Mixture models for analysis of the taxonomic composition of metagenomes

P. Meinicke, K.P. Aßhauer, T. Lingner

University of Göttingen, Institute of Microbiology and Genetics, Department of Bioinformatics.
Goldschmidtstr. 1, 37077 Göttingen, Germany.
Phone: +49 551 39-14925
(http://gobics.de/peter/taxy)

Supplementary Figures, Tables and Text.

SUPPLEMENTARY RESULTS

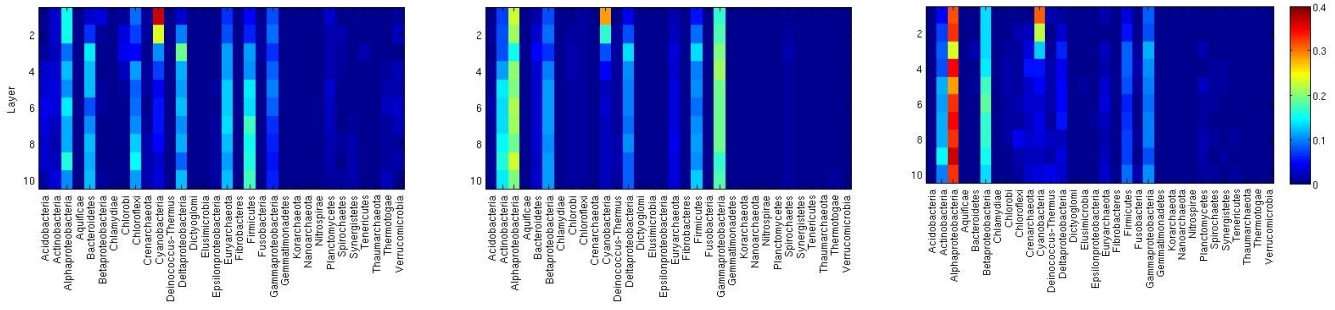
Comparison of profiles. Among publicly available data sets, the hypersaline microbial mat samples introduced in [1] provide a versatile test case for the comparison of taxonomic profiling methods. The data set includes 129,147 unassembled Sanger sequencing reads of approximately 700 base pairs (bp) in read length from ten samples according to different depth layers of the mat. In addition to the functional gradients, which have been reported in the original study [1], the taxonomic gradients across the ten depth layers are also valuable for investigation. For the comparison of methods, we focused on three complementary approaches, which cover the whole range of available methods. In addition to the Taxy method, we also included the Phymm [2] tool for signature-based short read classification and a Galaxy [3] analysis for classical homology-based profiling (see also Methods). In Supplementary Figure 1 the resulting profiles of the three methods are shown as three color-coded image matrices, where rows correspond to different depth layers, and columns represent different taxonomic categories at the subphylum level.

A common feature of all of the predictions is a strong Cyanobacteria gradient over the top four layers of the microbial mat. This gradient is supported by the original study [1], which, for the top layers, found a strong protein domain-based gradient for the Cyanobacteria-specific DUF820 family (PF05685). Interestingly, over the top layers, the Deltaproteobacteria and Cyanobacteria frequencies showed a negative correlation. This is in agreement with the original study in which the topmost layers were characterized by an inverse proportion of photosynthesis and sulfatase-related genes, with Deltaproteobacteria containing most of the known sulfate-reducing bacteria species. Additional millimeter-scale gradients were predicted by the Taxy and Phymm tools. Taxy predicted a bottom-heavy Firmicutes gradient, which was not apparent in the Phymm and BLAST-based predictions, while Phymm predicted a bottom-heavy Actinobacteria gradient. Below the top two layers, previous studies ([4][5]) identified Alphaproteobacteria, Bacteroidetes, Chloroflexi and Planctomycetes as the most abundant subphyla by phylotyping and Alphaproteobacteria, Gammaproteobacteria, Deltaproteobacteria and Bacteroidetes by a 16S survey. Among the three tested prediction tools, only Taxy identified the Bacteroidetes phylum as moderately abundant in all layers below the uppermost two.

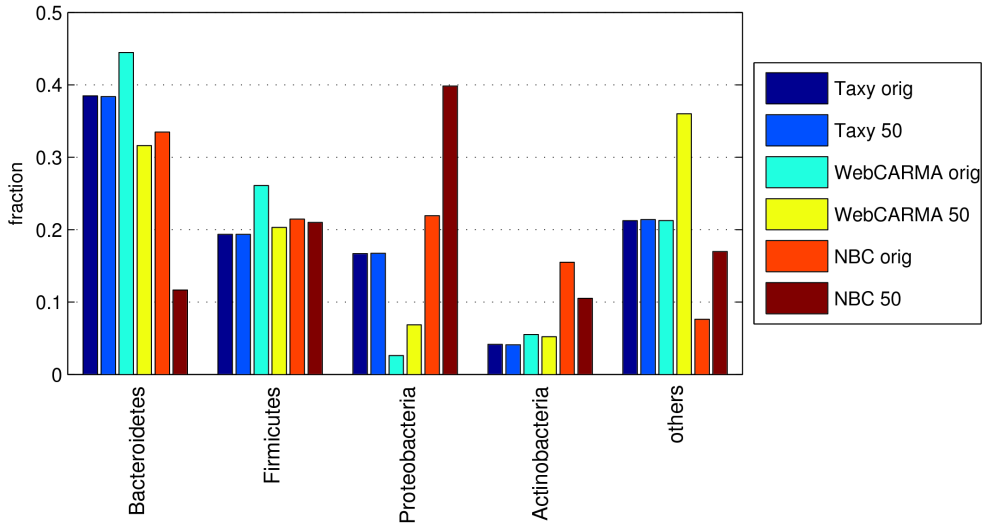
Inclusion of eukaryotes. The Taxy method can in principle be extended to profiling of mixed DNA from prokaryotic and eukaryotic organisms. As a preliminary study, we added 28 eukaryotic organisms to the signature database (see Supplementary Table 3) and performed taxonomic profiling analysis on an insect herbivore microbiome dataset [6] using Taxy, Galaxy and WebCARMA [7]. Taxy estimated a ~50% fraction of eukaryotes using the limited amount of additional signatures. On the same data, WebCARMA and Galaxy (BLAST) predicted an 70% and 83% fraction, respectively, using the full spectrum of eukaryotic genomes in current databases. Remarkably, among the eukaryotic Galaxy assignments more than 10% matched against mammalian genomes. We assume that the eukaryotic fraction predicted by Taxy would grow if additional signatures were added.

SUPPLEMENTARY FIGURES

Supplementary Figure 1: Taxonomic profile matrices for the hypersaline microbial mat metagenome [1] as obtained from the Taxy (left), Phymm (middle) and Galaxy (right) analyses. Rows correspond to different layers of the mat according to ten depth-specific samples. Columns correspond to subphylum categories. Matrix elements show color-coded abundances as indicated by the rightmost color bar.



Supplementary Figure 2: Taxonomic profiles of a human gut sample from [8] as obtained from Taxy, WebCARMA [7] and NBC [9] for original and fragmented sequences. The original sequences ("orig") comprise reads and assembled contigs with an average length of ~1250 bp while the fragmented data ("50") provided sequences with an average ~50 bp length. For an overview, only fractions of the four most dominant phyla of this sample are shown (see also Supplementary Methods).



SUPPLEMENTARY TABLES

Supplementary Table 1: Profile divergence (in percentage points, see Methods section in main manuscript) on phylum level between profiles obtained from full-length and fragmented reads for all ten depth layers of the hypersaline microbial mat [1]. The original sequence data with an average 700 bp read length were fragmented to simulate average read lengths of 350, 175, and 80 bp. Taxy was compared with Phymm [2] and Galaxy [3].

Method	Frag.Ln.	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7	Layer 8	Layer 9	Layer 10
Taxy	350 bp	0.12	0.23	0.22	0.19	0.18	0.22	0.22	0.19	0.25	0.29
	175 bp	0.25	0.32	0.32	0.42	0.32	0.49	0.26	0.31	0.22	0.47
	80 bp	0.44	0.79	0.67	0.70	0.72	0.56	0.69	0.65	0.47	0.70
Galaxy	350 bp	2.84	3.54	7.20	5.13	4.87	4.05	6.45	7.95	6.60	6.48
	175 bp	4.96	6.20	7.46	7.06	4.31	4.26	6.18	5.91	8.01	6.50
	80 bp	13.99	9.19	9.54	8.39	7.44	5.82	6.96	4.98	8.95	6.44
Phymm	350 bp	4.32	2.63	2.71	2.12	2.61	2.29	2.00	2.35	2.40	2.12
	175 bp	10.91	7.43	6.21	5.54	5.04	5.29	5.04	6.19	5.58	5.74
	80 bp	19.32	13.19	11.59	9.65	8.96	5.29	9.59	6.19	9.65	9.79

Supplementary Table 2: Comparing the profiles of full-length and fragmented sequences of 50 bp length for a human gut sample from [8]. Values indicate the deviations (in percentage points) of Taxy, WebCARMA [7] and NBC [9] for phylum level fractions.

	Taxy	WebCARMA	NBC
Actinobacteria	0.08	0.31	4.99
Bacteroidetes	0.11	12.85	21.81
Firmicutes	0.00	5.78	0.46
Proteobacteria	0.04	4.22	17.91
Others	0.15	14.72	9.36

Supplementary Table 3: List of eukaryotic organisms which have been used in the preliminary study of the insect herbivore microbiome dataset [6].

Species	Phylum
Anopheles gambiae	Arthropoda
Arabidopsis thaliana	Streptophyta
Ashbya gossypii	Ascomycota
Aspergillus clavatus NRRL1	Ascomycota
Aspergillus fumigatus Af293	Ascomycota
Aspergillus niger	Ascomycota
Aspergillus oryzae	Ascomycota
Aspergillus terreus	Ascomycota
Caenorhabditis briggsae	Nematoda
Caenorhabditis elegans	Nematoda
Candida albicans	Ascomycota
Candida glabrata	Ascomycota
Chaetomium globosum	Ascomycota
Cryptococcus neoformans	Basidiomycota
Debaryomyces hansenii	Ascomycota
Dictyostelium discoideum	Mycetozoa
Drosophila melanogaster	Arthropoda
Drosophila pseudoobscura pseudoobscura	Arthropoda
Encephalitozoon cuniculi	Microsporidia
Kluyveromyces lactis	Ascomycota
Lodderomyces elongisporus	Ascomycota
Ostreococcus lucimarinus	Chlorophyta
Ostreococcus tauri	Chlorophyta
Saccharomyces cerevisiae (budding yeast)	Ascomycota
Scheffersomyces stipitis	Ascomycota
Theileria annulata	Apicomplexa
Theileria parva	Apicomplexa
Yarrowia lipolytica	Ascomycota

SUPPLEMENTARY METHODS

Galaxy and Phymm analysis. Homology-based profiling was realized with Mega-BLAST against the NCBI nt database (as of Dec 2010) through the Galaxy web server [3]. Mega-BLAST was run with standard parameters (word length: 28, E-Value: 0.001, identity: 90 percent). For robust taxonomic profiling only diagnostic read assignments were considered, utilizing the Galaxy function “Find lowest diagnostic rank”. All hits in eukaryotic or viral genomes were disregarded. The taxonomic assignment of the reads was obtained according to the NCBI classification. For signature-based profiling the Phymm tool [2] was used with 985 organism specific models according to an installation from January 2010.

Simulated metagenome. We used the “simHC” data [10] to compare the accuracy of different profiling methods. The sequence data as obtained from the “FAMeS” web site (http://fames.jgi-psf.org/) included 116,771 Sanger sequencing reads with an approximate average read length of 1000 base pairs (bp) from a collection of 113 organisms. To simulate the realistic situation that very close homologs to genomic sequences in current databases are hardly found in

metagenomic data, we removed overlaps up to genus level between training or reference organisms and the 113 simHC organisms. The removal of genus level overlaps was applied to yield the same 654 remaining reference/training organisms for Taxy and Phymm. To ensure comparability, also Galaxy was forced to use only BLAST hits to these reference organisms for taxonomic profiling.

Read length dependence. To analyze the profile divergence arising from a varying read length we fragmented full-length (Sanger) sequencing reads to simulate data sets with a reduced average read length. The divergence was measured by comparing the profile from the original sequences with that of the reduced length sequences. The complete original sequences were split in a way to ensure the closest approximation of all fragments to the simulated target read length without generating too short sequences. Target fragment lengths for the microbial mat data [1] were 350, 175 and 80 bp. Profiles of fragmented and original reads were compared at the phylum level using 32 prokaryotic phyla and Proteobacteria classes.
For the human gut data set (sample "In-D, Healthy Human Adult Male" from [8]), where we used a 50 bp fragment length for comparison, we limited our analysis to the four most dominant phyla (Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria). Values for the remaining phyla were subsumed in the category "others" (see supplementay Figure 1 and Table 2). Note that the human gut data set not only contains the assembled contigs but also a considerable number of original reads which could not be assembled. Because of the widely varying sequence length we compared the amount of DNA (bp) attributed to phylum level categories and not the number of sequences assigned to these categories. However, this procedure still underestimates the abundant phyla since it does not take into account the number of original reads which have been assembled. Unfortunately, we did not have the original data to determine the number of assembled reads for a particular contig. Therefore, our simulation of the corresponding short read data only provides a coarse approximation of the profile variation that results from the use of assembled data for taxonomic profiling.

Taxy tool prototype. Besides the platform independent Taxy toolbox for Matlab/Octave we also provide a freely available tool implementation for Microsoft Windows (XP and higher, http://gobics.de/peter/taxy). This tool also includes precomputed taxonomic profiles of 256 metagenomes based on sequence data obtained from the CAMERA website [11]. The total size of the data is 25.8 Gb, and it took the Taxy method about eight minutes to process all of the sequences on a single standard PC (Intel 2.66 GHz).
Within the tool the taxonomic profiles can be used for comparative analysis. The program functionality allows the user to inspect the sample meta data and the taxonomic profile as estimated by the mixture modeling approach. For the phylum, class and order levels, the user can display the taxonomic profile as a bar plot. On the species level, the predicted reference genome weights can be further analyzed by means of an exportable list. The user can load metagenomic DNA sequence files of any size in multiple FASTA format. User-supplied metagenomes are integrated into the list of compared metagenomes.
In addition to the estimated taxonomic profile, Taxy also displays a hierarchical clustering for a comparison of the user-supplied metagenome with the preloaded collection of database samples. For the profile clustering distance on the phylum level, the user can choose between city-block and Euclidian metrics. The clustering algorithm can perform either complete or average linkage. The clustering is displayed in a tree with the user-supplied metagenome highlighted. Compared samples in the tree can be selected either individually or on the basis of a FOU error threshold (see Methods section in main manuscript).

References

1. V. Kunin and J. Raes and J.K. Harris and J.R. Spear and J.J. Walker and N. Ivanova and C. von Mering and B.M. Bebout and N.R. Pace and P. Bork and P. Hugenholtz Millimeter-scale genetic gradients and community level molecular convergence in a hypersaline microbial mat. *Molecular Systems Biol.* **4**, 198 (2008).

2. Brady, A. and Salzberg, S. L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* **6**, 673--676 (2009).
3. Kosakovsky Pond, S. and Wadhawan, S. and Chiaromonte, F. and Ananda, G. and Chung, W. Y. and Taylor, J. and Nekrutenko, A. Windshield splatter analysis with the Galaxy metagenomic pipeline. *Genome Res.* **19**, 2144--2153 (2009).
4. Ley, R. E. and Harris, J. K. and Wilcox, J. and Spear, J. R. and Miller, S. R. and Bebout, B. M. and Maresca, J. A. and Bryant, D. A. and Sogin, M. L. and Pace, N. R. Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. *Appl. Environ. Microbiol.* **72**, 3685--3695 (2006).
5. Spear, J. R. and Ley, R. E. and Berger, A. B. and Pace, N. R. Complexity in natural microbial ecosystems: the Guerrero Negro experience. *Biol. Bull.* **204**, 168--173 (2003).
6. Suen G., Scott J.J., Aylward F.O., Adams S.M., Tringe S.G., Pinto-Tomás A.A., Foster C.E., Pauly M., Weimer P.J., Barry K.W., Goodwin L.A., Bouffard P., Li L., Osterberger J., Harkins T.T., Slater S.C., Donohue T.J., Currie C.R. An insect herbivore microbiome with high plant biomass-degrading capacity. *PLoS Genetics* **6**, (2010).
7. Gerlach W, Jünemann S, Tille F, Goesmann A, Stoye J. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads.. *BMC Bioinformatics* **10**, (2009).
8. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP, Taylor TD, Noguchi H, Mori H, Ogura Y, Ehrlich DS, Itoh K, Takagi T, Sakaki Y, Hayashi T, Hattori M. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Research* **6**, 169-81 (2007).
9. Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* **27**, 127-9 (2011).
10. Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A.C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P., Hugenholtz, P., Kyrpides, N.C. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods* **4**, 495--500 (2007).
11. Seshadri, R. and Kravitz, S. A. and Smarr, L. and Gilna, P. and Frazier, M. CAMERA: a community resource for metagenomics. *PLoS Biol.* **5**, e75 (2007).