Supporting Information

Kelly et al. 10.1073/pnas.1403319111

SI Materials and Methods

Sample Collection. Samples were collected at 22 sites distributed across the 11 major atolls in the Line Island archipelago, 1–4 sites per atoll. Seawater samples of ~100 L were collected at the surface of representative benthos (including within crevices, when present) across ~20 m^2 of reef using a modified bilge pump. Samples were collected directly into low-density polyethylene collapsible bags (19 L; Cole-Parmer) and transported to the research vessel within 2 h. Before sampling, all containers, bilge pumps, and tubing were washed once with 1% bleach and 0.1 M NaOH, three times with freshwater, and once with 100 kDa filtered seawater. Samples were filtered through 20 μm Nitex to remove large eukaryotes. The filtrate was concentrated to <500 mL using a 100-kDa tangential flow filter, which retained the unicellular eukaryotes, microbes, and virus-like particles. The microbial fraction was collected by passing this concentrated sample through 0.45-μm Sterivex filters (Millipore, Inc.) and the filters were then stored at −80 °C.

While sampling the reef surface (above), water was also collected for nutrient analysis directly above the same reef area, within 20 cm of the reef surface, using diver-deployable polycarbonate Niskin bottles. Water samples were filtered through 0.2-μm Nuclepore Track-Etched membrane filters (Whatman) into 20 mL high-density polyethylene scintillation vials with cone-shaped plastic lined lids (Fisher Scientific) and then stored at −20 °C. Inorganic nutrient (nitrate + nitrite, nitrite, and phosphate) concentrations were measured using a QuikChem 8000 flow injection analyzer (Lachat Instruments) at the Marine Science Institute Analytical Laboratory (University of California, Santa Barbara).

Characterization of the benthic community was completed using photoquadrats (1). Two 25-m transect lines were quantified per site and ten 0.72-m^2 quadrats were assessed along each transect line using digital underwater photographs. Images were analyzed using the program Photogrid 1.0, where 100 stratified random points were identified to determine benthic community composition at each site. All organisms were characterized to the finest level of resolution possible (genus level for corals and macroalgae and functional group for turfing and crustose coralline algae). All surveys took place at 10 m depth on the fore-reef habitat of each atoll.

DNA Extraction and Metagenomic Library Construction. DNA was extracted and purified using a column purification protocol (NucleoSpin Tissue; Macherey-Nagel), modified to complete the lysis steps in the Sterivex filters. Lysates were removed from the Sterivex filters using a 3-mL Luer-Loc syringe. The rest of the extraction procedure was performed according to the manufacturer's recommendations. Metagenomic libraries were prepared using a GS FLX Titanium Rapid Library Preparation Kit (Roche Applied Sciences) and pyrosequenced using a 454 GS-FLX platform at San Diego State University.

Sequence Library Quality Control and Bioinformatics Analyses. Metagenomic sequence reads were filtered for quality using the Preprocessing and Information of Sequences tool, PRINSEQ (2), uploaded to the MG-RAST server (http://metagenomics.nmpdr.org/metagenomics.cgi), and compared with the SEED protein database using BLASTx (3). For taxonomic annotation, sequences with significant similarities ($E < 10^{-5}$) were assigned to the closest identified microbial representative. For functional annotation, sequences were assigned the function of the closest identified protein and these functions were then grouped into metabolic pathways according to the subsystems in the SEED database (4). These sequences are publically available through the MG-RAST server under the project name Pacific Reef Microbiomes (http://metagenomics.anl.gov/linkin.cgi?project=9220).

Statistics. Nonmetric multidimensional scaling (nMDS) analyses were used with the annotated metagenome data to visualize between-atoll similarity in terms of two discrete response variables: community structure and community metabolism. Community structure was determined by comparing the relative abundances of 19 higher-rank microbial taxa (to limit the number of taxonomic categories to avoid type I errors associated with loss of statistical power in multiple comparisons; see Table S7 for clarification of taxonomic groups), averaged by atoll. Similarly, community metabolism was determined by comparing the relative abundance of 20 level 1 subsystem categories in the SEED database (http://theseed.org/wiki/Main_Page). Significant groupings of atolls depicted by the nMDS were quantified using a similarity profile test based on the Bray–Curtis algorithm ($P < 0.01$) (similarity profile analysis or SIMPROF) (5), using the clustsig package (6) for R (R Development Core Team). Analyses were based on 10,000 random permutations of the annotated metagenomic data. These significant groupings designated by SIMPROF were then superimposed upon the nMDS plots. Individual variables that might be responsible for driving group differences in multivariate space were investigated by calculating Spearman's rank correlations and those with strong correlations (in this study, ≥0.6) plotted as vectors in the nMDS plots.

For an initial exploration of potential correlations between the three predictor variables and either microbial community structure or metabolism, a canonical correspondence analysis (CCA) was performed using the R package, vegan (7). The results from this analysis were visualized by plotting the CCA loading vectors. To formally quantify how much variation in the microbial communities or their metabolism could be explained by the predictors measured (continuous variables), a permutational distance-based multivariate linear model (DistLM) (8) was used in PERMANOVA+ (www.primer-e.com/permanova.htm). To determine their suitability for use in a linear model, collinearity of the predictor variables was tested by calculating pairwise Pearson correlation coefficients. No two predictors exceeded a correlation of 0.75 (Table S6); therefore, all were included in the model. Model selection (balancing performance with parsimony) was based on Akaike's information criterion (AIC) (9) with a second-order bias correction applied (AICc) (10). Significance was determined by comparing the model results obtained with the original data structure to those obtained with 10,000 random permutations of the raw data. Statistical analyses were performed using R Version 2.15.1 (R Development Core Team, www.r-project.org) (11) unless otherwise stated.

The program Xipe (12) was used to determine lower level taxa and level 3 subsystem metabolic pathways that were significantly different ($P < 0.05$; 1,000 iterations) between metagenomic libraries sampled from all 11 atolls. Its bootstrapping technique allows comparison of thousands of gene categories between two metagenomic libraries with a designated confidence threshold (e.g., 95%). The Xipe findings were further tested for correlation with distance from the equator, nutrient concentration, and percentage cover of the seven benthic functional groups by calculating Pearson correlation coefficients (r) in SPSS (IBM Corporation).

1. Sandin SA, et al. (2008) Baselines and degradation of coral reefs in the northern Line Islands. *PLoS ONE* 3(2):e1548.
2. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27(6):863–864.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
4. Meyer F, et al. (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386.
5. Clarke KR, Somerfield PJ, Gorley RN (2008) Testing of null hypotheses in exploratory community analyses: Similarity profiles and biota-environment linkage. *J Exp Mar Biol Ecol* 366(1-2):56–69.
6. Whitaker D, Christman M (2010) clustsig: Significant cluster analysis. R package Version 1.0. Available at www.r-project.org. Accessed June 20, 2014.
7. Oksanen JF, et al. (2012) vegan: Community ecology package. R package Version 2.0-4. Available at http://cran.r-project.org/index.html. Accessed June 20, 2014.
8. McArdle BH, Anderson MJ (2001) Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* 82(1):290–297.
9. Akaike H (1973) Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika* 60(2):255–265.
10. Hurvich CM, Tsai CL (1989) Regression and time-series model selection in small samples. *Biometrika* 76(2):297–307.
11. Gentleman R, Ihaka R (1996) The R language. *Proceedings of 28th Symposium on the Interface of Computing Science and Statistics (Graph-Image-Vision)*, eds Billard L, Fisher NI (Interface Found, Sydney, Australia), pp 326–330.
12. Rodriguez-Brito B, Rohwer F, Edwards RA (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics* 7:162.

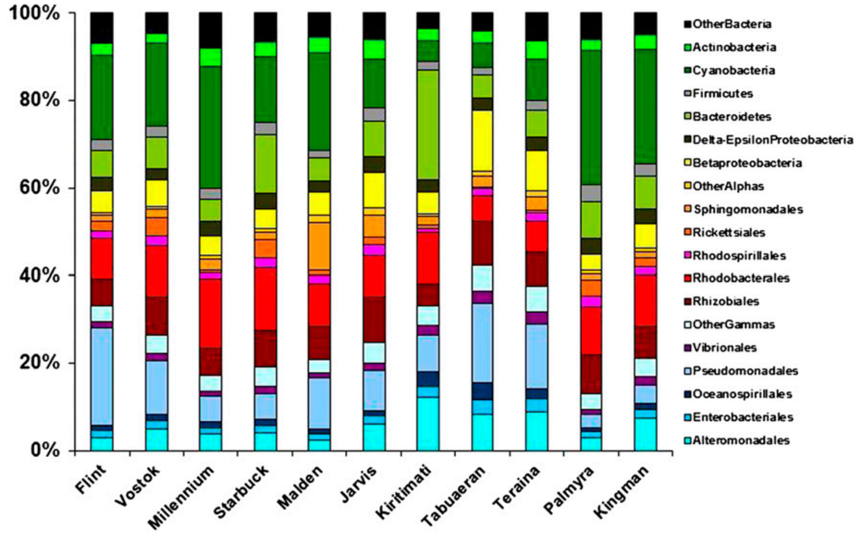

Fig. S1. The relative abundance of bacterial groups across the LI. Reads in the 22 metagenomes were taxonomically annotated by comparison with the SEED database and averaged by atoll. Atolls on the *x* axis are ordered south to north, left to right.

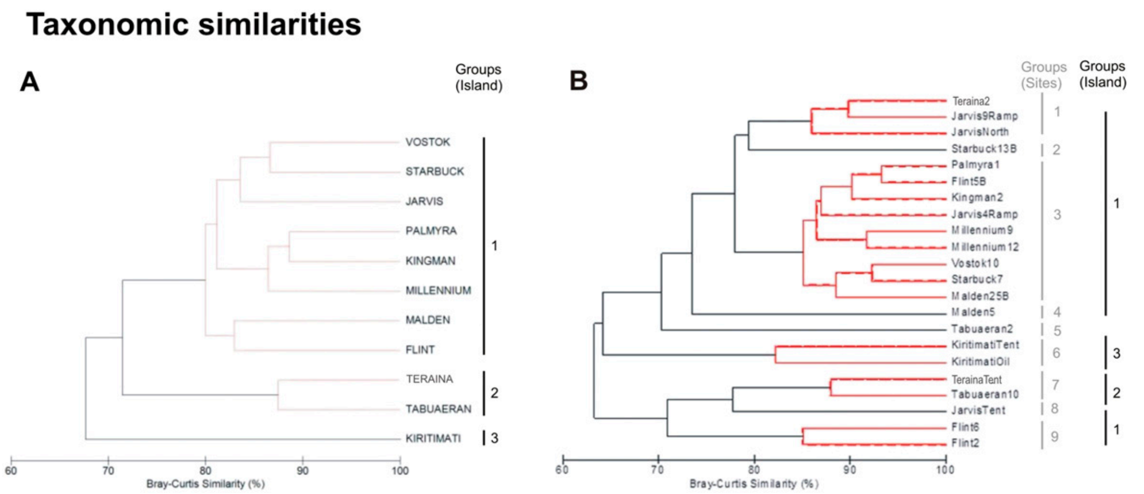

Fig. S2. Multivariate structure for the relative abundance of taxonomic similarities averaged by island (*A*) and at site level (*B*) analyzed using SIMPROF ($P < 0.01$).

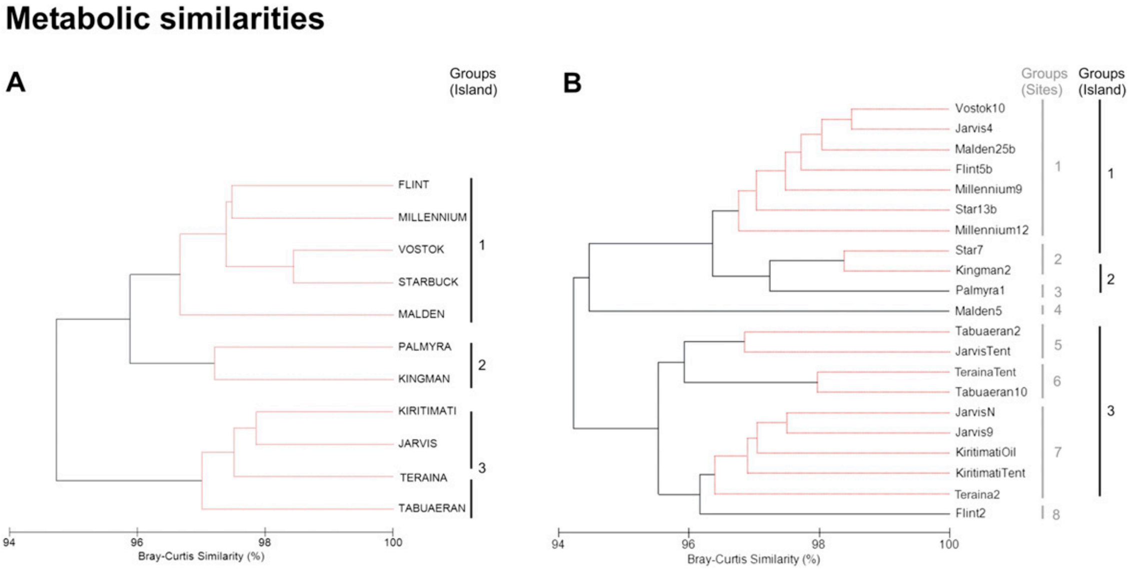

Fig. S3. Multivariate structure for the relative abundance of metabolic groupings averaged by island (*A*) and at site level (*B*) analyzed using SIMPROF ($P < 0.01$).

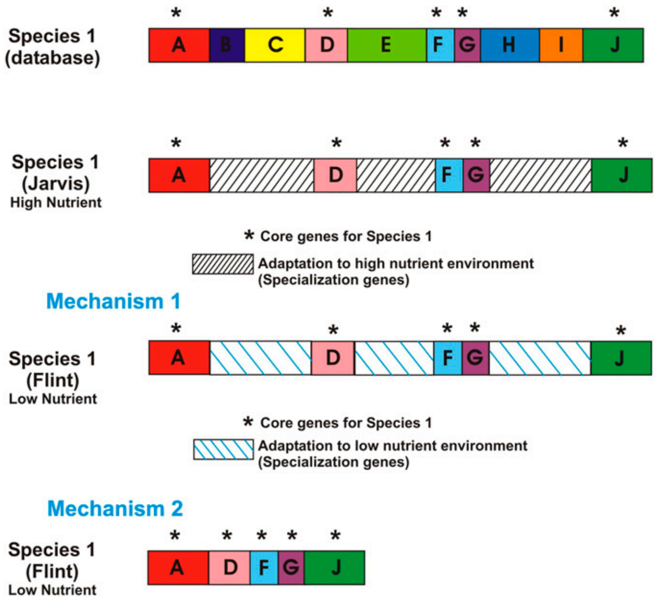
Species 1 (database)
A
B
C
D
E
F
G
H
I
J
Species 1 (Jarvis) High Nutrient
A
D
F
G
J
* Core genes for Species 1
Adaptation to high nutrient environment (Specialization genes)
Mechanism 1
Species 1 (Flint) Low Nutrient
A
D
F
G
J
* Core genes for Species 1
Adaptation to low nutrient environment (Specialization genes)
Mechanism 2
Species 1 (Flint) Low Nutrient
A
D
F
G
J

Fig. S4. Conceptual model depicting variation in genome size between strains due to different complements of environment-dependent specialization genes.

Table S1. Metagenomic libraries

Sample name	Date collected	Total no. of reads	Average read, bp	Latitude, °	Longitude, °	% GC content	Total no. of taxon similarities	Total no. of metabolism similarities
Flint 2	03/30/09	24,111	345.71	−11.41924	−151.82739	51	8,454	10,167
Flint 5	03/29/09	98,284	399.55	−11.43911	−151.81964	47	28,931	37,301
Flint 6	03/31/09	39,070	430	−11.44423	−151.81709	47	17,316	20,252
Jarvis 4	04/04/10	171,749	400.17	−0.38188	−159.99800	48	49,941	61,987
Jarvis 9	04/02/10	235,984	407.55	−0.36537	−160.00600	50	63,207	77,340
Jarvis North	11/13/10	66,808	384.81	−0.36902	−160.00819	52	20,340	23,781
Jarvis Tent	11/12/10	49,774	397.8	−0.369017	−160.00819	50	22,465	24,945
Kingman 2	10/31/10	225,914	381.44	6.387	−162.38600	52	47,187	61,909
Kiritimati Oil	11/21/10	156,251	393.74	1.99095	−157.48251	51	54,015	62,427
Kiritimati Tent	11/20/10	30,131	387.29	2.0085833	−157.48945	50	11,457	12,895
Malden 25	04/11/09	164,564	381.72	−4.03326	−154.95094	52	42,411	51,614
Malden 5	04/10/09	48,258	349.25	−3.99531	−154.94452	57	10,993	12,902
Millennium 12	04/19/09	26,895	357.99	−9.90774	−150.19974	53	7,801	9,190
Millennium 9	04/17/09	39,933	373.89	−9.91672	−150.21072	54	13,032	15,772
Palmyra 1	10/25/10	170,135	386.57	5.86646	−162.11346	37	53,623	71,606
Starbuck 13	04/05/09	29,347	401.56	−5.66441	−155.87346	46	11,874	15,052
Starbuck 7	04/06/09	83,014	431.87	−5.62220	−155.88002	42	34,058	45,652
Tabuaeran 10	11/04/10	104,845	396.09	3.82595	−159.34957	56	39,697	46,930
Tabuaeran 2	11/06/10	73,712	411.89	3.84085	−159.36047	55	31,874	36,241
Teraina 2	11/09/10	42,317	385.46	4.70242	−160.39212	53	12,035	14,199
Teraina Tent	11/08/10	285,841	412.73	4.6867167	−160.42023	51	86,972	101,107
Vostok 10	04/01/09	83,219	357.47	−10.05835	−152.30954	44	32,232	41,533
Total	—	2,250,156	—	—	—	—	699,915	854,802

Metadata and library details for the 22 metagenomes generated from the 22 sites sampled at 11 atolls.

Table S2. Predictor variable categories used for CCA and DistLM

Sample name	Hard coral	Crustose coralline algae	Calcified macroalgae	Soft coral	Macroalgae	Turf	Other*	NO$_3^-$ + NO$_2^-$, µM	PO$_4^3$, µM	NH$_4^+$	Distance from equator[†]
Flint 2	75.85	13.10	2.00	0.00	1.15	7.65	0.25	1.09	0.291	4.32	−11.41924
Flint 5	83.00	9.00	0.20	0.00	0.95	6.60	0.25	0.82	0.161	2.46	−11.43911
Flint 6	ND	ND	ND	ND	ND	ND	ND	0.79	0.147	2.13	−11.44423
Jarvis 4	46.30	27.20	0.70	0.00	7.90	17.70	0.30	4.65	0.384	ND	−0.38188
Jarvis 9	57.90	9.30	0.00	0.00	3.40	29.20	0.20	4.54	0.427	ND	−0.36537
Jarvis North	10.70	31.90	1.00	0.00	4.90	50.30	1.30	4.50	0.441	ND	−0.36902
Jarvis Tent	57.65	12.00	0.15	0.35	6.90	21.65	0.95	3.28	0.392	0.289	−0.369017
Kingman 2	14.55	51.20	12.70	0.40	0.95	18.25	1.95	1.44	0.247	0.203	6.387
Kiritimati Oil	2.21	0.00	6.68	0.00	4.00	82.47	4.63	2.26	0.241	1.15	1.99095
Kiritimati Tent	23.36	2.50	7.77	1.64	0.14	58.86	5.73	2.33	0.291	0.436	2.0085833
Malden 25	73.63	5.37	1.84	0.00	1.79	15.95	1.42	3.90	0.264	1.67	−4.03326
Malden 5	86.67	4.56	0.00	0.00	0.22	8.11	0.44	2.82	0.253	1.24	−3.99531
Millennium 12	65.30	11.00	12.30	0.00	1.10	10.10	0.20	2.28	0.216	0.674	−9.90774
Millennium 9	69.30	6.20	7.70	0.00	0.30	15.70	0.80	2.10	0.191	0.803	−9.91672
Palmyra 1	45.70	16.30	5.60	2.00	1.40	23.90	0.60	0.52	0.195	0.365	5.86646
Starbuck 13	25.55	12.35	57.95	0.00	0.10	0.20	3.85	2.87	0.247	1.93	−5.66441
Starbuck 7	21.68	49.42	21.84	0.00	1.21	4.47	1.21	4.83	0.254	3.18	−5.6222
Tabuaeran 10	22.08	30.08	34.62	0.00	8.69	3.92	0.46	2.78	0.299	0.630	3.82595
Tabuaeran 2	39.23	20.18	16.32	0.00	2.41	17.86	4.00	1.60	0.185	0.936	3.84085
Teraina 2	20.96	21.96	0.00	0.76	32.12	20.44	3.76	2.24	0.279	0.458	4.70242
Teraina Tent	8.64	44.93	6.64	6.36	2.36	30.79	0.29	1.92	0.278	0.348	4.6867167
Vostok 10	81.40	14.40	0.35	0.00	0.15	3.70	0.00	1.75	0.158	1.91	−10.05835
Average	44.4	18.7	9.35	0.548	3.91	21.3	1.55	2.51	0.266	1.32	—

Benthic coverage for each functional group is shown as percent cover. Nutrient concentrations are calculated as micromoles per liter. NH$_4^+$, ammonium; NO$_3^-$, nitrate + nitrite; PO$_4^3$, phosphate.

*Other benthic organisms.

[†]Distance from equator as absolute value of the latitude in decimal degrees.

Table S3. Taxa within the seven functional groups used to classify benthic macroorganisms

Hard coral
Acropora, Astreopora, Cyphastrea, Cycloseris, Echinophyllia, Favia, Favities, Fungia, Gardineroseris, Halomitra, Herpolitha, Hydnophora, Leptastrea, Leptoseris, Lobophyllia, Montastrea, Montipora, Pavona, Platygyra, Pocillopora, Porites, Psammocora, Sandolitha, Scapophyllia, Sytlophora, Tubastrea, Turbinaria
Calcified macroalgae
Galaxaura, Halimeda, Neomeris, Peyssonellia
Soft coral
Cladiellqa, Dendronephtya, Lobophytum, Pachyclavularia, Sarcophyton, Sinularia, Stereonephthya
Fleshy macroalgae
Avrainvillea, Brown crust, *Caulerpa, Dictyosphaeria, Dictyota, Hypnea, Lobophora, Valonia*
Other benthic organisms
Cyanobacteria, Heteractis, Holothurian, Hydroid, *Millepora, Rhodactis,* Sand, Sponge, *Stylaster, Tridacna,* Tunicate, Zoanthid

Crustose coralline algae and fleshy turf algae were identified as functional groups only.

Table S4. Summarized results of a DistLM for associations of microbial community structure (Taxa) and metabolic function (Metabolism)

Variable	AICc	SS, trace	Pseudo-*F*	*P*	Prop., %*	res.df
Taxa						
Hard coral	129.52	1,438.1	3.4065	0.0215	15.2	19
Metabolism						
Distance from equator	53.323	47.953	4.2767	0.0147	18.4	19

Prop., proportion of variance; res.df, degrees of freedom for the residual; SS, sum of squares. The total number of predictors included equals 10 (the percent cover of benthic functional groups, distance from the equator, and nutrient availability).

*The best-fit models are shown, along with the proportion of variability in the multivariate response explained by that variable (Prop.).

Table S5. Significance test for the linear correlations of metabolic pathway abundance with phosphate concentration

Metabolic pathway	*r*	*P*
Conjugative transfer	0.863	**0.001**
Bacterial chemotaxis	0.598	0.052
Nitrate and nitrite ammonification	0.628	**0.038**
Cobalt–zinc–cadmium resistance	0.637	**0.035**
Multidrug resistance	0.617	**0.043**
Ton and Tol transport	0.650	**0.03**
Chlorophyll biosynthesis	−0.552	0.079
Photosystem II	−0.534	0.091
Ribosome SSU bacterial	−0.620	**0.042**

P values $<$ 0.05 are shown in bold.

Table S6. Collinearity among predictor variables using Pearson's coefficient, *r*

	Hard coral	Crustose coralline algae	Other calcified algae	Soft coral	Macroalgae	Turf algae	Other benthic	Nitrate	Phosphate
Hard coral									
Crustose coralline algae	−0.536								
Other calcifying algae	−0.373	0.193							
Soft coral	−0.355	0.339	−0.089						
Macroalgae	−0.287	0.120	−0.151	0.001					
Turf algae	−0.533	−0.196	−0.297	0.213	0.051				
Other benthic	−0.548	0.196	0.294	−0.030	0.211	0.538			
Nitrate	−0.161	0.163	0.038	−0.025	0.105	0.077	−0.077		
Phosphate	−0.266	0.155	−0.175	−0.017	0.258	0.299	−0.088	**0.708**	
Distance from equator	0.487	−0.072	0.054	−0.071	−0.250	−0.531	0.316	**−0.641**	**−0.741**

The correlations between distance from equator and nutrient concentrations are shown in bold.

Table S7. The 19 bacterial taxa included in analyses of community structure

Phyla	Classes	Orders
Actinobacteria		
Bacteroidetes		
Cyanobacteria		
Firmicutes		
Proteobacteria	Alphaproteobacteria	Rickettsiales
		Rhizobiales
		Rhodobacterales
		Rhodospirillales
		Sphingomonadales
		Other
	Betaproteobacteria	
	Gammaproteobacteria	Alteromonadales
		Enterobacteriales
		Oceanospirillales
		Pseudomonadales
		Vibrionales
		Other
	Deltaproteobacteria	
	Epsilonproteobacteria	
Other		

Bacterial taxa were categorized at the phylum level except for the Proteobacteria (which made up 48–87% of the bacterial community). Rarer phyla (those that made up <5% of the relative abundance across all libraries) and unclassified bacteria were designated as "Other bacteria." Because of their higher abundances, Proteobacteria were categorized by class, and Gammaproteobacteria and Alphaproteobacteria (representing 13–64% and 12–36% of the bacterial communities, respectively) were further categorized by order. Rarer Alphaproteobacteria and Gammaproteobacteria orders that made up <1% across all libraries were combined and designated as "Other Alphaproteobacteria" and "Other Gammaproteobacteria."