Supplementary Table 5

Supplementary Table 5: Search for the best parameter settings for the SOM and TETRA-method. The results of the SOM and the binning of fragments by tetranucleotide-based correlation coefficients for the dominant sample populations are shown for the Sargasso sea reference set of ssu rRNA assigned high-coverage contigs (and contigs linked to those via a common scaffold). Dark blue bars indicate the determined parameter setting. For the TETRA-method, a correlation coefficient of 0.8 was chosen as the cut-off, the lowest setting with the same specificity as *PhyloPythia*. For the SOM, assigning a phylotype based on the overall counts for this phylotype for all 1kb fragments of a contig was found to fare slightly better. The SOMs results give counts for 25 manually defined phylotypes found at the position of the SOM where an analyzed 1 kb fragment was placed. The 25 phylotype groupings of the SOM are non-overlapping groups, but at different taxonomic ranks of NCBI Taxonomy. For a comparison with our technique, we give the results derived for contigs at class level for the subset of sequences, where it applies, and at phylum level, where for each lower-level clade a parent is defined. '# Contigs' gives the number of contigs that are part of each clade in the reference. The notation 'x (y)' denotes for each of the clades the number of correctly assigned fragments *x* and the number of false positive assignments *y*. 'A' denotes the percentage of correctly assigned items (note that unassigned items are counted 'false'), and *Sp.* the specificity of assignments. For the higher-level *Sp.*-values, assignments are counted as 'false' that are not assigned to a node at a level, but are annotated with a parent node that disagrees with the current level assignments (e.g. Proteobacteria as an assignment for an archaeal node).

Dominant sample pops.	**A**	**Sp. (%)**	**Prochlorococcus**
# Contigs	46		8
TETRA*0.6	0.50	82	1 (0)
TETRA*0.7	0.46	95	0 (0)
TETRA*0.8	**0.39**	**100**	0 (0)
TETRA*0.9	0.20	100	0 (0)

Class-level	**A**	**Sp. (%)**	**Cyanobacteria**
# Contigs	46		8
$\mathrm{SOM}_{\mathrm{byContigCounts}}$	**0.2**	100	0 (0)
$\mathrm{SOM}_{\mathrm{FragmentMajorityVote}}$	0.17	89	0 (0)
Phylum-level			**Cyanobacteria**
# Contigs	47		8
$\mathrm{SOM}_{\mathrm{byContigCounts}}$	**0.19**	82	0 (0)
$\mathrm{SOM}_{\mathrm{FragmentMajorityVote}}$	0.19	82	0 (0)

Dominant sample pops.	**Burkholderia**	**Shewanella**	**Other organisms**
# Contigs	8	13	32
TETRA*0.6	7 (0)	8 (0)	27 (n.a.)
TETRA*0.7	7 (0)	7 (0)	31 (n.a.)
TETRA*0.8	7 (0)	7 (6)	32 (n.a.)
TETRA*0.9	6 (0)	3 (0)	32 (n.a.)

Class-level	**Betaproteobacteria**	**Gammaproteobacteria**
# Contigs	8	30
$\mathrm{SOM}_{\mathrm{byContigCounts}}$	7 (0)	2 (0)
$\mathrm{SOM}_{\mathrm{FragmentMajorityVote}}$	6 (1)	2 (0)
Phylum-level	**Proteobacteria**	**Other organisms**
# Contigs	38	1
$\mathrm{SOM}_{\mathrm{byContigCounts}}$	9 (1)	0 (1)
$\mathrm{SOM}_{\mathrm{FragmentMajorityVote}}$	9 (1)	0 (1)