Photosynth Res
DOI 10.1007/s11120-014-0038-0

REGULAR PAPER

The *Prochlorococcus* carbon dioxide-concentrating mechanism: evidence of carboxysome-associated heterogeneity

Claire S. Ting · Katharine H. Dusenbury · Reid A. Pryzant · Kathleen W. Higgins · Catherine J. Pang · Christie E. Black · Ellen M. Beauchamp

Received: 19 March 2014 / Accepted: 28 August 2014

Abstract The ability of *Prochlorococcus* to numerically dominate open ocean regions and contribute significantly to global carbon cycles is dependent in large part on its effectiveness in transforming light energy into compounds used in cell growth, maintenance, and division. Integral to these processes is the carbon dioxide-concentrating mechanism (CCM), which enhances photosynthetic CO_2 fixation. The CCM involves both active uptake systems that permit intracellular accumulation of inorganic carbon as the pool of bicarbonate and the system of HCO_3^- conversion into CO_2. The latter is located in the carboxysome, a microcompartment designed to promote the carboxylase activity of Rubisco. This study presents a comparative analysis of several facets of the *Prochlorococcus* CCM. Our analyses indicate that a core set of CCM components is shared, and their genomic organization is relatively well conserved. Moreover, certain elements, including carboxysome shell polypeptides CsoS1 and CsoS4A, exhibit striking conservation. Unexpectedly, our analyses reveal that the carbonic anhydrase (CsoSCA) and CsoS2 shell polypeptide have diversified within the lineage. Differences in *csoSCA* and *csoS2* are consistent with a model of unequal rates of evolution rather than relaxed selection. The *csoS2* and *csoSCA* genes form a cluster in *Prochlorococcus* genomes, and we identified two conserved motifs directly upstream of this cluster that differ from the motif in marine *Synechococcus* and could be involved in regulation of gene expression. Although several elements of the CCM remain well conserved in the *Prochlorococcus* lineage, the evolution of differences in specific carboxysome features could in part reflect optimization of carboxysome-associated processes in dissimilar cellular environments.

Keywords Chlorophyll *b*-containing cyanobacteria · Genomic diversity · Global comparative genomics · Oxychlorobacteria · Microcompartment · Carbonic anhydrase

Introduction

For photosynthetic prokaryotes inhabiting variable environments, a central challenge involves optimizing metabolic flexibility within the crowded space of a single cell. The fact that key cellular processes can be incompatible adds layers of complexity to this evolutionary challenge. In light of this, the cyanobacterial CO_2^--concentrating mechanism (CCM) is an exquisite system that has evolved to enhance photosynthetic carbon dioxide fixation reactions (Price et al. 2008; Kupriyanova et al. 2013). This mechanism effectively concentrates CO_2 near ribulose-1,5 bisphosphate carboxylase oxygenase (Rubisco), which can utilize either CO_2 or O_2 in the reaction that it catalyzes involving ribulose 1,5-bisphosphate. One of the key components of the CCM is an active uptake system that permits inorganic carbon to accumulate within cells as the bicarbonate pool (Badger et al. 2006; Price et al. 2008). Together with the passive movement of CO_2 into cells, the recycling of leaked CO_2, and the maintenance of cytoplasmic pH levels, these essential features of the CCM ensure the accumulation of cellular inorganic carbon.

Electronic supplementary material The online version of this article (doi:10.1007/s11120-014-0038-0) contains supplementary material, which is available to authorized users.

C. S. Ting (✉) · K. H. Dusenbury · R. A. Pryzant · K. W. Higgins · C. J. Pang · C. E. Black · E. M. Beauchamp
Department of Biology, Williams College, Thompson Biology Lab 214, Williamstown, MA 01267, USA
e-mail: cting@williams.edu

Published online: 06 September 2014

A major operational component of the CCM involves the carboxysome, a microcompartment in which the carboxylase activity of Rubisco is optimized by the accumulation of CO_2 in its vicinity (Shively et al. 1973; Cannon et al. 2001; Yeates et al. 2008; Kerfeld et al. 2010; Espie and Kimber 2011; Kupriyanova et al. 2013; Rae et al. 2013a, b). Carboxysomes are polyhedral and are surrounded by a thin protein shell (3–6-nm thick), whose polypeptides are organized with apparent icosahedral symmetry (Iancu et al. 2007; Schmid et al. 2006; Yeates et al. 2008; Kinney et al. 2011; Sutter et al. 2013; Espie and Kimber 2011; Keeling et al. 2014). The shell can function as a permeability barrier and is thought to mediate the movement of molecules into and out of the carboxysome (Kerfeld et al. 2005; Espie and Kimber 2011; Kinney et al. 2011; Sutter et al. 2013). Although the carboxysome shell is permeable to protons (Menon et al. 2010), work by Dou et al. (2008) suggests that the carboxysome shell acts as a diffusional barrier to CO_2 and permits its accumulation in the carboxysome. Thus, with the functioning of carbonic anhydrase which co-localizes with Rubisco within carboxysomes and catalyzes the conversion of bicarbonate (HCO_3^-) and protons to CO_2 and water, CO_2 molecules are concentrated near the active site of Rubisco and are prevented from leaking out of the cell (Kerfeld et al. 2005; Price et al. 2008; Yeates et al. 2008; Espie and Kimber 2011).

Furthermore, structural studies have revealed that a large fraction of carboxysome shell proteins, which contain a bacterial microcompartment (BMC) domain (Pfam00936), are organized into hexameric units that constitute the facets of the icosahedron (Kerfeld et al. 2005; Tsai et al. 2007). These hexamers are associated with a central pore characterized by a large positive electrostatic potential (Kerfeld et al. 2005; Tsai et al. 2007). This central pore, as well as the gaps between hexamers, could be involved in regulating metabolite flux and in particular, could facilitate the movement of negatively charged molecules, such as bicarbonate (Kerfeld et al. 2005). Moreover, other carboxysome shell proteins form pentameric units, which could constitute the vertices of the icosahedral shell (Tanaka et al. 2008; Sutter et al. 2013; Keeling et al. 2014). These pentamers are also associated with a positively charged pore (Kinney et al. 2011; Sutter et al. 2013); however, because of the fewer number of pentamers compared to hexamers associated with an icosahedral shell, and the narrow pore diameter (about 4 Å), Sutter et al. (2013) have suggested that it is unlikely the pentamers have a significant role in metabolite flux across the shell. A unique class of trimeric carboxysome shell polypeptides (CsoS1D) has also recently been studied (Klein et al. 2009; Kinney et al. 2011). These exhibit pseudohexameric symmetry and are discussed in greater detail ahead.

Cyanobacterial carboxysomes are classified as either α- or β-carboxysomes, depending on whether they contain Form 1A or Form 1B Rubisco (Badger et al. 2002; Rae et al. 2013a). *Prochlorococcus* and some marine *Synechococcus* have α-carboxysomes, which are also found in several chemoautotrophic proteobacteria (Cannon et al. 2002; Badger and Price 2003; Badger et al. 2006; Espie and Kimber 2011). Studies using cryoelectron microscope tomography have revealed that in the lumen of the α-carboxysomes of *Halothiobacillus neapolitanus* and marine *Synechococcus* WH8102, Rubisco is organized in roughly concentric layers (Schmid et al. 2006; Iancu et al. 2007). Approximately 232 ± 18 Rubisco oligomers were observed in each marine *Synechococcus* WH8102 α-carboxysome, with about half of the Rubisco oligomers occurring in a layer next to the carboxysome shell (Iancu et al. 2007). Regions of direct contact between Rubisco and shell proteins were not observed (Iancu et al. 2007) and past work has suggested that a loose association exists between Rubisco and the caboxysome shell (Shively et al. 1973; So et al. 2004). Recently, it has been suggested that CsoS2 shell proteins could have a role in Rubisco organization in α-carboxysomes and might be involved in attaching Rubisco oligomers to the inner shell (Rae et al. 2013a).

Prochlorococcus, an ecologically important marine cyanobacterium that numerically dominates subtropical and tropical open ocean regions, represents a significant system in which to study the CCM. Recent work has revealed the evolution of striking diversity in this lineage, both in structure and function, and in particular, the photosynthetic apparatus and physiology (Moore et al. 1998; Moore and Chisholm 1999; Bibby et al. 2003; Ting et al. 2007, 2009). This diversification in part reflects ecotype-associated adaptation to abiotic and biotic factors, including utilization of key resources, such as light and nutrients. *Prochlorococcus* ecotypes involve isolates belonging to specific clades that share similarities at the physiological and genetic levels and examples include eMIT9312, eMED4, eNATL2A, eMIT9313, eSS120, and eMIT9211 (Coleman et al. 2006; Johnson et al. 2006). As field studies by Johnson et al. (2006) demonstrated, *Prochlorococcus* ecotype distributions can differ throughout the open ocean water column; for example, whereas eMIT9312 is present at high cell concentrations from the surface to 150 m, eMIT9313 is abundant only deeper in the water column (50–200 m).

A key element of the CCM that is of particular interest is the carbonic anhydrase, which serves to concentrate CO_2 within the carboxysome by catalyzing the conversion of HCO_3^- and protons to CO_2 and water (De Araujo et al. 2014; Kimber 2014; Nishimura et al. 2014). Thus, a basic operational feature of the CCM, involving the interconversion of inorganic carbon species inside the cell, is

ensured by the carbonic anhydrase. The *Prochlorococcus* carbonic anhydrase (CsoSCA) initially was identified as belonging to the novel ε-class, which includes the carbonic anhydrases of *Halothiobacillus neapolitanus* and marine *Synechococcus* WH8102, and was proposed to be distinct from the α-, β-, and γ-classes (So et al. 2004). However, structural studies indicated that the ε-class is actually a novel subclass of β-carbonic anhydrases (Sawaya et al. 2006). In members of this subclass, the two fused protein domains containing the pair of active sites in classic β-carbonic anhydrases have diverged to such an extent that only one domain still has an active site (Sawaya et al. 2006). The turnover number (k_{cat}) of the carbonic anhydrase present in carboxysomes purified from *Prochlorococcus* strain MED4 was reported to be less than (about half) that of the *H. neapolitanus* carbonic anhydrase and of other recombinant β-carbonic anhydrases (Roberts et al. 2012). Interestingly, however, the K_{CO_2} (295.1 ± 31.9 μM) of MED4 Rubisco is almost two times greater than that of the *H. neapolitanus* Rubisco (Roberts et al. 2012), and the K_{CO_2} (750 μM) of the MIT9313 Rubisco (cloned and expressed in *E. coli*) is the highest recorded for a Form I Rubisco (Scott et al. 2007). In light of these data, the *Prochlorococcus* carbonic anhydrase and its impact on the carboxylation activity of Rubisco in intact carboxysomes clearly require further characterization in different strains.

Recent studies on the *Prochlorococcus* carboxysome have also focused on CsoS1D, a novel shell polypeptide (Klein et al. 2009; Roberts et al. 2012). CsoS1D is a tandem BMC-domain protein (i.e., it contains a fusion of BMC domains), and CsoS1D trimers can dimerize to form pseudohexamers (Klein et al. 2009; Kinney et al. 2011). A distinct pore is present at the three-fold axis, and the conformation of neighboring Glu120 and Arg121 residues is thought to dictate whether the pore is in an open (~14 Å, diameter) or closed position, thus functioning as a potential gating mechanism (Klein et al. 2009; Kinney et al. 2011). It has been suggested that the relatively large size of this pore could permit metabolites, such as ribulose bisphosphate (RuBP), to pass freely, and that a gating mechanism could function in preventing loss of other key metabolites and/or the entrance of inhibitors (Klein et al. 2009; Kinney et al. 2011). However, it is unknown how metabolite specificity would be established or how structural changes associated with the open and closed pore conformations might be accommodated by the surrounding carboxysome shell proteins (Espie and Kimber 2011; Kimber 2014). Recent work involving production of carboxysomes in a heterologous host suggests that CsoS1D might have an important role in carboxysome architecture and assembly (Bonacci et al. 2012; Kimber 2014). The isolation of intact carboxysomes from *Prochlorococcus* strain MED4 revealed that CsoS1D is a low abundance shell protein (Roberts et al. 2012). In this same work, Western analyses were used to confirm that CosS1 (10.7 kDa), CsoS2 (101 kDa), and Rubisco (53.5 kDa) are associated with the MED4 carboxysome (Roberts et al. 2012).

In this study, we present a comparative analysis of several proteins that have a central role in the *Prochlorococcus* carbon dioxide-concentrating mechanism. While one might expect the CCM to be well conserved among ecotypes, it is conceivable differences in specific elements might evolve in response to factors, such as microheterogeneity in the physico-chemical properties of the water column and/or the need to optimize mechanisms for specific cellular/physiological environments. Previous work reported that carboxysome size is not conserved within the *Prochlorococcus* lineage (Ting et al. 2007), and some strains (i.e., those that are deeply branched, such as MIT9313, MIT9303, SS120, MIT9211, NATL1A, NATL2A) have an additional carboxysome gene (*csoS1E*) (previously named CsoS1-2, in Ting et al. 2007). CsoS1E has since been found in all marine cyanobacteria except *Prochlorococcus* strains belonging to the large clade of recently differentiated lineages (i.e., MED4, AS9601, MIT9215, MIT9301, MIT9312, MIT9515, MIT9202; Roberts et al. 2012). The precise function of CsoS1E is unknown, and although it has a C-terminal BMC domain, its N-terminus does not share homology to known protein domains (Kinney et al. 2011). Through our analyses, we have established that within the *Prochlorococcus* lineage, fundamental components of the CCM are shared and are even highly conserved between strains, which represent different ecotypes. Interestingly, however, specific elements have diversified between strains, including the carbonic anhydrase, an integral component of the CO_2-concentrating mechanism.

Materials and methods

Culture conditions

Prochlorococcus strain MIT9313 was grown in batch culture in an artificial sea water medium supplemented with $NaHCO_3$ (2 mM, final concentration) at 21 ± 1 °C and 20 μmol photons m^{-2} s^{-1}. Illumination was supplied by cool white fluorescence lights on a 14 h light: 10 h dark cycle. Cell growth was monitored by measuring changes in Chlorophyll *a* fluorescence over time using an Aquafluor fluorometer (Turner Designs), and cells were sampled for transmission electron microscopy during the exponential growth stage.

Genome sequence data and analyses

The annotated files containing the complete genome sequences of 12 *Prochlorococcus* strains (MIT9312, MED4, MIT9515, MIT9301, MIT215, AS9601, SS120, MIT9211, NATL1A, NATL2A, MIT9303, MIT9313), as well as other cyanobacteria, were extracted from GenBank (http://www.ncbi.nlm.nih.gov) and were used in constructing our in-house database system as described in Ting et al. (2009). Gene and protein sequences for specific CCM components were extracted from our in-house database, GenBank (http://www.ncbi.nlm.nih.gov), and Cyanobase (http://genome.microbedb.jp/cyanobase/).

The chromosomal organization of specific genes was visualized using Artemis (Rutherford et al. 2000). EMBOSS and CLUSTAL W or Omega were used for local/global pairwise sequence alignments and for multiple sequence alignments, respectively, and all alignments were conducted with the BLOSUM scoring matrix and default gap values. Analyses of the similarity of the MIT9303 genome (reference sequence) to nine other *Prochlorococcus* genomes and the relative location of genes encoding proteins involved in the carbon dioxide-concentrating mechanism were conducted with the BLAST Ring Image Generator (BRIG version 0.95; Alikhan et al. 2011). The phylogenetic footprint-discovery program from Regulatory Sequence Analysis Tools (RSAT) was used to detect putative regulatory motifs associated with the *Prochlorococcus* carboxysome gene cluster (Defrance et al. 2008; Thomas-Chollier et al. 2008). For the RSAT analyses, *Prochlorococcus* strain SS120 was designated as the query organism, and both dyad filtering and predicting operon leader genes were selected as analysis options. Transmembrane helices associated with the putative bicarbonate transporters of *Prochlorococcus* were predicted with the MEMbrane protein Structure and Topology (MEMSAT3) program (Jones et al. 1994; Jones 2007).

Molecular evolutionary analyses were conducted using the MEGA 5 (versions 5.05 and 5.2.1) software package (Kumar et al. 2004). Gene sequence alignments were manually edited so that they were in frame and consistent with protein sequence alignments. The maximum-likelihood method with a WAG model was used for protein phylogenetic analyses (Whelan and Goldman 2001). Nucleotide maximum-likelihood analyses were conducted using the Tamura–Nei model, with all codon positions included (Tamura et al. 2011). For both protein and gene maximum-likelihood analyses, all gaps were deleted, and bootstrap analyses involved 1,000 resamplings. The number of synonymous (d_S) and nonsynonymous (d_N) substitutions per site and tests of purifying selection were calculated using the modified Nei–Gojobori method (Jukes–Cantor), with a transition/transversion ratio of 0.89 for *csoSCA* and 0.82 for *csoS2*. Nonparametric relative rate tests were conducted using Tajima's general method (Tajima 1993), with one degree of freedom and a significance level of 5 %.

Two-step chemical fixation and electron microscopy

Cells were harvested by centrifugation, resuspended in a small volume (<10 ml) of the original culture media, and then repelleted in an Eppendorf Micro Centrifuge (1,675×*g*, 10 min). Following the removal of all sea water medium, cells were resuspended in the primary fixation buffer, which contained 2 % glutaraldehyde (Ted Pella, Inc., CA) and 0.25 M sucrose in a 0.1 M sodium phosphate buffer (~pH 7) and were fixed for at least 90 min at 4 °C and in darkness. Following primary fixation, cells were rinsed twice with 0.1 M sodium phosphate buffer (~pH 7) containing 0.25 M sucrose, and once with 0.1 M sodium phosphate buffer alone. Following post-fixation in 2 % potassium permanganate (Mallinckrodt) for about 2 h at 4 °C, samples were subjected to a graded ethanol dehydration series (30–100 %) and embedded in Spurr embedding medium (Ted Pella, Inc.). Ultrathin sections (55 nm) were stained with uranyl acetate (~6 %) and lead citrate, and examined in a Philips CM-10 transmission electron microscope (TEM).

Results and discussion

Visualization of *Prochlorococcus* carboxysomes

When visualized using transmission electron microscopy, *Prochlorococcus* carboxysomes appear as prominent, electron-dense, polygonal inclusions that have rounded rather than sharp corners and appear to lack an internal crystalline substructure (Fig. 1a). The carboxysomes tend to cluster in the central cytoplasmic space and are rarely positioned between the intracytoplasmic membranes and plasma membrane (Fig. 1a). This clustering is consistent with that observed in *Prochlorococcus* MED4 and MIT9313 cells visualized in a near-native state using cryoelectron microscope tomography, and we previously suggested that this cellular organization might facilitate carbon fixation (Ting et al. 2007). Carboxysome clustering has also been reported in *Halothiobacillus neapolitanus* cells visualized using electron cryotomography (Iancu et al. 2010). Previous studies indicate that *Prochlorococcus* carboxysome diameters can differ by as much as 40 nm between strains [i.e., approximately 90 nm in MED4 vs. 130 nm in MIT9313 (Ting et al. 2007)]. Overall carboxysomes from *Prochlorococcus* are within the size range reported for other (cyano)bacteria (*Synechococcus*

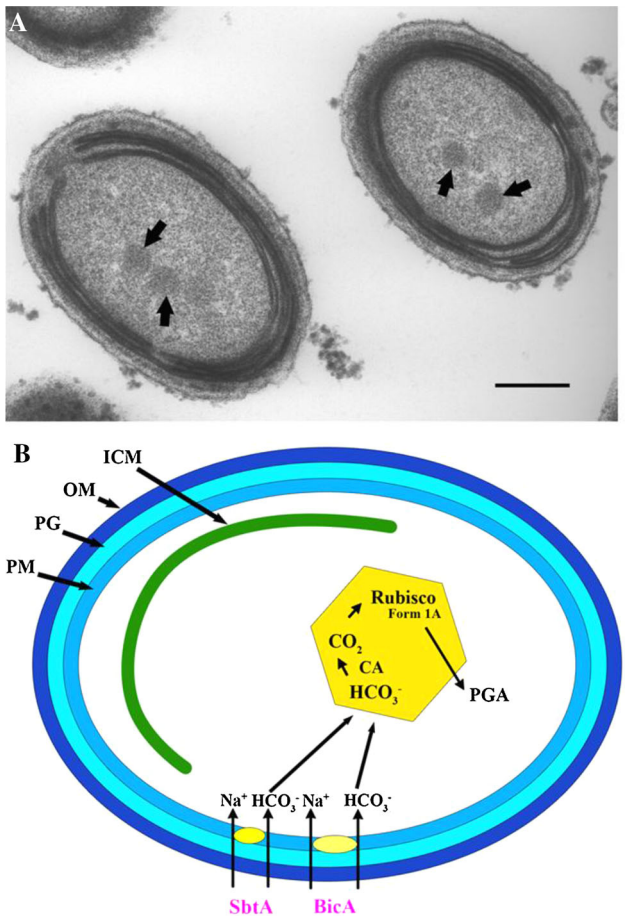

Fig. 1 Architecture of the *Prochlorococcus* cell and visualization of elements of the CCM. **a** Transmission electron micrograph of *Prochlorococcus* cells (strain MIT9313) preserved by chemical fixation. Carboxysomes (*black arrows*) are visible as electron dense structures in the central cytoplasmic space. Note that the intracytoplasmic lamellae are visible toward the cell periphery and are present in bands of approximately three layers. *Scale bar*, 0.25 μm **b** Illustration of a *Prochlorococcus* cell depicting CCM elements, including the carboxysome (*yellow hexagonal*) and putative HCO_3^- transporters (BicA, SbtA, *yellow circles*). The carboxysome-associated carbonic anhydrase (CA) catalyzes the conversion of HCO_3^- to CO_2. *PM* plasma membrane, *PG* peptidoglycan, *OM* outer membrane, *ICM* intracytoplasmic lamellae, *PGA* phosphoglyceric acid

WH8109, 92–116 nm, Dai et al. 2013; *Synechococcus* WH8102, 114–137 nm, Iancu et al. 2007; *Halothiobacillus neapolitanus*, average size 134 ± 8 nm, Iancu et al. 2010). The relationship between the carboxysome and putative components of the CCM in *Prochlorococcus* is depicted in the illustration shown in Fig. 1b.

Genomic organization of CCM components

In *Prochlorococcus*, genes encoding carboxysome proteins are part of a highly conserved cluster that includes *rbcL* and *rbcS*, and in the deeply branched MIT9313 and MIT9303 strains, the chromosomal location of this cluster relative to genes (*sbtA*, *bicA1/2*) encoding putative bicarbonate transporters is also conserved (Fig. 2; Tables S1, S2). Notably, in these two strains, a putative *bicA1* and *sbtA* gene cluster is located upstream of and nearby the carboxysome genes, while *bicA2* is found further downstream (and on the reverse strand in MIT9303) (Fig. 2). However, in most strains (MED4, MIT9312, MIT9515, AS9601, MIT9215, MIT9301, SS120, MIT9211, NATL1A, NATL2A), the *bicA1* and *sbtA* cluster is located at a greater distance upstream from the carboxysome gene cluster; in addition, while the carboxysome genes are located on the forward strand, *bicA1* and *sbtA*, as well as *bicA2*, are found on the reverse strand (shown in Fig. 2 for eight strains, Tables S1, S2).

Eight carboxysome-related genes (*csoS1D*, *csoS1*, *rbcL*, *rbcS*, *csoS2*, *csoSCA*, *csoS4A*, *csoS4B*) exhibit synteny in all *Prochlorococcus* genomes (Fig. 2), with a gene encoding a HAM1 family protein occurring on the opposite strand between *csoS1D* and *csoS1* in all strains (not shown in Fig. 2). The deeply branched MIT9303 strain has gene insertions that are not present in other strains, and these occur directly upstream of *csoS1*, *csoSCA*, and *csoS1E*. These insertions involve genes of unknown function (P9303_08061, P9303_08111), as well as a predicted pseudogene (annotated as a pseudogene derived from P9313_15181). Furthermore, as our laboratory (Ting et al. 2009) and as others (Roberts et al. 2012) have reported, deeply branched *Prochlorococcus* strains (SS120, MIT9211, NATL2A, NATL1A, MIT9313, MIT9303) retain an additional putative carboxysome gene (*csoS1E*) which is located directly downstream of *csoS4B* (Fig. 2). The *csoS1E* gene is also found in *Synechococcus*, including marine strains such as WH8102, WH7803, and WH8109 (Roberts et al. 2012). In light of this, it is most parsimonious to presume that *csoS1E* was present in the original ancestor of the *Prochlorococcus* lineage and was subsequently lost from the common ancestor that gave rise to the large clade of recently differentiated lineages that include strains MED4, MIT9515, MIT9312, MIT9301, MIT9215, and AS9601.

In the cyanobacterium *Synechococcus* WH5701, it has been suggested that key components of the CCM were acquired via lateral gene transfer (Rae et al. 2011). In particular, bicarbonate transporters (*cmpABCD*, *sbtAB 1*) were found to be present in a genomic island and occurred in association with a transposable element (Rae et al. 2011). In *Prochlorococcus*, genomic islands have been predicted in strains MIT9312 and MED4 based on comparative analyses of regions that do not exhibit synteny (Coleman et al. 2006) and in strains MED4, SS120, and MIT9313 (Dufresne et al. 2008) based on methods modified from Hsiao et al. (2005) and Rusch et al. (2007). Tables S3A and S3B present complete lists of the genomic islands identified in these studies, as well as those predicted by IslandViewer (http://www.

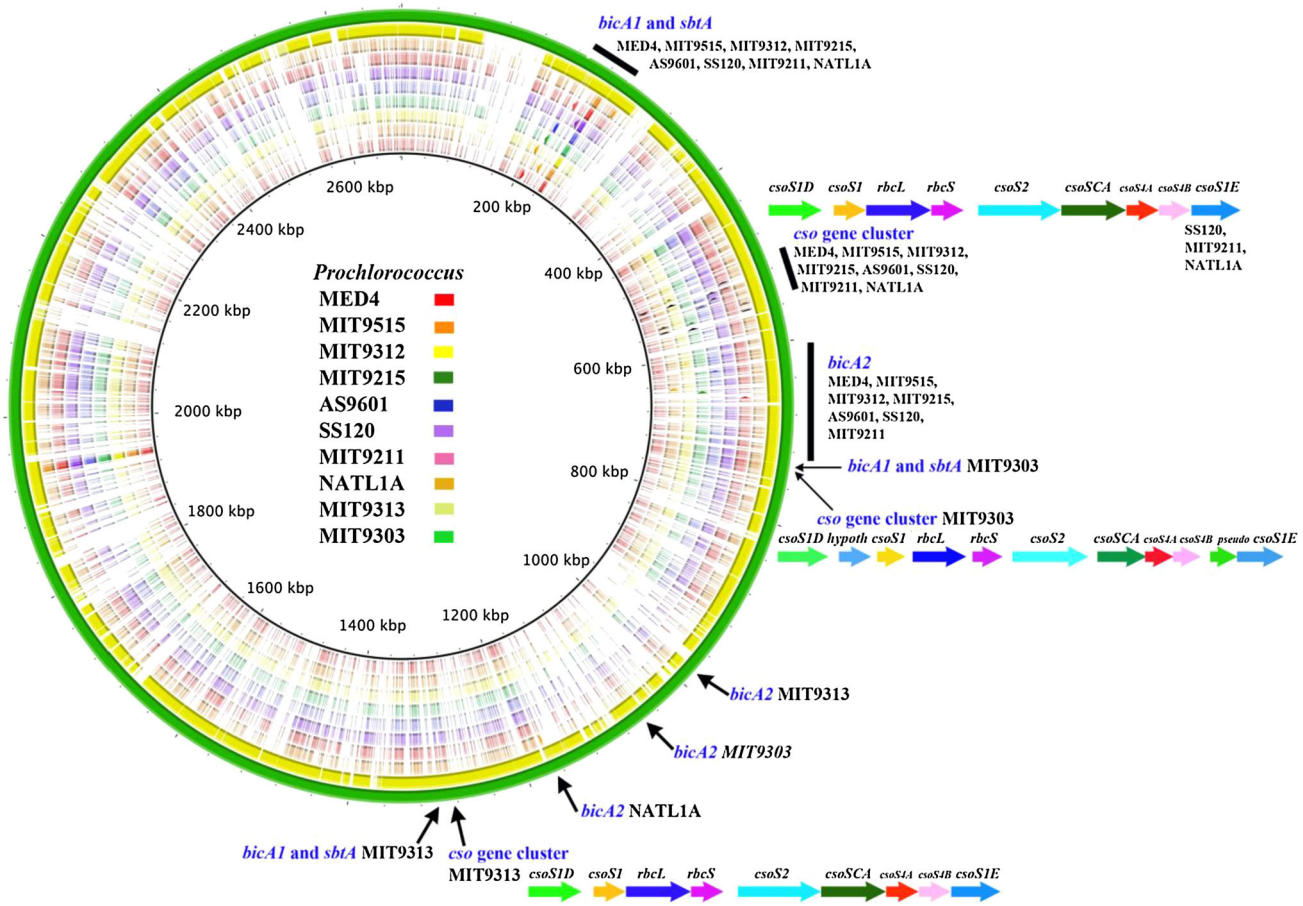

Fig. 2 Genomic context and organization of genes encoding predicted proteins involved in the CCM in *Prochlorococcus* presented with the genomes of ten strains visualized using the BLAST Ring Image Generator (BRIG version 0.95; Alikhan et al. 2011). Information about gene ID numbers and chromosomal start sites for each *Prochlorococcus* strain is provided in Tables S1 and S2. The MIT9303 genome (bright green, outermost ring) was used as the reference sequence in the BRIG diagram, and the genomes of other *Prochlorococcus* strains are represented by *colors* indicated in the *center* of the *ring*. Differences in the relative intensities of the *colored lines* in a *single ring* correlate with the similarity between a gene in a particular strain and that same gene in MIT9303 (i.e., the more the intense a color, the greater the similarity between the genes). A gene encoding a predicted HAM1 family protein is present on the opposite strand between *csoS1D* and *csoS1* in all *Prochlorococcus* strains but has not been included in the diagram. Note that strain MIT9303 has gene insertions (*hypoth*, *pseudo*) present within the carboxysome gene cluster that is absent in other strains. In this strain, a short gene (108 bp, P9303_08111, not shown) encoding a putative hypothetical protein has also been indentified between *csoS2* and *csoSCA* and overlaps in sequence with *csoSCA*

pathogenomics.sfu.ca/islandviewer/query.php), which integrates the IslandPick, SIGI-HMM, and IslandPath/DIMOB programs (Langille et al. 2008). Although there is consistency between the different approaches, it should be noted that none of the methods yielded completely identical results (Tables S3A, S3B). In *Prochlorococcus*, *bicA1*, *bicA2*, and *sbtA*, as well as other genes (*csoS1D*, *csoS1*, *rbcL*, *rbcS*, *csoS2*, *csoSCA*, *csoS4A*, *csoS4B*) encoding components of the CCM are not associated with the island regions predicted by any of these approaches.

Conservation of putative bicarbonate transporters

Previous analyses of cyanobacterial CO_2-concentrating-mechanisms by Price et al. (2008) suggest that *Prochlorococcus* strains lack the gene encoding the BCT1 high-affinity bicarbonate (HCO_3^-) transporter, as well as genes encoding the Ndh-1_4 and Ndh-1_3 CO_2 transporters. However, they do possess genes encoding putative BicA low-affinity bicarbonate transporters and a putative SbtA high-affinity bicarbonate transporter (Price et al. 2008), which exhibits low (22–25 %) sequence identity with the SbtA from *Synechocystis* PCC6803. Both BicA and SbtA are Na^+/HCO_3^- symporters, and BicA belongs to the SulP family of anion transporters found in prokaryotes and eukaryotes, and are often annotated as sulfate transporters in cyanobacterial genomes (Price et al. 2004, Price 2011; Kupriyanova et al. 2013).

Our comparative genomic analyses of 12 strains confirm that *Prochlorococcus* lacks genes encoding BCT1 and

Ndh-1_4/Ndh-1_3. However, all strains possess genes encoding putative BicA1, BicA2, and SbtA transporters, which exhibit comparable sequence identities (BicA1, 72–99 %; BicA2, 62–98 %; SbtA, 61–98 %) within the *Prochlorococcus* lineage (Tables S4A, S4B). While *Prochlorococcus* BicA1 exhibits greater sequence identity (68–69 %) with the *Synechococcus* WH8102 NP_898027 protein, BicA2 exhibits greater sequence identity (65–80 %) with the NP_896932 protein (Table S4A), both of which have been annotated as sulfate transporters. Although an additional gene annotated as BicA (NP_897617) is also present in the WH8102 genome, the predicted protein sequence exhibits lower sequence identity with the *Prochlorococcus* BicA1/A2 proteins (Table S4A).

Previous studies by Shelden et al. (2010) on BicA (566 amino acids) from *Synechococcus* PCC7002 demonstrated experimentally that this protein has 12 membrane-spanning regions. These authors also reported that a highly conserved sequence motif (NSNKELIGQGLGN) associated with members of the SulP family of proteins occurs in the loop region connecting helices eight and nine (Shelden et al. 2010). Notably, all transmembrane helix prediction programs tested, except for MEMSAT3 and SCAMPI-msa, suggested incorrectly that this loop region is a membrane-spanning domain (Shelden et al. 2010).

BicA1 has a predicted length of 550–555 amino acids in most *Prochlorococcus* strains, except for MIT9211 (577 amino acids) and MIT9313/MIT9303 (573 amino acids). In contrast, BicA2 is smaller and consists of 514–527 amino acids. Using MEMSAT3, we predicted that BicA1 from *Prochlorococcus* has ten transmembrane helices, and BicA2 has 11 transmembrane helices in all strains except for NATL1A/NATL2A (ten helices). Moreover, we were able to establish that in the loop region connecting helices eight and nine of BicA1, a highly conserved motif is present (NSDRELIGQGIGN) in *Prochlorococcus* that is similar to the one present in *Synechococcus* PCC7002. The loop region from MIT9215 is the only sequence to differ by one amino acid in the fourth position (NSDKELIGQGIGN). Similarly, in the loop connecting helices eight and nine in BicA2, a well-conserved motif is present (NKNKEARGQGIAN). However, the fourth residue of this sequence is more variable, and among deeply branched strains, this residue is a V (NATL2A, NATL1A, MIT9211, MIT9313, MIT9303) or a T (SS120); in addition, the 11th residue has been mutated to an M in one strain (NATL1A), and the 12th residue has been mutated to a G in NATL1A/NATL2A.

SbtA was characterized originally in *Synechocystis* PCC6803 as a high-affinity Na^+-dependent bicarbonate transporter (Shibata et al. 2002) and is thought to function as a tetramer (Zhang et al. 2004; Price 2011). A single subunit of the *Synechocystis* PCC6803 SbtA (slr1512) consists of 374 amino acids and has ten membrane-spanning domains (Price 2011). The putative SbtA of *Prochlorococcus* is smaller (330–341 amino acids) than the *Synechocystis* PCC6803 SbtA and shares low sequence identity (22–25 %) with it. Use of MEMSAT3 enabled us to predict that the putative *Prochlorococcus* SbtA also has ten transmembrane helices in all strains.

Comparative genomics of *Prochlorococcus* carboxysome proteins

Conservation of CsoS1, CsoS1E, CsoS1D, CsoS4A, and CsoS4B

Comparative genomic analyses of predicted carboxysome-associated polypeptide sequences indicate that CsoS1, CsoS1E, CsoS1D, CsoS4A, and CsoS4B are relatively well conserved within the *Prochlorococcus* lineage. CsoS1, a major shell protein containing one bacterial microcompartment domain (BMC), is one of the most highly conserved proteins. Notably, comparisons among 12 *Prochlorococcus* strains indicate that sequence identities for CsoS1 range from 97 to 100 % (Table 1). This extent of sequence conservation across the *Prochlorococcus* lineage for a protein is striking and not common. Furthermore, comparisons of CsoS1 between *Prochlorococcus* and marine *Synechococcus* WH8102 (98–100 %, Table 1), as well as other *Synechococcus* strains including WH7803 (98–100 %), RCC307 (91–96 %), CC9605 (98–100 %), and CC9902 (98–100 %), indicate that this protein is also highly conserved between genera.

The crystal structure of *Halothiobacillus neapolitanus* CsoS1A has been solved to 1.4 Å resolution (Tsai et al. 2007). The predicted amino acid sequence of *H. neapolitanus* CsoS1A shares 83 % sequence identity with that of its *Prochlorococcus* counterparts. CsoS1A is thought to assemble into a molecular layer consisting of distinct hexameric units, each associated with a central pore of about 4 Å (at the narrowest region) through which small-charged molecules, such as HCO_3^-, might pass (Tsai et al. 2007). The high level of sequence conservation between *H. neapolitanus* CsoS1A and *Prochlorococcus* CsoS1, as well as *Synechococcus*, suggests that this polypeptide is a key building block of α-carboxysomes in different organisms and that its major structural and functional roles are conserved.

Earlier reports indicated that the deeply branched *Prochlorococcus* strains MIT9313, SS120, NATL2A, and MIT9211 have a gene encoding CsoS1E (previously called CsoS1-2) which shares the highest (77–79 %) sequence identity with CsoS1 (previously called CsoS1-1) and is absent from strains belonging to the large clade of recently

Table 1 Identity matrix for CsoS1 (bold, above the diagonal) and CsoS1E (below the diagonal) of 12 *Prochlorococcus* (*Pro*) strains and marine *Synechococcus* (*Syn*)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
(1) *Pro* MED4		**100**	**99**	**99**	**99**	**99**	**98**	**98**	**98**	**98**	**98**	**97**	**98**	**98**
(2) *Pro* MIT9515	–		**99**	**99**	**99**	**99**	**98**	**98**	**98**	**98**	**98**	**97**	**98**	**98**
(3) *Pro* MIT9301	–	–		**100**	**100**	**100**	**98**	**98**	**98**	**98**	**98**	**97**	**98**	**98**
(4) *Pro* AS9601	–	–	–		**100**	**100**	**98**	**98**	**98**	**98**	**98**	**97**	**98**	**98**
(5) *Pro* MIT9215	–	–	–	–		**100**	**98**	**98**	**98**	**98**	**98**	**97**	**98**	**98**
(6) *Pro* MIT9312	–	–	–	–	–		**98**	**98**	**98**	**98**	**98**	**97**	**98**	**98**
(7) *Pro* NATL1A	–	–	–	–	–	–		**100**	**99**	**99**	**100**	**98**	**99**	**99**
(8) *Pro* NATL2A	–	–	–	–	–	–	99		**99**	**99**	**100**	**98**	**99**	**99**
(9) *Pro* SS120	–	–	–	–	–	–	60	60		**100**	**99**	**99**	**100**	**100**
(10) *Pro*MIT9211	–	–	–	–	–	–	58	59	66		**99**	**99**	**100**	**100**
(11) *Pro* MIT9313	–	–	–	–	–	–	54	54	58	55		**98**	**99**	**99**
(12) *Pro* MIT9303	–	–	–	–	–	–	54	55	60	54	90		**99**	**99**
(13) *Syn* WH8102	–	–	–	–	–	–	54	56	57	62	60	60		**100**
(14) *Syn* WH7803	–	–	–	–	–	–	52	52	53	57	63	61	62	

differentiated lineages (Ting et al. 2007). Comparisons between strains indicate that CsoS1E amino acid sequences share identities ranging from 54 to 66 %, and are not as well conserved as CsoS1 (Table 1). The only exceptions involve comparisons between the strains NATL1A/NATL2A (99 %) and MIT9313/MIT9303 (90 %) (Table 1). The greatest variability in CsoS1E exists in the N-terminal region of the polypeptide sequence, where there are large insertions and deletions. Removal of this region results in a significant increase in CsoS1E sequence identity scores (82 to 88 %), and these scores are again higher for NATL1A/NATL2A (100 %) and MIT9313/MIT9303 (98 %) comparisons. Although the structure and function of CsoS1E are not yet known, this protein has a predicted C-terminal bacterial microcompartment (BMC) domain (Kinney et al. 2011). Furthermore, microarray-based gene expression studies by Tolonen et al. (2006) indicated that *csoS1E* was expressed in MIT9313 grown at 10 umol photons m^{-2} s^{-1} under nutrient replete conditions; however, its expression levels were much lower (5–12×) relative to other carboxysome genes (*csoS4B*, *csoS4A*, *csoSCA*, *csoS2*, and *rbcS*) (Tolonen et al. 2006, Supplementary Material).

All *Prochlorococcus* strains also possess genes encoding CsoS1D which has been annotated as a hypothetical protein in the genomes. This protein has been identified as a low abundance shell polypeptide in MED4 (Roberts et al. 2012), and structural studies have established that this tandem BMC protein forms trimers which dimerize into hexamers that are associated with a central pore (14 Å) (Klein et al. 2009). The opening of this pore is likely gated by an Arg side chain, and structural analyses suggest that the convergence of three Arg side chains could result in pore closure (Klein et al. 2009). In the *Prochlorococcus* lineage, the predicted amino acid sequence of CsoS1D exhibits relatively high sequence identities (77–100 %) among all strains (Table S5A).

Genes encoding CsoS4A and CsoS4B are located in a relatively tight cluster that includes *csoS2* and *csoS3* (Fig. 2). While *csoS4A* is separated from *csoSCA* by only 0–3 nucleotides in all strains, *csoS4A* and *csoS4B* overlap by one nucleotide in strains such as MIT9313 and MIT9303 or are separated by 5–32 nucleotides in strains belonging to the large clade of recently differentiated lineages. Notably, CsoS4A exhibits high (86–100 %) sequence identity among all *Prochlorococcus* strains (Table S5B). In contrast, sequence identities are the highest for CsoS4B in comparisons between strains belonging to the large clade of recently differentiated lineages (92–100 %) and are lower (78–85 %) in comparisons between this group and deeply branched strains. CsoS4A and CsoS4B contain a EutN (Pfam03319) domain in the N-terminal region of the polypeptide sequence that has also been identified in the β-carboxysome CcmL protein, and structural studies indicate that CsoS4A subunits associate to form pentamers, which function as vertices in the carboxysome shell (Tanaka et al. 2008; Kinney et al. 2011). In individual pentamers, the C-terminal regions of neighboring CsoS4A subunits are tightly associated, and these pentamers are characterized mainly by a positive electrostatic potential and have a central pore of approximately 3.5 Å (Tanaka et al. 2008; Kinney et al. 2011).

Table 2 Identity matrix for the carbonic anhydrase protein (CsoSCA, bold, above the diagonal) and gene (*csoSCA*, below the diagonal) sequences of 12 *Prochlorococcus* (*Pro*) strains and marine *Synechococcus* (*Syn*)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
(1) *Pro* MED4		**87**	**86**	**85**	**85**	**85**	**58**	**58**	**62**	**61**	**57**	**56**	**58**	**57**
(2) *Pro* MIT9515	83		**94**	**93**	**94**	**93**	**58**	**58**	**63**	**61**	**59**	**57**	**58**	**57**
(3) *Pro* MIT9301	82	91		**98**	**96**	**97**	**58**	**58**	**62**	**60**	**59**	**59**	**58**	**59**
(4) *Pro* MIT9215	82	91	97		**95**	**96**	**58**	**58**	**62**	**60**	**58**	**58**	**58**	**57**
(5) *Pro* MIT9312	82	92	94	94		**95**	**58**	**58**	**62**	**60**	**58**	**58**	**58**	**59**
(6) *Pro* AS9601	82	91	97	96	94		**58**	**58**	**62**	**60**	**59**	**59**	**58**	**59**
(7) *Pro* NATL1A	68	67	67	66	67	67		**99**	**66**	**67**	**66**	**65**	**59**	**60**
(8) *Pro* NATL2A	68	67	67	67	67	67	100		**66**	**67**	**66**	**64**	**59**	**61**
(9) *Pro* SS120	67	68	66	67	68	68	69	69		**74**	**69**	**69**	**64**	**67**
(10) *Pro* MIT9211	67	65	65	66	66	67	69	69	75		**69**	**68**	**66**	**66**
(11). *Pro* MIT9313	63	62	63	62	63	63	67	67	69	69		**97**	**69**	**70**
(12) *Pro* MIT9303	62	62	62	62	62	64	66	66	69	69	97		**68**	**70**
(13) *Syn* WH7803	63	62	62	62	62	61	63	63	65	67	69	68		**72**
(14) *Syn* WH8102	58	58	60	59	59	59	62	62	63	64	67	67	70	

Divergence of CsoSCA and CsoS2 within the Prochlorococcus lineage

A central element of the carbon dioxide-concentrating mechanism is the carbonic anhydrase (CA), which functions to concentrate CO_2 in the carboxysome by catalyzing the conversion of HCO_3^- and protons to CO_2 and water. In *Prochlorococcus*, the *csoSCA* (formerly *csoS3*) gene encodes a carbonic anhydrase (So et al. 2004). Interestingly, comparisons of CsoSCA/*csoSCA* between members of the large clade of recently differentiated lineages (MED4, MIT9515, MIT9301, MIT9215, MIT9312, AS9601) indicate that MED4 exhibits lower (85–87 %, protein; 82–83 %, gene) pairwise identities than other strains (93–98 %, protein; 91–97 %, gene, Table 2). Moreover, comparisons between strains within this large clade and those that are deeply branched (NATL1A, NATL2A, SS120, MIT9211, MIT9313, MIT9303) indicate that CsoSCA/*csoSCA* sequence identities are generally low (56–63 % protein; 62–68 %, gene) and are comparable to those involving marine *Synechococcus* (57–70 %, protein; 58–69 %, gene, Table 2).

Analyses of the number of synonymous (d_S) and nonsynonymous (d_N) substitutions per site for *csoS3* indicate that $d_S > d_N$ in all strain comparisons and that this gene is under purifying selection (ρ-value is significant at the 5 % level) in all except three strain comparisons (Table S6). In two of these comparisons (MIT9313/MIT9303 vs. MIT9215), evolutionary distances could not be estimated (Table S6), and the third comparison involved the NATL1A and NATL2A pair (ρ-value = 0.052). For this latter pair, the codon-based test of neutrality indicated that the null hypothesis of strict neutrality ($d_N = d_S$) could not be rejected at the 5 % level (ρ-value = 0.101).

Alignments of the derived CsoSCA sequences from different *Prochlorococcus* strains indicate that while there is high conservation among strains in the central catalytic region of the sequence, as well as in the C-terminus, significant variability exists at the beginning (residues 1 to approximately 47) of the N-terminal region (Fig. 3). It has been suggested that the N-terminal domain might be the region involved in interactions with the carboxysome shell or Rubisco (Sawaya et al. 2006). In *Prochlorococcus*, the beginning of this region is characterized by insertions and deletions, as well as numerous point mutations. Structural studies on CsoSCA from *Halothiobacillus neapolitanus* have identified the remaining regions of the N-terminal domain to consist of four α-helices (Sawaya et al. 2006), and these helices, as well as their intervening loop regions, are well conserved among *Prochlorococcus* strains (Fig. 3). In addition, these structural studies on the *H. neapolitanus* CsoSCA have suggested that in the active site of the enzyme, a zinc ion is coordinated by Cys-173, His-242, and Cys-253 (Sawaya et al. 2006). These three amino acid residues are conserved in all *Prochlorococcus* CsoSCA sequences, and their positions are indicated in the catalytic domain of the sequence in Fig. 3 (shown in light blue). Moreover, in *H. neapolitanus,* CsoSCA Asp-175 and Arg-177 are thought to have a role in key catalytic steps, including substrate binding (Sawaya et al. 2006). Both Asp-175 and Arg-177 are also conserved in all *Prochlorococcus* CsoSCA sequences, and they are located in a loop region of the catalytic domain between neighboring β-sheets (Fig. 3, shown in light blue).

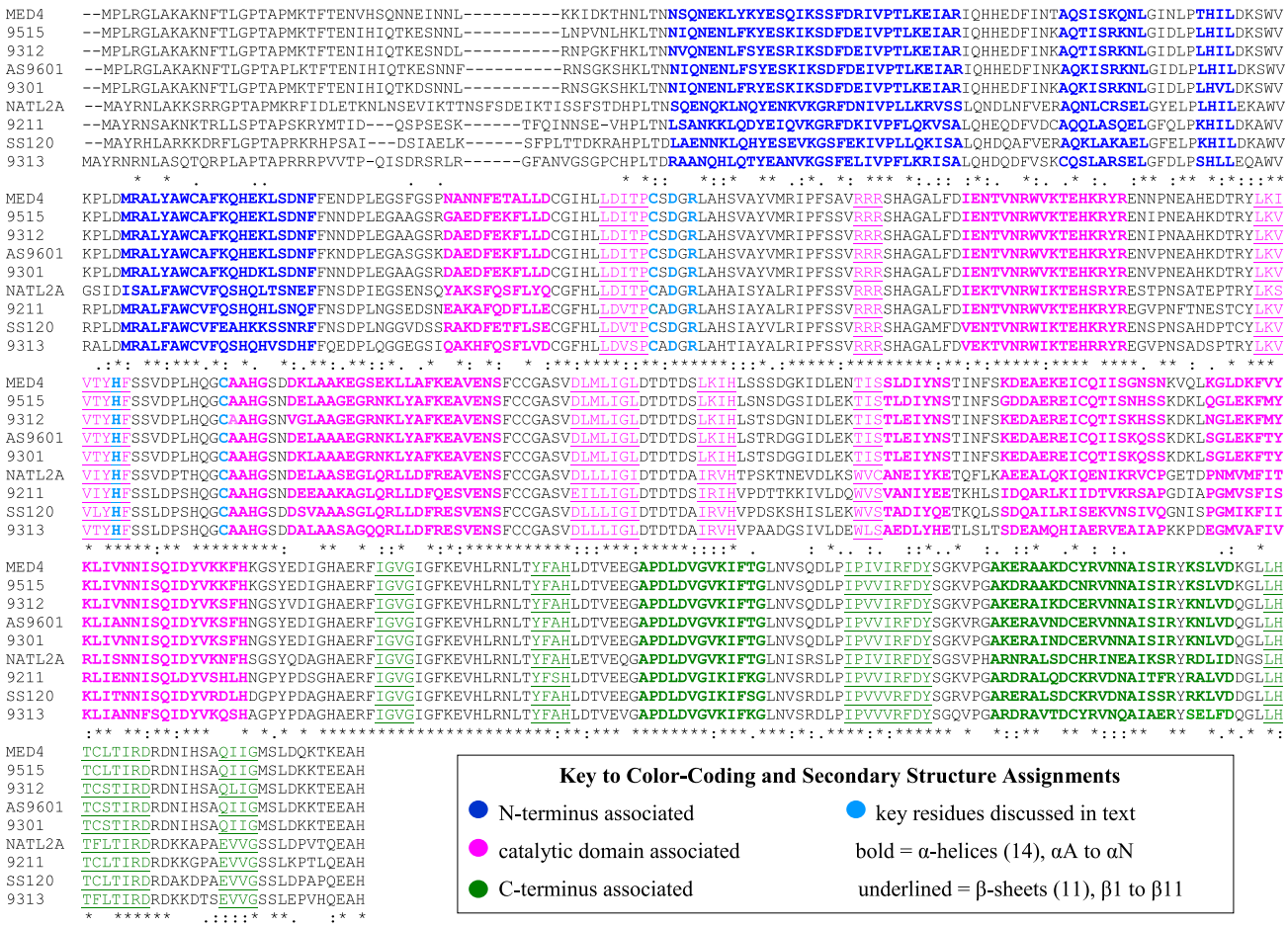

Fig. 3 Alignment of the predicted amino acid sequences of CsoSCA from nine *Prochlorococcus* strains. Residues associated with the N-terminus are shown in blue, those associated with the C-terminus are shown in green, and those associated with the catalytic domain are shown in pink. Predictions of α-helices (14) and *β*-sheets (11) are based on structural studies on CsoSCA from *H. neapolitanus* (Sawaya et al. 2006), and residues associated with these secondary structures are in *bold* (α-helices) or are *underlined* (*β*-sheets). Identical residues are designated with an *asterisk*, conserved residues are indicated with a *semicolon*, and semiconserved residues are marked with a *period*

Phylogenetic trees constructed from *csoSCA* (Fig. 4a) and CsoSCA (Fig. 4b) reveal that the clustering of isolates is generally congruent with 16S rDNA-based phylogenetic groupings. As in ribosomal trees, MIT9313 and MIT9303 are very closely related and form the basal branch of the *Prochlorococcus* clade (Fig. 4a). However, in phylogenetic trees constructed with either *csoSCA* (Fig. 4a) or CsoSCA (Fig. 4b), MED4 alone forms the basal branch of the large clade of recently differentiated lineages. This branching pattern is strongly supported (bootstrap value = 100) in both the *csoSCA* gene (Fig. 4a) and CsoSCA protein (Fig. 4b) trees and differs from the branching pattern observed in trees constructed from 16S rDNA or 16S–23S rRNA ITS region sequences, or in trees constructed using genome-based phylogenies (Kettler et al. 2007). In these latter phylogenetic trees, MED4 and MIT9515 group separately from MIT9312, MIT9215, AS9601, and MIT9301 (Kettler et al. 2007).

The above observations led us to ask whether the MED4 CsoSCA/*csoSCA* sequence is accumulating mutations at a different rate compared to other *Prochlorococcus* strains. In order to address this we conducted relative rate tests with *Synechococcus* WH8102 serving as the outgroup. The data in Table 3, as well as Table S7, indicate that in tests with MED4 and members of the large clade of recently differentiated lineages (MIT9515, MIT9312, MIT9301, MIT9215, AS9601), mutations in CsoSCA/*csoSCA* (all positions, as well as first and second nucleotide positions) appear to be accumulating at approximately the same rate (χ^2 was not significant at the 5 % level).

Notably, however, relative rate tests for CsoSCA/*csoSCA* between MED4 and deeply branched *Prochlorococcus* strains (SS120, MIT9211, NATL2A, NATL1A, MIT9313, MIT9303) revealed that it is possible to reject the null hypothesis of equal rates of evolution (Table 3). This was true for tests involving all nucleotides positions,

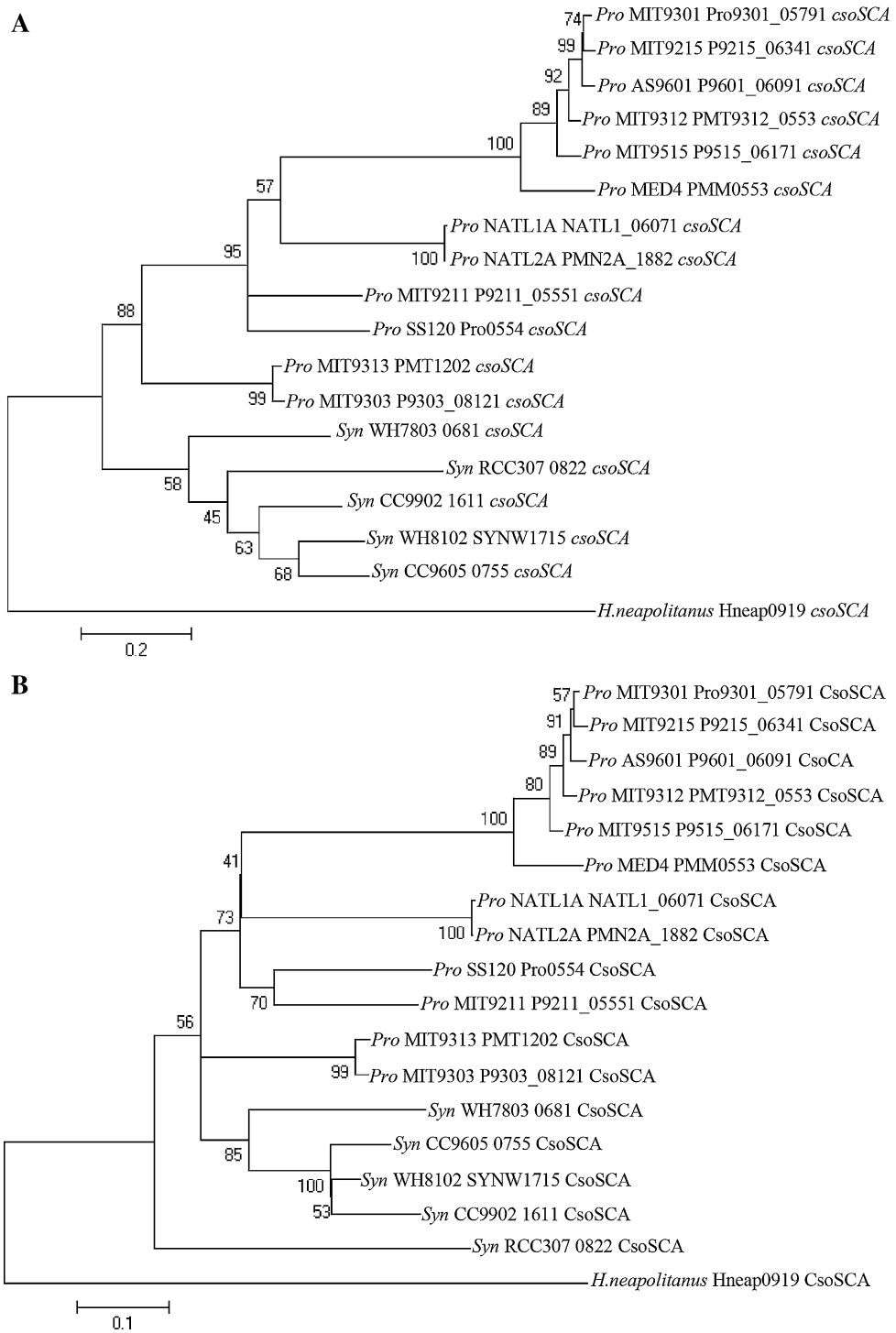

Fig. 4 Phylogenetic analysis of **a** *csoSCA* and **b** CsoSCA from *Prochlorococcus* and *Synechococcus* using maximum-likelihood methods. Sequences from *Halothiobacillus neapolitanus* were used as the outgroup in these analyses. Support values for internal branches are displayed at each node and represent percentage *bootstrap values* from 1,000 resamplings

as well as the first and second nucleotide positions. Furthermore, this was the case when the entire amino acid sequence was used and even when the more variable region at the beginning of the N-terminal domain was removed (Table 3). Thus, these data suggest that CsoSCA is not evolving at the same rate in all *Prochlorococcus* strains. In particular, this is true for deeply branched strains. The data in Table 3 indicate that the MIT9313/MIT9303 CsoSCA is accumulating mutations at a different rate compared to CsoSCA from other deeply branched strains (SS120, MIT9211, NATL2A). It should also be pointed out that the results of the relative rate tests between SS120, MIT9211,

Table 3 Nonparametric relative rate test scores for the carbonic anhydrase gene (*csoSCA*) and protein (CsoSCA) sequences of *Prochlorococcus* (*Pro*)

A	B	Nucleotide[a] (all positions)	Nucleotide[a] (1st, 2nd positions)	Amino acid[a]	Amino acid[a,b] (-N-terminus)
Pro MED4	*Pro* SS120	25.02 (0.000)	53.08 (0.000)	20.60 (0.000)	19.60 (0.000)
Pro MED4	*Pro* MIT9211	36.91 (0.000)	52.20 (0.000)	17.78 (0.000)	18.60 (0.000)
Pro MED4	*Pro* NATL2A	11.31 (0.001)	19.36 (0.000)	6.19 (0.013)	7.04 (0.008)
Pro MED4	*Pro* NATL1A	11.65 (0.001)	19.17 (0.000)	5.76 (0.016)	6.58 (0.010)
Pro MED4	*Pro* MIT9313	69.37 (0.000)	85.96 (0.000)	35.76 (0.000)	34.62 (0.000)
Pro MED4	*Pro* MIT9303	67.64 (0.000)	90.16 (0.000)	34.95 (0.000)	33.80 (0.000)
Pro MED4	*Pro* MIT9515	2.10 (0.147)	3.95 (0.047)	0.89 (0.346)	1.47 (0.225)
Pro MIT9313	*Pro* SS120	16.12 (0.000)	8.60 (0.003)	4.57 (0.033)	4.38 (0.036)
Pro MIT9313	*Pro* MIT9211	9.54 (0.002)	11.00 (0.001)	6.87 (0.009)	5.56 (0.018)
Pro MIT9313	*Pro* NATL2A	26.21 (0.000)	27.44 (0.000)	14.40 (0.000)	13.44 (0.000)
Pro MIT9303	*Pro* SS120	16.12 (0.000)	10.94 (0.001)	4.38 (0.036)	4.19 (0.041)
Pro MIT9303	*Pro* MIT9211	9.41 (0.002)	13.40 (0.000)	6.21 (0.013)	4.95 (0.026)
Pro MIT9303	*Pro* NATL2A	25.89 (0.000)	30.15 (0.000)	13.76 (0.000)	12.81 (0.000)
Pro SS120	*Pro* NATL2A	2.36 (0.125)	7.93 (0.005)	3.95 (0.047)	3.28 (0.070)
Pro MIT9211	*Pro* NATL2A	6.41 (0.011)	6.44 (0.011)	2.51 (0.113)	2.58 (0.108)

[a] *Synechococcus* WH8102 was used as the outgroup in all tests. Numbers represent χ^2 values and numbers in parentheses represent the ρ-value for the indicated analysis. Low ρ-values (<0.05) were used to reject the null hypothesis of equal rates of evolution between A and B

[b] Amino acid (-N-terminus) indicates analyses conducted using CsoSCA protein sequences lacking the initial N-terminus region (first 43 amino acids for MED4, MIT9515, MIT9312, AS9601, MIT9301, MIT9215, MIT9211; first 44 amino acids for SS120; first 48 amino acids for MIT9313, MIT9303; first 53 amino acids for NATL1A, NATL2A)

and NATL2A/NATL1A were dependent on whether the nucleotide (all positions or first and second positions only) or amino acid sequence was being examined (Tables 3, S7). For example, relative rate tests between MIT9211 and NATL2A indicated that although CsoSCA is accumulating mutations an equal rate, *csoSCA* (all positions, as well as first and second positions) is not (Table 3).

As relative rate tests conducted in the absence of the most variable region of the N-terminal domain did not alter these conclusions (Table 3), the data suggest that the accumulation of point mutations in other regions of CsoSCA might also have a role. For example, alignments of CsoSCA from several *Prochlorococcus* strains reveal a region within the catalytic domain where sequence conservation is lower than in neighboring regions (Fig. 3). In particular, regions between β-sheets five and six and α-helices J and K possess lower sequence conservation, as does α-helix J.

In all *Prochlorococcus* strains, *csoSCA* is clustered with *csoS2,* and these genes are separated by only seven nucleotides. The *csoS2* gene encodes a putative carboxysome shell polypeptide, and this structural protein has not yet been characterized in detail (Kinney et al. 2011). Our analyses of predicted CsoS2 sequences indicate that low (55–59 %) identities exist between strains belonging to the large clade of recently differentiated lineages and those that are more deeply branched (Table S8). Analyses of the number of synonymous (d_S) and nonsynonymous (d_N) substitutions per site for *csoS2* indicate that $d_S > d_N$ in all strain comparisons and that this gene is under purifying selection (ρ-value is significant at the 5 % level) in all except one comparison (NATL1A/NATL2A, Table S9). For NATL1A/NATL2A, the codon-based test of neutrality indicated that the null hypothesis of strict neutrality ($d_N = d_S$) could not be rejected at the 5 % level (ρ-value = 0.357).

In order to examine whether mutations might be accumulating at unequal rates in CsoS2 in different strains, we conducted relative rate tests using CsoS2 from *Synechococcus* WH8102 as the outgroup. Tests involving members of the large clade of recently differentiated lineages (MED4, MIT9312, MIT9301, MIT9215, AS9601) indicate that mutations in CsoS2 appear to be accumulating at approximately the same rate (χ^2 was not significant at the 5 % level, Table S10A). However, relative rate tests involving deeply branched strains (NATL1A, NATL2A, SS120, MIT9211, MIT9313, MIT9303) reveal that evolutionary rates are not equivalent in all strains (Table S10B, note comparisons between MIT9313/MIT9303 and SS120/NATL1A/NATL2A, and between MIT9211 and NATL1A/NATL2A). Moreover, comparisons between members of the large clade of recently differentiated lineages (MED4, MIT9312, MIT9301, MIT9215, AS9601) and deeply branched strains also indicate that mutations in CsoS2 are

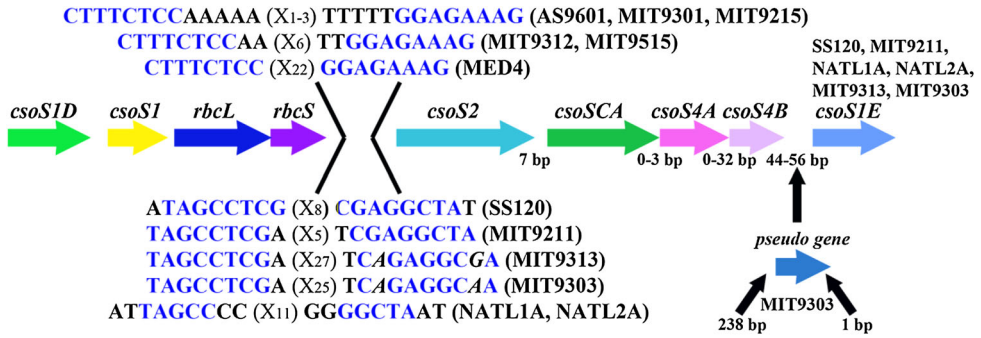

Fig. 5 Identification of putative hairpin-forming regulatory motifs directly upstream of *csoS2* and *csoSCA* in 12 *Prochlorococcus* genomes. Strains (MED4, MIT9312, MIT9515, AS9601, MIT9301, MIT9215) belonging to the large clade of recently differentiated lineages share a conserved motif and are shown above the carboxysome gene cluster. Deeply branched strains (SS120, MIT9211, MIT9303, MIT9313, NATL1A, NATL2A) share a different motif and are shown below the carboxysome gene cluster. The length of the intervening region ($X_{1–27}$) in the motif is indicated in the figure and differs between strains. A gene encoding a HAM1 family protein is present on the opposite strand between *csoS1D* and *csoS1* in all *Prochlorococcus* strains and is not included in the diagram. Notably, marine *Synechococcus* strains possess a different motif (GAG-CCCTGA-$X_{5–9}$–TCAGGGCTC) located in the same region upstream of *csoS2* and *csoSCA*

not accumulating at equal rates between strains (Table S10B). In summary, although the majority of carboxysome genes/proteins are well conserved in the *Prochlorococcus* lineage, our data suggest that *csoSCA* and *csoS2* constitute points of diversification among strains.

A putative regulatory motif unique to the *Prochlorococcus* carboxysome gene cluster

The highly conserved genomic organization of *Prochlorococcus* carboxysome genes discussed previously led us to examine whether putative regulatory motifs are associated with this gene cluster and could be identified in individual *Prochlorococcus* genomes. Phylogenetic footprint analyses conducted using Regulatory Sequence Analysis Tools (RSAT, http://rsat.ulb.ac.be/rsat/; Defrance et al. 2008; Thomas-Chollier et al. 2008) resulted in the identification of a putative hairpin-forming motif located directly upstream of *csoS2* (Fig. 5). Interestingly, although this motif is located in a similar region of the genome in strains belonging to the large clade of recently differentiated lineages (AS9601, MIT9301, MIT9215, MIT9312, MIT9515, MED4) and in deeply branched strains (SS120, MIT9211, MIT9313, MIT9303, NATL1A, NATL2A), the sequence of the motif is not conserved between these groups (Fig. 5).

For strains AS9601, MIT9301, MIT9215, MIT9312, MIT9515, and MED4, a relatively well-conserved motif was identified (Fig. 5). A region of eight nucleotides, CTTTCTCC-$X_{1–3/6/22}$–GGAGAAAG, is identical in the motifs present in each of these strains. Differences in the overall length of this motif between strains are due to the number of terminal adenines/thymines (two or five) and to the length of the intervening region ($X_{1–3/6/22}$) (Fig. 5). In *Prochlorococcus* strain MED4, this motif lacks all terminal adenines/thymines but has a longer intervening region ($X = 22$ nucleotides) (Fig. 5).

Among strains (SS120, MIT9211, NATL2A, NATL1A, MIT9313, MIT9303) that are more deeply branched within the *Prochlorococcus* lineage, a putative hairpin-forming motif was also identified in the same region upstream of *csoS2* (Fig. 5). The overall length of the motif, minus the intervening region, is nine base pairs in all of the deeply branched strains, and differences between strains are mainly due to the presence of one or four additional adenines/thymines and/or guanines/cytosines (Fig. 5). Although there is more variability in this motif among strains, a region of five nucleotides, TAGCC-$X_{5/8/11/25/27}$–GGCTA, is conserved in all motifs. Furthermore, in two strains (SS120, MIT9211), an additional three nucleotides are also conserved, TAGCCTCG-$X_{5/8/25/27}$–CGAGGCTA (Fig. 5). Strains MIT9313 and MIT9303, which are most deeply branched within the *Prochlorococcus* lineage, also have these three additional nucleotides, and their motif contains an insertion (an adenine) in this region in the second part of the motif (Fig. 5). Moreover, the thymine present in the second part of the motif has been mutated to an adenine/guanine in MIT9313 and MIT9303 (Fig. 5), and the length of the intervening region ($X = 25$ or 27) is more than twice that found in other strains.

Future work will establish whether this motif has a role in regulating gene expression, and thus, whether it might impact strain-specific differences in carboxysome gene expression. The motifs identified within the *Prochlorococcus* lineage are unique and are not present in marine *Synechococcus*. Phylogenetic footprint analyses conducted using RSAT resulted in the identification of a different putative hairpin-forming motif in marine *Synechococcus* that is located in the same region upstream of *csoS2* and

csoSCA. Notably, in contrast to what was observed in the *Prochlorococcus* lineage, this motif is highly conserved among *Synechococcus* strains that belong to different clades and were isolated from habitats such as the open ocean (*Synechococcus* WH8102 (clade III), WH7803 (clade V)) and California current (*Synechococcus* CC9311(clade I), CC9605 (clade II), CC9902 (clade IV)). This motif consists of nine nucleotides (GAGCCCTGA-X_{5-9}–TCAGGGCTC) that are identical in all five *Synechococcus* strains and an intervening region (X_{5-9}) of five to nine base pairs.

Conclusions

Cyanobacterial carbon dioxide-concentrating and carbon fixation reactions are associated intimately with the overall physiology and photosynthetic strategies of a cell. Our work indicates that *Prochlorococcus* strains share a core set of CCM elements, whose genomic context and organization are relatively well conserved. However, while certain components of the carboxysome, such as the CsoS1 and CsoS4A shell polypeptides, exhibit striking conservation, major proteins, such as the carbonic anhydrase (CsoSCA) and CsoS2 shell polypeptide, have diversified within the *Prochlorococcus* lineage. Our results indicate that differences in *csoSCA* and *csoS2* between strains are consistent with a model of unequal rates of evolution rather than relaxed selection. It will be important for future studies to address the impact of these differences in primary structure on carbonic anhydrase activity and the possibility of a structural/functional relationship between CsoSCA and CsoS2. The *csoS2* and *csoSCA* genes form a tight cluster in all *Prochlorococcus* genomes, and we identified two motifs upstream of this cluster. Interestingly, marine *Synechococcus* strains possess a different, yet highly conserved, motif in the same genomic context upstream of *csoS2-csoSCA*. As this putative hairpin-forming motif could be linked to strain-specific differences in gene expression, its role will be important to investigate in *Prochlorococcus*. While fundamental elements of the CCM are shared within the *Prochlorococcus* lineage, one cannot rule out the possibility that strain/ecotype-specific differences have evolved for optimizing carboxysome-associated function within specialized cellular environments.

Acknowledgments This work was supported by the National Science Foundation, Award Number MCB-0850900 to C.S. Ting and by Williams College (C.S.T., K.H.D., R.A.P., K.W.H., C.J.P., C.E.B., E.M.B.). The authors would like to thank the anonymous reviewers of this manuscript for their insightful suggestions and helpful comments.

References

Alikhan N-F, Petty NK, Zakour NLB, Beatson SA (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genom 12:402

Badger MR, Price GD (2003) CO_2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. J Exp Bot 54:609–622

Badger MR, Hanson D, Price GD (2002) Evolution and diversity of CO_2 concentrating mechanisms in cyanobacteria. Funct Plant Biol 29:161–173

Badger MR, Price GD, Long BM, Woodger FJ (2006) The environmental plasticity and ecological genomics of the cyanobacterial CO_2 concentrating mechanism. J Exp Bot 57:249–265

Bibby TS, Mary I, Nield J, Partensky F, Barber J (2003) Low-light-adapted *Prochlorococcus* species possess specific antennae for each photosystem. Nature 424:1051–1054

Bonacci W, Teng PK, Afonso B, Niederholtmeyer H, Grob P, Silver PA, Savage DF (2012) Modularity of a carbon-fixing protein organelle. Proc Natl Acad Sci USA 109:478–483

Cannon GC, Bradburne CE, Aldrich HC, Baker SH, Heinhorst S, Shively JM (2001) Microcompartments in prokaryotes: carboxysomes and related polyhedral. Appl Environ Microbiol 67:5351–5361

Cannon GC, Heinhorst S, Bradburne CE, Shively JM (2002) Carboxysome genomics: a status report. Funct Plant Biol 29:175–182

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF, Chisholm SW (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311:1768–1770

Dai W, Fu C, Raytchevea D, Flanagan J, Khant HA, Liu X, Rochat RH, Haase-Pettingell C, Piret J, Ludtke SJ, Nagayama K, Schmid MF, King JA, Chiu W (2013) Visualizing virus assembly intermediates inside marine cyanobacteria. Nature 502:707–710

De Araujo C, Arefeen D, Tadesse Y, Long BM, Price GD, Rowlett RS, Kimber MS, Espie GS (2014) Identification and characterization of a carboxysomal γ-carbonic anhydrase from the cyanobacterium *Nostoc* sp. PCC7120. Photosyn Res 121:135–150

Defrance M, Janky R, Sand O, van Helden J (2008) Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. Nature Prot 3:1589–1603

Dou Z, Heinhorst S, Williams EB, Murin CD, Shively JM, Canon GC (2008) CO_2 fixation kinetics of *Halothiobacillus neapolitanus* mutant carboxysomes lacking carbonic anhydrase suggest the shell acts as a diffusional barrier for CO_2. J Biol Chem 283:10377–10384

Dufresne A, Ostrowski M, Scanlan DJ, Garczarek L, Mazard S, Palenik BP, Paulsen IT, Tandeau de Marsac N, Wincker P, Dossat C, Ferriera S, Johnson J, Post AF, Hess WR, Partensky F (2008) Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. Genome Biol 9:R90

Espie GS, Kimber MS (2011) Carboxysomes: cyanobacterial RubisCO comes in small packages. Photosyn Res 109:7–20

Hsiao WWL, Ung K, Aeschliman D, Bryan J, Finlay BB, Brinkman FSL (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. PLoS Genet 1:e62

Iancu CV, Ding HJ, Morris DM, Dias DP, Gonzales AD, Martino A, Jensen GJ (2007) The structure of isolated *Synechococcus* strain WH8102 carboxysomes revealed by electron cryotomography. J Mol Biol 372:764–773

Iancu CV, Morris DM, Dou Z, Heinhorst S, Canon GC, Jensen GJ (2010) Organization, structure, and assembly of α-carboxysomes

determined by electron cryotomography of intact cells. J Mol Biol 396:105–117

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science 311:1737–1740

Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics 23:538–544

Jones DT, Taylor WR, Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochem 33:3038–3049

Keeling TJ, Samborska B, Demers RW, Kimber MS (2014) Interactions and structural variability of β-carboxysomal shell protein CcmL. Photosyn Res 121:125–133

Kerfeld CA, Sawaya MR, Tanaka S, Nguyen CV, Phillips M, Beeby M, Yeates TO (2005) Protein structures forming the shell of primitive bacterial organelles. Science 309:936–938

Kerfeld CA, Heinhorst S, Cannon GC (2010) Bacterial microcompartments. Annu Rev Microbiol 64:391–408

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet 3:2515–2528

Kimber MS (2014) Carboxysomes – Sequestering RubisCO for efficient carbon fixation. In: (MF Homann-Marriott, Ed) The Structural Basis of Biological Energy Generation. Advances in Photosynthesis and Respiration. Springer, Dordrecht, The Netherlands, pp.133–148

Kinney JN, Axen SD, Kerfeld CA (2011) Comparative analysis of carboxysome shell proteins. Photosyn Res 109:21–32

Klein MG, Zwart P, Bagby SC, Cai F, Chisholm SW, Heinhorst S, Cannon GC, Kerfeld CA (2009) Identification and structural analysis of a novel carboxysome shell protein with implications for metabolite transport. J Mol Biol 392:319–333

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Briefings in Bioinformatics 5:150–163

Kupriyanova EV, Sinetova MA, Cho SM, Park Y-I, Los DA, Pronina NA (2013) CO_2-concentrating mechanism in cyanobacterial photosynthesis: organization, physiological role, and evolutionary origin. Photosyn Res 117:133–146

Langille MGI, Hsiao WWL, Brinkman FSL (2008) Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinformatics 9:1–10

Menon BB, Heinhorst S, Shively JM, Cannon GC (2010) The carboxysome shell is permeable to protons. J Bact 192:5881–5886

Moore LR, Chisholm SW (1999) Photophysiology of the marine cyanobacterium *Prochlorococcus*: ecotypic differences among cultured isolates. Limnol Oceanogr 44:628–638

Moore LR, Rocap G, Chisholm SW (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature 393:464–467

Nishimura T, Yamaguchi O, Takatani N, Maeda S, Omata T (2014) In vitro and in vivo analyses of the role of the carboxysomal β-type carbonic anhydrase of the cyanobacterium *Synechococcus elongatus* in carboxylation of ribulose-1,5-bisphosphate. Photosyn Res 121:151–157

Price GD (2011) Inorganic carbon transporters of the cyanobacterial CO_2 concentrating mechanism. Photosyn Res 109:47–57

Price GD, Woodger FJ, Badger MR, Howitt SM, Tucker L (2004) Identification of a SulP-type bicarbonate transporter in marine cyanobacteria. Proc Natl Acad Sci USA 101: 18228–18233

Price GD, Badger MR, Woodger FJ, Long BM (2008) Advances in understanding the cyanobacterial CO_2-concentrating mechanism (CCM): functional components, C_i transporters, diversity, genetic regulation and prospects for engineering into plants. J Exp Bot 59:1441–1461

Rae BD, Forster B, Badger MR, Price GD (2011) The CO_2-concentrating mechanism of *Synechococcus* WH5701 is composed of native and horizontally-acquired components. Photosyn Res 109:59–72

Rae BD, Long BM, Badger MR, Price GD (2013a) Functions, compositions, and evolution of the two types of carboxysomes: polyhedral microcompartments that facilitate CO_2 fixation in cyanobacteria and some proteobacteria. Microbiol Mol Biol Rev 77:357–379

Rae BD, Long BM, Whitehead LF, Forster B, Badger MR, Price GD (2013b) Cyanobacterial carboxysomes: microcompartments that facilitate CO_2 fixation. J Mol Microbiol Biotechnol 23:300–307

Roberts EW, Cai F, Kerfeld CA, Cannon GC, Heinhorst S (2012) Isolation and characterization of the *Prochlorococcus* carboxysome reveal the presence of the novel shell protein CsoS1D. J Bacteriol 194:787–795

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcon LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, et al. (2007) The Sorcerer II global ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific. PLoS Biol 5:e77

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

Sawaya MR, Cannon GC, Heinhorst S, Tanaka S, Williams EB, Yeates TO, Kerfeld CA (2006) The structure of the *β*-carbonic anhydrase from the carboxysomal shell reveals a distinct subclass with one active site for the price of two. J Biol Chem 281:7546–7555

Schmid MF, Paredes AM, Khant HA, Soyer F, Aldrich HC, Chiu W, Shively JM (2006) Stucture of *Halothiobacillus neapolitanus* carboxysomes by cryo-electron tomography. J Mol Biol 364:526–535

Scott KM, Henn-Sax M, Harmer TL, Longo DL, Frame CH, Cavanaugh CM (2007) Kinetic isotope effect and biochemical characterization of form IA RubisCO from the marine cyanobacterium *Prochlorococcus* marinus MIT9313. Limnol Oceanogr 52:2199–2204

Shelden MC, Howitt SM, Price GD (2010) Membrane topology of the cyanobacterial bicarbonate transporter, BicA, a member of the SulP (SLC26A) family. Mol Mem Biol 27:12–23

Shibata M, Katoh H, Sonoda M, Ohkawa H, Shimoyama M, Fukuzawa H, Kaplan A, Ogawa T (2002) Genes essential to sodium-dependent bicarbonate transport in cyanobacteria – function and phylogenetic analysis. J Biol Chem 277:18658–18664

Shively JM, Ball FL, Kline BW (1973) Electron microscopy of the carboxysomes (polyhedral bodies) of *Thiobacillus neapolitanus*. J Bact 116:1405-1411

So AK-C, Espie GS, Williams EB, Shively JM, Heinhorst S, Cannon GC (2004) A novel evolutionary lineage of carbonic anhydrase (ε- Class) is a component of the carboxysome shell. J Bact 186:623–630

Sutter M, Wilson SC, Deutsch S, Kerfeld CA (2013) Two new high-resolution crystal structures of carboxysome pentamer proteins reveal high structural conservation of CcmL orthologs among distantly related cyanobacterial species. Photosyn Res 118:9–16

Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599–607

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739

Tanaka S, Kerfeld CA, Sawaya MR, Cai F, Heinhorst S, Canon GC, Yeates TO (2008) Atomic-level models of the bacterial carboxysome shell. Science 319:1083–1086

Thomas-Chollier M, Sand O, Turatsinze J-V, Janky R, Defrance M, Vervisch E, Brohee S, van Helden J (2008) RSAT: regulatory sequence analysis tools. Nucleic Acids Res 36:W119–W127

Ting CS, Hsieh C, Sundararaman S, Mannella C, Marko M (2007) Cryo-electron tomography reveals the comparative three-dimensional architecture of *Prochlorococcus*, a globally important marine cyanobacterium. J Bacteriol 189:4485–4493

Ting CS, Ramsey ME, Wang YL, Frost AM, Jun E, Durham T (2009) Minimal genomes, maximal productivity: comparative genomics of the photosystem and light-harvesting complexes in the marine cyanobacterium, *Prochlorococcus*. Photosyn Res 101:1–19

Tolonen AC, Aach J, Lindell D, Johnson ZI, Rector T, Steen R, Church GM, Chisholm SW (2006) Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. Mol Syst Biol 2: Article 53

Tsai Y, Sawaya MR, Cannon GC, Cai F, Williams EB, Heinhorst S, Kerfeld CA, Yeates TO (2007) Structural analysis of CsoS1A and the protein shell of the *Halothiobacillus neapolitanus* carboxysome. PLoS Biol 5:1345–1354

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699

Yeates TO, Kerfeld CA, Heinhorst S, Cannon GC, Shively JM (2008) Protein-based organelles in bacteria: carboxysomes and related microcompartments. Nat Rev Microbiol 6:681–691

Zhang PP, Battchikova N, Jansen T, Appel J, Ogawa T, Aro EM (2004) Expression and functional roles of the two distinct NDH-1 complexes and the carbon acquisition complex NdhD3/NdhF3/CupA/Sll1735 in *Synechocystis* PCC6803. Plant Cell 16:3326–3340