Journal Pre-proofs

Database

MicroPhenoDB Associates Metagenomic Data with Pathogenic Microbes, Microbial Core Genes, and Human Disease Phenotypes

Guocai Yao, Wenliang Zhang, Minglei Yang, Huan Yang, Jianbo Wang, Haiyue Zhang, Lai Wei, Zhi Xie, Weizhong Li

PII: S1672-0229(20)30169-8
DOI: https://doi.org/10.1016/j.gpb.2020.11.001
Reference: GPB 455

To appear in: *Genomics, Proteomics & Bioinformatics*

Received Date: 16 December 2019
Revised Date: 30 July 2020
Accepted Date: 12 November 2020

Please cite this article as: G. Yao, W. Zhang, M. Yang, H. Yang, J. Wang, H. Zhang, L. Wei, Z. Xie, W. Li, MicroPhenoDB Associates Metagenomic Data with Pathogenic Microbes, Microbial Core Genes, and Human Disease Phenotypes, *Genomics, Proteomics & Bioinformatics* (2020), doi: https://doi.org/10.1016/j.gpb.2020.11.001

MicroPhenoDB Associates Metagenomic Data with Pathogenic Microbes, Microbial Core Genes, and Human Disease Phenotypes

Guocai Yao[1,#], Wenliang Zhang[1,#], Minglei Yang[1], Huan Yang[1], Jianbo Wang[1], Haiyue Zhang[1], Lai Wei[3], Zhi Xie[2,3], Weizhong Li[1,2,4,*]

[1] Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China

[2] Center for Precision Medicine, Sun Yat-sen University, Guangzhou 510080 China

[3] State Key Lab of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 500060, China

[4] Key Laboratory of Tropical Disease Control of Ministry of Education, Sun Yat-Sen University, Guangzhou 510080, China

[#] Equal contribution

* Corresponding author

E-mail: liweizhong@mail.sysu.edu.cn (Li W).

Running title: *Yao G et al / MicroPhenoDB Associates Metagenomics with Disease*

Total word counts: 3680

Reference: 56

Total figures: 7

Total tables: 4

Total supplementary figure: 1

Total supplementary tables: 2

Abstract

Microbes play important roles in human health and disease. The interaction between microbes and hosts is a reciprocal relationship, which remains largely under-explored. Current computational resources lack manually and consistently curated data to connect metagenomic data to pathogenic microbes, microbial core genes, and disease phenotypes. We developed the MicroPhenoDB database by manually curating and consistently integrating microbe-disease association data. MicroPhenoDB provides 5677 non-redundant associations between 1781 microbes and 542 human disease phenotypes across more than 22 human body sites. MicroPhenoDB also provides 696,934 relationships between 27,277 unique clade-specific core genes and 685 microbes. Disease phenotypes are classified and described using the Experimental Factor Ontology (EFO). A refined score model was developed to prioritize the associations based on evidential metrics. The sequence search option in MicroPhenoDB enables rapid identification of existing pathogenic microbes in samples without running the usual metagenomic data processing and assembly. MicroPhenoDB offers data browsing, searching, and visualization through user-friendly web interfaces and web service application programming interfaces. MicroPhenoDB is the first database platform to detail the relationships between pathogenic microbes, core genes, and disease phenotypes. It will accelerate metagenomic data analysis and assist studies in decoding microbes related to human diseases. MicroPhenoDB is available through http://www.liwzlab.cn/microphenodb and http://lilab2.sysu.edu.cn/microphenodb.

KEYWORDS: Pathogenic microbes; Metagenomic data; Disease phenotypes; Microbe-disease association; COVID-19

Introduction

The human body feeds a large number of microbes, mainly composed of bacteria, followed by archaea, fungi, viruses, and protozoa. Microbes, inhabiting various organs of the human body, mainly in the gastrointestinal tract, as well as in the respiratory tract, oral cavity, stomach, and skin, play important roles in human health and disease [1−3]. Microbial gene products have rich biochemical and metabolic activities in the host [4−6]. Microorganisms usually form a healthy symbiotic relationship with the host. However, when the microbial content becomes abnormal or exogenous microbes infect the host, the balance of host microecology can be broken, which in turn can possibly cause various diseases [7,8]. Tripartite network analysis in patients with irritable bowel syndrome demonstrated that the gut microbe *Clostridia* is significantly associated with brain functional connectivity and gastrointestinal sensorimotor function [9]. Strati et al. reported that Rett syndrome is substantially associated with a dysbiosis of both bacterial and fungal components of the gut microbiota [10]. The alteration of microbial communities on psoriatic skin is different from those on healthy skin and has a potential role in Th17 polarization to exacerbate cutaneous inflammation [11]. The ongoing pandemic of coronavirus disease 2019 (COVID-19) has affected more than 220 countries, areas, or territories worldwide by November 2020. Lung injury has been reported in most patients with confirmed severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection [12].

The interaction between microbes and hosts is a reciprocal relationship and remains largely under-explored [13]. Accurate relationship information between microbes and diseases can greatly assist studies in human health [14]. With the wide application of next-generation sequencing (NGS) technology, microbiological analysis methods and standards are being rapidly developed, such as metagenomic approaches [15]. As a result, a large amount of experimental data has been published [16]. Thus accurate database platforms are greatly needed to utilize these experimental data, determine the composition of pathogenic microbes in hosts, clarify microbial-disease relationships, and provide standardized high-quality annotation for clinical uses [17].

Due to the functional and clinical significance of microbes, several public databases have been established to collect microbe-disease association data, such as the Human Microbe-Disease Association Database (HMDAD) [18], Disbiome [19],

the Virulence Factor Database (VFDB) [20], and the Comprehensive Antibiotic Resistance Database (CARD) [21]. HMDAD and Disbiome collate text-mining-based microbe–disease association data from peer-reviewed publications and describe the strength of the associations based on the credibility of the data sources. VFDB provides up-to-date knowledge of the virulence factors (VFs) of various bacterial pathogens; CARD contains high-quality reference data on the molecular basis of antimicrobial resistance with an emphasis on genes, proteins, and mutations involved. Data in VFDB and CARD help to explain the relationship between pathogenic microbial genes and the health status of hosts. In addition, to assist physicians and healthcare providers to quickly and accurately diagnose infectious diseases in patients, a guideline for utilization of the microbiology laboratory for diagnosis of infectious diseases was developed and is being regularly updated by the Infectious Diseases Society of America (IDSA) and the American Society for Microbiology (ASM) [22]. The curation and analysis of microbe-disease association data are essential for expediting translational research and application. However, these computational resources lack manually and consistently curated data to connect metagenomic data to pathogenic microbes, microbial core genes, and disease phenotypes.

To bridge this gap, we developed the MicroPhenoDB database (http://www.liwzlab.cn/microphenodb) by manually curating and consistently integrating microbe-disease association data. We collected and curated the microbe-disease associations from the IDSA guideline [22], the National Cancer Institute (NCI) Thesaurus OBO Edition (NCIT) [23], and the HMDAD [18] and Disbiome [19] databases, and also connected microbial core genes derived from the MetaPhlAn2 dataset [24] to pathogenic microbes and human diseases. A refined score model was adopted to prioritize the microbe-disease associations based on evidential metrics [18,25]. In addition, a sequence search web application was also implemented to allow users to query sequencing data to identify pathogenic microbes in metagenomic samples, as well as to retrieve the disease-related information of virulence factors and antibiotic resistances. MicroPhenoDB allows users to browse, search, access, and analyze data through user-friendly web interfaces, visualizations, and web service application programming interfaces (APIs).

Data collection and processing

Data collection and manual annotation

To ensure data quality, we integrated the association data with annotations from HMDAD and Disbiome and manually collated and curated microbe-disease association data from the IDSA guideline and NCIT (**Figure 1**). The IDSA guideline provides criteria for clinical identification of infectious microbes, while NCIT is a reference terminology that provides comprehensive information for infectious microbes. To enrich the annotation for disease-microbe associations, we manually traced the relevant literature in HMDAD and Disbiome; we also provided the microbes with annotation at the resolution of species levels, such as taxonomies and official names. Association data between infectious microbes and diseases in IDSA were extracted. Relevant information about disease phenotypes and microbes in the microorganism notes from NCIT were extracted as well. The collected and integrated association data include information about microbe symbols, disease symbols, the increased or decreased impacts of the microbes, PubMed identifiers, and validation methods.

Controlled vocabulary and ontology to describe microbes and diseases

In MicroPhenoDB, several standard terminology and controlled vocabulary resources were adopted to consistently annotate microbes and diseases (Figure 1). Different tools and reference databases might give different taxonomies for microbes. To avoid this discrepancy, the official names of microbes were taken from NCIT [23], and the taxonomy identifiers were adopted from the National Center for Biotechnology Information (NCBI) [26] and UniProt [27]. The relationships between core genes and microbes were annotated using the MetaPhlAn2 tool [28], the microbial gene functions were annotated using the InterProScan tool [29], and the virulence factors and the drug resistance information of microbes were retrieved respectively from the databases of VFDB [20] and CARD [21]. The disease phenotypes were annotated with official names, experimental factor terms, definitions, classifications, and cross-references using the Experimental Factor Ontology (EFO) [30]. EFO provides a systematic description of many experimental variables across the European Bioinformatics Institute (EMBL-EBI) databases and the National Human Genome Research Institute (NHGRI) genome-wide association study (GWAS) catalog [31]; it

also combines parts of several popular ontologies, such as Orphanet Rare Disease Ontology [32], Human Phenotype Ontology [33], and Monarch Disease Ontology [34]. The versions or releases of databases and tools used in the MicroPhenoDB construction are detailed in Table S1.

Association score model

One of the main problems in exploiting extensive collections of aggregated microbiome data is how to prioritize the associations. According to the previous studies by Ma et al. [18] and Pinero et al. [25], we refined the association score model to prioritize the microbe-disease associations using additional evidential metrics, including the number of sources that report the association, the type of curation of each source, and the number of supporting publications in the manual curation.

For every disease *i* and every microbe *j*, the raw score of their relationship Raw_score_{ij} was defined as:

$$\mathrm{Raw\_score}_{ij} = (W_{IDSA} + W_{NCIT} + W_{Literature}) \times \log(N/n_j) \quad (1)$$

In Equation (1), W_{IDSA} is the weight of the association source from the IDSA guideline, W_{NCIT} is the weight of the association source from NCIT, and $W_{Literature}$ is the weight of the association source from literature publications. N is the number of all diseases in MicroPhenoDB, and n_j is the number of diseases associated with microbe *j*. $Log(N/n_j)$ is computed to increase Raw_score_{ij} for the microbes that areassociated explicitly with few diseases or decrease Raw_score_{ij} for the microbes globally associated with several diverse diseases.

In Equations (2)−(4), MicroPhenoDB assigns different weights to different evidential sources according to their reliabilities (**Table 1**) [25]. If the association is curated from literature publications, $W_{Literature}$ is initially assigned as 0.25, otherwise assigned as 0. If the association is curated from NCIT [23], W_{NCIT} is initially assigned as 0.5, which is double that of $W_{Literature}$, otherwise assigned as 0. If the association is curated from IDSA [22], W_{IDSA} is initially assigned as 1.0, which is double that of W_{NCIT}, otherwise assigned as 0. The three weights also depend on the direction of the abundance change of a microbe in a disease and the number of supporting publications. D_{ij} ($D_{ij} \in \{1, -1\}$) represents the direction of the abundance change of microbe *j* in disease *i*. If the microbe *j* is increased in the case of disease *i*, D_{ij} equals

1; if the microbe *j* is decreased in the case of disease *i*, D_{ij} equals −1. n_p is the number of publications in which an association between a disease and a microbe has been reported. From the distribution of numbers of evidence, we found n_p was less than 16 and mostly ranged from 1 to 2 (Figure S1).

$$W_{Literature} = \begin{cases} D_{ij} \times 0.25 \times n_p & \text{Association from literature} \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

$$W_{NCIT} = \begin{cases} D_{ij} \times 0.5 & \text{Association from NCIT} \\ 0 & \text{Otherwise} \end{cases} \quad (3)$$

$$W_{IDSA} = \begin{cases} D_{ij} \times 1.0 & \text{Association from IDSA} \\ 0 & \text{Otherwise} \end{cases} \quad (4)$$

$$Score_{ij} = \frac{2}{1 + \frac{1}{e^{Raw\_score_{ij}}}} - 1 \quad (5)$$

Finally, the sigmoid function was used to normalize Raw_score_{ij} to limit the range of the final association score $Score_{ij}$ from −1 to 1. In Equation (5), '*e*' represents the natural constant *e*. $Score_{ij}$ can be used to judge the confidence of the relationship between a microbe and a disease phenotype. Please see the score distribution in **Figure 2**. A $Score_{ij}$ more than 0 indicates that the occurrence of the disease correlates with an increase of the microbial abundance, and a $Score_{ij}$ less than 0 indicates that the occurrence of the disease correlates with a decrease of the microbial abundance. The greater the absolute value of $Score_{ij}$, the higher the number of previous reports of the respective microbe-disease association; the closer the score is to zero, the lower the number of previous reports of the respective microbe-disease association. By investigating the $Score_{ij}$ distribution, most associations were found with $Score_{ij}$ between −0.3 and 0.3, and the two peaks with $Score_{ij}$ more than 0.3 were involved in high confidence associations from NCIT and IDSA (Figure 2). This suggested that the score points of −0.3 and 0.3 would be the highly reliable thresholds to assess the confidence level of an association.

Implementation

The web applications in MicroPhenoDB were implemented in Java language by using the model-view-controller model and the SpringBoot framework and were deployed on an Apache Tomcat web server. The association data of microbes and disease phenotypes were stored in a MySQL database. Data access, search, and visualization

were implemented by using the Ajax API technology. The frontend interface was visualized by using the Vue.js framework. The sequence search tool was implemented using the EMBL-EBI tool framework [35].

Database content and usage

Database content

MicroPhenoDB collated 7449 redundant associations between 1781 microbes and 542 human disease phenotypes across more than 22 human body sites (**Table 2**). Of the 7449 associations, 29.7% were manually curated from the IDSA guideline (1196, 16.1%) [22], NCIT (849, 11.4%) [23], and peer-reviewed publications for human respiratory infection virus (164, 2.2%), and the others were consistently integrated with annotation from HMDAD (673, 9.0%) [18] and Disbiome (4567, 61.3%) [19] (**Figure 3**A). Multiple publications might support the same association between a microbe and a disease phenotype. After removing data redundancy based on the supporting publications, MicroPhenoDB produced 5677 non-redundant microbe-disease phenotype associations (Table 2). The number of non-redundant associations was over 11-fold (5677/483) of that in HMDAD. Each non-redundant association was assigned with a unique accession number (*e.g.*, MBP00000900) and an association score. For the microbe distribution, MicroPhenoDB contained 1497 bacteria in a broad sense (including 1474 bacteria in a narrow sense, 11 *Rickettsia*, 6 *Chlamydia*, 4 *Ehrlichia*, and 2 *Mycoplasma*), 183 viruses, 58 fungi, and 43 parasites (Table 2). Approximately 88.3% (5014/5677), 8.5% (481/5677), 2.0% (116/5677), and 1.2% (66/5677) of the associations were related to bacteria, viruses, fungi, and parasites respectively (Figure 3B). The top six frequent disease-associated bacteria phyla were Firmicutes, Proteobacteria, Bacteroidetes, Actinobacteria, Spirochaetes, and Fusobacteria. The top disease-associated fungal phylum was Ascomycota. Firmicutes included 271 genus/species in 4 classes (Bacilli, Clostridia, Erysipelotrichia, and Negativicutes) (Figure 3C). The microbes were mainly distributed in the body sites of the gastrointestinal tract (37.3%), oral cavity (9.5%), respiratory tract (6.9%), skin and soft tissue (4.2%), urinary tract (3.5%), vagina (2.5%), and central nervous system (2.0%) (**Table 3**). The disease phenotypes were classified and described by EFO [30]. Many diseases were associated with pathogenic microorganisms, such as bacterial, digestive, nervous, and autoimmune diseases (Figure 3D).

In total, 27,277 unique clade-specific core genes of 685 bacteria and viruses were retrieved from the dataset in MetaPhlAn2 and were annotated with gene functions using InterProScan (Table 2). In addition, 4204 virulence factor genes and 2522 drug resistance genes were also included from VFDB [20] and CARD [21], respectively. A small percentage ((4.3%, 65/1497) and (4.4%, 66/1497)) of bacteria was annotated with virulence factor information and antimicrobial resistance information, respectively (Table 2).

Web interface

The MicroPhenoDB website (http://www.liwzlab.cn/microphenodb) provides user-friendly web interfaces to enable users to search, browse, prioritize, and analyze the microbe-disease association data in the database (**Figure 4**). The website offers multiple optional search applications of microbes, diseases, and associations to acquire prioritized association data with body site and microbe type filters. The prioritized microbe-disease associations can be downloaded as a CSV file for further analysis. The hierarchical structure of microbes and diseases are respectively displayed on the 'Browse' web page. Information regarding the increasing or decreasing tendency of microbial abundance in a disease, virulence factor, and antibiotic resistance of the microbes, along with its core gene information, are available on the 'Browse' web page. In addition, MicroPhenoDB provides the web service APIs for programmatical access of the association data and produces an output in the JSON format. All the association data and the API documentation are available on the website. Users are also encouraged to submit their data of newly published microbe-disease associations. Once checked by our professional curators and approved by the submission review committee, the submitted record will be included in an updated release.

Applications of association data

MicroPhenoDB sequence search to explore metagenomics data

In MicroPhenoDB, microbes were connected with diseases through 5677 non-redundant associations and linked to unique clade-specific core genes via 696,934 relationships (**Figure 5**). Core genes could serve as a hub to connect metagenomic sequencing data to microbes and their associated diseases (Figure 5). A sequence search application was implemented on the MicroPhenoDB website (http://www.liwzlab.cn/microphenodb/#/tool) to allow users to query their metagenomic sequencing data against the MicroPhenoDB sequence datasets through

the sequence alignment tools BLAST [36] and Bowtie2 [37] (Figure 5). The application can directly identify the composition of pathogenic microorganisms in metagenomic samples and can suggest potential disease phenotypes that may be caused without running the usual metagenomic sequencing data processing and assembly, which are both time and resource consuming. Functional annotation for microbial core genes by the application includes gene ontology and pathway information. Searching against the sequence datasets of microbial pathogenic factors and drug resistance genes allows identifying homologous genes and proteins related to virulence factors and antibiotic resistance (Figure 5).

To assess the sequence search usability, we used the sequence search application to analyze an existing metagenomic dataset downloaded from the Genome Sequence Archive (accession: PRJCA000880) [38]. The dataset contained metagenomics data of lung biopsy tissues from 20 patients with pulmonary infection [39]. Our results identified pathogenic microbes in 95% (19 of 20) of patients, significantly higher than the 75% identification rate (15 of 20) found through the original metagenomic NGS (mNGS) analysis [39]. In addition, our search identified 37 pathogenic microbes in patients, while the mNGS method only identified 29 (Table S2). Of the 37 microbes, 23 were identical to those by mNGS analysis. It was hard to estimate the false positives of the other 14 microbes, but we found that they may cause infections in patients with underlying diseases such as immunodeficiency. Therefore, this comparison suggested that the MicroPhenoDB sequence search application could screen metagenomic data for effective identification of pathogenic microbes. Due to the large size of metagenomic data and the need for a broadband network, we provide a software package of the search application for users to download and run locally. We also encourage users to upload the microbial abundance information to the online application for further analysis and visualization.

Distinguish clinical phenotypes of SARS-CoV-2 infection from different viral respiratory infections

The single-stranded RNA coronavirus SARS-CoV-2 can infect humans and cause COVID-19 disease [40]. Its structure is similar to those of viruses causing severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) [41] . At present, the diagnosis of SARS-CoV-2 infection is mainly based on clinical phenotypes, chest computed tomography (CT), and nucleic acid testing. Compared with CT and nucleic acid testing, clinical phenotype monitoring has significant

advantages, such as a short turnaround time, low cost, and convenience [42]. To distinguish clinical phenotypes of SARS-CoV-2 infection from different viral respiratory infections, we searched MicroPhenoDB and obtained association data that contained 63 disease phenotypes and 14 respiratory tract infection viruses, such as human rhinovirus, parainfluenza virus, respiratory syncytial virus, metapneumovirus, and coronaviruses. The data were then imported into the Cytoscape software [43] for network analysis. The output network (**Figure 6**) indicated that SARS-CoV-2 shares the clinical phenotype of pneumonia with the majority of other respiratory infection viruses, as well as the clinical phenotypes of dry-cough, headache, fever, myalgia, vomiting, diarrhea, and respiratory disease syndrome (underlined in green) with several influenza viruses and other coronaviruses. Importantly, the network also showed that dyspnea, fatigue, lymphopenia, anorexia, and septic shock (underlined in blue) were common clinical phenotypes of SARS-CoV-2 infection distinguished from other viral respiratory infections [12,44,45]. Bear in mind that these phenotypes of SARS-CoV-2 infection might be frequent complications of other diseases and treatments. For example, dyspnea is a frequent complication of chronic respiratory diseases [46], lung cancer [47], and hepatopulmonary syndrome [48]; septic shock is a complication of pneumococcal pneumonia, chronic corticosteroid treatment, and current tobacco smoking [49]; fatigue is a complication of multi-type cancers [50,51] and Parkinson's disease [52]; lymphopenia is a complication of human immunodeficiency viral infection [53]. However, our results suggest that these common clinical phenotypes could distinguish SARS-CoV-2 infection from infections by SARS-CoV, MERS-CoV, and other respiratory viruses.

Association network in different body sites

The microbe-disease association data can be downloaded and used for further analysis. To generate a network to explore the reliable connections between the microbial changes and the diseases in multiple body sites, we obtained the association data of body sites such as the vagina, urinary tract, and genitals using the reliable association score thresholds mentioned above (> 0.3 and < -0.3). The resulting association data were imported into the Cytoscape software [43] for network analysis. The output network (**Figure 7**) indicated that the decreasing abundance of *Lactobacillus* (underlined in red) was related to vaginal inflammation and bacterial vaginosis in the vagina, while the increasing abundance of *Chlamydia* (underlined in

green) resulted in lymphogranuloma venereum in the genitals. Moreover, the network showed that the increasing abundance of *Mycoplasma genitalium* (underlined in blue) was associated with multiple diseases, which involve genitals, such as pelvic inflammatory disease, nongonococcal urethritis, and nonchlamydial nongonococcal urethritis. Furthermore, the network showed that a microbe abnormality could be associated with diseases involving different body sites. For example, the increasing abundance of *Neisseria gonorrhoeae* (underlined in purple) was associated with two diseases, each in the genitals and urinary tract. For users to assess the microbial pathogenicity, it is recommended to filter the data by using the association scores and follow the supporting publications for further investigation. Users can follow our step-by-step guidelines on the website (http://www.liwzlab.cn/microphenodb/#/guideline) to perform similar association analyses and generate Cytoscape networks.

Concluding remarks

Microbes play important roles in human health and disease. The curation and analysis of microbe-disease association data are essential for expediting translational research and application. In this study, we developed the MicroPhenoDB database by manually curating and consistently integrating microbe-disease association data. As far as we are aware, MicroPhenoDB is the first database platform to detail the relationships between pathogenic microbes, core genes, and disease phenotypes. In terms of data coverage, scoring models, and web applications, MicroPhenoDB outperformed data resources that contain similar association data (**Table 4**). For example, the numbers of associations, microbes, disease phenotypes, and supporting evidence in MicroPhenoDB were approximately 11.1, 6.1, 13.9, and 18.9-fold of those in HMDAD, respectively. Compared with both HMDAD and Disbiome, MicroPhenoDB refined the confidence scoring model using additional evidential metrics with different weights; it standardized the association annotations by manual curation and included pathogenic data of virulence factors, microbial core genes, and antibiotic resistance gens. Moreover, MicroPhenoDB implemented web applications and APIs for pathogenic microbe identifications in metagenomic data.

In MicroPhenoDB, many associations with confident scores came from our manual curation of the up-to-date clinical guidelines supported by IDSA and ASM. MicroPhenoDB assigned higher weight values to the associations derived from the

guidelines and lower weight values to the associations from other literature data and databases. The original model for scoring confidence of the disease-microbe associations in HMDAD was based on a single literature evidence. Our MicroPhenoDB score model rated different supporting evidence according to the credibility of related sources and provided a score to evaluate a disease-microbe association.

By integrating unique, clade-specific microbial core genes and using the data from MetaPhlAn2, the MicroPhenoDB sequence search application enables rapid identification of existing pathogenic microorganisms in metagenomic samples without running the usual sequencing data processing and assembly. However, the resulting associations from the sequence search do not guarantee microbial pathogenicity but provide clues for further investigation. The annotated core genes are also limited in size and cannot represent all microbial species. To consistently analyze the important functions of microbes, other data or tools are also recommended, such as UniRef clusters [54], MetaCyc [55], HUMAnN2 [56], and pan-genomic data.

To serve the research community, we will update the database every six months and constantly improve it with more features and functionalities. As a novel and unique resource, MicroPhenoDB connects pathogenic microbes, microbial core genes, and disease phenotypes; therefore, it can be used in metagenomic data analyses and assist studies in decoding microbes associated with human diseases.

Data availability

To access the association data, the online applications, and the software package, please visit http://www.liwzlab.cn/microphenodb/#/download.

CRediT author statement

Guocai Yao: Methodology, Software, Visualization, Writing - original draft preparation. **Wenliang Zhang:** Methodology, Data curation, Writing - original draft preparation. **Minglei Yang:** Web server, Visualization. **Huan Yang:** Validation. **Jianbo Wang:** Formal analysis. **Haiyue Zhang:** Investigation, Writing - reviewing & editing. **Lai Wei:** Resources, Writing - reviewing & editing. **Zhi Xie:** Resources, Writing - reviewing & editing. **Weizhong Li:** Conceptualization, Resources, Methodology, Supervision, Project administration, Writing – reviewing & editing. All

authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant Nos. 2016YFC0901604 and 2018YFC0910401) and the National Natural Science Foundation of China (Grant No. 31771478) to WL.

ORCID

0000-0002-9869-5124 (Guocai Yao)

0000-0003-0454-6935 (Wenliang Zhang)

0000-0001-9957-7608 (Minglei Yang)

0000-0001-8197-1041 (Huan Yang)

0000-0003-3916-3678 (Jianbo Wang)

0000-0002-7143-9550 (Haiyue Zhang)

0000-0002-3300-8506 (Lai Wei)

0000-0002-5589-4836 (Zhi Xie)

0000-0002-9003-7733 (Weizhong Li)

References

[1] Sender R, Fuchs S, Milo R. Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. Cell 2016; 164:337−40.

[2] Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med 2016; 8:51.

[3] Ghaisas S, Maher J, Kanthasamy A. Gut microbiome in health and disease: Linking the microbiome–gut–brain axis and environmental factors in the pathogenesis of systemic and neurodegenerative diseases. Pharmacol Therapeut 2016; 158:52−62.

[4] Cho I, Blaser MJ. The human microbiome: At the interface of health and disease. Nature reviews. Genetics 2012; 13:260−70.

[5] Kundu P, Blacher E, Elinav E, Pettersson S. Our gut microbiome: The evolving inner self. Cell 2017; 171:1481−93.

[6] Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S,

et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol 2019; 4:293−305.

[7] Jackson MA, Verdi S, Maxan M, Shin CM, Zierer J, Bowyer RCE, et al. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. Nat Commun 2018; 9:2655.

[8] Schmidt TSB, Raes J, Bork P. The human gut microbiome: From association to modulation. Cell 2018; 172:1198−215.

[9] Labus JS, Osadchiy V, Hsiao EY, Tap J, Derrien M, Gupta A, et al. Evidence for an association of gut microbial Clostridia with brain functional connectivity and gastrointestinal sensorimotor function in patients with irritable bowel syndrome, based on tripartite network analysis. Microbiome 2019;7: 45.

[10] Strati F, Cavalieri D, Albanese D, De Felice C, Donati C, Hayek J, et al. Altered gut microbiota in Rett syndrome. Microbiome 2016;4: 41.

[11] Chang HW, Yan D, Singh R, Liu J, Lu X, Ucmak D, et al. Alteration of the cutaneous microbiome in psoriasis and potential role in Th17 polarization. Microbiome 2018; 6:154.

[12] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020; 395:497−506.

[13] Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 2016; 352:565−9.

[14] Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, et al. Population-level analysis of gut microbiome variation. Science 2016; 352:560−4.

[15] Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. Genome Med 2011; 3:14.

[16] Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature 2019; 568:499−504.

[17] Maffert P, Reverchon S, Nasser W, Rozand C, Abaibou H. New nucleic acid testing devices to diagnose infectious diseases in resource-limited settings. Eur J Clin Microbiol 2017; 36:1717−31.

[18] Ma W, Zhang L, Zeng P, Huang C, Li J, Geng B, et al. An analysis of human microbe–disease associations. Brief Bioinform 2017; 18:85−97.

[19] Janssens Y, Nielandt J, Bronselaer A, Debunne N, Verbeke F, Wynendaele E, et al. Disbiome database: Linking the microbiome to disease. BMC Microbiol 2018;18:50.

[20] Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: Hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Res 2016; 44:D694−7.

[21] Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res 2017; 45:D566−73.

[22] Miller JM, Binnicker MJ, Campbell S, Carroll KC, Chapin KC, Gilligan PH, et al. A guide to utilization of the microbiology laboratory for diagnosis of infectious diseases: 2018 update by the Infectious Diseases Society of America and the American Society for Microbiology. Clin Infect Dis 2018; 67:813−6.

[23] Sioutos N, Coronado SD, Haber MW, Hartel FW, Shaiu W, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007; 40:30−43.

[24] Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 2015; 12:902−3.

[25] Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford) 2015; 2015:v28.

[26] Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2019; 47:D23−8.

[27] The UniProt Consortium. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res 2019; 47:D506−15.

[28] Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 2012; 9:811−4.

[29] Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, et al. InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 2019; 47:D351−60.

[30] Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics 2010; 26:1112−8.

[31] MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 2017; 45:D896−901.

[32] Perez-Riverol Y, Ternent T, Koch M, Barsnes H, Vrousgou O, Jupp S, et al. OLS client and OLS dialog: Open source tools to annotate public omics datasets. Proteomics 2017; 17:1700244.

[33] Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res 2019; 47:D1018−27.

[34] Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, et al. The Monarch Initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res 2017; 45:D712−22.

[35] Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res 2015; 43:W580−4.

[36] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinformatics 2009; 10:421.

[37] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012; 9:357−9.

[38] Zhang Z, Zhao W, Xiao J, Bao Y, He S, Zhang G, et al. Database resources of the national genomics data center in 2020. Nucleic Acids Res 2020; 48:D24−33.

[39] Li H, Gao H, Meng H, Wang Q, Li S, Chen H, et al. Detection of pulmonary infectious pathogens from lung biopsy tissues by metagenomic Next-Generation sequencing. Front Cell Infect Microbiol 2018; 8:205.

[40] Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. Int J Antimicrob Agents 2020; 55:105924.

[41] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. Lancet 2020; 395:565−74.

[42] Guan WJ, Zhong NS. Clinical characteristics of covid-19 in China. Reply. N Engl J Med 2020; 382:1861−2.

[43] Shannon P. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 2003; 13:2498−504.

[44] Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. Lancet 2020; 395:507−13.

[45] Lake MA. What we know so far: COVID-19 current clinical knowledge and research. Clin Med (Lond) 2020; 20:124−7.

[46] Dubé BP, Vermeulen F, Laveneziana P. Exertional dyspnoea in chronic respiratory diseases: from physiology to clinical application. Arch Bronconeumol 2017; 53:62−70.

[47] Henshall CL, Allin L, Aveyard H. A systematic review and narrative synthesis to explore the effectiveness of exercise-based interventions in improving fatigue, dyspnea, and depression in lung cancer survivors. Cancer Nurs 2019; 42:295−306.

[48] Gorgy AI, Jonassaint NL, Stanley SE, Koteish A, DeZern AE, Walter JE, et al. Hepatopulmonary syndrome is a frequent cause of dyspnea in the short telomere disorders. Chest 2015; 148:1019−26.

[49] Garcia-Vidal C, Ardanuy C, Tubau F, Viasus D, Dorca J, Liñares J, et al. Pneumococcal pneumonia presenting with septic shock: host- and pathogen-related factors and outcomes. Thorax 2010; 65:77−81.

[50] Baguley BJ, Skinner TL, Jenkins DG, Wright ORL. Mediterranean-style dietary pattern improves cancer-related fatigue and quality of life in men with prostate cancer treated with androgen deprivation therapy: A pilot randomised control trial. Clin Nutr 2020; S0261−5614:30250−8.

[51] Desai J, Deva S, Lee JS, Lin C, Yen C, Chao Y, et al. Phase IA/IB study of single-agent tislelizumab, an investigational anti-PD-1 antibody, in solid tumors. J Immunother Cancer 2020; 8:e000453.

[52] Schrag A, Hommel ALAJ, Lorenzl S, Meissner WG, Odin P, Coelho M, et al. The late stage of Parkinson's -results of a large multinational study on motor and non-motor complications. Parkinsonism Relat Disord 2020; 75:91−6.

[53] Pothlichet J, Rose T, Bugault F, Jeammet L, Meola A, Haouz A, et al. PLA2G1B is involved in CD4 anergy and CD4 lymphopenia in HIV-infected patients. J Clin Invest 2020; 130:2872−87.

[54] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2015; 31:926−32.

[55] Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 2016; 44:D471−80.

[56] Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods 2018; 15:962−8.

Figure legends

Figure 1 Workflow demonstrating the construction and curation of the MicroPhenoDB database

CARD, Comprehensive Antibiotic Resistance Database; EFO, Experimental Factor Ontology; HMDAD, Human Microbe-Disease Association Database; IDSA, Infectious Diseases Society of America; NCIT, National Cancer Institute Thesaurus; VFDB, Virulence Factor Database.

Figure 2 The distribution of association scores in MicroPhenoDB

Figure 3 Data content and distribution in MicroPhenoDB

A. The association data collected from different resources. **B.** The distribution of different microbe types. **C.** The number of bacterial species in different phyla. **D.** The disease distribution in MicroPhenoDB.

Note: HMDAD, Human Microbe-Disease Association Database; IDSA, Infectious Diseases Society of America; NCIT, National Cancer Institute Thesaurus.

Figure 4 The MicroPhenoDB web interface

Figure 5 The MicroPhenoDB sequence search connects microbes, core genes and disease phenotypes

BLAST, Basic Local Alignment Search Tool; SQL, Structure Query Language.

Figure 6 The Cytoscape network illustrates different clinical phenotypes across different viral respiratory infections

The diamonds represent the respiratory infection viruses. The red circles represent the disease phenotypes. Lager size of a circle or a diamond indicates more connections to a disease phenotype or a virus. The solid connection lines represent the associations between clinical phenotypes and viruses. Underlines indicate the clinical phenotypes discussed in the main text.

Figure 7 The Cytoscape network illustrates the associations between clinical phenotypes and microbes at different body sites

The diamonds represent clinical phenotypes resulted from a microbial abnormality at different body sites. The red circles represent the microbes. Lager size of a circle or a diamond indicates more connections to a clinical phenotype or a virus. The solid connection lines represent the associations between diseases and microbes with an increase in microbial abundance, and the dash connection lines represent the associations between diseases and microbes with a decrease in microbial abundance. Underlines indicate the microbes discussed in the main text.

Tables

Table 1 The weight of different evidential sources according to their reliabilities

Table 2 Data scope and scale in MicroPhenoDB

Table 3 The top ten body sites of disease-associated microbes in MicroPhenoDB

Table 4 Data content and web applications of MicroPhenoDB compared with

HMDAD and Disbiome

Supplementary material

Figure S1 The distribution of numbers of supporting publications

The blue histogram represents the frequency of the number of supporting publications.

Table S1 The version or release of databases and tools used in the MicroPhenoDB construction

Table S2 The analysis result by MicroPhenoDB sequence search in an existing metagenomic dataset (GSA: PRJCA000880)

Table 1 The weight of different evidential sources according to their reliabilities

Evidence	Evidence definition	Weight
Literature	Author statement supported by traceable literature used in manual assertion	0.25
Data resource	Statement supported by manual curate data resource such as NCIT	0.5
Guideline	Consensus statement supported by the IDSA guideline	1.0

Note: IDSA, Infectious Diseases Society of America; NCIT, National Cancer Institute Thesaurus OBO Edition.

Table 2 Data scope and scale in MicroPhenoDB

Data scope	Data scale
Association	5677 non-redundant microbe-disease associations
Microbe	1781 microbe species including 1497 bacteria in a broad sense (including 1474 bacteria in a narrow sense, 11 *Rickettsia*, 6 *Chlamydia*, 4 *Ehrlichia,*

	and 2 *Mycoplasma*), 183 viruses, 58 fungi, and 43 parasites
Disease phenotype	542 disease phenotypes annotated with EFO across more than 22 body sites
Core gene	685 bacteria and viruses annotated with 27,277 unique clade-specific core genes
Virulence factor	65 (4.3% of 1497) bacteria annotated with information of more than 4204 virulence factors, including pathogenic species, virulence factor gene name, characteristics structure, and pathogenic mechanism
Antibiotic resistance data	66 (4.4% of 1497) bacteria annotated with information of more than 2522 antimicrobial resistances, including resistance genes, resistance mechanisms, and related antibiotics

Table 3 The top ten body sites of disease-associated microbes in MicroPhenoDB

Body site	Association number	Percentage
Gastrointestinal tract	2119	37.3%
Oral cavity	542	9.5%
Respiratory tract	391	6.9%
Skin and soft tissue	239	4.2%
Urinary tract	197	3.5%
Vagina	143	2.5%
Central nervous system	114	2.0%
Nasal cavity	85	1.5%
Bloodstream	83	1.5%
Throat	69	1.2%

Table 4 Data content and web applications of MicroPhenoDB compared with HMDAD and Disbiome

Data content & web applications	MicroPhenoDB	HMDAD	Disbiome	MicroPhenoDB/HMDAD	MicroPhenoDB/Disbiome
Association data	7449	673	4567	11.1	1.6
Microbe organism	1781	292	1292	6.1	1.4

Microbe standardized annotation	1041	-	-	-	-
Organism taxonomy	1032	-	-	-	-
Disease phenotype	542	39	282	13.9	1.9
Disease standardized annotation	446	-	-	-	-
Supporting evidence	1150	61	822	18.9	1.4
Association score	Yes	None	None	-	-
Virulence factor	4204	-	-	-	-
Antibiotic resistance gene	2522	-	-	-	-
Core gene	696,934	-	-	-	-
Sequence alignment	Yes	None	None	-	-
Identification of pathogenic microbe	Yes	None	None	-	-
Web service API	Yes	None	None	-	-

Note: HMDAD, Human Microbe-Disease Association Database; API, Application Programming Interface.