The following supplement accompanies the article

Marine bacterioplankton consortia follow deterministic, non-neutral community assembly rules

Kevin L. Vergin, Nicholas Jhirad, Jonathon Dodge, Craig A. Carlson, Stephen J. Giovannoni*

*Corresponding author: steve.giovannoni@oregonstate.edu

Aquatic Microbial Ecology 79: 165–175 (2017)

Supplementary material

Table S1. Time-series samples amplified for 16S rRNA pyrosequencing as indicated by an X. The month of deepest mixing is indicated in **bold**; X indicate samples used for the analysis of residuals for non-normal distribution after seasonal differencing.

Year	Month	BATS#	Surface	200	MLD(m)
1991	8	35	X	X	18
	9	36	X	X	38
	10	37	X	X	24
	11	38	X	X	56
	12	39	X	X	86
1992	1	40	X	X	132
	2	41	X	X	**238**
	3	42	X	X	32
	4	43	X	X	40
	5	44	X	X	20
	6	45	X	X	16
	7	46	X	X	22
	8	47	X	X	24
	9	48	X	X	16
	10	49	X	X	60
	11	50	X	X	64
	12	51	X		84
1993	1	52	X		104
	2	53	X	X	132
	3	54	X	X	**210**
	4	55	X	X	106
	5	56	X	X	12
	6	57	X	X	28
	7	58	X	X	10
	8	59	X	X	32
	9	60			30
	10	61	X	X	34
	11	62	X	X	56
	12	63	X	X	58
1994	1	64		X	114
	2	65	X	X	132
1997	9	108	X	X	38
	10	109	X	X	32
	11	110	X	X	40
	12	111	X	X	102
1998	1	112	X	X	76
	2	113	X	X	144
	3	114	X	X	**212**
	3	114 A	X	X	92
	4	115 A	X	X	116
	5	116	X	X	38
	6	117	X	X	16

	7	118	X	X	16
	8	119		X	22
	9	120	X	X	40
	12	123	X	X	76
1999	1	124	X	X	108
	2	125	X	X	128
	3	126	X		122
	4	127			**208**
	5	128			
	6	129			26
	7	130			16
	8	131	X	X	30
	9	132	X	X	36
	10	133	X	X	48
	11	134	X	X	70
	12	135	X		
2000	1	136	X	X	171
	2	137	X	X	138
	2	137 A	X	X	247
	3	138	X	X	**248**
	4	139	X	X	46
	5	140	X	X	19
	6	141	X	X	23
	7	142	X		26
	8	143	X	X	9
	9	144	X	X	32
	10	145	X	X	47
	11	146	X	X	76
	12	147	X	X	96
2001	1	148	X	X	146
	2	149	X	X	91
	2	149 A	X	X	55
	3	150	X	X	127
	3	150 A	X		**158**
	4	151	X	X	52
	5	152	X	X	45
	6	153	X	X	8
	7	154	X	X	28
	8	155	X	X	22
	9	156	X	X	39
	10	157	X	X	38
	11	158	X	X	64
	12	159	X	X	88
2002	1	160	X	X	
	1	160 A	X	X	
	2	161	X	X	**158**
	2	161 A			128

	3	162	X	X	152
	3	162 A	X	X	
	4	163	X	X	49
	5	164	X	X	27
	6	165	X	X	25
	7	166	X		19
	8	167	X	X	13
	9	168	X	X	48
	10	169	X	X	35
	11	170	X	X	65
	12	171	X	X	75
2003	1	172	X	X	112
	2	173	X	X	**215**
	2	173 A	X	X	
	3	174	X	X	90
	3	174 A	X	X	104
	4	175	X	X	31
	5	176	X	X	47
	7	177	X	X	19
	7	178	X	X	17
	8	179	X	X	25
	9	180	X	X	26
	10	181	X	X	53
	12	183	X	X	99
2004	1	184 A	X	X	**193**

Depths (Surface and 200) are in meters below the surface. MLD = Mixed Layer Depth, calculated as the depth where sigma-*t* was equal to sea surface sigma-*t* plus an increment in sigma-*t* equivalent to a 0.2°C temperature decrease (Sprintall & Tomczak 1992). The month of deepest mixing is indicated in bold font. Red X's indicate samples used for the analysis of residuals for non-normal distribution after seasonal differencing. PERMANOVA analyses of the period before and after the 1994-1997 gap for both surface and 200 m samples did not show convincing evidence for sample composition differences compared to other tests of equal sized samples from different time periods although differences in NTU distributions appeared to drive some sample separation.

Table S2. NTUs with the greatest number of connections.

Network	NTU	Number of Connections	Phylogenetic group
Surface	519	243	Gammaproteobacteria
	521	215	Gammaproteobacteria
	933	203	Alphaproteobacteria
	791	200	Gammaproteobacteria
	524	191	Gammaproteobacteria
	497	188	Gammaproteobacteria/SAR86
	420	184	Gammaproteobacteria/SAR86
	455	184	Gammaproteobacteria/SAR86
200 m	455	268	Gammaproteobacteria/SAR86
	457	268	Gammaproteobacteria/SAR86
	467	255	Gammaproteobacteria/SAR86
	456	249	Gammaproteobacteria/SAR86
	420	235	Gammaproteobacteria/SAR86
	418	223	Gammaproteobacteria/SAR86
	792	209	Gammaproteobacteria
	415	208	Gammaproteobacteria/SAR86

NTU (nodal taxonomic unit) refers to nodes on the reference phylogenetic tree to which sequences were binned. Nodes were numbered sequentially so values closer together have greater phylogenetic relatedness.

Table S3. PERMANOVA results for phylogenetic groups within the ∂ matrix

	Phylum/Class		Clade	
Network	Pseudo-F[a]	Pairwise[b]	Pseudo-F	Pairwise
Surface	5.74	129/231	10.4	34/36
200 m	1.04	121/210	1.07	36/36

a – Pseudo-F statistic from the main test comparison. All pseudo-p values are < 0.001.

b – Number of significant pairwise tests compared to the total number of tests.

Table S4. Phylogenetic classifications for NTU's shown in Figure 1.

Figure 1 panel	NTU number (Range from 1 to 2840)	Clade	Phylum (Class for Proteobacteria)
A	440	SAR86 IIIa	Gammaproteobacteria
A	822	Roseobacter Oct lineage	Alphaproteobacteria
A	1085	SAR116 Ib	Alphaproteobacteria
A	1135	SAR116 IIIa	Alphaproteobacteria
A	1257	SAR11 Ib	Alphaproteobacteria
A	1531	Undesignated	Bacteroidetes
A	1611	Undesignated	Bacteroidetes
A	1786	Undesignated	Bacteroidetes
B	132	Undesignated	Gammaproteobacteria
B	510	Undesignated	Gammaproteobacteria
B	934	Undesignated	Alphaproteobacteria
B	1015	Undesignated	Alphaproteobacteria
B	1292	SAR11	Alphaproteobacteria
B	1843	SAR406	Deltaproteobacteria
B	1935	Undesignated	Verrucomicrobia
B	2499	Undesignated	Clostridiales
C	534	Undesignated	Gammaproteobacteria
C	721	OM43	Betaproteobacteria
C	1069	SAR116 Ib	Alphaproteobacteria
C	1132	SAR116 IIIa	Alphaproteobacteria
C	1583	Undesignated	Bacteroidetes
C	1625	Undesignated	Bacteroidetes
C	1651	Undesignated	Bacteroidetes
C	1956	Stramenopiles	Cyanobacteria
D	336	Undesignated	Gammaproteobacteria
D	343	Undesignated	Gammaproteobacteria
D	522	Xanthomonadales	Gammaproteobacteria
D	746	Xanthomonadales	Gammaproteobacteria
D	936	Rhizobiales	Alphaproteobacteria
D	1230	SAR11 Ia	Alphaproteobacteria
D	1918	Planctomyces	Planctomyces
D	1989	Plastid	Cyanobacteria
E	358	OM182	Gammaproteobacteria
E	455	SAR86	Gammaproteobacteria
E	535	Undesignated	Gammaproteobacteria
E	908	Undesignated	Alphaproteobacteria
E	1125	SAR116 IIIb	Alphaproteobacteria
E	1232	SAR11 Ia	Alphaproteobacteria
E	1272	SAR11 IIa	Alphaproteobacteria
E	1545	Undesignated	Bacteroidetes
F	785	Rhodocyclales	Betaproteobacteria
F	930	Undesignated	Alphaproteobacteria
F	1170	Undesignated	Alphaproteobacteria

F	1259	SAR11 Ic	Alphaproteobacteria
F	1460	SAR276	Deltaproteobacteria
F	1511	Undesignated	Deltaproteobacteria
F	1893	Undesignated	Unclassified Bacteria
F	1919	Undesignated	Planctomyces
G	342	Undesignated	Gammaproteobacteria
G	594	Undesignated	Betaproteobacteria
G	894	Roseobacter	Alphaproteobacteria
G	906	Roseobacter	Alphaproteobacteria
G	1016	Rhodobacterales	Alphaproteobacteria
G	1165	Undesignated	Alphaproteobacteria
G	1430	SAR324 I	Deltaproteobacteria
G	1817	Undesignated	Bacteroidetes
H	409	SAR86 II	Gammaproteobacteria
H	497	SAR86 SPOTS	Gammaproteobacteria
H	852	Roseobacter	Alphaproteobacteria
H	874	Roseobacter	Alphaproteobacteria
H	1071	SAR116 Ib	Alphaproteobacteria
H	1093	SAR116 Ia	Gammaproteobacteria
H	1104	SAR116 IIb	Alphaproteobacteria
H	1240	SAR11 Ia	Alphaproteobacteria

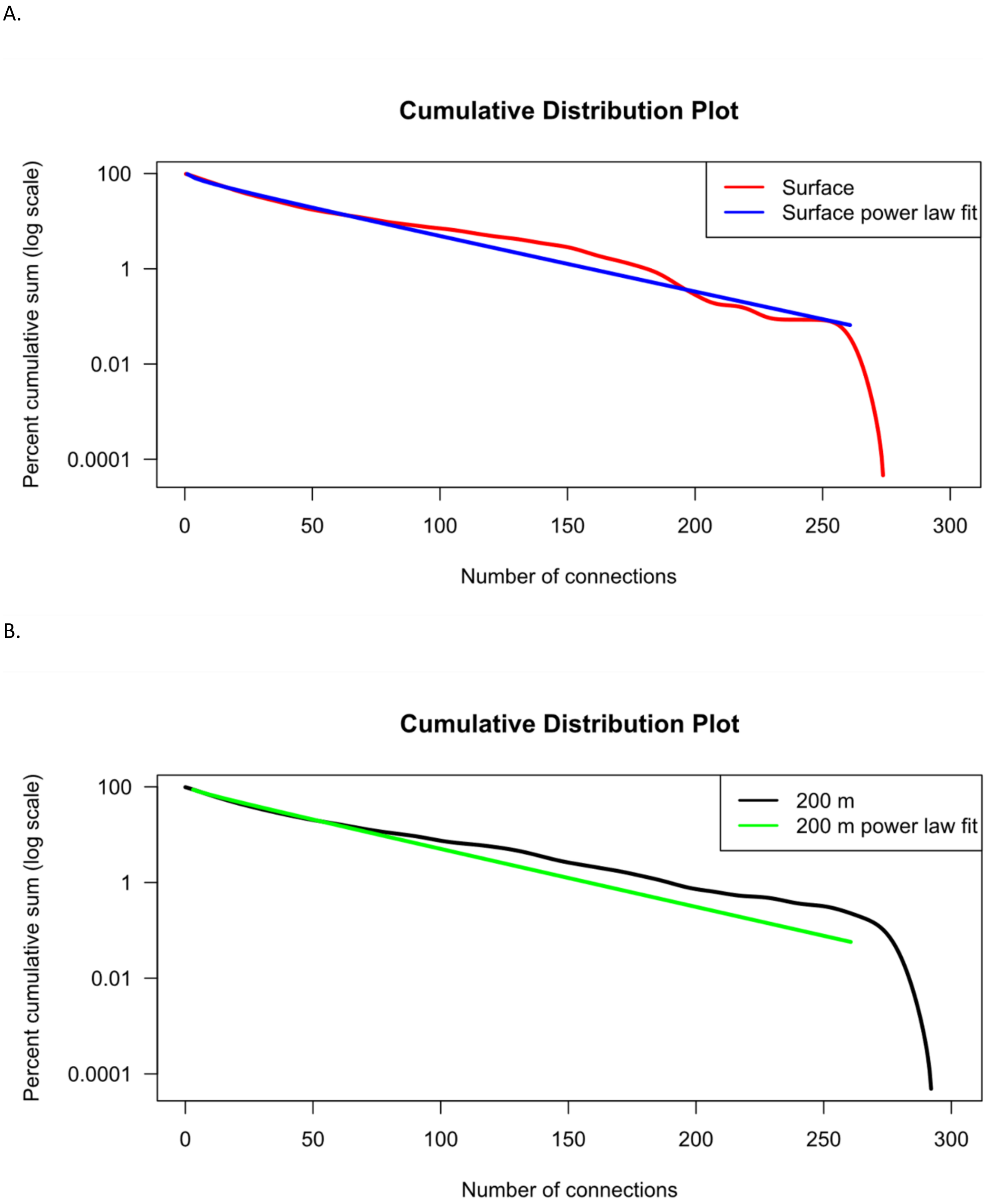

Figure S1. Cumulative distribution plots for the number of connections for each NTU (X axis). Cumulative percent is shown on the Y axis. Surface network (A) NTUs are shown in red and 200 m network (B) NTUs are shown in black. Truncated power law fits for surface (blue) and 200 m (green) are included. Power law fits were calculated from $F(x)= x^{-a}e^{-bx} + c$.

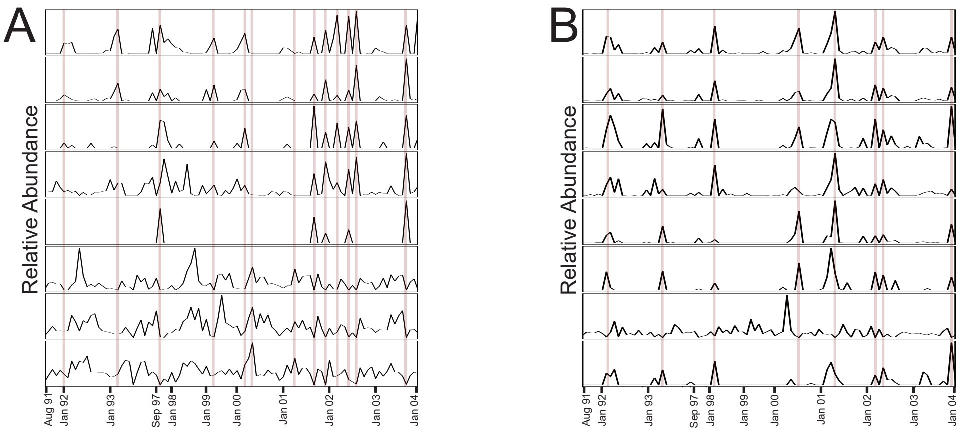

Figure S2. Relative abundance plots for the eight NTUs with the greatest number of connections at the surface (A) and 200 m (B). The most connected NTUs are frequently present in the system, especially during deep mixing events. Relative abundance of amplicon sequences (Y axis) for individual NTU's over the entire concatenated time series (X axis) are shown. The first months for each calendar year are marked on the X axis. The rose colored lines indicate periods of relative abundance for several NTUs

Supplementary methods for “Network pruning using diagnostic filtering to remove potentially spurious correlations”

In a linear regression model for the distributions of two NTU’s, the residuals should be independent and normally distributed. The distribution of the residuals may not be normal if there is seasonality or other structure in the distribution of the NTU. To simulate the extent of this non-normality, we used an R simulation (written by Charlotte Wickham, Oregon State University, and used with her permission). Two random data series are generated. The arima.sim function in the stats package can impose structure on the data by inputting values other than 0 for auto-regression or moving average. The command order = c(0,0,0) produces white noise. If X and Y are both white noise, repeating the simulation 1000 times usually results in fewer than 50 rejections (p less than 0.05) of the null hypothesis when the null hypothesis is false (type I error). ARIMA values for one data series generally still have a p value less than 0.05. However, structure in both series results in much higher p values and may result in higher confidences for spurious correlations.

```
#both white noise
one_sim <- function(){
  x <- arima.sim(list(), n = 100)
  y <- 1 + arima.sim(list(), n = 100)
  fit <- arima(y, order = c(0, 0, 0), xreg = x)

  abs( fit$coef["x"] /
     sqrt(diag(fit$var.coef)["x"]) )  1.96
}

reject <- replicate(1000, one_sim())
table(reject)
reject
FALSE  TRUE
  954   46

#one white noise, one structured
one_sim <- function(){
  x <- arima.sim(list(order = c(1,0,0), ar = 0.7), n = 100)
  y <- 1 + arima.sim(list(), n = 100)
  fit <- arima(y, order = c(0, 0, 0), xreg = x)

  abs( fit$coef["x"] /
     sqrt(diag(fit$var.coef)["x"]) )  1.96
}

reject <- replicate(1000, one_sim())
table(reject)
reject
FALSE  TRUE
  950   50
```

```
#both structured
one_sim <- function(){
  x <- arima.sim(list(order = c(1,0,0), ar = 0.7), n = 100)
  y <- 1 + arima.sim(list(order = c(1,0,0), ar = 0.9), n = 100)
  fit <- arima(y, order = c(0, 0, 0), xreg = x)

  abs( fit$coef["x"] /
      sqrt(diag(fit$var.coef)["x"]) )  1.96
}

reject <- replicate(1000, one_sim())
table(reject)
```

```
reject
FALSE  TRUE
  663   337
```

Supplementary Text

Explanation of Phylogenetically Weighted Connectivity (∂)

To understand this strategy, consider the following example. Suppose a network of university professors was made based on who knows whom. All professors have a list of professors that they know (also known as a profile). Suppose we wanted to calculate similarities between lists. A faculty member in microbiology who studies marine microorganisms works with oceanographers, and a particular oceanographer (“X”) is on their list. Now consider a faculty member in physics who studies ocean wave dynamics and works with another oceanographer (“Y”). When we calculate a similarity between the microbiologist and the physicist (who are not necessarily connected), the oceanographers (“X” and “Y”) are not a match but we can assign a weight to the relationship because they are both in the same department. Thus, the similarity between the microbiologist and physicist is higher because both have connections to oceanographers. By analogy, we infer that they play more similar roles in the university ecosystem because they both study oceanographic topics, although the overall similarities may not be high because they may not work with the same people. Subsequent analyses may use department affiliations as categories in the same way that phylogentic classifications are used for the BATS network.

Explanation of the spectral clustering strategy

To understand this approach, and to contrast it with the previous weighted approach, again consider the hypothetical university professor network. In this case, only the similarity of direct connections is considered. A cluster may be formed from the microbiologist, oceanographer “X”, a zoologist who studies coral disease, a fisheries biologist who studies marine fish ecology, and an ecologist who studies fungal interactions. This cluster could be described as a life science community, whereas the physicist might be in a separate physical science community. Both communities might include members from many different departments, and the communities would function in parallel to each other.