MicroCOSM: Microbial Clade-Oriented Sequence Markers for Phylogenetic Classification of Metagenomic Data

Dylan Chivian[1,2,*], Paramvir S. Dehal[1,2], Adam P. Arkin[1,2,3] (*DCChivian@lbl.gov, [1]VIMSS, [2]LBNL, [3]UC, Berkeley)

ESPP2 Environmental Stress Pathway Project

DOE GENOMICS:GTL ACCELERATING DISCOVERY FOR ENERGY AND ENVIRONMENT — OFFICE OF SCIENCE U.S. DEPARTMENT OF ENERGY

VIMSS Virtual Institute for Microbial Stress and Survival | Massachusetts Institute of Technology | University of Washington | Montana State University | Southern Illinois University Carbondale | Sandia National Laboratories | Oak Ridge National Laboratory | Berkeley University of California

http://vimss.lbl.gov/

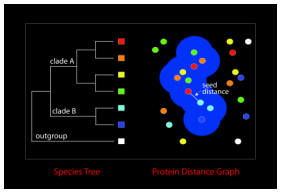
ABSTRACT

The VIMSS/ESPP2 project requires understanding of the microbial communities at contaminated field sites and, among other methods, will employ metagenomics in this endeavor. Metagenomics projects that seek to elucidate the population structure of microbial ecosystems are faced with the related computational challenges of classifying the sequences obtained and quantifying which organisms are present within a sample. Individually low-proportion species usually make up a large fraction of microbial communities, complicating their classification and quantification using traditional phylogenetic marker approaches. Such species usually don't yield sufficient read depth to assemble into longer sequences, leaving fragments that rarely contain traditional markers such as the small subunit (SSU) rRNA gene. BLAST-based approaches for analysis of metagenomic sequences [1] compensate for this rarity of traditional markers, but may be confounded by genes that are subject to horizontal transfer or duplication. Another approach instead makes use only of reliable non-transferred single-copy genes [2] to classify and quantify the organisms present within a sample, but the application has so far been limited to the use of a fairly small set of universal genes found in all organisms. In this work, we have extended the latter approach, boosting the set of reliable marker genes from only about 30-40 universal genes to several hundred by identifying sets of single-copy genes that are not subject to inter-clade horizontal transfer through investigation of finished bacterial and archaeal genomes. These clade-oriented sequence markers allow for a method, which we have named "MicroCOSM", that greatly increases the probability that a marker will be found in any given sequence and therefore offers improved coverage for phylogenetic classification and quantification of microbial types in an environmental sample.

[1] Huson D.H., Auch A.F., Qi J., Schuster S.C. (2007) "MEGAN analysis of metagenomic data." *Genome Res.* **17**(3):377-86.
[2] von Mering C., Hugenholtz P., Raes J., Tringe S.G., Doerks T., Jensen L.J., Ward N., Bork P. (2007) "Quantitative phylogenetic assessment of microbial communities in diverse environments." *Science* **315**(5815):1126-30.

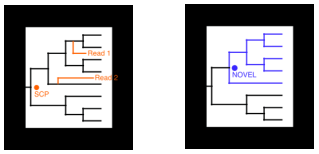
DEVELOPMENT OF COSMs

Clade-oriented sequence marker (COSM) gene families are built by clustering of BLAST-detected homologous sequences. Clusters must not include genes that belong to species outside of the clade, determined by the largest nearest-neighbor distance between any two members of the cluster.



SPECIAL COSM TYPES: SCP AND NOVEL

While the COSMs may be used to assign membership to a clade, it is single-copy prevalent (SCP) genes that are best for branch placement within the clade, as they are not subject to duplication nor horizontal transfer. Additionally, novel families (NOVEL) that are only observed within a given clade may represent the introduction of protein families to that lineage.



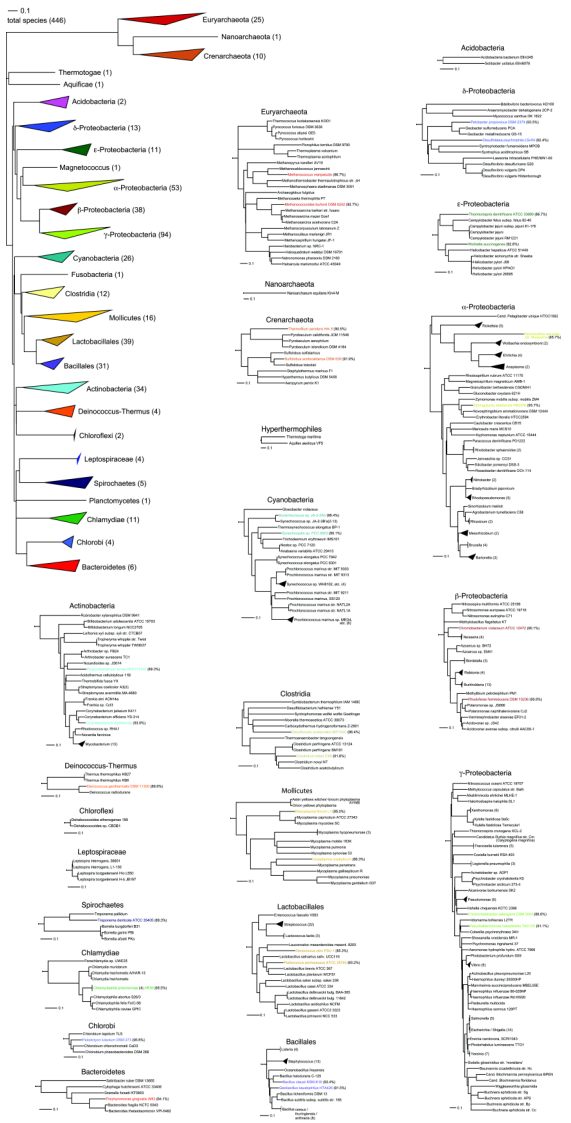
SCAN METAGENOMIC SEQUENCE WITH COSMs

In order to scan the metagenomic sequence with the clade-oriented sequence markers, profiles must be built and thresholds for membership in the protein family determined.

1. Make a multiple sequence alignment (MSA) of COSM proteins and build profile.
2. Determine lowest-scoring sequence for membership threshold.
3. Scan metagenomic reads with COSM profile and accept hits above threshold.
4. If COSM is also single copy prevalent (SCP), then fit into tree.

SPECIES TREE

Markers were developed from 446 species of Archaea and Bacteria with complete genomes. The 31 species that were removed for testing are shown in color.



COSM COUNTS

COSMs, SCP and Novel counts for several high level taxonomic groups (HLTG), specific to that node (not including children counts). Counts with the requirement of presence within >60% and >75% of clade member species are also reported.

High Level Taxonomic Group (HLTG)	COSM total	COSM >60% present	COSM >75% present	SCP >60% present	SCP >75% present	Novel total	Novel >60% present	Novel >75% present
Bacteria + Archaea	N/A	N/A	N/A	54	37	N/A	N/A	N/A
Archaea	1218	206	135	152	101	308	66	55
* Euryarchaeota	328	23	17	16	11	94	7	7
* Crenarchaeota	400	58	32	36	19	109	15	11
Bacteria	738	184	134	83	63	180	40	32
* Hyperthermophiles	27	27	27	26	26	1	1	1
* Proteobacteria + Acido	1205	20	14	0	0	195	1	0
** Acidobacteria	397	397	397	317	317	103	103	103
** Proteobacteria	2283	35	11	1	0	557	2	1
*** Delta	211	16	8	6	2	27	1	1
*** Epsilon	336	144	118	13	2	101	40	35
*** Alpha	114	32	24	1	0	20	6	4
*** Beta + Gamma	2647	28	15	2	0	895	4	2
**** Beta	151	24	19	0	0	35	4	3
**** Gamma	421	13	9	10	7	121	3	2
* Cyanobacteria	522	140	122	108	96	250	80	68

High Level Taxonomic Group (HLTG)	COSM total	COSM >60% present	COSM >75% present	SCP >60% present	SCP >75% present	Novel total	Novel >60% present	Novel >75% present
* Firmicutes	858	18	12	13	4	285	7	5
** Clostridia	224	24	11	9	4	47	5	1
** Mollicutes	8	0	0	0	0	0	0	0
** Bacilli	567	34	23	27	17	207	10	4
*** Lactobacillales	332	21	9	18	7	97	5	2
*** Bacillales	172	22	15	13	11	60	10	7
* Supergroup 1	206	13	9	5	1	23	0	0
** Actinobacteria	159	13	8	4	3	17	1	1
** Deinococcus-Thermus	207	200	193	175	152	37	37	36
** Chloroflexi	366	366	366	340	340	112	112	112
* Supergroup 2	186	4	1	4	0	33	0	0
** Spirochaetes + Lepto	46	28	17	24	15	10	4	2
*** Leptospiraceae	1159	1157	1037	1126	1013	498	448	448
*** Spirochaetes	68	65	65	63	63	19	18	18
** Chlamydiae	174	171	163	166	159	52	50	45
** Bacteroidetes-Chlorobi	100	34	27	20	18	17	2	2
*** Chlorobi	262	225	180	203	149	73	60	52
*** Bacteroidetes	92	41	30	19	10	22	9	4

COVERAGE AND ACCURACY

We removed 31 species that were of varying distance by 16S rRNA gene similarity to the closest remaining species and built a test set of COSMs. We then examined the fraction of each of these 31 genomes that could accurately be assigned to high-level taxonomic groups (e.g. phyla) using 16S rRNA genes, the 31 COGs used by von Mering *et al.*, our test COSMs and SCP-COSMs (both with >75% prevalence requirement), and a BLAST comparison with the proteomes of the remaining species. We also examined more fine-grained taxonomic assignment at the sister level with BLAST as the only method roughly comparable to phylogenetic placement.

Species	HLTG	Nearest Relative 16S(%)	Genome Size (Mbp)	16S Cov(%)	31COG Cov(%)	SCP HLTG Cov(%)	SCP HLTG Acc(%)	COSM HLTG Cov(%)	COSM HLTG Acc(%)	BLAST HLTG Cov(%)	BLAST HLTG Acc(%)	BLAST Sister Cov(%)	BLAST Sister Acc(%)
Methanococcus maripaludis	Euryarchaeota	86.7	1.66	0.25	1.5	6.6	100	4.3	99.1	85.6	87.5	85.6	66.5
Methanococcoides burtonii DSM 6242	Euryarchaeota	92.7	2.58	0.17	1	5.4	98.2	6	88.9	77	88.8	77	80.1
Thermofilum pendens Hrk 5	Crenarchaeota	90.5	1.81	0.083	1.4	5	99.2	2.6	92.5	75.6	55.6	75.6	21.4
Sulfolobus acidocaldarius DSM 639	Crenarchaeota	93.8	2.23	0.066	1.2	14.9	99.3	14.8	98.7	81.5	94.8	81.5	66.1
Desulfotalea psychrophila LSv54	Delta	82.4	3.66	0.45	0.59	3.3	93.1	1.9	61.5	78.9	42.3	78.9	17.5
Pelobacter propionicus DSM 2379	Delta	93.5	4.24	0.15	0.53	4.2	95.7	2.7	75.1	82.9	66.9	82.9	55.9
Thiomicrospira denitrificans ATCC 33889	Epsilon	86.7	2.2	0.26	1	4.2	92.7	6.3	91.5	88.4	53.6	88.4	40
Wolinella succinogenes	Epsilon	92.8	2.11	0.42	1.1	3.7	100	1.1	88	90.9	64.9	90.9	43.9
Neorickettsia sennetsu str. Miyayama	Alpha	85.7	0.86	0.17	2.6	7.7	100	1	100	75	91.8	75	72.9
Sphingopyxis alaskensis RB2256	Alpha	93.7	3.37	0.044	0.7	4.7	98	3.3	95.5	86.6	72.8	86.6	61.1
Chromobacterium violaceum ATCC 12472	Beta	90.1	4.75	0.25	0.48	2.2	84.8	0.7	59	83	56.6	83	11.4
Rhodoferax ferrireducens DSM 15236	Beta	95.8	4.97	0.062	0.47	2.1	89.3	1.8	71.8	86.6	78.1	86.6	40.8
Chromohalobacter salexigens DSM 3043	Gamma	88.6	3.7	0.2	0.61	3	97.2	1.2	74.7	87	71.2	87	10.6
Pseudoalteromonas haloplanktis TAC125	Gamma	91.1	3.85	0.36	0.59	3.3	93.3	1.2	77.7	84	87.8	84	24.2
Synechococcus sp. PCC 7803	Cyanobacteria	90.1	3.95	0.15	0.58	4.1	98.3	3.5	93.6	81.6	83.4	81.6	62.3
Synechococcus sp. JA-3-3Ab	Cyanobacteria	96.4	2.93	0.09	0.79	5.7	98.3	3.7	95.7	83.3	97.9	83.3	95.2
Desulfurudis audaxviator MP104C	Clostridia	86.4	2.35	0.14	0.95	3.9	95.5	0.44	44.7	80.7	65.3	80.7	17.2
Clostridium tetani E88	Clostridia	91.6	2.87	0.63	0.74	3.5	94.4	1.2	67.4	81	82.1	81	38.2
Ureaplasma urealyticum	Mollicutes	86.3	0.75	0.79	3	5.9	97.1	0.25	29.8	80.7	83.4	80.7	33.9
Mesoplasma florum L1	Mollicutes	95	0.79	0.19	2.9	9.8	98.3	3.8	95.7	86.6	83.2	86.6	77.8
Oenococcus oeni PSU-1	Lactobacillales	85.3	1.78	0.18	1.3	5	100	1.3	88	79	91.1	79	46.1
Pediococcus pentosaceus ATCC 25745	Lactobacillales	93.2	1.83	0.43	1.3	4.9	100	1.2	100	84.9	92.6	84.9	54.1
Geobacillus kaustophilus HTA426	Bacillales	91.5	3.59	0.39	0.62	5.1	96.1	3.5	80.7	80.9	77.3	80.9	59.4
Bacillus clausii KSM-K16	Bacillales	93.4	4.3	0.25	0.55	3.7	97	2.6	91	83.1	82.8	83.1	47.5
Propionibacterium acnes KPA171202	Actinobacteria	89.3	2.56	0.18	0.91	5.1	98.4	2.4	96.7	82.8	81.4	82.8	22.8
Corynebacterium diphtheriae	Actinobacteria	93.8	2.51	0.56	0.94	7.2	92.9	6.7	98.8	83.8	91.1	83.8	68.2
Deinococcus geothermalis DSM 11300	Deinococcus-Thermus	89.8	3.04	0.2	0.76	7.2	97.5	5.7	88.6	87	78.6	87	73.5
Treponema denticola ATCC 35405	Spirochaetes	89.3	2.84	0.11	0.82	3.8	97.4	3.1	48.1	80.7	39	80.7	37.5
Chlamydophila pneumoniae AR39	Chlamydiae	95.5	1.23	0.12	1.9	26.1	100	22.7	100	82.3	97.6	82.3	96.3
Pelodictyon luteolum DSM 273	Chlorobi	95.6	2.36	0.12	0.96	8.2	95.8	7.1	90.2	87.3	82	87.3	68.9
Porphyromonas gingivalis W83	Bacteroidetes	84.1	2.34	0.25	0.98	3.5	93.7	0.58	61.1	76	84	76	77
AVERAGE		90.7	2.71	0.25	1.1	5.7	96.5	3.8	88	82.7	64.9	82.7	51.2

Percent of genome classified: CORRECTLY / INCORRECTLY

CONCLUSIONS

We have improved the coverage provided by traditional universal sequence markers by identifying clade-oriented sequence markers. Using stringent requirements for inclusion in the gene family, coverage increases from the ~1% of universal COGs to approximately 3-6% with COSMs, accompanied by a small loss in accuracy. Greater coverage but less accuracy is expected using looser thresholds. Comparison with pair-wise BLAST-based methods shows COSMs to be far more accurate, albeit with less coverage. Lastly, unlike BLAST-based methods, SCP-COSMs permit placement explicitly on the tree, and permit greater coverage with which to access the population structure of a microbial community than universal COGs alone. We expect that more comprehensive sequencing in poorly sampled clades, such as Clostridia, will improve results in future versions of the method.

ACKNOWLEDGEMENTS

ESPP2 (MDCASE) is part of the Virtual Institute for Microbial Stress and Survival (VIMSS) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL Program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.