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], 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 and it is important to understand what those organisms are doing and who they are. In many environments a majority of the microbial community members cannot be cultured. Metagenomics is a powerful tool to directly probe uncultured genomes and understand the diversity of microbial communities by using only their DNA.

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, including MetaPhlAn (Segata et al. 2012), RAIphy (Nalbantoglu et al. 2011), MG-RAST (Meyer et al. 2008), and PhymmBL (Brady and Salzberg 2011). These applications classify sequences individually, and generate a taxonomic profile by summing the bins. 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. 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 MetaPhlAn, RAIphy, PhymmBL, Taxy, and MG-RAST.

Methods

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 (Marçais and Kingsford 2011), reducing the number of dimensions (Strous et al. 2012), and *k*-mer counts were normalized by the sum of frequencies.

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, one real dataset of human oral cavity diseased samples (MG-RAST accession 4447943.3) (Belda-Ferre et al. 2012) was selected as a test case.

Non-Negative Least Squares (NNLS)

NNLS is useful to solve problems like metagenome profiling that cannot have negative values for the fitted parameters.

The NNLS problem is defined as:

Given a matrix A ϵ $\mathbb{R}$ and b ϵ $\mathbb{R}$, where m $\geq$ n, find a non-negative vector x ϵ $\mathbb{R}$ 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 its *k*-mer frequencies. FOCUS uses non-negative least squares to compute the set of *k*-mer frequencies x that explains the 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.

Jackknife resampling of the data

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

Results and Discussion

Evaluation and comparison with other tools

All tools were run using default parameters and their default reference database, either online (MG-RAST) or using one core on a server with 24 processors x 6 cores Intel(R) Xeon(R) CPU X5650 @ 2.67GHz and 189 GB RAM.

For the real data, one metagenomic dataset of the human oral cavity from diseased conditions was used. MetaPhlAn apparently over predicted the genera *Veillonella*, and Taxy did not predict *Prevotella* hits (see Figure 1) as described in (Belda-Ferre et al. 2012). FOCUS was able to profile the organ-isms in only 38 seconds. Taxy took about 45 seconds, MetaPhlAn took about 3 minutes, RAIphy took 1 hour and 40 minutes, MG-RAST took 3 days, PhymmBL took 1 week and 6 days. Using different subsets for the oral metagenome, we tested the tools scalability and showed that FOCUS and Taxy profile metagenomes in constant time (see Figure 2).

For the simulated data, we removed species from the reference dataset that are present in this dataset and tried to predict the gen-era 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 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 (see Figure 3). FOCUS had the fastest performance (45 seconds) in comparison; RAIphy took about 2 hours, while PhymmBL took approximately 5 days. See Figure 4 for the species resolution.

For the SimHC simulated metagenomes, the genus 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 RAIphy that over predicted the top two classes (see Figure 5). Again, FOCUS was the fastest tool (30 seconds) in comparison RAIphy took about 1 hour and 50 minutes, and PhymmBL that took about 4 days.

The very small standard deviations observed after jackknife re-sampling indicate the robustness of our results.

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.

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 organ-isms present in metagenomic samples that does not rely on map-ping individual reads, but instead determines the taxonomic com-position 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.

Acknowledgements:

We thank Dr. Peter Blomgren for help with the Advanced Numerical Analysis.

References

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.

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.

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).

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

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.

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.

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

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

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

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

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

Figure 1

Figure 1

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.

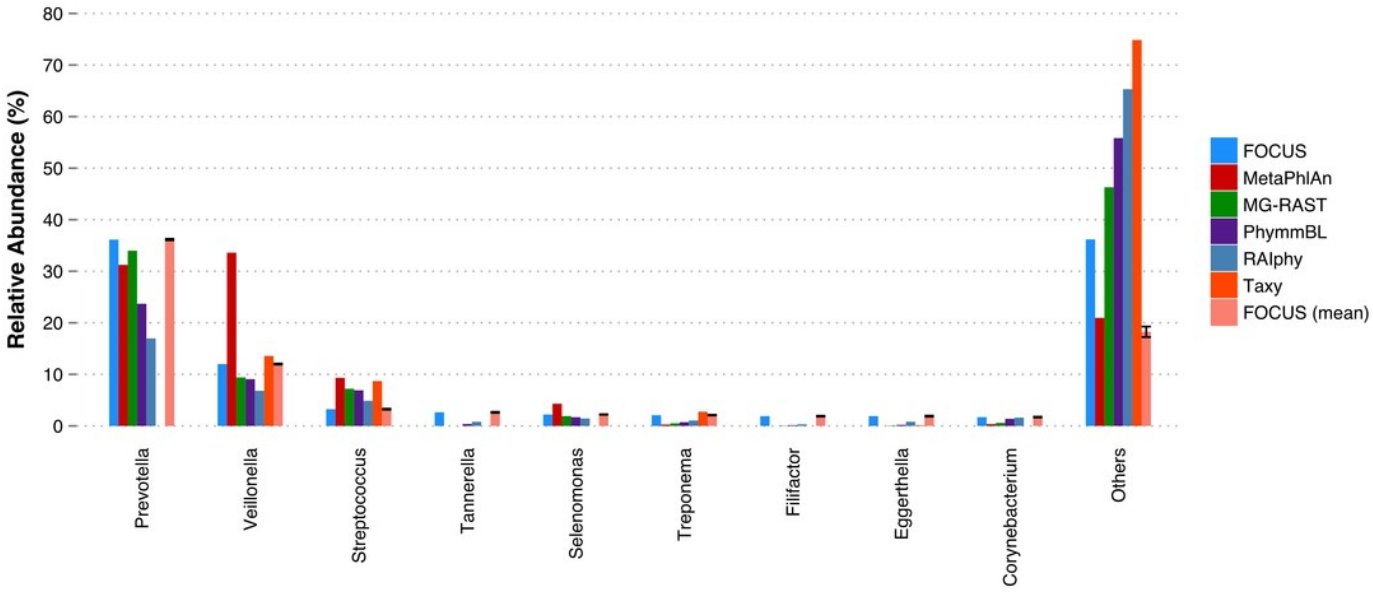

Figure 2

Figure 2

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

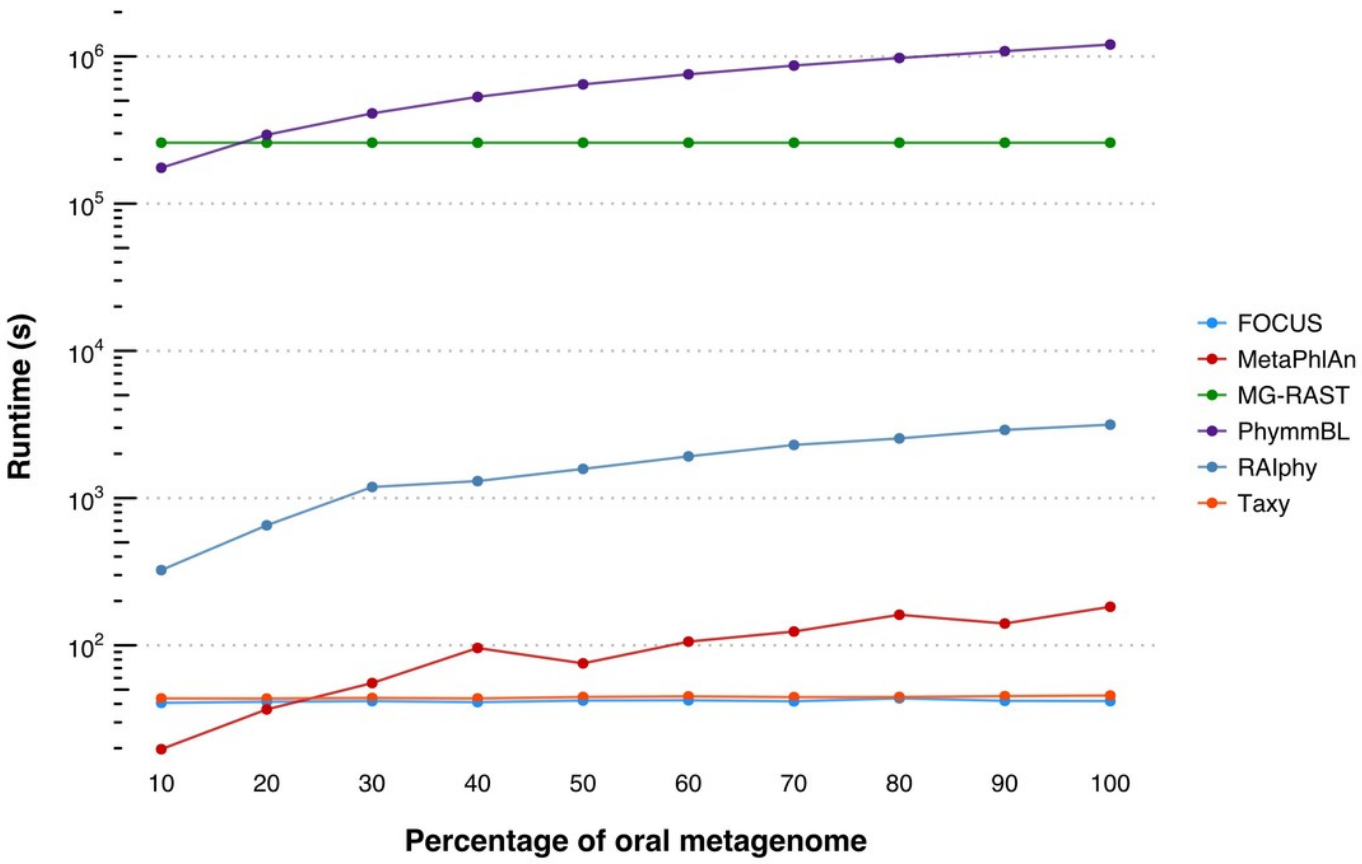

Figure 3

Figure 3

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

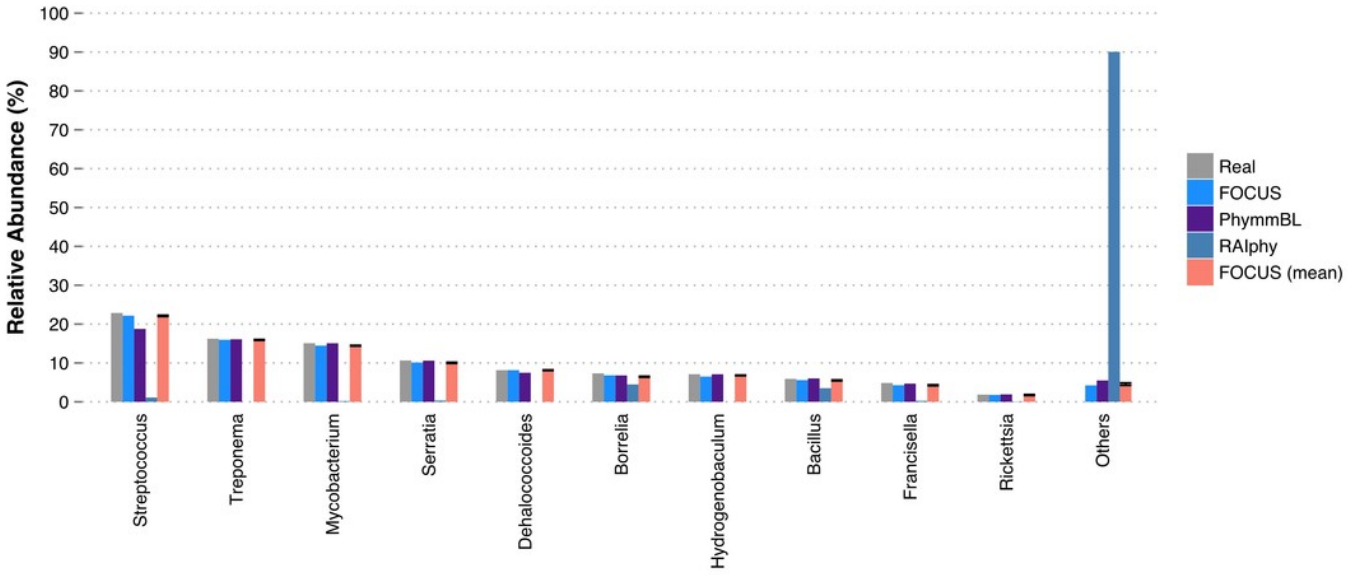

Figure 4

Figure 4

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

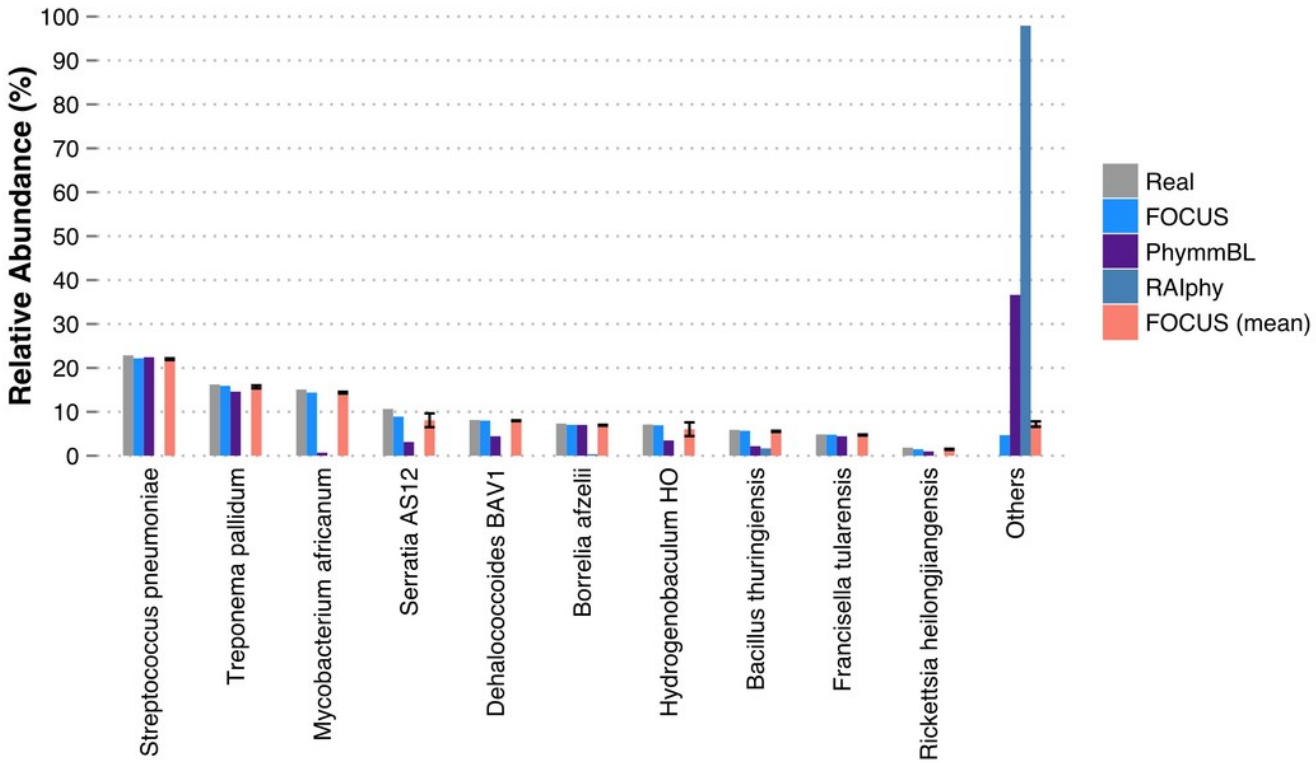

Figure 5

Figure 5

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