundance and prevalence levels, enabling users to browse the distribution of microbes across the human body intuitively. The distributions of microbes of interest in both healthy and diseased sites are shown side by side. In our database, we identified *S. mitis* in 22 body sites and associated them with 55 diseases (https://mbodymap.microbiome.cloud/#/taxon/species/28037). The relative abundances and prevalence of *S. mitis* were higher in multiple body sites of the diseased population than in the corresponding healthy sites, which is consistent with the characterization of *S. mitis* as a pathogenic bacterium (Figure 3).

FUTURE DIRECTIONS

In addition to continuously collecting new metagenomic data of various human body sites over the next few years, we plan to add new contents to mBodyMap, including (but not limited to) viral abundances, functional profiles, and metabolic pathway profiles of the collected samples. We also plan to include more functions that allow users to perform on-site cross-sample comparisons, differential abundance analyses, and mathematical modeling. Furthermore, we will aim to identify body site-specific or enriched species and microbial disease markers and compare them across datasets and projects. We have used the LEfSe (linear discriminant analysis effect size) (50) method to identify the marker microbes between health control and diseases in certain projects and visualized them on the web page; see the 'in-depth analysis' section of the following page for an example: https://mbodymap.microbiome.cloud/#/data/project/PRJNA275918. This feature will be available for all projects in the future. These developments should promote the reusability and accessibility of human metagenomic data further and help users better understand the relationship between the dysbiosis of microbiota at multiple body sites and human diseases.

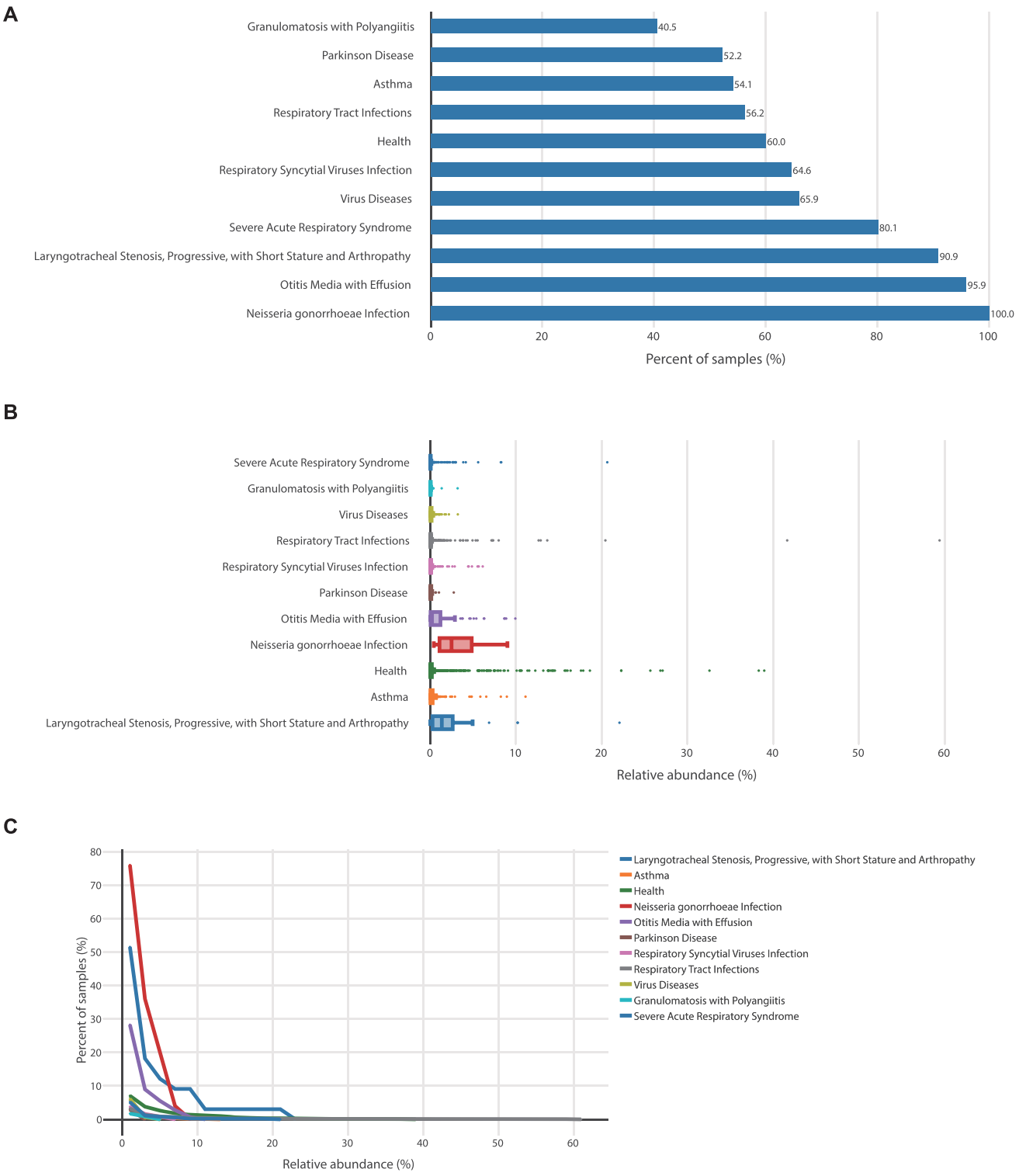

Figure 2. Graphical representation of the abundances, prevalence, and distributions within health and diseases of a selected taxon. Here, *Haemophilus parainfluenzae* at the upper respiratory tract is used as an example. (**A**) Its prevalence across health and multiple diseases. The Y-axis represents health and various diseases, and the X-axis denotes the proportion of the samples comprising this health or disease. (**B**) The box plot's Y-axis representing health and other diseases and its X-axis denoting relative abundances. (**C**) Its distributions among health and various diseases.

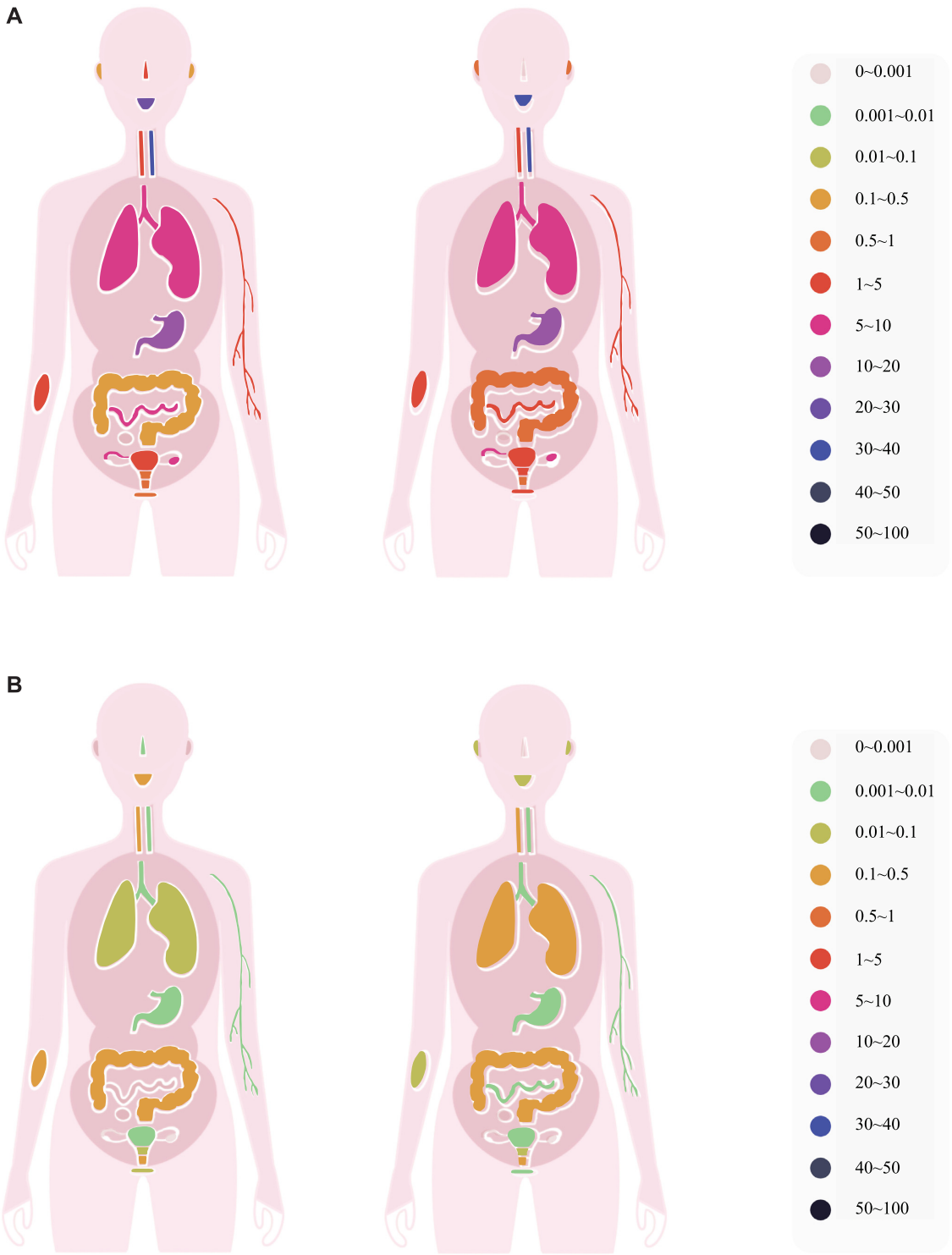

Figure 3. Distribution of *Streptococcus mitis*, a known disease-causing bacterium, across body sites in mBodyMap. Display of the relative abundance (**A**) and prevalence (**B**) of *S. mitis* in various sites of healthy and diseased human bodies. *S. mitis* was isolated in significant abundances in multiple body sites of the diseased population, which is consistent with its characterization as a pathogenic bacterium.

CONCLUSION

This article introduces mBodyMap, a curated database for microbes across the human body and their associations with health and diseases. So far, mBodyMap contains 63 148 runs, including 14 401 metagenomes and 48 747 amplicons relating to health and 56 human diseases, linked to 22 human body sites across 136 projects. We aim to provide a central resource for curated and consistently annotated microbes from various human body sites, which would allow users to quickly find microbes of interest and visualize their distributions across the human body and facilitate the identification of site- and/or disease-specific marker microbes. We collected the metagenomic datasets of human samples from multiple sources, manually curated their meta-data, and annotated their microbial contents using state-of-the-art toolsets. We then stratified samples according to the human health or diseases and body sites they are linked to and pre-computed species/genus relative abundances and prevalence. As compared with existing databases on microbe-human disease associations, mBodyMap focuses on metagenomics data and highlights the important roles of the microbial community as a whole in health and diseases. In the future, we will add more data and functions to mBodyMap.

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 view data and associated information from many web pages and download all data from the 'Data downloads' section of the 'Help' page. Programmable access through REST APIs is also supported: detailed instructions on using R and Python to access our data can be found at the 'Programmable access' section of the 'Help' page or our GitHub page: https://github.com/whchenlab/mBodymap/tree/main/programmable-access.

FUNDING

National Key Research and Development Program of China [2019YFA0905600 to W.H.C.]; National Key R&D Program of China [2020YFA0712403 to X.M.Z., in part]; National Natural Science Foundation of China [61932008, 61772368 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. Qin,J., Li,R., Raes,J., Arumugam,M., Burgdorf,K.S., Manichanh,C., Nielsen,T., Pons,N., Levenez,F., Yamada,T. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.
2. Turnbaugh,P.J., Ley,R.E., Mahowald,M.A., Magrini,V., Mardis,E.R. and Gordon,J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1031.
3. Ghaisas,S., Maher,J. and Kanthasamy,A. (2016) Gut microbiome in health and disease: Linking the microbiome-gut-brain axis and environmental factors in the pathogenesis of systemic and neurodegenerative diseases. *Pharmacol. Ther.*, **158**, 52–62.
4. Cho,I. and Blaser,M.J. (2012) The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.*, **13**, 260–270.
5. Tojo,R., Suárez,A., Clemente,M.G., de los Reyes-Gavilán,C.G., Margolles,A., Gueimonde,M. and Ruas-Madiedo,P. (2014) Intestinal microbiota in health and disease: role of bifidobacteria in gut homeostasis. *World J. Gastroenterol.*, **20**, 15163–15176.
6. The Integrative HMP (iHMP) Research Network Consortium. (2019) The integrative human microbiome project. *Nature*, **569**, 641–648.
7. Human Microbiome Project Consortium. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.
8. Ursell,L.K., Clemente,J.C., Rideout,J.R., Gevers,D., Caporaso,J.G. and Knight,R. (2012) The interpersonal and intrapersonal diversity of human-associated microbiota in key body sites. *J. Allergy Clin. Immunol.*, **129**, 1204–1208.
9. Tito,R.Y., Chaffron,S., Caenepeel,C., Lima-Mendez,G., Wang,J., Vieira-Silva,S., Falony,G., Hildebrand,F., Darzi,Y., Rymenans,L. *et al.* (2019) Population-level analysis of Blastocystis subtype prevalence and variation in the human gut microbiota. *Gut*, **68**, 1180–1189.
10. Brial,F., Chilloux,J., Nielsen,T., Vieira-Silva,S., Falony,G., Andrikopoulos,P., Olanipekun,M., Hoyles,L., Djouadi,F., Neves,A.L. *et al.* (2021) Human and preclinical studies of the host–gut microbiome co-metabolite hippurate as a marker and mediator of metabolic health. *Gut*, **70**, 2105–2114.
11. Forslund,K., Hildebrand,F., Nielsen,T., Falony,G., Chatelier,E.L., Sunagawa,S., Prifti,E., Vieira-Silva,S., Gudmundsdottir,V., Pedersen,H.K. *et al.* (2015) Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature*, **528**, 262–266.
12. Vieira-Silva,S., Falony,G., Belda,E., Nielsen,T., Aron-Wisnewsky,J., Chakaroun,R., Forslund,S.K., Assmann,K., Valles-Colomer,M., Nguyen,T.T.D. *et al.* (2020) Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature*, **581**, 310–315.
13. Wu,H., Esteve,E., Tremaroli,V., Khan,M.T., Caesar,R., Mannerås-Holm,L., Ståhlman,M., Olsson,L.M., Serino,M., Planas-Fèlix,M. *et al.* (2017) Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat. Med.*, **23**, 850–858.
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. Huang,S., Haiminen,N., Carrieri,A.-P., Hu,R., Jiang,L., Parida,L., Russell,B., Allaband,C., Zarrinpar,A., Vázquez-Baeza,Y. *et al.* (2020) Human skin, oral, and gut microbiomes predict chronological age. *Msystems*, **5**, e00630-19.
16. Caporaso,J.G., Lauber,C.L., Costello,E.K., Berg-Lyons,D., Gonzalez,A., Stombaugh,J., Knights,D., Gajer,P., Ravel,J., Fierer,N. *et al.* (2011) Moving pictures of the human microbiome. *Genome Biol.*, **12**, R50.
17. Zhernakova,A., Kurilshikov,A., Bonder,M.J., Tigchelaar,E.F., Schirmer,M., Vatanen,T., Mujagic,Z., Vila,A.V., Falony,G., Vieira-Silva,S. *et al.* (2016) Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, **352**, 565–569.
18. Van Rossum,T., Ferretti,P., Maistrenko,O.M. and Bork,P. (2020) Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.*, **18**, 491–506.
19. Hildebrand,F., Moitinho-Silva,L., Blasche,S., Jahn,M.T., Gossmann,T.I., Huerta-Cepas,J., Hercog,R., Luetge,M., Bahram,M., Pryszlak,A. *et al.* (2019) Antibiotics-induced monodominance of a novel gut bacterial order. *Gut*, **68**, 1781–1790.
20. Hillebrand,G.G., Dimitriu,P., Malik,K., Park,Y., Qu,D., Mohn,W.W. and Kong,R. (2021) Temporal variation of the facial skin microbiome: a 2-year longitudinal study in healthy adults. *Plast. Reconstr. Surg.*, **147**, 50S–61S.
21. Li,Z., Bai,X., Peng,T., Yi,X., Luo,L., Yang,J., Liu,J., Wang,Y., He,T., Wang,X. *et al.* (2020) New insights into the skin microbial communities and skin aging. *Front. Microbiol.*, **11**, 565549.

22. Janssens,Y., Nielandt,J., Bronselaer,A., Debunne,N., Verbeke,F., Wynendaele,E., Van Immerseel,F., Vandewynckel,Y.-P., De Tré,G. and De Spiegeleer,B. (2018) Disbiome database: linking the microbiome to disease. *BMC Microbiol.*, **18**, 50.
23. Noguera-Julian,M., Rocafort,M., Guillén,Y., Rivera,J., Casadellà,M., Nowak,P., Hildebrand,F., Zeller,G., Parera,M., Bellido,R. *et al.* (2016) Gut microbiota linked to sexual preference and HIV infection. *EBioMedicine*, **5**, 135–146.
24. Kushugulova,A., Forslund,S.K., Costea,P.I., Kozhakhmetov,S., Khassenbekova,Z., Urazova,M., Nurgozhin,T., Zhumadilov,Z., Benberin,V., Driessen,M. *et al.* (2018) Metagenomic analysis of gut microbial communities from a Central Asian population. *BMJ Open*, **8**, e021682.
25. Rinninella,E., Raoul,P., Cintoni,M., Franceschi,F., Miggiano,G.A.D., 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.
26. Eisenstein,M. (2020) The hunt for a healthy microbiome. *Nature*, **577**, S6–S8.
27. Lloyd-Price,J., Arze,C., Ananthakrishnan,A.N., Schirmer,M., Avila-Pacheco,J., Poon,T.W., Andrews,E., Ajami,N.J., Bonham,K.S., Brislawn,C.J. *et al.* (2019) Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, **569**, 655–662.
28. 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.
29. Morgan,X.C., Tickle,T.L., Sokol,H., Gevers,D., Devaney,K.L., Ward,D.V., Reyes,J.A., Shah,S.A., LeLeiko,N., Snapper,S.B. *et al.* (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.*, **13**, R79.
30. Hall,A.B., Yassour,M., Sauk,J., Garner,A., Jiang,X., Arthur,T., Lagoudas,G.K., Vatanen,T., Fornelos,N., Wilson,R. *et al.* (2017) A novel *Ruminococcus gnavus* clade enriched in inflammatory bowel disease patients. *Genome Medicine*, **9**, 103.
31. Jiang,P., Wu,S., Luo,Q., Zhao,X.M. and Chen,W.H. (2021) Metagenomic analysis of common intestinal diseases reveals relationships among microbial signatures and powers multidisease diagnostic models. *Msystems*, **6**, e00112-21.
32. Jiang,P., Lai,S., Wu,S., Zhao,X.-M. and Chen,W.-H. (2020) Host DNA contents in fecal metagenomics as a biomarker for intestinal diseases and effective treatment. *BMC Genomics*, **21**, 348.
33. 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.
34. Kostic,A.D., Chun,E., Meyerson,M. and Garrett,W.S. (2013) Microbes and inflammation in colorectal cancer. *Cancer Immunol. Res.*, **1**, 150–157.
35. 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.
36. Arthur,J.C., Gharaibeh,R.Z., Mühlbauer,M., Perez-Chanona,E., Uronis,J.M., McCafferty,J., Fodor,A.A. and Jobin,C. (2014) Microbial genomic analysis reveals the essential role of inflammation in bacteria-induced colorectal cancer. *Nat. Commun.*, **5**, 4724.
37. Sedighi,M., Razavi,S., Navab-Moghadam,F., Khamseh,M.E., Alaei-Shahmiri,F., Mehrtash,A. and Amirmozafari,N. (2017) Comparison of gut microbiota in adult patients with type 2 diabetes and healthy individuals. *Microb. Pathog.*, **111**, 362–369.
38. Qin,J., Li,Y., Cai,Z., Li,S., Zhu,J., Zhang,F., Liang,S., Zhang,W., Guan,Y., Shen,D. *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55–60.
39. Ma,W., Zhang,L., Zeng,P., Huang,C., Li,J., Geng,B., Yang,J., Kong,W., Zhou,X. and Cui,Q. (2017) An analysis of human microbe-disease associations. *Brief. Bioinform.*, **18**, 85–97.
40. Yao,G., Zhang,W., Yang,M., Yang,H., Wang,J., Zhang,H., Wei,L., Xie,Z. and Li,W. (2021) MicroPhenoDB associates metagenomic data with pathogenic microbes, microbial core genes, and human disease phenotypes. *Genomics Proteomics Bioinformatics*, **18**, 760–772.
41. Harrison,P.W., Alako,B., Amid,C., Cerdeño-Tárraga,A., Cleland,I., Holt,S., Hussein,A., Jayathilaka,S., Kay,S., Keane,T. *et al.* (2019) The European Nucleotide Archive in2018. *Nucleic. Acids. Res.*, **47**, D84–D88.
42. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic. Acids. Res.*, **40**, D54–D56.
43. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
44. Kwon,S., Lee,B. and Yoon,S. (2014) CASPER: context-aware scheme for paired-end reads from high-throughput amplicon sequencing. *BMC Bioinformatics*, **15**(Suppl. 9), S10.
45. Rodrigues,Matias, Schmidt,J.F., Tackmann,T.S.B. and Mering,von, C. (2017) MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics*, **33**, 3808–3810.
46. Segata,N., Waldron,L., Ballarini,A., Narasimhan,V., Jousson,O. and Huttenhower,C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods*, **9**, 811–814.
47. 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.
48. 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 (Oxford)*, **2020**, baaa062.
49. Gao,N.L., Zhang,C., Zhang,Z., Hu,S., Lercher,M.J., Zhao,X.-M., Bork,P., Liu,Z. and Chen,W.-H. (2018) MVP: a microbe–phage interaction database. *Nucleic Acids Res.*, **46**, D700–D707.
50. 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.