FOCUS: an Alignment-free Model to Identify Organisms in Metagenomes Using Non-negative Least Squares

One of the major goals in metagenomics is to identify the organisms present in a microbial community from unannotated shotgun sequencing reads. Taxonomic profiling has valuable applications in biological and medical research, including disease diagnostics. Most currently available approaches do not scale well with increasing data volumes, which is important because both the number and lengths of the reads provided by sequencing platforms keep increasing. Here we introduce FOCUS, an agile composition based approach using non-negative least squares (NNLS) to report the focal organisms present in metagenomic samples and profile their abundances. FOCUS was tested with simulated and real metagenomes, and the results show that our approach accurately predicts the organisms present in microbial communities. FOCUS was implemented in Python. The source code and web-sever are freely available at http://edwards.sdsu.edu/FOCUS.

FOCUS: an Alignment-free Model to Identify Organisms in Metagenomes Using Non-negative Least Squares

Genivaldo Gueiros Z. Silva[1], Daniel A. Cuevas[1], Bas E. Dutilh [4, 5], and Robert A. Edwards [1, 2, 3, 5, 6] *

[1]Computational Science Research Center, [2]Department of Computer Science, and [3]Department of Biology, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA, [4]Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Centre, Geert Grooteplein 28, 6525 GA, Nijmegen, The Netherlands, [5]Department of Marine Biology, Institute of Biology, Federal University of Rio de Janeiro, Brazil, [6]Division of Mathematics and Computer Science, Argonne National Laboratory, 9700 S. Cass Ave, Argonne, IL 60439, USA

*For correspondence please contact Dr. Robert Edwards at redwards@mail.sdsu.edu.

Abstract

One of the major goals in metagenomics is to identify the organisms present in a microbial community from unannotated shotgun sequencing reads. Taxonomic profiling has valuable applications in biological and medical research, including disease diagnostics. Most currently available approaches do not scale well with increasing data volumes, which is important because both the number and lengths of the reads provided by sequencing platforms keep increasing. Here we introduce FOCUS, an agile composition-based approach using non-negative least squares (NNLS) to report the focal organisms present in metagenomic samples and profile their abundances. FOCUS was tested with simulated and real metagenomes, and the results show that our approach accurately predicts the organisms present in microbial communities. FOCUS was implemented in Python. The source code and web-sever are freely available at http://edwards.sdsu.edu/FOCUS.

Introduction

Microbes are more abundant than any other cellular organism (Whitman et al. 1998), and it is important to understand which organisms are present and what they are doing (Handelsman 2004). In many environments a majority of the microbial community members cannot be cultured, and metagenomics is a powerful tool to directly probe uncultured genomes and understand the diversity of microbial communities by using only their DNA (Sharon and Banfield 2013).

Understanding microbial communities is important in many areas of biology. For example, metagenomes can distinguish taxonomic and functional signatures of microbes associated with marine animals (Trindade-Silva et al. 2012) or disease states (Belda-Ferre et al. 2012). Large sequencing volumes, short read lengths, and sequencing errors make the task of

identifying the diversity of organisms present in metagenomes challenging (Mande et al. 2012). Many programs exist for this, and they are either homology- or composition-based.

Homology-based programs normally use the BLAST program (Altschul et al. 1997) to identify the best hit in a large database output. In MG-RAST (Meyer et al. 2008) sequences are aligned to a set of databases in order to classify the metagenomic sample. MetaPhlAn (Segata et al. 2012) and GenomePeek (McNair and Edwards) use a reduced database containing only marker genes, e.g., unique clades and housekeeping genes, allowing the BLAST search to be fast. PhymmBL (Brady and Salzberg 2011) improves the BLAST results using interpolated Markov models. GASiC (Lindner and Renard 2013) uses Bowtie (Langmead et al. 2009) and the reference genomes similarities to correct the observed abundance estimated. Parallel-Meta (Su et al. 2012) a fast program, which requires a GPU, uses megaBLAST (Zhang et al. 2000) and HMM (Hidden Markov Model) to improve the homology result. Most of these applications classify sequences individually, and generate a taxonomic profile by summing the bins.

In general, composition-based approaches use oligonucleotide (*k*-mer) frequencies. Taxy (Meinicke et al. 2011) uses oligonucleotide distribution in metagenomes and in reference genomes and uses mixture modeling to identify the organisms present in the metagenome, and RAIphy (Nalbantoglu et al. 2011) identifies organisms using oligonucleotides and relative abundance index.

We developed a new approach that reconstructs a taxonomic profile using an ensemble *k*-mer composition of the entire metagenome. We compute the optimal set of organism abundances using non-negative least squares (NNLS) to match the metagenome *k*-mer composition to organisms in a reference database. *K*-mers have previously been used to cluster unknown sequences (Teeling et al. 2004; McHardy et al. 2007) and NNLS has been used to identify the genera present in metagenomic samples based on variations in gene count (Carr et al. 2013). Here we combine these two approaches in FOCUS, an ultra fast, accurate, composition based approach

to identify the taxa present in a metagenome. We compare the performance of FOCUS to GASiC, MetaPhlAn, RAIphy, PhymmBL, Taxy, and MG-RAST.

Methods

FOCUS workflow is described in Figure 1. As in most composition-based approaches, a training set is pre-generated using the complete genomes information, and here the non-negative least squares (NNLS) is applied to compute the relative abundance of each organism in the database into the unknown data**.**

Reference dataset

FOCUS requires a group of reference genomes to model and identify the organisms present in a metagenome. 2,766 complete genomes were downloaded from the SEED servers (Aziz et al. 2012) on 20 December 2013 (see Supplementary Table 1). *K*-mer frequencies (k=6-8, default: k=7) were calculated for both strands using Jellyfish 1.1.6 (Marçais and Kingsford 2011), reducing the number of dimensions (Strous et al. 2012), and *k*-mer counts were normalized by the sum of frequencies. The user can also create their own training set, which is scalable to the quickly increasing number of available reference genomes because it also uses Jellyfish in the *k*-mer counting.

Simulated and real metagenomes

In order to evaluate FOCUS performance, a simulated dataset of short sequences (SimShort), containing 500,000 single 100 nt reads, was created using the supplied error model for Illumina GA IIx with TrueSeq SBS Kit v5−GA using GemSim (McElroy et al. 2012) (Supplementary Table 2). The previous published high complexity simulated dataset (SimHC) from FAMeS (Mavromatis et al. 2007) was also used in the evaluation. Moreover, real

metagenomic datasets were selected as test cases: one under healthy conditions, one under disease conditions (MG-RAST accession 4447943.3 and 4447192.3) (Belda-Ferre et al. 2012), one fecal sample from a healthy individual (MG-RAST accession 4440945.3) (Kurokawa et al. 2007), and three hundred datasets from the Human Micriobiome Project (HMP) (Consortium 2012) (Supplementary Table 3) were selected as a test case.

Non-Negative Least Squares (NNLS)

The estimation of a parameterized model to understand some data is a fundamental problem in data modeling. Nevertheless, the estimation is not always easy, e.g., in problems like metagenome profiling that cannot have negative values for the fitted parameters. In such case, a solution can be estimated using NNLS, which is defined as:

Given a matrix $A \in \mathbb{R}^{mxn}$ and a vector $b \in \mathbb{R}^{m}$, where $m \geq n$, find a non-negative vector $x \in \mathbb{R}^{n}$ to minimize the function (1).

$$f(x) = \frac{1}{2} ||Ax-b||^2, \quad where \quad x \geq 0 \quad and \quad \sum_{i=i}^{n} x_i = 1 \qquad (1)$$

In FOCUS, the reference matrix A is composed of m *k*-mer frequencies from n genomes, while a vector describing the user's metagenomic dataset is calculated from the *k*-mer frequencies of both strands from the dataset using Jellyfish. FOCUS uses non-negative least squares to compute the set of *k*-mer frequencies x that explains the optimal possible abundance of *k*-mers in the user's metagenome by selecting the optimal number of frequencies from the matrix A. We minimize the sum of squared differences (1) using the open source Scipy library (Jones et al. 2001) which has a module for the NNLS algorithm which solves the KKT (Karush-Kuhn-Tucker) conditions (Lawson and Hanson 1987). We added Tikhonov regularization (Garda and Galias 2012) to deal with genomes that have similar *k*-mer compositions.

115

116 **Jackknife resampling of the data**

117 We implemented a jackknife resampling strategy to assess the robustness of the results.
118 50% of the reads were randomly resampled 1000x, and the species frequencies recalculated. For
119 each species, these 1000 frequencies were averaged and the standard deviation calculated to
120 estimate the spread.

121 **Web-based and graphical user interface version**

122 As an alternative to the command line version of the program, we have created a user-
123 friendly web version and a graphical user interface (GUI) for Microsoft Windows users. The web
124 server and the GUI are available at http://edwards.sdsu.edu/FOCUS.

125 **Results and Discussion**

126

127 **Evaluation and comparison with other tools**

128 All tools were run using default parameters and their default reference database, either
129 online (MG-RAST) or using one core on a server with 24 processors x 6 cores Intel(R) Xeon(R)
130 CPU X5650 @ 2.67GHz and 189 GB RAM. We only compared GASiC to the SimHC dataset
131 which had the results previously published (Lindner and Renard 2013). We tried to run the tool;
132 however, it requires a large amount of storage during the process to save its output data.

133 For the real data, three hundred and three metagenomic datasets were selected. First, the
134 metagenomic sample of the human oral cavity from diseased conditions was used. MetaPhlAn
135 apparently over predicted the genera *Veillonella* due to the short genome, and Taxy did not
136 predict *Prevotella* hits (see Figure 2) as described in (Belda-Ferre et al. 2012). FOCUS was able
137 to profile the organisms in only 41 seconds. Taxy took about 45 seconds, MetaPhlAn took about
138 3 minutes, RAIphy took 52 minutes, MG-RAST took 3 days, and PhymmBL took 1 week and 6

days. Using random subsets for the oral metagenome, we tested the tools scalability and showed that FOCUS and Taxy profile metagenomes in constant time (see Figure 3).

The oral metagenome from the healthy condition was used. MetaPhlAn possibly over predicted the genera *Neisseria*, and Taxy was not able to predict *Rothia* hits (see Figure 4). FOCUS profiled the metagenome in only 35 seconds. Taxy took about 41 seconds, MetaPhlAn took about 2 minutes, RAIphy took 48 minutes, MG-RAST took 3 days, and PhymmBL took 9 days.

A fecal metagenome from a healthy individual was used. All the tools predicted that *Bifidobacterium* and *Enterococcus* were the two most abundant genera in the sample. However, RAIphy apparently under predicted the genera *Bifidobacterium* (see Figure 5). For this small dataset, FOCUS profiled the metagenome in 35 seconds. Taxy took about 40 seconds, MetaPhlAn took only 30 minutes, RAIphy took about 4 minutes, MG-RAST took 3 days, and PhymmBL took 2 days and 14 hours.

Three hundred metagenomic samples (254 GB total) from HMP were analyzed at all the taxonomy levels using FOCUS (Supplementary Table 4) in about 1 hour and 20 minutes and compared with the published results from MetaPhlAn's paper (Segata et al. 2012) by calculating the Euclidean distance between the results (see Figure 6). For most of the samples, FOCUS and MetaPhlAn have similar predictions at the genera level but vary at the species level. However, for some samples in the posterior fornix and most of the samples from the anterior nares there were differences at all levels which may reflect the additional genome sequencing of isolates from those passages that has occurred since 2012. Other tools were not included in the analysis due to the CPU processing time.

For the simulated data, we removed species from the reference dataset that are present in this dataset and tried to predict the genera present in the SimShort dataset. A major limitation of many of the approaches discussed here is that the underlying databases cannot be changed. Only

FOCUS, RAIphy, GASiC, and PhymmBL allow the end user to change their reference database. FOCUS and PhymmBL best predicted the correct genera while RAIphy could not correctly predict their abundance (Figure 7). FOCUS had the fastest performance (45 seconds); RAIphy took about 2 hours, while PhymmBL took approximately 5 days. Supplementary Figure 1 to 5 show the same comparison for other taxonomy resolutions.

For the SimHC simulated metagenomes, the genera present in the dataset were deleted from the training dataset, and we evaluated the class-level prediction. The tested tools correctly predicted the classes, except that RAIphy over predicted the top two classes (see Figure 8). Again, FOCUS was the fastest tool (30 seconds) in comparison to RAIphy, which took about 1 hour and 50 minutes, and PhymmBL, which took about 4 days. See Supplementary Figure 6 to 8 for other taxonomy levels.

Furthermore, for the SimHC dataset, we ran all the previously used tools and the GASiC published results to evaluate the genera-level prediction. GASiC and PhymmBL had the best predictions, and FOCUS failed in the prediction of 4 minor genera probably because many organisms present in the SimHC dataset were not included in the FOCUS database (see Figure 9). We did not compare the running time because we extracted the GASiC results from its paper; however, in the original paper it took 2 days and needed at least 500 GB of storage to analyze the SimHC simulated metagenome.

The very small standard deviations observed after jackknife re-sampling indicate the robustness of our results. Furthermore, in order to show a quantitative evaluation between the real and predicted abundance for the synthetic metagenomes, we computed the Euclidean distance between the real and predicted abundances for all the simulated data presented above (see Figure 10). For some of the tools, only genus level predictions are available, but for RAIphy, PhymmBL, and FOCUS we included all taxonomic levels. The data demonstrate that FOCUS had the best prediction in more than half of test cases.

These tests were performed on a server; however, FOCUS is also ultra fast on a simple computer. For example, we profiled the real dataset in 1 minute and 45 seconds using an Intel(R) Core(TM) i3 @2.53 GHz and 1GB RAM. In addition to the Web server, we have developed a stand-alone version that runs on the Windows® platform.

Limitations

As with other methods created to profile metagenome sequences, FOCUS depends on a curated database of microbial reference genomes in order to predict a specific genus. If a reference genome is absent, the tool will predict the closest reference available.

Conclusions

Here we present FOCUS, an agile solution to identify the organisms present in metagenomic samples that does not rely on mapping individual reads, but instead determines the taxonomic composition of the entire metagenome at once by using NNLS. This makes FOCUS an extremely fast and scalable solution to profile the focal taxa in a metagenome. FOCUS reports very similar species compositions as currently available, state of the art metagenome profiling tools.

Availability and requirements

Project name: FOCUS

Project and web server home page: http://edwards.sdsu.edu/FOCUS

Operating system: the program has a command line version that works on OS X and Unix, and a GUI for Microsoft Windows users.

Programming language: Python 2.7.

Other requirements: Numpy (http://www.numpy.org), Scipy (http://scipy.org), Jellyfish (http://www.cbcb.umd.edu/software/jellyfish), and Python programming language (http://www.python.org).

License: GNU GPL3.

Any restrictions to use by non-academics: no special restrictions.

Acknowledgements:

We thank Dr. Peter Blomgren for help with the Advanced Numerical Analysis, and the reviewers for their useful comments.

References

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, 1997: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402, doi:10.1093/nar/25.17.3389.

Aziz, R. K., and Coauthors, 2012: SEED Servers: High-Performance Access to the SEED Genomes, Annotations, and Metabolic Models. *PLoS ONE*, **7**, e48053, doi:10.1371/journal.pone.0048053.

Belda-Ferre, P., L. D. Alcaraz, R. Cabrera-Rubio, H. Romero, A. Simón-Soro, M. Pignatelli, and A. Mira, 2012: The oral metagenome in health and disease. *ISME J.*, **6**, 46–56, doi:10.1038/ismej.2011.85.

Brady, A., and S. Salzberg, 2011: PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nat. Methods*, **8**, 367–367, doi:10.1038/nmeth0511-367.

Carr, R., S. S. Shen-Orr, and E. Borenstein, 2013: Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution. *PLoS Comput Biol*, **9**, e1003292, doi:10.1371/journal.pcbi.1003292.

Consortium, T. H. M. P., 2012: Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214, doi:10.1038/nature11234.

Garda, B., and Z. Galias, 2012: Non-negative least squares and the Tikhonov regularization methods for coil design problems. *2012 International Conference on Signals and Electronic Systems (ICSES)*, 2012 International Conference on Signals and Electronic Systems (ICSES), 1–5.

Handelsman, J., 2004: Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiol. Mol. Biol. Rev.*, **68**, 669–685, doi:10.1128/MMBR.68.4.669-685.2004.

Jones, E., T. Oliphant, and P. Peterson, 2001: SciPy: Open source scientific tools for Python. *http://www.scipy.org/*,. http://www.scipy.org/Citing_SciPy (Accessed October 23, 2013).

Kurokawa, K., and Coauthors, 2007: Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes*, **14**, 169–181, doi:10.1093/dnares/dsm018.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25, doi:10.1186/gb-2009-10-3-r25.

Lawson, C., and R. J. Hanson, 1987: *Solving Least Squares Problems*. SIAM,.

Lindner, M. S., and B. Y. Renard, 2013: Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Res.*, **41**, e10, doi:10.1093/nar/gks803.

Mande, S. S., M. H. Mohammed, and T. S. Ghosh, 2012: Classification of metagenomic sequences: methods and challenges. *Brief. Bioinform.*, **13**, 669–681, doi:10.1093/bib/bbs054.

Marçais, G., and C. Kingsford, 2011: A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770, doi:10.1093/bioinformatics/btr011.

Mavromatis, K., and Coauthors, 2007: Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods*, **4**, 495–500, doi:10.1038/nmeth1043.

McElroy, K. E., F. Luciani, and T. Thomas, 2012: GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*, **13**, 74, doi:10.1186/1471-2164-13-74.

McHardy, A. C., H. G. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos, 2007: Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*, **4**, 63–72, doi:10.1038/nmeth976.

McNair, K., and R. Edwards, GenomePeek – A tool for prokaryotic genome and metagenome analysis.

Meinicke, P., K. P. Aßhauer, and T. Lingner, 2011: Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics*, btr266, doi:10.1093/bioinformatics/btr266.

Meyer, F., and Coauthors, 2008: The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386, doi:10.1186/1471-2105-9-386.

277 Nalbantoglu, O. U., S. F. Way, S. H. Hinrichs, and K. Sayood, 2011: RAIphy: Phylogenetic
278 classification of metagenomics samples using iterative refinement of relative abundance
279 index profiles. *BMC Bioinformatics*, **12**, 41, doi:10.1186/1471-2105-12-41.

280 Segata, N., L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, 2012:
281 Metagenomic microbial community profiling using unique clade-specific marker genes.
282 *Nat. Methods*, **9**, 811–814, doi:10.1038/nmeth.2066.

283 Sharon, I., and J. F. Banfield, 2013: Genomes from Metagenomics. *Science*, **342**, 1057–1058,
284 doi:10.1126/science.1247023.

285 Strous, M., B. Kraft, R. Bisdorf, and H. E. Tegetmeyer, 2012: The binning of metagenomic
286 contigs for microbial physiology of mixed cultures. *Front. Microbiol.*, **3**, 410,
287 doi:10.3389/fmicb.2012.00410.

288 Su, X., J. Xu, and K. Ning, 2012: Parallel-META: efficient metagenomic data analysis based on
289 high-performance computation. *BMC Syst. Biol.*, **6**, S16, doi:10.1186/1752-0509-6-S1-
290 S16.

291 Teeling, H., J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner, 2004: TETRA: a web-
292 service and a stand-alone program for the analysis and comparison of tetranucleotide
293 usage patterns in DNA sequences. *BMC Bioinformatics*, **5**, 163, doi:10.1186/1471-2105-
294 5-163.

295 Trindade-Silva, A. E., and Coauthors, 2012: Taxonomic and Functional Microbial Signatures of
296 the Endemic Marine Sponge Arenosclera brasiliensis. *PLoS ONE*, **7**, e39905,
297 doi:10.1371/journal.pone.0039905.

298 Whitman, W. B., D. C. Coleman, and W. J. Wiebe, 1998: Prokaryotes: The unseen majority. *Proc.
299 Natl. Acad. Sci.*, **95**, 6578–6583.

300 Zhang, Z., S. Schwartz, L. Wagner, and W. Miller, 2000: A greedy algorithm for aligning DNA
301 sequences. *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, **7**, 203–214,
302 doi:10.1089/10665270050081478.

303

Figure 1

Figure 1

Workflow of the FOCUS program.

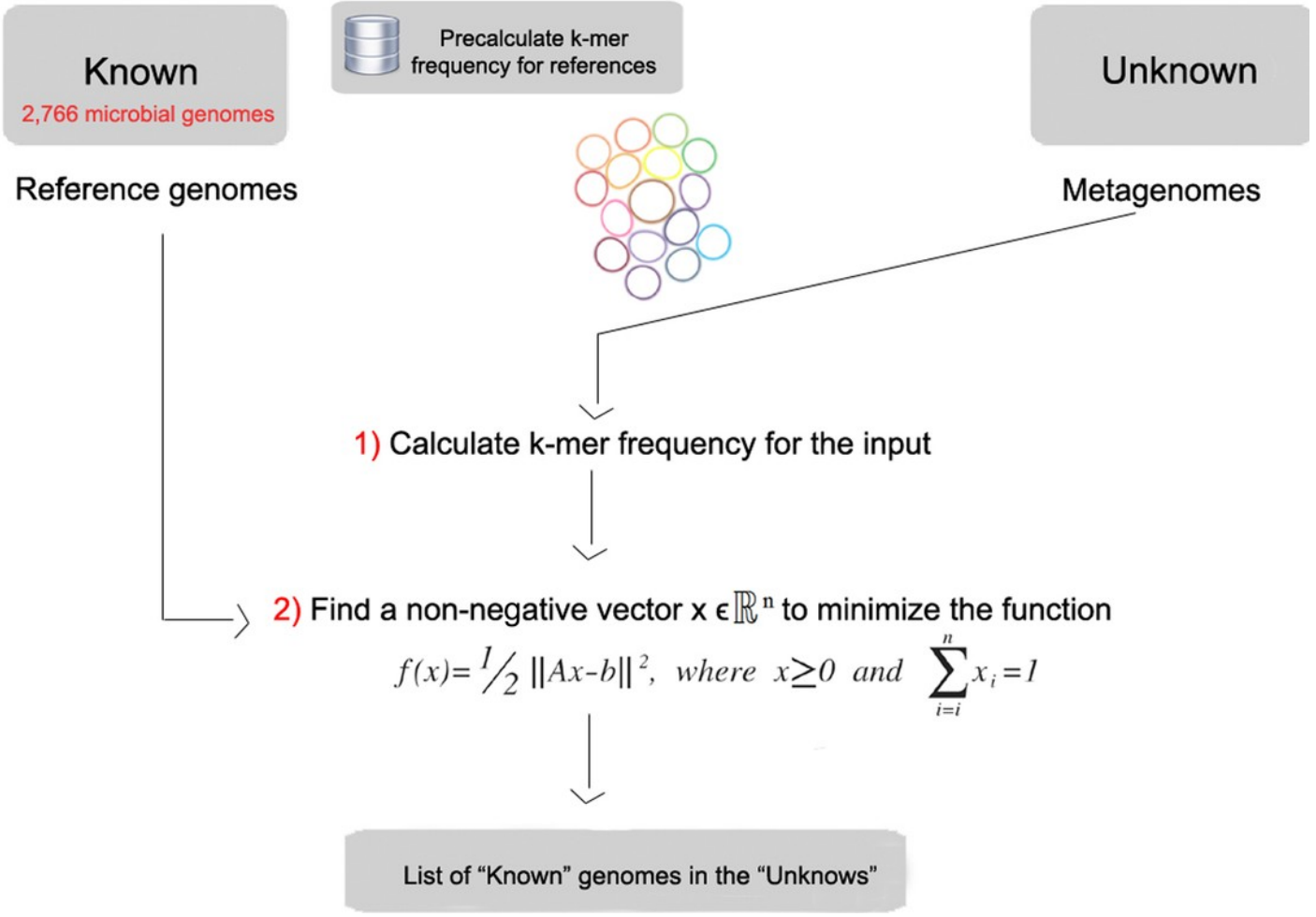

Figure 2

Figure 2

Genera-level taxonomy classification sorted by FOCUS prediction for the metagenome from a diseased human oral cavity using FOCUS, MetaPhlAn, MG-RAST, PhymnBL, RAIphy, Taxy, and FOCUS (mean). Error bars represent the standard deviation uncertainty in tested metagenome.

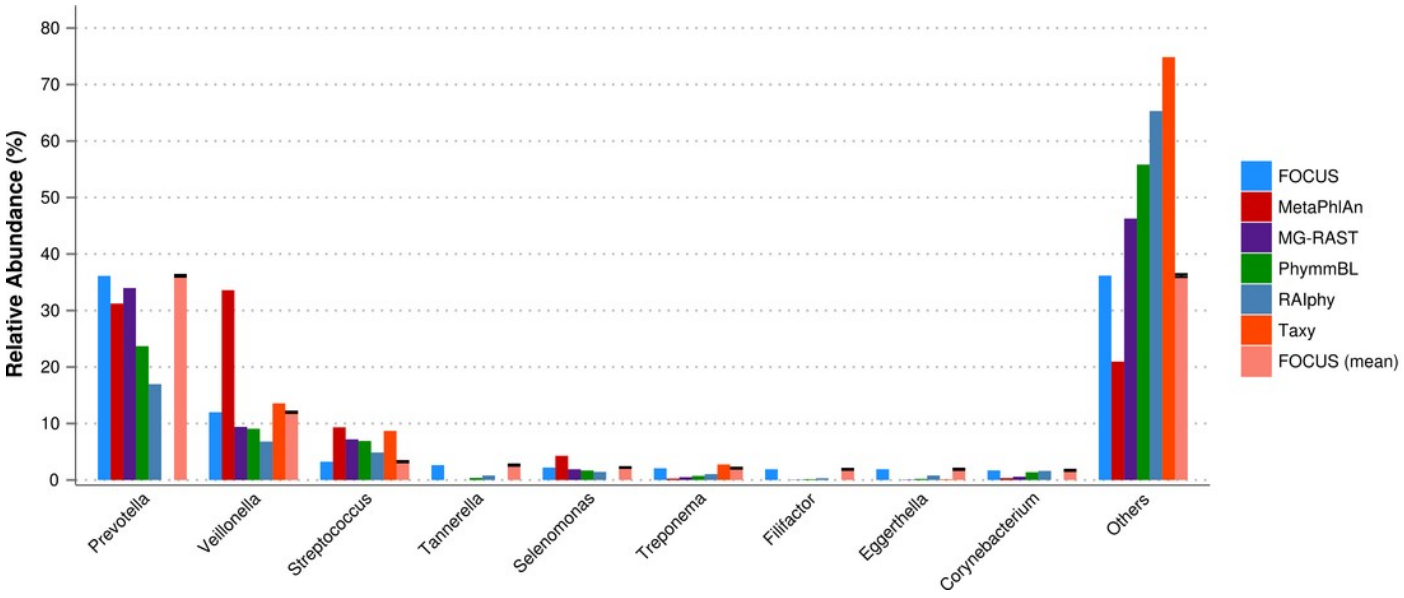

Figure 3

Figure 3

Scalability test using different sub-sets of the human oral cavity under disease metagenome using FOCUS, MetaPhlAn, MG-RAST, PhymnBL, RAIphy, Taxy.

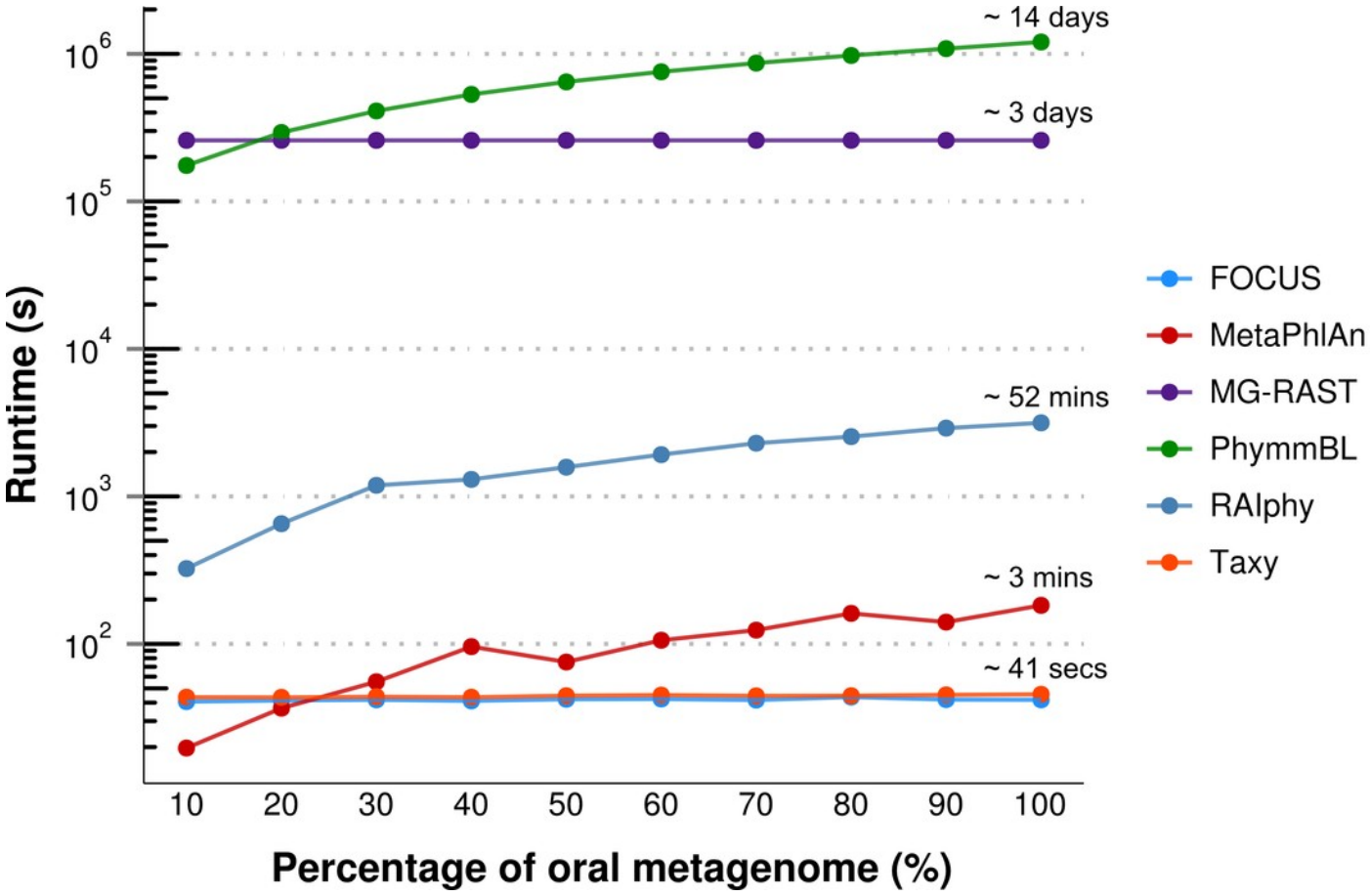

Figure 4

Figure 4

Genera-level taxonomy classification sorted by FOCUS prediction for the metagenome from a healthy human oral cavity using FOCUS, MetaPhlAn, MG-RAST, PhymnBL, RAIphy, Taxy, and FOCUS (mean). Error bars show the standard deviation uncertainty for the real metagenome.

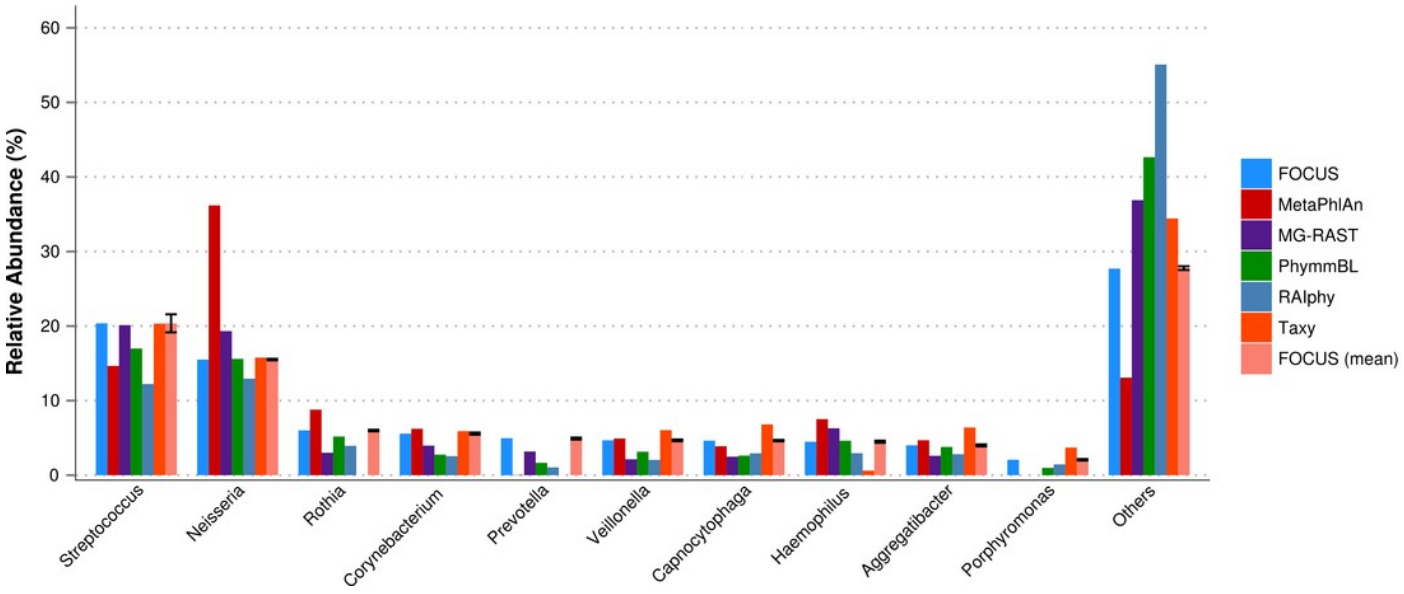

Figure 5

Figure 5

Genera-level taxonomy classification sorted by FOCUS prediction for the metagenome from a fecal metagenomic sample of a healthy human using FOCUS, MetaPhlAn, MG-RAST, PhymnBL, RAIphy, Taxy, and FOCUS (mean). Error bars show the standard deviation uncertainty for the real metagenome.

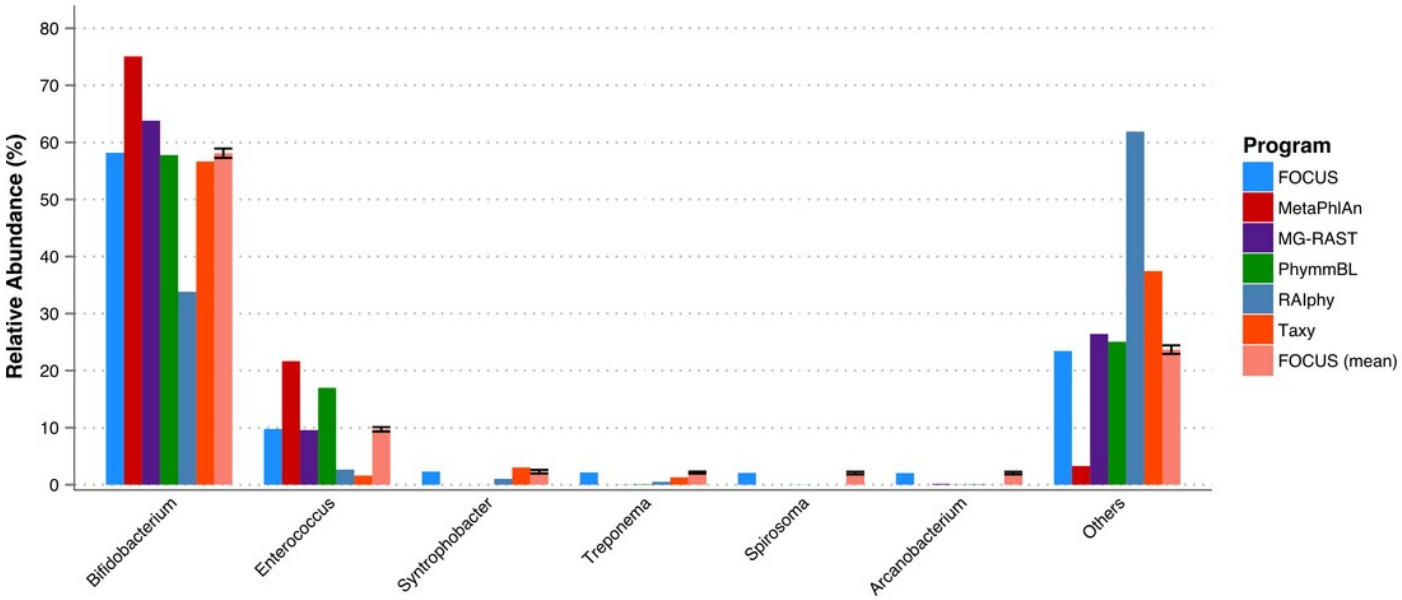

Figure 6

Figure 6

Heat-map representing the distance between the FOCUS and MetaPhlAn results for 300 metagenomes from the Human Microbiome Project across 15 body sites. The distance was computed using the Euclidean distance between the results of both tools.

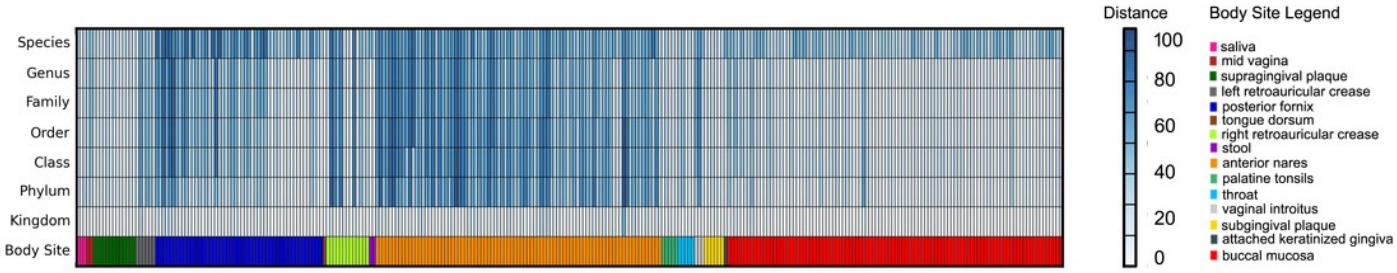

Figure 7

Figure 7

Genera-level taxonomy classification for the SimShort dataset using FOCUS, PhymnBL, RAIphy, and FOCUS (mean).

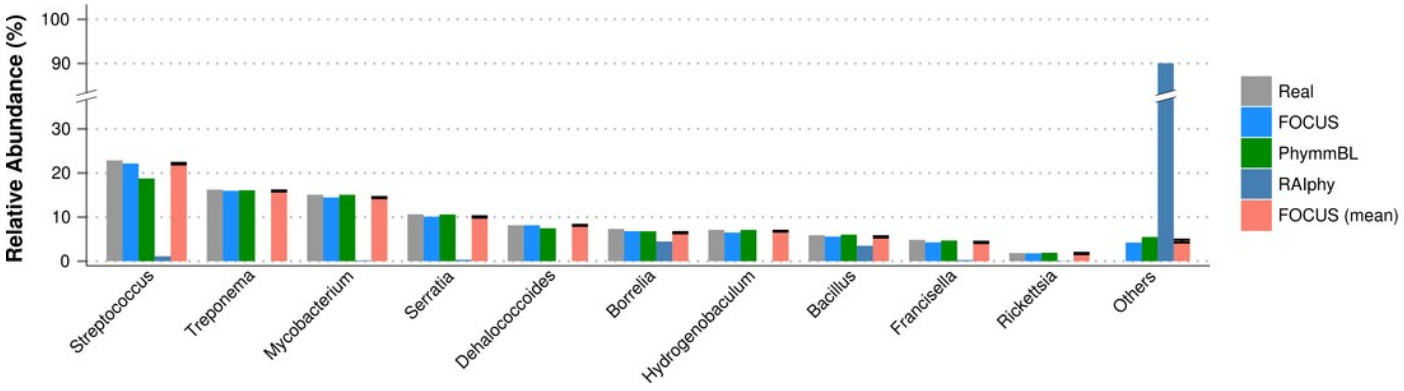

Figure 8

Figure 8

Class-level taxonomy classification for the SimHC dataset using FOCUS, PhymnBL, RAIphy, and FOCUS (mean).

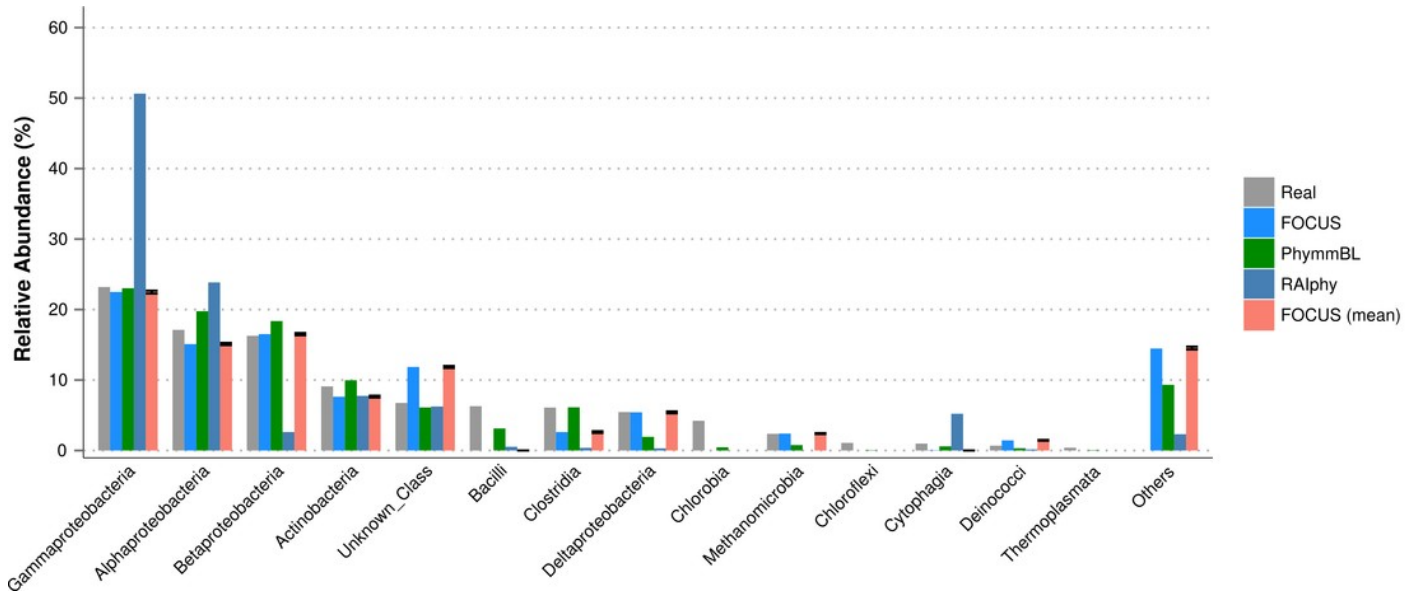

Figure 9

Figure 9

Genera-level taxonomy classification for the SimHC dataset using FOCUS, MetaPhlAn, MG-RAST, PhymnBL, RAIphy, Taxy, GASiC, and FOCUS (mean).

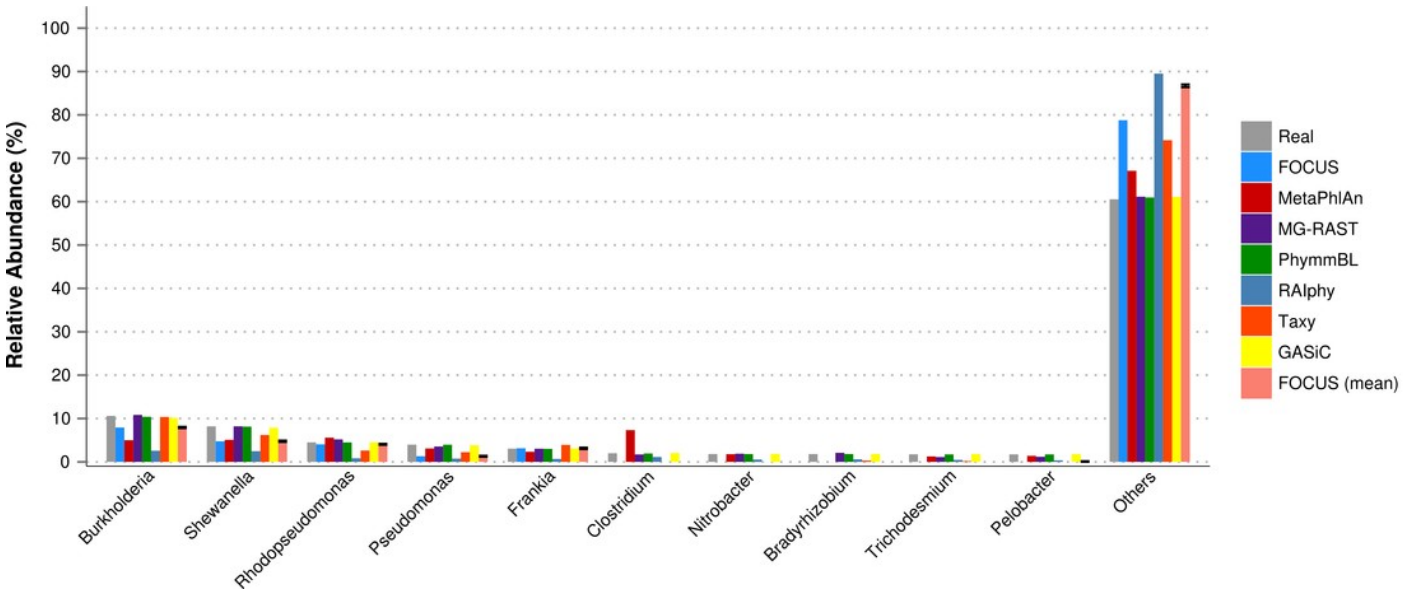

Figure 10

Figure 10

Numerical evaluation between the real and predicted abundance for the synthetic metagenomes computed by the Euclidean distance between the real and the predicted values.