

1-5-2018

The Microbe Directory: An annotated, searchable inventory of microbes' characteristics

Heba Shaaban
CUNY Hunter College

David A. Westfall
Weill Cornell Medicine

Rawhi Mohammad
CUNY College of Staten Island

David Danko
Weill Cornell Medical College

Daniela Bezdán
Weill Cornell Medical College

See next page for additional authors

[How does access to this work benefit you? Let us know!](#)

Follow this and additional works at: https://academicworks.cuny.edu/hc_pubs

 Part of the [Biochemistry, Biophysics, and Structural Biology Commons](#), and the [Cell and Developmental Biology Commons](#)

Recommended Citation

Shaaban, Heba; Westfall, David A.; Mohammad, Rawhi; Danko, David; Bezdán, Daniela; Afshinnekoo, Ebrahim; Segata, Nicola; and Mason, Christopher E., "The Microbe Directory: An annotated, searchable inventory of microbes' characteristics" (2018). *CUNY Academic Works*.

https://academicworks.cuny.edu/hc_pubs/380

Authors

Heba Shaaban, David A. Westfall, Rawhi Mohammad, David Danko, Daniela Bezdán, Ebrahim Afshinnekoo, Nicola Segata, and Christopher E. Mason



DATA NOTE

The Microbe Directory: An annotated, searchable inventory of microbes' characteristics [version 1; referees: 1 approved, 3 approved with reservations]

Heba Shaaban^{1-3*}, David A. Westfall^{1,2,4*}, Rawhi Mohammad ^{1,2,5}, David Danko^{1,2}, Daniela Bezdán^{1,2}, Ebrahim Afshinnekoo^{1,2,6}, Nicola Segata⁷, Christopher E. Mason ^{1,2,8}

¹Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, 10065, USA

²The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, 10065, USA

³CUNY Hunter College, New York, NY, 10065, USA

⁴School of Medicine, Weill Cornell Medicine, New York, NY, 10065, USA

⁵CUNY College of Staten Island, Staten Island, NY, 10314, USA

⁶School of Medicine, New York Medical College, Valhalla, NY, 10595, USA

⁷Centre for Integrative Biology, University of Trento, Trento, 38122, Italy

⁸The Feil Family Brain and Mind Research Institute, New York, NY, 10065, USA

* Equal contributors

v1 First published: 05 Jan 2018, 2:3 (doi: [10.12688/gatesopenres.12772.1](https://doi.org/10.12688/gatesopenres.12772.1))
Latest published: 05 Jan 2018, 2:3 (doi: [10.12688/gatesopenres.12772.1](https://doi.org/10.12688/gatesopenres.12772.1))

Abstract

The Microbe Directory is a collective research effort to profile and annotate more than 7,500 unique microbial species from the MetaPhlan2 database that includes bacteria, archaea, viruses, fungi, and protozoa. By collecting and summarizing data on various microbes' characteristics, the project comprises a database that can be used downstream of large-scale metagenomic taxonomic analyses, allowing one to interpret and explore their taxonomic classifications to have a deeper understanding of the microbial ecosystem they are studying. Such characteristics include, but are not limited to: optimal pH, optimal temperature, Gram stain, biofilm-formation, spore-formation, antimicrobial resistance, and COGEM class risk rating. The database has been manually curated by trained student-researchers from Weill Cornell Medicine and CUNY—Hunter College, and its analysis remains an ongoing effort with open-source capabilities so others can contribute. Available in SQL, JSON, and CSV (i.e. Excel) formats, the Microbe Directory can be queried for the aforementioned parameters by a microorganism's taxonomy. In addition to the raw database, The Microbe Directory has an online counterpart (<https://microbe.directory/>) that provides a user-friendly interface for storage, retrieval, and analysis into which other microbial database projects could be incorporated. The Microbe Directory was primarily designed to serve as a resource for researchers conducting metagenomic analyses, but its online web interface should also prove useful to any individual who wishes to learn more about any particular microbe.

Open Peer Review

Referee Status: 

	Invited Referees			
	1	2	3	4
version 1 published 05 Jan 2018				
	report	report	report	report

- David A. Coil**, University of California, Davis, USA
- James E. McDonald** , Bangor University, UK
- Nicole M. Vega** , Emory University, USA
- Elisabeth M. Bik** , uBiome, USA

Discuss this article

Comments (3)

Corresponding author: Christopher E. Mason (chm2042@med.cornell.edu)

Author roles: **Shaaban H:** Data Curation, Project Administration, Writing – Original Draft Preparation; **Westfall DA:** Data Curation, Project Administration, Software, Writing – Original Draft Preparation; **Mohammad R:** Software, Writing – Review & Editing; **Danko D:** Software, Visualization, Writing – Review & Editing; **Bezdán D:** Writing – Review & Editing; **Afshinnekoo E:** Conceptualization, Supervision, Writing – Original Draft Preparation; **Segata N:** Writing – Review & Editing; **Mason CE:** Conceptualization, Supervision, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

How to cite this article: Shaaban H, Westfall DA, Mohammad R *et al.* **The Microbe Directory: An annotated, searchable inventory of microbes' characteristics [version 1; referees: 1 approved, 3 approved with reservations]** Gates Open Research 2018, 2:3 (doi: [10.12688/gatesopenres.12772.1](https://doi.org/10.12688/gatesopenres.12772.1))

Copyright: © 2018 Shaaban H *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: We would like to thank the Epigenomics Core Facility at Weill Cornell Medicine, funding from the Irma T. Hirschl and Monique Weill-Caulier Charitable Trusts, Bert L and N Kuggie Vallee Foundation, the WorldQuant Foundation, The Pershing Square Sohn Cancer Research Alliance, NASA (NNX14AH50G, NNX17AB26G), the National Institutes of Health (R25EB020393, R01NS076465, R21AI129851), the Bill and Melinda Gates Foundation (OPP1151054), and the Alfred P. Sloan Foundation (G-2015-13964).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 05 Jan 2018, 2:3 (doi: [10.12688/gatesopenres.12772.1](https://doi.org/10.12688/gatesopenres.12772.1))

Introduction

With the advent of next-generation sequencing technologies, there has been a surge of metagenomic and microbiome studies in the last decade, ranging from studying the human microbiome¹ to the environment (water and soil)²⁻⁵, and city surfaces^{6,7}. All these studies depend heavily on bioinformatics analyses that translate the sequences they uncover to taxonomic profiles found in their samples. However, an immediate challenge from taxonomic outputs is the interpretation of the data. Learning more about a microorganism's properties, such as optimal pH and temperatures, presence in the human microbiome, ability to form spores or biofilms, and antimicrobial sensitivity, amongst many others, are key to understanding the biochemical and ecological dynamics of the microbiomes that can be found. Despite the presence of several databases that include some of this information, such as [MicrobeWiki](#), [PATRIC](#), [ARDB](#), and [IMG-JGI](#), these databases are either incomplete or focus on a specific characteristic (e.g. antimicrobial resistance). The Microbe Directory seeks to fill this gap with an online tool that aggregates these data and expands their annotations, which thus provides a useful tool for exploration of functional, medical, or biological traits found in any microbial community.

Methods

MetaPhlAn2 list of species

The list of distinct species that was subject to curation was generated from the [MetaPhlAn2 database](#), a computational tool for profiling the composition of microbial communities from sequencing data. MetaPhlAn2 works by relying on unique clade-specific marker genes identified from more than 16,000 reference genomes from NCBI and RefSeq⁸. It provides a 7-level (kingdom to strain) consistent taxonomic characterization of known domains of life and currently has identified >7,500 unique species in its database. This database was specifically chosen for the Microbe Directory due to its prevalent usage in microbiome and metagenomic studies⁹, allowing researchers to directly integrate the Microbe Directory into their research to learn more from the MetaPhlAn output¹⁰. Furthermore, there is a built-in capability for researchers to contribute and expand the Microbe Directory beyond the species currently curated in the database (see *Using the Microbe Directory*).

Selection and training of researchers

The Microbe Directory database was curated by a team of trained undergraduate, graduate, and medical students from City University of New York (CUNY) Hunter College, Macaulay Honors College, and Weill Cornell Medicine (see full list of students in *Acknowledgements*). The student-researchers were selected from a pool of applicants and underwent a three-hour training session that a) explained the objective of the research project and the desired outcome, b) provided a detailed and thorough explanation of each of the parameters that were the subject of research, and c) provided clear instructions on how

to curate the internet for the parameters for each species. They were also given a tutorial on how to conduct the research for a sample of 10 species. They were given a list of annotation-based websites to assist their research, but they were not limited to using only those sites. (see *Annotation Tutorial and Guidelines* in [Supplementary File 1](#)).

After every entry, students inserted citation links to the sources they utilized for the information they inputted. Each student-researcher independently worked 4–5 hours per week to curate parameters for 10 species per week, for a total of 20 weeks. To ensure that students were not making errors during curation, the first three weeks of the project were heavily monitored and entries were manually checked for inaccuracies by the project leads. After the first 3-week trial, only two randomly selected species were checked manually from every submitted entry of 10 species per week, per student. Considerable error rates (3 or more incorrect annotations out of 10 being the threshold) consequently meant the student had to resubmit the entire set of 10 species the following week. While there is always the potential for human error in manually curated databases, the Microbe Directory has a feature where anyone can make an account and submit edits and changes to the information hosted in the database. Thus, there is potential for the Microbe Directory to continue to grow and expand, but also ensure minimal errors in its database.

Building the microbe directory

[Table 1](#) defines the various microbial characteristics and categories of information that were curated to build the Microbe Directory. The parameters chosen were strictly objective features of microbes that are important to help interpret and understand the findings and context of whatever microbiome a researcher is studying. There is built-in potential to expand the Microbe Directory and for researchers to contribute more characteristics of these microbes, including native location, industrial applications, and associated symptoms/diseases; these features were considered to be included in the Microbe Directory but due to their subjective nature were omitted out to maintain proper quality control outlined above. Several databases were used to collect this information, including [COGEM](#), [MicrobeWiki](#), [BacMap](#), [ATCC](#), [PATRIC](#), [ARDB](#), [GOLD](#), [HOMD](#), and [BEI Resources](#) (see *Annotation Tutorial and Guidelines and Links* in [Supplementary File 1](#)). These peer-reviewed resources and databases have been well-established in the literature as reliable sources of information for researchers. Now, this information can be housed in one place, allowing for more efficient and comprehensive interpretation of microbiome analysis. [Figure 1](#) is a heatmap summarizing the current information hosted in the Microbe Directory's database across all species and parameters.

Pre-search. Before assignments were given to the student-researchers, the databases listed above were pre-searched in order to collect as much information as possible about the microbes.

Table 1. The Microbe Directory inventory parameters and descriptions.

Parameter	Definition and notes
Optimal pH	The optimal pH at which this species grows. If the species was not widely studied, the American Type Culture Collection (ATCC) was used to determine the optimal pH for storage. If two far ranges of pH were determined, the average was taken.
Optimal temperature	The optimal temperature at which this species grows. If the species was not widely studied, the ATCC was used to determine the optimal temperature for storage. If two far ranges of temperatures were determined, the average was taken.
COGEM pathogenicity rating	COGEM released a comprehensive database of pathogenicity assessment of around 2575 bacterial species in 2011 ¹⁰ . The database ranks the pathogenicity of species on a scale of 1 to 4 - 1 being not belonging to a recognized group of disease-invoking agents in humans or animals and having an extended history of safe usage and 4 being a species that can cause a very serious human disease, for which no prophylaxis is known.
Antimicrobial susceptibility	Are there any known antibiotics that this species is sensitive to? No = 0, Yes = 1
Spore-formation	Is the species spore-forming? No = 0, Yes = 1
Biofilm-formation	Is the species biofilm-forming? No = 0, Yes = 1
Extremophile	Extremophiles are organisms that live in extreme environments, as opposed to organisms that live in moderate (mesophilic) environments. This category includes acidophiles, thermophiles, osmophiles, halophiles, oligotrophs, and others. Mesophiles = 0, Extremophile = 1
Gram-stain	Negative = 0, Positive = 1, Indeterminate = 2
Found in human microbiome	Microbes that live anywhere in the human body and are not pathogenic to humans (i.e. capable of causing human disease) No=0, Yes=1
Plant pathogen	Does the species causes disease in plants? No = 0, Yes = 1
Animal pathogen	Does the species causes disease in animals? No = 0, Yes = 1

This was done using each website's search page. The species name was used as the search query, and the search results html page was parsed using regular expressions. The first search result that contained the microbe's binomial name and contained a link to the website's entry for that microbe was used as the pre-search's result. Such links for each microbe were compiled and given to each student with his or her weekly assignments. The student-researchers were only given the link to the entry, and they then had to manually find the relevant information (e.g. "optimal pH"). Such a system allowed the students to manually confirm that the pre-search identified the correct entry for the microbe and not just a microbe with a similar name. We also supplemented the manual curation by parsing MicrobeWiki for common keywords that could indicate particular features. We found that we could extract useful data for pathogenicity, biofilm-formation, microbe shape, halophilicity, spore formation, and metabolism. We were able to extract some subset of these features for 331 of the microbes that had been manually curated.

Text validation and normalization. Student-researchers filled out the columns for a given microbe using an Excel spreadsheet. Each entry was filled out as free-form text, so it was necessary to later normalize and validate the text. Valid column types included positive real numbers (e.g. optimal pH), ranges of positive real numbers (e.g. range of optimal pH values), series of ranges (e.g. multiple optimal pH ranges), binary values (e.g. spore

forming or non-forming), ternary values (e.g. Gram-positive, Gram-negative, Gram-indeterminate), and quaternary values (e.g. COGEM Classes 1-4). Regular expressions (RegEx) were used to ensure that a given column entry conformed to the correct type (i.e. validation); validated columns were then transformed to a common form (i.e. normalization). The common form for each entry is the form used in the database.

Using the Microbe Directory

The Microbe Directory can be accessed online at <https://microbe.directory>. This interface provides individual users a way to browse and search the directory's contents in an interactive format. Such a representation should prove useful for researchers who need information for a particular microbe. While viewing the page for a given microbe, registered users can also submit edits to that microbe's data. Individuals can register to contribute to the Microbe Directory by signing up [here](#). The edits are then put in a queue to be later reviewed by The Microbe Directory team (HS, DAW, RS).

In addition to the interactive web interface, the main website provides links to the project's [GitHub](#) and [BitBucket](#) repositories. From the GitHub repository, users can download the SQLite database used to power the website. Users will also find JSON and CSV (i.e. Excel) representations of the database, which are auto-generated from the SQLite database using Python scripts.

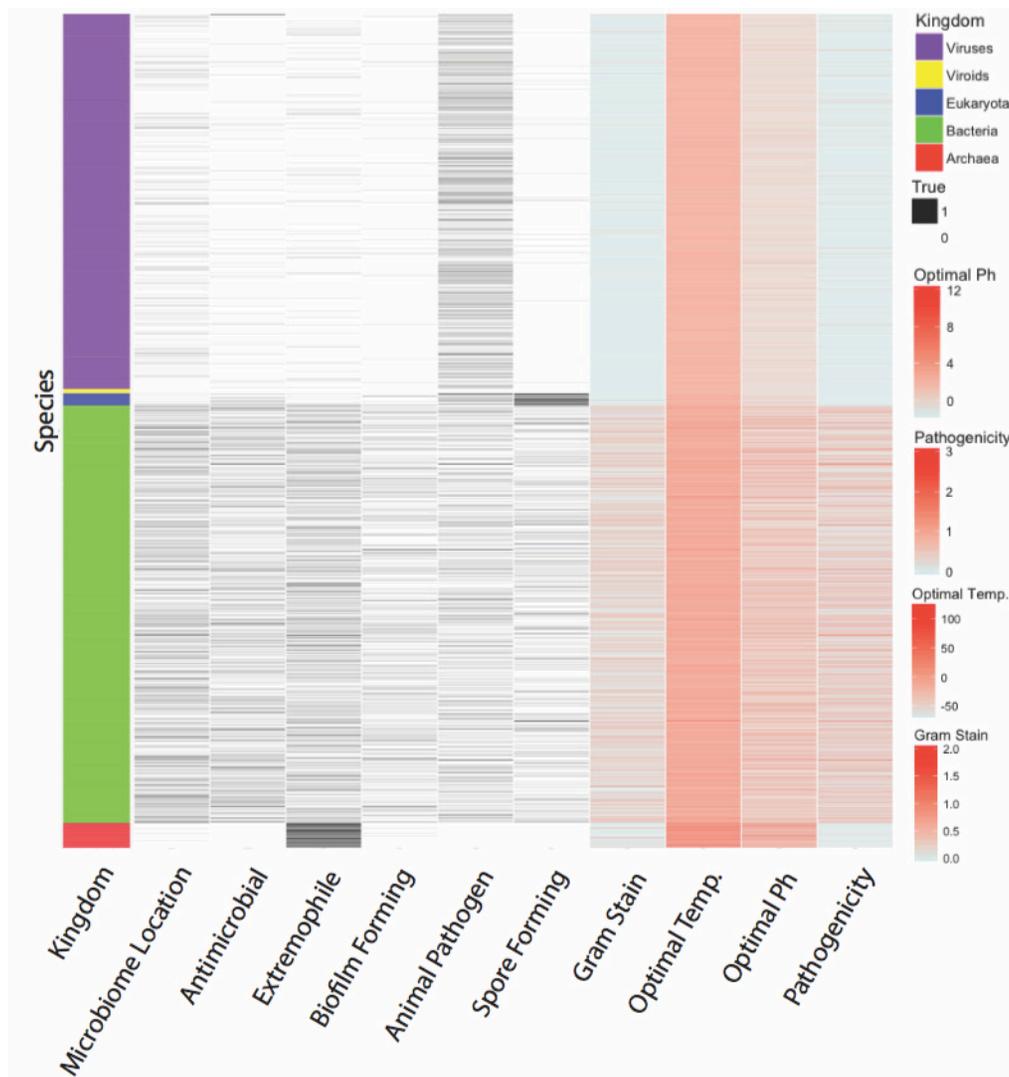


Figure 1. Microbe Directory heatmap. Annotation types (x-axis) are represented across the online database and the numbers of each category (y-axis, left side) are shown, including Viroids (purple), Viruses (yellow), Eukaryotes (blue), Prokaryotes (green), and Fungi (red). The scale for each of the types of metadata (right) are also shown for binary classifications (black, white) and quantitative traits (red scales). Heatmap was constructed using R (version 3) and Illustrator.

Since the Microbe Directory is meant to grow and expand over time, researchers wanting to make more substantial contributions can suggest changes to the database through our GitHub page. The requested changes will be merged as appropriate and could be incorporated into future releases. Moreover, there is a tutorial on the GitHub repository that shows users how they can use the JSON version of the database given a MetaPhlan2 output file. Finally, the website used to power the web interface can also be accessed and modified through a separate BitBucket repository, which can also be accessed through the main website.

The Microbe Directory was designed to help researchers in the microbiome and metagenomics fields to learn more about the organisms they are identifying through their bioinformatics

analyses. While this is only version 1.0 of the Microbe Directory, it is readily able to incorporate any contributions to the database to expand the microbial features included in our inventory. For more information on how to contribute to the project visit <https://microbe.directory/>.

Data availability

The web interface for the Microbe Directory can be found at <https://microbe.directory/>

The database and other files can also be found on the GitHub repository here: <https://github.com/microbe-directory/microbe-directory> and the BitBucket repository here: <https://bitbucket.org/account/signin/?next=/microbedb/microbedb>. Note: BitBucket

requires a login, but account generation is free and there are no restrictions for signing up.

Archived code as at time of publication:

GitHub: <https://doi.org/10.5281/zenodo.1069858>¹²

Bitbucket: <https://doi.org/10.5281/zenodo.1069860>¹³

License: MIT

Competing interests

No competing interests were disclosed.

Grant information

We would like to thank the Epigenomics Core Facility at Weill Cornell Medicine, funding from the Irma T. Hirsch and Monique Weill-Caulier Charitable Trusts, Bert L and N Kuggie Vallee Foundation, the WorldQuant Foundation, The Pershing Square Sohn Cancer Research Alliance, NASA (NNX14AH50G, NNX-17AB26G), the National Institutes of Health (R25EB020393, R01NS076465, R21AI129851), the Bill and Melinda Gates Foundation (OPP1151054), and the Alfred P. Sloan Foundation (G-2015-13964).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We would like to thank all the student-researchers who helped curate the data for the Microbe Directory without which this project would never be possible (student names are ordered based on their

contribution to the database): Sophie Dornbaum (Weill Cornell Medicine), Rabab Shaddoud (CUNY Hunter College), Sadia Chowdhury (CUNY Hunter College), Sarah Chebli (CUNY Hunter College), Christopher Chiang (CUNY Hunter College), Ellen Koag (CUNY Hunter College), Sophia Tam (CUNY Hunter College), Christopher Campbell (CUNY Hunter College), Timothy Lau (CUNY Hunter College), Camille Derderian (CUNY Hunter College), Elyas Amin (CUNY Hunter College), Nicole Rakhmanova (CUNY Hunter College), Amina Durakovic (CUNY Hunter College), Jereen Chowdhury (CUNY Hunter College), Catherine Ng (CUNY Hunter College), Jasmine Wong (CUNY Hunter College), Phuong Vo (CUNY Hunter College), Calvin Herman (CUNY Hunter College), Silva Baburyan (CUNY Hunter College), Kevin Londono (CUNY Hunter College), Julianna Romeo (CUNY Hunter College), Leah Katz (CUNY Hunter College), Valentina Bedoya (CUNY Hunter College), Juan Cambeiro (CUNY Hunter College), Amzad Chowdhury (CUNY Hunter College), Rangon Islam (CUNY Hunter College), Bibi Begum (CUNY Hunter College), Frances Chung (CUNY Hunter College), Mimi Fellner (CUNY Hunter College), Phillip Ye (CUNY Hunter College), Madeleine Winter (Poly Prep High School), Raghav Pant (Millburn High School), Kriti Devasenapathy (California Institute of Technology), Halime Sena Bastug (Istanbul University Cerrahpasa Medical Faculty), Chou Chou (Weill Cornell Medicine), Jasmine Sharron (Columbia Secondary School), Laolu Ogunnaike (Johns Hopkins University), Alina Sheikh (Adelphi University), Carol Apai (Rutgers New Jersey Medical School), Salama Chaker (Weill Cornell Medicine Qatar), Caleb Gordon (Bowdoin College), Michael Pineda (Arizona State), Dara Pierre (CUNY John Jay), Scott Kulm (Weill Cornell Medicine), Ike Lewis (Weill Cornell Medicine), Mustafa Hakyemezoglu (Weill Cornell Medicine).

Supplementary material

Supplementary File 1: Microbial Annotations - Tutorial and guidelines for student researchers, and useful links and tips.

[Click here to access the data.](#)

Supplementary File 2: Annotations_Automator.py – Python script used for automated research.

[Click here to access the data.](#)

References

1. The NIH HMP Working Group, Peterson J, Garges S, *et al.*: **The NIH Human Microbiome Project.** *Genome Res.* 2009; **19**(12): 2317–2323. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Gilbert JA, Jansson JK, Knight R: **The Earth Microbiome project: successes and aspirations.** *BMC Biol.* 2014; **12**: 69. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Thompson LR, Sanders JG, McDonald D, *et al.*: **A communal catalogue reveals Earth's multiscale microbial diversity.** *Nature.* 2017; **551**(7681): 457–463. [PubMed Abstract](#) | [Publisher Full Text](#)
4. Tighe S, Afshinnikoo E, Rock TM, *et al.*: **Genomic methods and microbiological technologies for profiling novel and extreme environments for the Extreme Microbiome Project (XMP).** *J Biomol Tech.* 2017; **28**(1): 31–39. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Yooseph S, Andrews-Pfannkoch C, Tenney A, *et al.*: **A metagenomic framework**

- for the study of airborne microbial communities. *PLoS One*. 2013; **8**(12): e81862. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Afshinnekoo E, Meydan C, Chowdhury S, *et al.*: **Geospatial resolution of human and bacterial diversity with city-scale metagenomics.** *Cell Syst*. 2015; **1**(1): 72–87. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 7. MetaSUB International Consortium: **The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report.** *Microbiome*. 2016; **4**(1): 24. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 8. Truong DT, Franzosa EA, Tickle TL, *et al.*: **MetaPhlan2 for enhanced metagenomic taxonomic profiling.** *Nat Methods*. 2015; **12**(10): 902–903. [PubMed Abstract](#) | [Publisher Full Text](#)
 9. McIntyre AB, Ounit R, Afshinnekoo E, *et al.*: **Comprehensive benchmarking and ensemble approaches for metagenomic classifiers.** *Genome Biol*. 2017; **18**(1): 182. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 10. Pasolli E, Schiffer L, Manghi P, *et al.*: **Accessible, curated metagenomic data through ExperimentHub.** *Nat Methods*. 2017; **14**(11): 1023–1024. [PubMed Abstract](#) | [Publisher Full Text](#)
 11. Van Belkum A: **COGEM Research Report: Classification of Bacterial Pathogens.** Department of Medical Microbiology and Infectious Diseases. The Netherlands, 2011.
 12. Shaaban H, Westfall DA, Mohammad R, *et al.*: **Microbe Directory Data v1.0.0 (Version v1.0.0) [Data set].** *Zenodo*. 2017. [Data Source](#)
 13. Shaaban H, Westfall DA, Mohammad R, *et al.*: **Microbe Directory Website v1.0.0 (Version v1.0.0).** *Zenodo*. 2017. [Data Source](#)

Open Peer Review

Current Referee Status:



Version 1

Referee Report 22 March 2018

doi:10.21956/gatesopenres.13832.r26308



Elisabeth M. Bik 

uBiome, San Francisco, CA, USA

Shaaban *et al.* describe The Microbe Directory, a database with more than 7,500 microbial species. This is a great initiative, in which a group of academic researchers, helped by a team of (under)graduate students have annotated bacteria, archaea, viruses and eukaryotic microbes taken from the MetaPhlAn2 database, with respect to pathogenicity, growth characteristics, and presence in the human microbiome. The paper is well written, and the initiative is very welcome. Although the initial list of fields is small, this database has the potential to grow, and there is an option for registered users to add missing data.

Comments on the manuscript:

1. A typo in the Introduction: "taxonomoic"
2. In the Methods, section "MetaPhlAn2 list of species", it reads "7-level (kingdom to strain)". However, based on the MetaPhlAn2 documentation, this should read "kingdom to species" (1, Kingdom; 2, phylum; 3, class; 4, order; 5, family; 6, genus; 7, species).
3. Figure 1.
 1. There is a discrepancy to the color codes described in the legend text and those in the key on the top right of the figure. Viruses, viroids have opposite colors, and the "prokaryotes" and fungi are not mentioned in the key.
 2. "Prokaryotes" is not a good term, as it defines the 2 groups by something they do not have, and suggests a common ancestor between archaea and bacteria. See e.g. Norm Pace's essay¹.
 3. What is meant by "Microbiome Location"? It has a binary value, suggesting that maybe "Found in Human Microbiome" (as used in Table 1) might be a better description. The "Location" suggests that this field stores which anatomical site this species has been found, which would also be a nice field to have, but not what is meant here.
 4. Similar question for "Antimicrobial" - does a yes mean that it is sensitive or resistant? Or that antimicrobial properties are known? Or that it makes an antimicrobial? Table 1 provides the answer to the question, but it might be worth addressing this here as well.
 5. "Optimal Ph" should be "Optimal pH".
 6. The order of the last 4 categories is different between the labels under the heatmap and the key on the right.
4. A possible and useful addition to the paper would be to describe some potential fields that could be added to the database. In its current form, the useability of the database is very limited, and it would probably be faster to just look up the information in e.g. Wikipedia. But the strength of this database is that it can grow, both in number of entries, as well as number of fields. Some

suggestions would be: a link to the draft genome of the organism, number of chromosomes, linear/circular chromosome, RNA/DNA virus.

5. The paper could also address that some information that appears to be simple at first glance, such as pathogenicity, might not be simple at all. For example, *Escherichia coli* and *Clostridium difficile* can both be a peaceful member of the human gut microbiota, or a human pathogen, depending on the presence of toxin genes. Herpes virus infections are so common in humans, and usually latent, that one could argue that it might be considered part of the human microbiome. The ability to form biofilms might be also more complicated than just a simple yes/no. The paper would be stronger if it acknowledges the difficulties of capturing these subtleties into simple binary answers.

Comments on the database:

1. Of course, this is just a first version, and the database will hopefully grow quickly, but the current data felt very sparse. For example, and as also pointed out by other reviewers, pH/temperature information was missing for well-studied microbes such as *Salmonella enterica*, *Agrobacterium tumefaciens*, *Candida albicans*, *Schizosaccharomyces pombe*, or *Yersinia pestis*. For others, there was an entry present in the database but all fields were empty (e.g. *Rhinovirus A*, *Bacillus cereus thuringiensis*).
2. There are important classification errors such as:
 1. *Yersinia pestis*, which causes disease in rodents, is not listed as an animal pathogen.
 2. *Magnaporthe oryzae*, causative agent of one of the most destructive diseases in rice, was not listed as a plant pathogen.
 3. *Candida albicans* was listed as not susceptible for antimicrobials, with a reference from 1999.
 4. Influenza B virus is classified as biofilm forming based on a paper that shows that Influenza A virus can disperse *Streptococcus pneumoniae* biofilms.
 5. *Agrobacterium tumefaciens*, a well-studied plant pathogen, is listed as an animal pathogen, although it does not infect animals in nature, only under laboratory conditions.
 6. Human herpesvirus 4 (Epstein-Barr virus, but that search did not bring up any results), is not listed as a human pathogen.
3. The “Found in the human microbiome” category is defined as “Microbes that live anywhere in the human body and are not pathogenic to humans (i.e. capable of causing human disease) No = 0, Yes = 1”. However, both *Escherichia coli* and *Clostridium difficile*, which can be both a pathogen as well as a symbiotic member of the microbiome, are classified as a “yes”. So maybe the definition of this should be refined and the part about not being pathogenic to humans should be taken out?
4. Links: Fields with data are marked with an “i”, which will lead to the source. Some of the i’s are yellow, while others are black/white. The yellow i’s lead to a URL that is not a hyperlink, while the white i’s are hyperlinks. As one of the other reviewers pointed out, it would be nice if all links would be hyperlinks. Also, in these links, it was very noticeable that the database was compiled of contributions by many different users, who all had their own specific way of adding links. In some cases, the yellow i’s give a citation but they are not very helpful. Examples:
 1. *Akkermansia muciniphila*: all information links give “Everard, Amandine, et al”, without doi, year, or working link.
 2. *Cryptococcus neoformans*: all i’s lead to “Todd W., Larimer, Frank W., Lippmeier, J. Casey, Lucas, Susan, Medina” - no link, year, doi. Which paper is that? I could probably find it, but that defeats the purpose of having a reference database.
 3. *Magnaporthe oryzae* Abx susceptibility is listed as an unhelpful “Choi J et al”. A Pubmed search for that author returns 19067 papers.
 4. *Escherichia coli*: the field for the optimal pH leads to a biotech company that sells culture media, but the link is broken.

5. All i's for *Candida albicans* refer to Staab, J. F. (1999)², which albeit a paper about *Candida*, is an older paper about a specific protein, not a general review paper.
 6. The "Pathogenicity" appears to always use the COGEM 2011 list as a source, but it is annotated in many different ways. It seems that each of the (under)graduate students used a different description for this field. Sometimes it is a working link to the COGEM document, but in other cases it is a cryptic and not-helpful pop-up text such as "CGM2011", "CGM PDF", "COGEM", "CGM 2011-07 Bijlage I Algemene text", without a hyperlink.
 7. *Pseudomonas aeruginosa*: the field Plant Pathogen has a link to a personal file (<file:///Users/catherineng/Downloads/54028.pdf>) that does not work for other users.
 8. *Haloferax denitrificans*: another personal link: <file:///C:/Users/Maddie/Downloads/35960.pdf>
 9. Human herpesvirus 4: None of the yellow I's appear to show any text.
 10. Some of the listed sources of information consists of 2 URLs separated by a space; it is tricky to correctly copy/paste these into a browser. E.g. *Methanobrevibacter smithii*:
https://microbewiki.kenyon.edu/index.php/Methanobrevibacter_Smithii
<http://bacmap.wishartlab.com/organisms/525>
 11. The i under *Bacillus anthracis* lists about 10 URLs but most are private search terms, so they are not useful for anyone. The complete list is:
http://medschool.creighton.edu/fileadmin/user/medicine/MMI/Files/Bacteria_Table.pdf ;
<http://www.life.umd.edu/classroom/bsci424/PathogenDescriptions/PathogenList.htm#D> ;
<https://www.patricbrc.org/portal/portal/patric/SpecialtyGeneList?cType=genome&cld=12978> ;
http://ardb.cbcb.umd.edu/cgi/search.cgi?db=R&term=Y_P_001373621 ;
http://ardb.cbcb.umd.edu/cgi/search.cgi?db=R&term=Z_P_02395450 ;
http://ardb.cbcb.umd.edu/cgi/search.cgi?db=R&term=Z_P_02391336 ;
http://ardb.cbcb.umd.edu/cgi/search.cgi?db=R&term=Z_P_03108029 ;
<https://microbewiki.kenyon.edu/index.php/Bacillus> ;
https://en.wikipedia.org/wiki/Bacillus_anthraxis ;
<https://gold.jgi.doe.gov/organisms?id=3251>
5. Layout:
1. I found the use of upper and lower case for the microorganism names a bit distracting. Bacteria are listed as completely lower case in the dark top bar, but in upper case in the bottom part. This might be great from a designer point of view, but it is not how most of us are used to write bacterial names.
 2. When browsing taxonomically, viruses are listed in non-alphabetical order, making it hard to find the correct entry. E.g. Picornaviridae-Enterovirus or genus *Bacillus* both lead to such a list.
 3. In "Optimal Temperature" there is a "P" missing
 4. The name of the category called "Microbiome location " suggests this field would contain a location, such as gut/mouth/skin), while it is a yes/no field.
6. Suggestion for a future additions:
1. a short line describing what the organism is known for.
 2. Additional categories: chromosome information (linear, circular, how many, number of ribosomal operon copies, RNA/DNA for viruses), use in food industry (brewing, bread making, probiotic)

References

1. Pace NR: Time for a change. *Nature*. 2006; **441** (7091): 289 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Staab JF, Bradway SD, Fidel PL, Sundstrom P: Adhesive and mammalian transglutaminase substrate properties of *Candida albicans* Hwp1. *Science*. 1999; **283** (5407): 1535-8 [PubMed Abstract](#)

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Partly

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: I work at uBiome, a microbial sequencing company.

Referee Expertise: Human microbiome analysis, biotech industry

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 15 March 2018

doi:[10.21956/gatesopenres.13832.r26309](https://doi.org/10.21956/gatesopenres.13832.r26309)



Nicole M. Vega 

Biology Department, Emory University, Atlanta, GA, USA

In this manuscript, the authors describe the creation and construction of the Microbe Directory, a resource for profiling and annotating species after large-scale metagenomic taxonomic analyses.

I very much like the idea of the Microbe Directory and think that this could be a valuable resource for the field. The manuscript describing the Directory's construction and curation to date was clear and understandable.

I think that the ability of researchers to add information to the database directly is a great feature. Is there also a plan and/or schedule for incorporating database updates from the sources described in the paper?

I did not download the database, but I did try the web interface and found it fairly intuitive to use. The Browse function was a little odd - it would be helpful if the options for each clade were presented alphabetically or in some other obvious order.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.**Referee Expertise:** Microbiology, microbial ecology**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Referee Report 15 March 2018

doi:[10.21956/gatesopenres.13832.r26238](https://doi.org/10.21956/gatesopenres.13832.r26238)**James E. McDonald**

School of Biological Sciences, Bangor University, Bangor, UK

Concept:

The microbe directory is an excellent concept and aims to provide phenotypic and ecological profiles of approx. 7500 microbial species represented in the MetaPhlAn2 database. Although some information is present in other repositories, the Microbe Directory aggregates information on the functional, biological or medical traits of these organisms into a single source where the profiles may be further expanded to represent a useful resource to better interpret the functional and ecological properties of taxonomic data. If the directory continues to grow and expand with additional information, it would be a fantastic and heavily-utilised resource for the wider community. In particular, integration of data for bacteria, archaea, viruses, fungi, and protozoa in the same database is a positive strategy.

An important feature of the Microbe Directory is that while there is always the potential for human error in manually curated databases, scientists can generate an account and submit edits and changes to the information hosted in the database. I hope that the wider scientific community engages with and contributes to the directory in order to enable it to reach its potential as an important resource for microbiologists.

Manuscript:

Abstract. The abstract focusses heavily on the application of the directory 'downstream of large-scale metagenomic taxonomic analysis' and 'designed to serve as a resource for researchers conducting metagenomic analysis', but perhaps this is too narrow a focus on the utility of the directory. I can see several other uses for the directory in other areas of microbiology; to inform/validate the potential phenotypic and ecological properties of a microbial isolate, or as an information source on a specific microorganism for an undergraduate student after a lab class, for example. Maybe it's worth re-wording this to broaden the potential for wider adoption of the resource.

Average values. Table 1 describes the microbial features currently listed in the directory. However, in instances where more than one optimal temperature and pH could be found for different strains of a

species, an average value has been taken. This would mask the range of optimal temperatures across the strains, which is useful information if you are using the resource to find out the best temperature(s) to grow a species at. Could the range of temperature recorded not also be provided as an additional source of information?

Website and sources of information:

The website looks good and was generally easy to navigate.

Information. I performed a few searches for microorganisms that we work on, some of which are not well-characterised at present, and was not surprised to see that for most of these many of the categories were blank. However, I then looked at some very well-studied organisms and also found lots of blank categories. For example, only 3/11 categories were complete for *Bacillus subtilis*, when a quick google search reveals primary research articles that provide information on several of these categories (e.g. that it is a spore former). *E. coli* also has no data for biofilm formation, which can again be verified with a quick google search. Going forward, additional buy-in will therefore been needed to ensure that the information is complete as possible.

For some species, the information links didn't work (I got an 'error 404' code) which made it difficult to find the source of information, but others worked fine.

Where the links did work, many of the sources of information were webpages (e.g. wiki pages) that did refer to primary literature that could be accessed to verify the information. However, in my view, if it was possible to provide links to more than one reference in the primary literature, and to allow others to add links to primary research articles, you could very quickly generate a set of primary literature that described those key attributes, which would be very beneficial as a reference source and for validation of the information.

Some links to information about microbial species took me to website articles that although were apparently peer-reviewed, had collated information from primary research articles. However, if it is possible to incorporate them, direct links to the research articles themselves would in my view be more useful.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Microbial ecology

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 12 January 2018

doi:[10.21956/gatesopenres.13832.r26191](https://doi.org/10.21956/gatesopenres.13832.r26191)



David A. Coil

Genome Center, University of California, Davis, Davis, CA, USA

I love the idea behind “The Microbe Directory”. I think this information will be of great value and I really like the way it was generated with the help of students. With the ability to expand the database, I think this could become an important resource. In particular I’m curious if the authors have considered working with the folks at Traitair? They have a really useful tool but their underlying phenotypic data is lacking and is heavily biased towards human pathogens in a way that the Microbe Directory appears not to be.

However, my attempts to use the website were somewhat frustrating. It’s not clear to me in a review of this nature whether I should be reviewing just the paper (basically fine) or the software itself (needs work before going live).

Paper

The paper is well-written and clear. I only had very minor suggestions (harder without line numbers!)

The phrase starting with “these features were considered” in the “Building the microbe directory section needs grammatical revision.

In the same section the sentence “These peer-reviewed resources and databases...” is a bit misleading since most of the listed resources are not peer-reviewed.

Not sure the “(RegEx)” abbreviation is useful since it’s never used again.

“The edits are then put in a queue to be later reviewed by the Microbe Directory team”. I would like to see a bit more about what the criteria for review are. Will they just be checking for spam or will they actually verify information using the reference(s)?

Website

The site is very clean and easy to navigate. However, when I attempted to do things I encountered some snags.

Firstly, I went and looked at some microbes. The first thing I noticed is that reference links aren’t clickable links, they are just html which is a bit off-putting and requires careful copying and pasting. But in a case where there are multiple links it becomes a mess (see [screenshot](#)). It seems like some better way to parse links is required, and having them be clickable would be awesome. The first one I tried also led to a 404 error, is there some way that links could be automatically checked?

So after I created an account I clicked on “Contribute” which took me to the login page, after logging in, it

dropped me back at the main page which seems not ideal... I then had to navigate back to the microbe I was interested in. Then after clicking on "Contribute", I'm faced with some fields... the first of which is the Microbe ID... but it doesn't say that anywhere, there's just a number there. A bit confusing. Perhaps this field could be labeled?

Then I wanted to add something about Gram staining... but there's no key for the "Values" field (see [screenshot](#)). I had to open a new window and pull up another organisms to know that "1" is what I wanted for Gram-positive. Is it possible to display the key for the values on the contribute page? Or do have a small key appear for the selected attribute? Some way for someone to have access to the key within the "Contribute" page.

I was surprised to see an entry like "*Porphyrobacter* sp AAP82". Is there a rationale for including isolates that don't have a species level identification? Not sure how useful this sort of thing is for the stated purpose of the database. For example in this case there are only two pieces of data, the COGEM listing (which since the reference "link" just says "COGEM" can't be verified) and the Gram stain field which links to the listing for that genus at bacterio.net which actually doesn't contain information about Gram staining and anyway wouldn't have a listing for an isolate like this that doesn't have a species name. How does something like this end up in the database?

In a similar vein, there are microbes in the database that have no information whatsoever, is this expected?

I really wanted to try adding a new species to the database, but couldn't find any way to do so. The paper sort of implies that this is possible but I didn't see any such option on the website? I then went to the "Contact" page figuring that I would send a ping with this question. But there's no general contact address for the project? Seems like that might be useful? Wasn't sure which specific person would be appropriate for this question.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: Our lab participated in the worldwide MetaSUB project which was run by Chris Mason and his lab. I wasn't the primary point of contact for the project, but I supervised the person that was. Our involvement was collecting samples and sending them to his lab (as did dozens of other labs). In addition, I once sent him some bacterial DNA that might or might not someday be part of a future publication.

Referee Expertise: Microbiology, Microbial Ecology, Bacterial Genomics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 04 Feb 2018

Heba Shaaban, Weill Cornell Medical Center, USA

Thank you for your review, Dr. Coil.

Paper:

The edits you suggested regarding the paper will be published on version 2 of the manuscript.

As for the edits that contributors will be making to the site, we will review them one-by-one to confirm that the sources/citations submitted are reliable and that the data inputted is verified by the reference. We have an administrative view that our team uses to accept/reject edits. We will expand on this in the paper, as suggested.

Website:

We are sorry that you had a difficult experience using our website. The links are now clickable if they are validated, and you should be redirected to the link right away. If more than one link was cited, you can click on the icon to see all the links. For now, copying and pasting the links (if multiple) is the only way to view them. We eventually want to change this, but due to the non-standard format of some of the citations, this is not currently possible. We will need to manually re-curate the citations in order to provide this functionality, and we are planning to do so in the near future. As for the link you clicked on being expired, the project has been ongoing for about two years now and some of the links are indeed broken. We plan to add timestamps to each citation, but this is currently not implemented. If a link is broken, it is possible for Microbe Directory users to manually submit edits using the "Contribute" interface.

We also programmed the site to redirect you to the microbe you were attempting to edit after you create an account. Thank you for bringing that to our attention. The microbe ID was actually an in-house cataloging number for admins and we understand why it might be confusing. It is just meant to make it easier to share species on social media without having to type in the species name. Also, the key for values will now appear on the contribute page. We also created a drop-down menu for each value, which should make it easier to edit.

As for including isolates that don't have a species-level identification, we made the decision to be compatible with MetaPhlan2, so researchers using this popular metagenomics analysis tool would have their code work "out-of-the-box." If in the future, researchers decide to work with these strains or species, they will already be cataloged in the database. We are aware that there are microbes for which there is no information. We wanted to include these to be conformant to MetaPhlan2 and also to provide a scaffold for users to fill in additional information.

Additionally, users can now add new species to the database by using the link on the "Contribute" page. We also updated the contact page to include our role descriptions in the project, so that we may be contacted accordingly. We also improved our "Help" page to include more instructions on contributing and to address some troubleshooting.

Lastly, this is just version 1 of the Microbe Directory. As the project expands, we plan to make changes to the site format and configuration as necessary in addition to maintaining proper quality-control. As you had mentioned with Traitar, we want to collaborate with organizations that have their own databases and incorporate them into the Microbe Directory. We think v1 of the Microbe Directory provides a scaffold for expansion, and we really want to see this project grow over time.

Thank you for your time and efforts in reviewing our project.

Competing Interests: No competing interests were disclosed.

Discuss this Article

Version 1

Reader Comment 15 Feb 2018

Qunfeng Dong, at Loyola University Chicago Medical School, USA

Very interesting work! I think that this resource can be potentially very helpful for metagenomics research. I applaud the research team's enormous efforts to organize students for this tedious yet important task. I really hope that the project can be sustained after the publication. Best wishes!

Competing Interests: None

Author Response 07 Feb 2018

Christopher Mason, Weill Cornell Medical Center, USA

We agree and are discussing this now with the editors.

Competing Interests: No competing interests were disclosed.

Reader Comment 05 Jan 2018

Daniel McDonald, University of California, San Diego, USA

The article states that "Each student-researcher independently worked 4–5 hours per week to curate parameters for 10 species per week, for a total of 20 weeks." Despite 80-100 hours of intellectual work, they only received an acknowledgement. That's quite a bit of time spent to construct the contribution to the field described by this manuscript. I hope for the authors reconsider their choice so that these students can receive credit appropriate for CVs (especially for undergraduates looking at graduate school).

Competing Interests: No competing interests were disclosed.

