Microbiome Search Engine 2: a Platform for Taxonomic and Functional Search of Global Microbiomes on the Whole-Microbiome Level

Gongchao Jing, Lu Liu, Zengbin Wang, Yufeng Zhang, Li Qian, Chunxiao Gao, Meng Zhang, Min Li, Zhenkun Zhang, Xiaohan Liu, Jian Xu, Xiaoquan Su

College of Computer Science and Technology, Qingdao University, Qingdao, China
Single-Cell Center, CAS Key Laboratory of Biofuels and Shandong Key Laboratory of Energy Genetics, Qingdao Institute of BioEnergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao, China
Shandong Institute of Energy Research, Qingdao, China
National Science Library, Chinese Academy of Sciences, Beijing, China
Department of Computer Science and Technology, Ocean University of China, Qingdao, China
Key Laboratory of Marine Drugs, the Ministry of Education of China, School of Medicine and Pharmacy, Ocean University of China, Qingdao, China

Gongchao Jing and Lu Liu should be regarded as joint first authors. Order of co-first author names was determined by mutual agreement.

ABSTRACT Metagenomic data sets from diverse environments have been growing rapidly. To ensure accessibility and reusability, tools that quickly and informatively correlate new microbiomes with existing ones are in demand. Here, we introduce Microbiome Search Engine 2 (MSE 2), a microbiome database platform for searching query microbiomes in the global metagenome data space based on the taxonomic or functional similarity of a whole microbiome to those in the database. MSE 2 consists of (i) a well-organized and regularly updated microbiome database that currently contains over 250,000 metagenomic shotgun and 16S rRNA gene amplicon samples associated with unified metadata collected from 798 studies, (ii) an enhanced search engine that enables real-time and fast (>0.5 s per query) searches against the entire database for best-matched microbiomes using overall taxonomic or functional profiles, and (iii) a Web-based graphical user interface for user-friendly searching, data browsing, and tutoring. MSE 2 is freely accessible via http://mse.ac.cn. For standalone searches of customized microbiome databases, the kernel of the MSE 2 search engine is provided at GitHub (https://github.com/qibebt-bioinfo/meta-storms).

IMPORTANCE A search-based strategy is useful for large-scale mining of microbiome data sets, such as a bird’s-eye view of the microbiome data space and disease diagnosis via microbiome big data. Here, we introduce Microbiome Search Engine 2 (MSE 2), a microbiome database platform for searching query microbiomes against the existing microbiome data sets on the basis of their similarity in taxonomic structure or functional profile. Key improvements include database extension, data compatibility, a search engine kernel, and a user interface. The new ability to search the microbiome space via functional similarity greatly expands the scope of search-based mining of the microbiome big data.

KEYWORDS amplicon, metagenome, microbiome, online service, search engine

Metagenomic approaches have been widely employed to probe microbiomes among various habitats by linking dynamics of microbial compositions and predicted functions to environmental changes (1, 2), human disease development (3–7), and drug responses (8, 9). With the rapid development of sampling strategies and sequencing technologies, an enormous volume of microbiome data sets, including...
both 16S rRNA gene amplicon-based and shotgun whole-genome sequencing (WGS)-based data sets, has been produced by individual, small-cohort projects or large-scale surveys, such as the Human Microbiome Project (10), the Earth Microbiome Project (11), the American Gut Project (12), and Tara Oceans (13). Most DNA sequence data are stored in either general-purpose DNA sequence repositories (e.g., NCBI SRA [14]) or microbiome- specific databases (e.g., MG-RAST [15] and EBI Metagenomics [16]). To support large-scale mining of the existing microbiome big data, tools have been introduced to organize metagenomes with unified sequence processing standard operating procedures (SOPs) (17), e.g., Qiita (18), gcMeta (19), and GMrepo (20). These tools typically support queries based on taxonomy terms (e.g., species name), sequence fragments, or Structured Query Language (SQL)-like metadata. To support the search of newly generated data sets against the existing microbiome big data based on taxonomic or functional similarity, Microbiome Search Engine (MSE) (21) was recently introduced, and it shows promise for search-based multiple-disease classification in a cross-cohort, sequence-platform-insensitive, and contamination-tolerant manner (22). However, it supports only amplicon sequencing-based data sets, which limits the queries to those probing taxonomical similarity of microbiomes (21).

To address this limitation, here we introduce Microbiome Search Engine 2 (http://mse.ac.cn), which enables the search of an amplicon or a shotgun WGS-based query microbiome against a large database based on the “functional” similarity of the microbiome (“taxonomical” similarity is also supported) (Fig. 1a). This platform, a significant improvement over the previous version (21), consists of three main components (Fig. 1b): (i) a well-maintained and regularly updated microbiome database that has been expanding since 2016 (see Fig. S1 in the supplemental material) and currently contains over 250,000 globally sampled (human, animal, marine, soil, etc.), curated microbiomes (both WGS- and amplicon-based samples) that are associated with a unified scheme of metadata from 798 studies; (ii) an enhanced search engine kernel that is compatible with both amplicon and shotgun WGS-based sequences and enables real-time searches against the database for best matches in microbiome taxonomy or function; and (iii) a Web-based graphical user interface that provides easy-use searching, data browsing, and tutoring.

**FIG 1** Design principles of MSE 2. (a) MSE 2 enables the search of a given amplicon- or shotgun-based microbiome against a large database, based on the whole-microbiome-level taxonomical or functional similarity between microbiomes. (b) The three key components of MSE 2 include a well-organized database, an enhanced search engine, and a Web-based interface.
 RESULTS

Microbiome database. (i) Data collection and curation. Clean sequences (refer to Materials and Methods) and their metadata were collected mainly from the Qiita (18), EBI (16), SRA (14), and MG-RAST (15) repositories. The common items of metadata for each study (e.g., project name, description, publication, etc.) (Table 1, project metadata) and sample (e.g., habitat, sequencing type, sampling year, etc.) (Table 1, sample metadata) were selected and manually integrated into a specific format, while the complete original metadata were also preserved. To ensure technical comparability and searchability among microbiome samples, sequences were preprocessed and profiled by unified methods according to sequence types (i.e., amplicon based or shotgun WGS based [Table 2]; for details, see Materials and Methods).

(ii) Database statistics. After the data preprocessing and curation (details are in Materials and Methods), a total of 250,273 microbiome samples from 798 projects/studies were included in the current MSE 2 database, including 14,957 shotgun WGS-based metagenomes and 235,334 16S rRNA gene amplicons. In terms of sampling source distribution (Fig. 2), human-associated habitats are the most frequent (52.8% in total; gut, 34.2%; skin, 9.1%; oral, 6.4%, etc.), followed by animal-associated habitats.
(23.7%), soil (6.4%), indoor environments (5.7%), and marine environments (2.7%) (for details, see Table S1 in the supplemental material).

(iii) Database organization and management. All microbiome samples are organized into two dimensions (Fig. 3). For Web-based data browsing (refer to the “Data browsing and download” section below), samples are arranged by studies and can be selected and filtered by the various metadata (e.g., habitat, sequence type, year, etc.). For searches based on taxonomical or functional similarity with microbiomes, samples were presorted by compositional features (e.g., operational taxonomic unit [OTU], species, or KEGG Orthology [KO] identifier [ID]) for indexing and searching (refer to the “Enhanced microbiome search engine” section below for details).
Enhanced microbiome search engine. (i) Whole-microbiome-level search. The search engine, as the kernel of MSE 2, was developed by C++ and optimized by OpenMP-based parallel computing. With a given query microbiome, MSE 2 searches it against the entire microbiome database for best-matched samples that have the highest taxonomical or functional similarity. The search results present the taxonomical or functional profiles of the matches, the quantitative similarity values (refer to Materials and Methods for more details) compared to the query, and their metadata information (refer to the "Microbiome search and interpretation of the search results" section below for more details). Compared to the previous version (21), which accepts only OTUs from 16S rRNA gene amplicons as the query, the search engine extended its capability by supporting OTU-based (via profiles derived from 16S rRNA gene amplicons), species-based (via profiles derived from shotgun WGS) searches, and metabolic function-based (via profiles derived from either shotgun WGS or 16S rRNA gene amplicons) searches (Fig. 4a).

(ii) Speed and scheduling. Benefited by a two-tier indexing and searching strategy (Fig. 4b; refer to Materials and Methods for details), this search engine is typically 1 to 2 orders of magnitude faster than exhaustive searches that directly compare the query to all the database samples. To test the indexing efficiency and searching speed of MSE 2, we performed OTU-based, species-based, and function-based searches against the entire database and compared the search time to that of an index-disabled exhaustive search (the exhaustive search is for in-house performance evaluation only and not provided in the public online service of MSE 2). Each process was repeated 10 times, and only the search running times (excluding the upload time, visualization time, and
The results showed that the indexing strategy accelerates the search speeds by up to 193 times, 15 times, and 605 times, respectively, for OTU-based, species-based, and KO ID-based searches (Fig. 4c and Table 3), corresponding to a real-time response of within 0.5 s for a whole-microbiome-level query against the over 250,000 samples. In addition, the online search service follows the “first come, first served” principle implemented by queue-based task scheduling, so that the computing resources are utilized efficiently.

**Graphical Web-based portal.**

(i) **Web-based user interface.** MSE 2 is freely accessible via [http://mse.ac.cn](http://mse.ac.cn) via Web browsers. Developed by PHP and MySQL under a Linux server, this website provides a user-friendly graphical interface (Fig. 5) for searching, data browsing, and data uploading/downloading. Tutorial materials are available for users to adjust the parameters for customized functions and result interpretation. Notifications of database updates, system maintenance, and other related information are regularly published. Users can also post any questions or bugs at the Help Desk and obtain replies via e-mail.

(ii) **Microbiome search and interpretation of the search results.** For microbiome searches, MSE 2 accepts the compositional features of a sample (OTUs, species, or KO IDs) as queries. Notably, query microbiomes should be preprocessed from sequences into compositional features in an identical way to that used for the database samples. Table 2 summarizes the recommended software for sequence processing for each sequence type, and the detailed analytical protocol is available via the "Search" or "Help" page. To submit a search, users first choose a search type from "Search by OTU," "Search by species," and "Search by function," depending on the type of query input (Fig. 5b). Then the query can be either uploaded from a tabular plain-text file or directly pasted into the text box of the Web page. Users can also specify other parameters, such as the maximum match number (the default is 10) and the cutoff similarity (the default is 0.6).

In the result page (Fig. 5c), top-matched samples from the database are listed with sample IDs, habitats, and similarity values relative to the query (nearest sequenced taxon index [NSTI] values are also provided for 16S rRNA gene-inferred functional profiles). Each sample ID is linked to its corresponding page with detailed full metadata (e.g., source study, sampling site, sequence type, etc.). The microbial compositions of the query and matched samples are visualized via both the bar chart and the Krona-based (23) interactive animation (Fig. 5d), so as to illustrate their links and distinctions in detail. Furthermore, all the above search results are packed for download in the result page for subsequent in-depth meta-analysis and data mining by users.

(iii) **Data browsing and download.** The MSE 2 online service provides two ways of sample browsing.

(a) **Browse by project.** In the project list page, samples are organized per project, and all projects are listed and sorted by project ID. Project pages can be accessed by clicking the project ID in the list or searched by metadata key words. Each project page contains the unified metadata (e.g., study title, publication, etc.) (Table 1, project metadata), original full list of metadata, links to samples in this project, and links to its data source.

(b) **Browse by sample.** In the sample list page, all samples are listed and sorted by sample ID, and samples can also be selected by a metadata filter for specific habitat, sequencing type, sampling year, etc. For a given sample in the database, all the unified metadata information (Table 1, sample metadata) can be displayed and the microbial taxonomy hierarchy visualized by Krona (23) by clicking on the sample ID.

<table>
<thead>
<tr>
<th>Search type</th>
<th>Index-based search time (s)</th>
<th>Index-disabled search time (s)</th>
<th>Avg speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>By OTU</td>
<td>0.241 ± 0.004</td>
<td>46.524 ± 1.029</td>
<td>193.345</td>
</tr>
<tr>
<td>By species</td>
<td>0.020 ± 0.001</td>
<td>0.300 ± 0.035</td>
<td>14.883</td>
</tr>
<tr>
<td>By function</td>
<td>0.101 ± 0.002</td>
<td>60.886 ± 0.189</td>
<td>605.076</td>
</tr>
</tbody>
</table>

*Each search procedure was repeated 10 times.*
In this work, we introduce Microbiome Search Engine 2 (MSE 2), which features (i) an expanded database of over 250,000 shotgun metagenomic and 16S rRNA gene amplicon samples associated with unified metadata collected from 798 studies and (ii) an enhanced search engine for real-time and fast (~0.5 s per query) searches for best-matched microbiomes via not just taxonomic but also functional profiles. The value of a search-based strategy has been demonstrated for defining the novelty of microbiome samples (21) and for cross-cohort disease diagnosis (22, 24). By adding a function-based dimension for these and related applications, MSE 2 should accelerate large-scale mining of the ever-expanding metagenome data space.

MATERIALS AND METHODS

Sequence preprocessing of the microbiome database. For shotgun sequences, MetaPhlAn2 (25) was used for species-level bacterial taxonomy assignment, and functional profiles were analyzed by HUMAnN2 (26) using "uniref90 gene families" and annotated on the basis of the KEGG Orthology (KO).
database (27). For 16S rRNA gene amplicon sequences, OTUs were picked against Greengenes reference data (version 13-8) (28) with 97% similarity by Parallel-META 3 (29), and the relative abundances of OTUs were corrected by amplicon copy numbers that were parsed from the IMG/M database (30). Then the functional profiles were translated into KEGG orthologs by PICRUSt 2 (31, 32), while the NSTI (nearest sequenced taxon index) values that measure the prediction reliability were also recorded.

**Curation of the microbiome database.** All the data sets in the MSE 2 microbiome database are “clean” DNA sequence reads, i.e., reads that have already passed the initial sequence quality control process, including primer and tag clipping, host DNA removal, filtering of low-quality reads, etc. Microbiome samples were further curated based on the comprehensiveness of metadata and the quality of taxonomical or functional profiling. Specifically, samples that lack source habitat information were excluded (e.g., the habitat domain, habitat type, or habitat details in Table 1). Moreover, those amplicon samples with either <500 total reads or >20% reads that cannot be annotated were removed from the microbiome database.

**Indexing strategy of the microbiome search engine.** MSE 2 performs a two-tier indexing strategy (21) in order to enable real-time speed searches of large microbiome databases (Fig. 4b). To build the database index, for each database entry sample, MSE 2 partitions its profile features so that the compositional complexity is reduced. For taxonomic profiles (OTUs and species), features are sorted and merged by family-level taxa; for functional profiles, KEGG orthologs that belong to the same KEGG BRITE level 2 pathways are combined. The merged features are treated as the index keys. When a given query microbiome is searched, in step 1, the search engine parses its index keys in the same way as the database entries and dynamically selects those “candidate matches” with the shortest distances to the query sample on the index keys. Then, in step 2, MSE 2 identifies the top matches via a pairwise comparison between the query and each of the “candidate matches” using the similarity algorithms (refer to the “Similarity algorithms of the microbiome search engine” section below for details). Since the index keys are from a particular partition of the profile features, this two-tier search is typically 1 to 2 orders of magnitude faster than exhaustive searches that directly compare the query to all the database samples (refer to the “Speed and scheduling” section above for details).

**Similarity algorithms of the microbiome search engine.** After indexing, MSE 2 calculates the whole-microbiome-level similarity (in taxonomy or function) between the query and each candidate selected by indexing. For OTU- and species-based searches, MSE 2 uses the Meta-Storms (33) and Dynamic Meta-Storms (34) algorithms, respectively, both of which employ phylogeny-based metrics to quantitatively assess the similarity between microbiome samples. For function-based searches, MSE 2 applies Bray-Curtis dissimilarity to compare query and database entries based on relative abundances of KEGG orthologs.

**Implementation of the MSE 2 system and the Web-based portal.** MSE 2 runs under a CentOS Linux (release version 7.4) operating system. The search engine, as the kernel of MSE 2, was implemented in C++ and optimized by OpenMP-based parallel computing. The online service system was designed and constructed based on the LAMP (consisted of Linux, Apache, MySQL, and PHP) architecture. In this system, Apache software provided the accessibility of the Web pages that were developed with the PHP language. The metadata of projects and samples were arranged using the MySQL database. All data of MSE 2 were stored in a RAID 5 (Redundant Arrays of Independent Disks 5) storage node for data safety and security, while the index of the search engine was kept on a solid-state drive (SSD) RAID for fast fetch.

**Source code and data availability.** The microbiome database, including microbiome taxonomy, function, and metadata, are open for download from the MSE 2 website. The kernel search engine software was released on GitHub (https://github.com/qibebt-bioinfo/meta-storms) for standalone use with user-defined microbiome databases.

**SUPPLEMENTAL MATERIAL**

Supplemental material is available online only.

**FIG S1**, PDF file, 0.1 MB.

**TABLE S1**, PDF file, 0.1 MB.

**ACKNOWLEDGMENTS**

We thank Yinghe Qiao of QIBEBT, CAS, for server system administration and hardware maintenance and the operations research community at QIBEBT, CAS, for inspiring discussions.

X.S. acknowledges the support of grants 31771463 and 32070086 from the National Natural Science Foundation of China. This project is also supported by grants from the Shandong Natural Science Foundation, China (ZR201807060158 for G.J., ZR2019BC087 for L.L., and ZR2016QZ004 for J.X.).

We have no conflicts of interest to declare.

**REFERENCES**


Taxonomic and Functional Search in Global Microbiomes


