Published online 3 October 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D1345–D1352***
https://doi.org/10.1093/nar/gkac851

microbioTA: an atlas of the microbiome in multiple disease tissues of *Homo sapiens* and *Mus musculus*

Ping Wang[1,†], Sainan Zhang[1,†], Guoyou He[1,†], Meiyu Du[1], Changlu Qi[1], Ruyue Liu[1], Siyuan Zhang[2], Liang Cheng[1,3,*], Lei Shi[3,*] and Xue Zhang[3,4,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, Heilongjiang, China, [2]Department of Anatomy, College of Basic Medical Sciences, Harbin Medical University, Harbin 150081, Heilongjiang, China, [3]NHC Key Laboratory of Molecular Probes and Targeted Diagnosis and Therapy, Harbin Medical University, Harbin 150028, Heilongjiang, China and [4]McKusick-Zhang Center for Genetic Medicine, State Key Laboratory of Medical Molecular Biology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China

Received August 05, 2022; Revised September 09, 2022; Editorial Decision September 18, 2022; Accepted September 30, 2022

ABSTRACT

***microbioTA* (http://bio-annotation.cn/microbiota) was constructed to provide a comprehensive, user-friendly resource for the application of microbiome data from diseased tissues, helping users improve their general knowledge and deep understanding of tissue-derived microbes. Various microbes have been found to colonize cancer tissues and play important roles in cancer diagnoses and outcomes, with many studies focusing on developing better cancer-related microbiome data. However, there are currently no independent, comprehensive open resources cataloguing cancer-related microbiome data, which limits the exploration of the relationship between these microbes and cancer progression. Given this, we propose a new strategy to re-align the existing next-generation sequencing data to facilitate the mining of hidden sequence data describing the microbiome to maximize available resources. To this end, we collected 417 publicly available datasets from 25 human and 14 mouse tissues from the Gene Expression Omnibus database and use these to develop a novel pipeline to re-align microbiome sequences facilitating in-depth analyses designed to reveal the microbial profile of various cancer tissues and their healthy controls. *microbioTA* is a user-friendly online platform which allows users to browse, search, visualize, and download microbial abundance data from various tissues along with corresponding analysis results, aimimg at providing a reference for cancer-related microbiome research.**

INTRODUCTION

Microorganisms colonize the gut, skin, oral cavity (OC), urine, and various other environments across the host body, interacting with their hosts in complex ways often playing important roles in maintaining the health of the host (1–3). This means that specific changes in the microbiome can be closely linked to both disease progress and therapeutic response (4). For example, gut microbiota are amongst the most important risk factors for developing inflammatory bowel diseases (IBD) (5–7) while several other studies have linked their metabolites to the development and progression of various cardiovascular diseases (8–10). In addition, our understanding of the association between specific diseases and microorganisms has increased the interest in microbiome research as these interactions clearly impact both disease and therapeutic outcomes (11,12).

With the rapid development of next-generation sequencing techniques, researchers have already been able to study the genomic characteristics of various diseased tissue and the microorganisms residing within these samples. Increasing evidence suggests that microbes' composition of these tissues may impact their susceptibility to certain cancers and influence the host's response to therapy (13–15). Poore *et al.* re-examined microbial reads from 33 types of cancer from The Cancer Genome Atlas (TCGA) and found that microbiomes derived from blood and other tissues could be applied to cancer diagnosis (16). Riquelme *et al.* analyzed the composition of the microbiome in patients with pancreatic adenocarcinoma (PDAC) according to their short-term survival (STS) and long-term survival (LTS) state. These evaluations revealed that the microbiome from the LTS group had higher alpha-diversity, and an intra-tumoral signature which included *Pseudoxanthomonas*, *Streptomyces*,

[*]To whom correspondence should be addressed. Tel: +86 153 0361 4540; Email: liangcheng@hrbmu.edu.cn
Correspondence may also be addressed to Xue Zhang. Email: xuezhang@hrbmu.edu.cn
Correspondence may also be addressed to Lei Shi. Email: shilei_0328@hrbmu.edu.cn

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Saccharopolyspora and *Bacillus clausii*, with this signature also being found to be highly predictive of long-term survivorship. This data also led these investigators to suggest that the microbiome from PDAC tissue undertook some degree of crosstalk with the gut microbiome, influencing the host immune response and disease treatment (17).

As our understanding of microbiome has deepened, the demand for microbial data resources has intensified. The *NIH Human Microbiome Project* (HMP) (18), which has been ongoing for over a decade, provides resources, methods and discoveries that link interactions between humans and the microbiome to health-related outcomes, while the *Microbiome Database* (MDB) (19) contains the sequencing resources and metadata from various ecological community samples to help researchers understand the variation in the gut microbiome across health and disease populations. The *expanded Human Oral Microbiome Database* (eHOMD) (20) provides comprehensive curated information on the bacteria found in the human mouth and aerodigestive tract, including the pharynx, nasal passages, sinuses, and esophagus, providing new insights into the nostril microbiome. *gutMDisorder* (21), a comprehensive manually literature-extracted resource for associations between gut microbes and phenotypes or interventions in *Homo sapiens* and *Mus musculus*, provides users references to help identify the functional connections between gut microorganisms and disease. The *data repository for Gut Microbiota* (GMrepo) (22,23) deposits curated data resources from consistently annotated human gut metagenomes, increasing the reusability and accessibility of human gut metagenomic data, and enabling cross-project and phenotypic comparisons. The *Human Microbiome Bodymap* (mBodyMap) (24) curates the collected microbial data identified via their associations with human diseases and body sites to enable cross-dataset integration and comparison. *Microbe-phage interaction database* (MVP) (25) provides a comprehensive catalog of phage-microbe interactions to assist users to find phages that target specific microbes of interest. However, most of these widely known microbial databases are gut-derived or oral-derived, few are tissue-derived. Moreover, there is no database providing a relatively comprehensive analysis and user-friendly interactive web resource for disease-related microbiome. Given this, we propose the development of the *microbioTA*, an atlas of the microbiome from cancer tissues of *H. sapiens* and *M. musculus*. This system is designed to be free to access http://bio-annotation.cn/microbiota and we hope that it will facilitate new discoveries in this developing field.

DATA COLLECTION AND DATABASE CONTENT

Collection of RNA-sequencing data from human and mouse tissue samples

An overview of *microbioTA* database is presented in Figure 1. We collected 417 publicly available datasets from the GEO database (26), comprising 302 datasets across 25 human tissues and 115 datasets across 14 mouse tissues. We then downloaded the Sequence Read Archive (SRA) (27) data and collected the corresponding metadata of the datasets allowing us to annotate each including source name, tissue, age, and gender of the host organism. We performed filtration on the metadata in advance because there was some loss of information and all the datasets were collected before March 2022.

Processing of raw sequencing data

We firstly decompressed the SRA data to fastq data using SRA-Tools (https://github.com/ncbi/sra-tools), then performed quality control and pre-processed the data using fastp software (28). Samples with lower quality were removed according to the following criteria: (i) samples without complete metadata; (ii) samples with very low sequence counts.

Alignment and abundance calculations and result filtration

Kraken2 (29,30) was used to complete the taxonomic classification of the sequencing data, and the alignment results were accurate to the species level. We used Bracken (31) to compute the abundance of microbes at the species level using the reads collected in Kraken2 and then used KrakenTools (https://github.com/jenniferlu717/KrakenTools/) to convert the microbiome abundance results from the kraken2 report into the MetaPhlAn (32) report format. We then combined all samples from the same dataset into one matrix and then used these to develop a set of filter criterion designed to guarantee the quality of the alignment result and minimize the noise of any contaminating sequences. Finally, any taxa whose abundances were <4 in over 80 samples from any individual dataset were removed.

Microorganism characterization

R package, vegan (v2.6–2) (https://github.com/vegandevs/vegan) was used to calculate microbial alpha diversity, including the observed, chao1 (33), Shannon (34) and Simpson (35) diversity indices. While python package, scikit-bio (v0.5.7) (https://github.com/biocore/scikit-bio) was used to calculate the various microbial beta diversity distance values, such as Euclidean, Jaccard, Bray Curtis and Jensen Shannon distance, before using these data to create the beta diversity distance matrix using principal coordinate analysis (PCoA) or non-metric multidimensional scaling (NMDS). Linear discriminant analysis of effect Size (LEfSe) (36) was used to identify the characteristic microorganisms from cancer and control samples, which in turn might help deepen our understanding of the relationship between certain microbes and cancers.

Database construction

microbioTA is freely accessible to the user community at http://bio-annotation.cn/microbiota and requires no registration or login. *microbioTA* was constructed using Vue (v3.3.37) and tested in Mozilla Firefox, Google Chrome, and Microsoft Edge browsers and part of the data was stored and queried using MySQL (v5.7.24), while the other data were stored locally in Microsoft Excel format. The interactive access form is implemented using the Python flask web framework and is visible via Apache ECharts and inhouse R scripts.

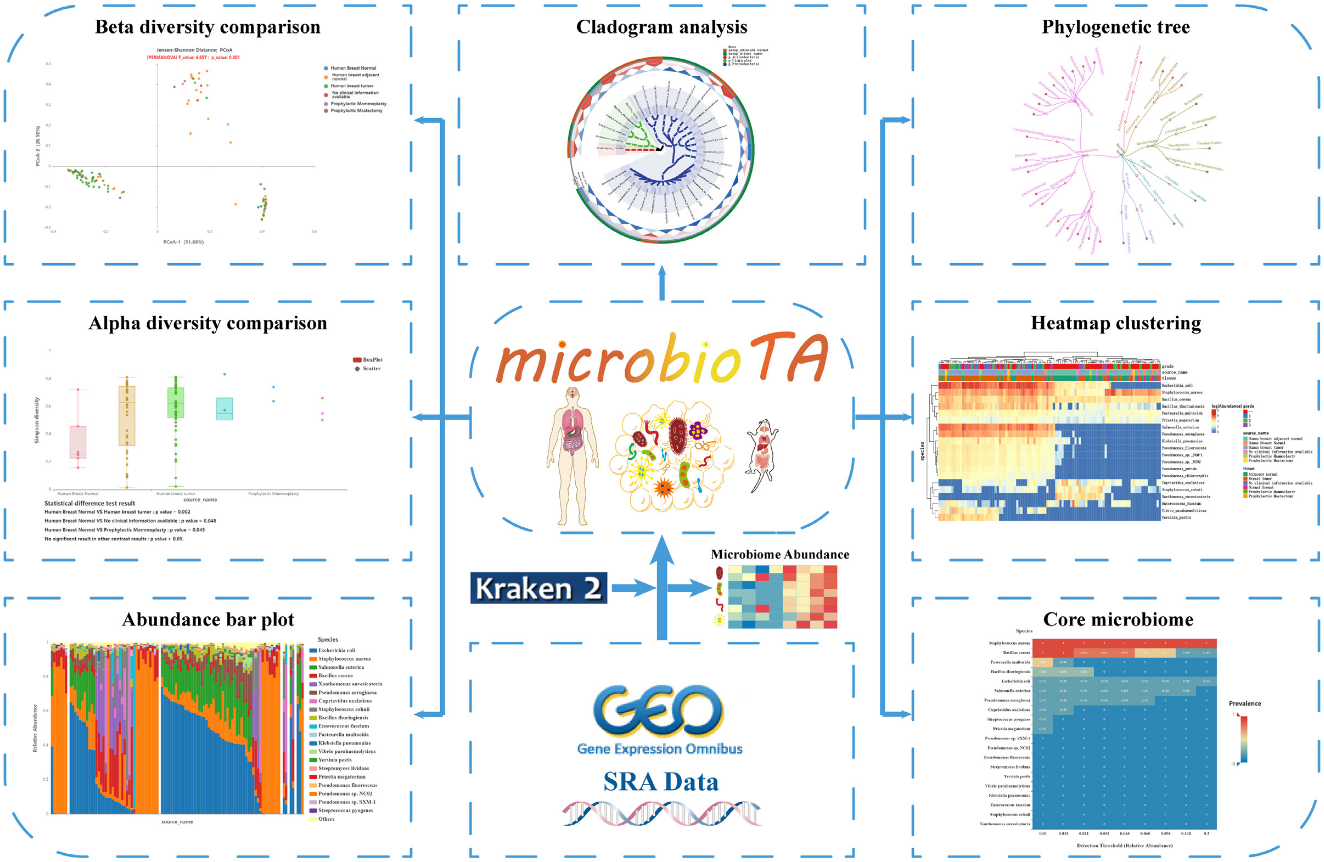


Figure 1. Overview of *microbioTA*. Raw RNA sequencing data were collected from the GEO database and then used to create a novel microbiome alignment pipeline to help re-align this RNA-Seq data with microbiome abundance information. We then used this to analyze the microbial characteristics of various microbiomes from different perspectives including the microbiome diversity and compositional difference analysis. The *microbioTA* interface was then constructed the *microbioTA* to visualize and store the analysis result of the microbiome resource.

DATABASE STATISTICS

As of July 2022, *microbioTA* includes the data from the 16 534 samples forming part of the 302 projects of *H. sapiens* projects identified in this study and 3973 samples from the 115 *M. musculus* projects (Figure 2A). These samples were taken from 69 cancer types across 25 *H. sapiens* tissues and 35 cancer types across 14 *M. musculus* tissues. Detailed statistics on the distribution of projects and samples among the different tissues are shown in Figure 2B. Finally, our evaluations identified a total of 6499 species across all projects, 6,468 from the *H. sapiens* samples and 6213 from *M. musculus* samples, respectively. Statistical analysis of the species distribution across different tissues is shown in Figure 2C.

USER INTERFACE

We created a user-friendly web platform for visualizing the microbiome features of various publicly available datasets. Users can browse these datasets by clicking the icons for different tissues and the hyperlinks for specific diseases under the tissue on the left portion of the 'Home' page or by clicking the hyperlinks of different tissues listed on the right portion of the 'Home' page (Figure 3A). These actions will then redirect the browser to the 'Organism' section of the 'Browse' page (Figure 3B) and once users select a dataset from this page, detailed information will be provided in table or picture format, including:

- **Detailed description**. The basic description of each of the datasets (Figure 4A) include the accession number (ID) and bio-project number (ID) from the GEO database, publication information, organism, tissue, disease, sample number, and species number of the selected dataset. We also provide hyperlinks to external public data resources and publication information.
- **Metadata.** The statistical information from the metadata category is shown using a pie plot (Figure 4B) and all metadata were manually filtered to remove unnecessary or ambiguous information. Users can select one metadata group to understand the distribution of the samples.
- **Relative abundance data.** Stack bar plots for microbial abundance are shown at different taxonomic levels as grouped by the metadata (Figure 4C) and users can choose the taxonomic level from phylum to species and select any one of the metadata categories to group the samples as well as the number of the top taxa they want to display. Each bar represents a sample, and different colors represent different taxa.

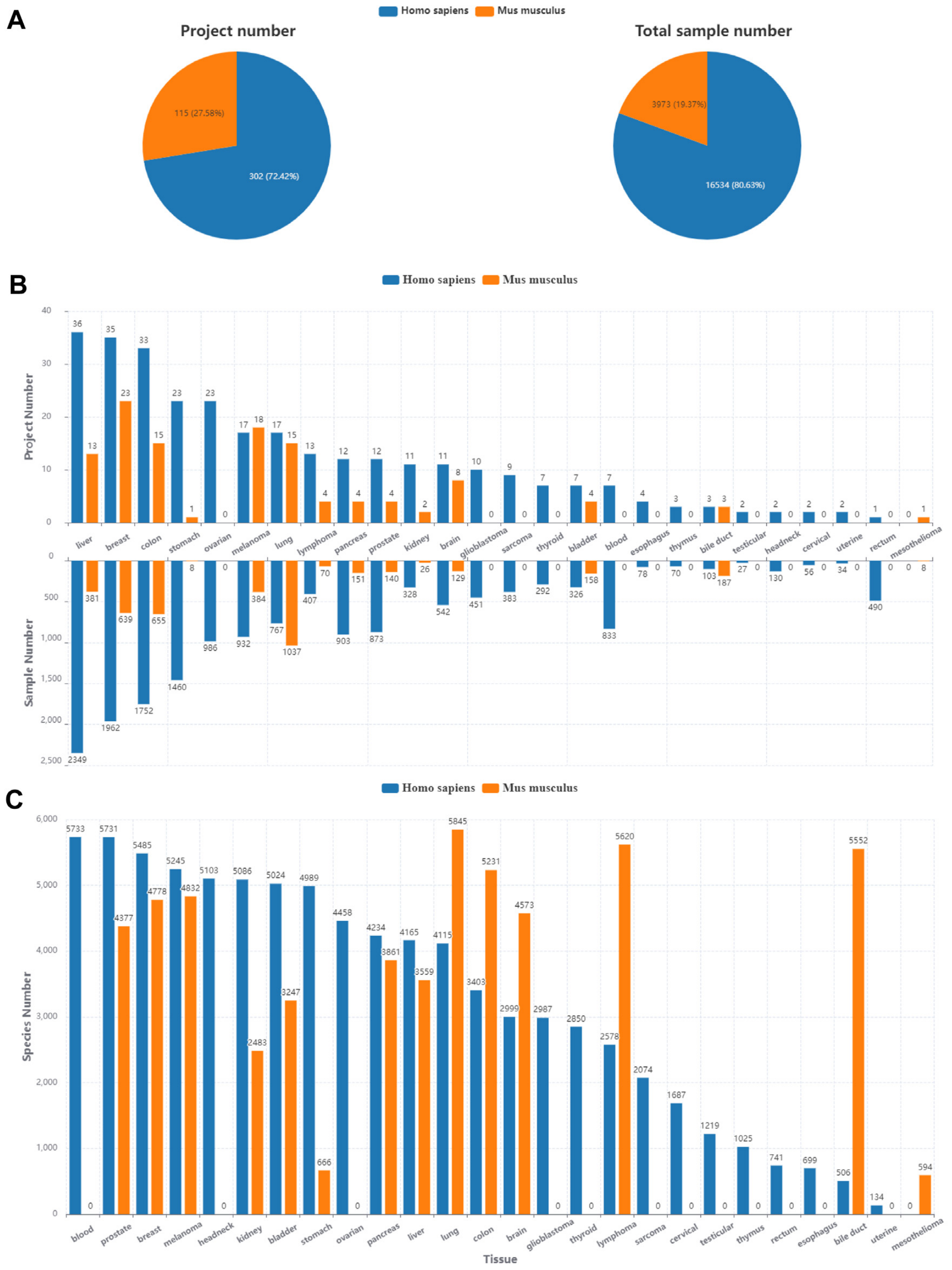
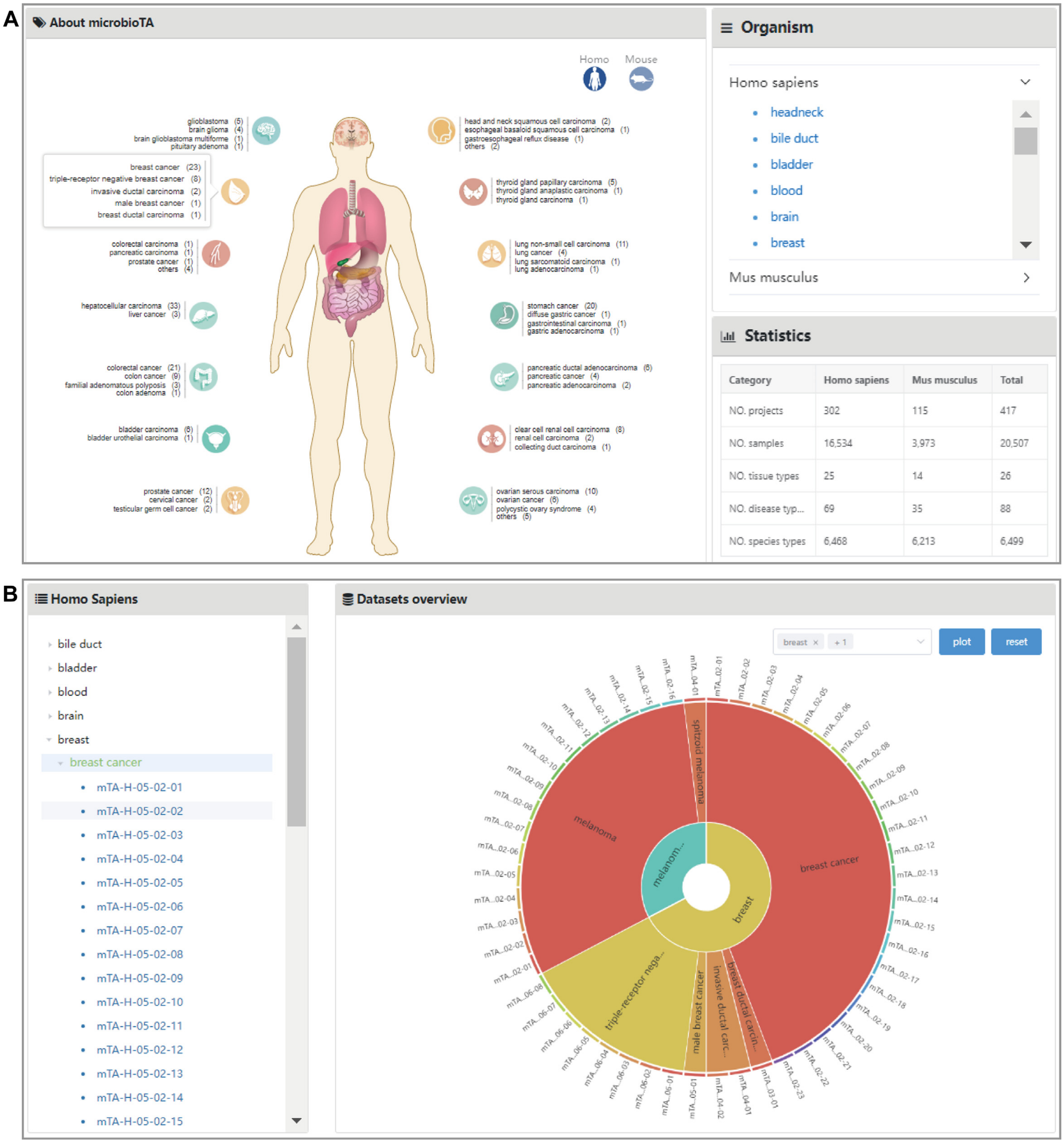


Figure 2. Statistics describing the data in *microbioTA*. (**A**) Total numbers of projects and samples collected from *Homo sapiens* and *Mus musculus*. (**B**) Numbers of projects (above) and samples (below) from different *Homo sapiens* and *Mus musculus* tissues. (**C**) Numbers of species detected in different *Homo sapiens* and *Mus musculus* tissues.

- **Phylogenetic tree. A** phylogenetic tree for each of the samples within the selected dataset is shown in Figure 4D. Users can choose one sample to plot the corresponding phylogenetic tree of the microbiome colonized in this sample. Branches of each phylum in the tree are highlighted in different colors.
- **Alpha diversity comparisons.** The microbial alpha diversity index at different taxonomic levels was compared among metadata groups and is presented as a boxplot (Figure 4E) with our platform providing four kinds of alpha diversity indices: observed, chao1, Shannon, and Simpson indices. The taxonomy level, metadata grouping category, and diversity indices are all optional, and users can choose the appropriate option according to their needs to compare the microbial alpha diversity among different groups of the dataset. The samples are grouped by metadata, and the scatters are colored by group. Statistical evaluations are also provided under the picture with a *P*-value less than 0.05 was considered significant.
- **Beta diversity comparisons.** Microbial beta diversity at different taxonomic levels were compared between metadata groups and are reported using the scatter plot



glioblastoma (5)
brain glioma (4)
brain glioblastoma multiforme (1)
pituitary adenoma (1)

breast cancer (23)
triple-receptor negative breast cancer (8)
invasive ductal carcinoma (2)
male breast cancer (1)
breast ductal carcinoma (1)

colorectal carcinoma (1)
pancreatic carcinoma (1)
prostate cancer (1)
others (4)

hepatocellular carcinoma (33)
liver cancer (3)

colorectal cancer (21)
colon cancer (9)
familial adenomatous polyposis (3)
colon adenoma (1)

bladder carcinoma (6)
bladder urothelial carcinoma (1)

prostate cancer (12)
cervical cancer (2)
testicular germ cell cancer (2)

head and neck squamous cell carcinoma (2)
esophageal basaloid squamous cell carcinoma (1)
gastroesophageal reflux disease (1)
others (2)

thyroid gland papillary carcinoma (5)
thyroid gland anaplastic carcinoma (1)
thyroid gland carcinoma (1)

lung non-small cell carcinoma (11)
lung cancer (4)
lung sarcomatoid carcinoma (1)
lung adenocarcinoma (1)

stomach cancer (20)
diffuse gastric cancer (1)
gastrointestinal carcinoma (1)
gastric adenocarcinoma (1)

pancreatic ductal adenocarcinoma (6)
pancreatic cancer (4)
pancreatic adenocarcinoma (2)

clear cell renal cell carcinoma (8)
renal cell carcinoma (2)
collecting duct carcinoma (1)

ovarian serous carcinoma (10)
ovarian cancer (6)
polycystic ovary syndrome (4)
others (5)

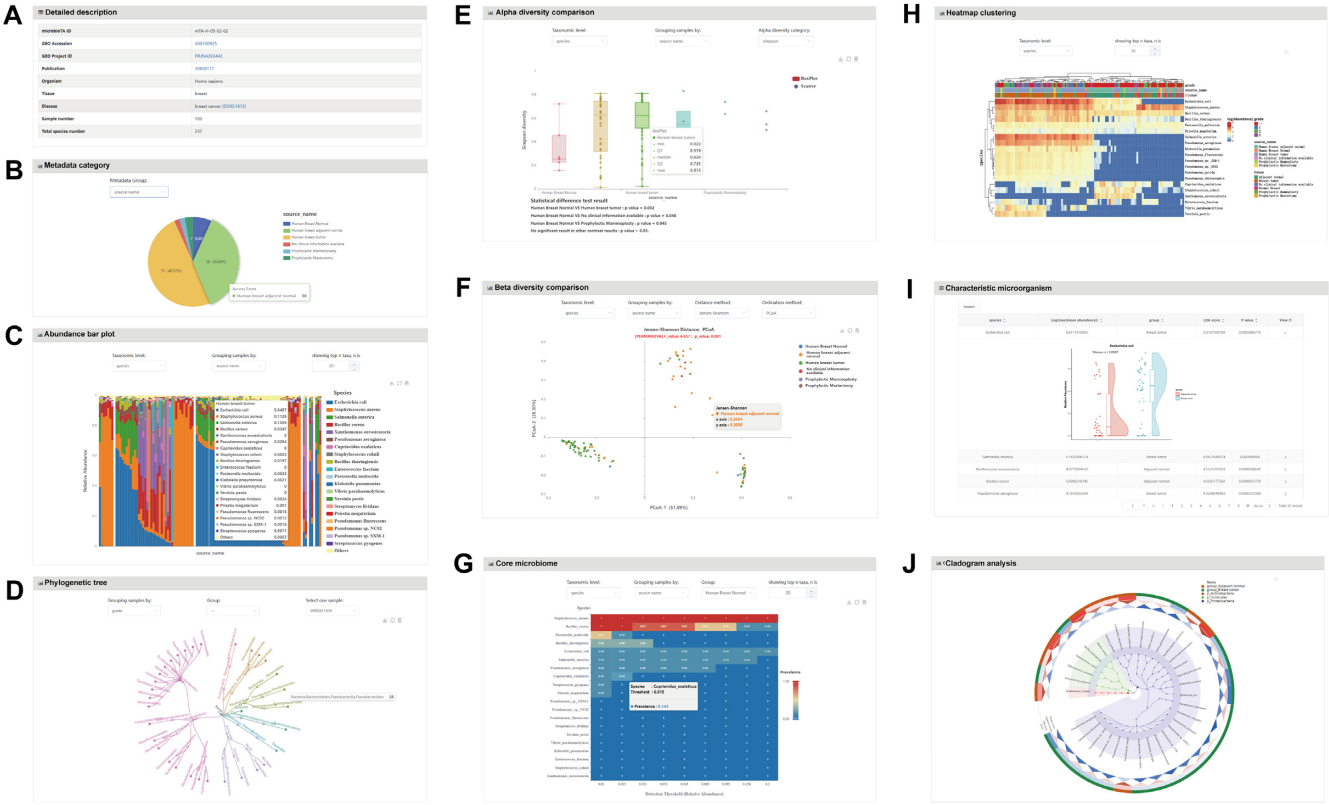
Category	Homo sapiens	Mus musculus	Total
NO. projects	302	115	417
NO. samples	16,534	3,973	20,507
NO. tissue types	25	14	26
NO. disease typ...	69	35	88
NO. species types	6,468	6,213	6,499

mTA-H-05-02-01
mTA-H-05-02-02
mTA-H-05-02-03
mTA-H-05-02-04
mTA-H-05-02-05
mTA-H-05-02-06
mTA-H-05-02-07
mTA-H-05-02-08
mTA-H-05-02-09
mTA-H-05-02-10
mTA-H-05-02-11
mTA-H-05-02-12
mTA-H-05-02-13
mTA-H-05-02-14
mTA-H-05-02-15

Figure 3. Home and browse pages from *microbioTA*. (**A**) Home page of *microbioTA*. (**B**) Browse page of *microbioTA*.

(Figure 4F). Here, we report four types of distance evaluations, including Bray-Curtis, Euclidean, Jaccard, and Jensen-Shannon, for comparison. In addition, we provide two ordination methods, principal coordinate analysis (PCoA) and non-metric multidimensional scaling (NMDS), to visualize the beta diversity differences and these values are all evaluated using a PERMANOVA function.

- **Core microbiome.** The prevalence of the microbes of interest in each of the selected metadata groups can be plotted using a clustering heatmap (Figure 4G), users can select the taxonomic level and the number of the top microbes with the highest mean abundance they want to display. The colors of the cells in the heatmap represent the prevalence of the microbe within each group.
- **Heatmap clustering.** The abundance of each of the microbes of interest in each of the samples with metadata at different taxonomic levels can be visualized using a heatmap (Figure 4H) and users can select both the taxonomic level and set the number of microbes with the



microbioTA ID	mTA-H-05-02-02
GEO Accession	GSE100925
GEO Project ID	PRJNA393443
Publication	30930117
Organism	Homo sapiens
Tissue	breast
Disease	breast cancer (DOID:1612)
Sample number	103
Total species number	237

Statistical difference test result

Human Breast Normal VS Human breast tumor : p value = 0.002

Human Breast Normal VS No clinical information available : p value = 0.048

Human Breast Normal VS Prophylactic Mammoplasty : p value = 0.045

No significant result in other contrast results : p value > 0.05.

Figure 4. Detailed information of *microbioTA*. (**A**) Detailed description of the selected dataset. (**B**) Pie plot describing the relevant Metadata categories. Each part represents a group of samples. (**C**) Stack bar plot of microbiome abundance (Relative abundance). (**D**) Phylogenetic tree for queried samples. Different colors represent differences in the cladogenesis of various phylum. (**E**) Box plot of the microbial alpha diversity index, each box represents one metadata group, and each node represents a sample. The result of the statistical difference analysis is provided in the left bottom corner of the image, with only *P* values <0.05 reported in this format. (**F**) Scatter plots of the microbial beta diversity. Samples are colored according to their group and statistical differences between different groups are recorded in the subtitle. (**G**) Heatmap clustering plot of microbiome prevalence showing the core microbiome of interest group in any dataset, the colors of the cells represent the prevalence of any microbiome under the threshold. (**H**) Heatmap clustering of microbiome abundance for each taxon of interest, the colors of the cells represent the normalized abundance of the microbiome. (**I**) The characteristic microbiome identified via LEfSe analysis comparing cancer and control samples. Clicking the '>' bottom under the 'View' column brings up a detailed comparison of the species distribution in the cancer and control groups with any statistically significant differences highlighted within the picture. (**J**) Cladogram plot describing the microbiome biomarkers identified in the LEfSe results. Moving from the outside to the inside, the circles represent the group to which each of the biomarkers belong, the LDA score, and the mean abundances of the biomarkers in any two groups, respectively. The phylogenetic tree in the insider of the circles is colored based on phylum.

highest mean abundance within the dataset they want to evaluate. Samples are then clustered using metadata on the x-axis and taxa on the y-axis.

- **Characteristic microorganisms.** LEfSe analysis can be performed allowing users to compare the disease and control groups (Figure 4I). This is then output as a table which contains the statistically significant microbiome biomarkers of the specific group in the dataset. The 'details' section then provides the boxplots for each of the characteristic microorganisms in each of the different groups.
- **Cladogram analysis.** This platform also provides a cladogram plot function allowing users to visualize the microbial biomarkers from their LEfSe evaluations (Figure 4J). The insider branches of these trees are colored according to phylum and as the rings moved from the outside to the inside, they represent the group of biomarkers belonging to the LDA score and the mean abundance of the biomarkers in the two groups, respectively.

In addition to the 'Organism' section on the 'Browse' page, *microbioTA* also provides detailed information on the species recorded in our database under the 'Taxonomy' section (http://bio-annotation.cn/microbiota/browse/taxonomy). Users can obtain taxon classification information from the phylum to species level, mean prevalence, and mean abundance among projects in both *H. sapiens* and *M. musculus* tissues. We also provide hyperlinks to descriptions of the taxa on the NCBI taxonomy web page.

We provide a 'Search' page (http://bio-annotation.cn/microbiota/search) for users to obtain the information of interest as quickly as possible via the summation of a simple search request via the 'Search projects by tissues or diseases' section. This function provides the summary information of the tissue and disease they want to learn about in table format and various hyperlinks also allow users to reach to the detailed information page for each of the projects. Users who follow the 'Search species in all projects' section can obtain a summary table for their taxon of interest at any taxonomy level with hyperlinks to the detailed information for each the species.

Moreover, all the images and tables can be downloaded by clicking on the download icon, and all the microbial abundance matrix data is also available via the 'Download' page.

SUMMARY AND FUTURE PERSPECTIVES

Associations between diseased tissue microbiomes and disease progression are coming, increasingly, under evaluation, creating a new niche for microbial-related research. Multiple studies have found that specific microbial metabolites participate in the pathogenetic process, with many microbes acting as potential biomarkers for cancer diagnosis. Most of the existing databases collect information on gut and oral microbiomes, but there remains a lack of public databases focused on curating data around cancer tissue microbiomes. Therefore, we developed this novel analysis pipeline and constructed a web database platform to store and help analyze the tissue-derived microbiome data. The current version of *microbioTA* includes 417 publicly available datasets describing the microbiome in 25 human and 14 mouse tissues. This database includes 6468 and 6213 unique microbial species from human and mouse samples, respectively and all datasets contain the relevant detailed information to compile a useful set of relative analysis results.

microbioTA is a user-friendly, interactive database that will be of particular interest and use to the general researchers and the broader life science community, providing a reference for cancer microbiome research. The future updated versions will include additional extensions focused on addressing some of the tool's current limitations. First, we will include the microbiome data from various chronic diseases and continue to collect the latest public datasets. Second, we will create useful online tools for users to analyze this extended data and third, we will continue to optimize our database to provide users with a better experience.

DATA AVAILABILITY

This database can be freely accessed via http://bio-annotation.cn/microbiota. The code is available at https://github.com/liangcheng-hrbmu/microbioTA.

FUNDING

Tou-Yan Innovation Team Program of the Heilongjiang Province [2019-15], National Natural Science Foundation of China [62222104, 61871160, 62172130]; Heilongjiang Postdoctoral Fund [LBH-Q20030]. Funding for open access charge: Tou-Yan Innovation Team Program of the Heilongjiang Province [2019-15]; National Natural Science Foundation of China [62222104, 61871160, 62172130]; Heilongjiang Postdoctoral Fund [LBH-Q20030].

Conflict of interest statement. None declared.

REFERENCES

1. Layeghifard,M., Hwang,D.M. and Guttman,D.S. (2017) Disentangling interactions in the microbiome: a network perspective. *Trends Microbiol.*, **25**, 217–228.
2. Klassen,J.L. (2018) Defining microbiome function. *Nat. Microbiol.*, **3**, 864–869.
3. Dominguez-Bello,M.G., Godoy-Vitorino,F., Knight,R. and Blaser,M.J. (2019) Role of the microbiome in human development. *Gut*, **68**, 1108–1114.
4. Young,V.B. (2017) The role of the microbiome in human health and disease: an introduction for clinicians. *BMJ*, **356**, j831.
5. Franzosa,E.A., Sirota-Madi,A., Avila-Pacheco,J., Fornelos,N., Haiser,H.J., Reinker,S., Vatanen,T., Hall,A.B., Mallick,H., McIver,L.J. *et al.* (2019) Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.*, **4**, 293–305.
6. 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.
7. Shan,Y., Lee,M. and Chang,E.B. (2022) The gut microbiome and inflammatory bowel diseases. *Annu. Rev. Med.*, **73**, 455–468.
8. Tang,W.H., Kitai,T. and Hazen,S.L. (2017) Gut microbiota in cardiovascular health and disease. *Circ. Res.*, **120**, 1183–1196.
9. Mozaffarian,D. (2019) The microbiome, plasma metabolites, dietary habits, and cardiovascular risk unravelling their interplay. *Circ. Res.*, **124**, 1695–1696.
10. Bjorkegren,J.L.M. and Lusis,A.J. (2022) Atherosclerosis: recent developments. *Cell*, **185**, 1630–1645.
11. Schwabe,R.F. and Jobin,C. (2013) The microbiome and cancer. *Nat. Rev. Cancer*, **13**, 800–812.
12. Sepich-Poore,G.D., Zitvogel,L., Straussman,R., Hasty,J., Wargo,J.A. and Knight,R. (2021) The microbiome and human cancer. *Science*, **371**, eabc4552.
13. Helmink,B.A., Khan,M.A.W., Hermann,A., Gopalakrishnan,V. and Wargo,J.A. (2019) The microbiome, cancer, and cancer therapy. *Nat. Med.*, **25**, 377–388.
14. Manor,O., Dai,C.L., Kornilov,S.A., Smith,B., Price,N.D., Lovejoy,J.C., Gibbons,S.M. and Magis,A.T. (2020) Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.*, **11**, 5206.
15. Britton,G.J. and Faith,J.J. (2021) Causative microbes in host-microbiome interactions. *Annu. Rev. Microbiol.*, **75**, 223–242.
16. Poore,G.D., Kopylova,E., Zhu,Q., Carpenter,C., Fraraccio,S., Wandro,S., Kosciolek,T., Janssen,S., Metcalf,J., Song,S.J. *et al.* (2020) Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*, **579**, 567–574.
17. Riquelme,E., Zhang,Y., Zhang,L., Montiel,M., Zoltan,M., Dong,W., Quesada,P., Sahin,I., Chandra,V., San Lucas,A. *et al.* (2019) Tumor microbiome diversity and composition influence pancreatic cancer outcomes. *Cell*, **178**, 795–806.
18. Integrative, H.M.P.R.N.C. (2019) The integrative human microbiome project. *Nature*, **569**, 641–648.
19. Li,J., Jia,H., Cai,X., Zhong,H., Feng,Q., Sunagawa,S., Arumugam,M., Kultima,J.R., Prifti,E., Nielsen,T. *et al.* (2014) An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.*, **32**, 834–841.
20. Escapa,I.F., Chen,T., Huang,Y., Gajare,P., Dewhirst,F.E. and Lemon,K.P. (2018) New insights into human nostril microbiome from the expanded human oral microbiome database (eHOMD): a resource for the microbiome of the human aerodigestive tract. *Msystems*, **3**,e00187-18.
21. 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.
22. Dai,D., Zhu,J., Sun,C., Li,M., Liu,J., Wu,S., Ning,K., He,L.J., Zhao,X.M. and Chen,W.H. (2022) GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Res.*, **50**, D777–D784.
23. 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.
24. Jin,H., Hu,G., Sun,C., Duan,Y., Zhang,Z., Liu,Z., Zhao,X.M. and Chen,W.H. (2022) mBodyMap: a curated database for microbes across human body and their associations with health and diseases. *Nucleic Acids Res.*, **50**, D808–D816.

25. 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.
26. Clough,E. and Barrett,T. (2016) The gene expression omnibus database. *Methods Mol. Biol.*, **1418**, 93–110.
27. International Nucleotide Sequence Database Collaboration, Leinonen,R., Sugawara,H. and Shumway,M. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
28. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
29. Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
30. Wood,D.E., Lu,J. and Langmead,B. (2019) Improved metagenomic analysis with kraken 2. *Genome Biol.*, **20**, 257.
31. Lu,J., Breitwieser,F.P., Thielen,P. and Salzberg,S.L. (2017) Bracken: estimating species abundance in metagenomics data. *Peer J Computer Science*, **3**, e104.
32. Beghini,F., McIver,L.J., Blanco-Miguez,A., Dubois,L., Asnicar,F., Maharjan,S., Mailyan,A., Manghi,P., Scholz,M., Thomas,A.M. *et al.* (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife*, **10**, e65088.
33. Chao,A. (1984) Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.*, **11**, 265–270.
34. Shannon,C.E. (1997) The mathematical theory of communication. 1963. *MD Comput.*, **14**, 306–317.
35. Simpson,E.H. (1949) Measurement of diversity. *Nature*, **163**, 688–688.
36. 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.