mockrobiota: a public resource for microbiome bioinformatics benchmarking

Nicholas A. Bokulich[1], Jai Ram Rideout[1], William G. Mercurio[1], Benjamin Wolfe[2], Corinne F. Maurice[3], Rachel J. Dutton[4], Peter J. Turnbaugh[5], Rob Knight[6,7,8], J. Gregory Caporaso[1,9#]

[1]Center for Microbial Genetics and Genomics, Northern Arizona University, Flagstaff, AZ, USA

[2]Department of Biology, Tufts University, Medford, MA, USA

[3]Department of Microbiology & Immunology Department, Microbiome and Disease Tolerance Centre, McGill University, Montreal, Quebec, Canada.

[4]Division of Biological Sciences, University of California, San Diego, CA, USA

[5]Department of Microbiology and Immunology, G.W. Hooper Foundation, University of California San Francisco, 513 Parnassus Ave, San Francisco, CA, USA

[6]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA

[7]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA

[8]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA

[9]Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA

#Corresponding author

Gregory Caporaso

Department of Biological Sciences

1350 S Knoles Drive

Building 56, 3rd Floor

Northern Arizona University

Flagstaff, AZ, USA

(303) 523-5485

(303) 523-4015 (fax)

Email: gregcaporaso@gmail.com

Abstract

Mock communities are an important tool for validating, optimizing, and comparing bioinformatics methods for microbial community analysis. We present mockrobiota, a public resource for sharing, validating, and documenting mock community data resources, available at https://github.com/caporaso-lab/mockrobiota. The materials contained in mockrobiota include dataset and sample metadata, expected composition data, which are annotated based on one or more reference taxonomies, links to raw data (e.g., raw sequence data) for each MC dataset, and optional reference sequences for mock community members. mockrobiota does not supply physical sample materials directly, but the dataset metadata included for each mock community indicate whether physical sample materials are available (and associated contact information). At the time of this writing, mockrobiota contains 11 mock community datasets with known species compositions (including bacterial, archaeal, and eukaryotic mock communities), analyzed by high-throughput marker-gene sequencing. The availability of standard, public mock community data will facilitate ongoing methods optimizations; comparisons across studies that share source data; greater transparency and access; and eliminate redundancy. This dynamic resource is intended to expand and evolve to meet the changing needs of the ‘omics community.

Introduction

An important step in the development of bioinformatics methods is the identification and acquisition of useful test datasets. For microbiome bioinformatics tools, test datasets frequently take the form of simulated data, data from natural microbial communities that is considered to

be well-understood, or “mock community” data. Each of these data types have their own pros and cons. With simulated data, a model is developed to computationally generate artificial data, e.g., marker-gene sequence reads. Since the developer of the simulated data has complete control over the model, the true values of the optimization criteria (e.g., the relative abundance of different species in a sample) are known with certainty. While this is very useful, optimizing a method on simulated data can result in fitting the method to work well on the results of the model used for simulation. This can be problematic if the model is not a good representation of reality. Natural microbial communities are on the opposite end of the spectrum. Assumptions are not made in generating the data, but are made about the true values of optimization criteria because the “right answer” isn’t necessarily known in advance. Mock microbial communities attempt to provide a balance between these two types of test data for microbiome methods benchmarking by providing data with a known biological composition and which is technologically relevant (i.e., represents actual experimental observations). It is important to stress that none of these approaches is perfect, and combining them (e.g., evaluating a bioinformatics method on both mock and natural communities) is common and likely to provide insight beyond evaluations using either data type on its own.

A mock community (MC) is a defined mixture of known microbial strains. To make a MC, axenic cultures are deliberately combined at precise ratios such that the species composition is known (a mock microbiome). If the genomes of these strains are sequenced, the expected collective gene content can be inferred, yielding a mock metagenome. This mixture is then processed as if it were a natural community sample, including DNA extraction, amplification of marker genes such as 16S rRNA (if applicable), and sequencing. This allows for generation of real sequence data (nullifying concerns about assumptions made during generation of simulated data sets),

and provides known optimization criteria (though in practice some uncertainty is still present). One limitation of mock communities, however, is that they often are composed of few taxa relative to natural microbial communities, so overfitting of methods to unrealistic conditions is still possible, emphasizing the importance of employing different types of test data. MCs have been widely used in microbiome methods development, including development of sequencing protocols (1, 2), validating sequence quality control (3-5), and evaluating and comparing bioinformatics methods for marker-gene (6, 7) and metagenomics sequencing (8).

We consider MCs to be composed of three parts: the physical sample materials (microbial cells, DNA, RNA, etc); expected composition data (e.g., taxonomic annotations and abundance); and raw data (e.g., raw sequence data obtained from marker-gene sequencing of the MC). MCs are a valuable community resource, and public sharing of standardized MC data will facilitate: 1) ongoing methods improvements for the 'omics community; 2) direct comparisons among studies that share source data; 3) greater transparency and access to source data; and 4) eliminate redundancy, as developers can bypass the time-consuming task of generating new MCs if appropriate MC datasets already exist. The use of multiple MCs is advisable to generalize method optimization across different conditions (e.g., taxonomic kingdoms, marker genes) and to avoid overfitting, underlining the value of shared, public MCs to accelerate bioinformatics methods development.

Results and Discussion

We present mockrobiota, a public resource for sharing, validating, and documenting MC resources. mockrobiota is open source and hosted on GitHub, an online software revision

control and collaboration tool. The materials contained in mockrobiota include dataset and sample metadata, expected composition data, which are annotated based on one or more reference taxonomies, links to raw data (e.g., raw sequence data) for each MC dataset, and optional reference sequences for mock community members. mockrobiota does not supply physical sample materials directly, but the dataset metadata included for each MC indicate whether physical sample materials are available from the contributor. If so, relevant contact information is listed for requesting that material directly from the contributor. Due to storage limits, raw sequence data are not stored in mockrobiota itself, but rather in other public resources such as FTP servers or the QIITA database (https://qiita.ucsd.edu/) and linked directly from the GitHub repository. These links to raw data are automatically validated regularly, as described below.

At the time of this writing, mockrobiota contains 11 MCs with known species composition (including bacterial, archaeal, and eukaryotic MCs), analyzed by high-throughput marker-gene sequencing (Table 1). Known taxonomies of these samples are annotated with Greengenes (9) and Silva (10) reference taxonomies for bacterial/archaeal samples, Silva for eukaryote samples, and UNITE (11) for fungal-only eukaryote samples. Translating from a MC developer's taxonomic description of a sample to relevant taxonomic database annotations can be time-consuming and error-prone. The availability of these annotations in mockrobiota will therefore save time and increase consistency across studies that use these data. MCs can be utilized in a few simple steps (Figure 1).

Attributes of each MC are summarized in dataset metadata tables viewable in the mockrobiota repository, facilitating navigation and selection of the MCs that best fit users' needs. From these

tables, users can directly access links for downloading MC data, metadata, and auxiliary files. The repository also contains guidelines for formatting and contributing new resources to mockrobiota (https://github.com/caporaso-lab/mockrobiota/blob/master/CONTRIBUTING.md).. All core MC resources are available in common file formats to facilitate universal access without any specific software requirements. This allows end users to “plug and play” their MCs of choice into analysis pipelines without software bottlenecks.

Importantly, mockrobiota makes use of Travis CI (https://travis-ci.org/) for continuous integration testing to ensure data integrity. Any time a change is proposed to any of the mockrobiota files (e.g., modification of an existing data set, or the addition of a new MC), a series of tests are run to validate all of the data. This includes confirming that raw data links are valid and accessible, that files are formatted correctly, and that expected taxonomic relative abundances in each sample sum to 100%. Together, these ensure that users can always access the data in mockrobiota (i.e., links are not outdated) and that all mock community data are reliable and available in consistent formats, facilitating analyses that involve multiple MCs. This model of using software testing approaches to validate community data resources would be very useful to generally adopt in bioinformatics, and as illustrated here is now simple to implement with free continuous integration testing systems.

Hosting mockrobiota on GitHub provides an additional major benefit: the data are not static. This resource will grow and evolve to meet the needs of the ‘omics community as more MCs are contributed, and to conform to changes in related resources, such as sequence reference databases and taxonomic annotations. Finally, hosting on GitHub invites community involvement to contribute, update, revise, and evaluate MC resources.

MCs provide many benefits for bioinformatics method benchmarking, complementing the use of other test data (e.g., natural and simulated communities). We anticipate that a public MC database will eliminate redundant effort, improve consistency across studies that use the same MCs, and thereby facilitate methods advances for the benefit of the entire microbiome research community. We hope that mockrobiota will fill this gap, and that community members will contribute to the growth and development of this resource.

Methods

Data availability

Links to raw data, database and sample metadata, expected composition data, and other useful resources are hosted in a public GitHub repository, which can be found at: https://github.com/caporaso-lab/mockrobiota.

mockrobiota is a data resource, and does not provide physical samples (e.g., DNA, RNA, cell mixtures) of MCs. However, contributors are encouraged to share physical samples of their mock communities as supplies permit. The dataset metadata included for each MC indicate whether physical sample materials are available to be shared, and if so, list relevant contact information for requesting that material directly from the contributor.

Expected Observation Data Generation

Expected observation data, representing the known composition of a MC, are provided in mockrobiota in two forms: “source” data and expected composition (taxonomy or gene

annotation) data. “Source” data provide a record of the original inputs to the MC as a list of microbial strains and their relative abundances. Ideally, a strain ID should be provided to identify a retrievable source strain. These data are generally created by the developer of the mock community, and taxonomic groups are not necessarily annotated with respect to any specific taxonomic reference database. “Expected composition” data represent the known composition of the MC (e.g., taxonomies or KEGG pathways), annotated according to a specific reference database. Compilation of expected composition data is not a trivial task, and requires careful review of database annotations to ensure that accurate annotations are applied to source data. Examples of source data and expected composition data are given in Tables 2 and 3.

Several issues may arise during database annotation that require careful attention, and hence careful manual curation of expected composition files is important:

1. Specific taxa may not be represented in a reference taxonomy to species level and must be annotated to the nearest common lineage. For example, *Streptococcus mutans* and *Streptococcus pneumoniae* are annotated as g__Streptococcus;s__ in the example above.
2. Multiple input strains, listed as separate entities in the “source” files, may need to be combined under common annotations in the “expected composition” files if they are not listed in the reference database. The relative abundance of an expected taxonomy will be equal to the sum of all members matching that taxonomy. For example, multiple strains may be combined as a single species, or species not listed in the reference database may be combined under a single genus; note the relative abundance of g__Streptococcus;s__ listed in the example above.
3. Reference databases may contain quirks that complicate annotation of expected composition files, such as listing strain IDs or different taxonomic lineages for multiple

entries of the same species. MC developers should carefully inspect reference database annotations and all expected composition files.

Expected composition data will consist of one of the following types:

1. Microbiome MC: expected taxonomic composition for a mixture of microbial cells. The taxonomic annotations present in the expected data will be specific to the database version that is used for analysis, and will be meaningless if used for different database versions. Likewise, they may not match the source annotation (i.e., the taxonomy of each strain to the best knowledge of the MC's creator) if taxonomic annotations have been revised or if the reference database being used does not contain a given taxonomy.
2. Metagenome MC: expected gene composition for a mixture of microbial cells/genomes. Gene annotations will be reference database specific, as for microbiome MCs above.

Other MC data types are theoretically possible, and could be included in mockrobiota, which only defines required information, files, and file formats. Expected data definitions can expand as other MC data types are contributed to mockrobiota.

The MCs currently deposited in mockrobiota are all microbiome MCs, representing known compositions of microbial species analyzed using marker-gene sequencing methods. Taxonomy strings for 16S rRNA MCs were generated using Greengenes 13_8 release (9) and Silva 119 release (10). Taxonomy strings for fungal ITS MCs were generated using the UNITE+INSD database (9-24-12 release) (11) prefiltered at 97% ID, and from which sequences with incomplete taxonomy strings and empty taxonomy annotations (e.g., "uncultured fungus") were

removed, as described previously (12). Taxonomy strings for the 18S rRNA MC were generated using Silva 119 release (10).

Raw Data Generation

Raw data for MCs fall into different types, corresponding to the MC types and expected composition data defined above:

1. Microbiome MC: raw data consist of marker-gene sequencing data.
2. Metagenome MC: raw data consist of shotgun metagenome sequences.

The raw data for each microbiome MC currently available in the repository were generated by 11 separate sequencing runs on the Illumina GAIIx (*n* = 1), HiSeq2000 (*n* = 6), and MiSeq (*n* = 4), as described previously (Table 1). These consisted of genomic DNA from known species isolates deliberately combined at defined rRNA copy-number ratios.

References

1. Human Microbiome Project Consortium. 2012. A framework for human microbiome research. Nature 486: 215–221.
2. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79(17):5112-20. doi: 10.1128/AEM.01043-13.

3. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature Methods 10: 57-59.

4. Schloss PD, Gevers D, Westcott SL. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS One 6:e27310. http://dx.doi.org/10.1371/journal.pone.0027310.

5. Mysara M, Leys N, Raes J, Monsieurs P. 2015. NoDe: a fast error-correction algorithm for pyrosequencing amplicon reads. BMC Bioinformatics 16:88. doi: 10.1186/s12859-015-0520-5.

6. Kopylova E, Navas-Molina JA, Mercier C, Xu ZZ, Mahe F, He Y, Zhou HW, Rognes T, Caporaso JG, Knight R. 2016. Open-source sequence clustering methods improve the state of the art. mSystems 1 (1) DOI: 10.1128/mSystems.00003-15

7. Bokulich NA, Rideout JR, Kopylova E, Bolyen E, Patnode J, Ellett Z, McDonald D, Wolfe B, Maurice CF, Dutton RJ, Turnbaugh PJ, Knight R, Caporaso JG. 2015. A standardized, extensible framework for optimizing classification improves marker-gene taxonomic assignments. PeerJ e1502: https://dx.doi.org/10.7287/peerj.preprints.934v2.

8. Peabody MA, Van Rossum T, Lo R, Brinkman FS. 2015. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC Bioinformatics 16:363. doi: 10.1186/s12859-015-0788-5.

9. McDonald D, Price MN, Goodric J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6: 610-618.

10. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucl. Acids Res. 42: D643-D648.

11. Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, Bates ST, Bruns TD, Bengtsson-Palme J, Callaghan TM, Douglas B, Drenkhan T, Eberhardt U, Dueñas M, Grebenc T, Griffith GW, Hartmann M, Kirk PM, Kohout P, Larsson E, Lindahl BD, Lücking R, Martín MP, Matheny PB, Nguyen NH, Niskanen T, Oja J, Peay KG, Peintner U, Peterson M, Põldmaa K, Saag L, Saar I, Schüßler A, Scott JA, Senés C, Smith ME, Suija A, Taylor DL, Telleria MT, Weiß M, Larsson KH. 2013. Towards a unified paradigm for sequence-based identification of Fungi. Molecular Ecology 22: 5271–5277.

12. Bokulich NA, Mills DA. 2013. Improved Selection of internal transcribed spacer-specific primers enables quantitative, ultra-high-throughput profiling of fungal communities. Applied and Environmental Microbiology 79: 2519-2526.

13. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences of the United States of America 108: 4516-4522.

14. Maurice CF, Haiser HJ, Turnbaugh PJ. 2013. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. Cell 152: 39–50.

Table 1. Marker-gene sequencing MCs currently available in mockrobiota.

Dataset Name	Target Region	Read Length (nt)	Method	Sample Count[a]	Strain Count	Original Citation
mock-1	16S	100	HiSeq	1 E	48	(3)
mock-2	16S	150	MiSeq	1 E	48	(3)
mock-3	16S	250	MiSeq	2 E 2 S	22	(3)
mock-4	16S	150	MiSeq	2 E 2 S	22	(3)
mock-5	16S	250	MiSeq	2 E 2 S	22	(7)
mock-6	16S	100	GAIIx	3 E	67	(13)
mock-7	16S	100	HiSeq	3 E	67	(14)
mock-8	16S	100	HiSeq	3 E	67	(7)
mock-9	ITS	100	HiSeq	3 E	16	(7)
mock-10	ITS	100	HiSeq	3 E	16	(7)

mock-11	18S	90	HiSeq	1 E	12	(3)

[a]Number of MC samples contained in MC dataset. E = samples with even abundance ratios among strains; S = staggered (uneven) abundance ratios.

Table 2. Example source composition.

#Taxonomy	sample1
Staphylococcus aureus ATCC BAA-1718	0.200
Staphylococcus epidermidis ATCC 12228	0.200
Streptococcus agalactiae ATCC BAA-611	0.200
Streptococcus mutans ATCC 700610	0.200
Streptococcus pneumoniae ATCC BAA-334	0.200

Table 3. Example expected composition, annotated with Greengenes 13_8 reference taxonomy.

#Taxonomy	sample1
k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Staphylococcaceae;g__Staphylococcus;s__aureus	0.200
k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Staphylococcaceae;g__Staphylococcus;s__epidermidis	0.200

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__	0.400
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__agalactiae	0.200

Figure Captions

Figure 1. Example usage of mockrobiota MC resource for marker-gene sequencing pipelines. MC datasets are selected based on multiple input criteria, including dataset metadata, sample metadata, and represented taxa. Raw data (e.g., fastq) are demultiplexed, sequences are dereplicated or clustered as OTUs, and taxonomy is assigned to representative sequences. Observed taxonomic assignments and abundances are compared to the expected composition (expected taxonomic assignments and abundances) of that MC, e.g., to generate precision and recall scores or correlations between observed/expected values.