Supplementary File 1: Tables and Figures

Supplementary Tables

Table S1: **Data sources**

DATA SOURCE	DATA TYPE	TAXONOMIC QUANTIFICATION METHOD	RAREFACTION DEPTH	DIVERSITY MEASURE	SAMPLING PERIOD	NUMBER OF SAMPLES	REFERENCE
MICROBIS	V6 Targeted 454 Sequence	*de novo* clustered OTUs; RDP classification of reads	4266 and 150	Richness and Shannon diversity	2000-2009	234	1
POMMIER2007	16S Targeted Sanger Sequence	RDP classification of *de novo* clustered OTUs	NA	Chao richness	2001-2003	9	2
FUHRMAN2008	ARISA	ARISA OTUs	NA	Richness	1988-2006	103	3
GOS	Metagenomic Sanger Sequence	RDP classification of reads	150	Richness	2003-2005	31	4

Table S2: **Environmental covariate metadata**

VARIABLE	METRIC	UNITS	DEPTH-SPECIFIC	SOURCE
Apparent Oxygen Utilization*	Monthly Mean	mL/L	Yes	5
Calcite Concentration	Annual Mean	mol/m^3	No	6
Chlorophyll A Concentration	Annual Mean	mol/m^3	No	6
Chlorophyll A Concentration*	Annual Range	mol/m^3	No	6
Chlorophyll A Concentration	Annual Standard Deviation	mol/m^3	No	7
Chlorophyll A Concentration*	Monthly Mean	mol/m^3	No	7
Cloud Fraction	Annual Mean	%	No	6
Cloud Fraction	Annual Standard Deviation	%	No	7
Cloud Fraction	Monthly Mean	%	No	7
Day Length*	Monthly Mean	h	No	8
Diffuse Attenuation	Annual Mean	m^{-1}	No	6
Dissolved Oxygen	Monthly Mean	mL/L	Yes	5
Distance from Land	NA	km	NA	9
Dust Flux	Annual Mean	g/m^2	No	10
Dust Flux	Annual Standard Deviation	g/m^2	No	10
Dust Flux	Monthly Mean	g/m^2	No	10
Insolation*	Annual Mean	W/m^2	No	6
Insolation*	Annual Standard Deviation	W/m^2	No	7
Insolation*	Monthly Mean	W/m^2	No	7
Latitude	NA	Degrees	NA	NA
Nitrate Concentration*	Annual Mean	µmol/L	No	6
Nitrate Concentration	Annual Standard Deviation	µmol/L	No	11
Nitrate Concentration*	Monthly Mean	µmol/L	Yes	11
Nitrate:Phosphate Ratio	Monthly Mean	NA	Yes	11
Ocean Depth*	NA	m	NA	7
Ocean Temperature	Annual Mean	°C	No	6
Ocean Temperature*	Annual Range	°C	No	12
Ocean Temperature	Annual Standard Deviation	°C	No	12
Ocean Temperature*	Monthly Mean	°C	Yes	12
Oxygen Saturation	Monthly Mean	NA	Yes	5
pH	Annual Mean	NA	No	6
Phosphate Concentration*	Annual Mean	µmol/L	Yes	6
Phosphate Concentration	Annual Standard Deviation	µmol/L	No	11
Phosphate Concentration*	Monthly Mean	µmol/L	Yes	11
Proximity to Thermocline*	Monthly Mean	log(m)	No	13
Pycnocline Depth	Annual Standard Deviation	m	No	13
Pycnocline Depth	Monthly Mean	m	No	13
Salinity*	Annual Mean	PSS	No	6
Salinity*	Annual Standard Deviation	PSS	No	14
Salinity*	Monthly Mean	PSS	Yes	14
Sea Ice Concentration	Annual Mean	%	No	9
Sea Ice Concentration	Annual Standard Deviation	%	No	7
Sea Ice Concentration*	Monthly Mean	%	No	9
Thermocline Depth	Annual Standard Deviation	m	No	13
Thermocline Depth	Monthly Mean	m	No	13
Sampling Depth	NA	m	NA	NA
Sampling Year	NA	Julian Day Number	NA	NA

Table S3: **Environmental covariate correlations**

Table S4: **Environmental covariate coverage**

VARIABLE	METRIC	GLOBALLY REPRESENTATIVE SAMPLE K-S	GLOBALLY REPRESENTATIVE SAMPLE SIZE	GLOBALLY REPRESENTATIVE SAMPLE P-VALUE	PROPORTION OF AREA WITH MESS <-20
Apparent Oxygen Utilization	Monthly Mean	0.2117	234	<0.001	0.0002
Calcite Concentration	Annual Mean	0.4456	234	<0.001	0.0006
Chlorophyll A Concentration	Annual Mean	0.4804	234	<0.001	0.0002
Chlorophyll A Concentration	Annual Range	0.4666	234	<0.001	0.0005
Chlorophyll A Concentration	Annual Standard Deviation	0.1832	91	<0.005	0.0003
Chlorophyll A Concentration	Monthly Mean	0.5136	229	<0.001	0.0001
Cloud Fraction	Annual Mean	0.3181	234	<0.001	0
Cloud Fraction	Annual Standard Deviation	0.2363	234	<0.001	0.0275
Cloud Fraction	Monthly Mean	0.3203	234	<0.001	0
Day Length	Monthly Mean	0.2495	234	<0.001	0
Diffuse Attenuation	Annual Mean	0.5183	234	<0.001	0.0001
Dissolved Oxygen	Monthly Mean	0.273	234	<0.001	0
Distance from Land	NA	0.6947	234	<0.001	0.1099
Dust Flux	Annual Mean	0.4569	234	<0.001	0.0051
Dust Flux	Annual Standard Deviation	0.0365	234	NS	0.0118
Dust Flux	Monthly Mean	0.0962	234	NS	0.0093
Irradiance	Annual Mean	0.3651	234	<0.001	0
Irradiance	Annual Standard Deviation	0.4599	234	<0.001	0.0002
Irradiance	Monthly Mean	0.0797	234	NS	0
Latitude	NA	0.587	234	<0.001	0
Nitrate Concentration	Annual Mean	0.2695	234	<0.001	0
Nitrate Concentration	Annual Standard Deviation	0.4436	234	<0.001	0.005
Nitrate Concentration	Monthly Mean	0.103	234	<0.05	0.0003
Nitrate:Phosphate Ratio	Monthly Mean	0.0007	234	NS	0.0074
Ocean Depth	NA	0.5811	234	<0.001	0.0001
Ocean Temperature	Annual Mean	0.3336	234	<0.001	0
Ocean Temperature	Annual Range	0.4911	234	<0.001	0.0006
Ocean Temperature	Annual Standard Deviation	0.4549	234	<0.001	0.0024
Ocean Temperature	Monthly Mean	0.3045	234	<0.001	0
Oxygen Saturation	Monthly Mean	0.1761	234	<0.001	0.0002
pH	Annual Mean	0.1759	234	<0.001	0.0161
Phosphate Concentration	Annual Mean	0.1954	234	<0.001	0
Phosphate Concentration	Annual Standard Deviation	0.3478	234	<0.001	0.0123
Phosphate Concentration	Monthly Mean	0.1037	234	<0.05	0.0004
Proximity to Thermocline	Monthly Mean	0.3355	234	<0.001	0
Pycnocline Depth	Annual Standard Deviation	0.1627	221	<0.001	0.047
Pycnocline Depth	Monthly Mean	0.2037	234	<0.001	0.0372
Salinity	Annual Mean	0.2895	234	<0.001	0
Salinity	Annual Standard Deviation	0.1074	234	NS	0.0045
Salinity	Monthly Mean	0.2389	234	<0.001	0
Sea Ice Concentration	Annual Mean	0.0804	234	NS	0.0001
Sea Ice Concentration	Annual Standard Deviation	0.073	234	NS	0
Sea Ice Concentration	Monthly Mean	0.0353	234	NS	0
Thermocline Depth	Annual Standard Deviation	0.2955	221	<0.001	0.0346
Thermocline Depth	Monthly Mean	0.2814	234	<0.001	0.0192

Table S5: **Modeling approaches**

ANALYSIS	SEQUENCE CLASSIFICATION METHOD	MODEL	RAREFACTION DEPTH	DIVERSITY MEASURE	TAXONOMIC SCOPE	TOPIC ADDRESSED
I	De Novo	Linear	4266	Richness	Bacteria	Primary analysis
II	RDP	Linear	4266	Richness	Bacteria	Effects of classification method on diversity patterns
III	De Novo	MARS	4266	Richness	Bacteria	Effects of modeling method on diversity patterns
IV	De Novo	Linear	150	Richness	Bacteria	Effects of sampling depth on diversity patterns
V	De Novo	Linear	4266	Shannon	Bacteria	Effects of diversity metric on diversity patterns
VI	De Novo	Linear	4266	Richness	Cyanobacteria	Phylum-level diversity patterns
VII	De Novo	Linear	4266	Richness	Alphaproteobacteria	Phylum-level diversity patterns
VIII	De Novo	Linear	4266	Richness	Actinobacteria	Phylum-level diversity patterns
IX	De Novo	Linear	4266	Richness	Gammaproteobacteria	Phylum-level diversity patterns
X	De Novo	Linear	4266	Richness	Bacteroidetes	Phylum-level diversity patterns
XI	De Novo	Linear	4266	Richness	Verrucomicrobia	Phylum-level diversity patterns
XII	RDP	Logistic	4266	Relative Abundance	Prochlorococcus/ Synechococcus	Range map
XIII	RDP	Logistic	4266	Relative Abundance	Pelagibacter	Range map
XIV	RDP	Logistic	4266	Relative Abundance	Polaribacter	Range map
XV	RDP	Logistic	4266	Relative Abundance	Sphingopyxis	Range map

Table S6: **Best models, Analysis I**

NUMBER OF PARAMETERS	STATISTIC MINIMIZED	COVARIATES[1]
0	CV R^2	--
0	AIC/BIC	--
1	CV R^2	DaylengthSq
1	AIC/BIC	DaylengthSq
2	CV R^2	DaylengthSq, PhosphateAnnualSq
2	AIC/BIC	DaylengthSq, PhosphateAnnualSq
3	CV R^2	DaylengthSq, PhosphateAnnualSq, ThermoclineDistance
3	AIC/BIC	DaylengthSq, PhosphateAnnualSq, ThermoclineDistance
4	CV R^2	DaylengthSq, PhosphateAnnualSq, ThermoclineDistance, ChlorophyllRange
4	AIC/BIC	DaylengthSq, PhosphateAnnualSq, ThermoclineDistance, ChlorophyllRange
5	CV R^2	PhosphateAnnualSq, ThermoclineDistance, TempRange, Insolation, InsolationStDev
5	AIC/BIC	PhosphateAnnualSq, ThermoclineDistance, TempRange, Insolation, InsolationStDev
6	CV R^2	PhosphateAnnual, ThermoclineDistance, Insolation, InsolationAnnual, ChlorophyllRange, NitrateAnnual
6	AIC/BIC	PhosphateAnnual, ThermoclineDistance, Insolation, InsolationAnnual, ChlorophyllRange, NitrateAnnual
7	CV R^2	PhosphateAnnual, ThermoclineDistance, Insolation, InsolationAnnual, ChlorophyllRange, NitrateAnnual, Salinity
7	AIC/BIC	PhosphateAnnual, ThermoclineDistance, Insolation, InsolationAnnual, ChlorophyllRange, NitrateAnnual, Salinity
8	CV R^2	PhosphateAnnual, ThermoclineDistance, Insolation, InsolationAnnual, ChlorophyllRange, NitrateAnnual, Salinity, Chlorophyll
8	AIC/BIC	PhosphateAnnual, ThermoclineDistance, Insolation, InsolationAnnual, ChlorophyllRange, NitrateAnnual, Salinity, SalinityStDev

[1]Abbreviations
Chlorophyll: mean chlorphyll a concentration in month sampled
ChlorophyllRange: mean annual range of chlorophyll a concentration
Daylength: mean daylength in month sampled
DaylengthSq: square of mean daylength in month sampled
Insolation: mean insolation in month sampled
InsolationAnnual: mean annual insolation
InsolationStDev: annual standard deviation of insolation
Nitrate: mean nitrate concentration in month sampled
NitrateAnnual: mean annual nitrate concentration
OceanDepth: depth of ocean
Oxygen: mean oxygen utilization in month sampled
Phosphate: mean phosphate concentration in month sampled
PhosphateAnnual: mean annual phosphate concentration
PhosphateAnnualSq: square of mean annual phosphate concentration
Salinity: mean salinity in month sampled
SalinityAnnual: mean annual salinity
SalinityStDev: annual standard deviation of insolation
Seaice: sea ice concentration in month sampled
Temp: mean ocean temperature in month sampled
TempRange: mean annual range of sea surface temperature
ThermoclineDistance: log of distance between sampling depth and mean thermocline depth in month sampled

Table S7: **Selected Models**

ANALYSIS	NUMBER OF PARAMETERS	COVARIATES[1]
I	3	DaylengthSq, PhosphateAnnualSq, ThermoclineDistance
II	4	DaylengthSq, PhosphateAnnualSq, ThermoclineDistance, ChlorophyllRange
IV	3	DaylengthSq, PhosphateAnnualSq, ThermoclineDistance
V	3	DaylengthSq, PhosphateAnnualSq, SalinityStDev
VI	5	ChlorophyllRange, InsolationAnnual, PhosphateAnnual, SalinityStDev, DaylengthSq
VII	2	DaylengthSq, PhosphateAnnual
VIII	7	ChlorophyllRange, Daylength, PhosphateAnnual, Seaice, Temp, OceanDepth, Salinity
IX	4	Insolation, NitrateAnnual, SalinityAnnual, PhosphateAnnualSq
X	3	InsolationAnnual, InsolationStDev, SalinityStDev
XI	5	Daylength, InsolationAnnual, NitrateAnnual, Salinity, PhosphateAnnualSq
XII	3	InsolationAnnual, SalinityAnnual, Temp
XIII	3	Insolation, PhosphateMean, OceanDepth
XIV	3	InsolationAnnual, Temp, DaylengthSq
XV	2	Salinity, TempRange

[1]Abbreviations
Chlorophyll: mean chlorphyll a concentration in month sampled
ChlorophyllRange: mean annual range of chlorophyll a concentration
Daylength: mean daylength in month sampled
DaylengthSq: square of mean daylength in month sampled
Insolation: mean insolation in month sampled
InsolationAnnual: mean annual insolation
InsolationStDev: annual standard deviation of insolation
Nitrate: mean nitrate concentration in month sampled
NitrateAnnual: mean annual nitrate concentration
OceanDepth: depth of ocean
Oxygen: mean oxygen utilization in month sampled
Phosphate: mean phosphate concentration in month sampled
PhosphateAnnual: mean annual phosphate concentration
PhosphateAnnualSq: square of mean annual phosphate concentration
Salinity: mean salinity in month sampled
SalinityAnnual: mean annual salinity
SalinityStDev: annual standard deviation of insolation
Seaice: sea ice concentration in month sampled
Temp: mean ocean temperature in month sampled
TempRange: mean annual range of sea surface temperature
ThermoclineDistance: log of distance between sampling depth and mean thermocline depth in month sampled

Supplementary Figures

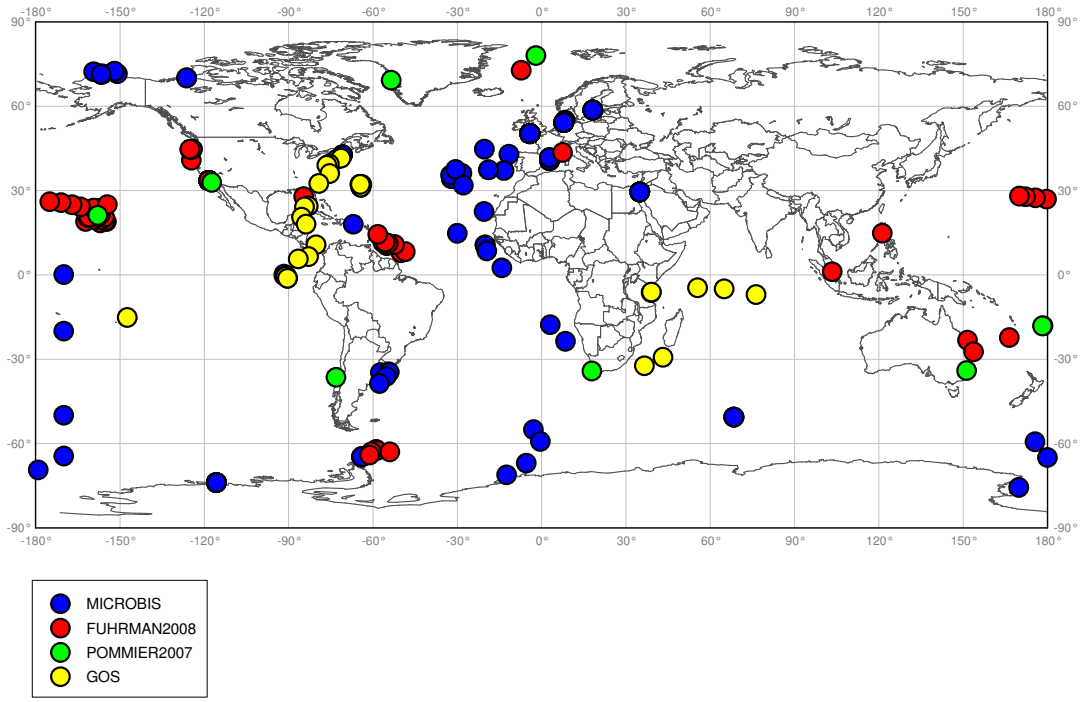


Figure S1: **Sampling locations for data used in constructing maps.** Models with zero to eight parameters were fitted using MICROBIS data. Predictive performance of the models was evaluated using both internal measures of model performance (AIC, BIC, and PRESS) and three independent data sets, collected at the locations shown in red, green, and yellow (see Table S1). Analyses were based on 377 samples (234 MICROBIS, 30 GOS, 9 POMMIER2007, 103 FUHRMAN2008) collected from 164 distinct locations.

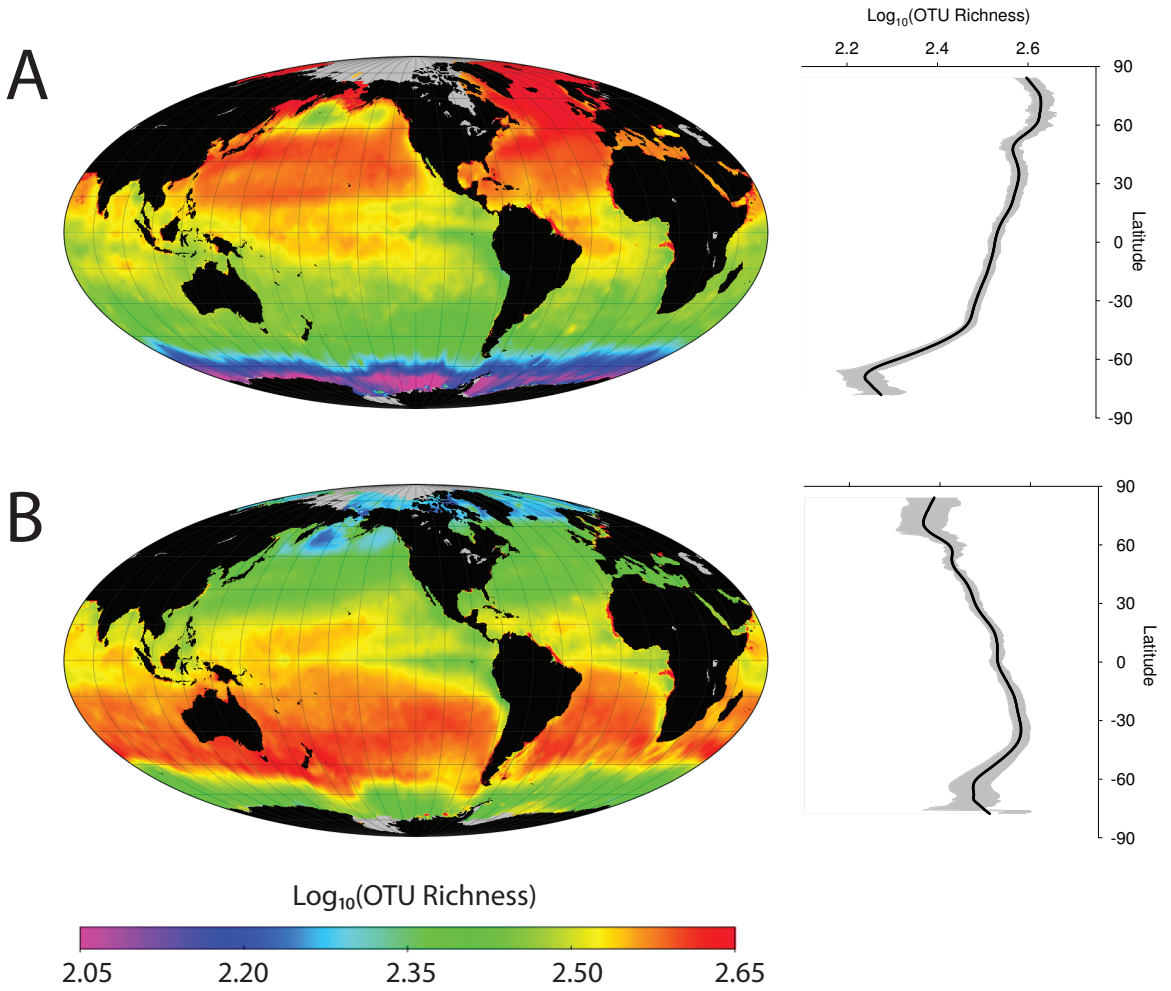


Figure S2: **Maps of global marine bacterial diversity.** Sequences are classified using RDP, and diversity is modeled using a linear model (Analysis II; see Table S5). (A) In December, OTU richness peaks in temperate latitudes in the Northern Hemisphere. (B) In June, OTU richness peaks in temperate latitudes in the Southern Hemisphere. Predicted richness during the spring and fall is intermediate, with roughly globally uniform richness near the equinoxes. These predictions hold regardless of the method used to classify reads (see Table S5 and Fig. 1)

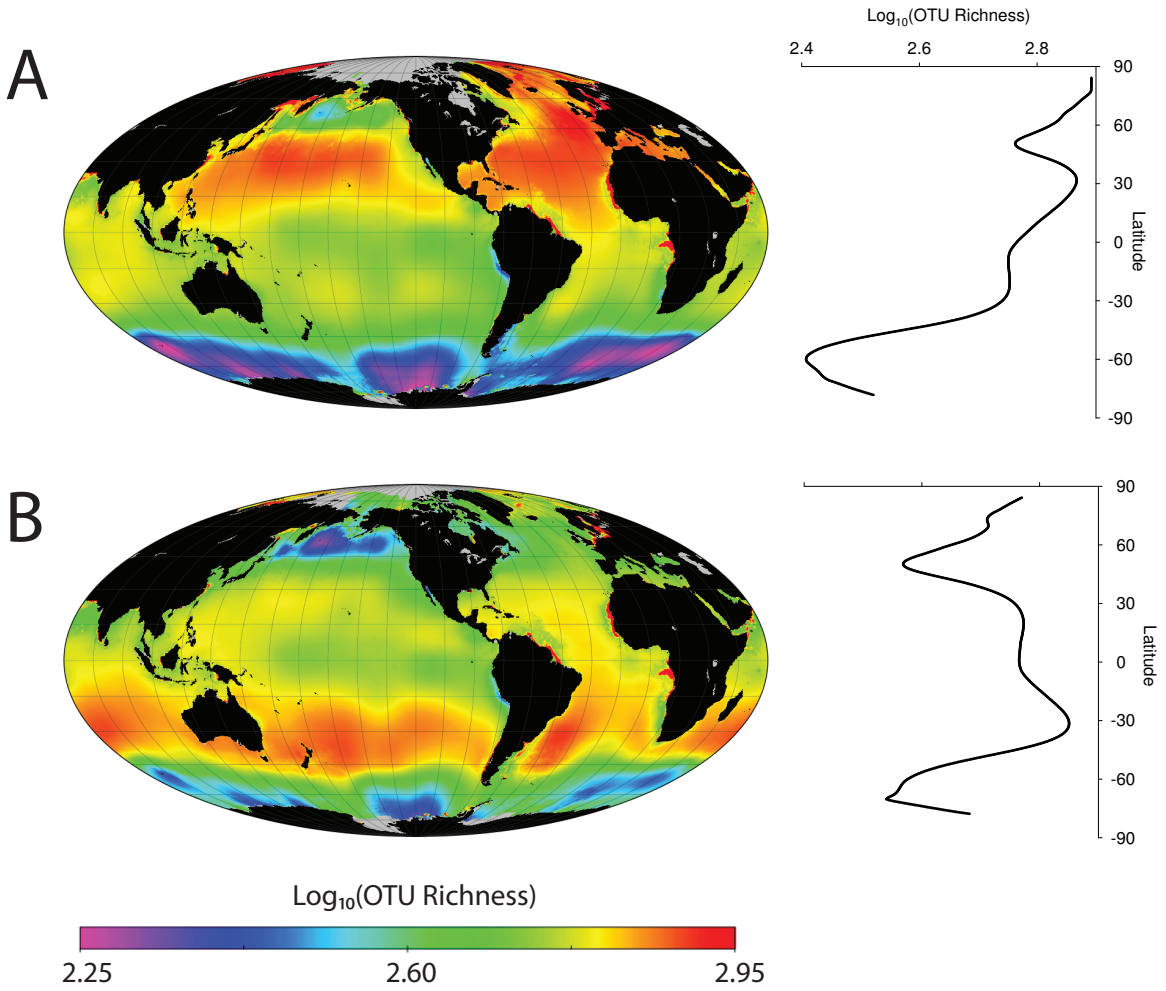


Figure S3: **Maps of global bacterioplankton diversity.** Sequences are classified using *de novo* clustering, and diversity is modeled using a multiple adaptive regression splines model (Analysis III; see Table S5). (A) In December, OTU richness peaks in temperate latitudes in the Northern Hemisphere. (B) In June, OTU richness peaks in temperate latitudes in the Southern Hemisphere. Predicted richness during the spring and fall is intermediate, with roughly globally uniform richness near the equinoxes. These predictions hold regardless of the modeling method (see Table S5 and Fig. 1).

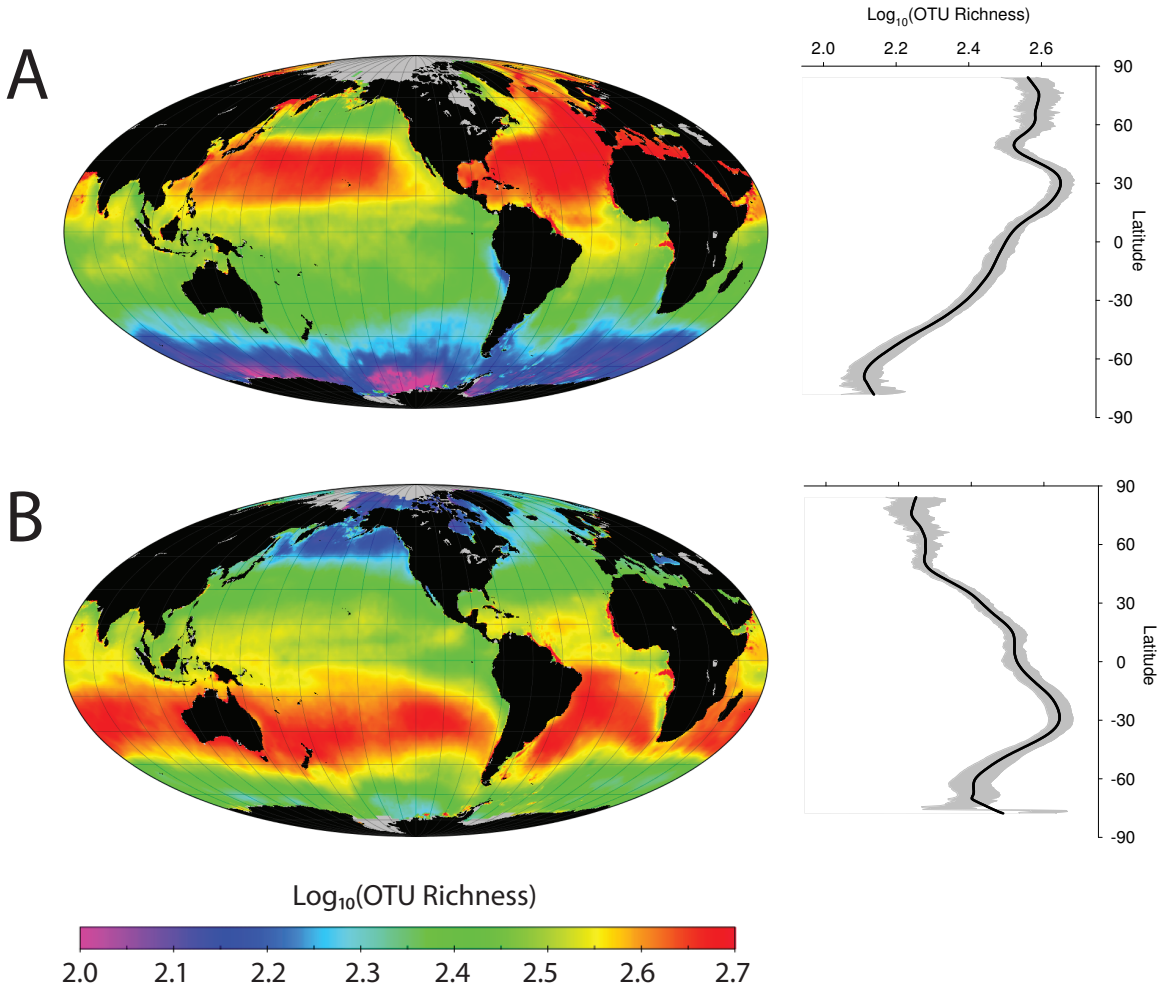


Figure S4: **Maps of global marine bacterial diversity.** Sequences are classified using *de novo* clustering and a linear model is used (see Table S5). A sub-optimal, seven-predictor model is used (Table S6 and Fig. S8). (A) In December, OTU richness peaks in temperate latitudes in the Northern Hemisphere. (B) In June, OTU richness peaks in temperate latitudes in the Southern Hemisphere. Predicted richness during the spring and fall is intermediate, with roughly globally uniform richness near the equinoxes. These predictions are robust to the exact predictors used; see the map in Fig. 1, which was generated using the optimal, three-parameter model given in Table S6.

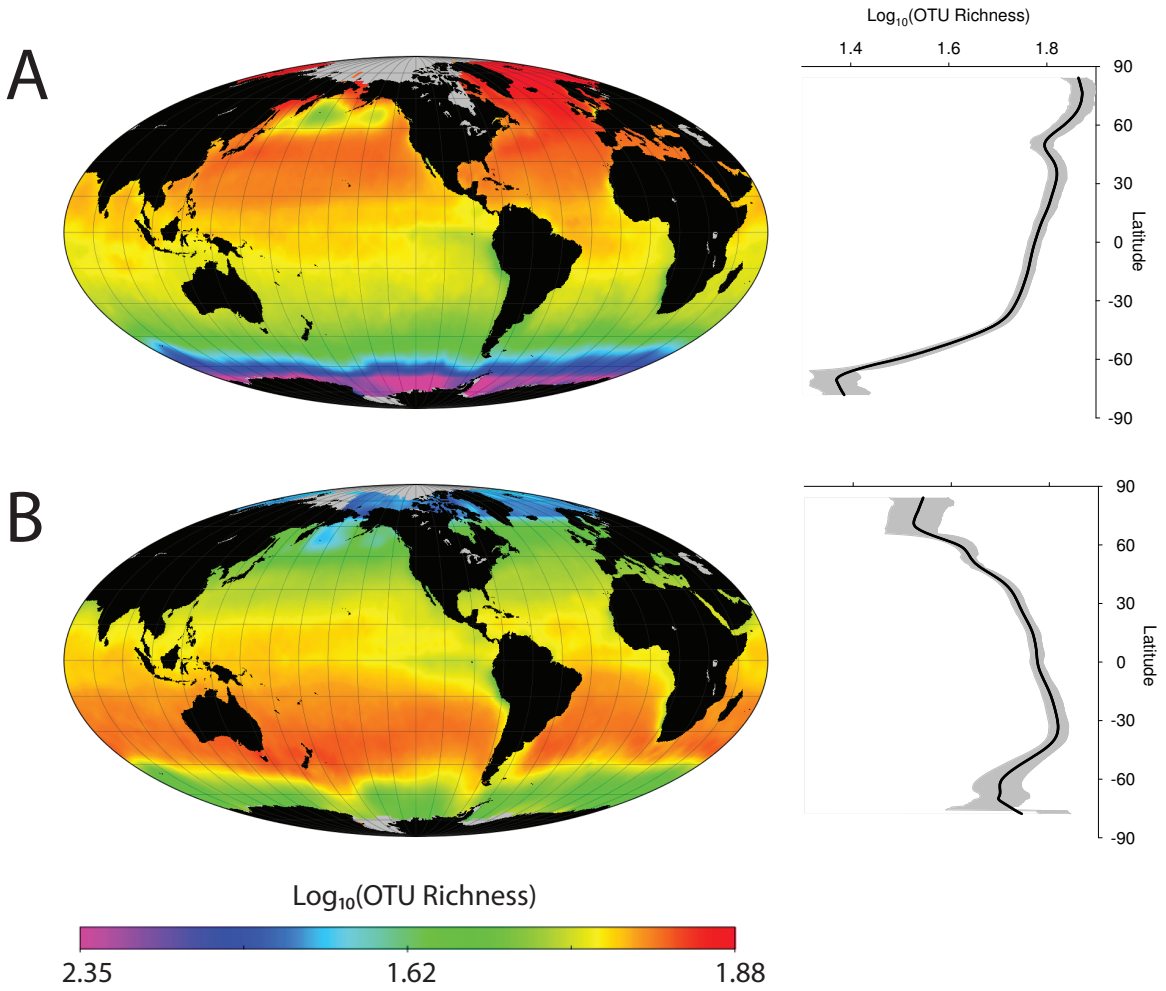


Figure S5: **Maps of global marine bacterial diversity.** Sequences are classified using *de novo* clustering, and diversity is modeled using a linear model, but sequencing depth is shallow (150 reads; Analysis IV; see Table S5). (A) In December, OTU richness peaks in temperate latitudes in the Northern Hemisphere. (B) In June, OTU richness peaks in temperate latitudes in the Southern Hemisphere. Predicted richness during the spring and fall is intermediate, with roughly globally uniform richness near the equinoxes. These predictions hold regardless of the rarefaction depth (see Table S5 and Fig. 1).

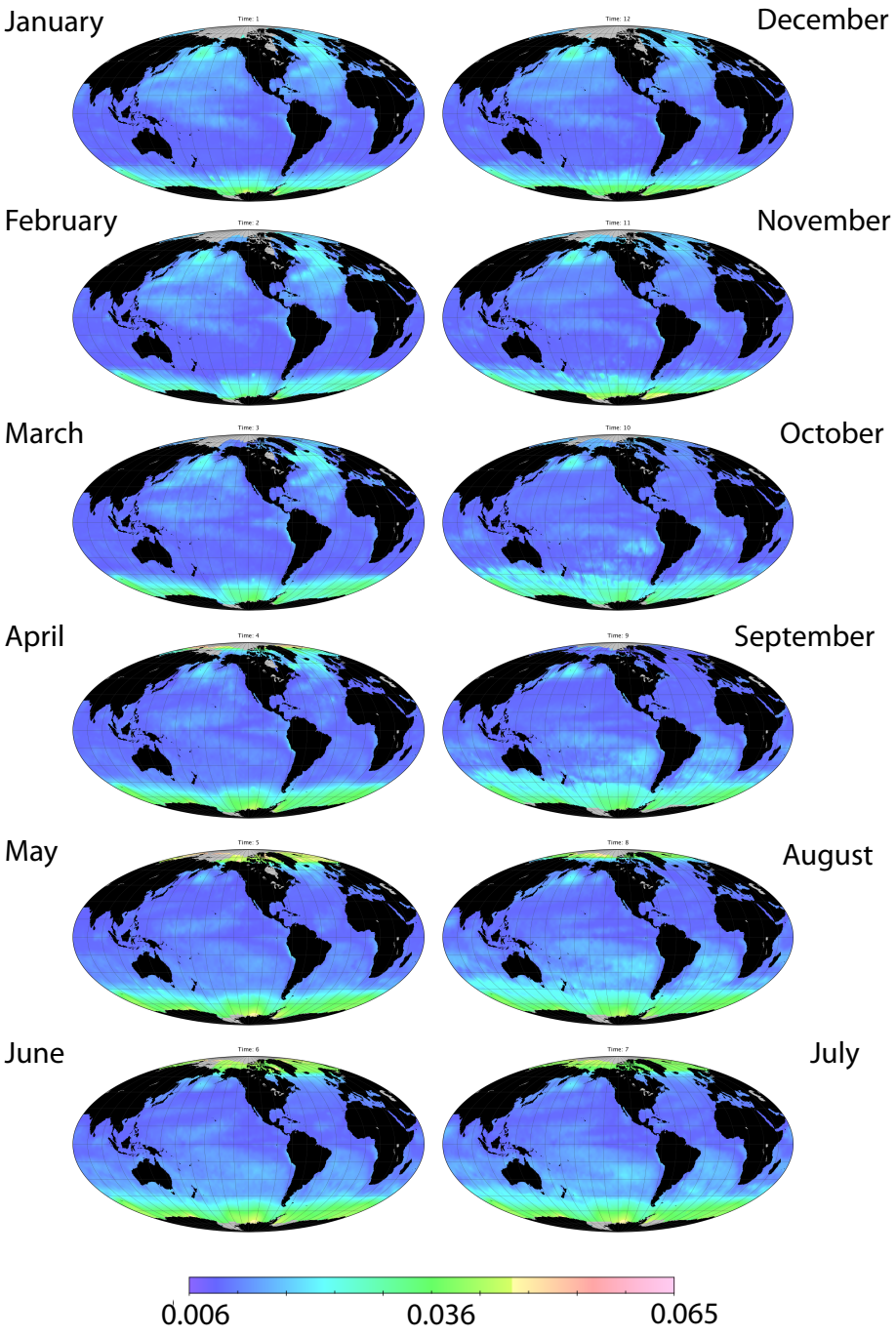


Figure S6: **Maps of standard errors of estimated richness for Analysis I.**

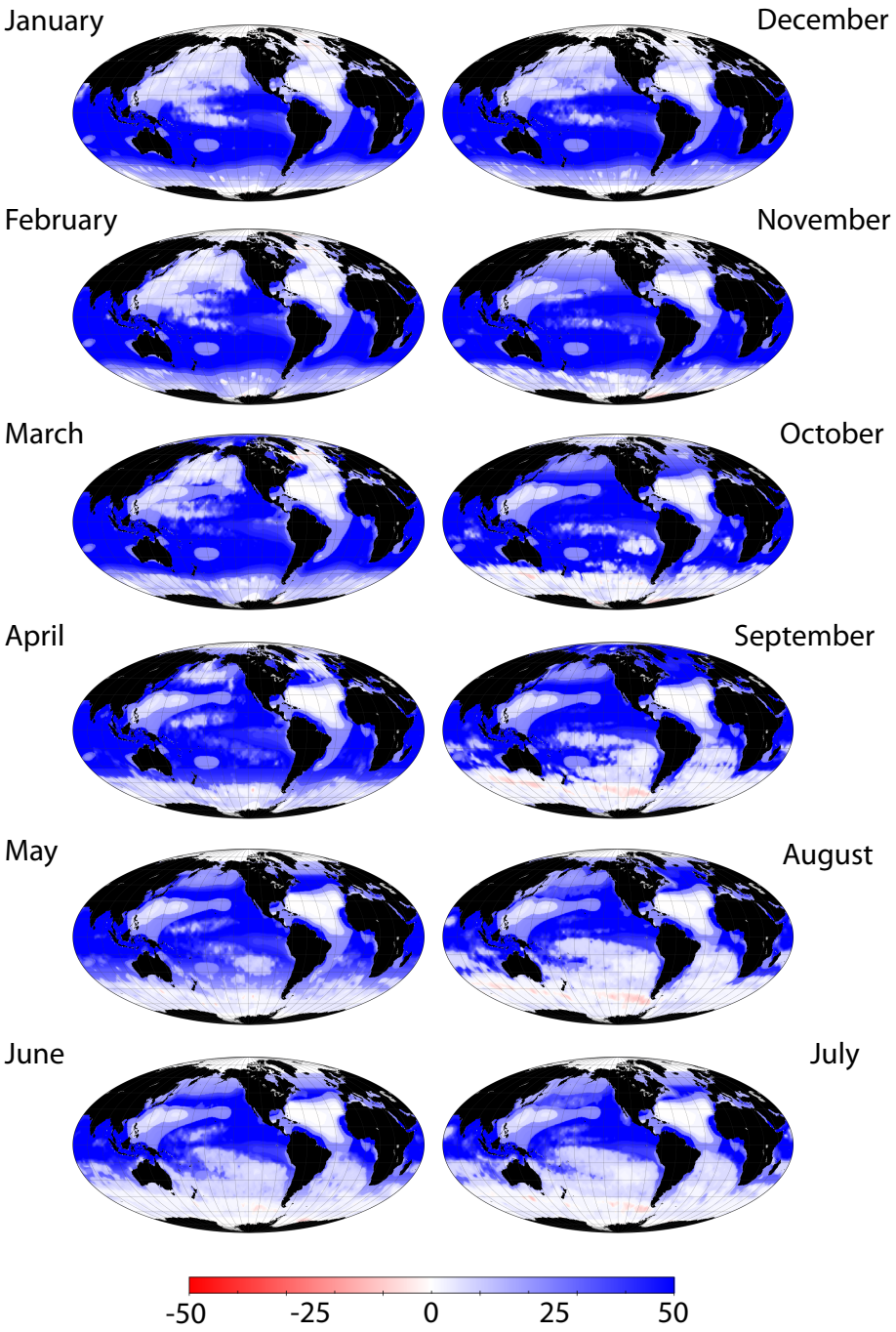


Figure S7: **MESS plots for Analysis I.** Sequences are classified using *de novo* clustering and a linear model is used (see Table S5). MESS plots give an estimate of where extrapolation, rather than interpolation, is necessary to project an SDM to geographic space. Negative MESS values indicate that at least one predictor variable is outside of the observed range. Overall, the variables used in the selected model were mostly within the observed range, indicating that maps were primarily based on interpolation rather than extrapolation.

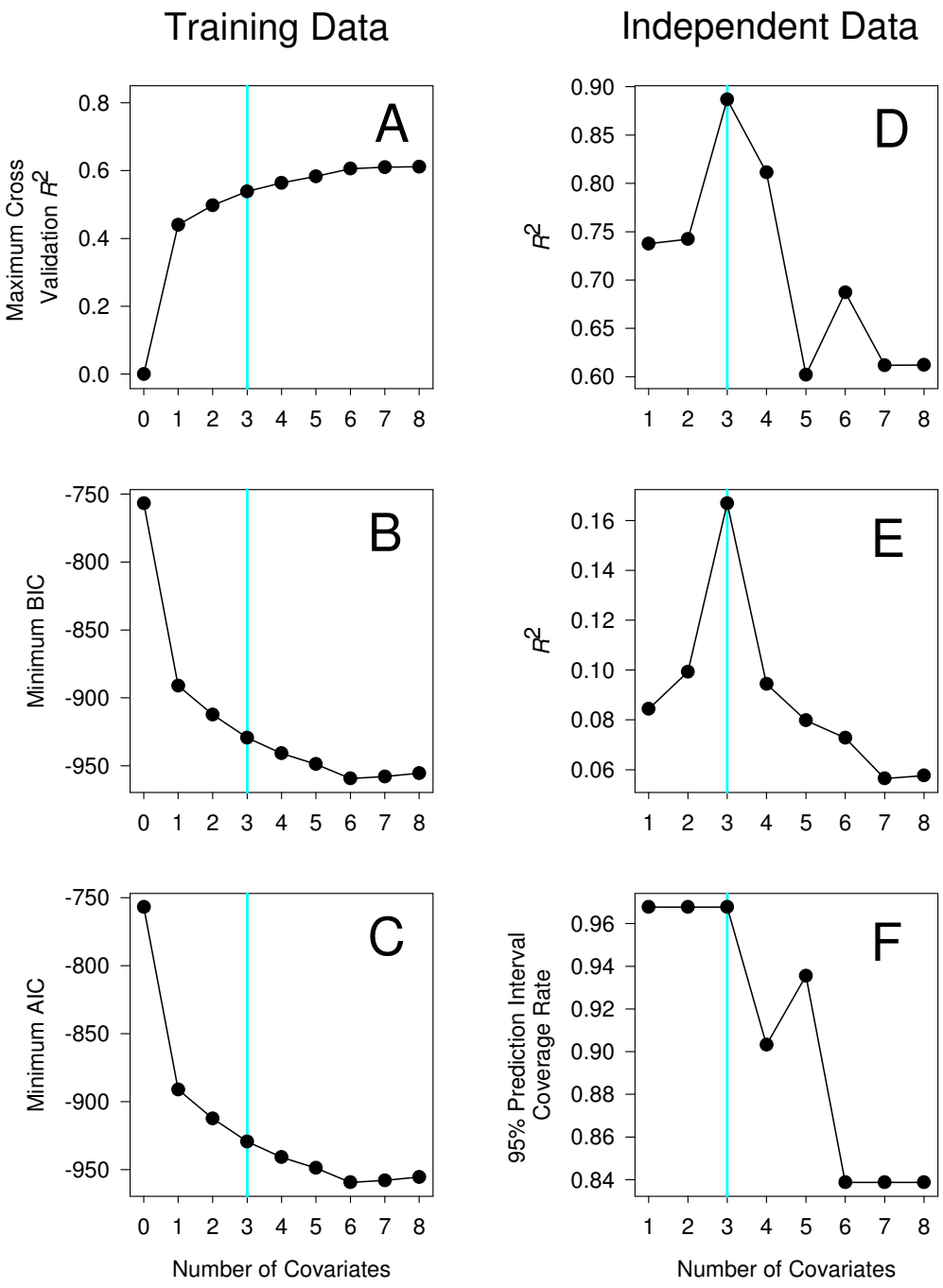


Figure S8: **Model selection for Analysis I.** Sequences are classified using *de novo* clustering and richness is modeled using a linear model (see Table S5). The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. (D) and (E) show the performance of the best cross validation model when used to predict independent data: POMMIER2007 and FUHRMAN2008, respectively. (F) Rate at which 95% prediction intervals cover observed values for the GOS data, when the model is fit using RDP-classified sequences at a rarefaction depth of 150 sequences. In all cases, the model was fit using MICROBIS data. All of training data measures (A, B, and C) point to a six-predictor model. However, the independent measures indicate that a six-predictor model is overfit, and that a three-predictor model is preferable. We, therefore, implemented the three-predictor model.

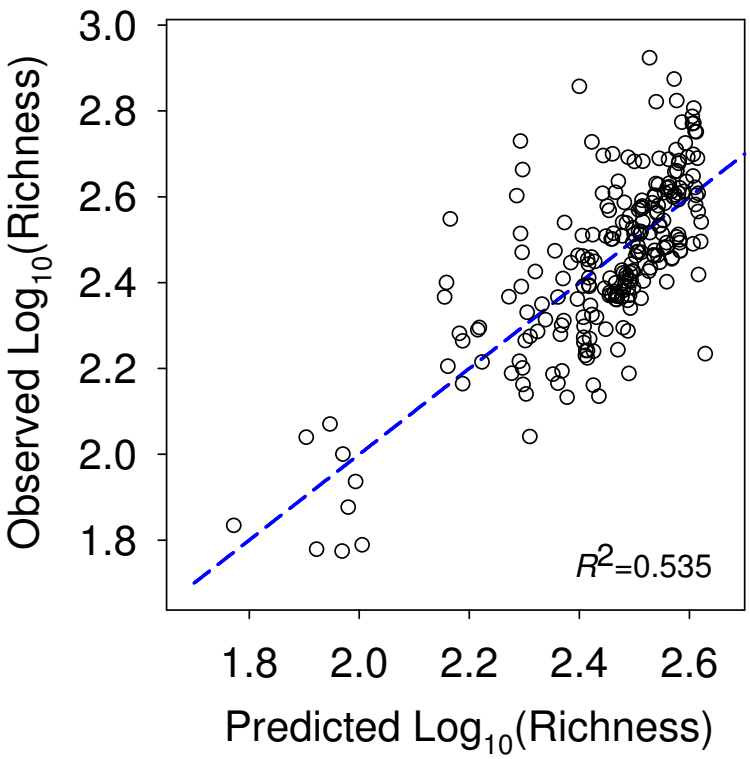


Figure S9: **Leave-one-out cross validation for Analysis I.** Sequences are classified using *de novo* clustering and a linear model is used.

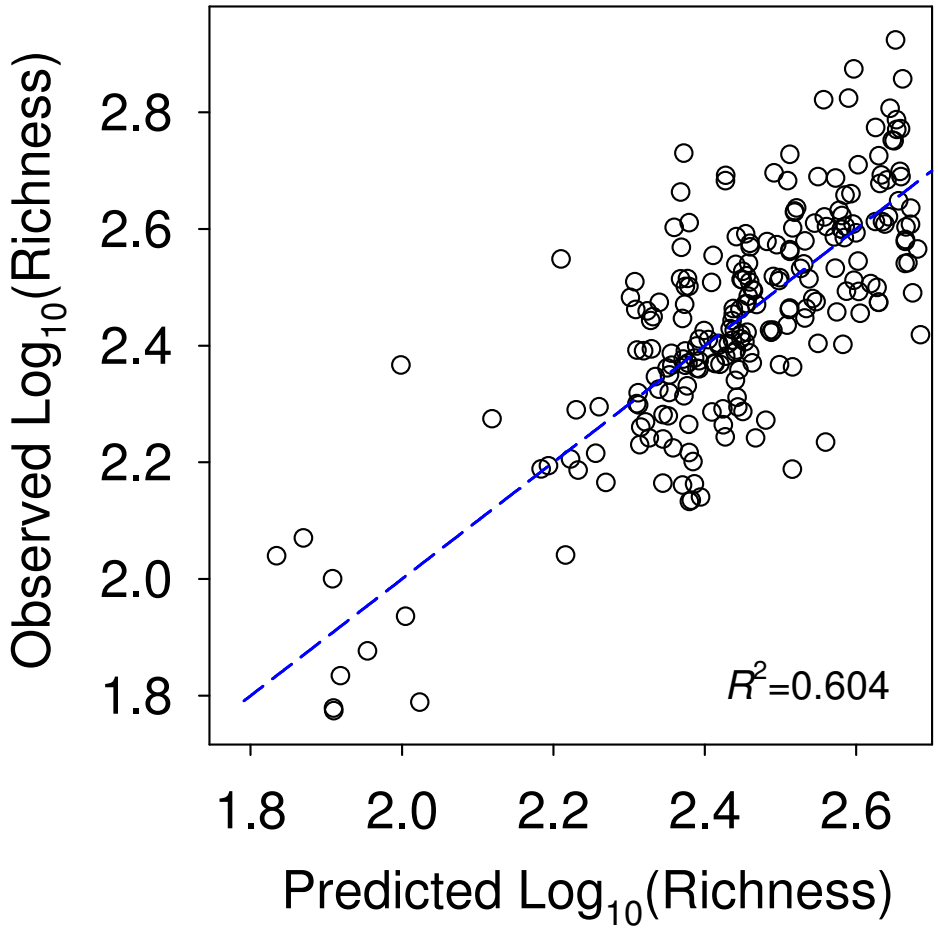


Figure S10: **Leave-one-out cross validation for Analysis III.** Sequences are classified using *de novo* clustering and a multiple adaptive regression splines model is used.

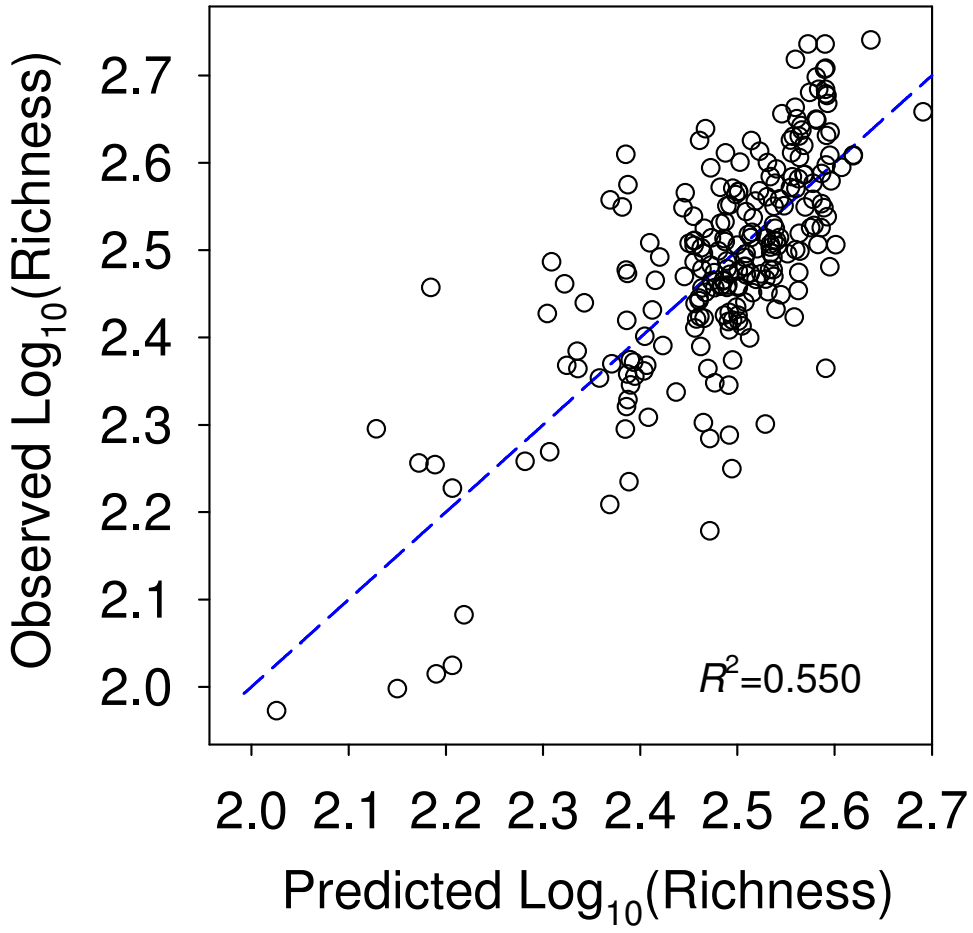


Figure S11: **Leave-one-out cross validation for Analysis II.** Sequences are classified using RDP and a linear model is used.

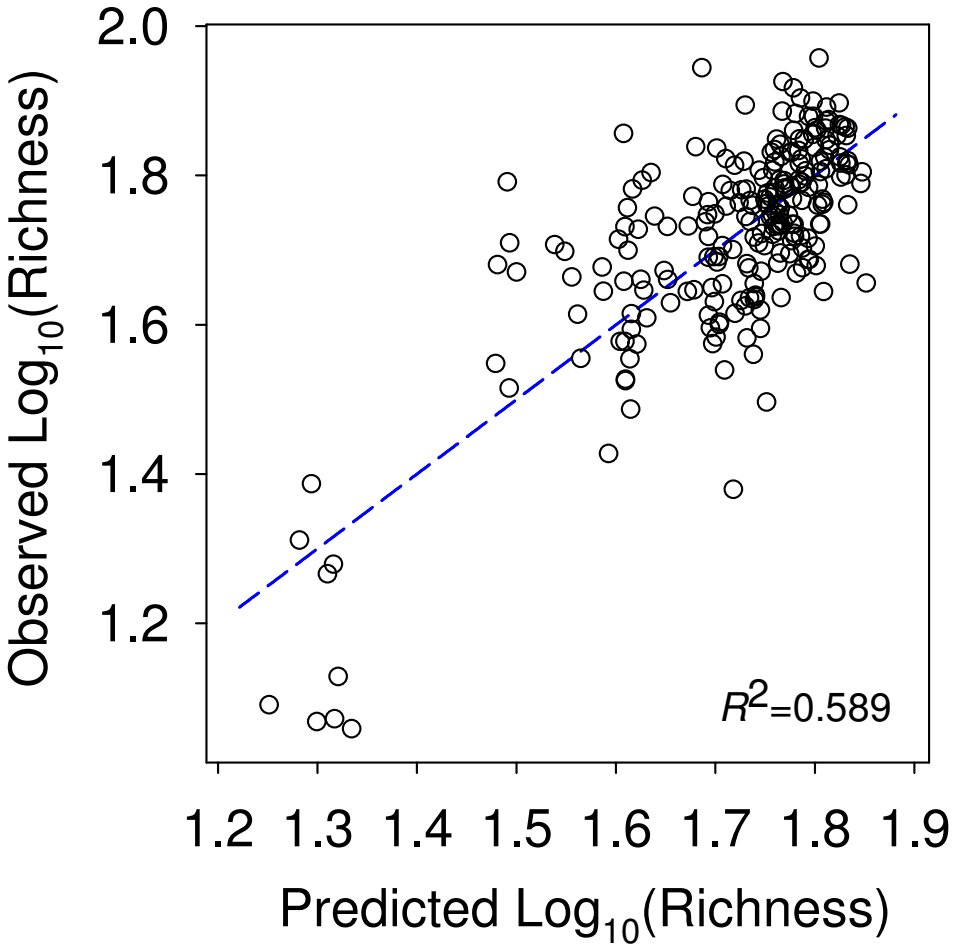


Figure S12: **Leave-one-out cross validation for Analysis IV.** Sequences are classified *de novo* clustering at shallow sampling depth and a linear model is used.

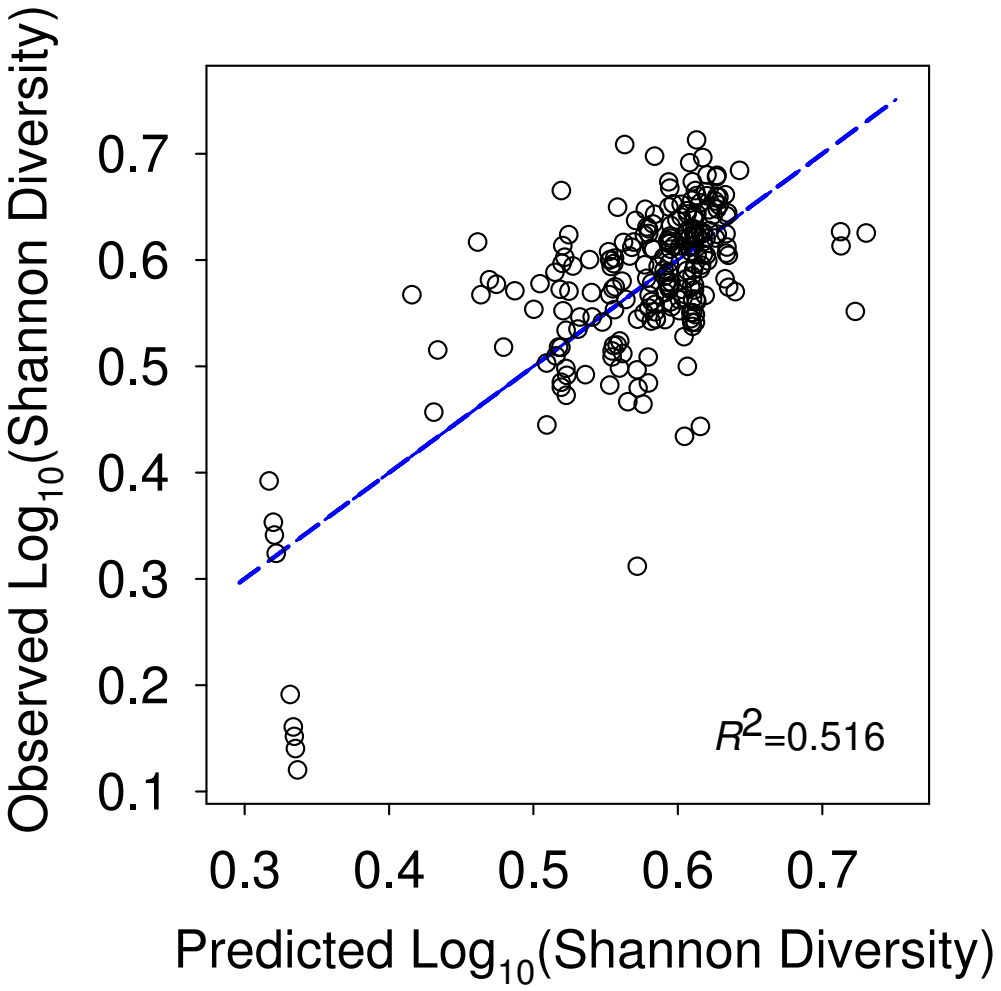


Figure S13: **Leave-one-out cross validation for the Shannon diversity analysis.**

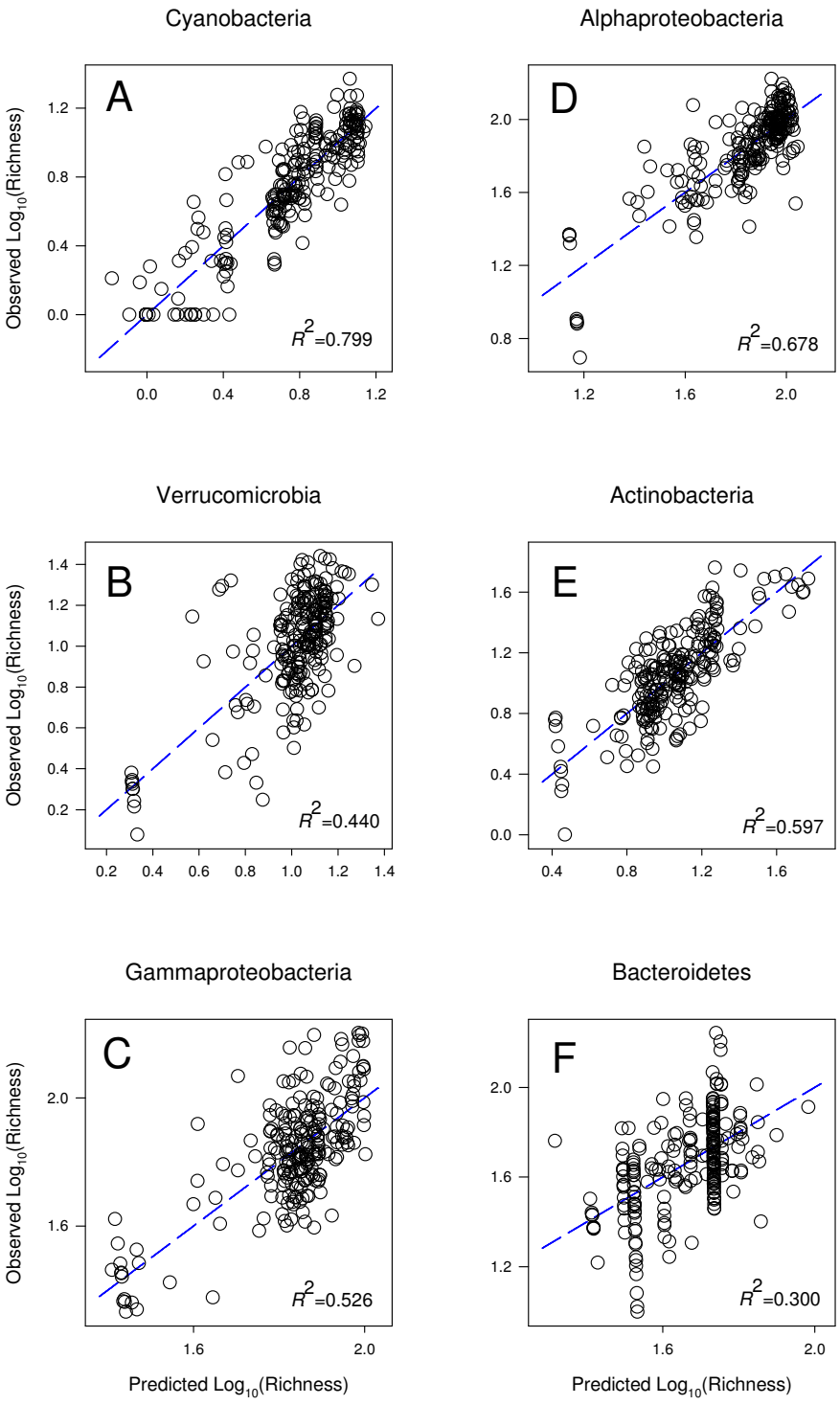


Figure S14: **Leave-one-out cross validation for models of diversity within phyla (Analyses VI to XI).** Panels are for (A) cyanobacteria, (B) verrucomicrobia, (C) gammaproteobacteria, (D) alphaproteobacteria, (E) actinobacteria, and (F) bacteroidetes.

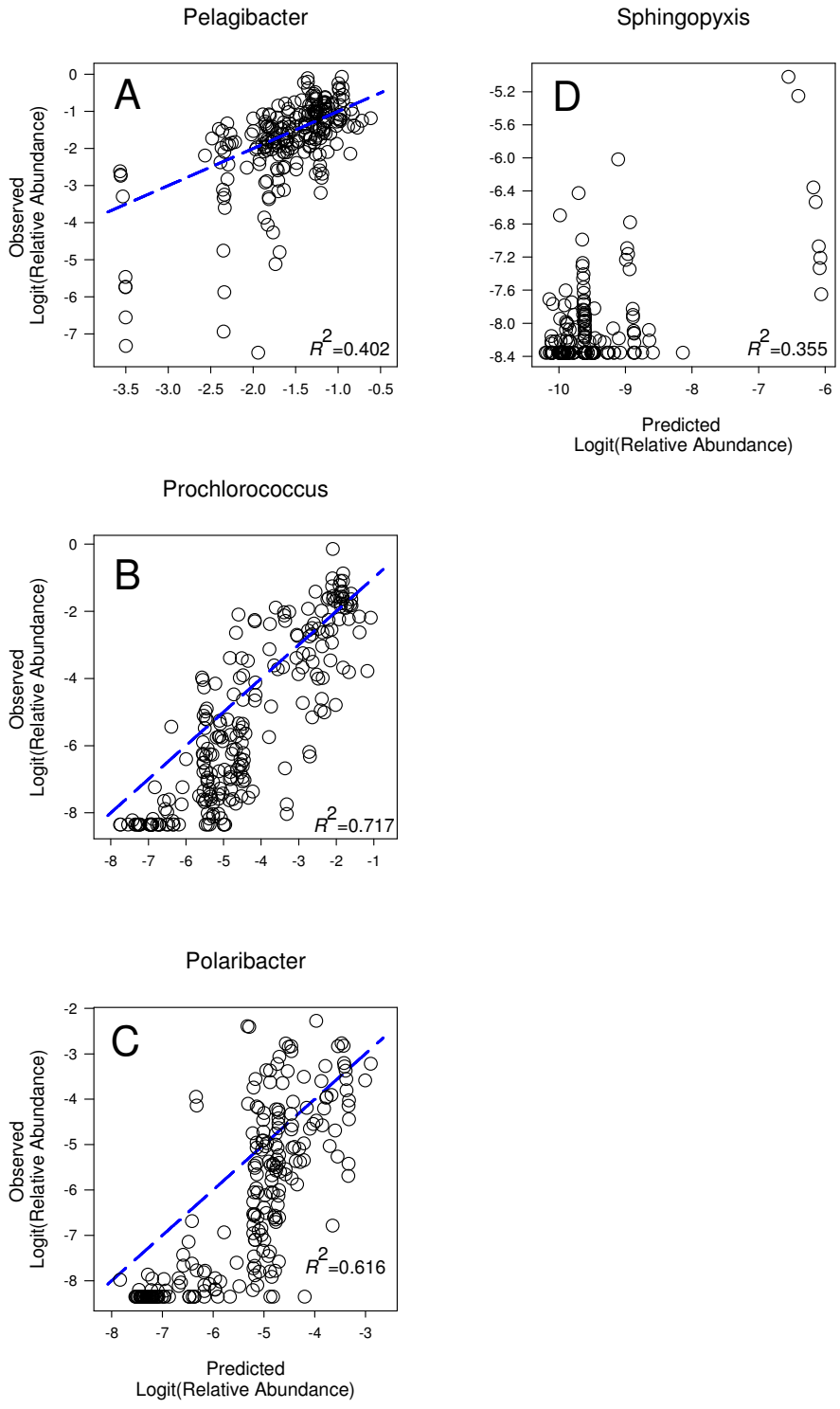


Figure S15: **Leave-one-out cross validation for models of relative abundance of select genera (Analyses XII to XV).** Panels are for (A) *Pelagibacter*, (B) *Prochorococcus* and *Synechococcus* (grouped together by RDP), (C) *Polaribacter*, and (D) *Sphingopyxis*.

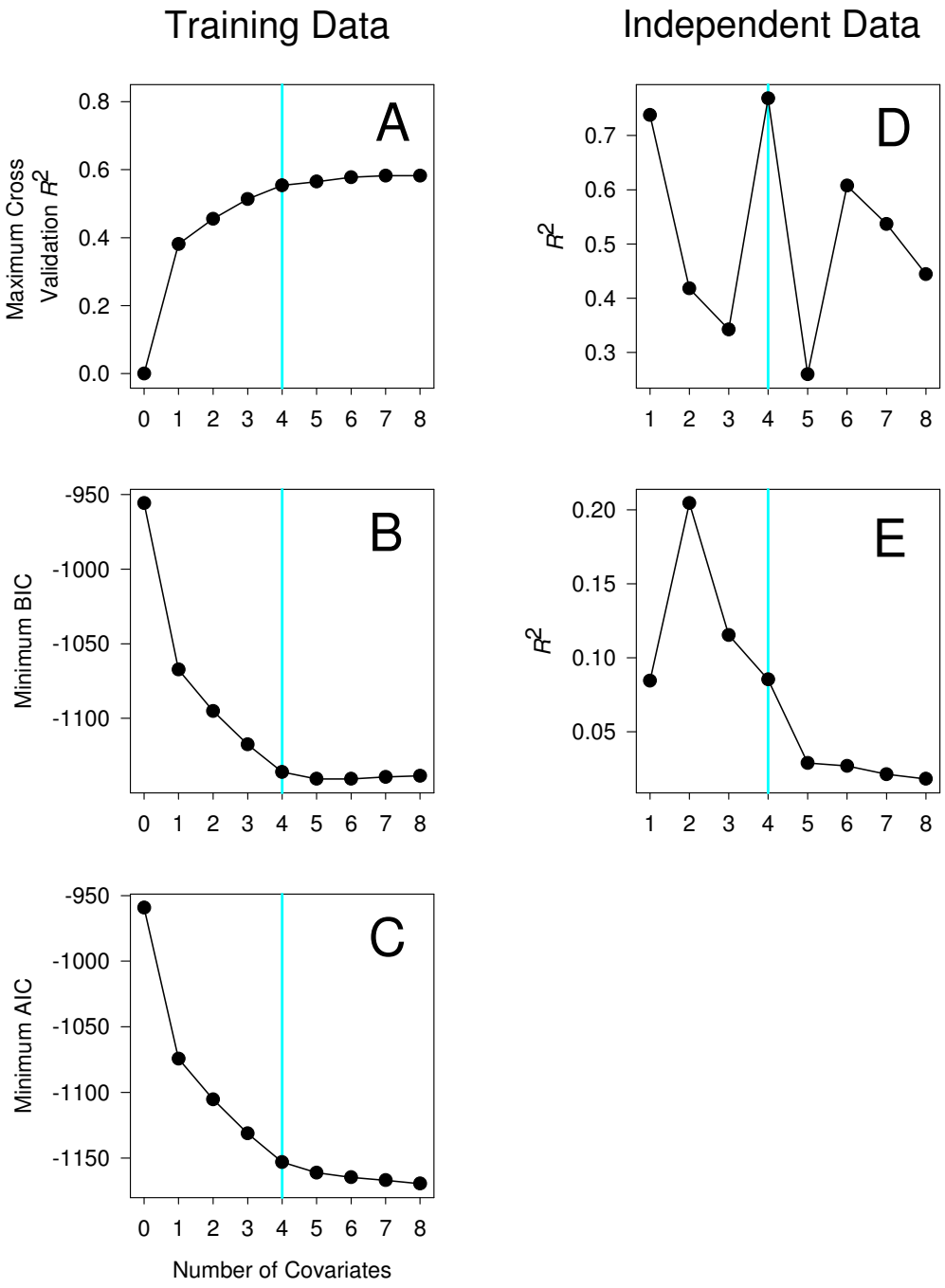


Figure S16: **Model selection for Analysis II.** Sequences are classified using RDP, and richness is modeled using a linear model (see Table S5). The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. (D) and (E) Performance of the best cross validation model when used to predict independent data: POMMIER2007 and FUHRMAN2008, respectively. All training data measures (A, B, and C) point to a six-, seven-, or eight-predictor model. However, the independent measures indicate that these models are overfit, and that a three-predictor model is preferable. We therefore implemented the three-predictor model.

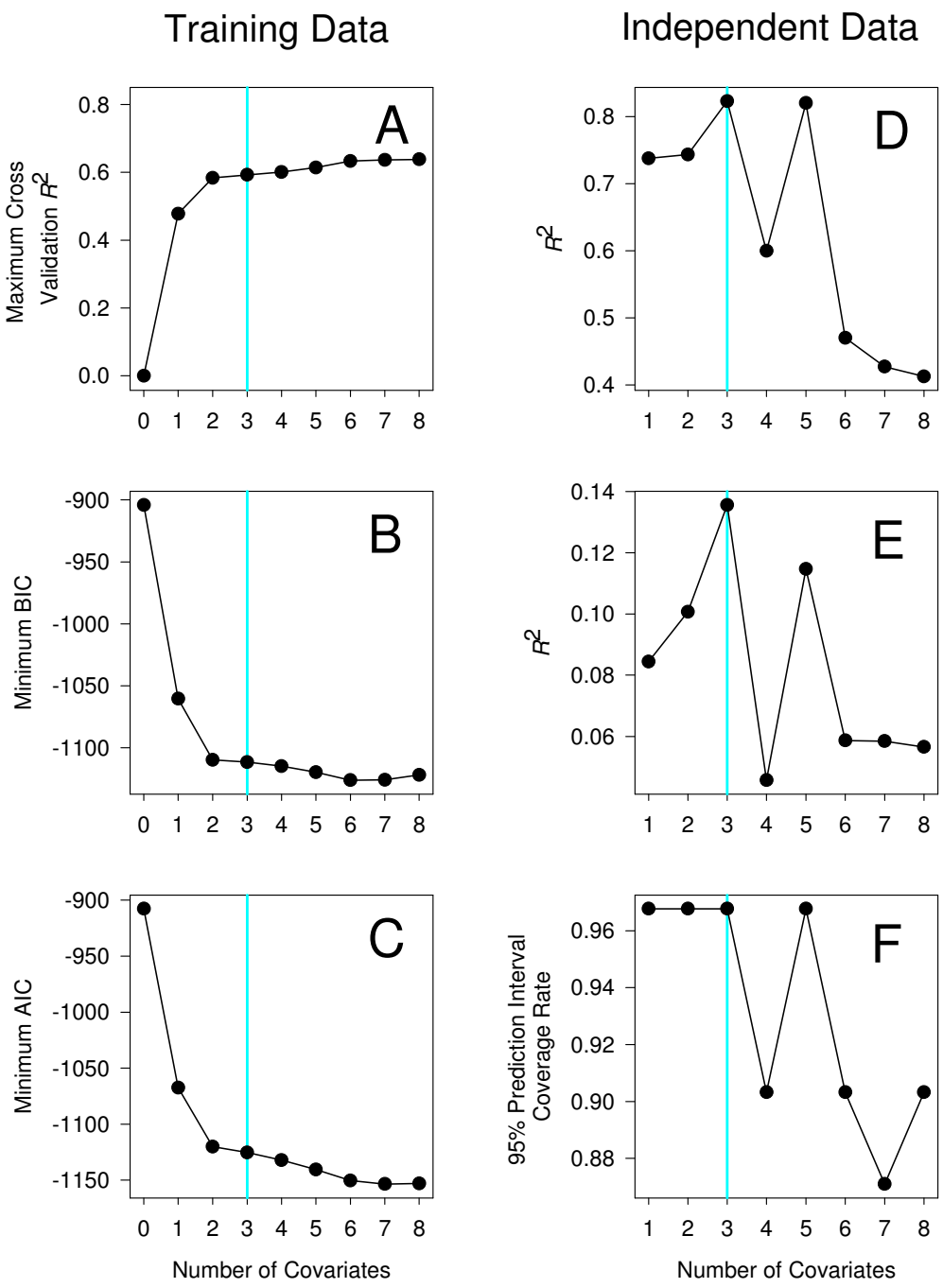


Figure S17: **Model selection for Analysis IV.** Sequences are classified using *de novo* clustering, and richness is modeled using a linear model, and sequencing depth is shallow (150 reads; see Table S5). The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. (D) and (E) Performance of the best cross-validation model when used to predict independent data: POMMIER2007 and FUHRMAN2008, respectively. (F) Rate at which 95% prediction intervals cover observed values for the GOS data, when the model is fit using RDP-classified sequences. In all cases, the model was fit using MICROBIS data. All training data measures (A, B, and C) point to a six- or seven-predictor model. However, the independent measures indicate that these models are overfit, and that a three-predictor model is preferable. We therefore implemented the three-predictor model.

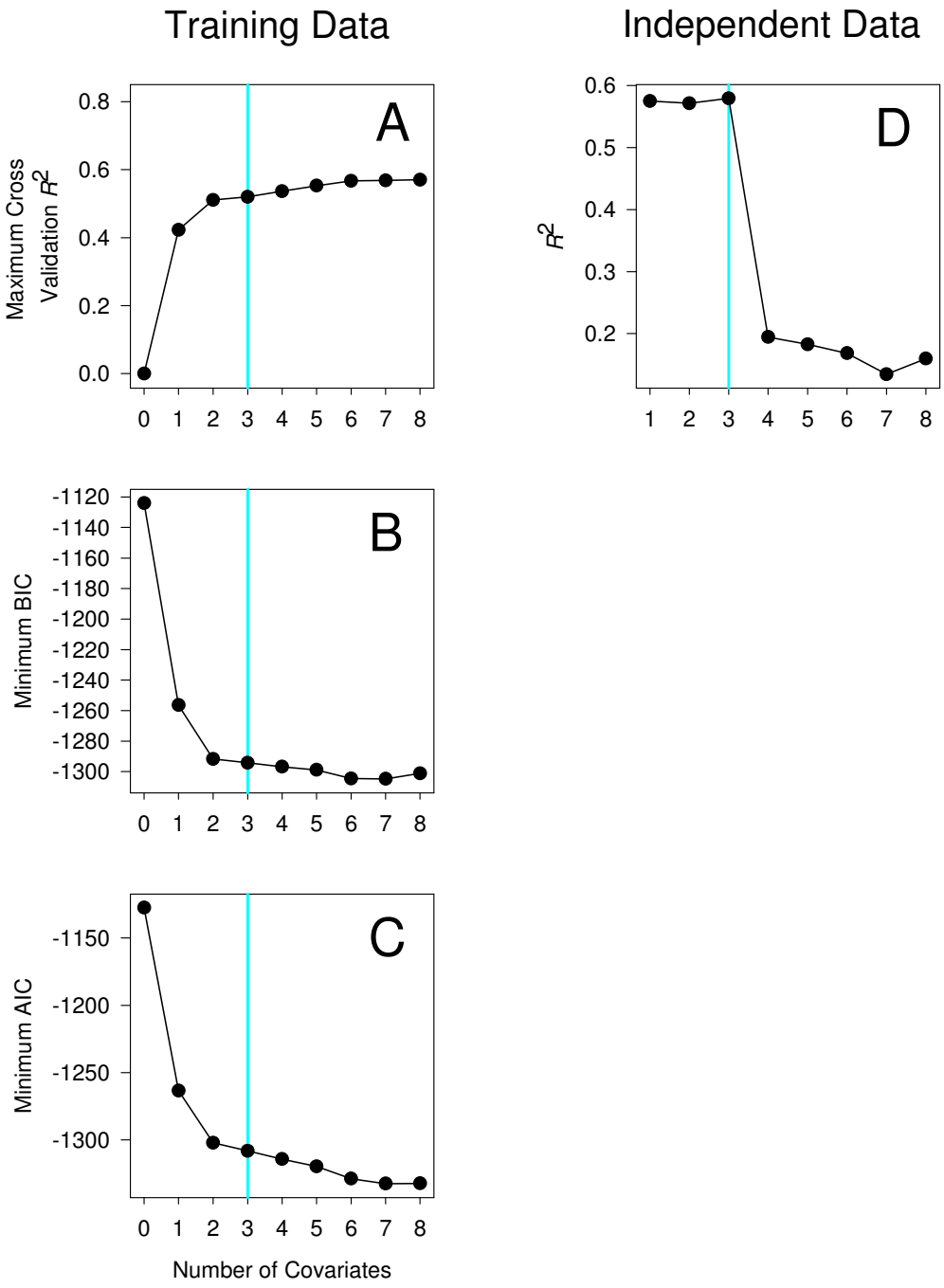


Figure S18: **Model selection for mapping Shannon diversity (Analysis V).** The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. (D) Performance of the best cross validation model when used to predict independent data: POMMIER2007. (The model is fit with MICROBIS data.) All training data measures (A, B, and C) point to a six- or seven-predictor model. However, the independent measure indicates that these models are overfit, and that a three-predictor model is preferable. We therefore implemented the three-predictor model.

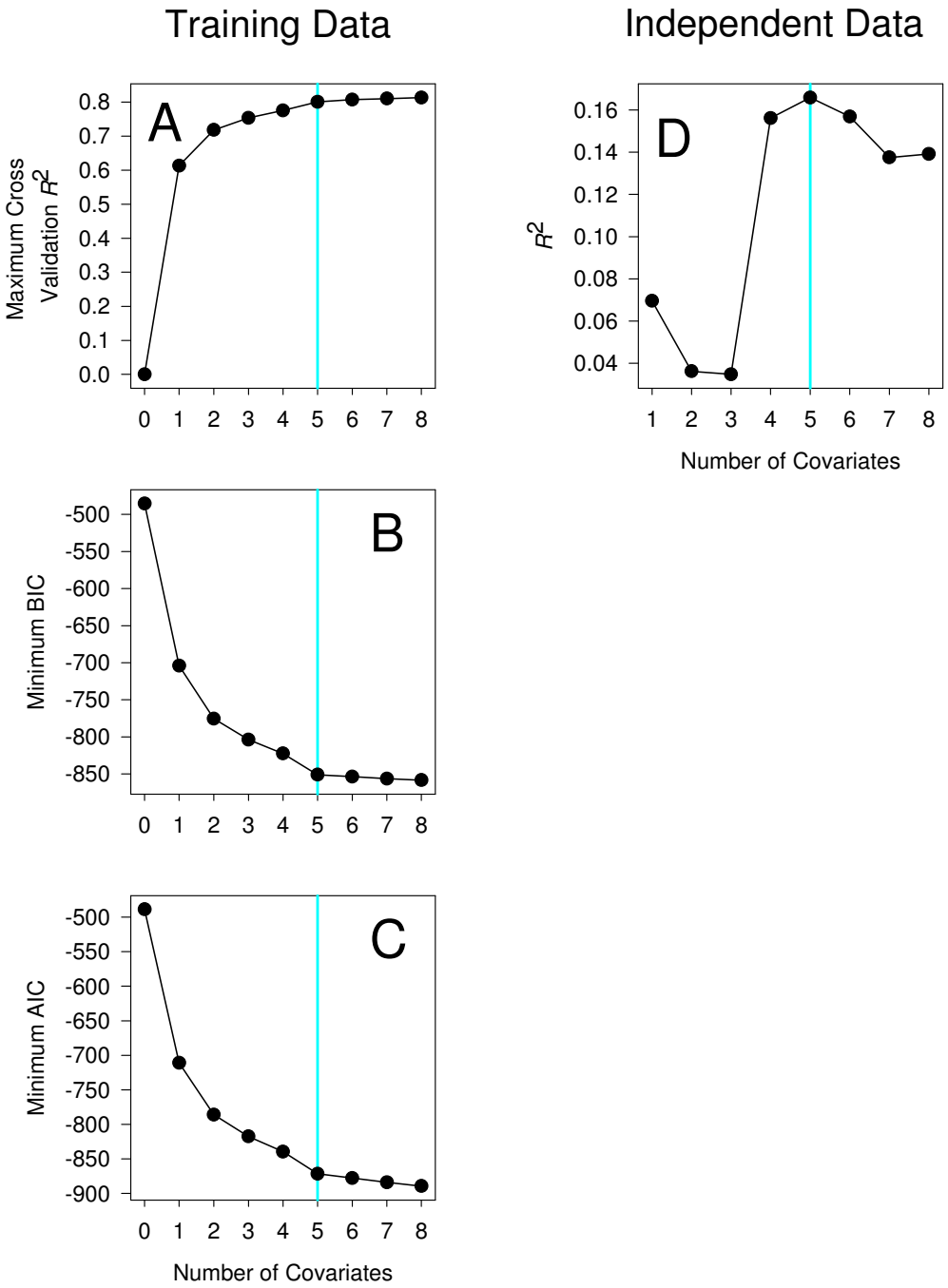


Figure S19: **Model selection for mapping richness within cyanobacteria (Analysis VI).** The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. (D) Performance of the best cross validation model when used to predict independent data POMMIER2007. (The model was fit using MICROBIS data.) All training data measures (A, B, and C) point to a eight-predictor model. However, the independent measures indicate that an eight-predictor model is overfit, and that a five-predictor model is preferable. We therefore implemented the five-predictor model.

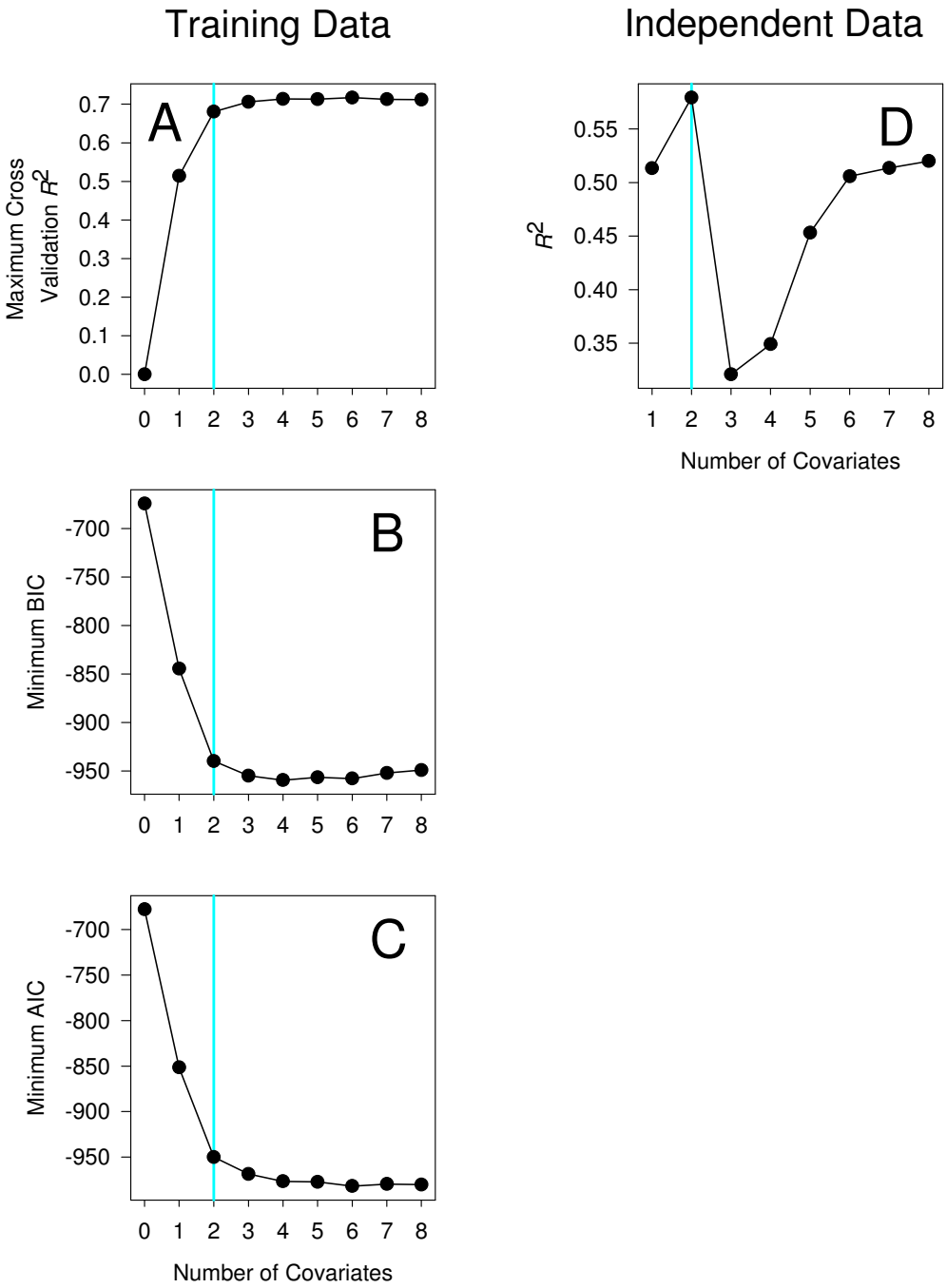


Figure S20: **Model selection for mapping richness within alphaproteobacteria (Analysis VII).** The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. (D) Performance of the best cross validation model when used to predict independent data POMMIER2007. (The model was fit using MICROBIS data.) All training data measures (A, B, and C) point to a six-predictor model. However, the independent measures indicate that a six-predictor model is overfit, and that a two-predictor model is preferable. We therefore implemented the two-predictor model.

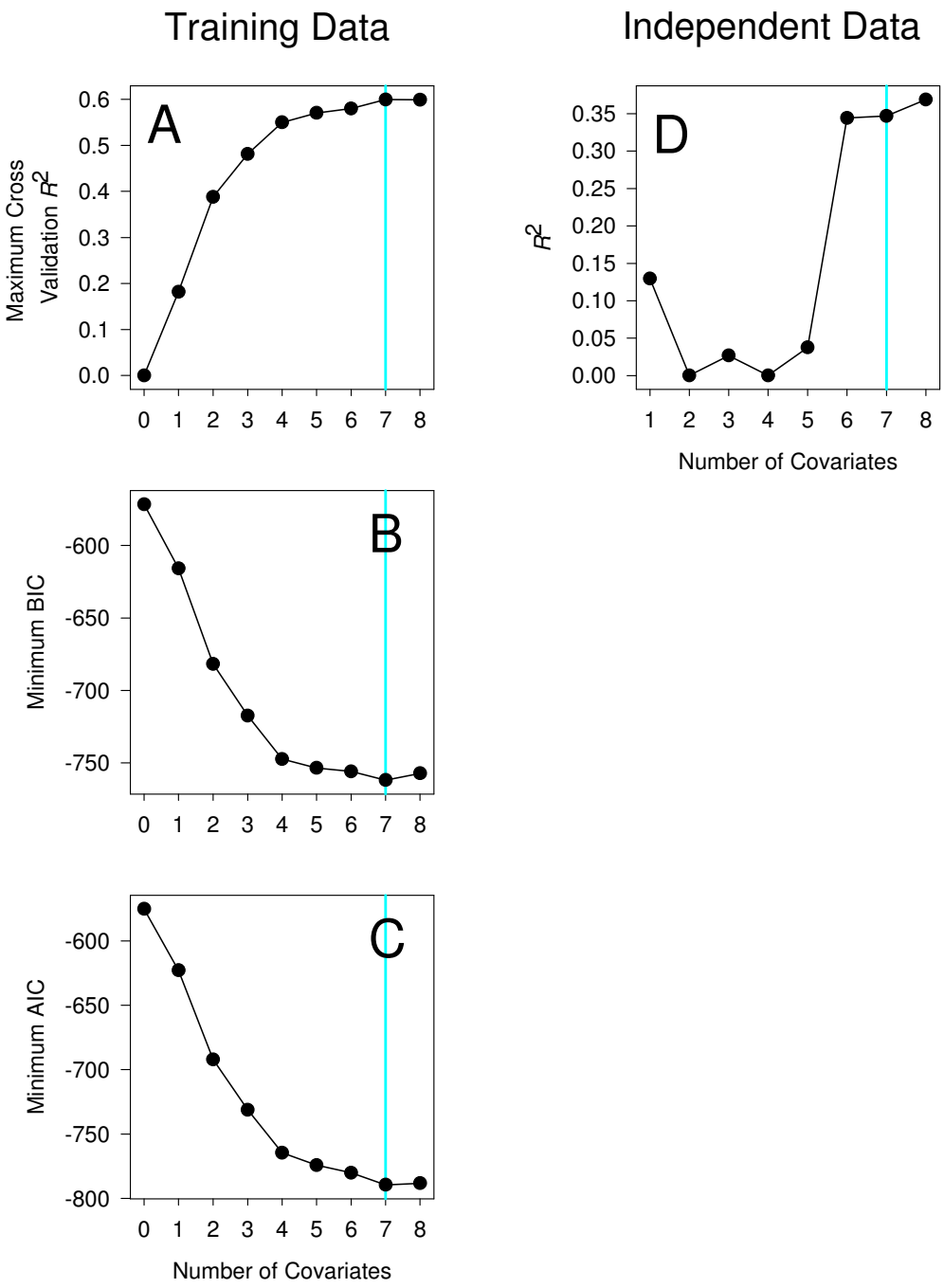


Figure S21: **Model selection for mapping richness within actinobacteria (Analysis VIII).** The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. (D) Performance of the best cross validation model when used to predict independent data POMMIER2007. (The model was fit using MICROBIS data.) All training data measures (A, B, and C) point to a seven-predictor model as does the independent data, so we implemented this model.

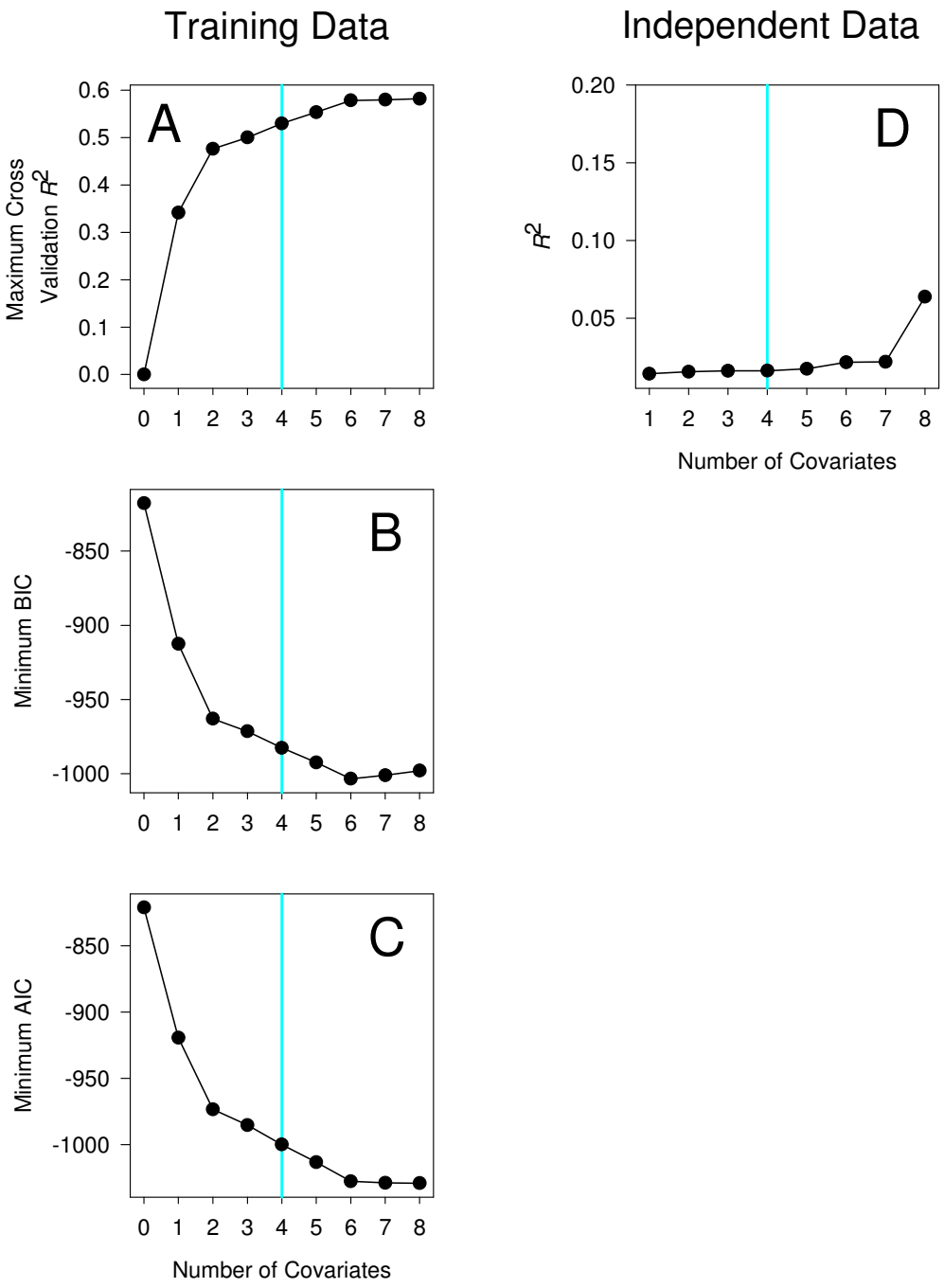


Figure S22: **Model selection for mapping richness within gammaproteobacteria (Analysis IX).** The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. (D) Performance of the best cross validation model when used to predict independent data POMMIER2007. (The model was fit using MICROBIS data.) All training data measures (A, B, and C) point to a six-predictor model. Although the independent data are uninformative here ($R^2 < 0.1$ for all models), based on the previous analyses, the six-predictor model is likely overfit. Thus, we implemented the four-predictor model, the largest model such that adding another predictor increases the cross validation R^2 less than 2.5% (see Supplementary Methods).

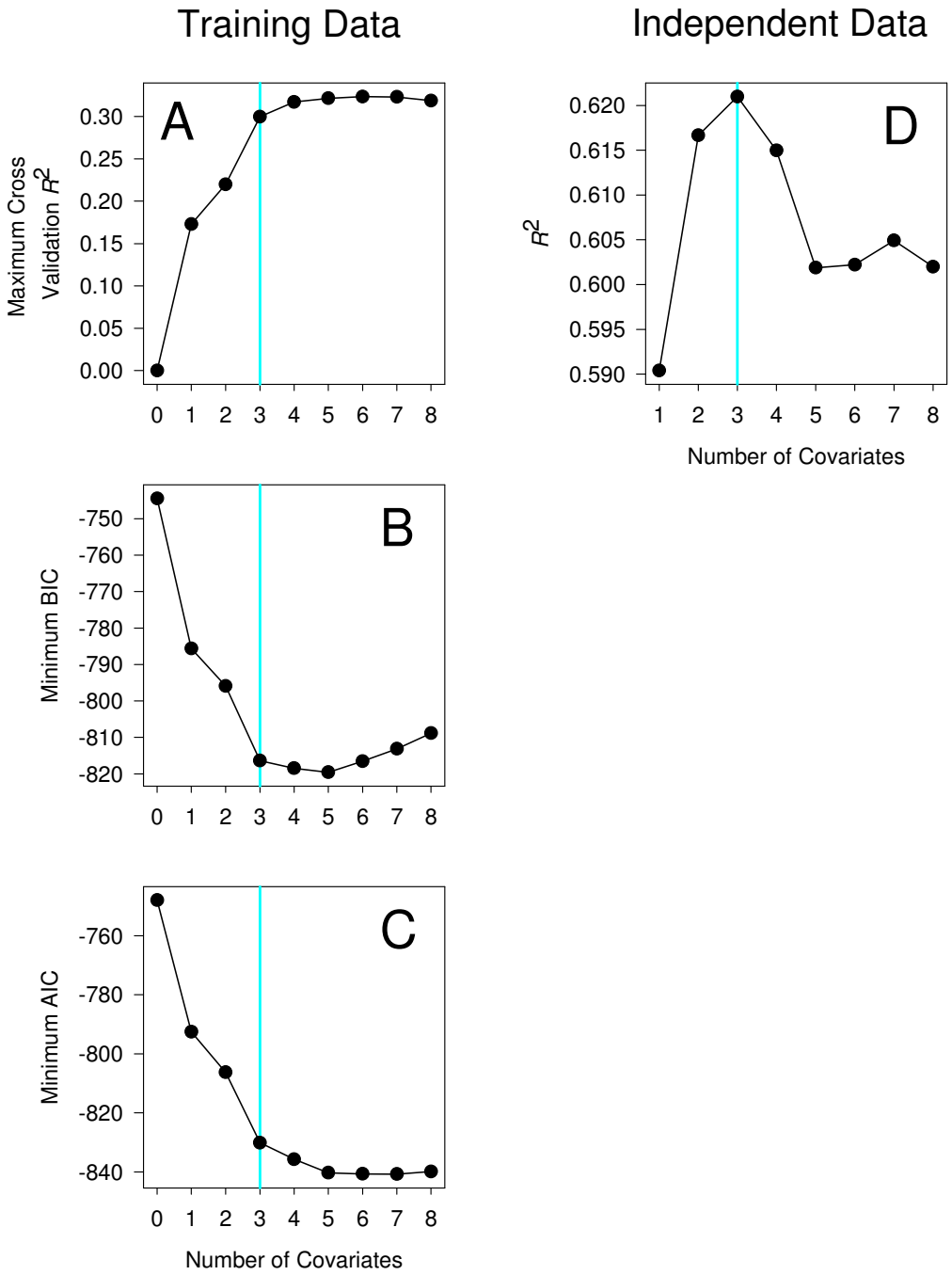


Figure S23: **Model selection for mapping richness within bacteroidetes (Analysis X).** The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. (D) Performance of the best cross validation model when used to predict independent data POMMIER2007. (The model was fit using MICROBIS data.) All training data measures (A, B, and C) point to a five-predictor model. However, the independent measures indicate that a five-predictor model is overfit, and that a three-predictor model is preferable. We therefore implemented the three-predictor model.

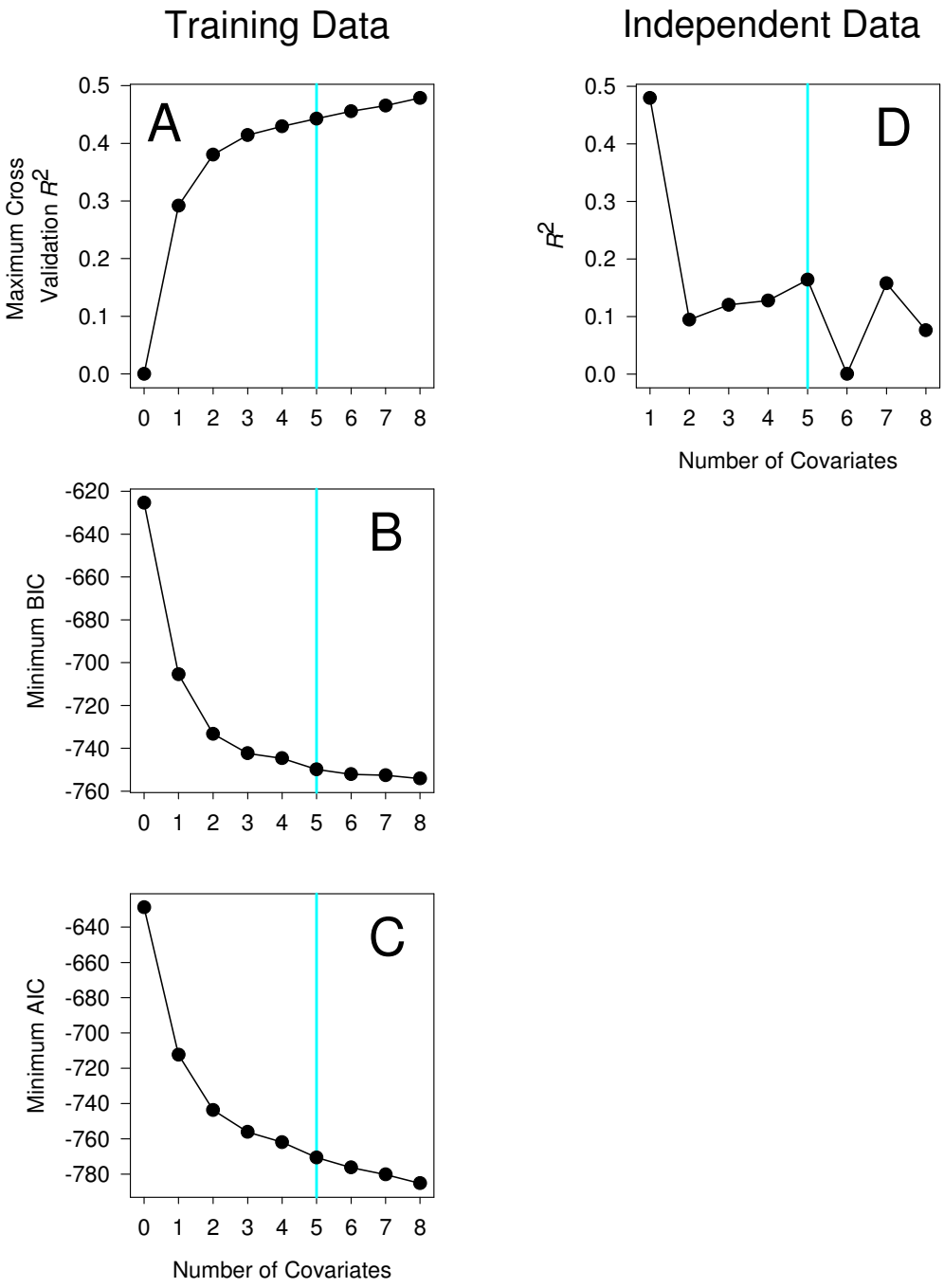


Figure S24: **Model selection for mapping richness within verrucomicrobia (Analysis XI).** The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. (D) Performance of the best cross validation model when used to predict independent data POMMIER2007. (The model was fit using MICROBIS data.) All training data measures (A, B, and C) point to a model with eight or more predictors. However, the independent measures indicate that a this model is overfit, and that a one- or five-predictor model is preferable. Because the five-predictor model had better performance with the training data, we implemented it.

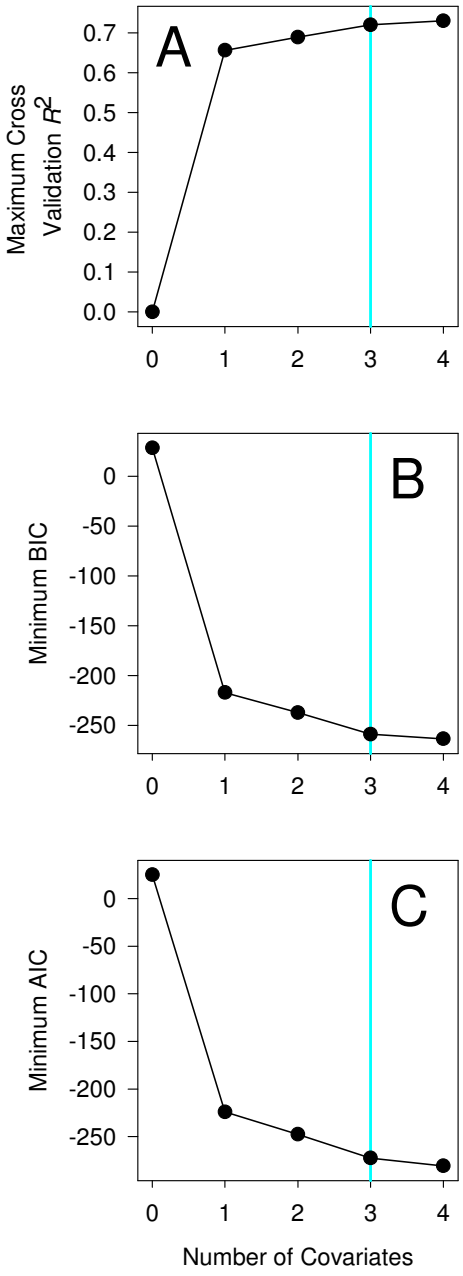


Figure S25: **Model selection for mapping the relative abundance of *Prochlorococcus* and *Synechococcus* (Analysis XII).** The taxa *Prochlorococcus* and *Synechococcus* were grouped together by RDP, so we analyzed them in a single analysis. The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. Although a small marginal improvement could be gained by using a more complex model, we chose a conservative three-predictor model based on the overfitting observed using the independent data in the other analyses (Supplementary Methods).

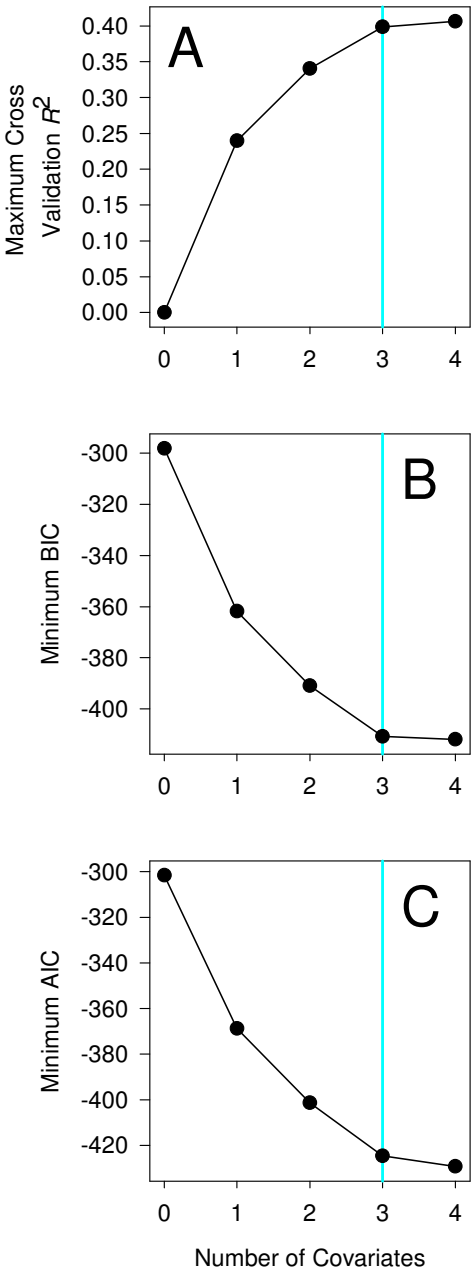


Figure S26: **Model selection for mapping the relative abundance of *Pelagibacter* (Analysis XIII).** The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. Although a small marginal improvement could be gained by using a more complex model, we chose a conservative three-predictor model based on the overfitting observed using the independent data in the other analyses (Supplementary Methods).

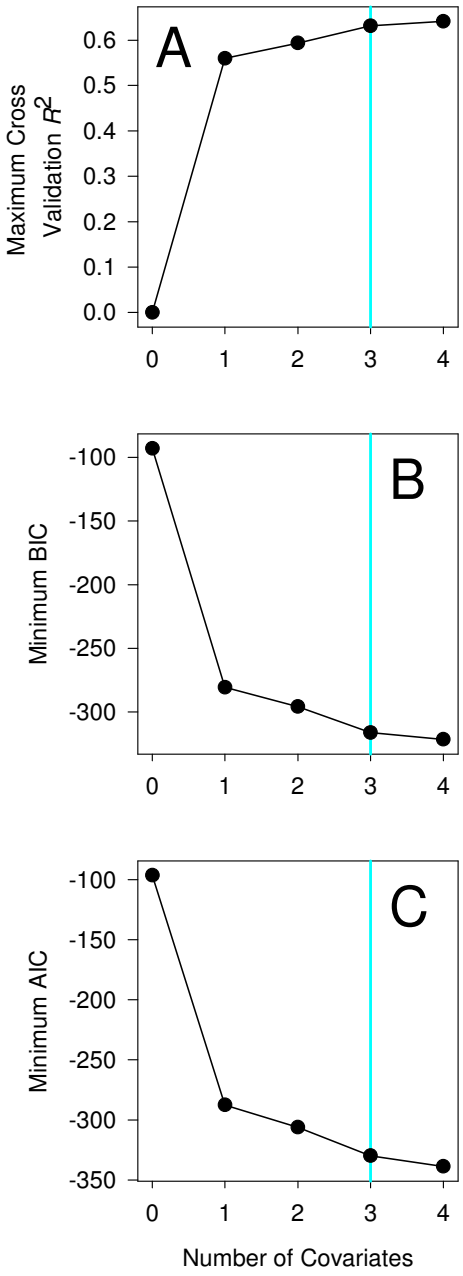


Figure S27: **Model selection for mapping the relative abundance of *Polaribacter* (Analysis XIV).** The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. Although a small marginal improvement could be gained by using a more complex model, we chose a conservative three-predictor model based on the overfitting observed using the independent data in the other analyses (Supplementary Methods).

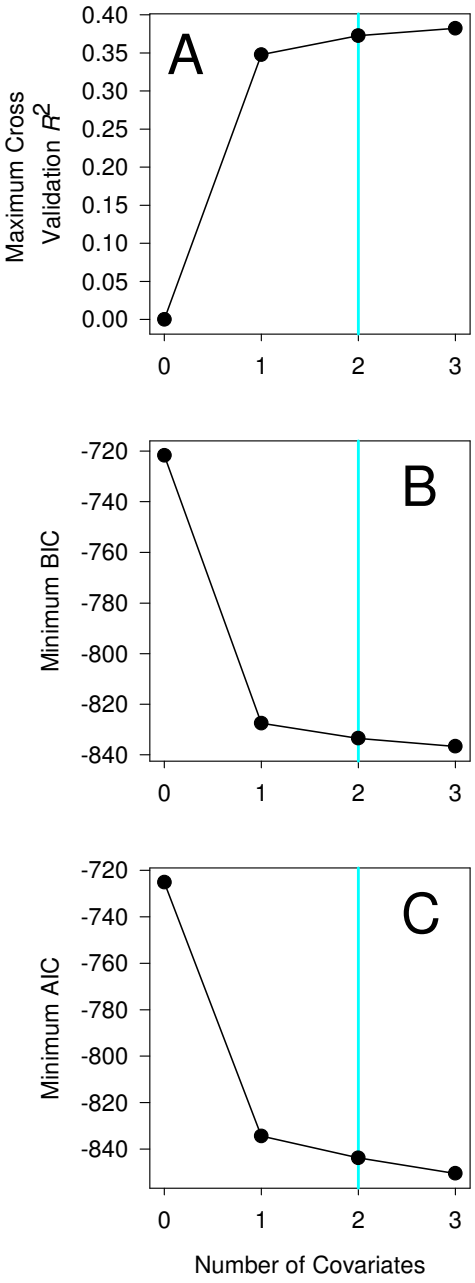


Figure S28: **Model selection for mapping the relative abundance of *Sphingopyxis* (Analysis XV).** The graphs in (A), (B), (C), respectively, show the best possible cross validation, BIC, and AIC values for the given number of predictors. Although a small marginal improvement could be gained by using a more complex model, we chose a conservative two-predictor model based on the overfitting observed using the independent data in the other analyses (Supplementary Methods).

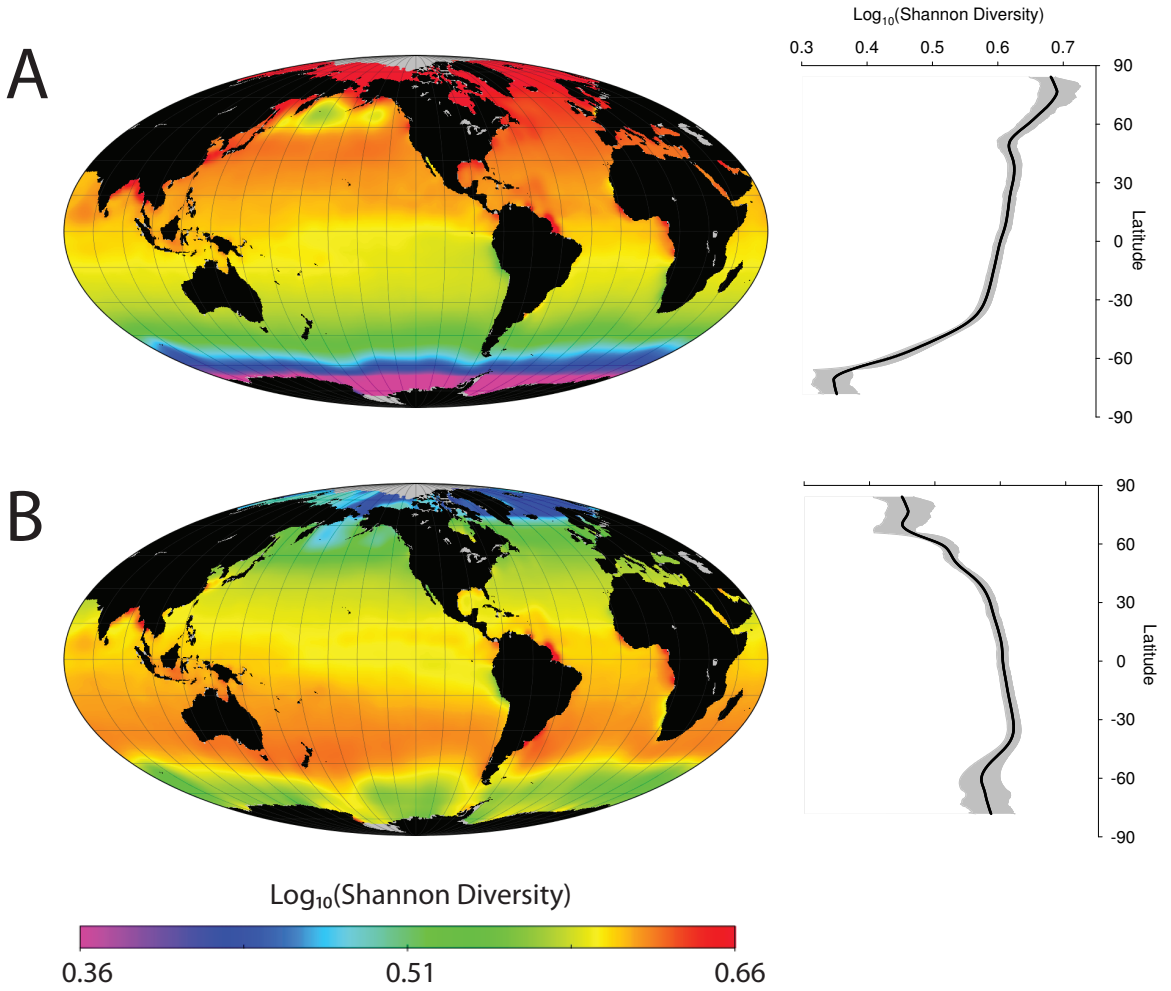
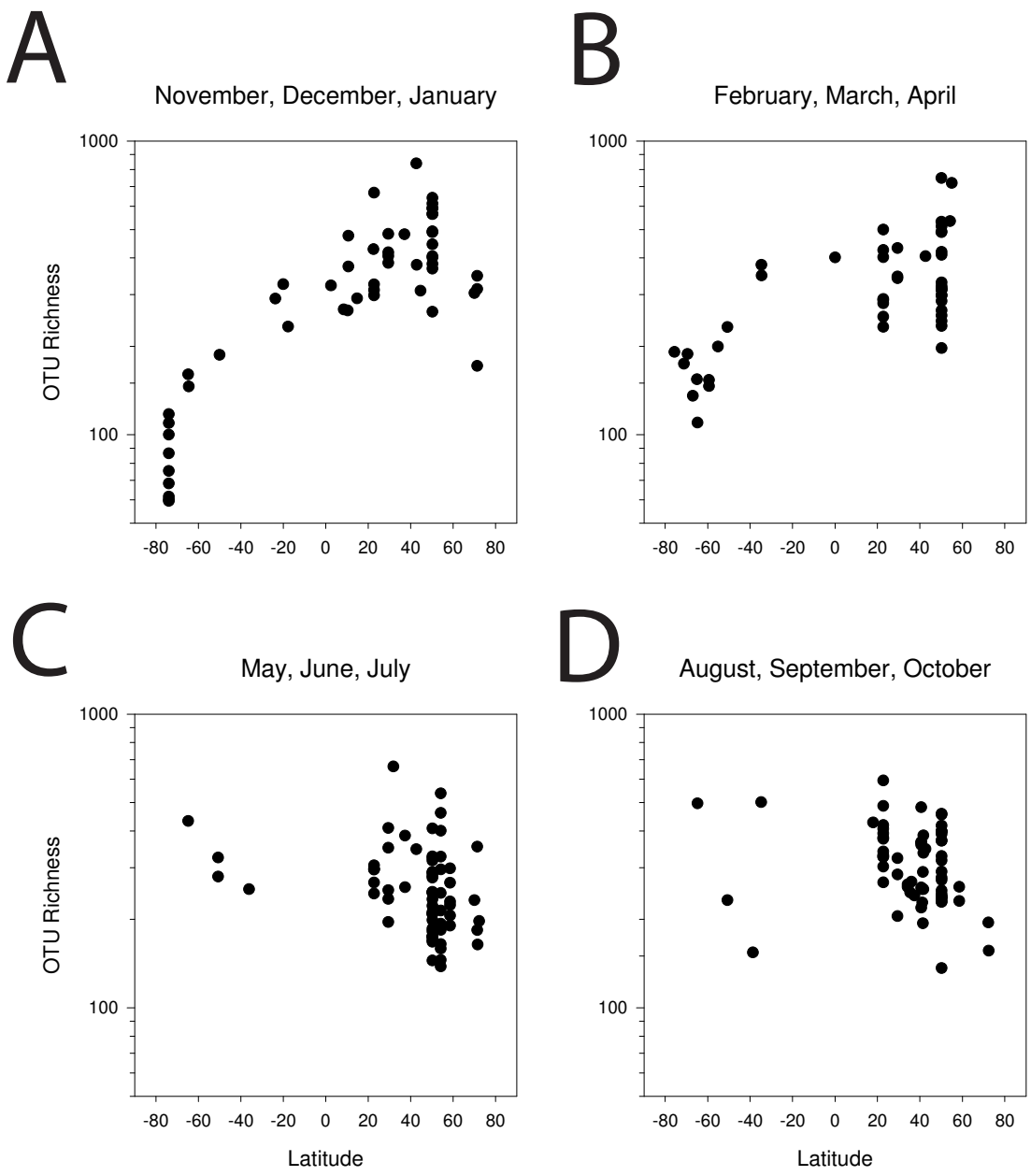


Figure S29: **Maps of global marine bacterial Shannon diversity.** Sequences are classified using *de novo* clustering, and diversity is modeled using a linear model (Analysis V; see Table S5). Shannon diversity is high when all OTUs are equally prevalent, and low when only a few OTUs dominate the community. (A) In December, Shannon diversity peaks in temperate latitudes in the Northern Hemisphere. (B) In June, Shannon diversity peaks in temperate latitudes in the Southern Hemisphere. Predicted Shannon diversity during the spring and fall is intermediate, with roughly globally uniform richness near the equinoxes. Predicted Shannon diversity patterns are qualitatively similar to predicted richness patterns (Fig. 1), and suggest that summertime blooms may contribute to global diversity patterns when measured by certain parameters.

Figure S30: **Latitudinal gradient in observed data.** OTU Richness (log scale) is plotted versus latitude for MICROBIS samples from each of four seasons. A strong latitudinal gradient is observed in the winter months, which generally reverses in summer. These plots do not account for distane from thermocline and other variables which affect OTU Richness and which were included in our SDM predictions. Their influence accounts for the fact that we predict a stronger reversed gradient in summer than is apparent in these plots.



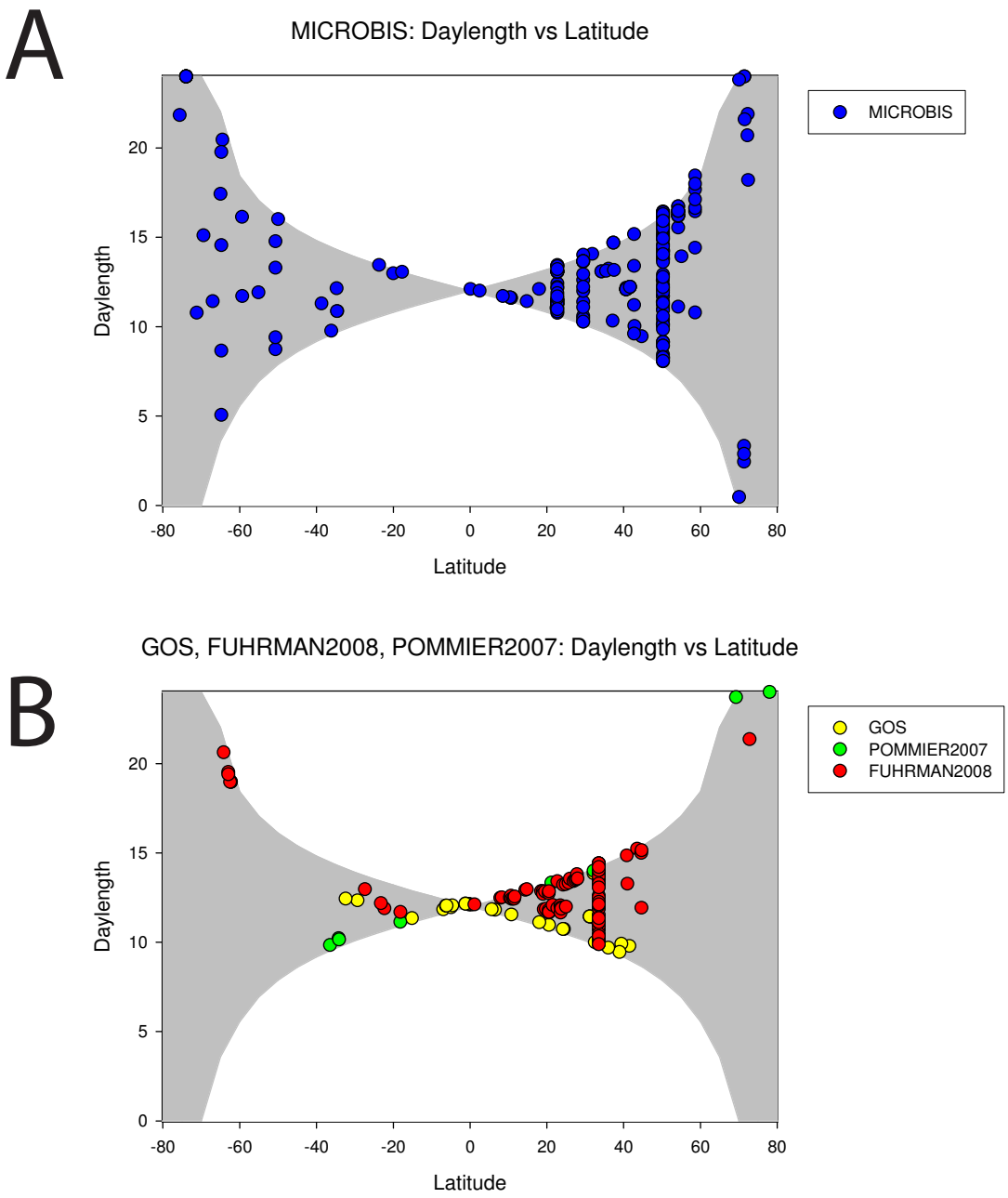


Figure S31: **Sampling bias in marine surveys.** The latitude of sampling locations from MICROBIS and the three validation data sets are plotted against the day length on the day of the survey. The range of daylengths at each latitude are shown in grey. (A) MICROBIS has more samples in the Northern versus Southern hemispheres, but nearly covers the range of possible latitude and daylength combinations. Specifically, MICROBIS sampled high latitudes in both winter and summer. (B) In contrast, GOS sampled only low latitudes, and POMMIER and FUHRMAN did not sample high latitudes in winter.

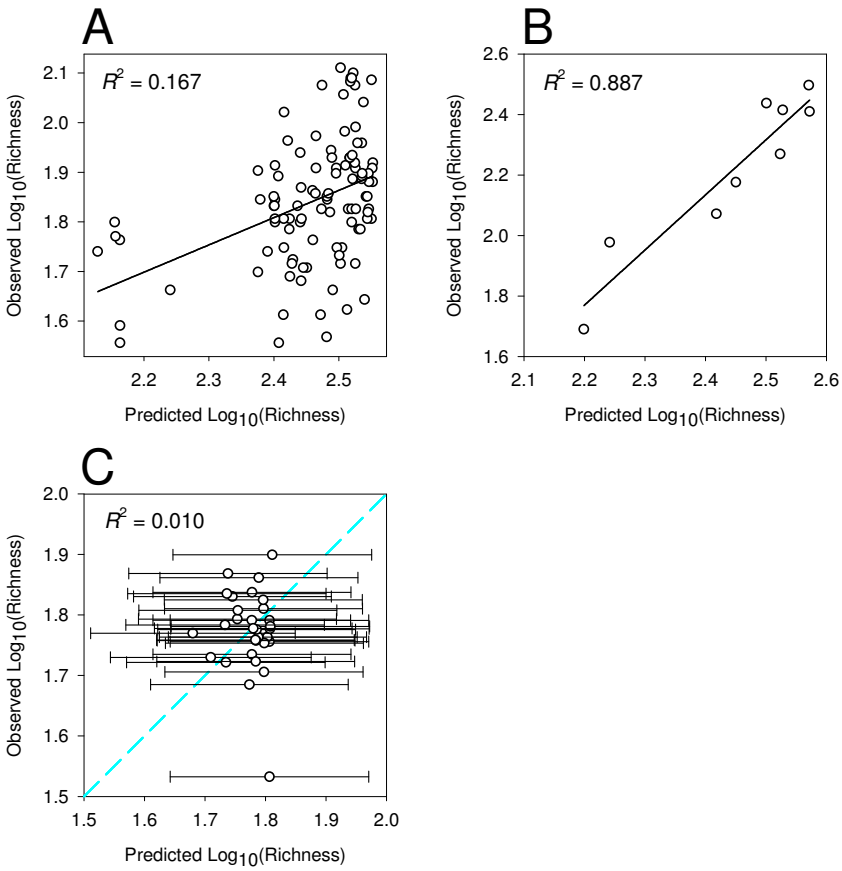


Figure S32: **Independent model validation.** In (A) and (B), a linear model was fit using MICROBIS sequences classified with a *de novo* clustering (Analysis I; see Table S5). This model was then used to predict observations in two independent data sets; "FUHRMAN2008" and "POMMIER2007" (Fuhrman et al., 2008; Pommier et al., 2007). In (A), the model fit with MICROBIS had good predictive power, similar to the predictive power to models fit using FUHRMAN2008 in the original publication (Fuhrman et al., 2008) ($R^2 = 0.2$). In (B) the model fit with MICROBIS has excellent predictive power. Thus, the model that we used has good predictive power, despite the independent data sets being collected at different locations at different times, using different methodology and with different sequencing depths S1. The model had relatively little predictive power for the Global Ocean Survey data, but (C) 95% prediction intervals covered observed richness values 96.8% of the time when the model is fit using RDP-classified MICROBIS sequences at a rarefaction depth of 150 sequences.

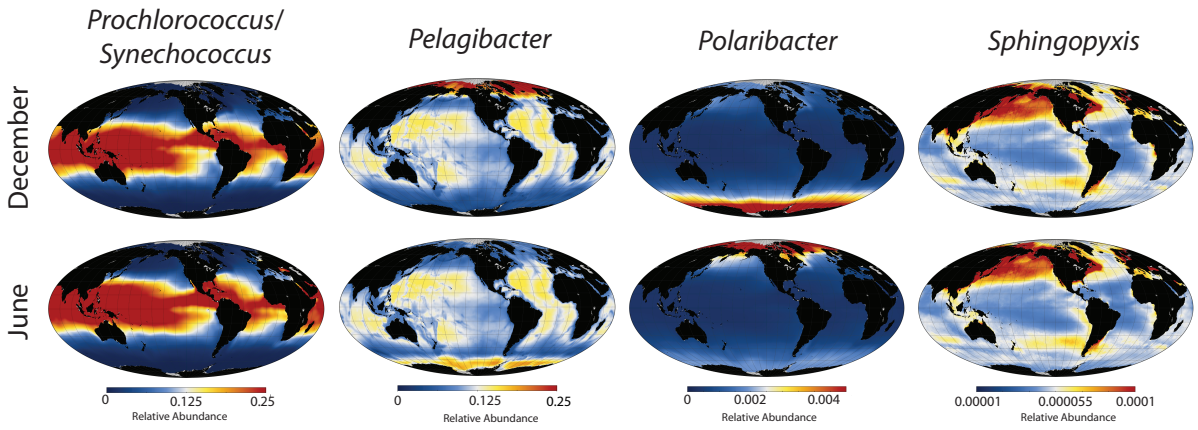


Figure S33: **Range maps of representative genera.** This is a color version of Figure 2 in the text.

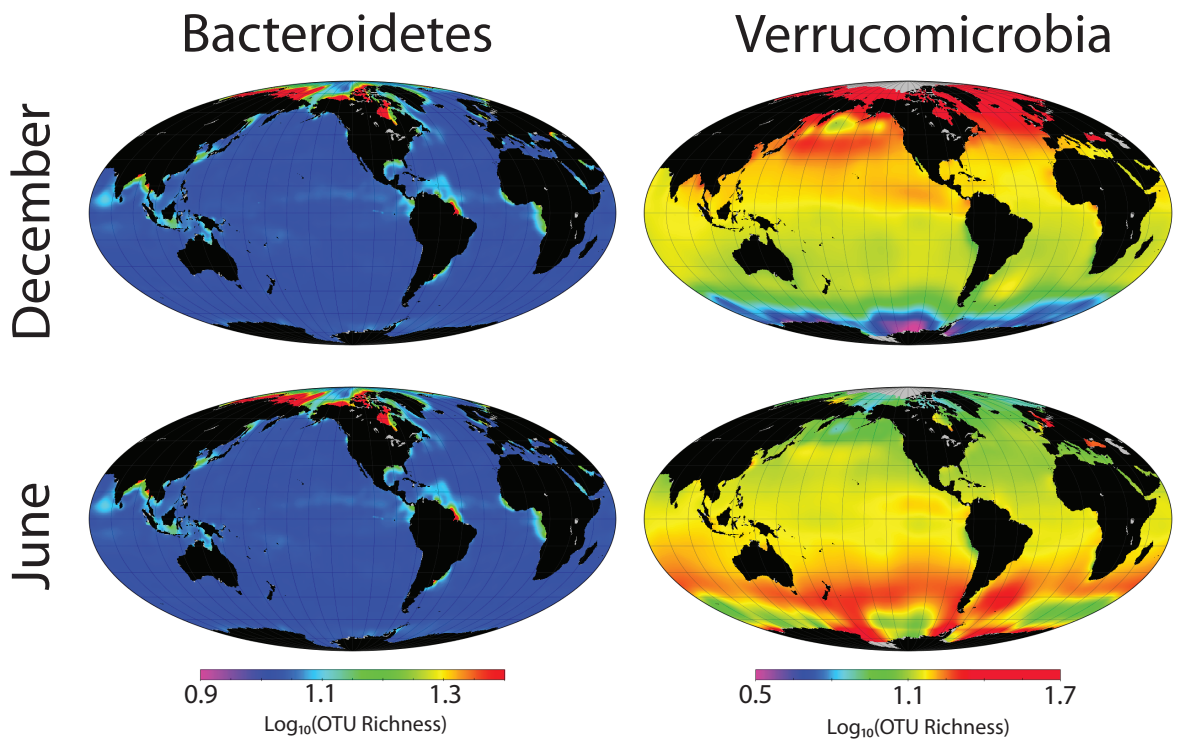


Figure S34: **Patterns of OTU richness within bacterial phyla.** Columns show maps for different phyla; rows for different seasons. Maps are for phyla not shown in Fig. 3; see caption for a complete description.

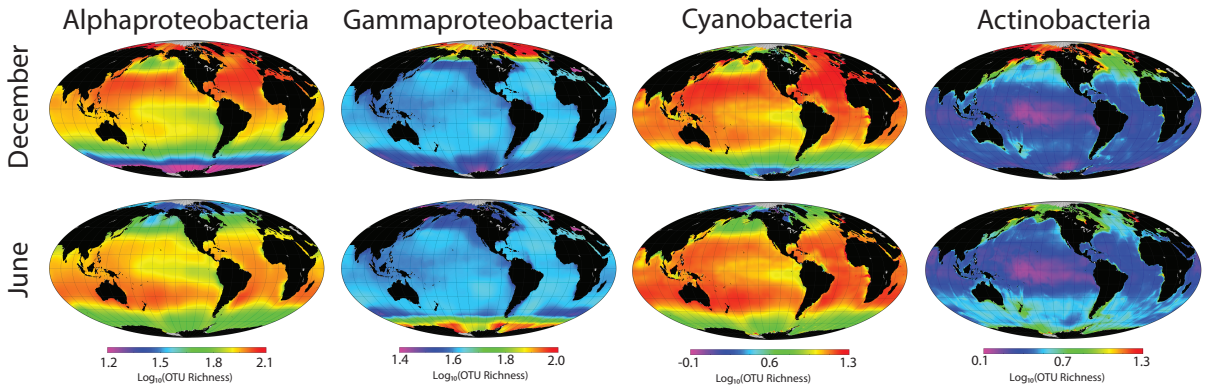


Figure S35: **Patterns of OTU richness within bacterial phyla.** This is a color version of Figure 3 in the text.

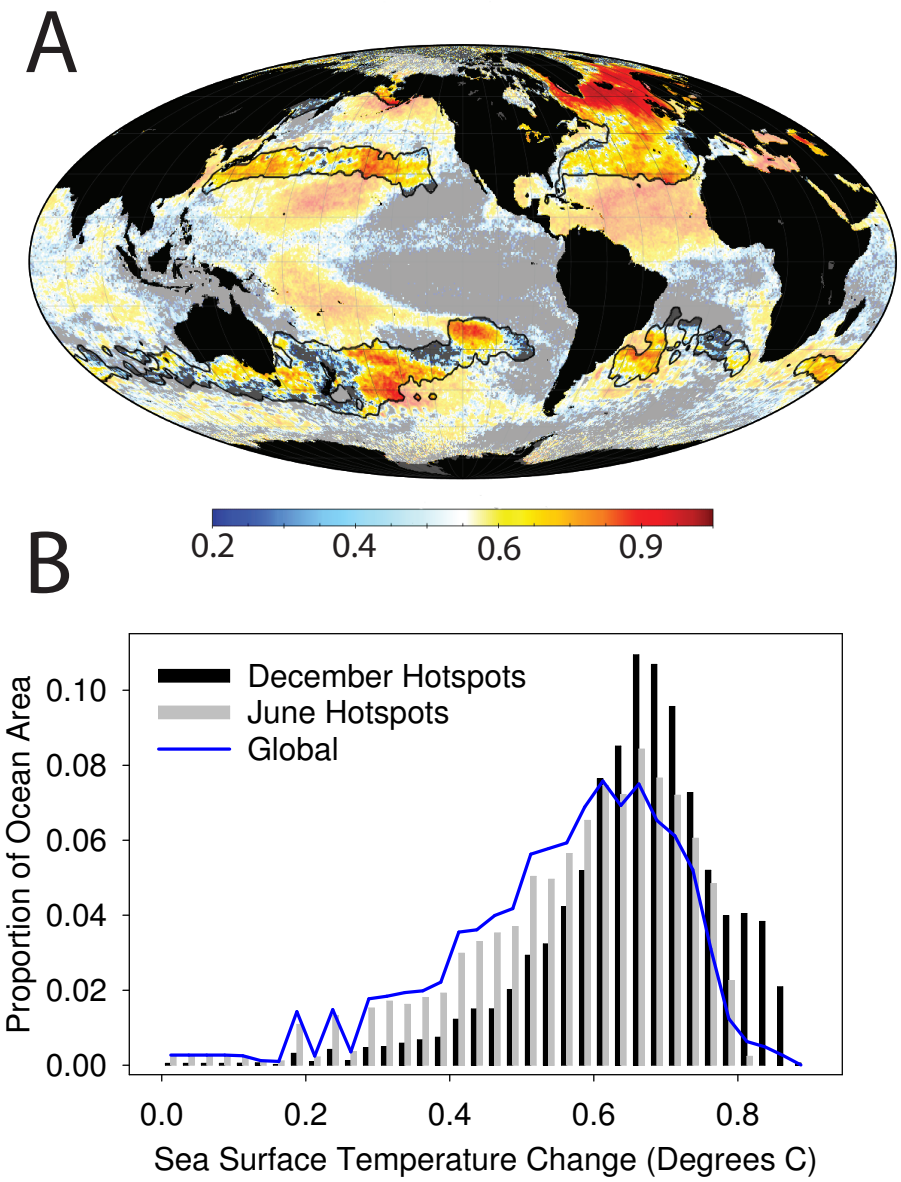


Figure S36: **Bacterial diversity hotspots and sea surface temperature.** (**a**) Hotspots of marine bacterial richness overlaid on a map of sea surface temperature (SST) increases, quantified by comparing weekly SST values between the late 1980s and the early 2000s (Halpern et al., 2008). Hotspots are outlined with black borders, and are defined as the 10% of ocean surface with the greatest diversity in December and June (primarily in the Northern and Southern hemispheres, respectively). (**b**) The distribution of SST increases across the entire ocean and within December and June diversity hotspots.

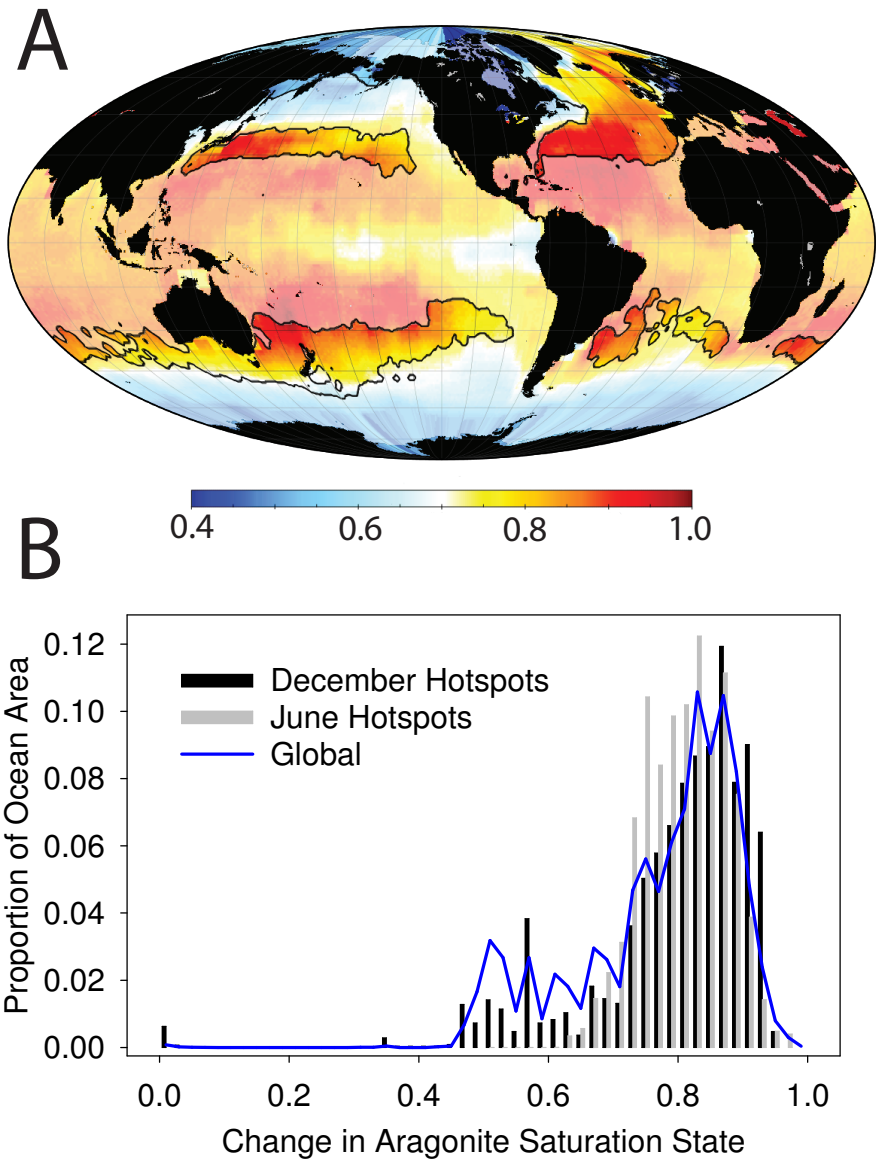


Figure S37: **Bacterial diversity hotspots and acidification.** (**a**) Hotspots of marine bacterial richness overlaid on a map of ocean acidification, measured as the change in aragonite saturation state (ASS) (Halpern et al., 2008). Hotspots are outlined with black borders, and are defined as the 10% of ocean surface with the greatest diversity in December and June (primarily in the Northern and Southern hemispheres, respectively). (**b**) The distribution of ASS values across the entire ocean and within December and June diversity hotspots.

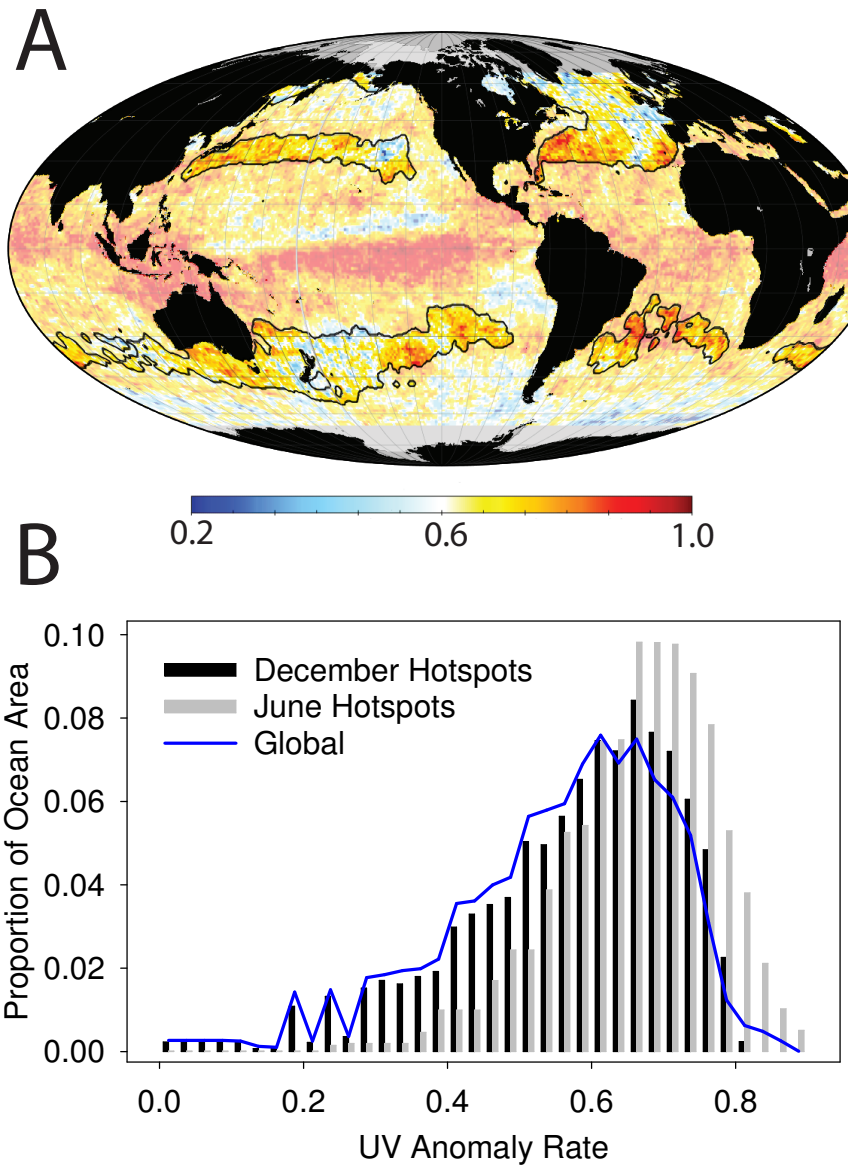


Figure S38: **Bacterial diversity hotspots and ultraviolet radiation.** (**a**) Hotspots of marine bacterial richness overlaid on a map of ultraviolet (UV) radiation anomalies, measured as the rate of deviations greater than one standard deviation from the average Joules per square meter per month (Halpern et al., 2008). Hotspots are outlined with black borders, and are defined as the 10% of ocean surface with the greatest diversity in December and June (primarily in the Northern and Southern hemispheres, respectively). (**b**) The distribution of UV anomaly values across the entire ocean and within December and June diversity hotspots.