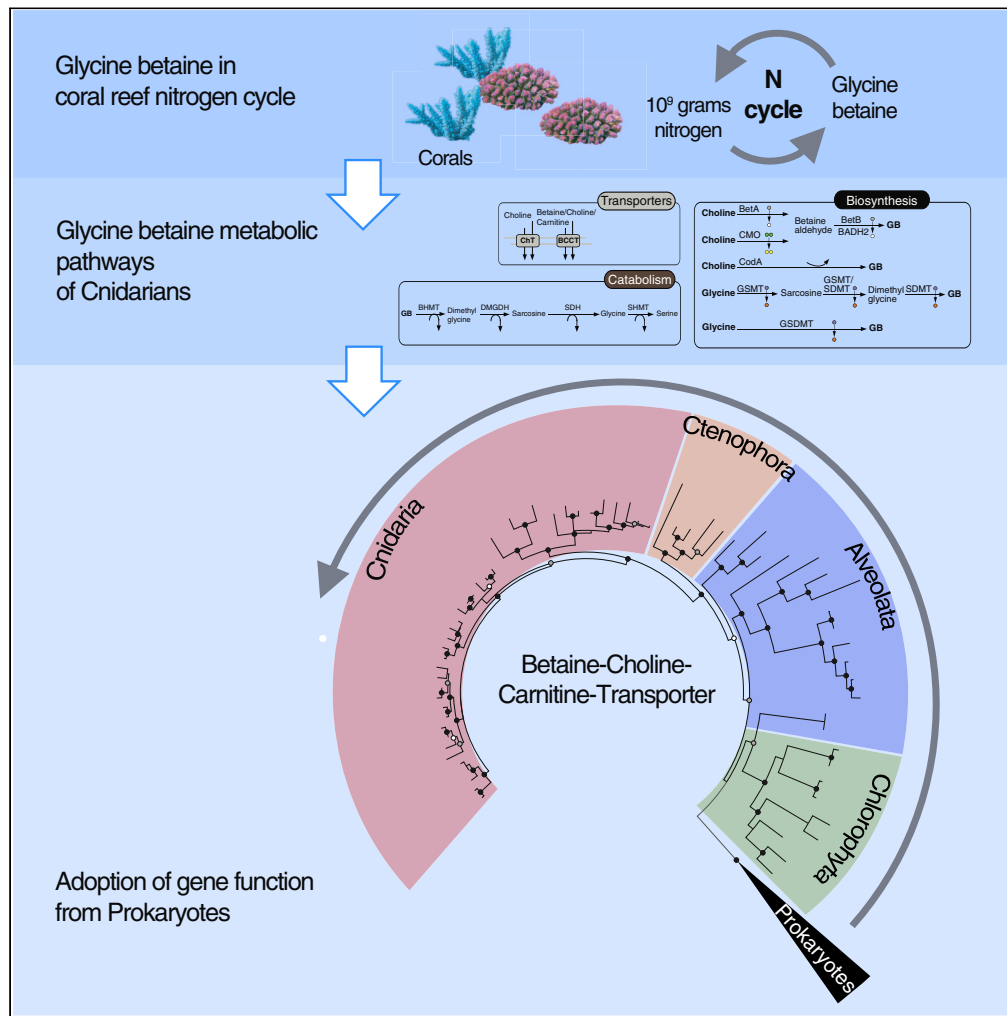# iScience

Article

Genomic Blueprint of Glycine Betaine Metabolism in Coral Metaorganisms and Their Contribution to Reef Nitrogen Budgets

David K. Ngugi,
Maren Ziegler,
Carlos M. Duarte,
Christian R.
Voolstra

david.ngugi@dsmz.de (D.K.N.)
christian.voolstra@
uni-konstanz.de (C.R.V.)

HIGHLIGHTS

Coral tissues contain high
concentrations of the
osmolyte glycine betaine

Corals and their microbial
symbionts can produce
and degrade glycine
betaine

High gene expression
patterns signifies role in
coral-microbial symbiosis

Glycine betaine is
estimated to encompass
16% of the coral's
nitrogen biomass

Article

Genomic Blueprint of Glycine Betaine Metabolism in Coral Metaorganisms and Their Contribution to Reef Nitrogen Budgets

David K. Ngugi,[1,2,5,*] Maren Ziegler,[2,3] Carlos M. Duarte,[2] and Christian R. Voolstra[2,4,*]

SUMMARY

The osmolyte glycine betaine (GB) ranks among the few widespread biomolecules in all three domains of life. In corals, tissue concentrations of GB are substantially higher than in the ambient seawater. However, the synthetic routes remain unresolved, questioning whether intracellular GB originates from _de novo_ synthesis or heterotrophic input. Here we show that the genomic blueprint of coral metaorganisms encode the biosynthetic and degradation machinery for GB. Member organisms also adopted the prokaryotic high-affinity carrier-mediated uptake of exogenous GB, rendering coral reefs potential sinks of marine dissolved GB. The machinery metabolizing GB is highly expressed in the coral model Aiptasia and its microalgal symbionts, signifying GB's role in the cnidarian-dinoflagellate symbiosis. We estimate that corals store between 10^6–10^9 grams of GB globally, representing about 16% of their nitrogen biomass. Our findings provide a framework for further mechanistic studies addressing GB's role in coral biology and reef ecosystem nitrogen cycling.

INTRODUCTION

The quaternary ammonium compound glycine betaine (GB) represents one of the most ubiquitous osmolytes on Earth (Roesser and Müller, 2001; Welsh, 2000; Yancey et al., 1982). GB is widespread in the cells and tissues from organisms of all three domains of life (King, 1988; Roesser and Müller, 2001; Welsh, 2000; Yancey et al., 1982), where it generally serves as a compatible solute (osmolyte) to counteract high salinity effects and helps protect membranes and proteins against abiotic stressors (Burg and Ferraris, 2008; Chen and Murata, 2002). Although reliable estimates of dissolved GB in the ocean water are scarce, evidence suggests that dissolved concentrations of betaine analogs (e.g., proline betaine and alanine betaine) are in the picomolar range [∼3–482 pM; (Muslin, 2017)], whereas GB bound in seawater (cellular) particulates are in the nanomolar to lower micromolar range [up to 0.48 μM; (Airs and Archer, 2010; Beale and Airs, 2016; Cree, 2015; Keller et al., 2004)]. Paradoxically, the extensive knowledge of GB in marine systems stems from studies of cultured phytoplankton species, mostly microalgae/diatoms such as _Emiliania huxleyi_, _Thalassiosira pseudonana_, and _Amphidinium carterae_ (Gebser and Pohnert, 2013; Keller et al., 1999a, 1999b; Spielmeyer et al., 2011; Spielmeyer and Pohnert, 2012). In the investigated phytoplankton cultures, maximum intracellular concentrations were higher under nitrogen-replete conditions [∼3–172 mM; (Keller et al., 1999b, 1999a)] than when nitrogen was limiting [4–15 mM; (Keller et al., 1999a)]. GB concentrations are also sensitive to elevated levels of temperature, CO_2, and salinity (Gebser and Pohnert, 2013; Spielmeyer and Pohnert, 2012) and can be strain-/species-specific and growth-stage dependent (Keller et al., 1999b, 1999a; Spielmeyer et al., 2011). Crucially, GB can comprise up to 20% of the cellular nitrogen pool in phytoplankton (Keller et al., 1999a). Once excreted or released from lysed cells, this organic compound represents a ubiquitous and dynamic constituent of oceanic dissolved organic matter readily taken up by marine heterotrophic bacteria (Kiene et al., 1998; Lidbury et al., 2015; Sun et al., 2011).

Quantitatively significant levels of GB also accumulate in the tissues of tridacnid clams, sponges, and scleractinian corals (33–215 mmol kg^{-1} wet tissue); concentrations of up to 90-fold higher than the alternative sulfur-containing osmolyte dimethylsulfoniopropionate [DMSP; (Hagedorn et al., 2010; Hill et al., 2017, 2010; Yancey et al., 2010)], in turn, suggesting a major role for GB in coral physiology. However, the biosynthesis of GB has not been studied in corals or marine invertebrates in general [see for example (Stephens-Camacho et al., 2015)], limiting our understanding of the synthetic routes for _de novo_ production and

[1]Leibniz Institute DSMZ – German Culture Collection for Microorganisms and Cell Cultures, Braunschweig, Germany

[2]Red Sea Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

[3]Department of Animal Ecology and Systematics, Justus-Liebig-University, Giessen, Germany

[4]Department of Biology, University of Konstanz, Konstanz 78457, Germany

[5]Lead Contact

*Correspondence:
david.ngugi@dsmz.de
(D.K.N.),
christian.voolstra@
uni-konstanz.de (C.R.V.)
https://doi.org/10.1016/j.isci.
2020.101120

sequestration of exogenous GB. Hypothetically, the higher intracellular concentrations of GB in reef corals (Hagedorn et al., 2010; Hill et al., 2017, 2010; Yancey et al., 2010), relative to the nanomolar concentrations reported in seawater (Airs and Archer, 2010; Beale and Airs, 2016), suggest an intrinsic capacity of corals to synthesize GB, the retention of extracellular GB obtained via heterotrophic input [e.g., feeding; (Houlbreque and Ferrier-Pages, 2009)], or both.

A two-enzyme oxidation pathway generally synthesizes GB from choline via the unstable intermediate betaine aldehyde in higher plants and bacteria (Figure 1A). The first oxidation step is catalyzed by an NAD^+-dependent choline dehydrogenase (BetA) in prokaryotes and by a flavin-dependent choline mono-oxygenase (CMO) in plants, whereas the second reaction is catalyzed by an NAD^+-dependent betaine aldehyde dehydrogenase (BetB) in all organisms (Chen and Murata, 2002). In some microorganisms, choline (and betaine aldehyde) is directly oxidized by choline oxidase (CodA) to GB with molecular oxygen as the electron acceptor (Lambou et al., 2013). GB can also be generated from glycine via a three-step methylation pathway involving glycine-sarcosine methyltranferase (GSMT) and sarcosine-dimethylglycine methyltransferase (SDMT) (Nyyssölä et al., 2000), with S-adenosylmethionine serving as the methyl group donor (Figure 1A). However, in a few organisms, all three successive methylation steps are catalyzed by a single enzyme called glycine sarcosine dimethylglycine N-methyltransferase (GSDMT), essentially a polypeptide with GSMT and SDMT domains (Kageyama et al., 2018; Lai et al., 2006). GB can be catabolized to serine via dimethylglycine, sarcosine, and glycine (Figure 1A) in reactions catalyzed by the enzymes betaine-homocysteine methyltransferase (BMHT), dimethylglycine dehydrogenase (DMGDH), sarcosine dehydrogenases (SoxA$_2$, SoxABCDG, or SDH), and serine hydroxymethyltransferase (SHMT), respectively. Here, we leveraged on available eukaryotic genomes and transcriptomes to elucidate the genetic basis for GB accumulation in marine invertebrates, with a focus on corals, where GB is postulated to play a role in thermotolerance and photoprotection (Hagedorn et al., 2010; Hill et al., 2017, 2010; Yancey et al., 2010).

RESULTS AND DISCUSSION
Eukaryotic Genomes Encode Multiple Synthetic Routes for GB Production

Bioinformatic analyses revealed that the genomes (n = 114) and transcriptomes (n = 16; Table S1) of an overwhelming majority of ecologically important marine invertebrates encode the biosynthetic pathway for GB via choline oxidation (CodA, CMO, BetA, and BADH2), glycine methylation (GSMT/SDMT and GSDMT), or both (Figures 1A and 1B). Genomes of eukaryotes inhabiting terrestrial and freshwater ecosystems exhibited a similar trend (Table S2). Crucially, 85% of the analyzed eukaryotic genomes possessed a putative *codA* gene (Figure 1B). Subject to further experimental validation, these findings suggest that the complementary pathway oxidizing choline directly to GB is widespread in marine eukaryotes.

Putative enzymes of the alternative, dual-enzyme pathways for oxidizing choline to GB using choline mono-oxygenase (CMO) or choline dehydrogenase (BetA)—for the first reaction—and betaine aldehyde dehydrogenases (BetB or BADH2)—for the second reaction were also present in the genomes/transcriptomes (Figure 1B; Table S2). Potential *CMO*, *betA*, and *BADH2* homologs with 40%–92% identity to validated proteins (Figure S1) were present in 13%, 23%, and 68% of the studied eukaryotes, respectively (Figure 1B). However, only two genomes (*Acanthaster planci* and *Sonneratia caseolaris*) possessed a putative *betB* homolog (Table S2). In-depth phylogenetic analyses showed that 88% of the predicted *BADH2* genes were homologous (>50% amino acid identity; Figure S1) to functionally characterized animal BADHs within aldehyde dehydrogenase (ALDH) family 9 (Julián-Sánchez et al., 2007), whereas the remainder were affiliated to validated plant BADH2s (Figure S2) within the ALDH family 10 proteins (Kirch et al., 2004). Co-occurrence analysis of predicted *CMO*, *betA*, and *BADH2* homologs in individual genomes/transcriptomes further showed that 11% and 43% of the analyzed eukaryotes possessed the complete CMO/BADH2- and BetA/BADH2-catalyzed pathways, respectively. However, the complete CMO/BADH2 pathway was predicted absent in Symbiodiniaceae and cnidarian genomes/transcriptomes (Figure 1B; Table S2). Notably, the putative "CMO"-like genes in both groups of organisms showed greater homology to carnitine monooxygenases (30%–40% identity) involved in carnitine metabolism than to ratified eukaryotic CMO proteins (≤25% identity). Crucially, these results indicate the possibility to oxidize choline to GB using the BetA/BADH2-catalyzed pathway in five out of eight cnidarians, which is absent in all Symbiodiniaceae genomes/transcriptomes analyzed (Figure 1B).

Glycine methylation pathways (Figure 1A) were generally underrepresented in the studied genomes/transcriptomes (Figure 1B); the complete GSMT/SDMT- and GSDMT-catalyzed pathways occurred in 18% and
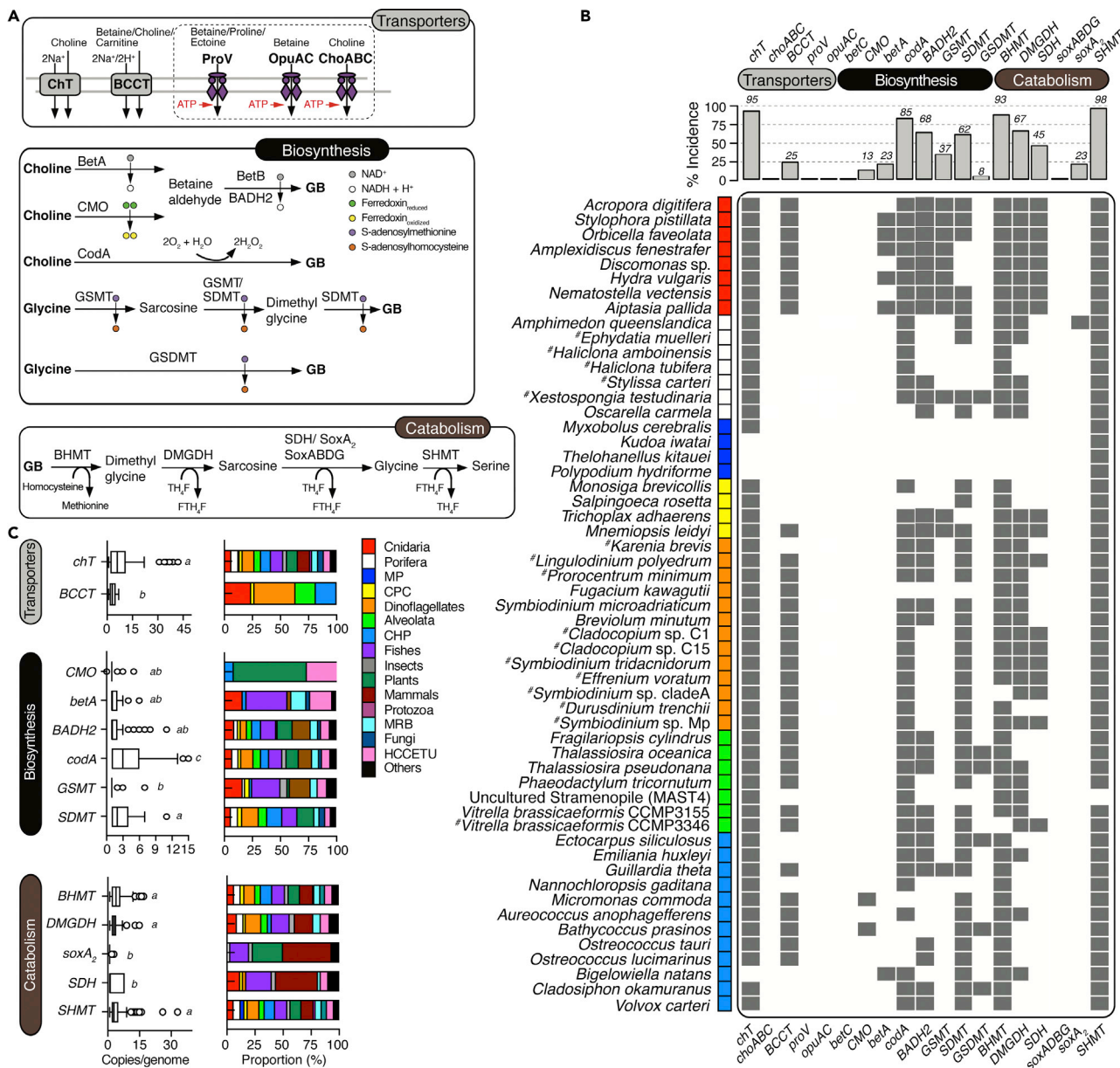
Figure 1. Incidence of Glycine Betaine (GB) Metabolic Pathways in Eukaryotes

(A) Schematic view of pathways for the uptake, biosynthesis (via oxidation and methylation), and catabolism of GB.

(B) Incidence of key genes in eukaryotic genomes (n = 114) and transcriptomes (n = 16) in the upper panel and diverse marine invertebrate lineages (and their close relatives) harboring BCCT carriers (lower panel). Data from transcriptomes are designated with "#". Additional information is provided in Tables S1 and S2.

(C) Copy numbers of predicted genes in each genome (excluding transcriptomes) and the taxonomic distribution of abundant metabolic pathways based on the presence-absence of key genes in the sampled genomes and transcriptomes. Statistically significantly different average copy numbers per genome are denoted by different letters as deduced by one-way ANOVA tests (p< 0.001) with two-stage step-up Benjamin-Krieger-Yekutieli's multiple comparison test. Enzymes: ChT, choline transporter; BCCT, betaine-choline-carnitine transporters; BetA, choline dehydrogenase; CMO, choline monooxygenase; BetB/BADH2, betaine aldehyde dehydrogenase; CodA, choline oxidase; GSMT, glycine-sarcosine methyltransferase; SDMT, sarcosine dimethyltransferase; GSDMT, glycine sarcosine dimethylglycine N-methyltransferase; BHMT, betaine-homocysteine methyltransferase; DMGDH, dimethylglycine dehydrogenase; SDH, eukaryotic sarcosine dehydrogenase; SoxA$_2$, monomeric sarcosine oxidase; SoxABDG, heterotetrameric sarcosine oxidase. Abbreviations: Na$^+$, sodium ion; H$^+$, proton; NAD$^+$ and NADH, oxidized and reduced nicotinamide adenine dinucleotide; TH$_4$F, tetrahydrofolate; FTH$_4$F, formyl-tetrahydrofolate; MP, Myxozoa and Polypodiozoa; CPC, Choanoflagellates, Placozoa and Ctenophora; CHP, Cryptophytes, Haptophytes, and Plantae (Algae); MRB, Mollusca, Rotifera, and Brachiopoda; HCCETU, Hemichordates, Cyclostomes, Cephalochordates, Echinoderms, Tunicates, and Urochordates. Species indicated as "Others" are listed in Table S2.

~8% of the genomes respectively. Interestingly, the GSMT/SDMT-catalyzed pathway was also predicted in five coral genomes but found lacking in Symbiodiniaceae (Figure 1B). Although further biochemical evidence is warranted to establish activities of all these pathways, the possession of the *GSMT* gene putatively encoding glycine-sarcosine methyltransferase by cnidarians, a demosponge (*Amphimedon queenslandica*), and many of the investigated mammalian genomes (Figure 1B; Table S2), suggests the potential to turnover glycine to dimethylglycine (Lai et al., 2006; Nyyssölä et al., 2001), which also counteracts osmotic and thermal stresses (Bashir et al., 2014). In summary, corals and their dinoflagellate endosymbionts alongside other marine invertebrates are deduced to possess at least one pathway for producing GB *de novo*, which points to the significance of GB in coral biology and underscores reef-building corals as a potential source of GB in the ocean.

The Ability to Scavenge Exogenous GB Is Taxonomically Constrained but Present in the Coral Host, Microalgae, and Bacteria of Coral Metaorganisms

Intriguingly, genomic profiling of transporters involved in GB metabolism demonstrates that the ability to scavenge extracellular GB in higher organisms is constrained to very few eukaryotic lineages, namely cnidarians, ctenophore, diatoms, dinoflagellates, ciliates, and cryptophytes (Figure 1B). The vast majority of eukaryotic genomes/transcriptomes investigated here encode the putative sodium-dependent high-affinity choline carrier ChT (30%–50% identity to validated proteins; Figure S1), which can take up choline for GB biosynthesis (Figures 1B and 1C; Table S2). However, only 25.4% of the eukaryotes harbored a potential GB uptake system (Figures 1B and 1C). Crucially, the eukaryotes that do were predicted to possess numerous copies of genes putatively encoding the betaine-choline-carnitine-transporter (BCCT) family (30%–40% identity to validated carriers; Figure S1). BCCT carriers are sodium- or proton-coupled transporters that are well characterized exclusively in prokaryotes, where they canonically mediate high-affinity uptake of betaines, choline, or carnitine (Ziegler et al., 2010). Homologs of BCCT carriers were also found in the genomes/transcriptomes of diverse Symbiodiniaceae species (Figures 1B and 2A), the intracellular dinoflagellate symbionts of cnidarians, suggesting the potential capacity to accumulate betaine, choline, or carnitine *in hospite*. Of evolutionary significance is the fact that *BCCT* homologs were found only in a few animals hosting microalgal endosymbionts (LaJeunesse et al., 2018) or close photosynthetic relatives, suggesting an important role in the cnidarian-dinoflagellate symbiosis.

It is noteworthy that the alternative ABC-type transport systems commonly found in prokaryotes for GB (OpuA and OpuC) or choline (OpuB, ProU, and ChoABC) (Ziegler et al., 2010) are lacking in the eukaryotes studied here. Also, only the diatom *Thalassiosira pseudonana*, the cryptomonad alga *Guillardia theta* and the coccolithophore *Emiliania huxleyi* encoded putative homologs of the choline and GB exporter EmrE (Tables S3 and S4), which exports naturally occurring quaternary cation compounds (Bay and Turner, 2012). Taken together, the systematic evaluation of potential GB uptake systems across genomes/transcriptomes of model and non-model metazoans depicts a confined distribution of BCCT and EmrE carriers to globally important marine invertebrates such as cnidarians and microalgae.

Cnidarians host a diverse suite of prokaryotes (Rohwer et al., 2002). Therefore, we sought to explore whether their bacterial associates also contribute to GB metabolism. Focusing on the ubiquitous *Endozoicomonas* species (*n* = 13), encompassing bacteria that seem to commonly colonize tissues of diverse cnidarians (Neave et al., 2016, 2017a, 2017b), our data show that not all species of these "obligate" endosymbionts have the putative machinery to synthesize and export GB though all possessed the ability to scavenge GB (Figure S3). These findings are also consistent with the incidence of profiled synthetic and uptake routes for GB across 52 metagenome-assembled prokaryotic genomes [MAGs; (Robbins et al., 2019)], representing the most abundant bacterial and archaeal taxa in the environmentally resilient reef-building coral *Porites lutea* (Table S5).

The majority of the genomes queried (56 out of 64) encoded BHMT proteins that can potentially convert GB to dimethylglycine, but most—all *Endozoicomonas* species and one-third of MAGs from *P. lutea*—lack putative enzymes for catabolizing dimethylglycine to sarcosine (Figure S3; Table S5). Thus, we speculate that hosting bacteria with reduced degradative capacity for GB maintains a high intracellular concentration of its precursor dimethylglycine—also an osmolyte, in coral tissues. Considering the ubiquity of *Endozoicomonas* (and other microbes) in coral microbiomes (Neave et al., 2016, 2017a, 2017b), we infer that the production of osmotically active compounds might be a cooperative mechanism supporting coral holobiont physiology (Ochsenkühn et al., 2017). Although speculative at this point, we propose that the coral host and its associated microbes—that is, Symbiodiniaceae and some bacteria—produce GB from choline or
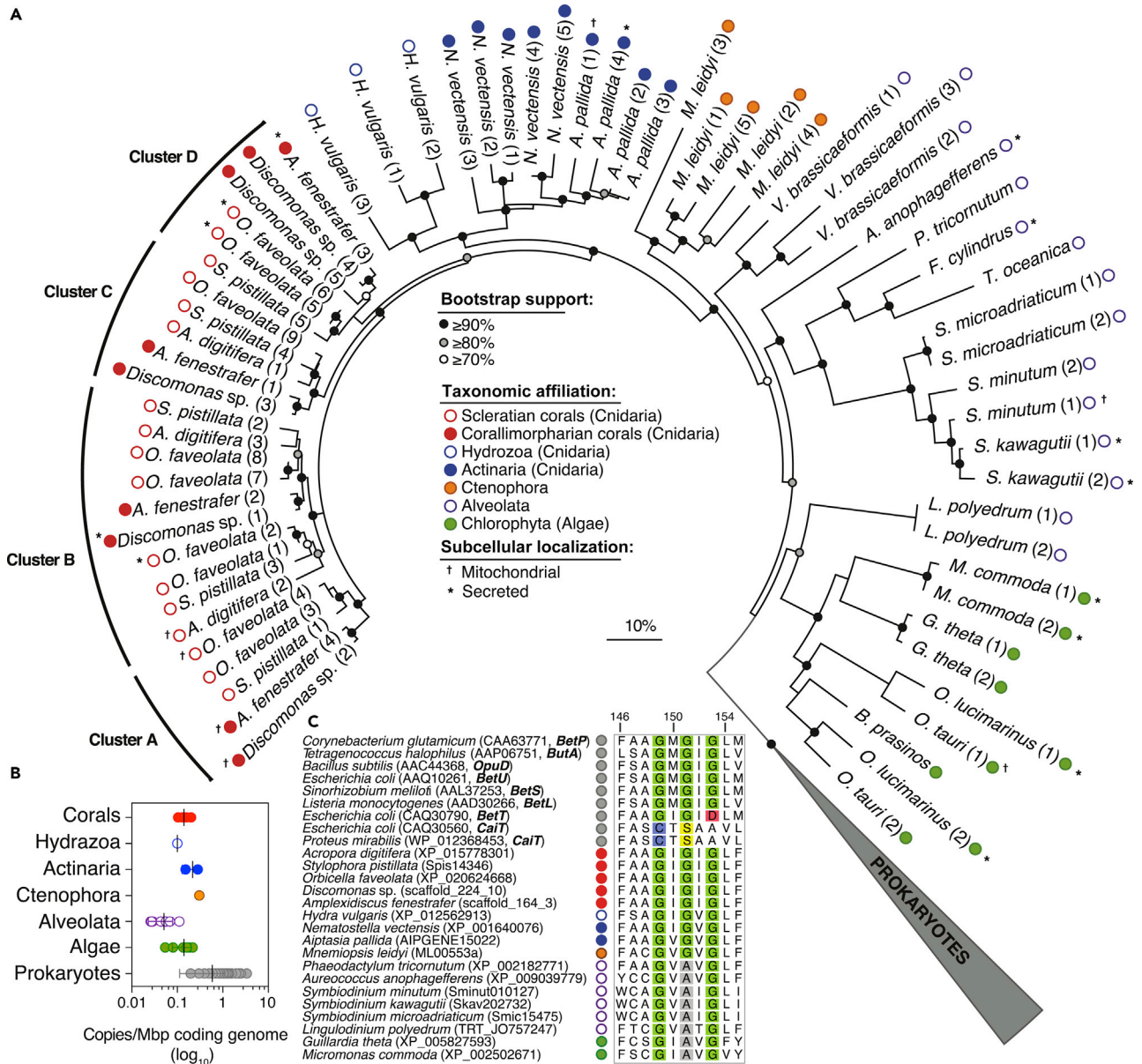
Figure 2. Evolutionary History of Predicted Eukaryotic BCCT Carriers

(A) An unrooted maximum-likelihood phylogenetic tree showing the affiliation of predicted coral BCCT proteins (red circles) next to their closest cnidarian relatives (blue circles) and the basal prokaryotic BCCT carriers (wedged gray symbol). Colored circles highlight taxonomic affiliation of eukaryotic BCCT proteins. Note that the coral BCCTs are divided into four clusters (≥80% bootstrap support), with some putatively located in the mitochondria.

(B) The average density of predicted *BCCT* genes in marine invertebrates and marine prokaryotes with fully sequenced genomes (*n* = 125; Table S6); difference in copy number per mega-base pair (Mbp) of coding genome was only significant (one-way ANOVA, p = 0.0012) for Prokaryotes (gray circles) relative to Alveolata (green circles).

(C) Multiple aminoacid sequence alignment showing the G-x-G-x-G motif (highlighted in green) found in validated prokaryotic BCCT carriers specific for GB (BetP, ButA, OpuD, BetU, and BetL), choline (BetT), and carnitine (CaiT). Glycine residues lacking in choline and carnitine carriers are colored in red and blue (and yellow) respectively, whereas those missing in Symbiodiniaceae proteins are shown in gray. The amino acid positions reflect those in BetP from *C. glutamicum*. For brevity, only representative eukaryotic proteins are shown; Figure S6 provides the complete data.

glycine and facilitate the exchange of GB scavenged from prey or the ambient seawater (by the host) or generated *in situ* (by the entire metaorganism) within the holobiont. This hypothesized model, however, requires experimental validation.

Eukaryotic BCCT Homologs Are Distant Relatives of Prokaryotic GB Carriers

The narrow distribution of putative *BCCT* genes in diverse eukaryotes studied here (Figure 1B) warrants elucidating the evolutionary history of the deduced transport system. Phylogenetic analyses indicate that the predicted eukaryotic BCCT carriers are distantly related to *bona fide* BCCT transporters from prokaryotes (Figure 2A). Moreover, these carriers group into discrete sequence clusters that reflect the metazoan evolutionary history (Derelle et al., 2015). Cnidarian sequences also form multiple protein clusters (A to D; Figure 2A), which are interspersed by proteins containing signal peptides or mitochondrial-like sequence motifs (Figure 2A), suggesting multiple subcellular localization of the carrier in coral tissues. Collectively, these results imply that the acquisition of BCCT carriers likely occurred independently *sensu stricto* before the diversification of metazoans.

Further analyses comparing copy numbers of predicted *BCCT* genes across different species indicate that the average copies ($\pm$ SD) scaled by the coding genome size (in mega base pairs, Mbp), were only significantly different (one-way ANOVA; p = 0.0012; Figure 2B) between marine prokaryotes (0.59 $\pm$ 0.48 copies Mbp^{-1}; *n* = 125; Table S6) and alveolates (0.03–0.11 copies Mbp^{-1}; *n* = 8). The genomic density of predicted *BCCT* genes is also similar among coral species with different life histories ($\sim$0.5 $\pm$ 0.2 copies Mbp^{-1}; *n* = 4–5 replicates per species) as deduced from metagenomic coding sequences of ten Red Sea scleractinian coral species (Figure S4; Table S7). However, the clustering of similar sequences at 95% nucleotide identity over 80% of the length of the shorter sequence reveals specificity of predicted BCCT copies to individual coral species (Figure S5). Taken together, this suggests that *BCCT* copy number and gene isoforms are highly conserved in cnidarians but evolved species specifically.

Coral BCCT Homologs Likely Transport GB as Their Cargo

Additional bioinformatic analyses strongly support the annotation of the predicted eukaryotic *BCCT* genes as GB transporters. Multiple sequence alignments of the predicted amino acid sequences and functionally characterized BCCT proteins with varied substrate specificity (Figures 2C, S6, and S7) revealed conservation of key active site glycine residues (G-x-G-x-G; Figures 2C and S6) contributing to sodium and substrate binding, as well as tryptophan and tyrosine residues (Figure S7) that catalyze GB binding in sodium-coupled betaine symporters (Perez et al., 2011; Ressl et al., 2009). Interestingly, the central glycine residue in the case of Symbiodiniaceae and some other algal species is substituted by alanine (G-x-A-x-G; Figures 2C and S6), in a region that renders the substrate-binding site accessible (Perez et al., 2011). Although further experimental validation is required, we hypothesize altered kinetics and reduced affinity for GB in Symbiodiniaceae BCCT carriers relative to host BCCT carriers, based on point mutation studies of the GB carrier of *Corynebacterium glutamicum* (Perez et al., 2011). It is noteworthy that betaine carriers can also transport small quantities of structurally related organic solutes such as proline betaine, dimethylsulfoniopropionate (DMSP), and dimethylglycine (Tandon et al., 2020; Ziegler et al., 2010) and that transportation is also fully reversible (Farwick et al., 1995).

High Expression of GB Metabolism in Aiptasia and Pelagic Symbiodiniaceae

To gain further insight on the relevance of the predicted synthetic and transport pathways for GB in coral physiology, we analyzed previously published transcriptomic data from the coral model system Aiptasia, encompassing different developmental and symbiotic states (Baumgarten et al., 2015). Gene expression analyses indicated that the pathways involved in the metabolism of GB were highly expressed in both symbiotic and aposymbiotic states (i.e., with and without microalgal symbionts) across larval and adult stages (Figure 3). For instance, the choline transporter ChT was highly expressed in aposymbiotic larvae, whereas the BCCT carrier was expressed similarly in larvae and adult stages. Genes putatively involved in the production of GB from choline oxidation (*betA*/*BADH2* and *codA*) were significantly highly expressed (one way ANOVA, p > 0.001) in adult anemones, especially in symbiotic colonies (e.g., *betA* and *codA*). In contrast, the glycine methylation pathway catalyzed by GSMT and SDMT was differentially highly expressed in symbiotic and aposymbiotic anemones, respectively (Figure 3). Crucially, transcription was in both cases, four-fold higher than choline oxidation pathways catalyzed by BetA/BADH2 and CodA. The degradative pathway catalyzed by BHMT, DMGDH, SDH, and SHMT, however, followed a trend of higher expression in symbiotic adult anemones (Figure 3), suggesting a role of Symbiodiniaceae in GB catabolism. Our results therefore suggest a greater contribution of GSMT-/SDMT-catalyzed pathway to GB production and the turnover of intracellular GB by Symbiodiniaceae.

We next evaluated whether free-living populations of the microalgal symbionts actively transcribed the metabolic pathway for GB in the global ocean by interrogating a catalog of 116 million expressed
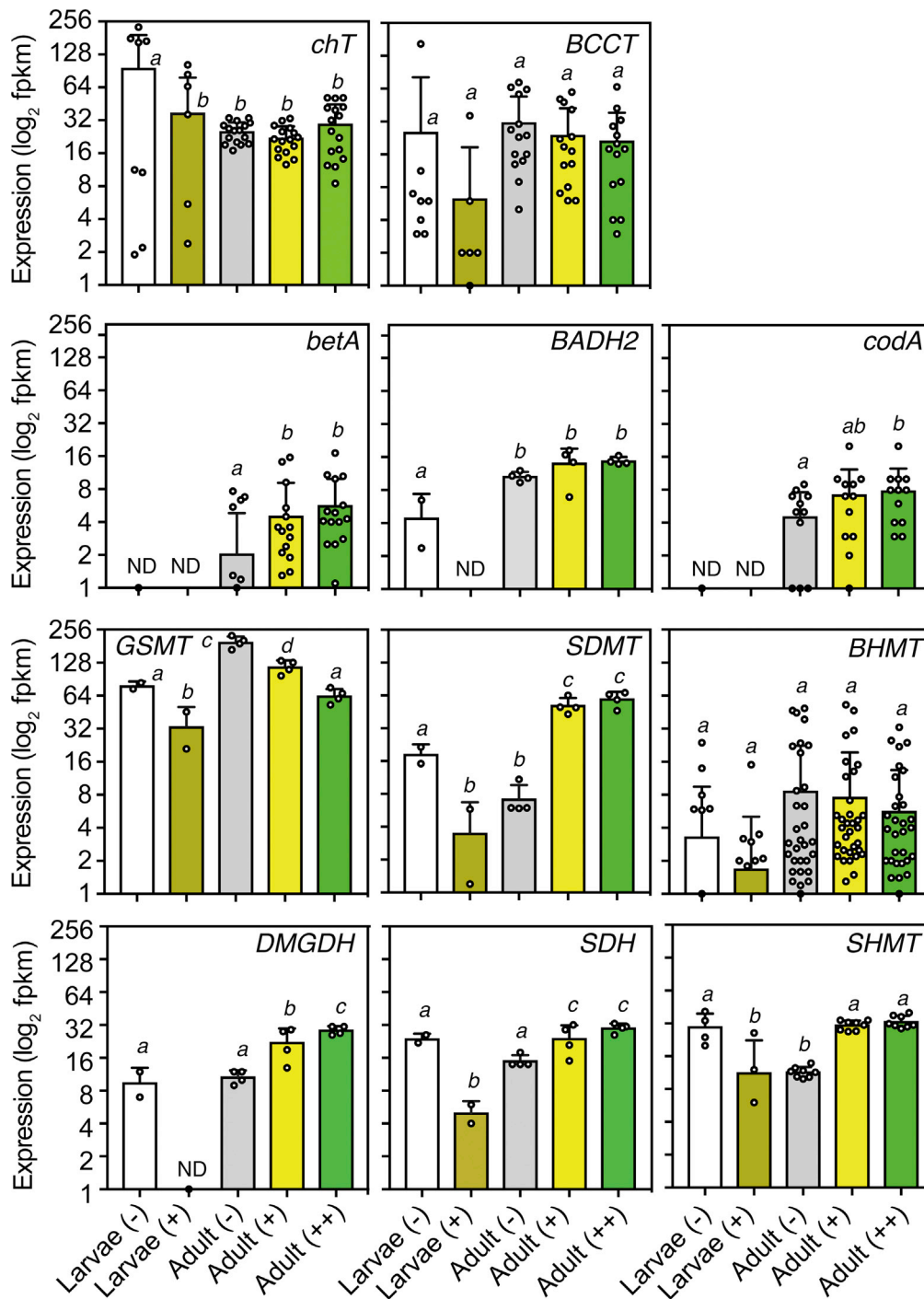
Figure 3. Transcriptional Activity of GB Metabolism in the Coral Model Aiptasia

Gene expression was determined from published whole tissue transcriptomes (Baumgarten et al., 2015) of aquacultured Aiptasia larvae and adults fed regularly on brine shrimps (n = 2–4 replicates per experiment). Transcriptomes encompass various developmental and symbiotic states. Animals without dinoflagellate endosymbionts are indicated with "–", whereas adults with intermediate or full endosymbiont levels are denoted with "+" and "++", respectively. Barplots indicate average ($\pm$ SD) expression in replicated samples and all predicted gene homologs (denoted by circular symbols). Statistically significantly different expression levels are denoted by different letters as deduced by one-way ANOVA tests ($p > 0.05$) with two-stage step-up Benjamin-Krieger-Yekutieli's multiple comparison test. ND, average expression was below ~1 FPKM.
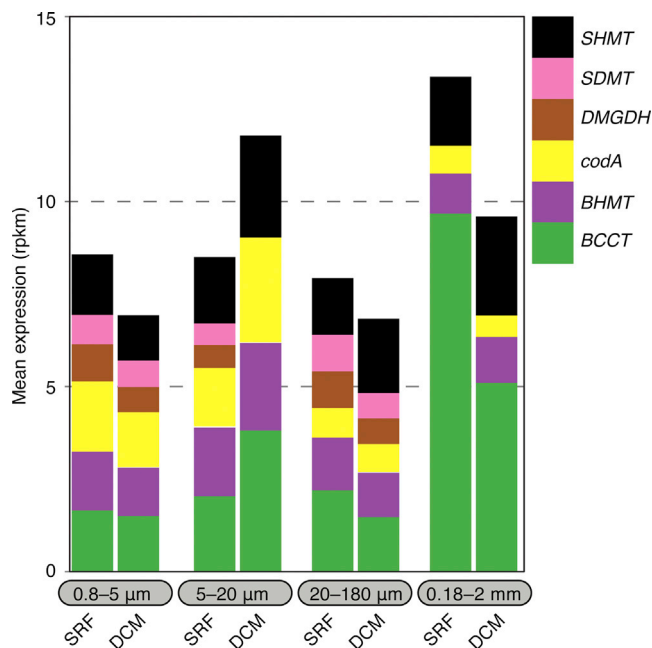
Figure 4. Expression of GB Metabolic Pathways in Free-Living Symbiodiniaceae

Bars shown mean expression in different size fractions and euphotic zone (subsurface, SRF and deep chlorophyll maximum, DCM) based on the eukaryotic unigene catalog of *Tara* Ocean from (Carradec et al., 2018). RPKM (reads per kilo base covered per million mapped reads) denotes normalized expression values.

eukaryotic genes of the *Tara* Oceans Expedition (Carradec et al., 2018) available at http://www.genoscope.cns.fr/tara/. We retrieved several unigenes with significant homology to Symbiodiniaceae genes (≥80% identity and alignment length of ≥60%) involved in GB metabolism, including *BCCT* (33), *codA* (1), *DMGDH* (3), *SDMT* (1), *SHMT* (6), and *BHMT* (31). The genes encoding for BCCT uptake and exchange system for GB were the most highly expressed among the six examined genes (Figure 4). Expression was highest in the mesoplankton size fraction (180–2000 μm)—suggested to correspond to Symbiodiniaceae in symbiosis with their host organisms (Decelle et al., 2018), especially in subsurface samples (Figure 4). In contrast, the *codA* gene encoding the one-step enzymatic route of GB production was highly expressed in the smaller (0.8–20 μm) size fractions (i.e., free-living Symbiodiniaceae). The expression levels of transport (BCCT) and synthesis (CodA) machineries suggest differences in the metabolism of GB between free-living and symbiotic lifestyles, with higher GB uptake potential inferred in symbiotic Symbiodiniaceae although experimental validation is needed.

A Putative Role for GB in Stress Resilience and Nitrogen Storage of Coral Metaorganisms

The "omics" evidence presented here strongly supports the metabolism of GB as a genetic facet of coral physiology and symbiotic interactions, especially given our estimates that GB comprises about 16% of the total nitrogen biomass of corals [based on the predicted biomass of corals of about 30 g C m^{-2} (Crossland et al., 1991), a median GB concentration in coral biomass of 6 μmol cm^{-2} (Hill et al., 2010), and a C:N ratio of 6.5 in coral tissues]. Despite its osmolytic properties, GB may have additional properties that supersede other osmolytes, such as DMSP. For instance, in plants GB inhibits thermal and photon stress by stabilizing extrinsic proteins vital for the photosynthetic and oxygen-evolving complex of photosystem II (Chen and Murata, 2002). A similar role is hypothesized in microalgae and corals (Hill et al., 2010), because corals located at shallower depths and locations exposed to high irradiance possess elevated pools of GB relative to colonies living in deeper waters and shaded areas (Hill et al., 2010). Moreover, although corals are osmoconformers (i.e., in osmotic equilibrium with their environment), their partnership with Symbiodiniaceae that formed at least 160 million years ago (LaJeunesse et al., 2018) probably generated osmotic constraints, whereby the coral host had to devise a suitable intracellular environment for its endosymbionts (Ochsenkühn et al., 2017).

In addition, we postulate that the thermodynamic constraints imposed by salinity (Oren, 1999) coupled with the inherent biochemical properties of GB rendered this osmolyte particularly suitable for a number of

reasons. Firstly, it is among the "cheapest" osmolytes to produce in nature (Oren, 1999). Secondly, the strategy of scavenging this ubiquitous marine osmolyte using ATP-independent carriers is also metabolically less costly than *de novo* biosynthesis. Thirdly, it is superior to other osmolytes, including successively methylated derivatives of glycine, regarding the degree and effectiveness of offsetting salt-induced stress on enzymes (Courtenay et al., 2000; Yancey et al., 1982). Fourthly, it represents a readily available, albeit significant internal nitrogen stock, which can be degraded or transformed into other N-containing compounds such as amino acids. Likely, GB can be replaced by DMSP under nitrogen limitation (Keller et al., 1999a), so that nitrogen is utilized for essential processes such as amino acid and protein synthesis. Lastly, the metabolism of GB might be a key determinant of one-carbon metabolism supporting methyl-dependent processes (such as DNA methylation and protein synthesis). Of note, the reaction catalyzed by BHMT (Figure 1) is also not compromised by the lack of B-type vitamins (Holm et al., 2007)—which are often-limiting nutrients in the ocean (Sañudo-Wilhelmy et al., 2014). Because carbon and nitrogen metabolism are generally closely linked, we hypothesize that the symbiosis between corals and their carbohydrate-producing microalgal symbionts may have selectively enhanced the sequestration and storage of GB to support metabolic activity by maintaining a suitable osmotic environment, while serving as a reservoir for nitrogen to accommodate the growth requirements of the endosymbionts.

Corals and Reef Ecosystems Are a Sink for Dissolved GB in the Sea

It is worthwhile to note that the uptake of GB by the high-affinity membrane carriers deduced here might be effective at the ambient dissolved (picomolar range) concentrations reported in the ocean (Muslin, 2017), especially given the high affinity for GB (K_m of up to 8.6 μM) in prokaryotic BCCT carriers (Farwick et al., 1995; Peter et al., 1996). However, the ingestion of plankton (Houlbreque and Ferrier-Pages, 2009) most likely comprises the main source of GB in coral tissues since heterotrophic intake (i.e., feeding) can deliver three to four times more nitrogen than dissolved sources in model coral species (Houlbreque and Ferrier-Pages, 2009). The fact that intracellular concentrations of GB make up a significant fraction (7%–23%) of the organic N in marine phytoplankton (Keller et al., 1999a; King, 1988) implies that incidental GB acquired from feeding might be important in the nitrogen economy of coral reefs. Following this notion, we reasoned that, if heterotrophic input via feeding is vital for a coral's nutritional ecology, then GB pool sizes in corals have to be placed in the context of their planktonic prey, which likely includes GB producers.

Therefore, as a first step, we profiled the biosynthetic pathways of GB in 189 completely sequenced microbial genomes [mean ($\pm$SD) completeness of 99 $\pm$ 1%] retrieved from the marine microbial reference (Mar-Ref) genome database (Klemetsen et al., 2018) encompassing diverse species from various oceanic provinces (Table S8). The inventory of genes encoding putative synthetic enzymes for GB showed a high incidence of choline-oxidizing pathways (in 6%–~55% of the genomes) and low occurrence of glycine-methylation-based pathways (up to ~17%; Figure S8A), possibly reflecting high energetic costs of methylation reactions, particularly the production of the methyl donor/acceptor *S*-adenosylmethionine (Luka et al., 2009; Nyyssölä et al., 2000). The co-occurrence of choline-oxidizing enzyme complements in individual genomes (Figure S8B) further revealed the predominance of the BetA/BADH2 and BetA/BetB-catalyzed pathways (in ~36% and 13%, respectively) relative to the pathway involving choline monooxygenase (CMO). Also, a low-average copy number per genome (Figure S8C) was observed for *betA* (1 copy; n = 71 genomes), *betB* (1 copy; n = 26), and *BADH2* (1 copy; n = 72) in comparison to *codA* (2.3 copies; n = 104), where the high coefficient of variation of 67% (versus 11% in *betA*) around the mean suggested significant species-dependent copy variations. Considering the high incidence of co-occurrence of single-copy genes for the BetA/BADH2-catalyzed pathway in reference marine microbial genomes (Figure S8B; Table S8) implies that *betA* and *BADH2* homologs are feasible proxies for estimating the relative abundance of GB producers from environmental metagenomes, because they have a greater likelihood of co-occurring in individual genomes. It is also noteworthy that two-thirds of the genomes with *codA* (55% of the genomes) also possessed the complete BetA/BADH2 pathway.

On this basis, we subsequently probed the abundance of *betA* and *BADH2* protein-coding genes by interrogating global ocean metagenomic sequence datasets and normalized counts of retrieved genes using the mean abundance of four conserved single-copy housekeeping genes present in microbial genomes (Table S9). The two respective genes were predicted to be present in 2.4% and 1.2% of prokaryotes sampled in the *Tara* Oceans marine metagenomes (Sunagawa et al., 2015), which encompass picoplankton-sized microorganisms (0.22–3 μm) from the photic and epipelagic layers (Table S9). The mean of roughly 2% suggests that the environmental abundance of picoplankton producing GB is twice that of

DMSP-producing microbes (Curson et al., 2017) in the pelagic ocean; however, the value could be higher, if we consider picoplankton encoding the one-step CodA-catalyzed pathway.

Based on the above estimates [i.e., a 2% proportion of GB-producing picoplankton, a cellular concentration of GB between 0.9 and 483 nM in seawater particulates >0.2 μm (Beale and Airs, 2016; Cree, 2015; Keller et al., 2004), $\sim 10^5$ cells per ml, and a total biomass of microbial picoplankton in the pelagic ocean of 3.6×10^{28} cells (Whitman et al., 1998)], we project a global GB stock in pelagic marine picoplankton in the range of ~ 1 to 410×10^{12} grams. Scaling these estimates to the areal cover of reefs of $\sim 0.1\%$ of the world's ocean surface (Spalding et al., 2001), and assuming the global average water depth in areas occupied by reefs of 20 m, suggests that between 0.1 to 41×10^9 grams of GB is potentially distributed within picoplanktonic particulates in the waters overlaying corals.

These estimates are similar to the predicted global stock of GB in corals of about 18×10^6 to 4.64×10^9 grams, based on the GB concentration in coral tissues of 0.05–13.2 μmol cm^{-2} (Hill et al., 2010), the molecular weight of GB (117.15 g mol^{-1}), and a mean areal coral reef cover of 0.3×10^6 km^2 (Spalding et al., 2001). However, the predicted global stock of GB present in corals may be significantly underestimated considering that much of the GB is likely catabolized by corals and their symbionts (e.g., via dimethylglycine to serine) or translocated through coral mucus ropes into the adjacent sediments and pore waters. Thus, at the global scale, coral reef ecosystems and their reef-building corals represent a rare, albeit significant and dynamic "hot spot" of microbial-derived GB in the pelagic ocean.

Conclusions

The systematic bioinformatic framework applied here combined with existing biochemical data strongly suggests a role of marine invertebrates in the biogeochemical cycling of GB in the ocean, adding corals to the roster of sinks and organisms producing GB. Because GB is a major reservoir of nitrogen and considering the general scarcity of nitrogen in the pelagic ocean (Bristow et al., 2017), we propose that GB serves both as an osmolyte and as a potent nitrogen reserve in marine environments hosting coral reefs. Moreover, given its photoprotective role in plants including alleviating oxidative stress (Chen and Murata, 2008), we speculate further that GB serves thermo- and photoprotective roles in the coral holobiont. This in turn implicates GB in the integrity of the coral-microalgal symbiosis and the resilience of coral reef ecosystems, further reiterating the notion that coral reefs are biological hotspots, acting as sinks for dissolved organic nitrogen in oligotrophic seas. Our findings open a path for a number of testable hypotheses requiring in-depth biochemical and experimental analysis for (1) the many uncharacterized genes and pathways we elucidated and postulated to play a role in regulating GB uptake, storage, and cycling, as well as (2) the hypothesized protective and nutritional roles of GB in the metabolism of coral holobionts.

Limitations of the Study

In the current study, we examined the genetic potential of corals to metabolize GB, which is present in their tissues at high concentrations. Most of the bioinformatic analyses are based on genomic/transcriptomic data, questioning whether the inference for GB production based on homologs is sufficient in the absence of biochemical validation of the deduced enzymes. A second limiting factor in robust interpretation of our results relates to the localization of the enzyme activities within the coral holobiont and the concentrations of GB in the studied animals species. These limitations call for ratification of the pathways that we deduced to function in GB metabolism in corals and their microbial symbionts and elucidating the contribution of coral holobiont partners in shaping the dynamics of intracellular GB pools.

Resource Availability

Lead Contact

david.ngugi@dsmz.de.

Materials Availability

All data needed to evaluate the results and conclusions are presented in the main text and/or are available in the Supplementary Materials. Additional data or scripts related to this paper are available from the corresponding authors.

Data and Code Availability

The Raw sequence data for the coral metagenomes have been deposited in NCBI under BioProject number PRJNA437202 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA437202).

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101120.

AUTHOR CONTRIBUTIONS

D.K.N. conceived and designed the study and performed the bioinformatic analyses. M.Z. and C.R.V. determined and provided the metagenomic data. C.M.D. provided the bioinformatic resources for the analyses. D.K.N. wrote the manuscript with contributions from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Airs, R.L., and Archer, S.D. (2010). Analysis of glycine betaine and choline in seawater particulates by liquid chromatography/electrospray ionization/mass spectrometry. Limnol. Oceanogr. Methods 8, 499–506.

Bashir, A., Hoffmann, T., Smits, S.H.J., and Bremer, E. (2014). Dimethylglycine provides salt and temperature stress protection to *Bacillus subtilis*. Appl. Environ. Microbiol. 80, 2773–2785.

Baumgarten, S., Simakov, O., Esherick, L.Y., Liew, Y.J., Lehnert, E.M., Michell, C.T., Li, Y., Hambleton, E.A., Guse, A., Oates, M.E., et al. (2015). The genome of *Aiptasia*, a sea anemone model for coral symbiosis. Proc. Natl. Acad. Sci. U S A 112, 11893–11898.

Bay, D.C., and Turner, R.J. (2012). Small multidrug resistance protein EmrE reduces host pH and osmotic tolerance to metabolic quaternary cation osmoprotectants. J. Bacteriol. 194, 5941–5948.

Beale, R., and Airs, R. (2016). Quantification of glycine betaine, choline and trimethylamine N-oxide in seawater particulates: minimisation of seawater associated ion suppression. Anal. Chim. Acta 938, 114–122.

Bristow, L.A., Mohr, W., Ahmerkamp, S., and Kuypers, M.M.M. (2017). Nutrients that limit growth in the ocean. Curr. Biol. 27, R474–R478.

Burg, M.B., and Ferraris, J.D. (2008). Intracellular organic osmolytes: function and regulation. J. Biol. Chem. 283, 7309–7313.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., et al. (2018). A global ocean atlas of eukaryotic genes. Nat. Commun. 1–13, https://doi.org/10.1038/s41467-017-02342-1.

Chen, T.H.H., and Murata, N. (2008). Glycine betaine: an effective protectant against abiotic stress in plants. Trends Plant Sci. 13, 499–505.

Chen, T.H.H., and Murata, N. (2002). Enhancement of tolerance of abiotic stress by metabolic engineering of betaines and other compatible solutes. Curr. Opin. Plant Biol. 5, 250–257.

Courtenay, E.S., Capp, M.W., Anderson, C.F., and Record, M.T. (2000). Vapor pressure osmometry studies of osmolyte−protein interactions:

implications for the action of osmoprotectants *in vivo* and for the interpretation of "osmotic stress" experiments *in vitro*. Biochemistry 39, 4455–4471.

Cree, C. (2015). Distributions of glycine Betaine and the Methylamines in Coastal Waters: Analytical Developments and a Seasonal Study (University of Plymouth).

Crossland, C.J., Hatcher, B.G., and Smith, S.V. (1991). Role of coral reefs in global ocean production. Coral Reefs 10, 55–64.

Curson, A.R.J., Liu, J., Martínez, A.B., Green, R.T., Chan, Y., Carrión, O., Williams, B.T., Zhang, S.-H., Yang, G.-P., Page, P.C.B., et al. (2017). Dimethylsulfoniopropionate biosynthesis in marine bacteria and identification of the key gene in this process. Nat. Microbiol. 2, 17009.

Decelle, J., Carradec, Q., Pochon, X., Henry, N., Romac, S., Mahé, F., Dunthorn, M., Kourlaiev, A., Voolstra, C.R., Wincker, P., and de Vargas, C. (2018). Worldwide occurrence and activity of the reef-building coral symbiont Symbiodinium in the open ocean. Curr. Biol. 28, 3625–3633.e3.

Derelle, R., Torruella, G., Torruella, G., Klimeš, V., Klimeš, V., Brinkmann, H., Brinkmann, H., Kim, E., Kim, E., Vlček, Č., et al. (2015). Bacterial proteins pinpoint a single eukaryotic root. Proc. Natl. Acad. Sci. U S A 112, E693–E699.

Farwick, M., Siewe, R.M., and Kramer, R. (1995). Glycine betaine uptake after hyperosmotic shift in Corynebacterium glutamicum. J. Bacteriol. 177, 4690–4695.

Gebser, B., and Pohnert, G. (2013). Synchronized regulation of different zwitterionic metabolites in the osmoadaption of phytoplankton. Mar. Drugs 11, 2168–2182.

Hagedorn, M., Carter, V.L., Ly, S., Andrell, R.M., Yancey, P.H., Leong, J.A.C., and Kleinhans, F.W. (2010). Analysis of internal osmolality in developing coral larvae, Fungia scutaria. Physiol. Biochem. Zool. 83, 157–166.

Hill, R.W., Armstrong, E.J., Florn, A.M., Li, C., Walquist, R.W., and Edward, A. (2017). Abundant betaines in giant clams (Tridacnidae) and western Pacific reef corals, including study of coral betaine acclimatization. Mar. Ecol. Prog. Ser. 576, 27–41.

Hill, R.W., Li, C., Jones, A.D., Gunn, J.P., and Frade, P.R. (2010). Abundant betaines in reef-building corals and ecological indicators of a photoprotective role. Coral Reefs 29, 869–880.

Holm, P.I., Hustad, S., Ueland, P.M., Vollset, S.E., Grotmol, T., and Schneede, J. (2007). Modulation of the homocysteine-betaine relationship by methylenetetrahydrofolate reductase 677 C->T genotypes and B-vitamin status in a large-scale epidemiological study. J. Clin. Endocrinol. Metab. 92, 1535–1541.

Houlbreque, F., and Ferrier-Pages, C. (2009). Heterotrophy in tropical scleractinian corals. Biol. Rev. 84, 1–17.

Julián-Sánchez, A., Riveros-Rosas, H., Martinez-Castilla, L.P., Velasco-Garcia, R., and Muñoz-Clares, R.A. (2007). Phylogenetic and structural relationships of the betaine aldehyde dehydrogenases. In Enzymology and Molecular Biology of Carbonyl Metabolism, W. Henry, M. Edmund, L. Ronald, and P. Bryce, eds. (Purdue University Press), pp. 64–76.

Kageyama, H., Tanaka, Y., and Takabe, T. (2018). Biosynthetic pathways of glycine betaine in Thalassiosira pseudonana; functional characterization of enzyme catalyzing three-step methylation of glycine. Plant Physiol. Biochem. 127, 248–255.

Keller, M.D., Kiene, R.P., Matrai, P.A., and Bellows, W.K. (1999a). Production of glycine betaine and dimethylsulfoniopropionate in marine phytoplankton. I. Batch cultures. Mar. Biol. 135, 237–248.

Keller, M.D., Kiene, R.P., Matrai, P.A., and Bellows, W.K. (1999b). Production of glycine betaine and dimethylsulfoniopropionate in marine phytoplankton. II. N-limited chemostat cultures. Mar. Biol. 135, 249–257.

Keller, M.D., Matrai, P.A., Kiene, R.P., and Bellows, W.K. (2004). Responses of coastal phytoplankton populations to nitrogen additions: dynamics of cell-associated dimethylsulfoniopropionate (DMSP), glycine

betaine (GBT), and homarine. Can. J. Fish. Aquat. Sci. 61, 685–699.

Kiene, R.P., Williams, L., and Walker, J.E. (1998). Seawater microorganisms have a high affinity glycine betaine uptake system which also recognizes dimethylsulfoniopropionate. Aquat. Microb. Ecol. 15, 39–51.

King, G.M. (1988). Distribution and metabolism of quaternary amines in marine sediments. In Nitrogen Cycling in Coastal Marine Environments, T.H. Blackburn and J. Sorensen, eds. (John Wiley and Sons Ltd), pp. 143–173.

Kirch, H.-H., Bartels, D., Wei, Y., Schnable, P.S., and Wood, A.J. (2004). The ALDH gene superfamily of Arabidopsis. Trends Plant Sci. 9, 371–377.

Klemetsen, T., Raknes, I.A., Fu, J., Agafonov, A., Balasundaram, S.V., Tartari, G., Robertsen, E., and Willassen, N.P. (2018). The MAR databases: development and implementation of databases specific for marine metagenomics. Nucleic Acids Res. 46, D692–D699.

Lai, M.-C., Wang, C.-C., Chuang, M.-J., Wu, Y.-C., and Lee, Y.-C. (2006). Effects of substrate and potassium on the betaine-synthesizing enzyme glycine sarcosine dimethylglycine N-methyltransferase from a halophilic methanoarchaeon Methanohalophilus portucalensis. Res. Microbiol. 157, 948–955.

LaJeunesse, T.C., Parkinson, J.E., Gabrielson, P.W., Jeong, H.J., Reimer, J.D., Voolstra, C.R., and Santos, S.R. (2018). Systematic revision of Symbiodiniaceae highlights the antiquity and diversity of coral endosymbionts. Curr. Biol. 28, 2570–2580.e6.

Lambou, K., Pennati, A., Valsecchi, I., Tada, R., Sherman, S., Sato, H., Beau, R., Gadda, G., and Latge, J.P. (2013). Pathway of glycine betaine biosynthesis in Aspergillus fumigatus. Eukaryot. Cell 12, 853–863.

Lidbury, I., Kimberley, G., Scanlan, D.J., Murrell, J.C., and Chen, Y. (2015). Comparative genomics and mutagenesis analyses of choline metabolism in the marine Roseobacter clade. Environ. Microbiol. 17, 5048–5062.

Luka, Z., Mudd, S.H., and Wagner, C. (2009). Glycine N-methyltransferase and regulation of S-adenosylmethionine levels. J. Biol. Chem. 284, 22507–22511.

Muslin, O. (2017). Quantification of Osmolytes in the Sargasso Sea Surface Layer Water Column (Oregon State University).

Neave, M.J., Apprill, A., Ferrier-Pages, C., and Voolstra, C.R. (2016). Diversity and function of prevalent symbiotic marine bacteria in the genus Endozoicomonas. Appl. Microbiol. Biotechnol. 100, 8315–8324.

Neave, M.J., Michell, C.T., Apprill, A., and Voolstra, C.R. (2017a). Endozoicomonas genomes reveal functional adaptation and plasticity in bacterial strains symbiotically associated with diverse marine hosts. Sci. Rep. 7, 40579.

Neave, M.J., Rachmawati, R., Xun, L., Michell, C.T., Bourne, D.G., Apprill, A., and Voolstra, C.R.

(2017b). Differential specificity between closely related corals and abundant Endozoicomonas endosymbionts across global scales. ISME J. 11, 186–200.

Nyyssölä, A., Kerovuo, J., Kaukinen, P., vonWeymarn, N., and Reinikainen, T. (2000). Extreme halophiles synthesize betaine from glycine by methylation. J. Biol. Chem. 275, 22196–22201.

Nyyssölä, A., Reinikainen, T., and Leisola, M. (2001). Characterization of glycine sarcosine N-methyltransferase and sarcosine dimethylglycine N-methyltransferase. Appl. Environ. Microbiol. 67, 2044–2050.

Ochsenkühn, M.A., Röthig, T., D'Angelo, C., Wiedenmann, J., and Voolstra, C.R. (2017). The role of floridoside in osmoadaptation of coral-associated algal endosymbionts to high-salinity conditions. Sci. Adv. 3, e1602047.

Oren, A. (1999). Bioenergetic aspects of halophilism. Microbiol. Mol. Biol. Rev. 63, 334–348.

Perez, C., Koshy, C., Ressl, S., Nicklisch, S., mer, R.K.A., and Ziegler, C. (2011). Substrate specificity and ion coupling in the Na+/betaine symporter BetP. EMBO J. 30, 1221–1229.

Peter, H., Burkovski, A., and Kramer, R. (1996). Isolation, characterization, and expression of the Corynebacterium glutamicum betP gene, encoding the transport system for the compatible solute glycine betaine. J. Bacteriol. 178, 5229–5234.

Ressl, S., Terwissscha van Scheltinga, A.C., Vonrhein, C., Ott, V., and Ziegler, C. (2009). Molecular basis of transport and regulation in the Na$^+$/betaine symporter BetP. Nature 458, 47–52.

Robbins, S.J., Singleton, C.M., Chan, C.X., Messer, L.F., Geers, A.U., Ying, H., Baker, A., Bell, S.C., Morrow, K.M., Ragan, M.A., et al. (2019). A genomic view of the reef-building coral Porites lutea and its microbial symbionts. Nat. Microbiol. 4, 2090–2100.

Roesser, M., and Müller, V. (2001). Osmoadaptation in bacteria and archaea: common principles and differences. Environ. Microbiol. 3, 743–754.

Rohwer, F., Seguritan, V., Azam, F., and Knowlton, N. (2002). Diversity and distribution of coral-associated bacteria. Mar. Ecol. Prog. Ser. 243, 1–10.

Sañudo-Wilhelmy, S.A., Gómez-Consarnau, L., Suffridge, C., and Webb, E.A. (2014). The role of B vitamins in marine biogeochemistry. Ann. Rev. Mar. Sci. 6, 339–367.

Spalding, M.D., Ravilious, C., and Green, E.P. (2001). World Atlas of Coral Reefs (University of California Press).

Spielmeyer, A., Gebser, B., and Pohnert, G. (2011). Dimethylsulfide sources from microalgae: improvement and application of a derivatization-based method for the determination of dimethylsulfoniopropionate and other zwitterionic osmolytes in phytoplankton. Mar. Chem. 124, 48–56.

Spielmeyer, A., and Pohnert, G. (2012). Influence of temperature and elevated carbon dioxide on the production of dimethylsulfoniopropionate and glycine betaine by marine phytoplankton. Mar. Environ. Res. *73*, 62–69.

Stephens-Camacho, N.A., Muhlia-Almazan, A., Sanchez-Paz, A., and Rosas-Rodriguez, J.A. (2015). Surviving environmental stress: the role of betaine aldehyde dehydrogenase in marine crustaceans. J. Invertebr. Pathol. *12*, 66–74.

Sun, J., Steindler, L., Thrash, J.C., Halsey, K.H., Smith, D.P., Carter, A.E., Landry, Z.C., and Giovannoni, S.J. (2011). One carbon metabolism in SAR11 pelagic marine bacteria. PLoS One 6, e23973.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Structure and function of the global ocean microbiome. Science *348*, 1261359.

Tandon, K., Lu, C.-Y., Chiang, P.-W., Wada, N., Yang, S.-H., Chan, Y.-F., Chen, P.-Y., Chang, H.-Y., Chiou, Y.-J., Chou, M.-S., et al. (2020). Comparative genomics: dominant coral-bacterium *Endozoicomonas acroporae* metabolizes dimethylsulfoniopropionate (DMSP). ISME J. *55*, 1290–1303.

Welsh, D.T. (2000). Ecological significance of compatible solute accumulation by micro-organisms: from single cells to global climate. FEMS Microbiol. Rev. *24*, 263–290.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998). Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. U S A *95*, 6578–6583.

Yancey, P.H., Clark, M.E., Hand, S.C., Bowlus, R.D., and Somero, G.N. (1982). Living with water stress: evolution of osmolyte systems. Science *217*, 1214–1222.

Yancey, P.H., Heppenstall, M., Ly, S., Andrell, R.M., Gates, R.D., Carter, V.L., and Hagedorn, M. (2010). Betaines and dimethylsulfoniopropionate as major osmolytes in Cnidaria with endosymbiotic dinoflagellates. Physiol. Biochem. Zool. *83*, 167–173.

Ziegler, C., Bremer, E., and Krämer, R. (2010). The BCCT family of carriers: from physiology to crystal structure. Mol. Microbiol. *78*, 13–34.

Supplemental Information

Genomic Blueprint of Glycine Betaine

Metabolism in Coral Metaorganisms and Their

Contribution to Reef Nitrogen Budgets

David K. Ngugi, Maren Ziegler, Carlos M. Duarte, and Christian R. Voolstra

Supplemental Information

Genomic blueprint of glycine betaine metabolism in coral metaorganisms and their contribution to reef nitrogen budgets

David K. Ngugi[1,2*], Maren Ziegler[2,3], Carlos M. Duarte[2] and Christian R. Voolstra[2,4*]

Lead Contact: david.ngugi@dsmz.de
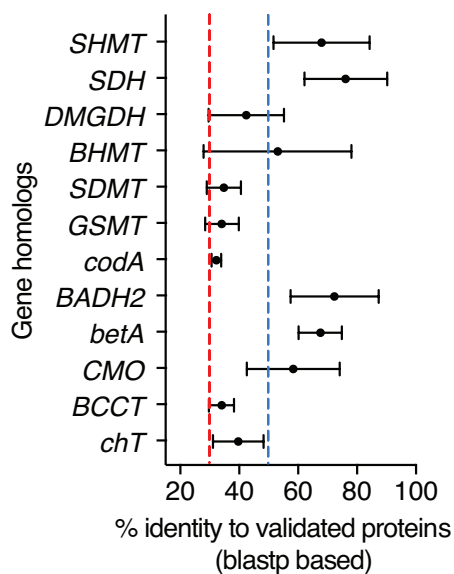
Figure S1. Related to Figure 1. Percentage identity of retrieved gene homologs to validated protein sequences from UniProt. The black dot and horizontal lines show the mean (± SD) identities. The dotted red and blue vertical lines denote identity cutoffs used for interrogating enzymes with no HMM models *CMO*, *codA*, *SDMT*, and *GSMT* (30%) and *BADH2* and *SDH* (50%) respectively. Identities for genes with HMM models (*chT*, *BCCT*, *betA*, *BHMT*, *DMGDH*, and *SHMT*) are based upon blast search of selected hmmsearch hits for each gene (see methods for details) and subsequent search of the homologs against validated proteins from UniProt.

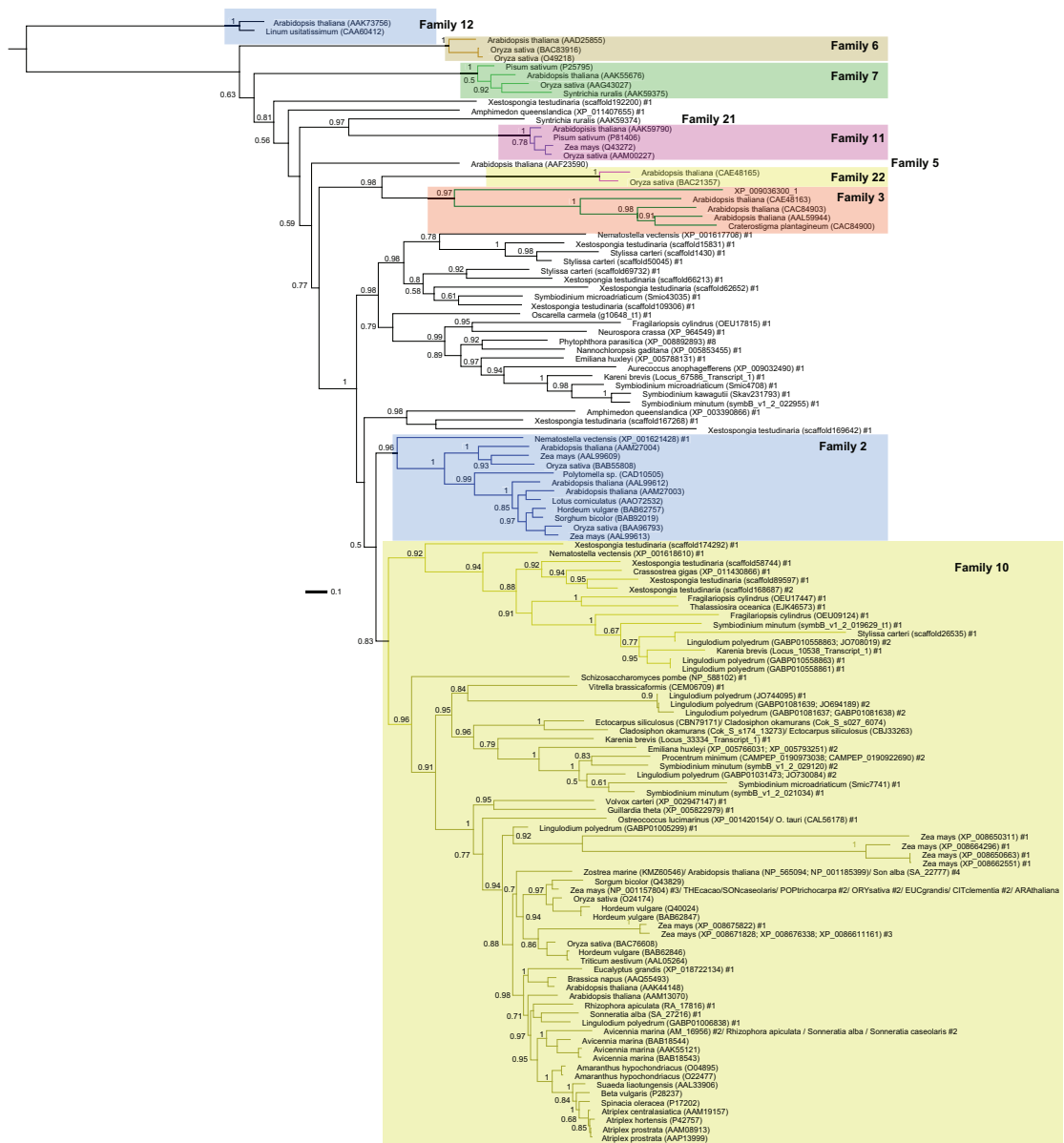**Figure S2**. Related to Figure 1. Phylogenetic analysis of betaine aldehyde dehydrogenase-like genes (*BADH2*-like) predicted in eukaryotic genomes studied here. The tree represents a maximum-likelihood tree, with all nodes having bootstrap values above 50% (out of 1,000 replicates) shown on the tree as proportions. The family designation of the different aldehyde dehydrogenases (ALDHs) follows the work of (Kirch et al., 2004) and are color-highlighted as follows: family 12 (the outgroup); family 6 (methylmalonate semialdehyde dehydrogenase), family 7 (antiquitin-related ALDHs); family 21 and 22 (characterized novel ALDHs); family 11(non-phosphorylating glyceladehyde-3-phosphate dehydrogenase); family 5 (succinic semialdehyde dehydrogenase); family 2 and 3 (ALDHs with variable substrates); family 10 (BADH2). The values in parenthesis are the accession numbers, while the values after "#" denote the number of redundant genes (70 % global identity over 60% of the protein) belonging to that sequence in the genome(s). The bar denotes 10% amino acid substitutions.
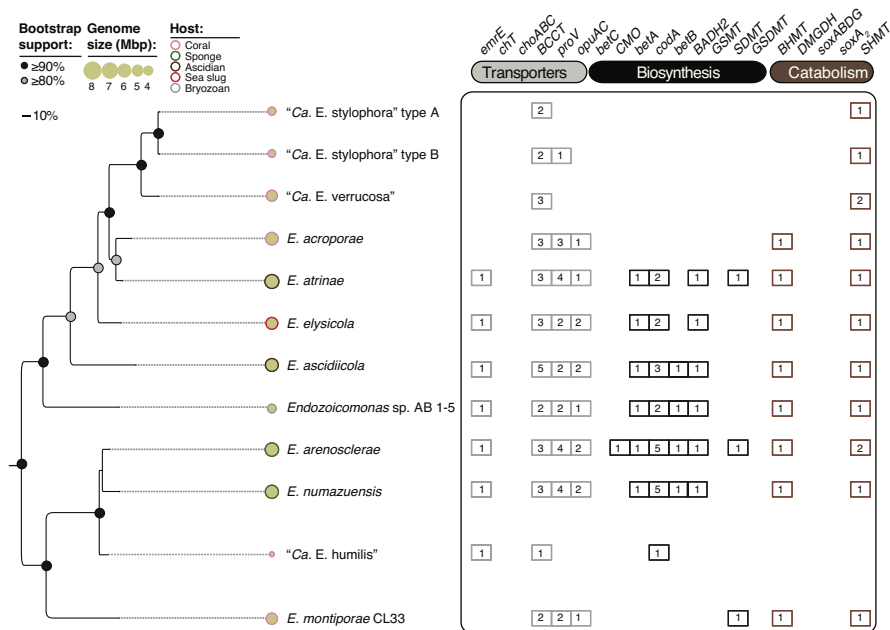
Figure S3. Related to Figure 1. Incidence of metabolic pathways for GB in twelve *Endozoicomonas* species associated with diverse marine invertebrates. The phylogenetic tree shows the relationship based on 20 bacterial single-copy genes inferred using the maximum-likelihood approach. The tree was rooted using *Pseudomonas aeroginosa* PA01 (Accession number NC_002516).
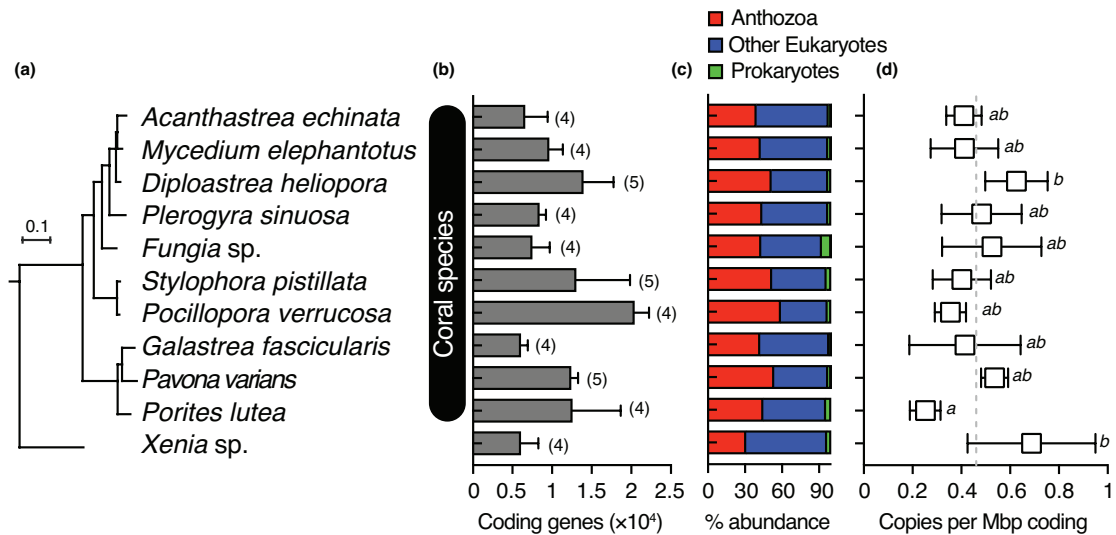
Figure S4. Related to Figure 2. Density of predicted *BCCT* homologs in diverse coral species based on metagenomics sequence data from the Red Sea. The maximum-likelihood tree on the left panel (a) is based on cytochrome oxidase (Cox2) gene present in representative assemblies of each species. (b) Bars show the average (± SD) counts of coding genes from metagenomes of each species, with values shown in brackets indicating the number of replicates per species. (c) Depicts the overall taxonomic breakdown of the coding genes present in replicates of each species (n = 4–5). (d) Shows the deduced copy number (per mega base pairs) of *BCCT* homologs present in the different coral species. Statistically significantly different average copy numbers per mega base pairs (Mbp) of coding genome are denoted by different letters as deduced by one-way ANOVA tests ($P < 0.001$) with two-stage step-up Benjamin-Krieger-Yekutieli's multiple comparison tests. The dotted grey line shows the overall average copy number per Mbp across all metagenomes.
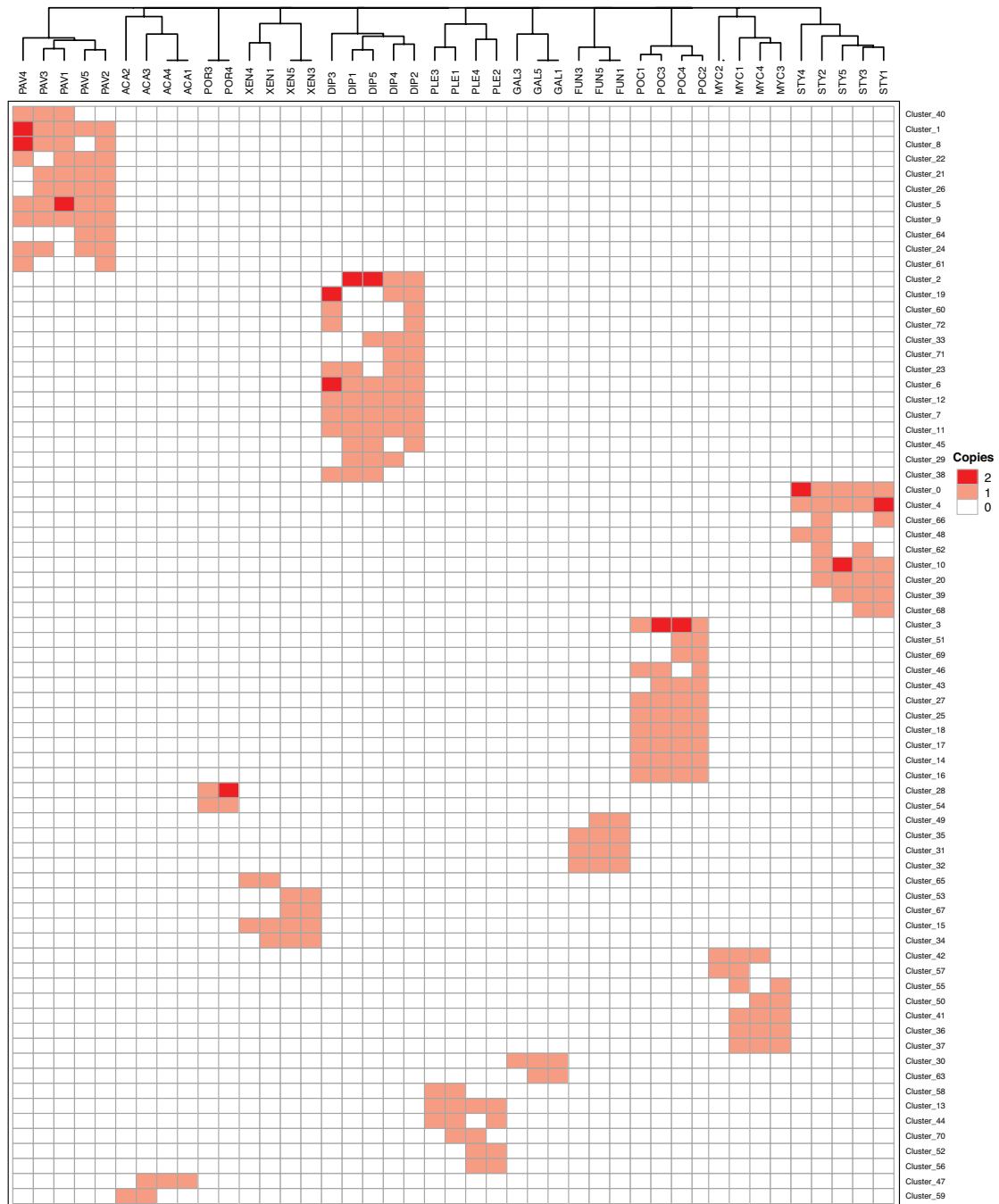
Figure S5. Related to Figure 2. Sequence diversity of *BCCT* genes predicted in Red Sea corals. The heatmap shows grouping of coral species (*n* = 4–5 replicates) by *BCCT* gene clusters. Similarity is defined at 95% global nucleotide identity over 80% of the shorter sequence length. Species and clusters are grouped based on a binary distance matrix and the average-linkage clustering approach. Clusters exclude singletons and sequences homologous to Symbiodiniaceae *BCCT*s.
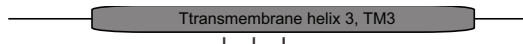
Ttransmembrane helix 3, TM3

140   150   160

Species (accession)	Sequence
Tetragenococcus halophilus (AAP06751, **ButA**)	FSRWGWF AML L SA GMG I GL VFWS VA EP I SHL
Bacillus subtilis (AAC44368, **OpuD**)	FGL L SW FAML FS AGMG IG LV FY GA AE PI SHY
Escherichia coli (AAQ10261, **BetU**)	F SY HSWF AML FS AGMG IG LMF FG VA EP VMHY
Sinorhizobium meliloti (AAL37253, **BetS**)	FR YL SW I AML FA AGMG IG LMY FA VG EP MT HF
Listeria monocytogenes (AAD30266, **BetL**)	Y S NK SW FAML FS AGMG IG LV FWGA A EP L SHY
Corynebacterium glutamicum (CAA63771, **BetP**)	FR TV SW I SMM FA AGMG IG LMF YGTT EP LT FY
Escherichia coli (CAQ30790, **BetT**)	F SL L SW AAML FA AG IG I DLMF F SV AE PV TQ Y
Escherichia coli (CAQ30560, **CaiT**)	FS TA SW I FMM FA SC TSA AVL FWGS IE - I YYY
Proteus mirabilis (WP_012368453, **CaiT**)	FS TA SW I FMM FA SC TSA AVL FWGS IE - I YYY
Acropora digitifera (XP_015778301, #3)	FND L TY FT ML FA AG IG IGL FY FG VA EP IWHY
Orbicella faveolata (XP_020624668, #7)	FND LT YF TML FA AG IG IGL FY FG VA EP VWHY
Orbicella faveolata (XP_020624717, #8)	FND LT YF TML FA AG IG IGL FY FG VA EP VWHY
Amplexidiscus fenestrafer (scaffold_164_3, #2)	FND AT YF TML FA AG IG IGL FY FG VA EP IWHY
Orbicella faveolata (XP_020604716, #1)	F SD AT YF TML FA AG IG VGL FY FG VA EP VWHY
Orbicella faveolata (XP_020620548, #5)	FND AT YF TML FA AG IG IGL FY FG VA EP IWHY
Stylophora pistillata (Spis14346, #1)	FKD AT YF TML FA AG IG IGL FY FG VA EP VWHY
Stylophora pistillata (Spis9846, #5)	FND AT Y FT ML FA AG IG IGL FY YG VA EP IWHY
Orbicella faveolata (XP_020608607, #3)	FKD AS YF TML FA AG IG IGL FY FG VA EP VWHY
Discomonas sp. (scaffold_224_10, #3)	FND AT YF TML FA AG IG IGL FY FG VA EP IWHY
Stylophora pistillata (Spis3098, #3)	F SD AT YF TML FA AG IG VGL FF FG VA EP VWHY
Amplexidiscus fenestrafer (scaffold_11_64, #1)	FND AT YF TML FA AG IG IGL FY FG VA EP IWHY
Orbicella faveolata (XP_020620549, #6)	FND AT YF TML FA AG IG IGL FY FG VA EP IWHY
Discomonas sp. (scaffold_64_15, #4)	F SD AT YF TML FS AG IG IGL FY FG VA EP IWHY
Amplexidiscus fenestrafer (scaffold_9_57, #4)	FND VS YF TML FA AG IG IGL FY FG VA EP IRHY
Orbicella faveolata (XP_020629092, #9)	FND AT YF TML FA AG IG IGL FY FG VG KP VY PT
Discomonas sp. (scaffold_16_115, #2)	FND VS YF TML FA AG IG IGL FY FG VA EP IRHY
Discomonas sp. (scaffold_64_16, #5)	FND AT YF TML FA AG IG IGL FY FG VA EP IF HY
Stylophora pistillata (Spis6993, #4)	F SD AS YF TML FA AG IG IGL FF FG VA EP IY HY
Amplexidiscus fenestrafer (scaffold_40_52, #3)	F SD AT YF TML FA AG IG IGL FY FG VA EP IWHY
Acropora digitifera (XP_015762628, #1)	FND VT YF TML FA AG IG IGL FY FG VA EP IY HY
Aiptasia pallida (AIPGENE2222, #2)	F SD AT YF TML FA AG IG VGL FY FG VA EP IXHY
Hydra magnipapillata (XP_012562913, #3)	F SD AS YF TML FSA GIG VGL FY FG VA EP VY HY
Hydra vulgaris (XP_012562913, #3)	F SD AS YF TML FSA GIG VGL FY FG VA EP VY HY
Hydra vulgaris (XP_002159349, #1)	Y SN SN YF MML FSA GV GIG L FY YG VA EP IY HY
Hydra vulgaris (XP_012556512, #2)	Y SD SA YF MML FSA GLG IGL FY YG VA EP IY HY
Nematostella vectensis (XP_001640076, #1)	FND GS ML FA AG IG VGL FY FG VA EP VL HY
Nematostella vectensis (XP_001640384, #2)	F SD G SY F SML FA AG IG VGL FY FG IA EP VQ HY
Nematostella vectensis (XP_001640389, #4)	Y SD AT YF TML FA AG IG VGL FY FG VA EP IY HY
Nematostella vectensis (XP_001640390, #5)	Y SD AT YF TML FA AG IG VGL FY FG VA EP IY HY
Aiptasia pallida (AIPGENE15022, #1)	F SD VT YF TML FA AG VG VGL FY FG VA EP IQ HY
Aiptasia pallida (AIPGENE2229, #3)	F SD AT YF TML FA AG IG VGL FY YG VA EP IY HY
Aiptasia pallida (AIPGENE3680, #4)	F SD TS YF TML FA AG VG VGL FY YG VA EP VY HY
Mnemiopsis leidyi (ML00553a, #1)	FND GT WF AML FAC GV GVG L FF FG VA EP LWHY
Mnemiopsis leidyi (ML071166a, #3)	F SD A EW F SML FAC GIG IGL FF FSV SE PL RHY
Mnemiopsis leidyi (ML131717a, #4)	Y PD GS WF CML FAC GV GVG L FY FG VA EP INHY
Mnemiopsis leidyi (ML296214a, #5)	YND AT WF AML YAC GIG VGL FF FY GV AE PMWHY
Vitrella brassicaeformis (CEM00714, #1)	F SLG SW FAML FAC GV GIG L FY FG VA EP IWHY
Vitrella brassicaeformis (CEM19281, #2)	FN SI SW FTML FCC GV GPG L FY FG VA EP IL HY
Vitrella brassicaeformis (CEM17339, #3)	FS TV SW F SML FCC GV GVG M FY FG VA EP IF HY
Thalassiosira oceanica (EJK50030)	F SDMTY FAML FSA GV GVG L FF FG VS EP LF HL
Fragilariopsis cylindrus (OEU16937)	F SN SS YF SML FA AG IG VGL FF YG VS EP LF HR
Symbiodinium minutum (Sminut025002, #2)	YGD LS YL AML WCS GV GIG L IF YGA SE PL MHA
Ostreococcus tauri (CEF96976, #1)	Y SD LV WF IL IF TT GL GT GI FY FG VS EP MY YY
Ostreococcus tauri (CEG01428, #2)	F SD YE WF SML FAC GIG VGL YT YG VA EP MWYY
Ostreococcus lucimarinus (XP_001416550, #1)	Y SD LV WF TL IF TT GL GT GI FY FG VS EP MY YY
Ostreococcus lucimarinus (XP_001422299, #2)	FD YY EW F SMM FAC GIG VGL YT F SV ME PI SY Y
Bathycoccus prasinos (XP_007508379)	F SD FA WF SML F SC GIG VG FY YGV SE PI Y HY
Stylophora pistillata (Spis15025, #2)	F SD LT Y FT ML FA AG VA IGL FY YG VA EP IWHY
Lingulodinium polyedrum (TRT_GABP01088257, #2)	FS TF AW FAML FT CG VA TG L FF FA VT EP QY YY
Lingulodinium polyedrum (TRT_JO757247, #1)	FS TF AW FAML FT CG VA TG L FF FA VT EP QY YY
Guillardia theta (XP_005827593, #1)	FS DF SW FV ML FC SG IA VG FY FY GV RE PL SY Y
Guillardia theta (XP_005835951, #2)	FS DF SW FV ML FC SG IA VG FY FY GV RE PL TY Y
Phaeodactylum tricornutum (XP_002182771)	FS TG AY FAML FSC GV GL FV FG VA EP LWHQ
Aureococcus anophagefferens (XP_009039779)	F SR VQ WF AML YCC GV AV GL FY YG VA EP MWHF
Micromonas commoda (XP_002502671, #1)	FNN IS WF SML F SC GIA VG VY T LG VA EP MY YY
Micromonas commoda (XP_002502672, #2)	- - - - - - - - - - MLF SC GIA VG VY T LG VA EP MY YY
Symbiodinium minutum (Sminut010127, #1)	YGD LT Y LAMV WC AG VA IG L IF YGA SE PL FHA
Symbiodium kawaguti (Skav218262, #2)	- - - - - - - - - - - - VA IG L IF FY GA SE PL LHA
Symbiodium kawaguti (Skav202732, #1)	- - - - - - - - - - - - AG VA IG L IF YGA SE PL FHA
Symbiodinium microadriticum (Smic154742, #2)	YGD LT Y L AMV WC AG VA IG L IF YG VS EP LT HS
Symbiodinium microadriticum (Smic15475, #1)	YGD LT Y L AMV WC AG VA IG L IF YG VS EP LT HS

Conservation

2 1 0 0 0 3 1 0 + 8 9 7 7 + 8 + 6 + 7 6 6 9 8 8 7 + * 8 2 8 6

Consensus

FS DA T Y FT ML FA AG IG IGL FY FG VA EP I+ HY

Figure S6. Related to Figure 2. Multiple sequence alignments of validated bacterial BCCT carriers and predicted eukaryotic homologs; the third transmembrane helix (TM3) region of the carriers is highlighted. Note the significant homology of predicted eukaryotic BCCT carriers in the region shown by arrows containing the conserved G-x-G-x-G motif present in biochemically defined GB carriers (e.g., BetP from *C. glutamicum*). Amino acid positions are based on BetP.

```
                                                                    TM4                                TM8
                                                               190        195             371              381
 1. Corynebacterium glutamicum (CAA63771, BetP)    H W T L H P W A I Y A     - - W A W W I S W S P F V
 2. Tetragenococcus halophilus (AAP06751, ButA)    H W G I H P W A I Y A     - - W A W W I S W S P F V
 3. Bacillus subtilis (AAC44368, OpuD)             H W G L H A W A I Y A     - - W A W W I S W S P F V
 4. Escherichia coli (AAQ10261, BetU)              H W G L H A W A I Y A     - - W G W W L S W S P F V
 5. Sinorhizobium meliloti (AAL37253, BetS)        H W G V H A W A I Y S     - - W A W W I S W S P F V
 6. Listeria monocytogenes (AAD30266, BetL)        H W G I   A W S I Y A     - - W A W W L S W S P F V
 7. Escherichia coli (CAQ30790, BetT)              H Y G L T G W S M Y A     - - W A W W V A W S P F V
 8. Escherichia coli (CAQ30560, CaiT)              H W G P L P W A T Y S     - - W A W W V I Y A I Q M
 9. Proteus mirabilis (WP_012368453, CaiT)         H W G P L P W A T Y S     - - W A W W V I Y A I Q M
10. Acropora digitifera (XP_015778301)             H W G I H G W I V Y V     - - W G W W I A W S P F V
11. Stylophora pistillata (Spis14346)              H W G I H G W I V Y V     - - W G W W I A W S P F V
12. Orbicella faveolata (XP_020624668)             H W G V H G W I V Y V     - - W G W W I A W S P F V
13. Discomonas sp. (scaffold_224_10)               H W G I H G W I V Y V     - - W G W W I A W S P F V
14. Amplexidiscus fenestrafer (scaffold_164_3)     H W G V H G W I V Y V     - - W G W W I A W S P F V
15. Hydra vulgaris (XP_012562913)                  H W G I H G W I V Y V     - - W G W W V S W S P F V
16. Nematostella vectensis (XP_001640076)          H W G I H G W I V Y T     - - W G W W I A W A P F V
17. Aiptasia pallida (AIPGENE15022)                H W G I H A W V V Y T     - - W G W W I A W S P F V
18. Thalassiosira oceanica (EJK50030)              H W G I A A W S P Y L     R C S T W R G G S A G L V
19. Fragilariopsis cylindrus (OEU16937)            H W G L A A W S P Y L     - - L A W W V A W A C F V
20. Ostreococcus tauri (CEF96976)                  H W G F H G W A P Y L     - - W A W W I T W A P F V
21. Bathycoccus prasinos (XP_007508379)            H W G L H G W I P Y I     - - W G W W I S W A P F V
22. Mnemiopsis leidyi (ML00553a)                   H W G I H G W I V Y V     - - W G W W I S W S P F V
23. Phaedactylum tricornutum (XP_002182771)        N W G I   G W A P Y L     - - Q A W W V S W S A F V
24. Aureococcus anophagefferens (XP_009039779)     H W G F H G W V P Y V     - - W G W W I S W A P F V
25. Symbiodinium minutum (Sminut010127)            H W G F M A W I V Y C     - - W G W W I A W A P F V
26. Symbiodinium kawaguti (Skav202732)             H W G F M A W I V Y S     - - W G W W I A W A P F V
27. Symbiodinium microadriaticum (Smic15475)       H W G F L A W I V Y A     - - W G W W I A W A P F V
28. Linulodinium polyedrum (TRT_JO757247)          H W G L H G W V V Y C     - - W G W W I A W A P F V
29. Guillardia theta (XP_005827593)                H W G I H A W A C Y I     - - W G W W I S W A P F V
30. Micromonas commoda (XP_002502671)              H W G L H A W G P Y I     - - W G W W V S W A P F V
```

Figure S7. Related to Figure 2. Multiple sequence alignments of validated bacterial BCCT carriers and predicted eukaryotic *BCCT* homologs illustrating residues involved in GB binding (orange and green boxes) based on (Ressl et al., 2009).
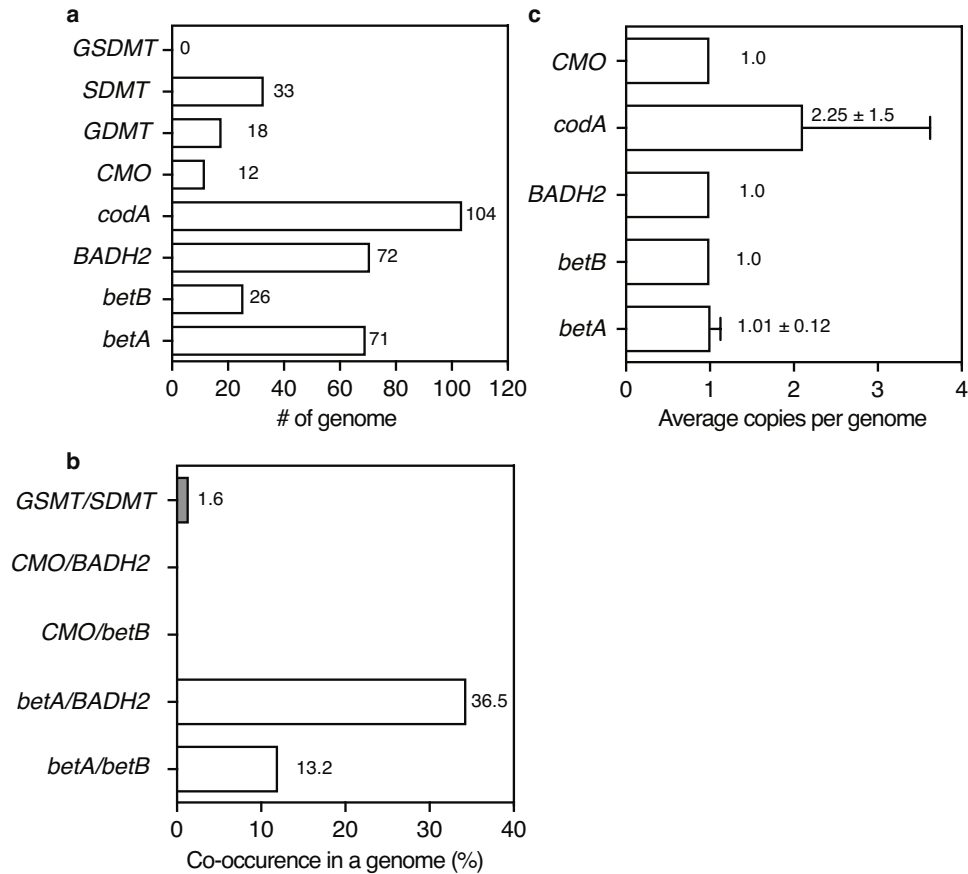
Figure S8. Related to Figure 2. The density and co-occurrence of synthetic routes for GB in completely sequenced marine prokaryotic genomes (*n* = 189). (a) Shows the overall number genomes possessing the retrieved gene homolog. (b) Co-occurrence incidence of genes involved choline oxidation pathways (in white) and GB formation via glycine methylation (in grey). (c) Indicates the average (± SD) copy number of predicted homologs per genome. Details of the genomes are provided in **Table S8**. Values beside each bar graph denote number of genomes (in a) or the proportion of genomes with the pathway (in b) out of 189 genomes; in (c) values indicate the average (± SD) copies across the genomes.

Transparent Methods

Genomic and transcriptomic sequence datasets

Information on GB metabolism is largely restricted to prokaryotes and plants. Hence, our genomic and transcriptomic survey strategy was tailored at capturing sequenced model and non-model organisms, including evolutionary unrelated higher organisms, such as those thriving in terrestrial habitats and marine-freshwater transition zones (e.g., mangrove plants and animals). Eukaryotic genomes (*n* = 114) and transcriptomes (*n* = 16) representing different metazoan lineages were retrieved from the NCBI/GenBank database or several dedicated databases (downloaded on March 2016; **Table S1**). Cnidarian and endosymbiotic Symbiodiniaceae sequences were retrieved from the reefgenomics database [http://reefgenomics.org/)(Liew et al., 2016)]. Additional reference Symbiodiniaceae transcriptomes (*n* = 8) from the Marine Microbial Eukaryote Transcriptome Sequencing Project [MMETSP; (Keeling et al., 2014)] were retrieved from the iMicrobe data server (http://datacommons.cyverse.org/) under project #104. Transcriptomic data for other dinoflagellates (*Karenia brevis, Lingulodinium polyedrum*, and *Prorocentrum minimum*) and other marine invertebrates of interest lacking genomes such as sponges (*Ephydatia muelleri*, *Haliclona* species, *Stylissa carteri*, and *Xestospongia testudinaria*) were retrieved from the comparative genomic platform for early branching metazoans (http://www.compagen.org/) as well as the iMicrobe data server. Three recently published mangrove plant genomes (*Avicennia marina*, *Sonneratia alba*, and *Sonneratia caseoralis*) were also included to cover plants in the marine realm [http://evolution.sysu.edu.cn/Sequences/)(Xu et al., 2017)]. Additional genomes such as fungus-cultivating termite (*Macrotermes natalensis*), the Japanese abalone (*Haliotis discus*), the estuarine crocodile (*Crocodylus porosus*), and the Chinese mitten crab (*Eriocheir sinensis*) were retrieved from the GigaScience Database (http://gigadb.org/site/index).

We also retrieved 189 completely sequenced reference prokaryotic genomes (**Table S6**) from the marine microbial reference (MarRef) genome database (Klemetsen et al., 2018). These genomes were filtered from a total of 839 genomes (Accessed on June 07th 2019) by selecting the analysis project type "whole genome sequencing" and environmental material "sea water" in the on-line server and quality assessment with CheckM v1.0.3 (Parks et al., 2015), retaining only "high-quality" genomes (≥90% completeness and <5% contamination) fulfilling the MIMAG standards for single-cell and metagenome-assembled genomes (Bowers et al., 2017).

Endozoicomonas genomes, representing a bacterial lineage prevalent in diverse marine invertebrates (Neave et al., 2017a), were retrieved from GenBank (downloaded February 2018). These included the following accession numbers: *E. acroporae* Acr-14 (GCF_002864045), *E. atrinae* DSM 22380 (GCF_001647025), *E. elysicola* DSM 22380 (GCF_000710775), *E. ascidiicola* AVMART05 (GCF_001646945), *E. arenosclerae* E-MC227 (GCA_001562005), *E. namuzuensis* DSM 25634 (GCF_000722635), *E. montiporae* CL-33 (GCF_001583435), and *Endozoicomonas* sp. AB 1-5 (GCA_001729985). Additionally, two single-cell genomes from *Stylophora pistillata* ("Ca. E. stylophora" type A and B) and

another two metagenome-resolved population genomes from *Acropora humilis* ("*Ca.* Endozoicomonas humilis") and *Pocillopora verrucosa* ("*Ca.* Endozoicomonas verrucosa") (Neave et al., 2017b) were retrieved from the RAST server with the following RAST IDs 6666666.127878, 6666666.127879, 305899.6, and 305899.13 respectively.

Another additional set of 52 metagenome-resolved prokaryotic genomes (MAGs), encompassing representative taxa of 12 phyla predominating the microbial microbiome of the Indo-Pacific reef-building coral *Porites lutea* (Robbins et al., 2019) were also retrieved from http://plut.reefgenomics.org/microbiome_download/ (**Table S5**).

Prediction of GB metabolism in eukaryotic and prokaryotic genomes

Profile Hidden Markov Models (profile-HMMs) of protein family alignments were used to query enzymes involved in the transport, biosynthesis, and catabolism of GB and other osmolytes in genomes and transcriptomes. A total of 52 profile-HMMs were retrieved from PFAM (Bateman et al., 2000) and TIGRFAM (Haft et al., 2003) databases (**Figure 1**; **Table S2** and **S3**). Protein-coding genes and full-length transcriptomic sequences were queried using the program hmmsearch within HMMER v3.2.1 (hmmer.org), applying default settings and an *E*-value threshold of 1×10^{-6} as well as the pre-calculated "gathering" and "noise" cutoffs defined for the 52 model HMMs to search for corresponding protein homologs in the genomes/transcriptomes.

However, eight enzymes lacked profiled HMMs, including those encoding choline oxidase (CodA), choline monooxygenase (CMO), glycine-sarcosine methyltranferase (GSMT), sarcosine-dimethylglycine methyltransferase (SDMT), glycine sarcosine dimethylglycine N-methyltransferase (GSDMT), the GB/choline exporter (EmrE), the eukaryotic sarcosine dehydrogenase (SDH), and betaine aldehyde dehydrogenase (BADH2). For these enzymes, we performed blast searches employing customized protein databases of experimentally characterized proteins available in Swissprot/Uniprot based on the following accession numbers: CodA (*Arthrobacter globiformis*, AAP68832 and Q7X2H8; *Aspergillus fumigata*, B0YA55); CMO (*Spinacia oleracea*, AAB52509; *Atriplex hortensis*, Q9LKN0; *Amaranthus tricolor*, Q93XE1; *Arabidopsis thaliana*, Q9SZR0; *Beta vulgaris*, BAE07177); GSMT (*Methanohalophilus portucalensis*, 5GWX_A; *Aphanothece halophytica*, Q83WC4; *Synechococcus* sp. WH8102, Q7U4Z8; *Halorhodospira halochloris*, Q9KJ22); SDMT (*H. halochloris*, Q9KJ21; *A. halophytica*, Q83WC3); GSDMT (*Actinopolyspora halophila*, Q9KJ20); EmrE (*Escherichia coli*, P23895); SDH (*Mus musculus*, Q99LB7; *Homo sapiens*, Q9UL12; *Rattus norvegicus*, Q64380); and BALDH2 (*A. thaliana*, Q9STS1 and Q9S795; *Oryza sativa*, Q84LK3; *Gadus callarias*, P56533; *Pseudomonas aeuroginosa*, Q9HTJ1; *Burkholderia mallei*, Q62CH7; *Amaranthus hypochondriacus*, AAB700100; *Gibberella zeae*, EAA74487; *Sinorhizobium meliloti*, P54222; *Bacillus subtilis*, P71016; *Homo sapiens*, P49189). Additional proteins sequences with 90% homology over 80% of the gene length from the UniProt's (UniRef90) reference database (https://www.uniprot.org) were consolidated into the database of GSDMT (for proteins with 474–564 amino acid sequence length; *n* = 140), EmrE (*n* = 100), and BADH2 (*n* = 22). All customized databases are available from the corresponding author. Protein-coding gene and

transcriptomic sequences were subsequently queried against these customized protein databases using diamond v0.9.24 (Buchfink et al., 2015) with the "sensitive" search mode and an *E*-value threshold of 10^{-10}. A minimum amino acid identity of 30% (*CMO*, *codA*, *GSMT*, and *SDMT*) or 50% (*GSMDT*, *emrE*, *SDH*, and *BADH2*) and alignment coverage of 60% to the reference sequence was employed for filtering of top hits. The 30% identity cut off was selected because pairwise alignment of ratified enzymes directly in UniProt showed that some of the genes (*CMO*, *codA*, *GSMT*, and *SDMT*) had pairwise identities as low as 30–40% with alignment coverage minimally around 60%. The 50% cut off was useful in the case of enzymes in which different subfamilies fell into substrate-specific lineages such as BADH2 (s.f. **Figure S2**) to avoid false positives. We however, contend that the cutoffs my exclude distantly related enzymes catalyzing the target reactions but with different kinetics.

Taxonomic assignment of the retrieved homologs was done based on their blastp matches against the non-redundant NCBI database (downloaded on June 2016) as described by (Ngugi et al., 2017), applying an *E*-value threshold of 1×10^{-6} and a minimum blast score value of 40. Because some eukaryotic genomes likely have microbial contaminants (Artamonova et al., 2015), we employed the lowest-common-ancestor (LCA) approach to flag hits matching prokaryotic genomes (**Table S2** and **S3**). If the taxonomic lineage of the retrieved gene homolog was not consistently supported by at least five blast matches to eukaryotic lineages, then that retrieved gene was completely ignored for downstream analyses. However, only few eukaryotic genomes presented this anomaly; for instance, some genomes (e.g., *Avicennia marina and Capitella teleta*) and transcriptomes (e.g., *Xestospongia testudinaria*) were found to have BCCT carriers that were 100% identical to their expected symbionts, indicative of microbial contamination.

Subcellular localization of retrieved protein homologs was predicted using the TargetP server (http://www.cbs.dtu.dk/ser vices/TargetP/) (Emanuelsson et al., 2007) with default "non-plant" settings allowing also for cleavage site prediction.

Phylogenetic analyses

All phylogenetic analyses were conducted in Geneious Pro v8.1.9 (https://www.geneious.com) and the best amino acid sequence substitution model for tree inference determined with ProTest3 (Darriba et al., 2011). The deduced eukaryotic BCCT carrier proteins (*n* = 72; 749 ± 185 amino acid residues) together with validated prokaryotic BCCT proteins (*n* = 22; 557 ± 69 amino acid residues) were aligned using MUSCLE (Edgar, 2004) with default settings. The alignment was reduced to 92 amino acid sequences by removing shorter proteins and then filtered using Gblocks (Castresana, 2000) to eliminate poorly aligned positions, reducing it to 438 characters based on the liberal settings that retain the maximum number of characters in the alignment. An un-rooted maximum-likelihood phylogenetic tree was reconstructed using RAxML v7.2.830 (Stamatakis, 2014), as implemented in Geneious Pro under the best-fit general time reversible (GTR) + GAMMA model (log-likelihood ratio of –33909.16), with 1,000 bootstraps and a maximum-likelihood search of the best tree topology.

Considering that aldehyde dehydrogenases encompass enzymes with variable substrates (Fitzgerald et al., 2009; Kirch et al., 2004), we were prompted to employ an iterative process to robustly identity genes that were phylogenetically affiliated with *bona fide* betaine aldehyde dehydrogenases (BADH2) among the 100s of matches (in total 1,985) obtained via blast search to our customized blast data (see above). The predicted matches from the genes were clustered into a non-redundant set using cd-hit (Li and Godzik, 2006) by applying global protein identity cutoff of 70% and alignment coverage of 60% for the shorter sequences. Next, the resultant non-redundant proteins (n = 927) were aligned together with 22 validated BADH2 amino acid sequences based on the work of Kirch et al. (Kirch et al., 2004) using MUSCLE. The aligned sequences were then used as the basis for a first-run phylogenetic tree with FastTree v.2.1.5 (Price et al., 2010) under the best-fit WAG + GAMMA model (log-likelihood ratio of – 34818.25) and using the approximately maximum likelihood method in the fastest mode. This enabled to prune sequences that were distantly affiliated to the validated plant BADH2 sequences, yielding a total of 87 putative BADH genes from the studied eukaryotic genes. To construct the final tree, we gathered 28 additional proteins from NCBI encompassing validated plant aldehyde dehydrogenases with variable substrates and biochemistry (Kirch et al., 2004). The final set of 127 proteins were aligned with MUSCLE resulting in an alignment with 3, 727 columns, which were further reduced to 492 column by removing poorly aligned sequence position using Gblocks as stated above. A maximum-likelihood tree was then constructed using RaxML (1000 bootstraps), under the model LG + GAMMA (log-likelihood ratio of –6582.72) selected using Protest3 to accommodate the additional sequences. The reconstructed tree was then manually populated adding the number of redundant sequences in each representative BADH-like gene including the names of organisms harboring that homolog based on the cd-hit information using Adobe Illustrator (available from http://adobe.com/products/illustrator). For the final presence/absence profile only genomes with a gene phylogenetically related to family 10 ALDHs (**Figure S2**) were scored as having a putative BADH2-like gene (**Figure 1b**) in addition to genes matching characterized ALDH family 10 proteins (via blast search).

Coral sampling

Samples of 10 scleractinian coral species (*Acanthastrea echinata*, *Diploastrea heliopora*, *Fungia* sp., *Galaxea fascicularis*, *Mycedium elephantotus*, *Pavona varians*, *Plerogyra sinuosa*, *Pocillopora verrucosa*, *Porites lutea*, *Stylophora pistillata*) and one octocoral (*Xenia* sp.) were collected at the ocean-facing side of Al Fahal reef (22.3034, 38.9602) in the central Red Sea, Saudi Arabia (n = 4 to 5 replicates per species; **Table S7**). The sampled species fall across many coral clades and life histories, thus attractive for evaluating metabolism of GB in diverse corals. Five replicate colonies per species were sampled on May 17 and 18 (2016) at depths of 3 to 18 m. One small piece (3–4 cm) was collected per colony using hammer and chisel, transferred into Whirlpack bags, and immediately flash-frozen upon return to the boat.

DNA extraction and whole-genome sequencing

DNA was extracted using a DNA All Prep kit (#80204) from Qiagen with modifications. Briefly, corals in Whirlpack bags were first defrosted and a

pressurized air gun was used to remove host tissue and putative Symbiodiniaceae cells away from the skeleton and into the Whirl Pak. Subsequently, 30 mg of air blasted tissue slurry was weighed out in a sterile weigh boat, where 600 μl of buffer RLT was added to the slurry. The mixture was homogenized with a 20-gauge needle on a 1-ml syringe, and the buffer and tissue mixture were placed in a new 2 ml tube. DNA was then extracted as recommended by the manufacturer. The extracted DNA was then quantified on a Qubit prior to library preparation.

100 ng of DNA per sample was sheared to approximately 250–300 bp (Covaris M2). Sheared DNA was end-repaired and A-tailed with a single A dNTP, the NEB adapter (NEB #E7645L) was then ligated to the end-repaired A-tailed DNA fragment. Size selection of the library was performed using Agencourt AMPure XP beads (Beckman Coulter) for size selection of 250 bp fragment inserts. Dual indexes were used in the library enrichment step to add Illumina TruSeq HT indexes (NEB #7600) to the DNA molecules using 6 cycles of PCR. Libraries were quality checked using the Bioanalyzer DNA 1000 Chip (Agilent Technologies), quantified using the Qubit BR DNA system (Invitrogen), and subsequently pooled in equimolar ratios to a final concentration of 10 nM. The pool was re-quantified using qPCR (KAPA Biosystems library quantification) with the ABI HT7900 (Applied Biosystems) and paired-end sequenced (2×300 bp) on the Illumina NextSeq 500 at 1.8 pM with 1 % PhiX.

Metagenomic sequence analyses

Raw read sequences were quality filtered and trimmed using Trimmomatic v0.32 (Bolger et al., 2014) to remove adapter sequences and leading and trailing bases with a quality score below 20, less than 50 base-pairs (bp) long as well as reads with an average per base quality of 20 over a 4-bp window. Paired-end (PE) reads were mapped onto run quality control sequences (PhiX genome) using BBmap v37.44 (http://jgi.doe.gov/data-and-tools/bbtools/). At each stage, the quality of read sequences was assessed using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

The resulting high-quality PE reads for each dataset (n = 4–5 per coral species) were *de novo* assembled independently with metaSPAdes v3.9.0 (Bankevich et al., 2012) using a kmer range of 21 to 127, while employing the error-correction mode and preset metagenomic options. The assembled contigs were then filtered to a minimum length of 500 bp followed by gene prediction using Prodigal v2.6.2 (Hyatt et al., 2010). Prodigal was run with the options "-p meta -n -m". **Table S7** summarizes the general statistics of read sequences, the assembled contigs, and predicted genes.

Individual metagenomic protein-coding genes were interrogated as described above but focusing on the BCCT carrier family (PF02028) regarding potential high-affinity GB transporters in different coral species from the Red Sea. Here as well, the resultant matches were searched against the nr protein database to clarify the origin of hits, and only those matching to anthozoans were retained for downstream analyses.

Gene density was computed from the counts of retrieved hits per metagenome scaled by the total length of coding sequences (in megabase-pairs) among the

different coral species. All predicted coding genes in each metagenome and the recovered transporters were taxonomically assigned as described above.

Gene expression analyses using available metatranscriptomes

Using ten recently published RNA-Seq data covering the larval–adult developmental cycle and two symbiotic states of the laboratory-raised sea anemone Aiptasia (Baumgarten et al., 2015), we tracked the transcriptional activity of GB metabolism in this model dinoflagellate-cnidarian symbiosis. Briefly, raw sequencing reads were downloaded from the short-reads archive under accession number SRX286028 and quality-trimmed, while removing adaptor sequences, using Trimmomatic v0.323 (Bolger et al., 2014) with the following parameters: ILLUMINACLIP::4:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:60. The internal sequencing standard PhiX 174 was subsequently removed by mapping the quality-trimmed reads against the PhiX 174 genome using Bowtie2 v2.2.45 (Langmead and Salzberg, 2012), with default settings. Potential errors in the reads were then corrected using SPAdes' implementation of the BayesHammer error-correction tool with default parameters (Bankevich et al., 2012). Ribosomal RNA-like reads were removed using SortMeRNA v1.99 (Kopylova et al., 2012). The resulting high-quality PE reads from each sample were mapped against the Aiptasia v1 gene models (http://reefgenomics.org/) using Bowtie2 (Langmead and Salzberg, 2012), with default settings. The mapped read counts per gene (and sample) were subsequently normalized into a common metric of gene expression—that is, fragments per kilobase of exon per million fragments mapped (FPKM) using an in-house perl script.

To evaluate whether genes involved in GB metabolisms are actively expressed in free-living Symbiodiniaceae, we interrogated a catalog of 116 million expressed eukaryotic unigenes of the *Tara* Ocean Expedition (Carradec et al., 2018) available at http://www.genoscope.cns.fr/tara/.

The available Pfam annotations of the unigene catalog were used to retrieve the expression values of selected four enzymes of GB metabolism, including PF02028 (BCCT), PF02574 (BHMT), PF16350 (DMGDH) and PF00464 (SHMT). In addition, local blast searches were performed as described above in the case of CodA, CMO and SDMT. In both cases, we retained genes taxonomically annotated as "Suessiaceae" at the lowest rank as well as translated blast matches with ≥80% identity and alignment length of ≥60% against reference genes from Symbiodiniaceae genomes and transcriptomes (this study). Expression values of homologs that showed significant similarity to Symbiodiniaceae genes were then plotted using the R package ggplot2 (Wickham, 2009).

Global estimates of GB stocks bound in marine picoplankton cells

To estimate the global stock of GB bound in marine picoplankton cells (0.2 to 3.0 μm in size), we interrogated metagenomic datasets from the *Tara* Ocean global circumnavigation expedition (Sunagawa et al., 2015) to estimate the fraction of marine prokaryotes capable of synthesizing GB using the abundance of *betA*, *betB*, *BADH2*, and *codA* homologs as marker genes for two different choline oxidation pathways (**Figure 1a**). Briefly, protein-coding genes from 201 (out of 243) *Tara* metagenomes covering photic to epipelagic depths (5–70 m) were retrieved from the European Nucleotide Archive

(http://www.ebi.ac.uk/ena/data/view/ERZ096909-ERZ097151). These metagenomes were probed based on profiled HMMs (BetA and BetB) and the customized BADH2 and CodA protein databases as described in the previous section. These four enzymes were selected because marine reference prokaryotic genomes possess a single copy of each gene homolog on average and also because they are widespread (**Figure S8**), making them ideal for estimating the abundance of microorganisms capable of producing GB in the ocean.

In parallel, we also retrieved homologs of four conserved single-copy genes (SCGs) in the same metagenomes—that is, *recA* (recombinase; TIGR02012), *ligA* (NAD-dependent DNA ligase; TIGR00575), *pyG* (CTP synthase; TIGR00337), *fmt* (methyl-tRNA formyltransferase; TIGR00460). Based on the retrieved counts of these SCGs, we then retained metagenomes with at least 15 RecA hits and removed those in which the coefficient of variation (CV) in the counts of all four SCGs was ≥ 25% (mean CV was 16 ± 5%). This yielded a final set of 108 metagenomes encompassing microorganisms with a size range of 0.22 μm to 3 μm (**Table S9**). Finally, the number of unique hits to *betA*, *betB*, *BADH2*, and *codA* genes were normalized to the average number of unique hits to the four SCGs in the metagenomes to estimate the environmental abundance of GB-producing picoplankton in the pelagic sea (**Table S9**). The *betA*, *betB*, *BADH2*, and *codA* homologs were estimated to be present in 2.4, 0.3, 1.2, and 34.1% (**Table S9**), respectively of the sampled picoplankton in the *Tara* metagenomes (*n* = 108).

To calculate the pool size of GB bound in marine pelagic picoplankton (0.2 to 3 μm in size), we first made four assumptions, namely that (1) a milliliter of seawater contains on average 10^5 pico-sized cells (Whitman et al., 1998); (2) the total biomass of prokaryotic picoplankton in the pelagic ocean is about 3.6×10^{28} cells (Whitman et al., 1998); (3) about 2% of marine picoplankton communities produce (and store) GB intracelullarly based on the environmental abundance of *betA* and *BADH2* genes encoding key synthetic routes for GB in marine metagenomes (**Table S9**)—equivalent to roughly 36×10^{26} cells in the upper 200 m globally; and (4) the concentrations of cellular-bound (particulate) GB quantified from 0.2-μm-filtered seawater particulates of about 0.9 to 483 nanomolar (Beale and Airs, 2016; Cree, 2015; Keller et al., 2004)—that is, GB bound in marine picoplankton cells.

Statistical Analyses

Significant differences between the average count of recovered genes among different groups and coral species was tested by conducting a one-way analysis of variance (one-way ANOVA) using GraphPad Prism v7.0a (GraphPad Software, Inc.). Multiple comparisons for controlling the false discovery rate (α = 0.05) was conducted with the Benjamini–Hochberg's two-stage step up method (Benjamini et al., 2006).

Data and materials availability

The Raw sequence data for the coral metagenomes have been deposited in NCBI under BioProject number PRJNA437202 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA437202). All data needed to evaluate the results and conclusions are presented in the main text and/or are available in the Supplementary Materials. Additional data or scripts related to

this paper are available from the corresponding authors.

Supplemental References

Artamonova, I.I., Lappi, T., Zudina, L., Mushegian, A.R., 2015. Prokaryotic genes in eukaryotic genome sequences: when to infer horizontal gene transfer and when to suspect an actual microbe. Environ. Microbiol. 17, 2203–2208. doi:10.1111/1462-2920.12854

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477. doi:10.1089/cmb.2012.0021

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., Sonnhammer, E.L., 2000. The Pfam protein families database. Nucleic Acids Res. 28, 263–266.

Baumgarten, S., Simakov, O., Esherick, L.Y., Liew, Y.J., Lehnert, E.M., Michell, C.T., Li, Y., Hambleton, E.A., Guse, A., Oates, M.E., Gough, J., Weis, V.M., Aranda, M., Pringle, J.R., Voolstra, C.R., 2015. The genome of *Aiptasia*, a sea anemone model for coral symbiosis. Proc. Natl. Acad. Sci. U.S.A. 112, 11893–11898. doi:10.1073/pnas.1513318112

Beale, R., Airs, R., 2016. Quantification of glycine betaine, choline and trimethylamine N-oxide in seawater particulates: Minimisation of seawater associated ion suppression. Anal. Chim. Acta 938, 114–122. doi:10.1016/j.aca.2016.07.016

Benjamini, Y., Krieger, A.M., Yekutieli, D., 2006. Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93, 491–507. doi:10.1093/biomet/93.3.491

Bolger, A.M., Bolger, A.M., Lohse, M., Lohse, M., Usadel, B., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., Tringe, S.G., Ivanova, N.N., Copeland, A., Clum, A., Becraft, E.D., Malmstrom, R.R., Birren, B., Podar, M., Bork, P., Weinstock, G.M., Garrity, G.M., Dodsworth, J.A., Yooseph, S., Sutton, G., Glöckner, F.O., Gilbert, J.A., Nelson, W.C., Hallam, S.J., Jungbluth, S.P., Ettema, T.J.G., Tighe, S., Konstantinidis, K.T., Liu, W.-T., Baker, B.J., Rattei, T., Eisen, J.A., Hedlund, B., McMahon, K.D., Fierer, N., Cochrane, G., Karsch-Mizrachi, I., Tyson, G.W., Rinke, C., Meyer, F., Knight, R., Finn, R., Lapidus, A., Yilmaz, P., Parks, D.H., Eren, A.M., Schriml, L., Banfield, J.F., Hugenholtz, P., Woyke, T., 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. 35, 725–731. doi:10.1126/science.1150427

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Meth. 12, 59–60. doi:10.1038/nmeth.3176

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D.J., Yoshikawa, G., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Acinas, S.G., Boss, E., Follows, M., Gorsky, G., Grimsley, N., Karp-Boss, L., Krzic, U., Pesant, S., Reynaud, E.G., Sardet, C., Sieracki, M., Speich, S., Stemmann, L., Velayoudon, D., Weissenbach, J., Jaillon, O., Aury, J.-M., Karsenti, E., Sullivan, M.B., Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., Vargas, C., Iudicone, D., Bowler, C., Wincker, P., 2018. A global ocean atlas of eukaryotic genes. Nat. Comm. 1–13. doi:10.1038/s41467-017-02342-1

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552. doi:10.1093/oxfordjournals.molbev.a026334

Cree, C., 2015. Distributions of glycine betaine and the methylamines in coastal waters: analytical developments and a seasonal study. University of Plymouth, Plymouth.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27, 1164–1165. doi:10.1093/bioinformatics/btr088

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. doi:10.1093/nar/gkh340

Emanuelsson, O., Brunak, S., Heijne, von, G., Nielsen, H., 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. 2, 953–971. doi:10.1038/nprot.2007.131

Fitzgerald, T.L., Waters, D.L.E., Henry, R.J., 2009. Betaine aldehyde dehydrogenase in plants. Plant Biology 11, 119–130. doi:10.1111/j.1438-8677.2008.00161.x

Haft, D.H., Selengut, J.D., White, O., 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. 31, 371–373.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. doi:10.1186/1471-2105-11-119

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., Beszteri, B., Bidle, K.D., Cameron, C.T., Campbell, L., Caron, D.A., Cattolico, R.A., Collier, J.L., Coyne, K., Davy, S.K., Deschamps, P., Dyhrman, S.T., Edvardsen, B., Gates, R.D., Gobler, C.J., Greenwood, S.J., Guida, S.M., Jacobi, J.L., Jakobsen, K.S., James, E.R., Jenkins, B., John, U., Johnson, M.D., Juhl, A.R., Kamp, A., Katz, L.A., Kiene, R., Kudryavtsev, A., Leander, B.S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A.M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M.A., Murray, S., Nadathur, G., Nagai, S., Ngam, P.B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M.C., Rengefors, K., Romano, G., Rumpho, M.E., Rynearson, T., Schilling, K.B., Schroeder, D.C., Simpson, A.G.B., Slamovits, C.H., Smith, D.R., Smith, G.J., Smith, S.R., Sosik, H.M., Stief, P., Theriot, E., Twary, S.N., Umale, P.E., Vaulot, D., Wawrik, B., Wheeler, G.L., Wilson, W.H., Xu, Y., Zingone, A., Worden, A.Z., 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the

functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 12, e1001889–6. doi:10.1371/journal.pbio.1001889

Keller, M.D., Matrai, P.A., Kiene, R.P., Bellows, W.K., 2004. Responses of coastal phytoplankton populations to nitrogen additions: dynamics of cell-associated dimethylsulfoniopropionate (DMSP), glycine betaine (GBT), and homarine. Can. J. Fish. Aquat. Sci. 61, 685–699. doi:10.1139/f04-058

Kirch, H.-H., Bartels, D., Wei, Y., Schnable, P.S., Wood, A.J., 2004. The ALDH gene superfamily of *Arabidopsis*. Trends Plant Sci. 9, 371–377. doi:10.1016/j.tplants.2004.06.004

Klemetsen, T., Raknes, I.A., Fu, J., Agafonov, A., Balasundaram, S.V., Tartari, G., Robertsen, E., Willassen, N.P., 2018. The MAR databases: development and implementation of databases specific for marine metagenomics. Nucleic Acids Res. 46, D692–D699. doi:10.1093/nar/gkx1036

Kopylova, E., Noé, L., Touzet, H., 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211–3217. doi:10.1093/bioinformatics/bts611

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Meth. 9, 357–359. doi:10.1038/nmeth.1923

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659. doi:10.1093/bioinformatics/17.3.282

Liew, Y.J., Aranda, M., Voolstra, C.R., 2016. Reefgenomics.Org - a repository for marine genomics data. Database 2016, baw152. doi:10.1093/database/baw152

Neave, M.J., Michell, C.T., Apprill, A., Voolstra, C.R., 2017a. *Endozoicomonas* genomes reveal functional adaptation and plasticity in bacterial strains symbiotically associated with diverse marine hosts. Sci. Rep. 7, 40579. doi:10.1038/srep40579

Neave, M.J., Rachmawati, R., Xun, L., Michell, C.T., Bourne, D.G., Apprill, A., Voolstra, C.R., 2017b. Differential specificity between closely related corals and abundant *Endozoicomonas* endosymbionts across global scales. ISME J. 11, 186–200. doi:10.1038/ismej.2016.95

Ngugi, D.K., Miyake, S., Cahill, M., Vinu, M., Hackmann, T.J., Blom, J., Tietbohl, M.D., Berumen, M.L., Stingl, U., 2017. Genomic diversification of giant enteric symbionts reflects host dietary lifestyles. Proc. Natl. Acad. Sci. U.S.A. 114, E7592–E7601. doi:10.1073/pnas.1703070114

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055. doi:10.1101/gr.186072.114

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490. doi:10.1371/journal.pone.0009490

Ressl, S., Terwisscha van Scheltinga, A.C., Vonrhein, C., Ott, V., Ziegler, C., 2009. Molecular basis of transport and regulation in the Na$^+$/betaine symporter BetP. Nature 458, 47–52. doi:10.1161/01.HYP.16.6.595

Robbins, S.J., Singleton, C.M., Chan, C.X., Messer, L.F., Geers, A.U., Ying, H., Baker, A., Bell, S.C., Morrow, K.M., Ragan, M.A., Miller, D.J., Forêt, S., ReFuGe2020 Consortium, Voolstra, C.R., Tyson, G.W., Bourne, D.G., 2019. A genomic view

of the reef-building coral *Porites lutea* and its microbial symbionts. Nat. Microbiol. 4, 2090–2100. doi:10.1038/s41564-019-0532-4

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. doi:10.1093/bioinformatics/btu033

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., Cornejo-Castillo, F.M., Costea, P.I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J.M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B.T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T.O., Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M.B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S.G., Bork, P., 2015. Structure and function of the global ocean microbiome. Science 348, 1261359–1261359. doi:10.1126/science.1261359

Whitman, W.B., Coleman, D.C., Wiebe, W.J., 1998. Prokaryotes: the unseen majority. Proc Natl Acad Sci USA 95, 6578–6583.

Wickham, H., 2009. ggplot2: elegant graphics for data analysis. Springer New York, New York, NY. doi:10.1007/978-0-387-98141-3

Xu, S., He, Z., Guo, Z., Zhang, Z., Wyckoff, G.J., Greenberg, A., Wu, C.-I., Shi, S., 2017. Genome-wide convergence during evolution of mangroves from woody plants. Mol. Biol. Evol. 34, 1008–1015. doi:10.1093/molbev/msw277