Bacterial community assembly based on functional genes rather than species

Catherine Burke[a,b], Peter Steinberg[c,d], Doug Rusch[e], Staffan Kjelleberg[a,f], and Torsten Thomas[a,1]

[a]School of Biotechnology and Biomolecular Sciences, [c]School of Biological, Earth and Environmental Sciences, Centre for Marine Bio-Innovation, University of New South Wales, Sydney, New South Wales 2052, Australia; [b]The iThree Institute, University of Technology, Ultimo, New South Wales 2007, Australia; [d]Sydney Institute of Marine Science, Mosman, New South Wales 2088, Australia; [e]The J. Craig Venter Institute, Rockville, MD 20850; and [f]Singapore Centre on Environmental Life Sciences Engineering, Nanyang Technological University, Singapore

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved July 14, 2011 (received for review January 30, 2011)

The principles underlying the assembly and structure of complex microbial communities are an issue of long-standing concern to the field of microbial ecology. We previously analyzed the community membership of bacterial communities associated with the green macroalga *Ulva australis*, and proposed a competitive lottery model for colonization of the algal surface in an attempt to explain the surprising lack of similarity in species composition across different algal samples. Here we extend the previous study by investigating the link between community structure and function in these communities, using metagenomic sequence analysis. Despite the high phylogenetic variability in microbial species composition on different *U. australis* (only 15% similarity between samples), similarity in functional composition was high (70%), and a core of functional genes present across all algal-associated communities was identified that were consistent with the ecology of surface- and host-associated bacteria. These functions were distributed widely across a variety of taxa or phylogenetic groups. This observation of similarity in habitat (niche) use with respect to functional genes, but not species, together with the relative ease with which bacteria share genetic material, suggests that the key level at which to address the assembly and structure of bacterial communities may not be "species" (by means of rRNA taxonomy), but rather the more functional level of genes.

lateral gene transfer | biofilm | ecological model

Metagenomic analysis of environmental microbial communities has revealed an enormous and previously unknown microbial diversity, and expanded our knowledge of their function in a variety of environments (1–5). Much still remains unknown, however, such as the principles underlying the assembly and structure of complex microbial communities, an issue of long-standing concern to the field of microbial ecology. To this aim, several recent studies have supported the "neutral hypothesis" (6–8), a largely stochastic model for community assembly, which assumes that species are ecologically equivalent and that community structure is determined by random processes (9, 10). However, there is also evidence that niche or deterministic processes play a role in community structure (11, 12); thus, both niche and neutral processes are likely to affect the assembly of complex microbial communities.

Support for these models is based on species abundance distributions, and critical functional aspects, such as the assumption of ecological equivalence, have for the most part not been tested. In this study, we examine the encoded functions of an algal-associated bacterial community and link patterns of function to patterns of community assembly. Following the results of an earlier study (13), we investigate these communities in the context of the lottery hypothesis, a model for community "assembly" derived from studies of eukaryotic communities, such as coral reef fish (14). This hypothesis incorporates both neutral and functional aspects and argues that ecological niches are colonized randomly from a pool of species with similar ecological function that can coexist in that niche (14, 15). Available space within that niche is colonized by whichever suitable species happens to arrive there first, meaning that colonization of space is random from within a functionally equivalent group of species. In the context of a bacterial community, this model implies that there are guilds of bacterial species, whose members are functionally equivalent with respect to their ability to colonize a particular niche (e.g., the surface of the seaweed *Ulva australis*), but that the composition of species in any particular community (e.g., a single *U. australis* individual) is determined stochastically by recruitment from within those guilds. Importantly, members of a guild can be phylogenetically related or unrelated. If this model is correct, different species from within these guilds should share functional traits, and a core suite of functional genes should be consistently present in all communities of a particular habitat, independent of the taxonomic or phylogenetic composition of its species.

In an earlier study (13) we characterized the bacterial phylogenetic diversity of seawater and *U. australis*, a member of a common green algal family often found in tidal rock pools or shelves around the world. We found that algal-associated communities were highly distinct from the surrounding seawater communities, but were also highly variable among individual algal samples, with only six operational taxonomic units (of a total of 528) at a 97% sequence identity cut-off occurring on all samples (13). This finding means that each *U. australis* sample hosts a unique assemblage of species (as defined by 97% 16S rRNA similarity). Given that the recruitment of new community members onto *U. australis* is most likely to come from the seawater, these results are somewhat contradictory with respect to dominant models of community assembly: the differences between seawater and algal communities imply selective mechanisms of assembly on the algal surface (niche partitioning), and the high variability between algal hosts is consistent with random colonization (e.g., neutral processes).

Here, we analyze the metagenomes of these communities to show that the algal-associated bacterial communities are functionally distinct from seawater communities, but contain a core of functional genes, which are represented across all algal samples. These functions are consistent with the ecology of surface- and host-associated bacteria, and importantly are distributed across a variety of taxa in individual communities, indicating functional redundancy across taxa. This mix of functional and random processes is consistent with the predictions of the com-

Author contributions: C.B., P.S., S.K., and T.T. designed research; C.B. and T.T. performed research; C.B., P.S., D.R., S.K., and T.T. analyzed data; and C.B., P.S., S.K., and T.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Shotgun sequencing data are available through the Community Cyber-infrastructure for AdvancedMicrobial Ecology Research and Analysis (accession: CAM_PROJ_BotanyBay).

[1]To whom correspondence should be addressed. E-mail: t.thomas@unsw.edu.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1101591108/-/DCSupplemental.

petitive lottery model for the assembly of complex microbial communities. Moreover, this functional (niche) partitioning with respect to genes, but not phylogeny, in these assemblages highlights the potential difficulty in using bacterial species to test hypotheses derived from eukaryotic ecology, which focus on species as the critical unit. Given the relative ease with which bacteria share genetic material, the key level at which to address the assembly of these bacterial communities may not be species, but rather the more functional level of genes or gene clusters.

Results and Discussion

Algal-Associated Bacterial Communities Encode a Distinct Functional Profile. We generated over 681 Mbp of metagenomic shotgun sequencing data from six algal (UA1–UA6) and eight seawater samples (SW3–SW10) (see *Materials and Methods* and Table S1). Environmental parameters were similar for each sample (see *Materials and Methods* and Table S2) and we found no evidence for increased viral or eukaryotic DNA between algal samples, indicating that these groups were a minor part of the microbial community at the time of sampling. Chao 1 estimates showed very similar levels of bacterial species diversity (ranging from 225 to 451) for the six algal communities (13). *U. australis* specimens were in the same developmental stage (i.e., fully matured) and the simple, two-cell layer alga is depauperate in bioactives found in many other algae, making it an ideal choice to minimize effects of host variability. Overall, these observations led us to the conclusion that bacterial communities on the *U. australis* samples collected existed under similar broad ecological conditions.

Sequencing data were used to create functional community profiles for each sample based on Clusters of Orthologous Groups (COG) (16) and SEED (17) annotations (details in Table S1). This process revealed that the algal-associated communities were functionally distinct from those in the surrounding seawater. Multidimensional scaling (MDS) plots show that the *U. australis* samples clustered together, and separately from seawater samples (Fig. 1*A*), and permutational multivariate analysis of variance (PERMANOVA) indicates that this difference is significant (P = 0.001). This distinctiveness was quantitative and multivariate, rather than qualitative, and there were few functions that were consistently present in one environment and consistently absent in the other. As such, the differences between the two environments lie predominantly in the relative abundance of particular functions as defined by COG and SEED annotations. Variation was also observed among seawater or *U. australis* samples, but this was primarily seen for those samples with the least (order-of-magnitude less) sequence data. When these samples were removed, the remaining algal samples still clustered together quite tightly (Fig. 1*B*). Bray-Curtis similarity shows that on average there was 15% similarity in species composition across samples (13), compared with 70% similarity in functional composition (COGs), indicating that despite large differences in species composition between hosts [as detected in our previous study (13)], many encoded functions are shared.

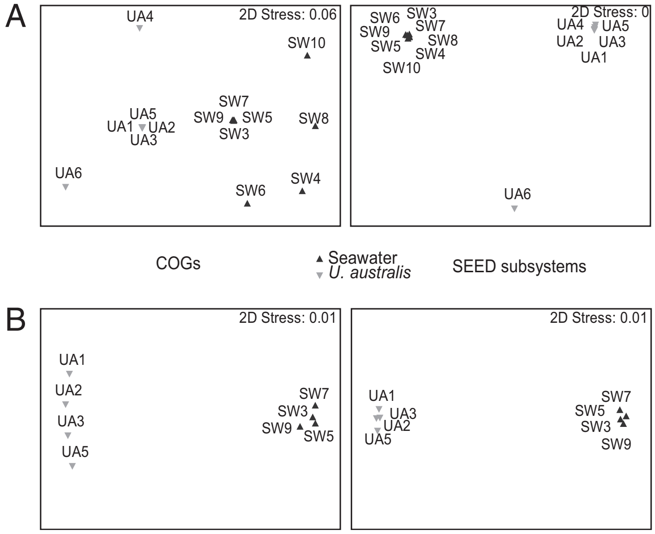

Fig. 1. MDS plots based on Bray-Curtis similarity of functional gene profiles based on COG and SEED annotations of *U. australis* and planktonic seawater metagenomes. (*A*) MDS plots for all samples, including UA4, UA6, SW4, SW6, SW8, and SW10, which contained an order-of-magnitude less sequencing than the remaining samples. (*B*) MDS plots with low coverage samples removed. Samples from each of the respective environments cluster together based on their functional profile.

Functional Core in *U. australis*-Associated Bacterial Communities. We observed a set of COG and SEED functions that contributed strongly to the difference between the two community types (Fig. S1 and *SI Materials and Methods*), and that were consistently abundant across *U. australis* samples. These were defined as the core functions of the *U. australis* associated bacterial communities (Fig. 2). These functions fit well with the current understanding of the ecology of an algal- or surface-associated bacterial community and could mostly be grouped into broader categories (Table S3) and are summarized here.

Detection and movement toward the host surface. Proteins associated with chemotaxis and flagellum-mediated motility were consistently abundant in the *U. australis* community, and are likely to be important for the detection and movement toward the algal host surface during colonization. Chemotaxis is essential for the development and maintenance of symbiotic, surface associations, as for example in symbiotic *Rhizobium* sp., which are chemotactically attracted to the flavenoids that induce the nodulation genes necessary for nitrogen fixation in the host plant (18). Flagella-mediated motility is also important for biofilm formation in a range of bacteria (19–22).

Attachment and biofilm formation. An array of proteins functionally assigned to attachment and biofilm formation were overrepresented across *U. australis* samples. Functions include homologs of the OmpA protein (COG2197), which is required for adhesion to both mammalian and fish epithelial cells in a range of *Proteobacteria* (23, 24), *Listeria* internalin-like proteins, which enhance attachment and biofilm formation (25, 26), and the widespread colonization island, which is essential for biofilm formation, colonization, and pathogenesis in a range of bacteria (27). Proteins related to the production and excretion of galactoglycan, or exopolysaccharide II, were more abundant in the *U. australis* community, and apart from forming part of the biofilm matrix, is also essential for the establishment and maintenance of symbiosis in several *Rhizobium* strains (28–32). GGDEF and EAL domain proteins, which are involved in the production and degradation, respectively, of bis-(3′-5′)-cyclic dimeric GMP (cyclic-di-GMP), were also detected at a higher abundance (33, 34). Cyclic-di-GMP is an important secondary messenger, regulating the transition from a motile planktonic to a surface-associated biofilm lifestyle in a range of bacteria (34, 35), for example by up-regulating the production of adhesins and biofilm matrix components (36–39) or downregulating motility genes (34).

Genes encoding for Cbb3-type cytochrome *c* oxidases, which have a high affinity for oxygen and are associated with microaerobic metabolism in oxygen-limited environments (40), were overrepresented, as well as the function of nitrite and nitrate ammonification, considered the highest energy-yielding respiration systems after oxygen has been depleted (41). The higher abundance of proteins assigned these functions may be related to survival in biofilms, which are known to be spatially heterogeneous, containing pockets of low or no oxygen in some areas (42).

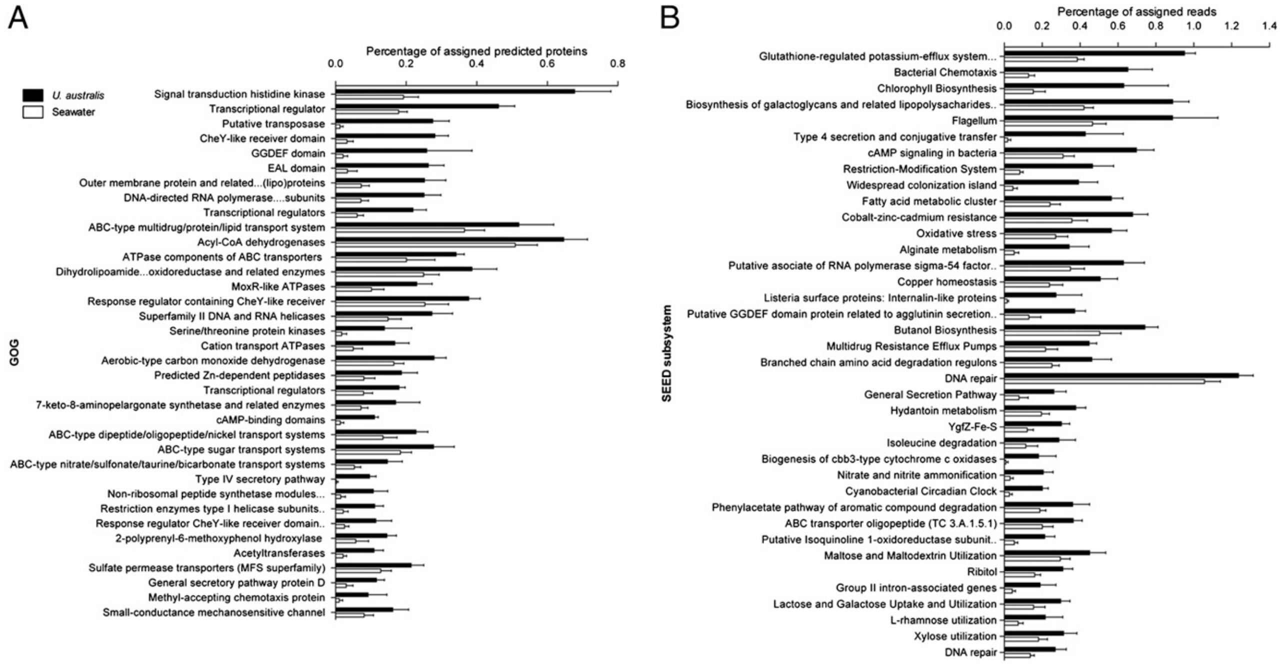

Fig. 2. COG (*A*) and SEED subsystems (*B*), which comprise the characteristic functions of the *U. australis* community, by comparison with planktonic seawater, with SD across the six algal and eight seawater samples analyzed. COGs and SEED subsystems are presented in the order of their contribution to difference (highest to lowest, *Top* to *Bottom*) as assessed by SIMPER analysis.

Response to the algal host environment. Some of the overrepresented functions can be related directly to *U. australis*' surface environment. For example, proteins associated with the metabolism of water-soluble polysaccharides produced by *Ulva* sp., such as rhamnose, xylose, glucose, mannose, and galactose (43), would enable bacteria to use these sugars as a source of carbon and energy and, hence, gain a competitive advantage in colonizing the host surface. Proteins associated with heat and osmotic stress (e.g., COG0668: small conductance mechanosensitive channel) (44–46) may be related to the desiccation of *U. australis* fronds, which occur in their intertidal habitats at low tide. Macroalgae also defend themselves against bacteria and herbivores by the release of reactive oxygen species, an oxidative burst (47), and the overrepresentation of proteins controlling oxidative stress might represent a protective mechanism for the surface community. Finally, *Ulva* sp. is known to take up and store heavy metals (48), and the overrepresentation of proteins associated with the export of heavy metals could thus be related to the presence of heavy metals in the algal host.

Regulation in response to environmental stimuli. Some of the most overrepresented functions found in the microbial community of *U. australis* involve regulatory mechanisms in response to environmental stimuli. The high proportion of environmental signal transducers and transcriptional regulators could be indicative of the need to respond to changes in the host environment (e.g., osmotic; see above), to control the complex steps of colonization or biofilm formation and to mediate the interactions with other community members. For example, there are many homologs of COG0642 (histidine kinase), members of which are known to be involved in osmoregulation (49), multidrug export (50), sporulation (51), nitrate reduction (52), cell differentiation (53), and plant virulence (54). Also overrepresented is COG0583 (transcriptional regulator), which contains proteins that regulate transcription in response to plant exudates (55), virulence, motility, and quorum sensing (56). There are many examples of plants and their metabolites affecting gene regulation of associated bacteria (57–61), and it is likely that such regulators play an important part in mediating interactions between *U. australis* and the surface community.

Lateral gene transfer. Type IV secretion system (T4SS), transposases, and intron-associated genes were overrepresented in *U. australis* samples. These functions are associated with lateral gene transfer, a source of dynamic genomic change that allows for rapid ecological adaptation (62), and which would provide a broad mechanism for facilitating the functional similarity of phylogenetically distinct bacteria on the surface of *U. australis* (see below). Biofilms are ideal environments for lateral gene transfer (63), and an abundance of transposases has been noted in other biofilm communities (64). A recent survey of the transposase families in different taxa suggested that transposases are most often transferred to other organisms within the same habitat and can be shared by distantly related taxa (65), and COG2801 (putative transposase) is a potential source of shared functional traits among different taxa in the *U. australis* bacterial community. Although T4SSs are also often associated with virulent host/bacterium interactions (66), this system can also mediate symbiotic interactions (67). For example, *Mesorhizobium loti* R7A encodes a "symbiosis island," containing a T4SS homologous to the Vir pilus from *Agrobacterium tumefaciens*. It is believed that this T4SS transfers effector proteins into host plant cells (68). Although speculative, it is plausible that T4SS could be used by the bacterial community of *U. australis* to mediate symbiotic interactions, such as the transfer of compounds inducing correct morphology of the alga (69).

Defense. Genes related to defense include multidrug transport, restriction modification systems, nonribosomal peptide synthase modules, and ABC transporters, which have reported links with virulence (70, 71), and are also abundant in the algal surface community. Algal-associated bacteria may protect the host by inhibiting the attachment of other bacteria and biofouling organisms through the production of secondary metabolites (71). Toxic and antibiotic compounds may be transported out of the cell via homologs of ABC transporters known to export multiple drugs (72). The presence of restriction modification systems indicates a need to minimize transduction or transformation, or to only allow for genetic exchange between bacteria that have similar restriction modification systems. All these functions can clearly contribute to the maintenance of the structural and genetic integrity and stability of the surface community, which is of particular importance in a system that is constantly exposed to a large number of potentially invading bacteria (e.g., planktonic secondary colonizers).

These functions are all consistent with the ecological role of the *U. australis* community and together provide support for the notion of a specific and stable core metagenome, which is functionally adapted to life on the alga's surface.

Colonization of *U. australis*: Does Taxonomy Reflect Function? In eukaryotic models of colonization via a competitive lottery, functional groupings (or "guilds") often reflect taxonomic groupings (15, 73–76). Alternatively, there may be taxonomic redundancy, in which any given function is distributed broadly across a variety of taxa as opposed to being associated with any particular taxonomic group. To address this question of taxonomic redundancy, amino acid sequences of proteins assigned to six functions, which contributed most to the difference between algal and seawater samples, were analyzed using phylogenetic tree comparisons in Unifrac (77) and a taxonomic last common ancestor algorithm (MEGAN) (78) (see *Materials and Methods*). A large degree of phylogenetic dissimilarity of the core functions was exemplified by pairwise comparisons of samples, displaying between 65% and 97% dissimilarity (Table S4), indicating that the protein phylogenies for each core function were distinct in each sample. Unifrac *P* values indicated that for five core functions, the phylogenetic distribution of proteins between samples was not significantly different to that expected by chance (*P* values were > 0.05). The only exception was the pairwise dissimilarities of as little as 52% at a *P* value of 0.03 for COG2801 (putative transposase). Transposases are part of mobile genetic elements and therefore are likely transferred between members of the different communities. This finding would make those communities more similar and the differences appear less random with respect to the function of COG2801.

Taxonomic assignment with MEGAN further showed that the core functions were generally widely distributed across major bacterial groups present on *U. australis*, namely the α- and γ*-proteobacteria*, *Bacteroidetes*, and *Planctomycetes*, suggesting that a broad range of bacteria possess the ability to carry out these functions (Figs. S2–S7). At the level of species or genus, proteins were distributed across a variety of taxa, which differed from sample to sample, and often a function was assigned to a species that was present in one sample only (although it should be noted that proteins are not necessarily from the specific species assigned, but have high similarity to the homologous protein from that species). This finding means that protein functions are derived from distinct lineages that can only be crudely assigned to very high-level taxa (e.g., phyla) or, when a low-level assignment (e.g., species/genus) is possible, there is little overlap between samples. Together the result of the phylogenetic and taxonomic analysis of proteins support the assertion that the core functions are not restricted to a particular taxonomic group. This result also means that different taxa provide the core functions to the community and that members from different taxa could form a functional guild.

Structure of Bacterial Communities: Assembly of Functional Genes or Assembly of Species? Although our samples are taken over a limited time scale, and thus do not fully accommodate potential successional or historic changes in these communities, the evidence presented here and in our previous study (13) is most consistent with a competitive lottery model for community assembly on the surface of *U. australis*. Originally proposed for coral reef fish (15, 73), and subsequently applied to plant (74, 75) and parasite communities (76), the lottery model combines functional (niche- or guild-based) and random components as drivers of community structure. Specifically, species with similar trophic or other ecological properties are able to occupy the same niche within an ecosystem, and the particular species that occupies a particular space is then determined by stochastic recruitment. This means that within a group of species with similar ecologies, the "lottery" for space is won by whoever gets there first (15, 79).

In our system, as long as a bacterium has the necessary functional characters (defined here as particular gene functions) to colonize or grow on *U. australis*, the specific assemblage of species present on any given algal surface is stochastic, determined by which members of the guild happen to be available to colonize from the water column when space becomes available. Although the lottery model was originally proposed for relatively phylogenetically narrow groups of organisms (e.g., families of coral reef fish), we argue here that it also explains community assembly when the species pool spans multiple bacterial phyla.

More broadly and independently of any specific theory of community assembly, our results imply that genes or gene clusters are as or more important than species for understanding community assembly in bacterial systems. Dawkins (80) has famously argued in the context of evolutionary theory that individuals, and by inference collections of individuals such as species, are essentially containers for collections of genes. This approach contrasts with ecological models for community assembly, in which species (or higher taxonomic levels) are the fundamental metric. Such models are largely drawn from studies of eukaryotes, and so this species-focused approach is not surprising, given the general assumption of substantial genetic coherence within eukaryotic species.

However, the utility of the species concept for bacteria has been challenged in a number of ways (81–83), ranging from the level of genetic similarity necessary to define a species, to the extent (or lack thereof) of the genomic coherence of "species," because of the occurrence of substantial genetic exchange among taxonomically distinct bacteria. Indeed, in our system this frequent genetic exchange was supported by the abundance of functions for horizontal gene transfer (e.g., type IV secretion and transposase). We also found that analysis of functional gene systems revealed a considerable biological pattern that was not evident by focusing only on patterns of species diversity (13). Similar observations have been made for the human microbiome project (84, 85). This finding has at least two implications for studies of microbial community ecology. First, tests of community assembly theory [e.g., neutral theory (8)] using species as the key parameter may be misleading, because, if function does not map onto taxonomy, then accumulation or assembly of species will always appear random. Second, almost all of the theory currently used to understand patterns of microbial diversity is derived from eukaryotic ecology. However, our results raise the general possibility that genes and their functions may be more useful in testing models of community ecology for bacterial communities. This may be one important way in which the ecology of bacteria differs from that of eukaryotes.

Conclusion

This study provides insight into the link between community structure and function for a complex, algal-associated bacterial community. This community contains a consistent functional profile, with features related to a host-associated lifestyle, and functional similarities exist within phylogenetically distinct members from different host individuals. Although the community members on *U. australis* contain functional similarities, we do not yet know whether they form guilds that are specific to *U. australis*, or whether they are more generally adapted to other living or inanimate surfaces. Nevertheless, our evidence is consistent with community assembly via a competitive lottery mechanism. This model encompasses both selective and neutral processes, and could apply to other complex host-associated microbial communities, such as the human microbiome, where a consistent core of functional genes is detected across hosts (85), but species-level community composition is highly variable (86). The lack of correspondence between function (as determined by functional gene systems) and phylogeny in this system suggests that genes, rather than species, may be the appropriate parameter for understanding patterns of diversity in many microbial communities.

Materials and Methods

Sampling. Seawater was collected from Botany Bay (SW3 and SW4: 33°59′S, 151°14′E) on the 3rd of January 2005 and on the 19th of January 2005 from Bare Island (SW5 and SW6: 33°59′S, 151°13′E) and 200 m away in Botany Bay (SW7 and SW8). Seawater was collected again from Bare Island (SW9 and SW10) at the 18th of October 2006 to coincide with the sampling of *U. australis* at this site. Two-hundred liters of seawater were collected for each sample (exception is 100 L each for SW9 and SW10) from a depth of 2 m, and immediately serially filtered through 20-, 3-, 0.8-, and 0.1-μm filters. *U. australis* thalli were collected (wet weight: 20 g per sample) from two different rock pools at Bare Island in October 2006 (UA1 and UA2: 33°59′S, 151° 13′E) and again on the 7th of February 2007 (UA3 and UA4). Thalli were also collected from two different rock pools ~9 km away at Shark Point, Clovelly, on the 7th of February 2007 (UA5 and UA6: 33°91′S, 151°26′E). Sampling was performed between 10:00 AM and 12:00 PM after outgoing tides to ensure rock pools were well flushed with the surrounding seawater, and only algae of the same approximate size were collected to ensure that they were in the same developmental stage of their life cycle. Water parameters were also measured (Table S2) and showed similar values across the years. The sampling design and environmental measurements suggest that each sample of *U. australis* was subjected to similar environmental conditions.

DNA Sequencing, Assembly, and Annotation. DNA was extracted from the six algal (UA1–UA6) and eight seawater samples (SW3–SW10), which correspond to samples from ref. 13. Bacterial DNA was extracted from the surface of the algal fronds as described previously (87), which leaves the algal host intact and extracts total DNA from the entire surface community. For the filtered seawater samples, DNA was extracted from the 0.1-μm filter, as previously described (88).

Large-scale shotgun sequencing was performed and sequences were quality-filtered, assembled, and annotated (details in *SI Materials and Methods*). Each ORF was searched against the COG database (16) and a matrix of raw counts per COG and per sample was generated. Data were also submitted to the MG-RAST server (89) in the form of unassembled individual reads. Matches to the SEED database (17) with an e-value of less than 10^{-20} and a minimum alignment length of 100 amino acids were used in a matrix of counts per SEED subsystem per sample.

Statistical Analysis of Metagenomic Datasets and Core Functions. Matrices of raw counts of COG and SEED functional annotations per *U. australis* or seawater sample were standardized to account for the unequal sequence coverage between samples. Bray-Curtis similarity matrices were calculated and used to generate MDS plots. PERMANOVA (90) were carried out to compare samples from each environment. Similarity percentage analyses (SIMPER) (91) were carried out to determine the contribution of each COG or SEED subsystem to similarities within, and difference between, environments. The PRIMER-6 package was used for all multivariate statistical analysis (92).

The top 50 COG and SEED subsystems which contributed most to the differences between the two environments (as assessed by SIMPER), were selected for further analysis. After removal of those, which were not consistently more abundant in the algal dataset, the remaining 36 COG and 38 SEED subsystems were defined as comprising the core functions of the *U. australis* bacterial community. These functions were grouped into broader categories with respect to a surface-associated lifestyle with a living algal host (see *SI Materials and Methods* for details).

Assessment of Phylogenetic Similarity and Taxonomic Orgin of Core Functions. Protein sequences from each *U. australis* sample for the six COGs, which contributed most to the difference from seawater libraries, were aligned in ClustalW 1.83 (93) and maximum likelihood trees were built in RAxML 7.0.4 (94) using 100 bootstraps. Tree files were analyzed with the Unifrac Webserver (http://bmf2.colorado.edu/unifrac/) (77) to calculate the proportion of branch length, which is unique or shared in each environment. The unweighted algorithm was used as each protein sequence was unique. Sequences were also compared against the National Center for Biotechnology Information nonredundant protein database using BLASTP (95) and analyzed with MEGAN 3.7 (78), which uses a last common ancestor algorithm to assign a likely taxonomic origin to each sequence. The predicted origins of proteins were compared to determine if particular functions were associated with particular taxa. Only sequences from samples UA1, UA2, UA3, and UA5 were used in MEGAN analysis; samples UA4 and UA6 contained fewer than 10 sequences for each function because of the lower depth of sequence coverage obtained.

ACKNOWLEDGMENTS. We thank Matt DeMaere for help with the annotation and Alastair Poore for help with the statistical analysis. This work was funded by the Australian Research Council; the Gordon and Betty Moore Foundation; the US Department of Energy, Office of Science, Office of Biological and Environmental Research (DE-FG02-02ER63453); The Center for Marine Bio-Innovation; and the J. Craig Venter Institute. This is contribution #0053 from the Sydney Institute of Marine Science.

MICROBIOLOGY

1. Debroas D, et al. (2009) Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget—France). *Environ Microbiol* 11:2412–2424.
2. Hewson I, et al. (2009) Metagenomic potential of microbial assemblages in the surface waters of the central Pacific Ocean tracks variability in oceanic habitat. *Limnol Oceanogr* 54:1981–1994.
3. Ram RJ, et al. (2005) Community proteomics of a natural microbial biofilm. *Science* 308:1915–1920.
4. van Elsas JD, et al. (2008) The metagenomics of disease-suppressive soils— Experiences from the METACONTROL project. *Trends Biotechnol* 26:591–601.
5. Yooseph S, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PLoS Biol* 5(3):e16.
6. Ofiteru ID, et al. (2010) Combined niche and neutral effects in a microbial wastewater treatment community. *Proc Natl Acad Sci USA* 107:15345–15350.
7. Sloan WT, et al. (2006) Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ Microbiol* 8:732–740.
8. Woodcock S, et al. (2007) Neutral assembly of bacterial communities. *FEMS Microbiol Ecol* 62(2):171–180.
9. Hubbell SP (2001) *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univeristy Press, Princeton).
10. Hubbell SP (2006) Neutral theory and the evolution of ecological equivalence. *Ecology* 87:1387–1398.
11. Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH (2010) Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME J* 4:337–345.
12. Yang H, Schmitt-Wagner D, Stingl U, Brune A (2005) Niche heterogeneity determines bacterial community structure in the termite gut (*Reticulitermes santonensis*). *Environ Microbiol* 7:916–932.
13. Burke C, Thomas T, Lewis M, Steinberg P, Kjelleberg S (2011) Composition, uniqueness and variability of the epiphytic bacterial community of the green alga *Ulva australis*. *ISME J* 5:590–600.
14. Sale PF (1976) Reef fish lottery. *Nat Hist* 85:60–65.
15. Munday PL (2004) Competitive coexistence of coral-dwelling fishes: The lottery hypothesis revisited. *Ecology* 85:623–628.
16. Tatusov RL, et al. (2003) The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4:41.
17. Overbeek R, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33:5691–5702.
18. Munoz Aguilar JM, et al. (1988) Chemotaxis of *Rhizobium leguminosarum* biovar *phaseoli* towards flavonoid inducers of the symbiotic nodulation genes. *J Gen Microbiol* 134:2741–2746.
19. Hossain MM, Tsuyumu S (2006) Flagella-mediated motility is required for biofilm formation by *Erwinia carotovora* subsp. *carotovora*. *J Gen Plant Pathol* 72:34–39.
20. Houry A, Briandet R, Aymerich S, Gohar M (2010) Involvement of motility and flagella in *Bacillus cereus* biofilm formation. *Microbiology* 156:1009–1018.
21. Lemon KP, Higgins DE, Kolter R (2007) Flagellar motility is critical for *Listeria monocytogenes* biofilm formation. *J Bacteriol* 189:4418–4424.
22. Pratt LA, Kolter R (1998) Genetic analysis of *Escherichia coli* biofilm formation: Roles of flagella, motility, chemotaxis and type I pili. *Mol Microbiol* 30:285–293.
23. Namba A, et al. (2008) OmpA is an adhesion factor of *Aeromonas veronii*, an optimistic pathogen that habituates in carp intestinal tract. *J Appl Microbiol* 105:1441–1451.
24. Serino L, et al. (2007) Identification of a new OmpA-like protein in *Neisseria gonorrhoeae* involved in the binding to human epithelial cells and in vivo colonization. *Mol Microbiol* 64:1391–1403.
25. Chen BY, Kim TJ, Silva JL, Jung YS (2009) Positive correlation between the expression of inlA and inlB genes of *Listeria monocytogenes* and its attachment strength on glass surface. *Food Biophys* 4:304–311.
26. Franciosa G, Maugliani A, Scalfaro C, Floridi F, Aureli P (2009) Expression of internalin A and biofilm formation among *Listeria monocytogenes* clinical isolates. *Int J Immunopathol Pharmacol* 22(1):183–193.
27. Tomich M, Planet PJ, Figurski DH (2007) The tad locus: Postcards from the widespread colonization island. *Nat Rev Microbiol* 5:363–375.

28. Borthakur D, et al. (1986) A mutation that blocks exopolysaccharide synthesis prevents nodulation of peas by *Rhizobium leguminosarum* but not of beans by *R. phaseoli* and is corrected by cloned DNA from *Rhizobium* or the phytopathogen *Xanthomonas*. *Mol Gen Genet* 203:320–323.
29. González JE, Reuhs BL, Walker GC (1996) Low molecular weight EPS II of *Rhizobium meliloti* allows nodule invasion in *Medicago sativa*. *Proc Natl Acad Sci USA* 93:8636–8641.
30. Hotter GS, Scott DB (1991) Exopolysaccharide mutants of *Rhizobium loti* are fully effective on a determinate nodulating host but are ineffective on an indeterminate nodulating host. *J Bacteriol* 173:851–859.
31. Pellock BJ, Cheng HP, Walker GC (2000) Alfalfa root nodule invasion efficiency is dependent on *Sinorhizobium meliloti* polysaccharides. *J Bacteriol* 182:4310–4318.
32. Rolfe BG, et al. (1996) Defective infection and nodulation of clovers by exopolysaccharide mutants of *Rhizobium leguminosarum* bv trifolii. *Aust J Plant Physiol* 23:285–303.
33. Ryjenkov DA, Tarutina M, Moskvin OV, Gomelsky M (2005) Cyclic diguanylate is a ubiquitous signaling molecule in bacteria: Insights into biochemistry of the GGDEF protein domain. *J Bacteriol* 187:1792–1798.
34. Simm R, Morr M, Kader A, Nimtz M, Römling U (2004) GGDEF and EAL domains inversely regulate cyclic di-GMP levels and transition from sessility to motility. *Mol Microbiol* 53:1123–1134.
35. Jenal U, Malone J (2006) Mechanisms of cyclic-di-GMP signaling in bacteria. *Annu Rev Genet* 40:385–407.
36. Kulasekara HD, et al. (2005) A novel two-component system controls the expression of *Pseudomonas aeruginosa* fimbrial cup genes. *Mol Microbiol* 55:368–380.
37. Lee VT, et al. (2007) A cyclic-di-GMP receptor required for bacterial exopolysaccharide production. *Mol Microbiol* 65:1474–1484.
38. Römling U (2002) Molecular biology of cellulose production in bacteria. *Res Microbiol* 153:205–212.
39. Tischler AD, Camilli A (2004) Cyclic diguanylate (c-di-GMP) regulates *Vibrio cholerae* biofilm formation. *Mol Microbiol* 53:857–869.
40. Pitcher RS, Brittain T, Watmough NJ (2002) Cytochrome cbb(3) oxidase and bacterial microaerobic metabolism. *Biochem Soc Trans* 30:653–658.
41. Strohm TO, Griffin B, Zumft WG, Schink B (2007) Growth yields in bacterial denitrification and nitrate ammonification. *Appl Environ Microbiol* 73:1420–1424.
42. de Beer D, Stoodley P, Roe F, Lewandowski Z (1994) Effects of biofilm structures on oxygen distribution and mass transport. *Biotechnol Bioeng* 43:1131–1138.
43. Lahaye M, Axelos MAV (1993) Gelling properties of water soluble polysaccharides from proliferating marine green seaweeds (*Ulva* spp). *Carbohydr Polym* 22:261–265.
44. Hurst AC, et al. (2008) MscS, the bacterial mechanosensitive channel of small conductance. *Int J Biochem Cell Biol* 40:581–585.
45. Miticka H, et al. (2003) Transcriptional analysis of the rpoE gene encoding extracytoplasmic stress response sigma factor sigmaE in *Salmonella enterica* serovar Typhimurium. *FEMS Microbiol Lett* 226:307–314.
46. Vanaporn M, Vattanaviboon P, Thongboonkerd V, Korbsrisate S (2008) The rpoE operon regulates heat stress response in *Burkholderia pseudomallei*. *FEMS Microbiol Lett* 284:191–196.
47. Pohnert G (2004) Chemical defense strategies of marine organisms. *Chemistry of Pheromones and Other Semiochemicals I, Topics in Current Chemistry*, ed Schulz S (Springer-Verlag Berlin, Berlin), Vol 239, pp 179–219.
48. Gaudry A, et al. (2007) Heavy metals pollution of the Atlantic marine environment by the Moroccan phosphate industry, as observed through their bioaccumulation in *Ulva lactuca*. *Water Air Soil Pollut* 178:267–285.
49. Yoshida T, Phadtare S, Inouye M (2007) Functional and structural characterization of EnvZ, an osmosensing histidine kinase of *E. coli*. *Two-Component Signaling Systems, Pt B, Methods in Enzymology*, eds Simon MI, Crane BR, Crane A, (Elsevier Academic Press Inc, San Diego), Vol 423, pp 184–202.
50. Nagakubo S, Nishino K, Hirata T, Yamaguchi A (2002) The putative response regulator BaeR stimulates multidrug resistance of *Escherichia coli* via a novel multidrug exporter system, MdtABC. *J Bacteriol* 184:4161–4167.
51. Perego M, Cole SP, Burbulys D, Trach K, Hoch JA (1989) Characterization of the gene for a protein kinase which phosphorylates the sporulation-regulatory proteins Spo0A and Spo0F of *Bacillus subtilis*. *J Bacteriol* 171:6187–6196.
52. Chiang RC, Cavicchioli R, Gunsalus RP (1992) Identification and characterization of narQ, a second nitrate sensor for nitrate-dependent gene regulation in *Escherichia coli*. *Mol Microbiol* 6:1913–1923.
53. Iniesta AA, McGrath PT, Reisenauer A, McAdams HH, Shapiro L (2006) A phospho-signaling pathway controls the localization and activity of a protease complex critical for bacterial cell cycle progression. *Proc Natl Acad Sci USA* 103:10935–10940.
54. Jin SG, Roitsch T, Ankenbauer RG, Gordon MP, Nester EW (1990) The VirA protein of *Agrobacterium tumefaciens* is autophosphorylated and is essential for vir gene regulation. *J Bacteriol* 172:525–530.
55. Cai T, et al. (2009) Host legume-exuded antimetabolites optimize the symbiotic rhizosphere. *Mol Microbiol* 73:507–517.
56. Maddocks SE, Oyston PCF (2008) Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology* 154:3609–3623.
57. Mark GL, et al. (2005) Transcriptome profiling of bacterial responses to root exudates identifies genes involved in microbe-plant interactions. *Proc Natl Acad Sci USA* 102:17454–17459.
58. Pothier JF, Wisniewski-Dyé F, Weiss-Gayet M, Moënne-Loccoz Y, Prigent-Combaret C (2007) Promoter-trap identification of wheat seed extract-induced genes in the plant-growth-promoting rhizobacterium *Azospirillum brasilense* Sp245. *Microbiology* 153:3608–3622.
59. Spaepen S, Das F, Luyten E, Michiels J, Vanderleyden J (2009) Indole-3-acetic acid-regulated genes in *Rhizobium etli* CNPAF512. *FEMS Microbiol Lett* 291:195–200.
60. Spaepen S, Vanderleyden J, Remans R (2007) Indole-3-acetic acid in microbial and microorganism-plant signaling. *FEMS Microbiol Rev* 31:425–448.
61. Yuan ZC, Haudecoeur E, Faure D, Kerr KF, Nester EW (2008) Comparative transcriptome analysis of *Agrobacterium tumefaciens* in response to plant signal salicylic acid, indole-3-acetic acid and gamma-amino butyric acid reveals signalling cross-talk and *Agrobacterium*–plant co-evolution. *Cell Microbiol* 10:2339–2354.
62. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.
63. Wuertz S, et al. (2001) In situ quantification of gene transfer in biofilms. *Microbial Growth in Biofilms, Biological Aspects, Methods in Enzymology*, ed Doyle RJ (Academic Press Inc., San Diego), Vol 336, pp 129–143.
64. Brazelton WJ, Baross JA (2009) Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. *ISME J* 3:1420–1424.
65. Hooper SD, Mavromatis K, Kyrpides NC (2009) Microbial co-habitation and lateral gene transfer: What transposases can tell us. *Genome Biol* 10(4):R45.
66. Juhas M, Crook DW, Hood DW (2008) Type IV secretion systems: Tools of bacterial horizontal gene transfer and virulence. *Cell Microbiol* 10:2377–2386.
67. Sullivan JT, et al. (2002) Comparative sequence analysis of the symbiosis island of *Mesorhizobium loti* strain R7A. *J Bacteriol* 184:3086–3095.
68. Hubber A, Vergunst AC, Sullivan JT, Hooykaas PJ, Ronson CW (2004) Symbiotic phenotypes and translocated effector proteins of the *Mesorhizobium loti* strain R7A VirB/D4 type IV secretion system. *Mol Microbiol* 54:561–574.
69. Nakanishi K, Nishijima M, Nomoto AM, Yamazaki A, Saga N (1999) Requisite morphologic interaction for attachment between *Ulva pertusa* (Chlorophyta) and symbiotic bacteria. *Mar Biotechnol (NY)* 1(1):107–111.
70. Liu ZY, Jacobs M, Schaff DA, McCullen CA, Binns AN (2001) ChvD, a chromosomally encoded ATP-binding cassette transporter-homologous protein involved in regulation of virulence gene expression in *Agrobacterium tumefaciens*. *J Bacteriol* 183:3310–3317.
71. Wenzel SC, Müller R (2005) Formation of novel secondary metabolites by bacterial multimodular assembly lines: Deviations from textbook biosynthetic logic. *Curr Opin Chem Biol* 9:447–458.
72. van Veen HW, Konings WN (1998) The ABC family of multidrug transporters in microorganisms. *Biochim Biophys Acta* 1365(1-2):31–36.
73. Sale PF (1979) Recruitment, loss and coexistence in a guild of territorial coral-reef fishes. *Oecologia* 42:159–177.
74. Laurie H, Mustart PJ, Cowling RM (1997) A shared niche? The case of the species pair *Protea obtusifolia Leucadendron meridianum*. *Oikos* 79:127–136.
75. Lavorel S (1999) Ecological diversity and resilience of Mediterranean vegetation to disturbance. *Divers Distrib* 5:3–13.
76. Janovy J, Jr., Clopton RE, Percival TJ (1992) The roles of ecological and evolutionary influences in providing structure to parasite species assemblages. *J Parasitol* 78:630–640.
77. Lozupone C, Hamady M, Knight R (2006) UniFrac—An online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* 7:371.
78. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386.
79. Kelley SE (1989) Experimental studies of the evolutionary significance of sexual reproduction. 5. A field-test of the sib-competition lottery hypothesis. *Evolution* 43:1054–1065.
80. Dawkins R (1976) *The Selfish Gene* (Oxford University Press, Oxford).
81. Doolittle WF, Zhaxybayeva O (2009) On the origin of prokaryotic species. *Genome Res* 19:744–756.
82. Gevers D, et al. (2005) Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol* 3:733–739.
83. Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* 361:1929–1940.
84. Qin JJ, et al.; MetaHIT Consortium (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464(7285):59–65.
85. Turnbaugh PJ, Gordon JI (2009) The core gut microbiome, energy balance and obesity. *J Physiol* 587:4153–4158.
86. Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res* 19:1141–1152.
87. Burke C, Kjelleberg S, Thomas T (2009) Selective extraction of bacterial DNA from the surfaces of macroalgae. *Appl Environ Microbiol* 75:252–256.
88. Shaw AK, et al. (2008) It's all relative: Ranking the diversity of aquatic bacterial communities. *Environ Microbiol* 10:2200–2210.
89. 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.
90. Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecol* 26(1):32–46.
91. Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. *Austral Ecol* 18(1):117–143.
92. Clarke KR, Gorley RN (2006) *PRIMER v6: User Manual/Tutorial* (PRIMER-E, Plymouth).
93. Chenna R, et al. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 31:3497–3500.
94. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
95. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.

Supporting Information

Burke et al. 10.1073/pnas.1101591108

SI Materials and Methods

Samples, DNA Sequencing, and Assembly. Samples of *Ulva australis* and seawater were collected as described in Burke et al. (1), and DNA extracted as described in Burke et al. (2) for alga and in Rusch et al. (3) for seawater. Clone libraries were produced and sequenced from environmental DNA samples as described in Rusch et al. (3). Once samples had passed several quality control steps, including gel electrophoresis of DNA samples to ensure adequate size distribution, as well as test-sequencing of a small proportion of the clone library followed by analysis for potential clone contaminations (e.g., *Escherichia coli* DNA), large-scale sequencing was performed on ABI3730XL sequencers on eight seawater samples (SW3, SW4, SW5, SW6, SW7, SW8, SW9, SW10), which correspond to the samples used for preparation of 16S rRNA gene libraries (1), and six *U. australis* samples (UA1, UA2, UA3, UA4, UA5, UA6), again corresponding to samples from Burke et al. (1). Shotgun sequences were assembled as described in Rusch et al. (3), with the following modifications. The read fragments were assembled with Celera Assembler software version 5.1 from the public repository (4). A 12% error was permitted in the unitigger (utgErrorRate) with 14% error allowed in the overlapper (ovlErrorRate), consensus (cnsErrorRate), and scaffolder (cgwErrorRate) module. Seed length (merSizeOvl) was set to 14. Overlap trimming, extended clear ranges, and surrogates were turned on. Fragment correction and bubble popping were turned off.

Annotation of Metagenomic Sequences. To remove any possible eukaryotic contaminations, such as DNA derived from mitochondria, the assembled data were filtered by searching all DNA fragments against the National Center for Biotechnology Information nucleotide database, using BLASTN. Based on these search results, sequences were classified into taxonomic groups using the last common ancestor algorithm in MEGAN (5), with parameters "min score" set at 30 and "top score" set at 10%. Manual evaluation confirmed that this procedure effectively removes scaffolds and singletons derived from mitochondrial DNA. For all samples, the amount of sequencing classified as "Eukaryotes" was less than 1% of the total assembled nucleotides. All ORFs associated with those putative eukaryotic DNA fragments were disregarded from the functional comparison.

Scaffolds and singletons were processed with Metagene (6) to identify ORFs. The algorithm implemented in Metagene is specifically designed for prokaryotic (i.e., bacterial and archaeal) gene detection from metagenomic datasets. For functional annotation, each ORF was searched against the Clusters of Orthologous Groups (COG) (7) database using HMMER version 2.3.2 (8) applying a confidence cutoff of 10^{-20}. The functional annotation of each ORF was multiplied by the average coverage of the DNA fragment to estimate the abundance of each functional assignment; a similar strategy was recently used by Tringe et al. (9). COG hits with an e-value of less than 10^{-20} were tallied and placed in a matrix of raw counts per COG per sample.

Each shotgun sequencing dataset was also submitted to the MG-RAST server (10) in the form of unassembled individual reads. Matches to the SEED database with an e-value of less than 10^{-20} and a minimum alignment length of 100 amino acids were extracted for each sample and placed in a matrix of counts per SEED subsystem per sample.

Core Functions Within the *U. australis* Metagenome. To determine whether there was a set of functions that defined the *U. australis* community in comparison with the seawater, COG and SEED annotations were extracted, which were more abundant in *U. australis*, and expressed as a proportion of the total annotated ORFs for each sample. These annotations were ranked by their percent-contribution to the difference between *U. australis* and seawater environments, as determined by the similarity percentage analyses (SIMPER). The contribution to difference was plotted against the ranked functions, and the top 50 COGs and SEED subsystems were chosen for further analysis. These 50 COGs accounted for 10.7% of the overall difference caused by abundant COGs of the algal dataset, and the remaining 1,316 abundant algal COGs accounted for 37.9% of the total difference. The top 50 SEED subsystems contributed 25.5% to the difference between *U. australis* and seawater samples, and the remaining 464 SEED subsystems accounted for 25.8% of the total difference. This finding means that on average, the top 50 COG and SEED annotations each contributed ~7.4 and 9.3 times more to difference, respectively, than the remaining COG and SEED annotations. COG and SEED annotations that were not abundant in all *U. australis* samples were removed, leaving 36 COGs and 38 SEED subsystems, defined here as "core functions" of the *U. australis* bacterial community.

1. Burke C, Thomas T, Lewis M, Steinberg P, Kjelleberg S (2011) Composition, uniqueness and variability of the epiphytic bacterial community of the green alga *Ulva australis*. *ISME J* 5:590–600.
2. Burke C, Kjelleberg S, Thomas T (2009) Selective extraction of bacterial DNA from the surfaces of macroalgae. *Appl Environ Microbiol* 75:252–256.
3. Rusch DB, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5(3):e77.
4. Denisov G, et al. (2008) Consensus generation and variant detection by Celera Assembler. *Bioinformatics* 24:1035–1040.
5. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386.
6. Noguchi H, Park J, Takagi T (2006) MetaGene: Prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* 34:5623–5630.
7. Tatusov RL, et al. (2003) The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4:41.
8. Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14:755–763.
9. Tringe SG, et al. (2008) The airborne metagenome in an indoor urban environment. *PLoS ONE* 3:e1862.
10. 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.

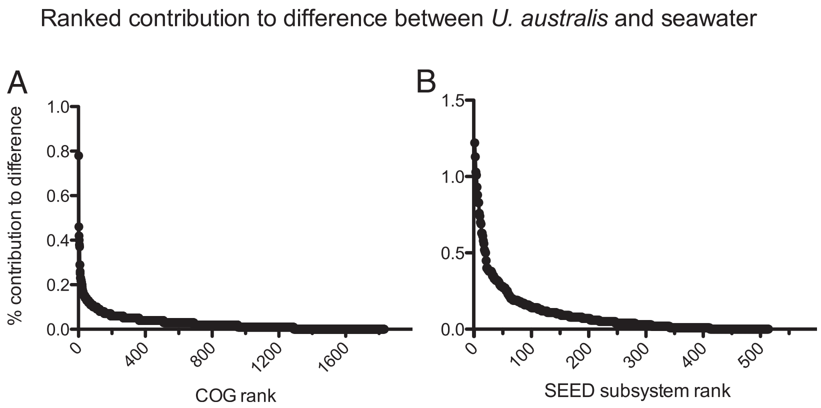

Fig. S1. The contribution to differences between environments for (*A*) COG and (*B*) SEED annotations, which were more abundant in *U. australis* metagenomes. COG and SEED categories are ranked in order of their percentage contribution to difference, as assessed by SIMPER.

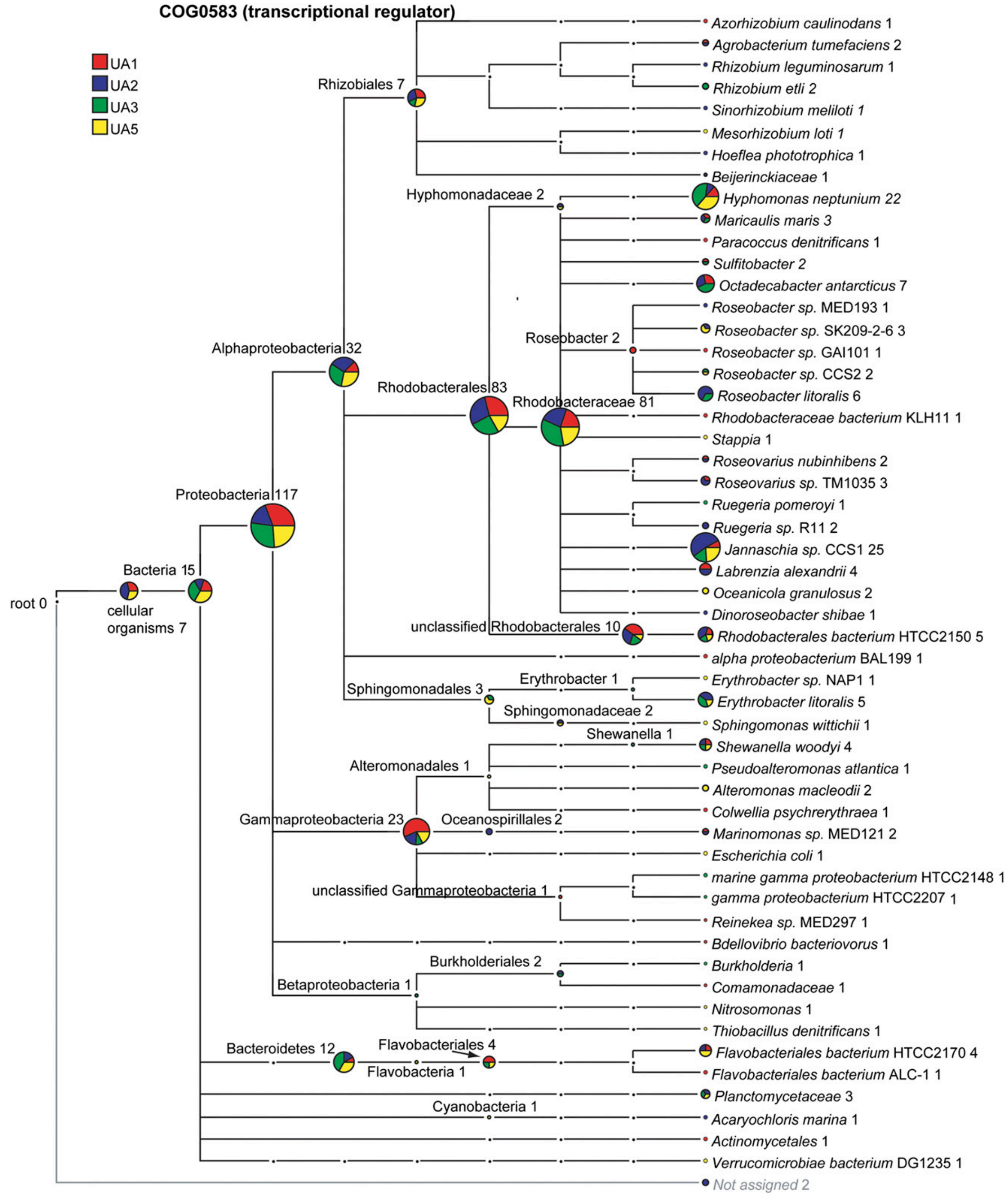

Fig. S2. MEGAN analysis of predicted proteins assigned to COG0583 (transcriptional regulator). The size of the circles is relative to the number of proteins assigned to each node (also indicated in numbers), and taxonomy is displayed with the lowest level predicted.

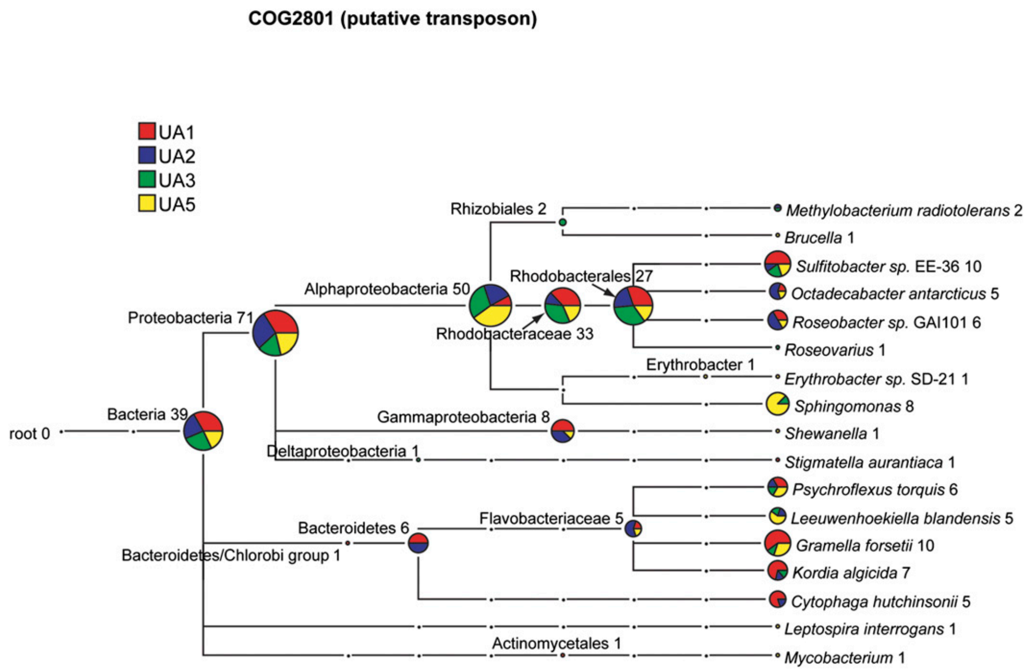

Fig. S3. MEGAN analysis of predicted proteins assigned to COG2801 (putative transposase). The size of the circles is relative to the number of proteins assigned to each node (also indicated in numbers), and taxonomy is displayed with the lowest level predicted.

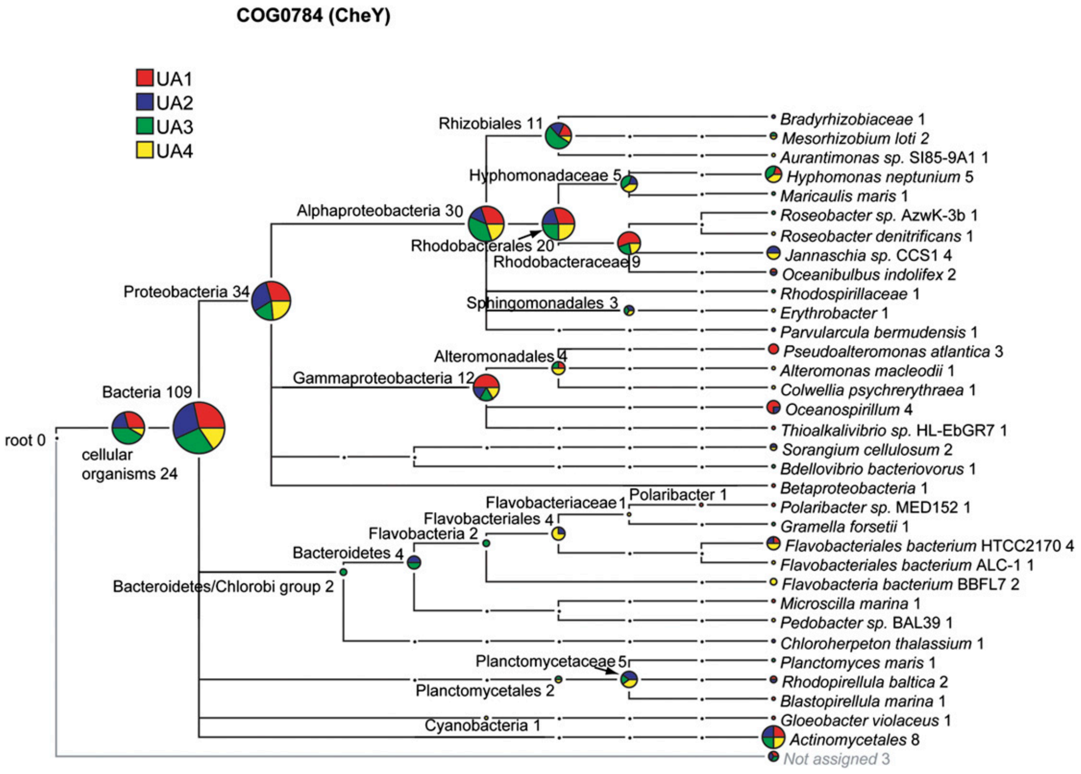

Fig. S4. MEGAN analysis of predicted proteins assigned to COG0784 (CheY). The size of the circles is relative to the number of proteins assigned to each node (also indicated in numbers), and taxonomy is displayed with the lowest level predicted.

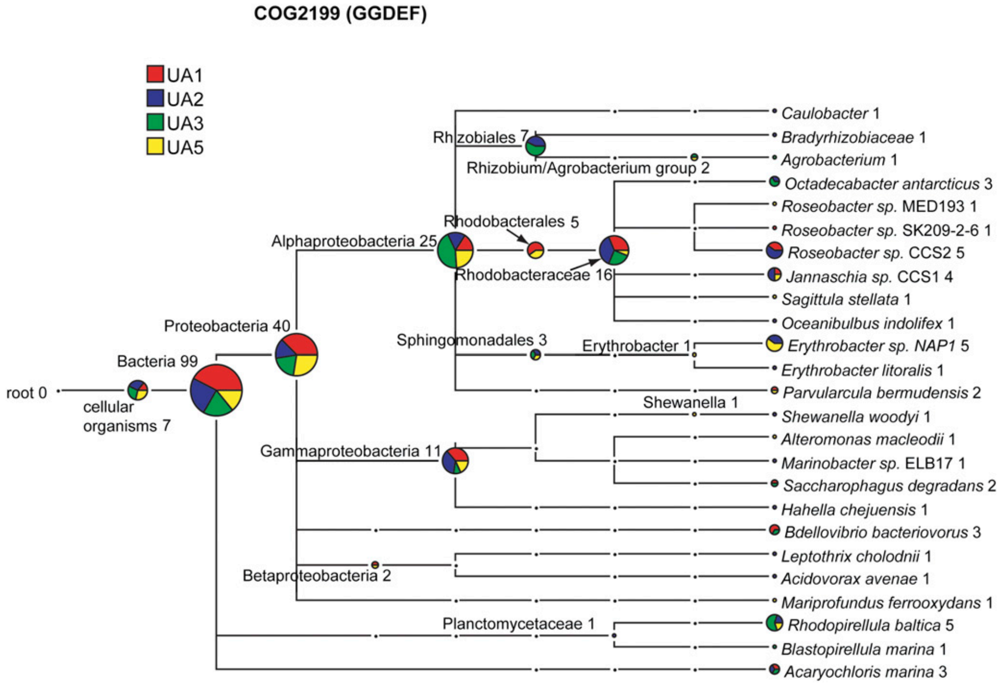

Fig. S5. MEGAN analysis of predicted proteins assigned to COG2199 (GGDEF domain). The size of the circles is relative to the number of proteins assigned to each node (also indicated in numbers), and taxonomy is displayed with the lowest level predicted.

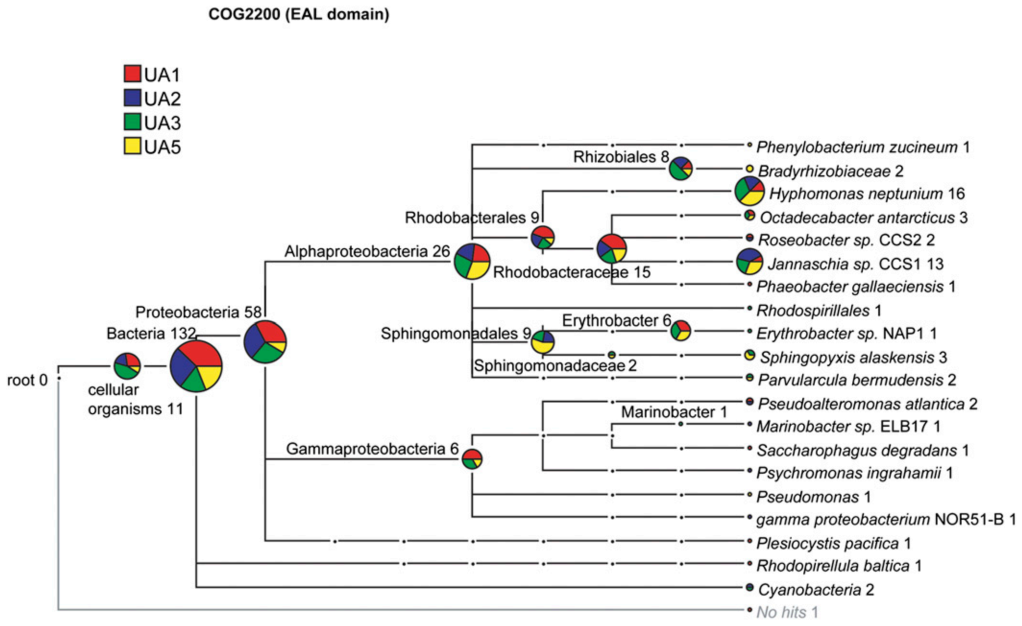

Fig. S6. MEGAN analysis of predicted proteins assigned to COG2200 (EAL domain). The size of the circles is relative to the number of proteins assigned to each node (also indicated in numbers), and taxonomy is displayed with the lowest level predicted.

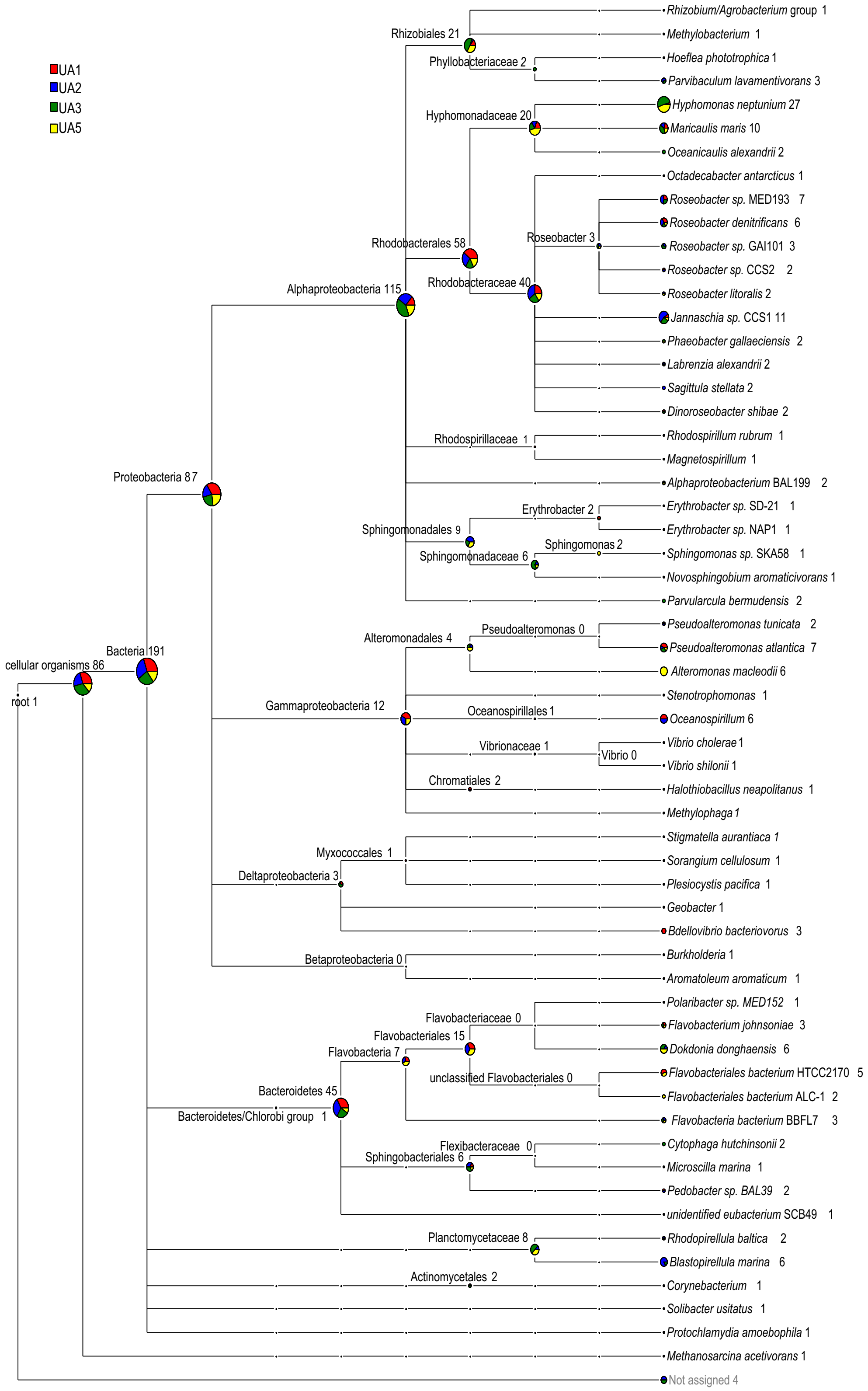

Fig. S7. MEGAN analysis of predicted proteins assigned to COG0642 (histidine kinase). The size of the circles is relative to the number of proteins assigned to each node (also indicated in numbers), and taxonomic affiliations are shown down to species level (if possible).

Table S1. Number of sequencing reads, base pairs, ORFs, or reads assigned to COGs and SEED subsystems and the percentage of total ORFs and reads

Sample	No. sequencing reads	Total bp	Number of ORFs assigned to COGs	% of total predicted ORFs	Number of reads assigned to SEED subsystems	% of total reads
UA1	94,505	75,822,569	44,439	31.33	39,953	42.28
UA2	91,719	71,921,435	43,732	30.93	38,366	41.83
UA3	90,404	67,092,023	48,762	31.40	37,879	41.90
UA4	9,861	7,547,408	6,113	31.86	4,237	42.97
UA5	93,434	67,121,596	41,213	32.14	40,171	42.99
UA6	9,425	7,259,709	5,852	30.46	3,767	39.97
SW3	89,783	71,742,472	57,580	37.14	48,613	54.14
SW4	9,863	7,943,725	7,934	37.26	5,362	54.36
SW5	162,760	128,450,836	100,002	37.55	86,173	52.95
SW6	10,066	8,257,922	8,585	39.65	5,721	56.83
SW7	95,637	71,150,001	62,485	39.53	53,207	55.63
SW8	10,210	8,702,705	8,886	39.74	5,995	58.72
SW9	88,784	77,019,320	49,913	36.89	48,929	54.77
SW10	13,512	11,105,167	5,886	33.92	5,665	41.93

Table S2. Seawater parameters measured at time of sampling of SW3-10 and UA1-6

Analysis type	SW3	SW4	SW5	SW6	SW7	SW8	SW9	SW10
Date	Jan. 3, 2005	Jan. 3, 2005	Jan. 19, 2005	Jan. 19, 2005	Jan. 19, 2005	Jan. 19, 2005	Oct. 18, 2006	Oct. 18, 2006
Temperature (°C)	21.4	21.4	21.4	21.4	21.8	21.8	18.3	17.1
pH	8.2	8.2	8.09	8.09	na	na	8.06	8.05
Salinity	36.0	36.0	35.0	35.0	36.0	36.0	36.6	36.7
Dissolved oxygen (mg/L)	5.8	5.8	4.65	4.65	5.87	5.87	7.00	5.95
Chlorophyll (mg/L)	1.82	1.82	0.61	0.61	2.68	2.68	1.44	0.82
Total dissolved nitrogen (μmol N/L)	10.208 ± 0.221	10.208 ± 0.221	9.102 ± 0.109	9.102 ± 0.109	16.273 ± 0.796	16.273 ± 0.796	11.805 ± 0.056	9.373 ± 0.61
Nitrate (μmol/L)	1.002 ± 0.002	1.002 ± 0.002	0.979 ± 0.059	0.979 ± 0.059	1.048 ± 0.002	1.048 ± 0.002	2.946 ± 0.003	2.082 ± 0.004
Urea (μmol N/L)	3.699 ± 0.010	3.699 ± 0.010	4.559 ± 0.100	4.559 ± 0.100	5.801 ± 0.959	5.801 ± 0.959	6.470 ± 0.010	5.490 ± 0.049
Phosphate (μmol/L)	na	na	na	na	na	na	0.608 ± 0.000	0.468 ± 0.013
Silicon (μmol/L)	1.633 ± 0.006	1.633 ± 0.006	0.281 ± 0.012	0.281 ± 0.012	0.293 ± 0.008	0.293 ± 0.008	1.338 ± 0.034	1.008 ± 0.002

Note that UA1 and UA2 were sampled at the same time as SW9 and SW10. na, Not applicable.

Table S3. Core functions (COGs and SEED subsystems), grouped into functional categories, which are informative as to their role in the *U. australis* epiphytic community

Functional category	COG or SEED subsystem	Description
Detection and movement toward the host surface	COG0840: Methyl-accepting chemotaxis protein COG0784: CheY-like receiver domain SEED subsystem: Bacterial chemotaxis	These proteins mediate chemotaxis, which has been demonstrated to play a role in the development of symbiosis, attracting bacteria toward the host.
	SEED subsystem: Flagellum	Flagella-mediated motility may be induced as a result of chemotaxis, which facilitates movement to the host surface. Motility also been demonstrated to be important for biofilm formation in several species.
Attachment and biofilm formation	COG2197: OmpA	OmpA has a demonstrated role in the adhesion to living surfaces (including plants) and biofilm formation
	SEED subsystem: *Listeria* internalin-like proteins	Required for virulence and internalization into host cells in *Listeria*, and truncated forms of internalin A (InlA) have been shown to enhance biofilm formation.
	SEED subsystem: widespread colonization island	Consists of the Tad (tight adherence) cluster and is important for biofilm formation, colonization and pathogenesis in several bacteria.
	SEED subsystem: Biosynthesis of galactoglycans and related lipopolysacharides	Contains genes related to the production of exopolysaccharide II, related to biofilm formation, and essential for symbiosis in several *Rhizobium* sp.
	COG2199: GGDEF domain proteins COG2200: EAL domain proteins SEED subsystem: Putative GGDEF domain protein related to agglutinin secretion	Function as cyclic-di-GMP synthase (GGDEF domain) and phospho diesterase (EAL domain), respectively. Have been shown to regulate the transition from a motile to a surface associated lifestyle.
	SEED subsystem: Glutathione-regulated potassium-efflux system and associated functions	Approximately one-third of proteins assigned to this subsystem contained GGDEF and EAL domains
	SEED subsystem: Biogenesis of cbb3-type cytochrome *c* oxidases Nitrate and nitrite ammonification	High-affinity oxygen binding protein, and nitrate ammonification associated with microaerobic metabolism in oxygen-limited and anoxic environments, such as cell clusters within biofilms.
Response to the algal host environment	SEED subsystems: Maltose and maltodextrin utilization Ribitol, Xylitol, Arabitol, Mannitol, and Sorbitol utilization Lactose and galactose uptake and utilization L-rhamnose utilization Xylose utilization Alginate metabolism COG3839: ABC-type sugar transport systems, ATPase components	Utilization of carbohydrates, which are found in *U. australis.*
	COG1595: RpoE σ subunit homolog COG0668: Small conductance mechanosensitive channel	RpoE is a specialized subunit of RNA polymerase, which regulates the cell response to heat and environmental stress; the small conductance mechanosensitive channel is a protein, which helps control osmoregularity in the cell. Heat and high salt concentrations are stresses, which would both be encountered on the surface of an intertidal pool alga.
	SEED subsystem: Oxidative stress	This subsystem contains proteins related to oxidative stress in bacteria. The release of reactive oxygen species is a known defense mechanisms of algae, and may be a stress encountered by the *U. australis* community.
	COG2217: Cation transport ATPases SEED subsystems: Copper homeostasis Cobalt-zinc-cadmium resistance	Proteins assigned to these subsystems are related to the export of heavy metals. *Ulva* sp. are known for their ability to uptake and store heavy-metal pollution from the environment, and bacteria on the algal surface may be exposed to them.
Regulation in response to environmental stimuli	COG0642: Signal transduction histidine kinase COG0583: Transcriptional regulator COG1609: Transcriptional regulators COG2204: Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains COG1522: Transcriptional regulators COG0515: Serine/threonine protein kinases COG0664: cAMP-binding domains SEED subsystem: cAMP signaling in bacteria	Environmental sensor proteins and transcriptional regulator proteins, homologs of which are known to be involved in osmoregulation, multidrug export, sporulation, nitrate reduction, cell differentiation and plant virulence. Some homologs are known to respond to plant exudates, and others regulate a range of features, including virulence, motility and quorum sensing. There may be a role for genetic regulation in response to the host environment, such as algal exudates, in the *U. australis* community.

Table S3. Cont.

Functional category	COG or SEED subsystem	Description
	SEED subsystem: Cyanobacterial circadian clock	Almost all proteins assigned to this subsystem were CikA, a phytochrome histidine kinase, which resets the cyanobacterial circadian clock in response to light.
Lateral gene transfer	COG2801: putative transposase SEED subsystem: Group II intron-associated genes	Transposases and introns are part of mobile genetic elements, which can be horizontally transferred. Lateral transfer genes have been found in high abundance in biofilm communities, and may result in a higher degree of lateral gene transfer in the *U. australis* community.
	COG0630: Type IV secretory pathway, VirB11 SEED subsystem: Type IV secretion and conjugative transfer	Type IV secretion is used for both DNA and protein transfer between bacteria (F plasmid in *E. coli*) and between bacteria and hosts (e.g., *Agrobacterium tumifaciens*). Type IV secretion has also been associated with symbiotic interactions between bacteria and plants, and potentially plays a role in symbiotic interactions between *U. australis* and its bacterial community.
Defense	COG1020: Nonribosomal peptide synthetase (NRPS) modules and related proteins COG1132: ABC-type multidrug/protein/lipid transport system, ATPase component SEED subsystem: Multidrug resistance efflux pumps	Proteins are involved in the production (NRPS) and export of drugs, toxins, and secondary metabolites, which may be produced as a competitive in the *U. australis* microbial community.
	COG0610: Restriction enzymes type I helicase subunits and related helicases SEED subsystem: Restriction modification system	Consists of mostly type I and some type III restriction modifications systems, which have been linked with protection of the cell from foreign DNA. May prevent phage transduction with the biofilm.
Miscellaneous		
COGS	COG1249: Dihydrolipoamide dehydrogenase/ glutathione oxidoreductase COG1960: Acyl-CoA dehydrogenases COG0714: MoxR-like ATPases COG0513: Superfamily II DNA and RNA helicases COG0612: Predicted Zn-dependent peptidases COG0156: 7-keto-8-aminopelargonate synthetase COG0654: FAD-dependent oxidoreductases COG1670: Acetyltransferases, COG1529: Aerobic-type carbon monoxide dehydrogenase, large subunit CoxL/CutL homologs COG1450: General secretory pathway protein D	Significance relating to biofilm mode of life on *U. australis* is unclear.
SEED subsystems	Chlorophyll biosynthesis Fatty acid metabolic cluster Putative associate of RNA polymerase sigma-54 factor RpoN Branched chain amino acid degradation regulons DNA repair, bacterial YgfZ-Fe-S cluster Isoleucine degradation Putative Isoquinoline 1-oxidoreductase subunit, Butanol biosynthesis General secretion pathway	

A description of the role is provided, and supporting references can be found in *Results and Discussion*.

Table S4. Environmental matrices showing the degree of difference between *U. australis* samples for proteins assigned to the six COGs, which contributed most to the difference between algae and seawater samples

	UA1	UA2	UA3	UA4	UA5	UA6
COG0642 (histidine kinase)						
UA1	0.00	0.84	0.94	0.99	0.93	0.97
UA2		0.00	0.90	0.97	0.91	0.97
UA3			0.00	0.96	0.88	0.96
UA4				0.00	0.96	0.97
UA5					0.00	0.96
UA6						0.00
COG0583 (transcriptional regulator)						
UA1	0.00	0.70	0.74	0.84	0.78	0.88
UA2		0.00	0.72	0.86	0.72	0.83
UA3			0.00	0.84	0.71	0.83
UA4				0.00	0.84	0.85
UA5					0.00	0.82
UA6						0.00
COG2801 (putative transposon)						
UA1	0.00	0.52	0.60	0.75	0.64	0.75
UA2		0.00	0.59	0.70	0.63	0.72
UA3			0.00	0.67	0.62	0.74
UA4				0.00	0.66	0.65
UA5					0.00	0.75
UA6						0.00
COG0784 (CheY)						
UA1	0.00	0.76	0.81	0.87	0.76	0.88
UA2		0.00	0.78	0.85	0.77	0.89
UA3			0.00	0.86	0.77	0.91
UA4				0.00	0.81	0.76
UA5					0.00	0.87
UA6						0.00
COG2199 (GGDEF domain)						
UA1	0.00	0.83	0.88	0.96	0.89	0.91
UA2		0.00	0.85	0.94	0.86	0.90
UA3			0.00	0.94	0.87	0.91
UA4				0.00	0.91	0.85
UA5					0.00	0.88
UA6						0.00
COG2200 (EAL domain)						
UA1	0.00	0.66	0.75	0.77	0.74	0.87
UA2		0.00	0.68	0.74	0.61	0.89
UA3			0.00	0.77	0.65	0.88
UA4				0.00	0.73	0.78
UA5					0.00	0.88
UA6						0.00

0 = no difference; 1= complete difference.