Life in a Drop of Water

by

Sarah Catherine Bagby

B.A. (Hons), Physiological Sciences
Oxford University, 2002

B.S., Biological Chemistry; B.A., Chemistry; B.A., Philosophy
University of Chicago, 2000

SUBMITTED TO THE DEPARTMENT OF BIOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2009

Signature of Author: ______________________________
Department of Biology
22 May 2009

Certified by: ______________________________
Sallie W. Chisholm
Professor of Biology
Lee and Geraldine Martin Professor of Environmental Studies
Thesis Supervisor

Accepted by: ______________________________
Tania A. Baker
E. C. Whitehead Professor of Biology and Investigator, Howard Hughes Medical Institute
Co-chairperson, Graduate Committee

Life in a Drop of Water

by

Sarah Catherine Bagby

Submitted to the Department of Biology
on 22 May 2009 in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Biology

ABSTRACT

The last century of biology brought a revolution to our understanding of life at the molecular level; the last decade, a widening re-evaluation of the claim that understanding gained *in vitro* could reflect the true complexities *in vivo* and *in situ.* I present the results of two projects, one grounded in each strain of biological thought. In the first, I statistically analyze and biochemically map the functional groups making intramolecular interactions that permit activity in an *in vitro*-evolved RNA enzyme, the class I ligase ribozyme; although this ribozyme, derived from random sequence, has never been a part of any organism, understanding its structure and biochemistry is a key step on one of the few relatively well-defined paths to understanding the origins of life. I identify key residues in the ribozyme and present biochemical evidence in support of its proposed catalytic mechanism. In the second study, the interactions at issue are those between an organism, the cyanobacterium *Prochlorococcus*, and its environment, the oligotrophic ocean. *Prochlorococcus* has been found living in very different oxygen regimes in the open ocean; I hypothesized that these different oxygen levels might primarily affect *Prochlorococcus* growth through the competition between oxygen and carbon dioxide for binding to the carbon-fixing enzyme Rubisco. I characterize the transcriptional and growth response of *Prochlorococcus* strain MED4 to limitations on its supply of oxygen and inorganic carbon, finding indications that oxygen contributes to the health of the carbon-limited cell through two photoprotective pathways. I discuss these responses in the context of both the studied responses of *Prochlorococcus* to other extreme environmental stressors and the normal modulations necessary for life amid daily flux.

Thesis Supervisor: Sallie W. Chisholm
Title: Professor of Biology and Lee and Geraldine Martin Professor of Environmental Studies

TABLE OF CONTENTS

I.

In vitro

INTRODUCTION

The problem of how life began is essentially intractable so long as we stipulate that life as it began bore any great resemblance to life as it is. With the benefit of billions of years of evolution, life as it is puts the lie to any notion of the essential efficiency of nature's approach to problem-solving; cycles and epicycles of reaction and regulation permit robust homeostatic responses to a wide array of threats, but do not offer an easy way in for a protocell just emerging from the inorganic world. Replication of all known living organisms requires not just many molecules but many molecules of particular chemistries, sequences, and structures: the DNA genome; the protein enzymes encoded by that genome that unwind and replicate the DNA; and implicitly the informational, adaptor, and catalytic RNAs and the additional proteins required to make the proteins that make the DNA. It is not entirely inconceivable that such a system could have built itself from scratch in one go, but the probabilities involved, to say the least, are not on its side.

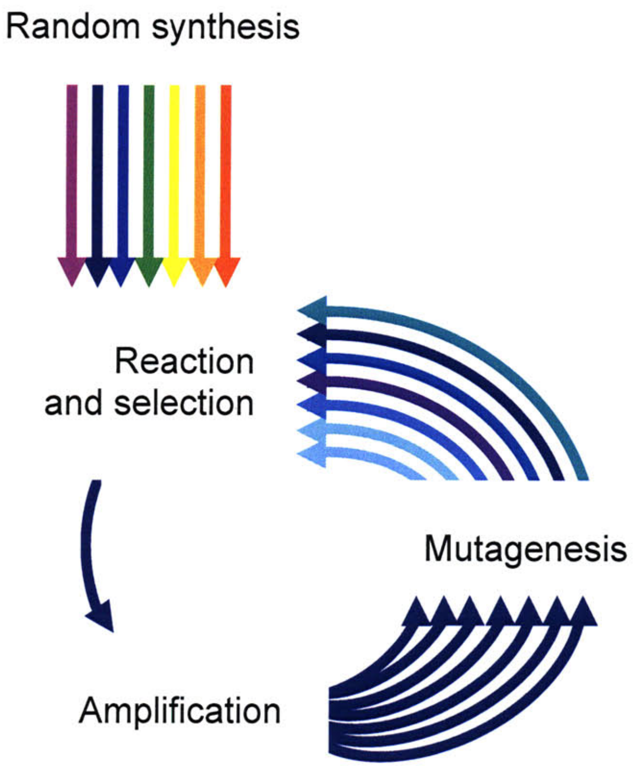
Life's having begun becomes less astronomically absurd if we step back and think about other ways a proto-organism might become self-replicating and subject to selection. (I do not discuss the capacity for homeostasis here; the need for some form of encapsulation is discussed in, *e.g.*, (1).) Such an organism needs a genome, and it needs a genome replicase. One class of macromolecules could play both roles. This simplifying idea entered the discussion in the 1960s, with the proposal that RNA could be that macromolecule coming variously from Alex Rich, Carl Woese, Leslie Orgel, and Francis Crick (2-5). This proposal drew some support from the demonstration by Alex Rich and colleagues that transfer RNA was a highly structured molecule

(6), able to form the sort of tertiary contacts that would let a macromolecule build up a stable active site. Nonetheless, it was not a particularly active area of research until the independent demonstrations by Altman (7) and Cech (8) that RNA was capable of catalysis. Together with the observation that derivatives of RNA act as essential cofactors in many pathways central to metabolism (9, 10), proof of the catalytic capacity of RNA lent considerable clout to the idea of an "RNA world" early in the evolution of life.

But it is not enough for RNA to be able to perform some reaction, any reaction. For the RNA world hypothesis to carry lasting weight, RNA must be shown to be capable of the particular reaction a primitive genome would have relied on: replication. That is, there must be an RNA-dependent RNA polymerase (RdRP) RNA enzyme: an RNA enzyme, or ribozyme, able to bind a template and primer, bind the templated nucleotide triphosphate (NTP), catalyze formation of a new 3',5'-phosphodiester bond between the 3'-hydroxyl group of the primer and the 5' α-phosphate, eject the pyrophosphate leaving group, push or pull the primer-template duplex to move the newly added residue into the attacking position, and repeat this process until the entire template has been copied, first to a reverse complement and then to a daughter strand identical to the parent. At a minimum, the replicase ribozyme must be able to copy a sequence as long as the ribozyme itself, to permit its own passage to the next generation. This is no mean feat.

Ribozyme researchers have used two main approaches to demonstrate such an activity. The first looks to nature: starting with ribozymes that have evolved naturally to perform phosphoryltransferase reactions, can we engineer structural changes that will allow the ribozyme replicase activity to emerge? Following the observation that a shortened version of the *Tetrahymena thermophila* group I self-splicing intron sequence could catalyze oligonucleotide disproportionation reactions (11), Been and Cech demonstrated that the intron could moreover add single nucleotides, activated at the 5'-position with a guanylyl residue rather than a triphosphate, to a growing primer; although this activity did seem to involve a "template-like region" of the group I intron, addition of nucleotides was not templated *per se* (12). Subsequent work introduced templating to the system, although fidelity was still poor (13).

The second approach, starting from random sequence, has so far been more productive. *In vitro* evolution takes as its starting material a very large pool of synthesized RNA molecules of different sequences and attempts to apply a selective biochemical pressure over repeated rounds of amplification and mutagenesis so as to optimize inchoate enzymatic activity and let inactive sequences fall by the wayside. In this system, fitness is simply competence for amplification; access to reverse transcription primers is guarded by some chemical shibboleth (for instance, an affinity tag), and only those RNA sequences capable of catalyzing the desired reaction (for instance, binding themselves covalently to that affinity tag) can pass. With genetic variation introduced among the successful sequences by random mutagenesis, iterating the selective scheme can uncover robust activities from previously unexplored regions of sequence space (Fig. 1). The kernel of a polymerase ribozyme emerged in this way from random sequence in 1993, when Bartel and Szostak obtained a ligase ribozyme that regioselectively forms a phosphodiester bond between the 3'-hydroxyl group of an oligonucleotide substrate and its own 5' α-phosphate (14, 15).



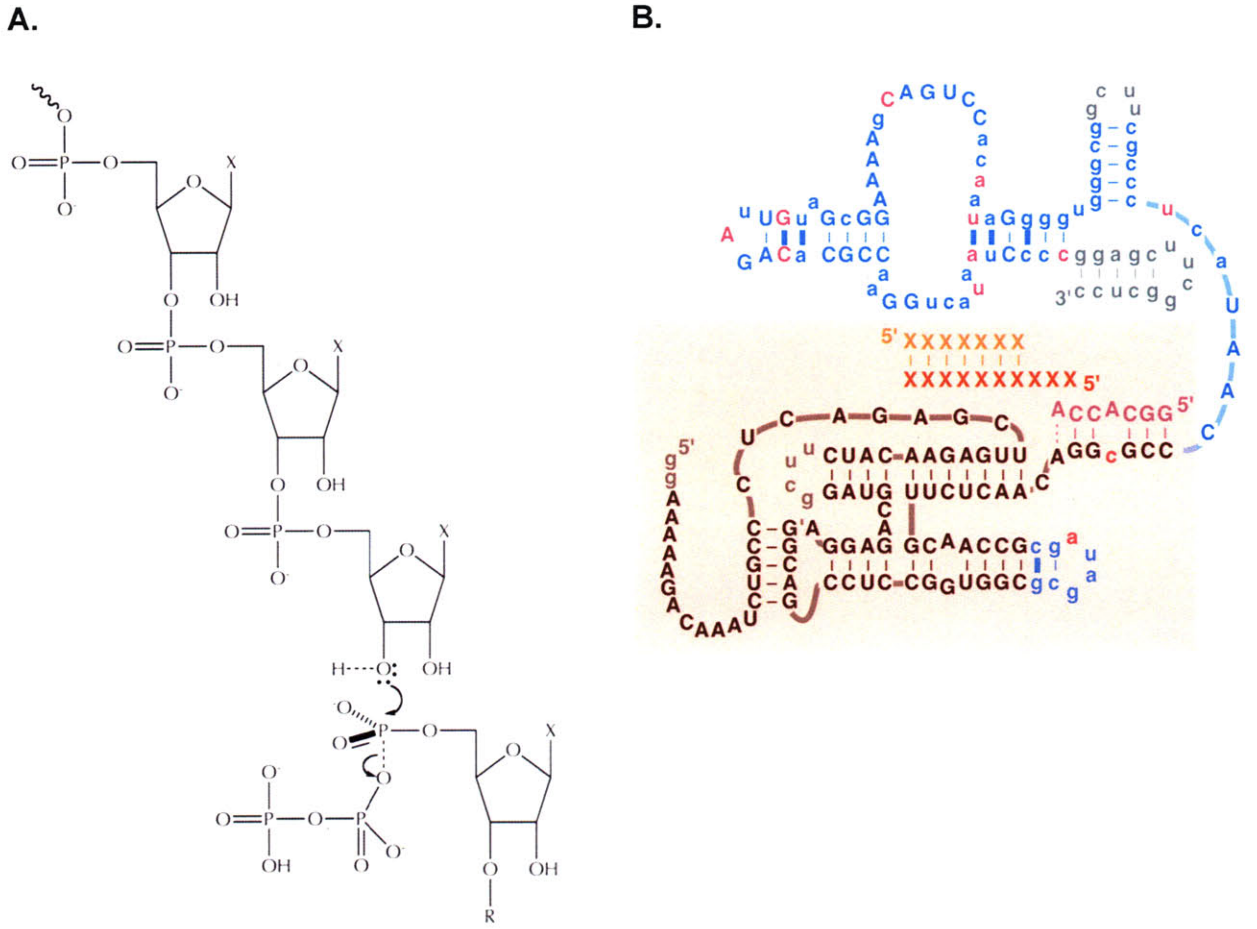
Adapted from (31).

Figure 1. The logic of *in vitro* evolution experiments. Diverse sequences are synthesized and placed under conditions that should permit the desired reaction to take place. It is highly unlikely that the starting pool will contain sequences of outstanding catalytic power, and no more than a handful of sequences are expected to be active at all. Successful members of the pool emerge from the reaction to be amplified by reverse transcription and the polymerase chain reaction; mutagenesis in the latter serves to re-introduce sequence diversity—now, variation in the neighborhood of known active sequences, rather than scattered across sequence space—to the new pool. In a well-designed selection, no pool members that have failed to perform the desired reaction will be competent for amplification, so the population should rapidly come to be dominated by active ribozymes. Periodically, the pool is assayed for detectable activity. When pool improvement approaches saturation, pool sequences are cloned, sequenced, and assayed individually. Adapted from (31).

The first ribozyme to be selected *in vitro* from random sequence, this class I ligase catalyzes a reaction whose chemistry is essentially identical to that of a single turnover of an RNA replicase ribozyme (Fig. 2). The ligase was later engineered to act as a primitive RNA polymerase (16), one with 40-fold better template fidelity than the *Tetrahymena* group I intron-based ribozymes had had. But it was selection, rather than engineering, that led to the next major breakthrough, the development by Johnston *et al.* of a class I ligase-based polymerase ribozyme that could bind to a primer-template duplex in trans and catalyze high-fidelity templated primer extension of up to 14 nucleotides (17) (Fig. 2). Subsequent work with ligase-based selections for an RNA replicase ribozyme have led to a broader diversity of known RNA polymerase ribozymes and to an RNA polymerase ribozyme capable of more than 20 nucleotides to a growing primer strand (18-20).

Since its isolation, the class I ligase ribozyme sequence has been the springboard for a number of other studies of non-natural ribozymes and evolution, being subjected to continuous evolution under approximately constant conditions (21), at changing pH (22), at decreasing Mg^{2+} concentration (23), and in the presence of a "predator" DNA enzyme (24), among other conditions. It was the class I ligase, too, that formed the basis of the pioneering *in vitro* compartmentalization work of Levy *et al.* (25) that yielded the first ribozyme directly selected for multiple-turnover activity. In part, the attraction of the class I ligase as an experimental system is the nature of the reaction it performs; because the single-turnover reaction leaves successful ribozymes covalently tagged with their substrate oligonucleotides, plucking successful

Figure 2. (A) The chemistry of an RNA ligase reaction is identical to the chemistry of an RNA polymerase reaction; the only difference is the identity of the R group borne by the 3' oxygen of the nucleotide that is attacked. In a polymerase reaction, that nucleotide is a free nucleotide triphosphate, with the R group a hydrogen atom; in the ligase reaction, that nucleotide is the 5'-terminal nucleotide triphosphate of an RNA chain, the rest of which constitutes the R group. In both reactions, a 3' oxygen, made a better nucleophile by partial deprotonation, attacks the α-phosphorus of the target nucleotide. The reaction proceeds via an 'in-line' attack, with the attacking nucleophile and the bridging oxygen of the departing pyrophosphate group aligned as the axial ligands of a trigonal bipyramidal transition state. Metal ions (not shown) typically assist with both deprotonation of the nucleophile and stabilization of developing negative charge at the electrophile (32). **(B)** The RNA polymerase ribozyme isolated by (17). The template oligonucleotide is shown in red and the primer in orange. The ribozyme catalyzes serial formation of 3',5'-phosphodiester bonds joining the 3' end of the primer strand to incoming nucleotide triphosphates complementary to the next exposed residue of the template strand. The domain highlighted in light red derives from the class I ligase ribozyme catalytic core; the remainder is an accessory domain, derived during the selection in (17), that is thought to contribute to polymerase binding of nucleotide triphosphates and perhaps the primer-template duplex. Reprinted from (17) (with addition of light red highlighting) with permission from AAAS.



From (17).
Reprinted with permission from AAAS.

ribozymes from a mutagenized pool is relatively straightforward, depending only on the ability of the ligase to perform its reaction despite the affinity tag the substrate bears. And in part, the class I ligase ribozyme is studied because it is fast. At 60 mM Mg^{2+} and pH 8.0, the *cis*-acting construct b1-207 reacts with $k_c = 300$ min^{-1} (26), while the *trans*-acting construct b1-207t reaches $k_c = 375$ min^{-1} for the multiple turnover reaction (27). For the latter construct, k_{cat}/K_M is 7×10^7 M^{-1} min^{-1}, just over an order of magnitude shy of the diffusion-controlled limit.

As fast as class I ligase catalysis is, however, the chemical step remains slower than the folding reaction at pH ≤ 7 (26). Research in the Bartel lab capitalized on this separation of rate-limiting regimes for the folding and chemical steps to target the chemical step for improvement (28, 29). This targeted selection yielded a family of successful clones, many of them surprisingly inured to low concentrations of Mg^{2+}, a cation central to much of RNA biochemistry (30). Ligase variants with low [Mg^{2+}] dependence are of particular interest in the context of the polymerase ribozyme, whose voracious appetite for metal-binding requires of order hundred-millimolar Mg^{2+} in the polymerization reaction—conditions that are a very real threat to the integrity of the ribozyme backbone. Here, I present two studies of class I ligase activity: a statistical analysis of nucleotides contributing to improved ligase activity under the selection conditions and at different concentrations of Mg^{2+}; and several biochemical probes of the functional groups essential for activity in two generations of the class I ligase ribozyme.

WORKS CITED

1. UF Müller (2006). Re-creating an RNA world. *Cell Mol Life Sci* **63:**1278-93.
2. A Rich (1962). On the problems of evolution and biochemical information transfer. In *Horizons in Biochemistry.* M Kasha and B Pullman, eds. New York: Academic Press.
3. CR Woese (1967). *The genetic code: The molecular basis for genetic expression.* Harper & Row.
4. LE Orgel (1968). Evolution of the genetic apparatus. *J Mol Biol* **38:**381-393.
5. FHC Crick (1968). The origin of the genetic code. *J. Mol. Biol* **38:**367-379.
6. SH Kim, GJ Quigley, FL Suddath, A McPherson, D Sneden, JJ Kim, J Weinzierl, and A Rich (1973). Three-dimensional structure of yeast phenylalanine transfer RNA: Folding of the polynucleotide chain. *Science* **179:**285.
7. C Guerrier-Takada, K Gardiner, T Marsh, N Pace, and S Altman (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35:**849-57.
8. K Kruger, PJ Grabowski, AJ Zaug, J Sands, DE Gottschling, and TR Cech (1982). Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31:**147-57.
9. HB White (1976). Coenzymes as fossils of an earlier metabolic state. *J Mol Evol* **7:**101-4.
10. SA Benner, AD Ellington, and A Tauer (1989). Modern metabolism as a palimpsest of the RNA world. *Proc Natl Acad Sci U S A* **86:**7054-8.
11. AJ Zaug and TR Cech (1986). The intervening sequence RNA of *Tetrahymena* is an enzyme. *Science* **231:**470.

12. MD Been and TR Cech (1988). RNA as an RNA polymerase: Net elongation of an RNA primer catalyzed by the *Tetrahymena* ribozyme. *Science* **239:**1412-6.

13. DP Bartel, JA Doudna, N Usman, and JW Szostak (1991). Template-directed primer extension catalyzed by the *Tetrahymena* ribozyme. *Mol Cell Biol* **11:**3390-4.

14. DP Bartel and JW Szostak (1993). Isolation of new ribozymes from a large pool of random sequences. *Science* **261:**1411-8.

15. EH Ekland, JW Szostak, and DP Bartel (1995). Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* **269:**364-370.

16. EH Ekland and DP Bartel (1996). RNA-catalysed RNA polymerization using nucleoside triphosphates. *Nature* **382:**373-6.

17. WK Johnston, PJ Unrau, MS Lawrence, ME Glasner, and DP Bartel (2001). RNA-catalyzed RNA polymerization: Accurate and general RNA-templated primer extension. *Science* **292:**1319-25.

18. KE McGinness, MC Wright, and GF Joyce (2002). Continuous *in vitro* evolution of a ribozyme that catalyzes three successive nucleotidyl addition reactions. *Chem Biol* **9:**585-96.

19. MS Lawrence and DP Bartel (2005). New ligase-derived RNA polymerase ribozymes. *RNA* **11:**1173-80.

20. HS Zaher and PJ Unrau (2007). Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *RNA* **13:**1017-26.

21. MC Wright and GF Joyce (1997). Continuous *in vitro* evolution of catalytic function. *Science* **276:**614-7.

22. H Kühne and GF Joyce (2003). Continuous *in vitro* evolution of ribozymes that operate under conditions of extreme pH. *J Mol Evol* **57:**292-8.

23. T Schmitt and N Lehman (1999). Non-unity molecular heritability demonstrated by continuous evolution *in vitro*. *Chem Biol* **6:**857-69.

24. P Ordoukhanian and GF Joyce (1999). A molecular description of the evolution of resistance. *Chem Biol* **6:**881-9.

25. M Levy, KE Griswold, and AD Ellington (2005). Direct selection of trans-acting ligase ribozymes by *in vitro* compartmentalization. *RNA* **11:**1555-62.

26. ME Glasner, NH Bergman, and DP Bartel (2002). Metal ion requirements for structure and catalysis of an RNA ligase ribozyme. *Biochemistry* **41:**8103-12.

27. NH Bergman, WK Johnston, and DP Bartel (2000). Kinetic framework for ligation by an efficient RNA ligase ribozyme. *Biochemistry* **39:**3115-23.

28. NH Bergman (2001). *The reaction kinetics and three-dimensional architecture of a catalytic RNA.* MIT Ph.D. Thesis.

29. CC Yen (2002). *Optimization and characterization of the class I RNA ligase.* MIT M.S. Thesis.

30. AL Feig, OC Uhlenbeck (1999). The role of metal ions in RNA biochemistry. In *The RNA World.* RF Gesteland, TR Cech, and JF Atkins, eds. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

31. DP Bartel (1999). Re-creating an RNA replicase. In *The RNA World.* RF Gesteland, TR Cech, and JF Atkins, eds. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

32. TA Steitz and JA Steitz (1993). A general two-metal-ion mechanism for catalytic RNA. *Proc Natl Acad Sci U S A* **90:**6498-502.

Analysis of Sequence Conservation and Kinetic Variation Uncovered in Selection of an Improved Class I Ligase Ribozyme

This work is an analysis of an *in vitro* selection performed by Nicholas H. Bergman and Catherine Yen and of kinetic characterization of the selected clones by Catherine Yen. My contribution is exclusively analytical. I adopted the Monte Carlo approach to rate-constant analysis on the strength of useful conversations with Steven C. Wang of Williams College. In addition, this work benefits greatly from comparison of the sequence analysis to the class I ligase crystal structure solved in the Bartel lab by David M. Shechner and Robert A. Grant. Support for research in the Bartel lab comes from the National Institutes of Health and the Howard Hughes Medical Institute, of which David Bartel is an Investigator and I am a Pre-doctoral Fellow.

ABSTRACT

The class I ligase is the first RNA enzyme to have been derived from random sequences by *in vitro* selection, and it remains one of the fastest, achieving rate enhancements comparable to those of some protein enzymes (1). Further improvements to this ribozyme arose from a selection that targeted the kinetics of the chemical step of the ligase reaction (2). The improved ligase, the most successful of the clones to emerge from selection, not only achieves further rate enhancements under the selective conditions but also shows a sharp reduction in [Mg^{2+}]-dependence compared to the parent ribozyme. Analysis of the sequences and kinetics of successful clones suggests which mutations play the greatest part in these improvements. Comparison of these results to the recently solved crystal structure of the improved class I ligase ribozyme (3) suggests the structural logic behind the effects of many of these mutations, and the concordance of these lines of evidence indicates that the crystal structure is likely to represent an active conformation of the ligase ribozyme.

ABBREVIATIONS

Nt, nucleotide; P, helical RNA stem; J, joining region; L, loop region.

INTRODUCTION

Selection experiments face a trade-off between breadth and depth, the number of positions varied and the completeness with which sequence space can be covered at those positions. Previous selection experiments with the class I ligase ribozyme indicated that many of the unpaired residues within the ligase were important for activity, and suggested that single-stranded regions may play a role in defining the tertiary structure and active site of the ribozyme (4). Thus, the new selection turned away from the known stem regions and sought instead to explore sequence space deeply in the single-stranded stretches of the ligase.

Where conservation in previous selections suggested that the optimal residue in a joining or loop region was already in place, the pool was biased towards that residue but mutagenized at the level of 10% (Fig. 1), i.e., 90% of pool molecules contained the parental nucleotide, and 3.3% contained each of the other nucleotides. At the remaining positions in these regions, the pool was randomized completely in both length and sequence. Only one basepair in the ribosome's seven stems was allowed to vary: both bases in the non-canonical G88:A103 pair in stem P7 were mutagenized at the 10% level. Two basepairs were changed outright across the pool: G73:C84 was converted to an A:U pair, a change that improves class I-catalyzed primer extension (5), and helix P5 was extended by one basepair to improve its stability, in order to stabilize the background against which L5 optimization would take place. From this pool, Bergman and Yen selected specifically for rapid catalysis of the ligation reaction by performing the selective step in later rounds in a rapid-quench flow apparatus under increasingly stringent conditions. This apparatus enabled selection for ribozymes capable of performing ligation in as little as 0.2 s.

Trimmed of their reverse transcription primer-binding sites and presented with 5'-truncated all-RNA substrates, the 35 unique successful clones (Fig. 1) were next measured to determine what improvement over the parent ligase ribozyme their catalytic performance represented. Initial measurements of ligation activity were performed under the conditions of the final round of selection (pH 6.0, 10 mM Mg^{2+}, 200 mM KCl), revealing ligation rate constants ranging from 0.15 min^{-1} to over 2 min^{-1} (Table 1). The seven clones (101, 22, 96, 23, 91, 80, and 141) clustered at the top of this range performed ligation roughly twice as fast as did the parent ribozyme under these conditions.

Next, Bergman and Yen asked whether having performed the selection at Mg^{2+} levels known to be below saturation for the parent ligase had produced any change in Mg^{2+} dependence among the successful clones. The parent ligase has a steep dependence on [Mg^{2+}] in this range, its rate constant dropping 40-fold when the Mg^{2+} concentration is reduced from 10 mM to 1 mM. Among the selected clones, five show as steep a dependence as that of the parent; the remainder range shallower, with three clones (187, 66, and 23) dropping as little as 3.5–5.5-fold. Clone 23 was designated the 'improved' class I ligase and used for crystallographic studies of the ribozyme. I report the results of statistical analyses of the sequences and rate constants of all isolates, one method of shedding light on the changes that led to improved activity in general and reduced metal dependence in particular.

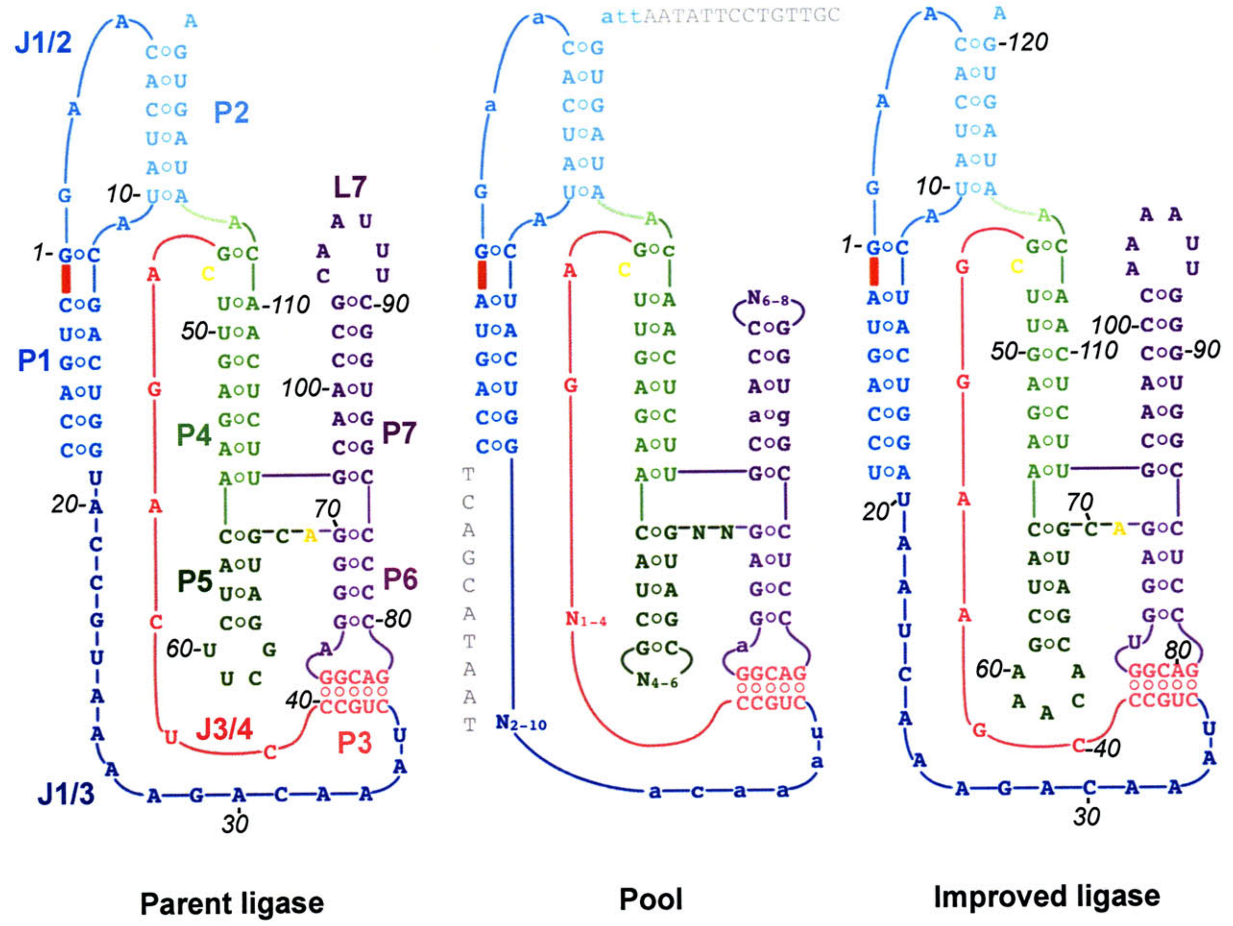


Figure 1. Secondary structures of the parent ligase, the pool of sequences used for selection, and the improved ligase (clone 23). Secondary structural elements are labeled on the parent ligase. Joining regions are named for the two paired regions they connect; thus, nucleotides 19–34, joining P1 and P3, constitute J1/3. The heavy red line indicates the ligation junction between the ribozyme 5' end and the oligonucleotide substrate; residue numbering begins at the 5' end of the ribozyme itself, i.e., at the 3' nucleotide of the ligation junction. Numbering given in the text is with reference to the improved ligase unless otherwise noted. Positions randomized in the pool are labeled N on the pool secondary structure; positions mutagenized at the 10% level are labeled in lowercase. Sequences in gray on the pool secondary structure are, at left, the DNA portion of the substrate used for selection and, at top, the reverse transcription primer binding site.

Table 1. Rate constants for the 35 isolated ligase variants, measured at pH 6.0 in 10 mM and 1 mM Mg^{2+}. For details of kinetic measurements, see references 2 and 9.

Clone	$k_{10\,mM}$ (min^{-1})	$k_{1\,mM}$ (min^{-1})
23	2.15	0.40
101	2.50	0.28
96	2.40	0.25
22	2.25	0.26
91	2.20	0.13
80	2.15	0.21
141	2.05	0.11
159	1.70	0.057
84	1.65	0.13
173	1.55	0.13
178	1.50	0.06
50	1.50	0.11
162	1.45	0.074
69	1.45	0.096
55	1.40	0.069
89	1.40	0.108
70	1.35	0.12
71	1.30	0.13
175	1.20	0.029
68	1.20	0.067
66	1.10	0.32
2	1.10	0.034
106	1.05	0.032
18	1.00	0.023
172	0.80	0.031
77	0.80	0.020
180	0.65	0.029
1	0.60	0.016
186	0.55	0.036
124	0.55	0.036
153	0.50	0.058
35	0.40	0.010
61	0.40	0.013
187	0.35	0.10
158	0.15	0.005

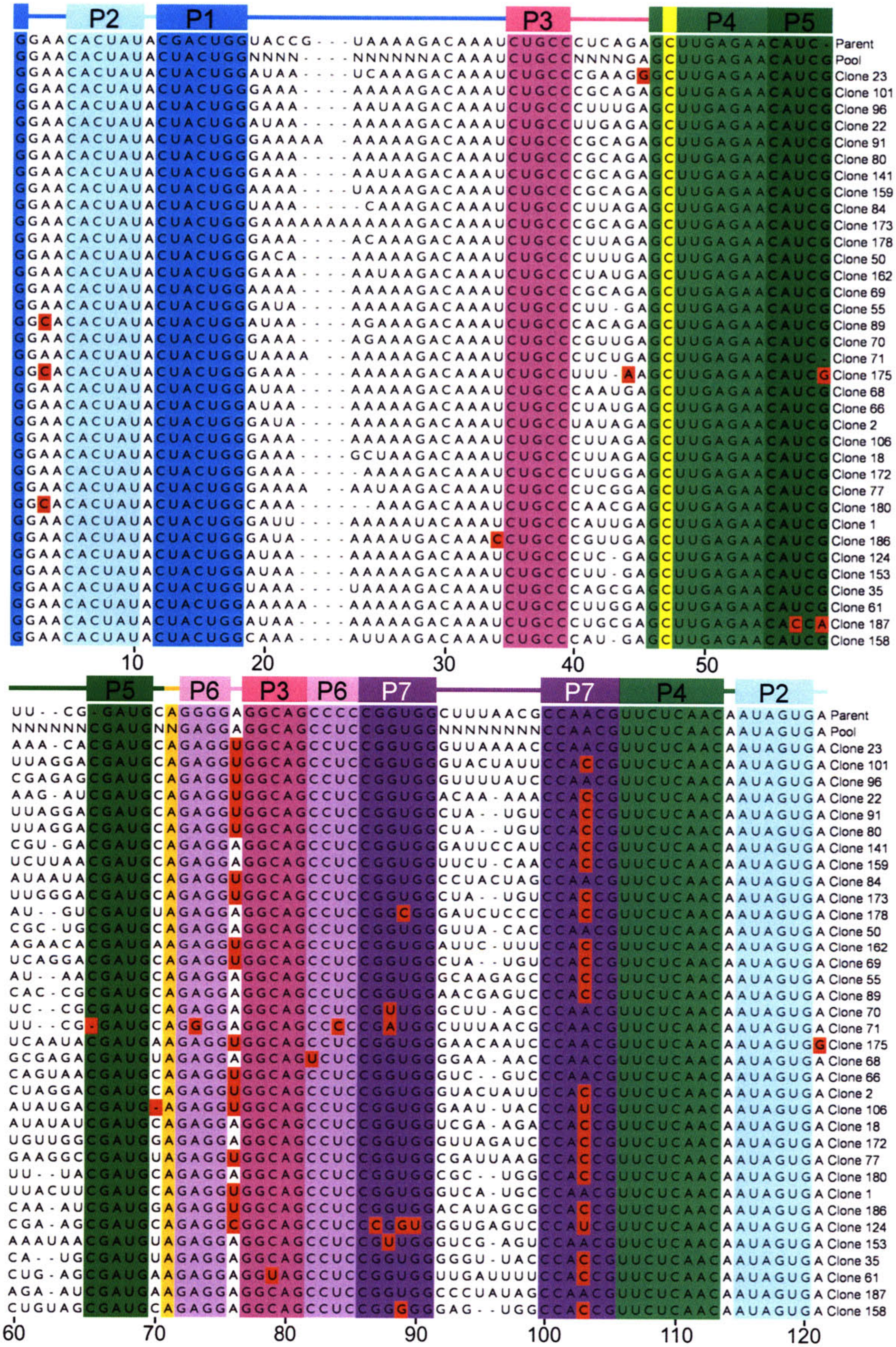
RESULTS AND DISCUSSION

Sequence analysis of improved class I ligase variants

We aligned the sequences of all isolates and compared the nucleotide distribution at each mutagenized position to that of the starting pool (Fig. 2, 3B). The first fully randomized region of the ribozyme, positions 19–28, lies in a long joining region, J1/3. Although this segment was varied randomly in the pool from 2 to 10 nt in length, the successful isolates were tightly clustered about 10 nt (Fig. 3C), with five clones even acquiring additional nucleotides in the course of evolution. Notably, this is one nucleotide shorter than the parent J1/3; it is not clear whether this change is connected to the 1-nt extension of P1 the modified ligase substrate makes possible. A strong nucleotide composition bias was also evident in this region of the isolates, with adenosine constituting 74% of the nucleotides in this region, compared to 50% in this region of the parent ligase. The structural underpinnings of the fitness advantage accruing to isolates with adenosine-rich J1/3 sequences are discussed in chapter 2. Remarkably, the *least* significant nucleotide enrichment in this region was the preference for G at position 19 with $p = 0.011$ (Fisher's exact test); positions 20–27 were all preferentially an A with p values ranging from 4×10^{-5} to 4×10^{-12}, and position 28 was preferentially a G with $p = 9 \times 10^{-11}$.

The 5' end of the joining segment J3/4 showed somewhat less sequence bias but a clear length preference; the starting pool varied from 1 to 4 nt here, but 30 of 35 successful clones maintained the parental length, 4 nt, and in the remaining 5 clones J3/4 was shortened only to 3 nt. The 5' nucleotide of J3/4, position 40, was strongly conserved as the parental C ($p = 3 \times 10^{-9}$) and nucleotide 41 was largely conserved as the parental U ($p = 0.004$), but the 3' positions

Figure 2. Alignment of the 35 ligase variants isolated by Bergman and Yen. Helix P1 is formed when the substrate oligonucleotide hybridizes to nucleotides 13–18. Red blocks highlight non-parental sequence in regions that were not fully randomized (i.e., in helices and at positions mutagenized at the 10% level) in the starting pool. Colors are as in Figure 1. Sequence numbering below the alignment is with respect to clone 23, the improved class I ligase. Adapted from (2).



```
GGAACACUAUACGACUGGUACCG---UAAAAGACAAAUCUGCCCUCAGAGCUUGAGAACAUC- Parent
GGAACACUAUACUACUGGNNNN----NNNNNNACAAAUCUGCCNNNNGAGCUUGAGAACAUCG Pool
GGAACACUAUACUACUGGAUAA----UCAAAGACAAAUCUGCCCGAAGGGCUUGAGAACAUCG Clone 23
GGAACACUAUACUACUGGGAAA----AAAAAGACAAAUCUGCCCGCAGAGCUUGAGAACAUCG Clone 101
GGAACACUAUACUACUGGGAAA----AAUAAGACAAAUCUGCCCUUUGAGCUUGAGAACAUCG Clone 96
GGAACACUAUACUACUGGAUAA----AAAAAGACAAAUCUGCCUUGAGAGCUUGAGAACAUCG Clone 22
GGAACACUAUACUACUGGGAAAAA -AAAAAGACAAAUCUGCCCGCAGAGCUUGAGAACAUCG Clone 91
GGAACACUAUACUACUGGGAAA----AAAAAGACAAAUCUGCCCGCAGAGCUUGAGAACAUCG Clone 80
GGAACACUAUACUACUGGGAAA----AAUAAGACAAAUCUGCCCGCAGAGCUUGAGAACAUCG Clone 141
GGAACACUAUACUACUGGAAAA----UAAAAGACAAAUCUGCCCGCAGAGCUUGAGAACAUCG Clone 159
GGAACACUAUACUACUGGUAAA-----CAAAGACAAAUCUGCCCUUAGAGCUUGAGAACAUCG Clone 84
GGAACACUAUACUACUGGGAAAAAAAAAAAAGACAAAUCUGCCCGCAGAGCUUGAGAACAUCG Clone 173
GGAACACUAUACUACUGGGAAA----ACAAAGACAAAUCUGCCCUUAGAGCUUGAGAACAUCG Clone 178
GGAACACUAUACUACUGGGACA----AAAAAGACAAAUCUGCCCUUUGAGCUUGAGAACAUCG Clone 50
GGAACACUAUACUACUGGGAAA----AAUAAGACAAAUCUGCCCUAUGAGCUUGAGAACAUCG Clone 162
GGAACACUAUACUACUGGAAAA----AAAAAGACAAAUCUGCCCGCAGAGCUUGAGAACAUCG Clone 69
GGAACACUAUACUACUGGGAUA----AAAAAGACAAAUCUGCCCUU-GAGCUUGAGAACAUCG Clone 55
GGCACACUAUACUACUGGAUAA----AGAAAGACAAAUCUGCCCACAGAGCUUGAGAACAUCG Clone 89
GGAACACUAUACUACUGGGAAA----AGAAAGACAAAUCUGCCCGUUGAGCUUGAGAACAUCG Clone 70
GGAACACUAUACUACUGGUAAAA --AAAAAGACAAAUCUGCCCUCUGAGCUUGAGAACAUC- Clone 71
GGCACACUAUACUACUGGGAAA----AAAAAGACAAAUCUGCCUUU-AAGCUUGAGAACAUCG Clone 175
GGAACACUAUACUACUGGAUAA----AAAAAGACAAAUCUGCCCAAUGAGCUUGAGAACAUCG Clone 68
GGAACACUAUACUACUGGAUAA----AAAAAGACAAAUCUGCCCUAUGAGCUUGAGAACAUCG Clone 66
GGAACACUAUACUACUGGGAUA----AAAAAGACAAAUCUGCCUAUAGAGCUUGAGAACAUCG Clone 2
GGAACACUAUACUACUGGGAAA----AAAAAGACAAAUCUGCCCUUAGAGCUUGAGAACAUCG Clone 106
GGAACACUAUACUACUGGGAAA----GCUAAGACAAAUCUGCCCUUAGAGCUUGAGAACAUCG Clone 18
GGAACACUAUACUACUGGGAAA-----AAAAGACAAAUCUGCCCUUGGAGCUUGAGAACAUCG Clone 172
GGAACACUAUACUACUGGGAAAA --AAUAAGACAAAUCUGCCCUCGGAGCUUGAGAACAUCG Clone 77
GGCACACUAUACUACUGGCAAA------AAAGACAAAUCUGCCCAACGAGCUUGAGAACAUCG Clone 180
GGAACACUAUACUACUGGGAUU----AAAAAUACAAAUCUGCCCAUUGAGCUUGAGAACAUCG Clone 1
GGAACACUAUACUACUGGGAUA----AAAAUGACAAACCUGCCCGUUGAGCUUGAGAACAUCG Clone 186
GGAACACUAUACUACUGGAUAA----AAAAAGACAAAUCUGCCCUC-GAGCUUGAGAACAUCG Clone 124
GGAACACUAUACUACUGGAUAA----AAAAAGACAAAUCUGCCCUU-GAGCUUGAGAACAUCG Clone 153
GGAACACUAUACUACUGGUAAA----UAAAAGACAAAUCUGCCCAGCGAGCUUGAGAACAUCG Clone 35
GGAACACUAUACUACUGGAAAAA---AAAAAGACAAAUCUGCCCUUGGAGCUUGAGAACAUCG Clone 61
GGAACACUAUACUACUGGAUAA----AAAAAGACAAAUCUGCCCUGCGAGCUUGAGAACACCA Clone 187
GGAACACUAUACUACUGGCAAA----AUUAAGACAAAUCUGCCCAU-GAGCUUGAGAACAUCG Clone 158
```

```
UU--CG-GAUGCAGGGGAGGCAGCCCCCGGUGGCUUUAACGCCAACGUUCUCAACAAUAGUGA Parent
NNNNNNCGAUGNNGAGGAGGCAGCCUCCGGUGGNNNNNNNNCCAACGUUCUCAACAAUAGUGA Pool
AAA-CACGAUGCAGAGGUGGCAGCCUCCGGUGGGUUAAAACCCAACGUUCUCAACAAUAGUGA Clone 23
UUAGGACGAUGCAGAGGUGGCAGCCUCCGGUGGGUACUAUUCCACCGUUCUCAACAAUAGUGA Clone 101
CGAGAGCGAUGCAGAGGUGGCAGCCUCCGGUGGGUUUUAUCCCAACGUUCUCAACAAUAGUGA Clone 96
AAG-AUCGAUGCAGAGGUGGCAGCCUCCGGUGGACAA-AAACCACCGUUCUCAACAAUAGUGA Clone 22
UUAGGACGAUGCAGAGGUGGCAGCCUCCGGUGGCUA--UGUCCACCGUUCUCAACAAUAGUGA Clone 91
UUAGGACGAUGCAGAGGUGGCAGCCUCCGGUGGCUA--UGUCCACCGUUCUCAACAAUAGUGA Clone 80
CGU-GACGAUGCAGAGGAGGCAGCCUCCGGUGGGAUUCCAUCCACCGUUCUCAACAAUAGUGA Clone 141
UCUUAACGAUGCAGAGGAGGCAGCCUCCGGUGGUUCU-CAACCACCGUUCUCAACAAUAGUGA Clone 159
AUAAUACGAUGCAGAGGUGGCAGCCUCCGGUGGCCUACUAGCCAACGUUCUCAACAAUAGUGA Clone 84
UUGGGACGAUGCAGAGGUGGCAGCCUCCGGUGGCUA--UGUCCACCGUUCUCAACAAUAGUGA Clone 173
AU--GUCGAUGUAGAGGAGGCAGCCUCCGGCGGGAUCUCCCCCACCGUUCUCAACAAUAGUGA Clone 178
CGC-UGCGAUGCAGAGGAGGCAGCCUCCGGUGGGUUA-CACCCAACGUUCUCAACAAUAGUGA Clone 50
AGAACACGAUGAAGAGGUGGCAGCCUCCGGUGGAUUC-UUUCCACCGUUCUCAACAAUAGUGA Clone 162
UCAGGACGAUGCAGAGGUGGCAGCCUCCGGUGGCUA--UGUCCACCGUUCUCAACAAUAGUGA Clone 69
AU--AACGAUGCAGAGGAGGCAGCCUCCGGUGGGCAAGAGCCCACCGUUCUCAACAAUAGUGA Clone 55
CAC-CGCGAUGCAGAGGAGGCAGCCUCCGGUGGAACGAGUCCCACCGUUCUCAACAAUAGUGA Clone 89
UC--CGCGAUGCAGAGGAGGCAGCCUCCGUUGGGCUU-AGCCCAACGUUCUCAACAAUAGUGA Clone 70
UU--CG-GAUGCAGGGGAGGCAGCCCCCGAUGGCUUUAACGCCAACGUUCUCAACAAUAGUGA Clone 71
UCAAUACGAUGAAGAGGUGGCAGCCUCCGGUGGGAACAAUCCCAACGUUCUCAACAAUAGUGG Clone 175
GCGAGACGAUGUAGAGGAGGCAGUCUCCGGUGGGGAA-AACCCAACGUUCUCAACAAUAGUGA Clone 68
CAGUAACGAUGCAGAGGUGGCAGCCUCCGGUGGGUC--GUCCCAACGUUCUCAACAAUAGUGA Clone 66
CUAGGACGAUGCAGAGGUGGCAGCCUCCGGUGGGUACUAUUCCACCGUUCUCAACAAUAGUGA Clone 2
AUAUGACGAUG-AGAGGUGGCAGCCUCCGGUGGGAAU-UACCCAUCGUUCUCAACAAUAGUGA Clone 106
AUAUAUCGAUGCAGAGGAGGCAGCCUCCGGUGGUCGA-AGACCACCGUUCUCAACAAUAGUGA Clone 18
UGUUGGCGAUGGAGAGGAGGCAGCCUCCGGUGGGUUAGAUCCCACCGUUCUCAACAAUAGUGA Clone 172
GAAGGCCGAUGUAGAGGUGGCAGCCUCCGGUGGCGAUUAAGCCACCGUUCUCAACAAUAGUGA Clone 77
UU--UACGAUGCAGAGGAGGCAGCCUCCGGUGGCGC--UGGCCACCGUUCUCAACAAUAGUGA Clone 180
UUACUUCGAUGCAGAGGUGGCAGCCUCCGGUGGGUCA-UGCCCAACGUUCUCAACAAUAGUGA Clone 1
CAA-AUCGAUGGAGAGGUGGCAGCCUCCGGUGGACAUAGCGCCACCGUUCUCAACAAUAGUGA Clone 186
CGA-AGCGAUGCAGAGGCGGCAGCCUCCCGGUGGGUGAGUCCCAUCGUUCUCAACAAUAGUGA Clone 124
AAAUAACGAUGUAGAGGAGGCAGCCUCCGUUGGGUCG-AGUCCAACGUUCUCAACAAUAGUGA Clone 153
CA--UGCGAUGUAGAGGAGGCAGCCUCCGGUGGGGGU-UACCCACCGUUCUCAACAAUAGUGA Clone 35
CUG-AGCGAUGAAGAGGAGGUAGCCUCCGGUGGUUGAUUUUCCACCGUUCUCAACAAUAGUGA Clone 61
AGA-AUCGAUGAAGAGGAGGCAGCCUCCGGUGGCCCUAUAGCCAACGUUCUCAACAAUAGUGA Clone 187
CUGUAGCGAUGCAGAGGAGGCAGCCUCCGGGGGGAG--UGGCCACCGUUCUCAACAAUAGUGA Clone 158
```

(nucleotides 42 and 43) showed no significant trends. The final joining segment to be randomized, J5/6, hewed to the parent ribozyme in both sequence and length. J5/6 is 2 nt long in the parent ligase and in 34 of 35 isolates, and 1 nt in the remaining isolate. Even that shortened variant maintains A71, which was absolutely conserved among successful clones ($p = 4 \times 10^{-12}$). The 5' position of J5/6, though not absolutely conserved, showed signifiant selection for the parental C ($p = 0.003$).

In contrast with the joining regions, the two terminal loops, L5 and L7, showed little conservation in either size or sequence, consistent with previous observations that perturbing these loops has little impact on ligase activity (6, 7). Only three positions in these two loops showed evidence of selection: nucleotides 62 and 64 in L5, and nucleotide 92 in L7. (Note that the improved ligase, clone 23, has a 5-nt L5, while 20 other clones have 6 nt here, with the inserted nucleotide falling between improved ligase nucleotides 62 and 64.) Adenosine was favored at both L5 sites, and guanosine at the L7 site; in none of these cases was the favored nucleotide selected to the exclusion of any other. In loops L5 and L7 alike, the first and last nucleotides of loops of the largest allowable size were often (in 12 of the 20 6-nt L5 sequences and 14 of the 18 8-nt L7 sequences) Watson-Crick or wobble complements, such that the paired stem could be extended by one basepair and the loop decreased by 2 nt. It is possible that the selective enrichment of G at position 92 is tied to a preference for extending P7 in this way; in all 19 clones bearing G92, the final nucleotide of L7 is either C or U. By contrast, the remaining 16 clones are evenly divided among those with potential for an 'extra' P7 pair and those without ($p = 5 \times 10^{-4}$).

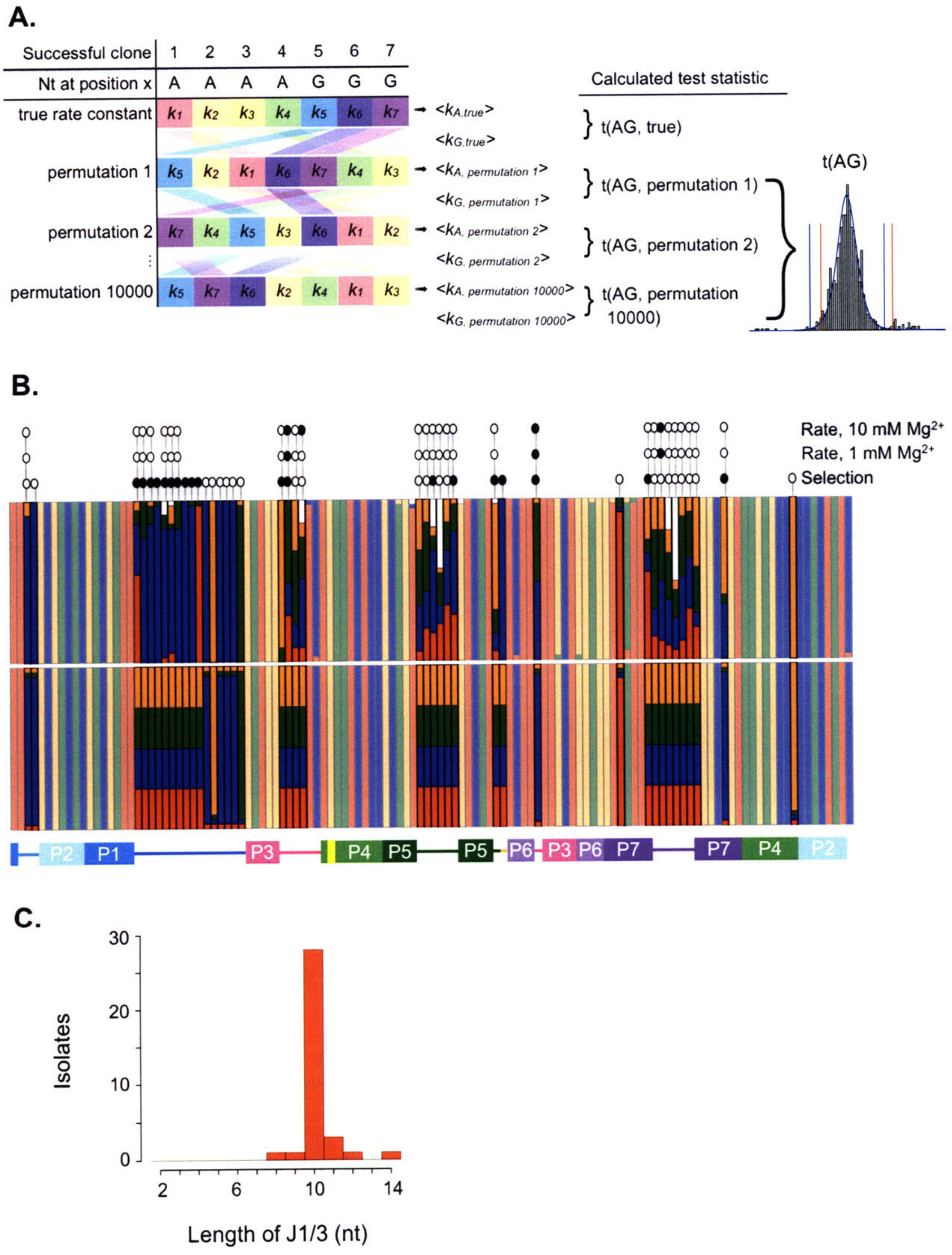
Among the 19 positions mutagenized at the 10% level, two showed significant movement away from the parental identity. The first of these, nucleotide 76, is the sole nucleotide of J6/3; an A in the parent ligase and in 90% of the pool, it remained an A in just 17 isolates, becoming a C in one and a U in the other 17 ($p = 2 \times 10^{-6}$, Fisher's exact test). The second such position, nucleotide 103, is part of the non-canonical G88:A103 basepair in helix P7. Remaining a G:A pair in 10 isolates, mutation of nucleotide 103 gave a G:U wobble in 2 isolates and a canonical G:C pair in 21 ($p = 1 \times 10^{-8}$). One clone lost the G:A mismatch only to gain an A:A mismatch by mutation of G88. Although G88, like A103, was also mutagenized at the 10% level, conversion of the G88:A103 mismatch to a U:A pair was observed only twice, suggesting that the selective advantage of canonical pairing at this site may be tied to the stability of the basepair.

Although Fisher's exact test can readily detect bias towards non-parental nucleotides at lightly mutagenized positions, the test lacks the statistical power to detect significant conservation of the parental nucleotides at such sites; given expected frequencies of 32 parental and 1 each non-parental nucleotide at a given position, the probability of observing the parental nucleotide in all 35 isolates is 0.239. The two positions discussed above, nucleotides 76 and 103, account for nearly all of the substitutions and gaps observed among the 19 lightly mutagenized sites; setting these two positions aside and considering the other 17 positions in the aggregate, we observed only 10 non-parental features in all 35 isolates. We performed one million simulated selections at 19 lightly mutagenized sites, discarding from each simulation the two positions with the most non-parental features. On average, the 17 sites remaining in our simulations contained 53 ± 7 non-parental features, significantly more than the 10 observed in our *in vitro* isolates ($p < 1 \times 10^{-6}$). Thus, the parental sequence appears to have been optimal at most of these 17 positions.

Connecting kinetic measurements to sequence variation

We next examined whether any mutations were specifically associated with rapid catalysis at 1 mM or 10 mM Mg^{2+}: at a given position, despite the changing background of sequence variation at different positions, do ribozymes bearing an A (for example) have on average a higher mean rate constant than those bearing a G? Although a t-statistic could be calculated for any two such means, the group sizes and variances in our dataset are frequently unequal, and we could not safely assume that the underlying distribution of t-statistics would closely resemble the canonical t-distribution, and thus we did not *a priori* know the critical values for our test statistic. We therefore took a Monte Carlo approach to determine the true distribution of t statistics for our data (Fig. 3A; Methods). Starting with the measured rate constants and known group sizes (e.g., at position 64, 12 isolates carry a G, 12 an A, 6 a U, and 5 a C), we randomly re-assigned rate constants to different groups. We could then calculate the mean rate constant in the "G" group and the mean in the "A" group, and from these mean values the t-statistic describing the difference between them. By repeating this process 10000 times, we generated t-statistics with a distribution approaching that characteristic of our data. This distribution allowed us to determine the probability that the means of two random subsets of ligase rate constants would give a t-statistic as extreme as that obtained from the true groupings.

Figure 3. **(A)** Outline of Monte Carlo analysis of the kinetic effects of different nucleotides at each position in the ribozyme. At each position subjected to analysis, the pairings of nucleotide identity and rate constant were shuffled randomly 10000 times and the mean rate constants newly associated with each nucleotide were calculated. The t statistic describing the difference in mean rate constants of ribozymes bearing, e.g., A and G residues is calculated for each permutation, revealing the underlying t distribution and the critical values to which the true t statistic can be compared. Note that, whereas the canonical t-distribution (blue curve) has symmetric tails and thus symmetric critical values (blue vertical lines), the Monte Carlo simulation can reveal a t-distribution with markedly asymmetric tails and critical values (red lines). **(B)** Observed (top) and expected (bottom) nucleotide frequencies in the ligase selection. Red, G; blue, A; green, U; orange, C; white, gap. Less saturated colors mark positions that were not deliberately varied in the pool. Above, the results of statistical analysis: Monte Carlo analysis of nucleotide identity effects on ribozyme kinetics at 10 mM and 1 mM Mg^{2+}, and Fisher's exact test to detect significant deviation of observed from expected nucleotide frequencies. Open ovals indicate that a test was performed but revealed no significant effect; filled ovals indicate significant effects. Colors in the secondary-structure schematic below are as in Figure 1. **(C)** Histogram of the observed lengths of J1/3 sequences among selected ligase variants. J1/3 was varied from 2 to 10 nt in the starting pool; how some variants acquired longer J1/3 sequences is unknown.


A.
Successful clone
1
2
3
4
5
6
7
Nt at position x
A
A
A
A
G
G
G
true rate constant
k_1 k_2 k_3 k_4 k_5 k_6 k_7
permutation 1
k_5 k_2 k_1 k_6 k_7 k_4 k_3
permutation 2
k_7 k_4 k_5 k_3 k_6 k_1 k_2
permutation 10000
k_5 k_7 k_6 k_2 k_4 k_1 k_3
$<k_{A,true}>$
$<k_{G,true}>$
$<k_{A,\ permutation\ 1}>$
$<k_{G,\ permutation\ 1}>$
$<k_{A,\ permutation\ 2}>$
$<k_{G,\ permutation\ 2}>$
$<k_{A,\ permutation\ 10000}>$
$<k_{G,\ permutation\ 10000}>$
Calculated test statistic
t(AG, true)
t(AG, permutation 1)
t(AG, permutation 2)
t(AG, permutation 10000)
t(AG)
B.
Rate, 10 mM Mg^{2+}
Rate, 1 mM Mg^{2+}
Selection
P2
P1
P3
P4
P5
P5
P6
P3
P6
P7
P7
P4
P2
C.
Isolates
30
20
10
0
2
6
10
14
Length of J1/3 (nt)

With this approach, we found several positions in the ligase at which nucleotide identity has a significant effect on rate. Two of these positions, 41 and 43, are in J3/4. Isolates bearing a G at position 41 have significantly faster rate constants at 1 mM Mg^{2+} than do isolates bearing an A (Bonferroni-adjusted $p < 0.018$), and at 10 mM Mg^{2+} than isolates bearing either an A or a U (Bonferroni-adjusted $p < 0.0066$ and 0.041, respectively). Surprisingly, although we did observe significant selection at nucleotide 41, the favored residue was U (8.75 U41 clones and 8.75 G41 clones expected; 18 U41 and 11 G41 clones observed). Because the pool was not pre-folded during selection but was for rapid kinetic assays, one possible interpretation is that G41 might promote catalysis but U41 might promote more optimal folding. At nucleotide 43, isolates bearing an A had significantly higher rate constants at 10 mM Mg^{2+} than do isolates with either a C (Bonferroni-adjusted $p < 0.002$) or a G (Bonferroni-adjusted $p < 0.05$), but the effect was not significant at 1 mM Mg^{2+}. A43 was somewhat but not significantly over-represented among successful clones (8.75 expected, 16 observed; $p = 0.079$, Fisher's exact test). In light of the unexpected result of Shechner *et al.* (3) that nucleotides 40 and 44 form a basepair that caps P3, with nucleotides 41–43 adopting a GNRA tetraloop-like configuration, the preferences at positions 41 and 43 likely reflect the energetics of this very short loop.

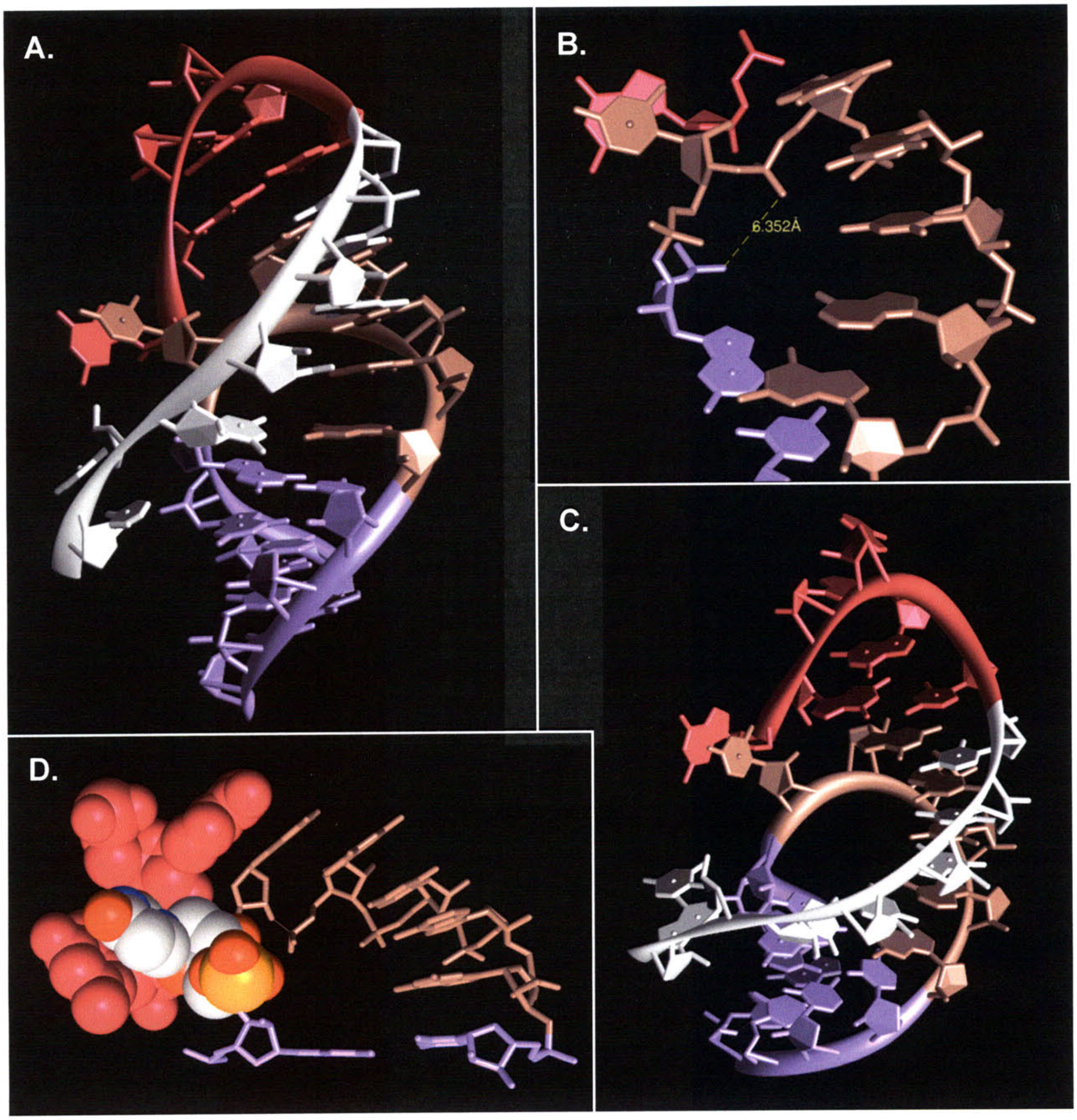
At two positions, the Monte Carlo analysis detected a significant effect at 10 mM Mg^{2+} that grew stronger at 1 mM Mg^{2+}. At the first of these, position 76, we have already seen that the parental identity was selected against strongly. Isolates bearing the parental adenosine have a mean ligation rate constant of 1.04 ± 0.55 min^{-1} at 10 mM Mg^{2+}, compared with 1.54 ± 0.64 min^{-1} among U76 clones ($p < 0.019$); at 1 mM Mg^{2+}, the relative gap widens to 0.06 ± 0.04 min^{-1} and

0.14 ± 0.12 min^{-1}, respectively ($p < 0.013$). A kinetic difference that widens with decreasing metal concentrations in this way could reflect either of two situations: a pre-existing, but weak, metal binding site might have its affinity improved by the favored substitution, allowing it to remain saturated at Mg^{2+} concentrations that would strip the site in the parent ligase; or the favored substitution might favor a non-metal-mediated interaction that compensates at low [Mg^{2+}] for a missing metal elsewhere in the ribozyme. Examination of the crystal structure of the improved ligase (3) suggests that U76 may fall into the former class.

Several structural elements come together in the neighborhood of nucleotide 76 (Fig. 4). Moving up helix P6 from the 3' end, the transition to the pseudoknot helix P3 is seamless; P3 simply continues the P7-P6 coaxial stack. But basepairing to form P3 occupies 5 of the 6 residues in what would otherwise be L6; the lone unpaired residue must negotiate the strand's return to P6 in register. That unpaired residue is nucleotide 76 (74 in the parent). In the improved ligase crystal structure, the U76 sugar and base are flipped out, allowing the base to stack favorably against G45 (Fig. 4). With the sugar flipped, the phosphate groups flanking U76 are brought closer together; the pro-R_p non-bridging phosphate oxygens at nucleotides 75 and 77 are separated by just 6.22 ± 0.19 Å (averaged across the two ribozyme chains in the crystallographic asymmetric unit), a significantly smaller gap than elsewhere in the P3/P6 junction (9.84 ± 1.54 Å; $p < 5 \times 10^{-5}$, unpaired t-test assuming unequal variances). As discussed in the next chapter, biochemical evidence suggests that these two oxygens, together with the pro-R_p oxygen of nucleotide 46, jointly ligand a structural Mg^{2+} ion. Notably, the U76/G45 stacking interaction and this metal-binding site are the sole points of structural communication between the P7-P6-P3 and P4-P5 domains (3). The flipped-out U76 base nearly fills the small cavity created by nucleotides 45–46 (Fig. 4D); a purine sidechain here would likely force small realignments of the backbone, movement that could affect the metal-binding affinity of the binding site formed by nucleotides 75, 77, and 46. In this model, the disposition of the parental adenosine constrains the backbone at positions 75 and 77 to a slightly less favorable conformation, more easily stripped of a bound Mg^{2+} ion. Conversely, mutation to U76 could improve the affinity of the metal-binding site, such that the kinetic benefit of this mutation is magnified at low Mg^{2+} levels.

At nucleotide 94, the widening kinetic gap at low [Mg^{2+}] seems to result not so much from one nucleotide's success as from another's failure; here, the failure belongs to a non-parental nucleotide. G94 is slightly but not significantly under-represented among the isolates; its disadvantage only becomes clear in kinetic assays, where G94 clones perform somewhat more poorly than U94 clones at 10 mM Mg^{2+} (0.49 ± 0.36 min^{-1} vs. 1.52 ± 0.55 min^{-1}; Bonferroni-adjusted $p < 0.031$) and much more poorly than U94 and A94 clones at 1 mM Mg^{2+} (0.01 ± 0.01 min^{-1}, 0.13 ± 0.11 min^{-1}, 0.11 ± 0.09 min^{-1}, respectively; Bonferroni-adjusted $p < 0.028$ for the G–U comparison, 0.035 for G–A). Thus, U94 clones slow by just 12-fold on average, and A94 clones by 14-fold, at 1 mM vs. 10 mM Mg^{2+}; by contrast, the nearly 40-fold drop for G94 clones is exactly on par with the drop seen in the parent ligase. The parent ligase bears a uracil at this position (parent U92), so the source of the parent's strong metal dependence must lie elsewhere; indeed, for the G94 clones to be no more sharply metal-dependent than the parent, they must have acquired another mutation elsewhere that can compensate for the cost of G94. With so few isolates to examine, however, we lack the statistical power to identify that beneficial change.

Figure 4. The shape of the nucleobase residue at position 76 may affect the docking of the P7-P6-P3 and P5-P4 domains. **(A)** The transition from P6 (light purple) to P3 (salmon, nucleotides 76–81; white, nucleotides 33–39) is smooth, with coaxial stacking between the two. J3/4 residues are shown in fuchsia. **(B, C)** The sugar and nucleobase at position 76 are flipped out of the pseudoknot arc, stacking the nucleotide 76 and 45 nucleobases and bringing the nucleotide 75 and 77 pro-R_p oxygens into unusually close proximity. **(D)** As a uridine residue, the nucleotide 76 base (in space-filling representation, CPK colors) closely abuts the nucleotide 45–46 base and backbone atoms (in space-filling representation, fuchsia). A larger purine residue here, like the parental adenosine, would more than fill this pocket, and in moving to resolve the steric clash would likely alter the position of the nucleotide 77 pro-R_p oxygen, potentially lowering the affinity of the metal-binding site it contributes to. Molecular graphics images produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081) (10).


A.
B.
6.352Å
C.
D.

CONCLUSIONS

Continued selection on the class I ligase ribozyme has shown that even this remarkable catalytic machine has room for optimization, with improvements in both rate constant and metal dependence appearing within just seven generations. The reduced metal dependence of the improved ligase is of particular interest in light of the close evolutionary relationship between this ribozyme and the polymerase ribozyme that is our current best guess as to the origin of life in an RNA world: a primordial ribozyme that requires high metal concentrations to saturate its binding sites is a ribozyme that is likely to decay before it can reproduce. Statistical analysis of the population of ribozymes that survives selection can identify residues under selection and candidate residues at which nucleotide identity may buffer the ribozyme against changing Mg^{2+} concentrations, and in some cases comparison with the crystal structure strengthens the argument that sequence matters at these positions. But the depth of understanding we can gain from such analysis is necessarily limited by the number of ribozyme clones it is practical to isolate and characterize. More direct approaches are needed to corroborate or falsify these candidate interactions, and to gain access to regions of the ribozyme at which the absolute dominance of one nucleotide closes the door to statistical analysis of the others.

MATERIALS AND METHODS

For details of ligase selection and kinetic assays, see chapter 4 of reference 2.

Fisher's exact test was used to detect positions at which the nucleotide composition of the successful isolates deviated significantly from that expected by chance. Analysis was performed in R version 2.7.1 using a 2x4 contingency table in the function fisher.test in the package stats (8). For pooled analysis of lightly mutagenized sites, 19 sets of 35 random numbers uniformly distributed between 0 and 1 were generated using the Mersenne-Twister random number generator as implemented in R; those numbers less than 0.1 were considered "non-parental". Simulated non-parental features were counted in each of the 19 sets, and the two sets with the most non-parental features were discarded. The remaining 17 sets' non-parental features were summed to give an aggregate count of *in silico* "observed" non-parental features. This simulation was repeated 10^6 times, generating a normally distributed set of *in silico* observations. The *in vitro* observation (10 non-parental features) was far smaller than the smallest of our 10^6 *in silico* observations (24 non-parental features, from a distribution with mean 53 ± 7), allowing us to conclude conservatively that the *in vitro* observation would arise by chance with $p < 10^{-6}$.

For Monte Carlo analysis, each ligase position x under scrutiny gave rise to a vector of length 35, whose ith element is the nucleotide identity of successful clone i at position x. Some ordered subset {A} of these elements will be adenosines, some subset {G} will be guanosines, some subset {C} cytosines, some subset {U} uracils, and some subset {E} gaps. The rate constants measured for the successful clones at each Mg^{2+} concentration were treated as another vector of length 35. R was used to re-order the components of the rate constant vector randomly 10000 times, producing a matrix of 35 rows by 10000 columns. Simulated mean rate constants for ribozymes bearing A, G, C, U, or a gap at nucleotide x were computed for each iteration of the Monte Carlo by averaging the elements in, respectively, the set of positions {A}, {G}, {C}, {U}, or {E} of column j of the matrix. Simulated standard deviations were computed in similar fashion. The t-statistic for each accessible pairwise comparison of means was then calculated in each column, where 'accessible' comparisons are those for which both nucleotide identities have at least 3 representatives. Thus, if all 5 nucleotide identities are represented in at least 3 clones each at position x, all 10 pairwise combinations—A/C, A/G, A/U, A/gap, C/G, C/U, C/gap, G/U, G/gap, and U/gap—were performed; if 4 nucleotide identities are represented, 6 pairwise comparisons were performed; if 3, 3; if 2, one; and if only one, the position was entirely inaccessible to this analysis.

At each position x in the ribozyme, this procedure yielded a histogram of 10000 t-statistics for each accessible pairwise nucleotide comparison. The true t-statistic was then calculated from the original vector of rate constants for each accessible comparison and compared to the appropriate histogram. Simulated t-statistics more extreme than the true t-statistic were counted, and the reported p value was determined as

$$p \leq \frac{m}{10000} \cdot 2 \cdot B$$

where m is the count of more extreme simulated t-statistics, the factor of 2 is a correction to give the two-tailed probability, and B is the appropriate Bonferroni penalty factor (10, 6, 3, or 1) for the number of comparisons performed.

WORKS CITED

1. DP Bartel and JW Szostak (1993). Isolation of new ribozymes from a large pool of random sequences. *Science* **261:**1411-8.
2. NH Bergman (2001). *The reaction kinetics and three-dimensional architecture of a catalytic RNA.* MIT Ph.D. Thesis.
3. DM Shechner, RA Grant, SC Bagby, and DP Bartel (2009). Crystal structure of the catalytic core of an RNA polymerase ribozyme. *Submitted manuscript.*
4. EH Ekland and DP Bartel (1995). The secondary structure and sequence optimization of an RNA ligase ribozyme. *Nucleic Acids Res* **23:**3231.
5. EH Ekland and DP Bartel, unpublished results.
6. NH Bergman, NC Lau, V Lehnert, E Westhof, and DP Bartel (2004). The three-dimensional architecture of the class I ligase ribozyme. *RNA* **10:**176-84.
7. DM Shechner, unpublished results.
8. R Development Core Team (2008). R: A language and environment for statistical computing. <http://www.R-project.org>.
9. CC Yen (2002). *Optimization and characterization of the class I RNA ligase.* MIT M.S. Thesis.
10. EF Pettersen, TD Goddard, CC Huang, GS Couch, DM Greenblatt, EC Meng, and TE Ferrin (2004). UCSF chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **25:**1605-1612.

Interference Mapping of Functional Interactions in the Class I Ligase Ribozyme

The experiments reported in this section benefited from discussions with Uli F. Müller, I-hong Shih, David M. Shechner, and Edward A. Curtis in the Bartel lab and with Piyali Chatterjee in the Caprara lab at Case Western Reserve University, and the interpretation gains greatly from the ligase crystal structure solved by David M. Shechner and Robert A. Grant. The interference maps are my own. Support for research in the Bartel lab comes from the National Institutes of Health and the Howard Hughes Medical Institute, of which David Bartel is an Investigator and I am a Pre-doctoral Fellow.

ABSTRACT

When the class I ligase emerged from random sequence, it was the fastest known all-RNA enyzme. Together with the prospect of improving the catalytic core of the RNA polymerase ribozyme, the combination of phylogenetic novelty with catalytic virtuosity makes the ligase ribozyme a desirable target for structural investigation. Extending this investigation across generations of ligase evolution offers the possibility of understanding the fitness advantage that accrued to the features that were favored in selection. We present backbone and nucleobase interference maps of the parent and improved ligase ribozymes; together with the newly solved crystal structure of the improved ligase (1), these analyses greatly extend our understanding of the functionally significant interactions underlying the outstanding catalytic ability of the class I ligase ribozyme.

ABBREVIATIONS

Nt, nucleotide; P, helical RNA stem; J, joining region; L, loop region NAIM, nucleotide analog interference mapping; DMS, dimethyl sulfate; NTP, nucleotide triphosphate; ATP, adenosine triphosphate; CTP, cytidine triphosphate; GTP, guanosine triphosphate; UTP, uridine triphosphate; EDTA, ethylene diamine tetraacetic acid; TBE, Tris-borate-EDTA; APM, N-acryloyl aminophenylmercuric acetate; HEPES, N-(2-Hydroxyethyl)piperazine-N′-(2-ethanesulfonic acid); MES, 2-(N-Morpholino)ethanesulfonic acid; DTT, dithiothreitol; AMV-RT, avian myeloblastoma virus reverse transcriptase.

INTRODUCTION

Previous structural characterization of the class I ligase ribozyme has established that loops L5 and L7 are adjacent in the active ribozyme and that J3/4 and P6 show extensive protection from hydroxyl radical cleavage (2). To begin to extend our structural understanding of the ligase to include functionally significant interactions, we performed a series of interference mapping experiments (3, 4). The hallmark of interference mapping is that the experimental signal derives exclusively from correctly folded, active ribozymes. The ligase folding step typically yields 70% active and 30% inactive ribozymes (5); the latter population is large enough to confound interpretation of structural assays that measure active and inactive ribozyme molecules indiscriminately. Given a selective step that can detect active ribozymes and silence the remainder, interference experiments essentially ask at which positions in a ribozyme a given chemical modification interferes enough with folding or catalysis to move the modified ribozyme from the population of active ribozymes to the inactive, 'silent' population. The ligase is eminently suited for interference mapping because its reaction can be the selective step: given the appropriate substrates, active ligase molecules can readily covalently modify themselves to bear a radiolabel or an affinity tag.

Functional contacts in the class I ligase might have changed in either of two ways over the course of selection of the improved ligase. The change of a base might directly alter key contacts between that base and other base or backbone atoms; it might also have indirect effects, if the altered base causes a shift in the positions of its backbone atoms and thus in their hydrogen-bonding or metal-binding potential. Two mapping techniques gave us access to the large majority of positions at which changes of either type might have occurred. First, nucleotide analog interference mapping (NAIM) of the parent and improved ligases allowed us to identify functionally significant backbone atoms in two classes, the non-bridging pro-R_p phosphate oxygens and the 2'-hydroxyl groups. The NAIM method (6-8) capitalizes on the chemical observation that, when a non-bridging phosphate oxygen in an RNA backbone is replaced with a sulfur atom, the backbone at that position becomes vulnerable to cleavage by iodine (9). T7 RNA polymerase can be coerced to make such substituted RNAs when the mixture of NTPs in an *in vitro* transcription reaction is doped with an α-phosphorothioate NTP, an NTP whose α-phosphate group has had one of its non-bridging oxygens replaced with sulfur (Fig. 1B) (10). The resulting RNAs contain sulfur substitutions only in the pro-R_p position (11, 12). Ligase molecules can be transcribed in such a reaction and then allowed to react with [^{32}P]-labeled substrate oligonucleotides; when the population is then subjected to I_2 cleavage and gel electrophoresis, positions at which the phosphorothioate substitution interferes with activity can be detected as missing bands on the cleavage ladder.

Sites of phosphorothioate interference are interesting in themselves as indicators of metal-binding sites, because sulfur has much lower affinity for Mg^{2+} than does oxygen (13-15). But the method can be extended greatly by performing transcription with T7 RNA polymerase mutants that tolerate a variety of chemical modifications in their substrate NTPs (16-20). Additional modifications to the α-phosphorothioate NTP—for instance, use of a 2'-deoxyribose sugar—permit interrogation of contacts made by other atoms, against the backdrop of interference effects seen for the phosphorothioate modification alone. It is important to note that

NAIM can return false positives; for instance, a phosphorothioate interference might reflect the steric effect of replacing an oxygen atom with a larger sulfur or the electrostatic effect of replacing an oxygen atom with a less electronegative sulfur, rather than a disrupted metal-binding site (21). Similarly, a 2'-deoxy interference effect might arise from the ramifications of a changed sugar pucker (22) as well as from disruption of a 2'-hydroxyl mediated hydrogen bond. False negatives are unlikely, so long as the selective step is performed under adequately stringent conditions.

Following NAIM, interactions made by the bases themselves were investigated by interference mapping with dimethyl sulfate (DMS) (23-26). DMS attacks and methylates the Watson-Crick face of adenosine residues (at N1) and cytosine residues (at N3), the favored nucleotides at 15 of the 18 sites that showed significant bias in ligase selection (Part I, chapter 1; Fig. 2B).[1] Following this treatment, the ribozymes are allowed to fold and react, and the ligase product is separated from the ribozymes that failed to react. Because these methyl modifications do not engender backbone lability, the purified products cannot simply be cleaved and resolved by electrophoresis as in NAIM. Instead, the assay is primer extension, with a labeled primer: methylated A and C residues cause reverse transcriptase to pause as it tries to cope with their modified Watson-Crick faces. The result is an accumulation of truncated primer extension products at positions where methylation still permits ribozyme activity and, as in NAIM, faint or missing bands at positions of interference. DMS is a useful tool for mapping secondary

[1] DMS also methylates guanosine N7, which makes the modified position labile to cleavage in sodium borohydride and aniline (27). However, this modification does not interrupt reverse transcription, and so it is transparent to the primer extension assay used here.

structure, because the methyl modifications disrupt Watson-Crick pairing; where secondary structure is already known, as it is for the ligase, DMS allows us to investigate regions of the ribozyme where selective pressure for a given nucleotide may signal a tertiary interaction of note.

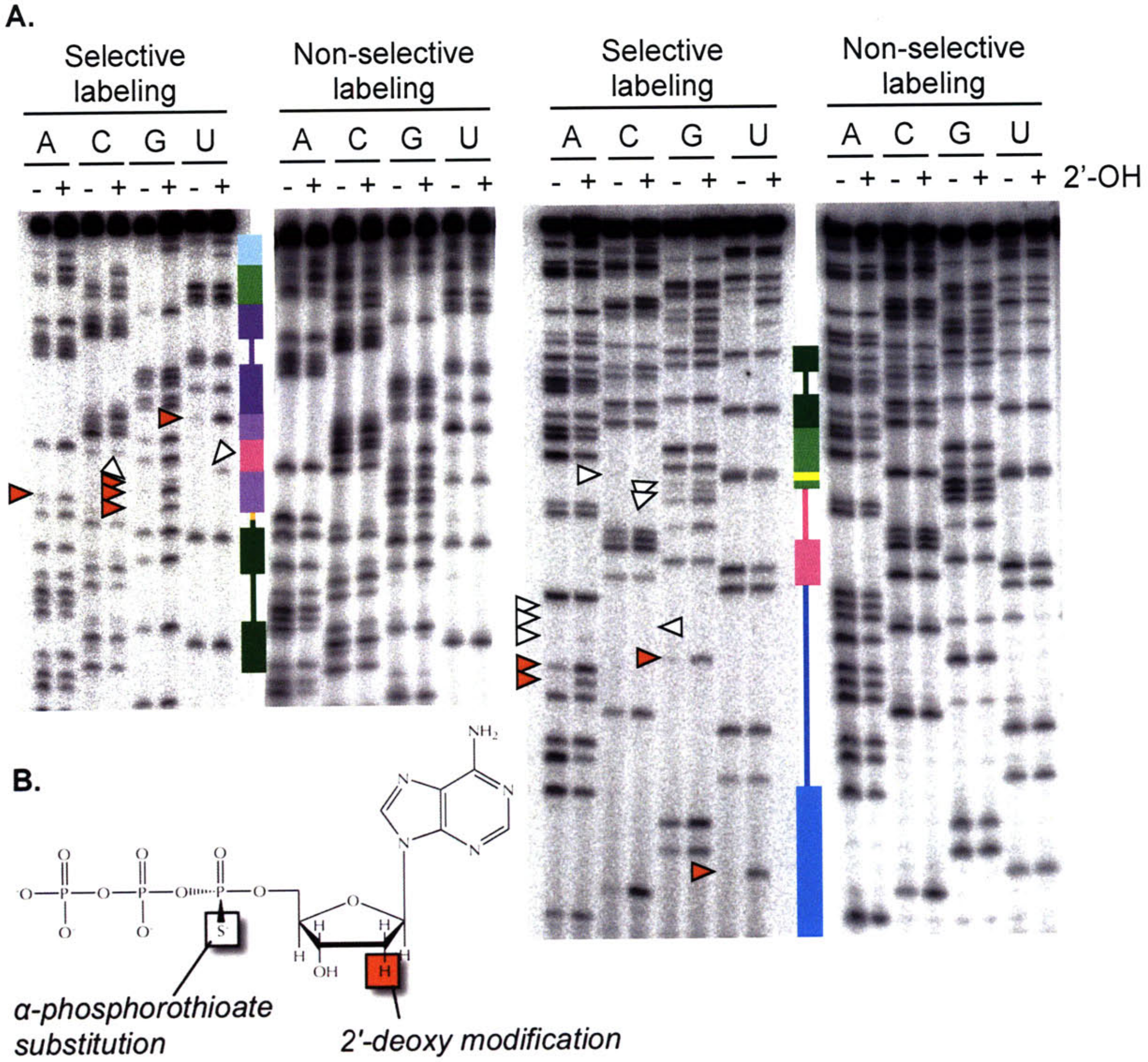
I report phosphorothioate, 2'-deoxy, and DMS interference maps for the parent and improved class I ligase ribozymes. These three interference maps present a coherent portrait of the ligase, one in which not only the key functional interactions but also the interactions whose importance to the ligase appears to be evolving are clustered in the largely single-stranded 5' end, with an island of interactions in helix P6 and another in P2. (It should be noted that all the modifications made here sometimes enhance ribozyme activity rather than interfering or leaving it unaffected; I report those results but do not in general attempt to interpret them.) In parallel with the work reported here, the crystal structure of the product of the improved ligase reaction was solved at 3.0 Å resolution (1). Independent interpretation of these results led to a number of concordant conclusions; in other cases, a pattern hidden in the interference data and the crystal structure considered separately came into focus when the observations were combined. Finally, at some positions the interference maps cannot be explained by the crystal structure, perhaps indicating moieties involved in ribozyme folding. I discuss the interference maps and their relation to the crystal structure below, as well as several telltale signs that some contacts particularly important to the polymerase are central to ligase activity as well, underscoring the argument that the ligase remains a good model for the polymerase core.

RESULTS AND DISCUSSION

Broad patterns of pro-R_p phosphorothioate interference

On the whole, the phosphorothioate profiles of the parent and improved ligases were similar (Fig. 1, 4, 5), although the magnitude of interference effects was typically greater in the parent ligase than in the improved ligase (Fig. 6). Among the 28 positions with a significant effect in either or both ligases, 11 showed interferences of equal strength and 1 an enhancement of equal strength; 3 showed enhancements that were significantly stronger in the parent than in the improved ligase and 2 an enhancement that was significantly stronger in the improved ligase than in the parent; and 9 showed interferences that were significantly stronger in the parent, while only one showed significantly stronger interference in the improved ligase. (At the remaining position, we have data only for the improved ligase, as the relevant band could not consistently be resolved on the parent ligase gels.) This pattern of weaker interference effects in the improved ligase is broadly consistent with its reduced Mg^{2+} dependence and suggests that the improved ligase followed a path to lower Mg^{2+} dependence that involved strengthening non-metal-mediated interactions such that a (presumably low-affinity) subset of existing metal-binding interactions became relatively dispensable for folding and function. Alternatively, it is formally possible that the improved ligase relies on the same metals as the parent but has tweaked the orientation of a series of phosphates and thereby positioned the NAIM-inaccessible pro-S_p oxygen for liganding, rather than the pro-R_p oxygen. But that such widespread backbone re-orientations should proceed and only once, at most, convert a parental pro-S_p interaction to a newly detectable pro-R_p one seems to us to be highly unlikely.

Figure 1. **(A)** Representative NAIM gels for quantification of phosphorothioate and 2'-deoxyribonucleotide effects in the improved ligase. Secondary-structure cartoons, colored as in the previous chapter, provide landmarks on the gels. 6% gels (left) were used to resolve the 3' half of the ligase, and 15% gels (right) were used to resolve the 5' half of the ligase. The range resolved by the two sets of gels overlaps in helix P5 (dark green boxes in secondary-structure cartoon). Non-selective labeling with ^{32}P by T4 polynucleotide kinase permitted quantification of the extent of incorporation of the experimental nucleotide analogs, while selective labeling revealed the extent of modification interference or enhancement. White arrowheads mark positions of particularly strong phosphorothioate interference; red arrowheads mark positions of particularly strong 2'-deoxy interference. **(B)** [α-Phosphorothioate]-2'-deoxyadenosine triphosphate (dATPαS), one of the eight nucleotide analogs used for NAIM. Like all the nucleotide analogs used, dATPαS bears an α-phosphorothioate substitution, the replacement of one non-bridging oxygen at the α-phosphate group with a sulfur atom. Use of the α-phosphorothioate-bearing ribonucleotides ATPαS, CTPαS, GTPαS, and UTPαS permits quantification of phosphorothioate interference effects and establishes a baseline for comparison with the α-phosphorothioate-bearing deoxyribonucleotides dATPαS, dCTPαS, dGTPαS, and dUTPαS to determine 2'-deoxy interference effects. The pattern of single and double bonds depicted at the α-phosphate follows the results of (40). The stereoisomer shown bears a pro-S_p sulfur substitution; this isomer is the only isomer recognized by T7 RNA polymerase, but because polymerization proceeds with inversion of stereochemistry (41), all sulfur substitutions in the resulting RNA are in the pro-R_p position.


A.
Selective labeling
Non-selective labeling
Selective labeling
Non-selective labeling
A C G U
- + - + - + - +
2'-OH
B.
NH2
O
P
S⁻
H
OH
α-phosphorothioate substitution
2'-deoxy modification

Each of the 20 sites of significant interference we detected has the potential to be an inner-sphere ligand of a functionally important Mg^{2+} ion. In principle, a metal rescue experiment, looking for phosphorothioate interferences that disappear in the presence of a thiophilic metal such as Cd^{2+}, could lend support to the role of some of these oxygens as inner-sphere ligands (28). However, the ligase is strongly inhibited by many transition metals, including the best candidates for metal rescue (29); pilot metal rescue experiments confirmed that this line of inquiry, always dogged by caveats (30, 31), was unlikely to be productive in our system (data not shown). We instead turned to our crystal structure of the product of the improved ligase reaction (1) to interpret our interference maps. Nearly half the phosphorothioate interference effects we observed can readily be understood in terms of metal-binding activity; reasonable hypotheses can be advanced for another third; and the remainder, together with most phosphorothioate enhancement effects, remain unexplained.

Candidate metal-binding sites in the ribozyme core

The phosphorothioate interference effects for which the crystal structure offers a clear interpretation largely cluster in and around the ligase active site, proposed by Shechner *et al.* to center on backbone atoms from nucleotides 29–30 and the C47 nucleobase (parent C48; Fig. 6) (1). As reported in that work, the strongest set of phosphorothioate interference effects lie at the 3' end of J1/3, at positions 29–32 (parent 30–33). At each of these positions, the phosphorothioate substitution reduced ligase activity by a factor of ≥6, the limit of the assay's reliable range (see Methods). Independent of the crystal structure, we hypothesized that this series of contiguous interference sites might indicate a bend in the backbone sharp enough that pro-R_p oxygens at adjacent residues would be positioned to ligate the same metal ion. This prediction was borne out by the backbone geometry at nucleotides 30–32, where J1/3 abruptly makes a right-angled turn (Fig. 3D,E) (1). This turn positions the pro-R_p oxygens of nucleotides 31–32 to coordinate a magnesium ion whose stable presence is supported by a strong peak in the difference Fourier map (1). That the difference map of the ligated product does not give evidence of a second metal bound at positions 29–30, despite the extremely strong interference effects observed, is consistent with a metal ion bound by these two phosphates that is important at an early stage of the reaction but is released later in the trajectory.

On the other side of the proposed active site, the J3/4 backbone closely abuts the backbone of the 5' strand of helix P6, with phosphate-phosphate distances between adjacent strands of just ~7 Å. A chain of well-ordered metal ions is strung between the J3/4 and P6 backbones, providing the electrostatic screening necessary to permit such close strand-strand packing. Loss of any of these metal-binding sites by phosphorothioate substitution should lead to repulsion between J3/4 and P6, presumably disrupting the positioning of the proposed catalytic nucleobase C47 in the active site. Indeed, as at nucleotides 29–32, we observe a ≥6-fold phosphorothioate interference effect at nucleotide 47 (parent 48); in the crystal structure, the pro-R_p oxygen of nucleotide 47 is well-positioned to be an inner-sphere ligand of one Mg^{2+} in the backbone-screening chain of ions. One position upstream, nucleotide 46 (parent 47) directs its pro-R_p oxygen towards the pro-R_p oxygens of P6 residues G75 and G77 (parent G73 and G75), with a spacing of ~6 Å between oxygens. Although these three residues all show strong (and, strikingly, nearly equal) phosphorothioate interference effects in both ribozymes, no connection between the three was expected before the crystal structure was solved. Examination of an early model of the crystal structure, before heteroatoms had been built, revealed that the pro-R_p oxygens of these three residues were well-placed to be inner-sphere ligands of a common metal ion. Subsequent refinement of the crystal structure confirmed the presence of an electron density peak in between the pro-R_p oxygens of positions 46, 75, and 77, the third apparent metal ion in the chain that links J3/4 with P6.

Hints of evolving interactions

The phosphorothioate interference map and the electron density map also confirm each other at position 115 (parent 113). Phosphorothioate substitution here produces a ≥6-fold drop in activity in the improved ligase (we could not compare this to the effect in the parent ligase, because the band could not be unambiguously identified on parent ligase gels). In the improved ligase, this pro-R_p oxygen is beautifully positioned to serve as an inner-sphere ligand of a metal ion whose outer-sphere ligands appear to include the pro-R_p oxygens of nucleotides 3–4 and the 2'-hydroxyl group of A114 (A112 in the parent). Consistent with a hydrogen bond from the A114 2'-hydroxyl group to a water molecule hydrating the metal, we observed a significant 2'-deoxy interference effect at A114 (see below), ≥3-fold stronger in the parent ligase than in the improved ligase. The simplest explanation is that the parent 2'-deoxy interference also arises from disruption of this metal, in which case the observed difference in magnitude would indicate that ligase evolution has made the ribozyme less reliant on this interaction.

A clear instance of evolving interference effects arises at nucleotide 20. Here, the parent ligase shows 3.5-fold phosphorothioate interference, twice as strong as the significant but mild 1.8-fold effect in the improved ligase. But the crystal structure gives no indication of a metal ion; instead, the pro-R_p oxygen is positioned well to serve as a hydrogen bond acceptor for the 2'-hydroxyl group of position 19. Such an interaction could be disrupted by the electrostatic perturbation of phosphorothioate substitution. Looking to the 2'-deoxy interference map, we find that removing the 2'-hydroxyl group of position 19 produces twice as strong an interference in the parent (2.1-fold) as in the improved ligase (1.1-fold, not significant). The concordant relative impacts of the phosphorothioate effect at position 20 and the 2'-deoxy substitution at position 19 in the two ligases strongly suggests that both effects arise from disrupting the hydrogen bond seen in the crystal structure, and further that the improved ligase may tolerate the loss of this bond more readily than can the parent. In both ribozymes, disrupting the pro-R_p oxygen partner of this hydrogen bond appears to come at a greater cost than disrupting the 2'-hydroxyl partner, perhaps indicating that additional steric penalties are incurred by the phosphorothioate substitution.

Probably not coincidentally, as we look at phosphorothioate effects that are weaker in the improved ligase, the structural source of these effects grows less clear. For instance, one might imagine that the 2.6-fold phosphorothioate interference at position 35 (parent 36) arises from disruption of a metal ion coordinated jointly by the pro-R_p oxygens of this position and nucleotide 76 (parent 74), which shows 2.3-fold interference in the improved ligase. At 4.7 Å, the spacing of the two pro-R_p oxygens is consistent with this idea, but no ordered metal occupies the site in the crystal structure. Further, the phosphorothioate interference at position 35 is a factor of two stronger in the parent than in the improved ligase, while the interference effects at position 76 are not significantly different across generations. This by no means falsifies the hypothesis, but we lack a route to confirmation.

The final pair of phosphorothioate effects for which we can offer any interpretation lie in the 5' end of J1/3. As discussed below, several lines of evidence indicate that some key structural difference in this region separates the improved ligase from the parent. Consistent with the picture of ongoing evolution in this region, positions 23 and 24 (parent 24 and 25) show no phosphorothioate interference in the improved ligase but 2.3- and 3.4-fold interference, respectively, in the parent ligase. Looking at the structure of the improved ligase with an archaeologist's eye, we find what could be the ruins of the interactions behind these parent interferences. At nucleotide 23, the pro-R_p oxygen is suggestively but sub-optimally placed in the improved ligase for a hydrogen-bonding interaction with the exocyclic amine of G17; local sequence changes during selection could easily have introduced both a slight but disruptive backbone shift here and a new hydrogen bond elsewhere to compensate. A shift in the backbone conformation at position 23 could also have forced the re-orientation of that nucleotide's pro-S_p oxygen, which in the improved ligase is poorly oriented but reasonably spaced for joint coordination of a metal ion with the pro-R_p oxygen of position 24 and the 2'-hydroxyl group of substrate residue –4. It is a coherent story, but in the absence of direct evidence of the parent structure here it remains only a story.

Lingering questions in the phosphorothioate map

At most of the remaining sites of significant phosphorothioate effects, the crystal structure offers no suggestion of the source of the effect. At three positions, however, the interpretation the structure suggests is belied by other aspects of the interference maps. First, based on the structure alone, the ~4-fold phosphorothioate interference effect at G45 (A46 in the parent) is difficult to connect to an interaction with any group other than the 2'-hydroxyl group of position 43; but a 2'-deoxy nucleotide at position 43 is neutral in the improved ligase and enhances activity 3-fold in the parent. Similarly, one might connect the parental phosphorothioate interference at position 71 (parent 69) to an interaction with the 2'-hydroxyl at position 70 (parent 68), but again a 2'-deoxy substitution at the latter gives significant enhancement of ligase activity. Finally, the structure suggests that the ~two-fold phosphorothioate interference at position 73 (parent 71) could reflect an interaction with the pro-R_p oxygen of nucleotide 107 (parent 105), when in fact the latter shows mild but significant phosphorothioate enhancement in both ligases. Such instances of seeming contradiction between crystallographic and biochemical data may simply mark positions where steric or electrostatic problems rather than interactions of note cause the phosphorothioate effect. However, it should be remembered that the interference maps reflect the early part of the reaction, from folding through the transition state, while the crystal structure reflects the product; differences between the two methods may have something to tell us about the changes that take place along the reaction coordinate.

A mixture of mutability and evolutionary stability in the 2'-deoxy interference map

Unlike the phosphorothioate interference maps, the 2'-deoxy interference maps show no marked asymmetries of effect strength between the two ribozymes (Fig. 1, 4). However, there are a number of positions at which the mean interference effects in the parent and improved ligases are not only significantly different according to a t-test but also differ by at least a factor of two (Fig. 4, 6). At one of these positions, nucleotide 84 (parent 82), significant interference is seen in both ribozymes but the effect is much stronger in the improved ligase; at four positions, nucleotides 12, 15, 17, and 114 (parent 112), only the parent ligase shows a significant effect; at one position, nucleotide 29 (parent 30), only the improved ligase does; and at two positions, nucleotides 25 and 116 (parent 26 and 114), one ligase shows a significant interference while the other shows a significant enhancement. The reliance of the ligase on the 2'-hydroxyl groups at positions 12, 15, and 17 was already weak enough in the parent ligase that no heroic selective pressure should have been required to uncover other interactions more beneficial to the ribozyme. But the pattern at positions 29 and 114 reveals that our selection gave the ribozyme scope for more dramatic shifts, even involving the handful of backbone groups whose perturbation reduces ligase activity to near-undetectable levels. Even more strikingly, the 2'-deoxy interference maps at position 116 show a more than 6-fold change, from >3-fold interference in the parent to >2-fold enhancement in the improved ligase; it is as though, in our seven rounds of selection, we had taken the ligase apart, put it back together, and then discovered that, for once, leaving out the proverbial spare parts made the system run more smoothly.

In the large majority of cases, though, the 2'-hydroxyl groups that matter to the parent ligase matter roughly as much to the improved ligase. The interactions at these positions in the parent ligase may represent global optima, or the fitness landscape in this region of sequence space may be too rugged to allow compensatory interactions elsewhere in the ribozyme to emerge. One hydroxyl group in particular, at position 16, has a remarkably persistent impact on ribozyme function; its 2'-deoxy interference effect is the strongest observed anywhere in either ligase, and the effect survives not only the seven generations that separate the improved ligase from the parent but also the 18 generations and altered connectivity that separate the polymerase ribozyme from the parent. Nucleotide 16 lies in the 5' strand of helix P1 in the ligase; strictly speaking, it is not part of the polymerase ribozyme at all, but part of the template oligonucleotide added in trans to the polymerase reaction. The two strands are directly analogous, serving in each case to align the attacking 3'-hydroxyl group with the 5'-triphosphate being attacked: in the ligase, the 5' strand of helix P1 pairs simultaneously with both the (primer-like) substrate oligonucleotide and with nucleotide G1, the triphosphorylated residue the substrate will attack, while the polymerase template strand pairs with both the growing primer and the incoming NTP. Because of the chemical similarity of the ligase and polymerase reactions, it has commonly been assumed that interactions between the (trans) polymerase primer:template duplex and the polymerase proper likely echo those between ligase helix P1 and the ligase core. The first direct evidence in support of that assumption comes from our 2'-deoxy interference maps at nucleotide 16.

In the terminology of Müller and Bartel (32), nucleotide 16 is at position –3 (with respect to the ligation junction) of the template-like strand of helix P1. Müller and Bartel, using the polymerase ribozyme and template strands with site-specific 2'-deoxy substitutions, found that a 2'-deoxynucleotide causes very strong interference at position –3. Just as the 2'-deoxy interference we observe at ligase nucleotide 16 is the strongest 2'-deoxy interference in either ligase ribozyme, the 2'-deoxy interference found at polymerase template nucleotide –3 is stronger than at any other site in the template strand. The robustness of this effect across the ligase family of ribozymes strongly suggests that the phosphoryl transfer reactions performed by all three ribozymes require this substrate-helix hydroxyl group to form a specific contact with the ribozyme core, a contact likely to contribute substantially to proper alignment of the substrate helix in the active site. The nature of this contact remains somewhat elusive. The crystal structure suggests that the nucleotide 16 2'-hydroxyl group may be involved in two hydrogen bonds, one to the 4'-oxygen of position 17 and one to the imine nitrogen (N1) of A25 (parent A26). Given that selection strongly favored adenosine at position 25, it is particularly tempting to think that the latter base-specific contact could explain the importance of the position 16 2'-hydroxyl. Yet a key hydrogen bonding interaction between a 2'-hydroxyl and an adenosine N1 should appear strongly in both the 2'-deoxy and DMS interference maps, and DMS interference at nucleotide 25 was variable enough in both ribozymes that the effect is significant in neither (Fig. 4, 5).

Interactions in need of further investigation

When considered side by side with the crystal structure and the DMS interference map, the 2'-deoxy interference map reveals a suite of tertiary interactions that join and brace key elements of secondary structure and explain the longstanding observation of a strong preference for adenosine residues in the 3' region of J1/3. I discuss these interactions in the context of the DMS map in the sections that follow, but turn first to the handful of interactions that raise more questions than they answer. The few 2'-deoxy interference effects that cannot be fully explained by the crystal structure serve as a further reminder that the crystal structure need not capture all the interactions that matter in the pre-ligation and transition state structures. For instance, as noted above, the 2'-deoxy interference effect at position 29 is more than twice as strong in the improved ligase (≥6-fold and significant) as in the parent (2.75-fold and not significant). The 2'-hydroxyl group of nucleotide 29 appears to form a hydrogen bond to the 2'-hydroxyl group of nucleotide 85 (parent 83). Consistent with this interaction, nucleotide 85 does show significant 2'-deoxy interference, but the effect is both mild (1.8-fold) and almost exactly equal in the two ribozymes. This pattern suggests that the active, unligated form of the improved ligase may have evolved to rely on a second interaction mediated by the 2'-hydroxyl group at position 29.

At nucleotide 76 (parent 74), too, comparison of the crystal structure and 2'-deoxy interference map suggests that we are still missing part of the story. Position 76 shows nearly equal 2'-deoxy interference in the parent and improved ligases (3.5-fold and 3.4-fold, though the latter is not significant). It is important to note, however, that the Monte Carlo analysis (Part I, chapter 1) showed the parental nucleotide at nucleotide 76 to be associated with significantly slower rate constants at low Mg^{2+} concentrations, a result we interpreted in terms of the parental purine over-filling a small cavity and forcing the backbone into a conformation less favorable for metal-binding. If local backbone alignment does indeed differ between the parent and the improved ligase, there is no reason to think that the parent has access to the same 2'-OH hydrogen bonding interactions as the crystal structure suggests the improved ligase has. Reasoning just about the improve ligase, then, we can say that its 2'-hydroxyl group appears to hydrogen-bond to that of nucleotide 44 (parent 45); and yet disruption of the latter 2'-hydroxyl group gives just 1.5–1.7-fold interference. Here, the rest of the story may be hydrogen bonds from the nucleotide 76 2'-

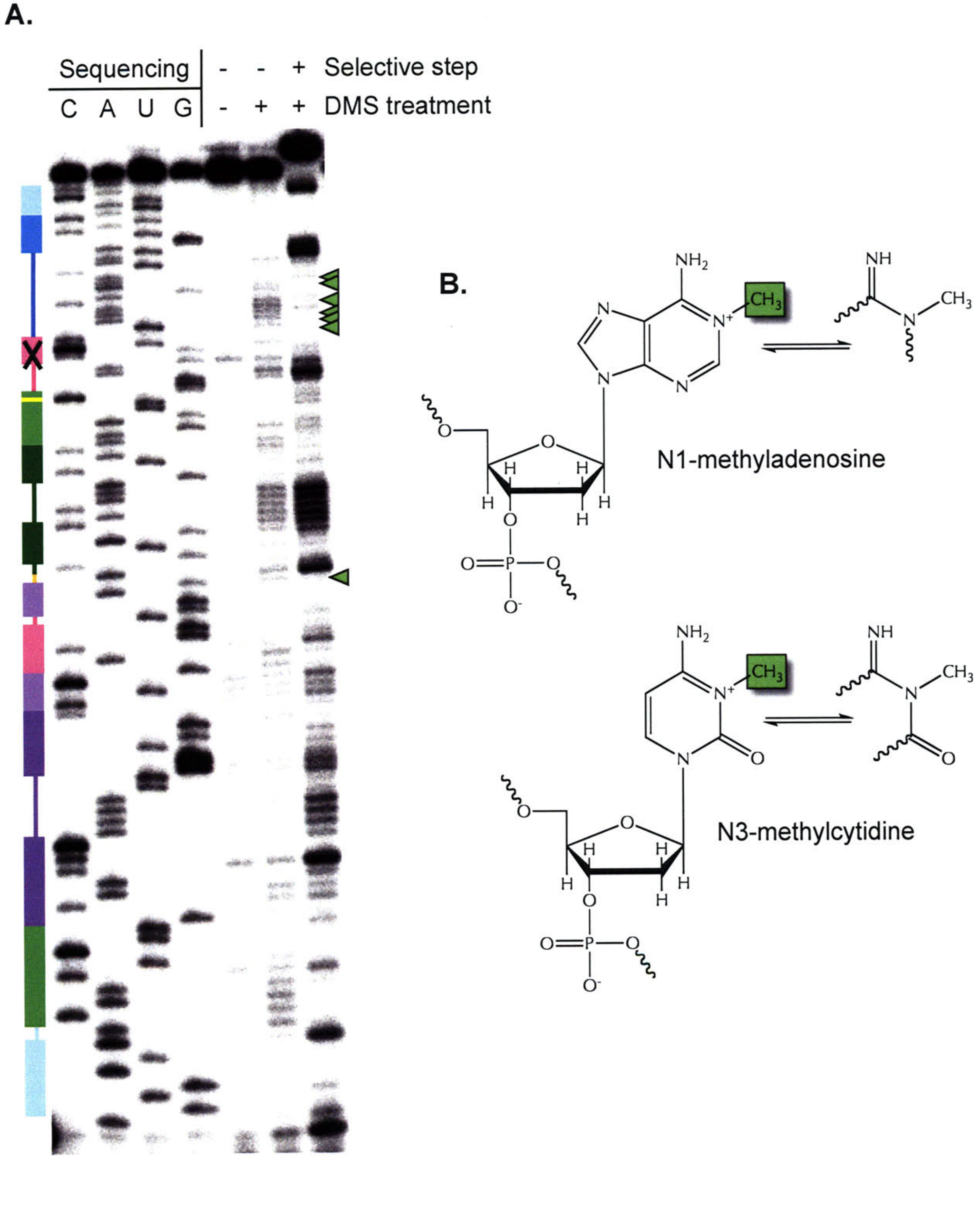
OH to the 4' and 5' oxygens of nucleotide 45 (parent 46), which appear in just one of the two ligase chains in the crystal asymmetric unit. The additional cost of losing these bonds would appear only when the 2'-hydroxyl group of position 76 was disrupted, accounting for the disparity between 2'-deoxy interference strength there and at position 44.

At a number of other positions showing significant interference in both ligases, the crystal structure predicts interactions with other atoms that are not probed by the interference maps presented. High backgrounds prevented the quantification of interference effects at the very 5' and 3' ends of either ligase; within those bounds, we have probed only the pro-R_p oxygens, the 2'-hydroxyl groups, and the N1 atoms of adenosines and N3 atoms of cytosines. Thus, we predict but cannot confirm that the 2'-deoxy interference at position 12 reflects disruption of a hydrogen bond to an ordered water molecule that coordinates a bound metal; the 2'-deoxy interference at position 75 (parent 73), disruption of a hydrogen bond to the 3' oxygen of position 34 (parent 35); the 2'-deoxy interference at position 113 (parent 111), disruption of hydrogen bonds to the imine (N1) and exocyclic amine (N6) groups of A3; and the 2'-deoxy interference at position 114 (parent 112), disruption of hydrogen bonds to the pro-R_p oxygen of A4 and to another metal-bound water.

Basepairing interactions in the dimethyl sulfate interference map

The primer extension assay used for DMS interference experiments (Fig. 2) gave rise to a somewhat noisier map than did the NAIM assay, including a number of positions at which one ribozyme showed significant and frequently very strong interference and the other a high mean

Figure 2. **(A)** A representative primer extension gel for quantification of DMS interference in the improved ligase, with sites of strong interference in single-stranded regions marked by green arrowheads. Whereas the label used to visualize NAIM gels is on the 5' end, the label used in primer extension is on the 3' end; thus, the ribozyme sequence on these gels appears upside-down with respect to the gels in Fig. 1. The secondary-structure cartoon is colored as in Fig. 1. Band identities were established using dideoxy sequencing ladders (from left, ddG, ddU, ddA, and ddC, yielding the positions of C, A, U, and G residues, respectively); note that there is a one-base offset between dideoxy sequencing ladders and experimental lanes, because chain termination by pausing at a methylation site will give a product one nt shorter than chain termination by incorporation of a dideoxynucleotide at that position. The gel compression that prevented analysis of nucleotides 34–41 is indicated by an X on the secondary-structure cartoon. Quantification includes normalizing over different whole-lane intensities. **(B)** The products of DMS-mediated methylation of adenosine and cytidine residues, with the added methyl groups marked by green boxes. In the wake of methylation, the equilibrium position of the enamine-imine tautomerism for these residues is expected to shift to the right.

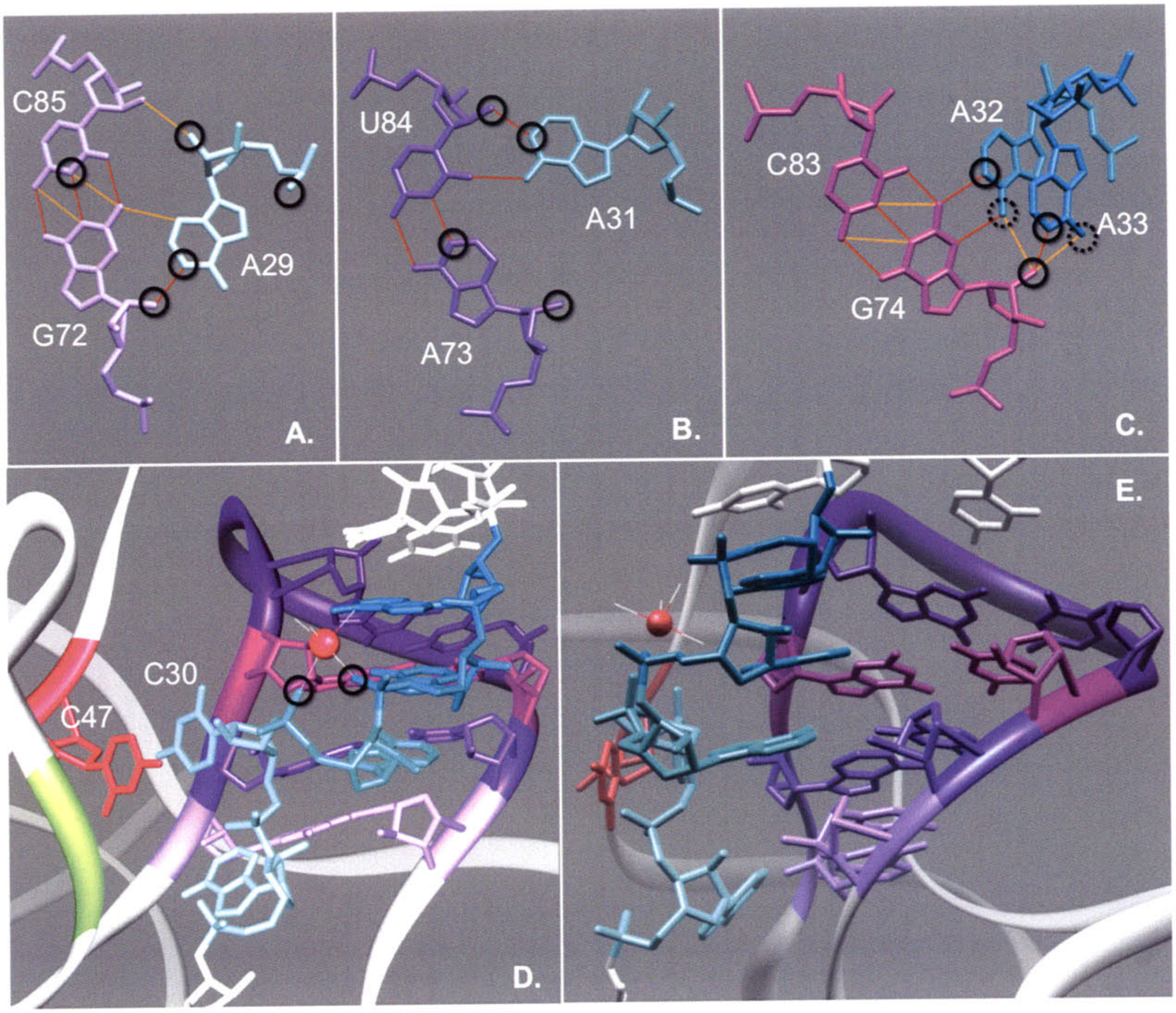

A.
Sequencing
C A U G
- - + Selective step
- + + DMS treatment
B.
NH2
CH3
NH
CH3
N+
N
O
H
O=P—O
O-
N1-methyladenosine
N3-methylcytidine

interference factor but also a large standard deviation, such that there was neither a significant interference in the latter nor a significant difference between that non-effect and the strong effect in the former construct (Fig. 4). We have not attempted to rationalize these occurrences, but we note them with an asterisk (*) in the discussion below. Neither can comparisons between the two ribozymes be made at C35 and C38–C40, within a compression artifact on the improved ligase gels; at positions 19, 20, 45, 73, 76, 84, 95, and 99 (parent 19, 20, 46, 71, 74, 82, 93, and 97), where only one construct bears a C or A residue; and at position 117 (parent 115), which could not be resolved on the parent ligase gels.

A number of strong interference effects in the DMS map are readily explained as the result of disrupting Watson-Crick basepairing within helices: A6, A9, C35, C38, *A51, A53–A54, A73, *C79, A80, *A102, *C108, C110–C113, and A117 fall into this category. But not all helical regions are equally important for ligase activity; it is known, for instance, that L5 and parts of P5 can be manipulated with little effect on activity, so it is reasonable that the DMS interference effect at A56 in this stem should be relatively weak (just 2.2-fold interference in the parent ligase, and no significant effect in the improved ligase).

At the other end of the spectrum, the DMS interference map independently predicted that the crystal structure would reveal an error in our previous secondary-structure model of the ligase. That model (32) predicted a non-canonical basepair between A11 and A114 (parent A112). A symmetric A:A N1-amino basepair would almost certainly give rise to strong DMS interference at both positions, and a symmetric A:A N7-amino basepair might do so as well, depending on the modulation of N6 hydrogen bonding by the N1 modification. We observe very strong

Figure 3. Interactions between J1/3 and P6, suggested by the crystal structure (1) and confirmed as functionally significant by the 2'-deoxy and DMS interference maps. **(A, B, C)** A-minor interactions made by adenosines 29, 31, and 32 docking into the minor groove of P6. Note that the geometry of each interaction is distinct. Predicted hydrogen bonds with canonical geometry are shown in red; hydrogen bonds with slightly relaxed geometric constraints are shown in orange. Black circles mark sites of significant DMS or 2'-deoxy interference; dotted black circles mark the adenosine N6 atoms, indirectly probed by DMS, that could contribute to the DMS interference measured at these positions. Of the probed functional groups contributing to the interactions shown, only the N3 atom of C83 and the 2'-hydroxyl group of C85 fail to show significant interference. **(D, E)** Two views of the 3' end of J1/3 docking into P6. Residues are colored as in (A)–(C). The viewpoint in (D) highlights the extruded residue C30, which forms a stacking interaction with the proposed catalytic residue C47 (fuchsia), and the metal-binding interaction that stabilizes the dramatic backbone kink at nucleotides 30–32. Black circles in (D) mark the pro-R_p oxygens of A31 and A32, both of which showed extremely strong phosphorothioate interference; this effect is rationalized by their position as inner-sphere ligands of the crystallographically resolved Mg^{2+} ion shown in fuchsia. The position of the ligation junction is highlighted in light green. Molecular graphics images produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081) (43).

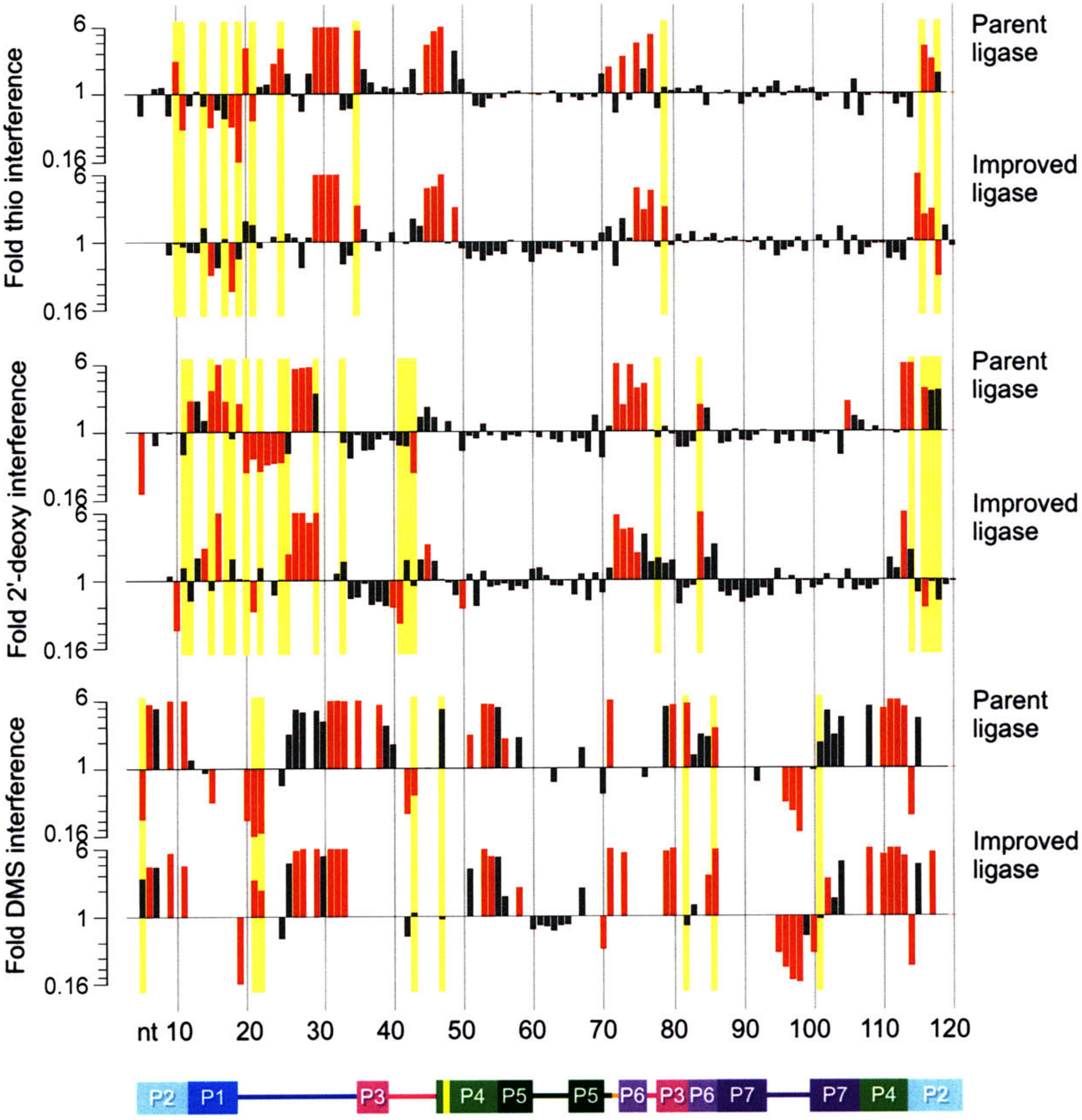

C85
A29
G72
A.
U84
A31
A73
B.
A32
C83
A33
G74
C.
C30
C47
D.
E.

interference at A11 in both ribozymes, and opposite but equally strong enhancement at A114, inconsistent with the formation of this basepair. The crystal structure bears this argument out: A11 forms a sheared G:A pair with G2, while A114 is splayed out to stack between A4 and G46 (parent G47). Methylation of A11 N1 would certainly interfere with the A11:G2 pair, and the A114 enhancement effect may indicate that, in the context of the A114 stacking interaction, the additional surface area for van der Waals interactions that is introduced by N1 methylation is beneficial.

Extensive concordance between dimethyl sulfate and 2'-deoxy interference mapping

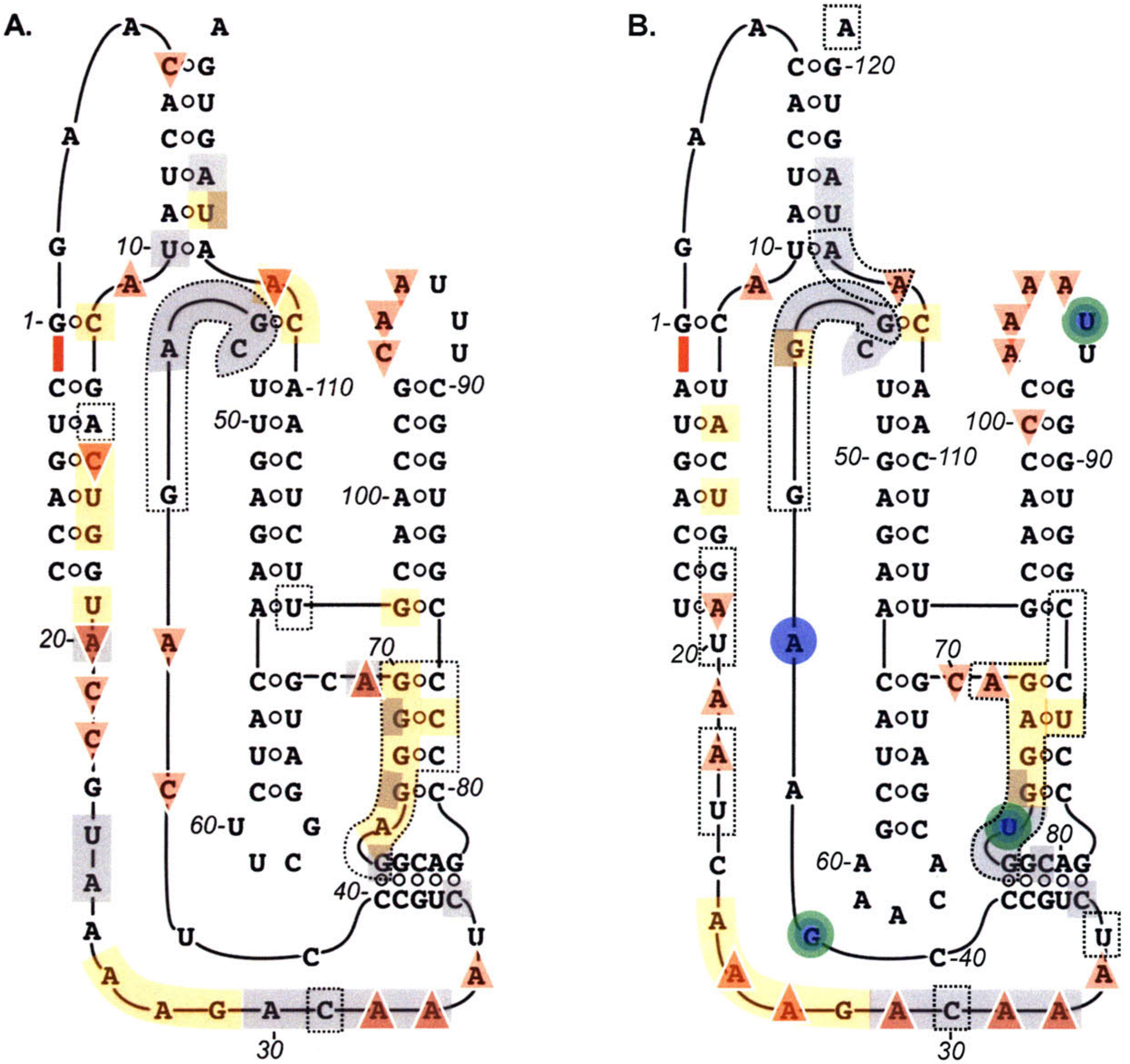
The DMS interference effects that arise in single-stranded regions are of particular interest, particularly those in J1/3, where the DMS interference pattern echoes both the strong 2'-deoxy interference effects at positions 26–29 and the strong phosphorothioate interference effects at positions 29–32 (Fig. 3–5). The crystal structure suggests that the DMS interference effects at nucleotides *26–*27 (parent 27–28) arise from interactions with the substrate oligo in helix P1. The N1 group of A26 is beautifully aligned to serve as a hydrogen bond acceptor for the imine nitrogen N2 of the guanosine residue at position –3, two nucleotides upstream of the ligation junction. At the same time, the N1 group of A27 may form a hydrogen bond with the 2'-hydroxyl group of the same substrate residue. Our 2'-deoxy interference maps do not encompass the ligase substrate, but we can look to Müller and Bartel's investigations of the effects of 2'-deoxy substitutions in the substrate strand used by the polymerase ribozyme. That work revealed a strong 2'-deoxy interference at the position analogous to ligase substrate nucleotide –3: by no means a direct confirmation of the proposed hydrogen bond, but a clear correlation.

Figure 4. Maps of mean phosphorothioate (thio), 2'-deoxyribonucleotide (2'-deoxy) and dimethyl sulfate (DMS) interference and enhancement effects in the parent and improved class I ligases. Positions of significant (95% confidence interval excludes 1) and strong (mean ≥ 2-fold) effects are shown in red. Yellow bars highlight positions at which the mean fold effect differs both significantly ($p < 0.05$, t-test) and substantially (by a factor of ≥2) between the two ribozymes. Note that the y-axis is a log scale. Phosphorothioate interferences obscure possible 2'-deoxy effects at several positions in both ligases. DMS mapping by primer extension yields data only for A and C residues. Ribozyme positions are numbered below, with the secondary-structure cartoon colored as before.


Fold thio interference
Fold 2'-deoxy interference
Fold DMS interference
6
1
0.16
Parent ligase
Improved ligase
nt 10
20
30
40
50
60
70
80
90
100
110
120
P2
P1
P3
P4
P5
P5
P6
P3
P6
P7
P7
P4
P2

We find direct confirmation of the interactions the crystal structure suggests a few positions downstream, at nucleotides *29 and 31–33 (parent 30, 32–34; Fig. 3A–C). These four nucleotides all show ≥6-fold interference in the improved ligase, and the parent ligase shows ≥5.9-fold interference at positions 31–33. The crystal structure suggests a stack of A-minor interactions (34), with each of these J1/3 adenosine residues packing into the minor groove of P6 in a slightly different geometry (1). A29 docks into the G72:C85 (parent G70:C83) pair at the base of P6, with a hydrogen bond forming between the A29 N1 and the G72 2'-OH, on the 5' strand of P6; as expected, we see very strong DMS interference at A29 and equally strong 2'-deoxy interference at G72. C30 (parent C31) is flipped out of the stack of J1/3-P6 interactions to stack with the proposed catalytic nucleobase C47; J1/3-P6 docking resumes with A31. This residue docks into the A73:U84 (parent G71:C82) pair, this time with a hydrogen bond forming between the A31 N1 and the position 84 2'-OH, on the 3' strand of P6; again we see very strong DMS interference and, in the improved ligase, equally strong 2'-deoxy interference. A32 forms an N3-amino, amino-N1 A:G pair with G74 (which also forms a Watson-Crick pair in P6 with C83; parent basepair G72:C81), forming hydrogen bonds between A32 N1 and G74 N2 and between A32 N6 and G74 N2, N3, and possibly 2'-OH. (DMS mapping does not directly probe adenosine N6, the exocyclic amine nitrogen, but methylation of N1 should affect the enamine-imine tautomerism equilibrium at N1, C6, and N6 (Fig. 2), potentially weakening hydrogen bonds made by N6.) And A33 caps this stack of J1/3–P6 interactions with a hydrogen bond from its N1 to the 2'-OH of G74 (parent G72); disrupting either end of this bond produces very strong interference in both ribozymes. A33 may further contribute to ligase activity by forming hydrogen bonds between its N6 and as many as three of the ordered water molecules that

Figure 5. Secondary-structure projections of the **(A)** parent and **(B)** improved class I ligase ribozymes, superimposed with the results of structural and sequence mapping. Dotted lines in (A) indicate the parent residues at which the C4' atom was protected from Fe-EDTA-generated hydroxyl radicals (2, 42); dotted lines in (B) indicate sites at which the C4' atom was calculated to be solvent-inaccessible in the improved ligase crystal structure (1). Phosphorothioate interference is shown in gray; 2'-deoxy interference is shown in orange; DMS interference is shown by upward-pointing salmon triangles and DMS enhancement by downward triangles. In (B), positions at which the Monte Carlo analysis (Part 1, chapter 1) revealed a significant association between nucleotide identity and ligase activity are highlighted by blue circles if the effect appears at 10 mM $MgCl_2$ and by green and blue rings if the effect appears at both 10 mM and 1 mM $MgCl_2$.


A.
1-
10-
20-
30
40-
50-
60-
70
-80
-90
100-
-110
B.
1-
10-
20
30
-40
50-
60-
70
80
-90
100-
-110
-120

coordinate the metal ion bound by the pro-R_p oxygens of nucleotides 31–32. The near-perfect agreement of the interference maps with each other and the crystal structure in this crucial region of the ligase serves to confirm that, whatever structural changes the ligase may undergo in the wake of reaction, the state of P6 and J1/3 captured by the crystal does still bear a very strong resemblance to the active conformation.

These J1/3 and P6 results also agree well with the hydroxyl radical footprinting data previously reported for the parent ligase (2). Those experiments revealed robust solvent protection of the ligase backbone on both strands of P6, even though only the 5' strand is buried against other helices in the catalytic core. But if J1/3 walks its way up a ladder of A-minor interactions involving backbone groups from both strands of P6, the observed solvent protection pattern is an obvious consequence. Notably, the solvent protection calculated from the improved ligase crystal structure accords with that measured in the parent ligase at nearly all positions (1), with a handful of intriguing differences (Fig. 5). One of these exceptions is nucleotide 84 (parent 82). In the parent, measured C4' solvent accessibility at this position was well below average; in the improved ligase, calculated C4' solvent accessibility is above average. In the parent, 2'-deoxy interference here is a meager 2-fold effect; in the improved ligase, it is ≥6-fold. As noted above, this basepair was changed from the parental G:C to an A:U during pool construction, in light of evidence that an A:U pair slightly improved function (35); perhaps that improvement is tied to a slight realignment of the helix backbone, permitting the improved ligase to further stabilize contact(s) that already play a significant role in the parent.

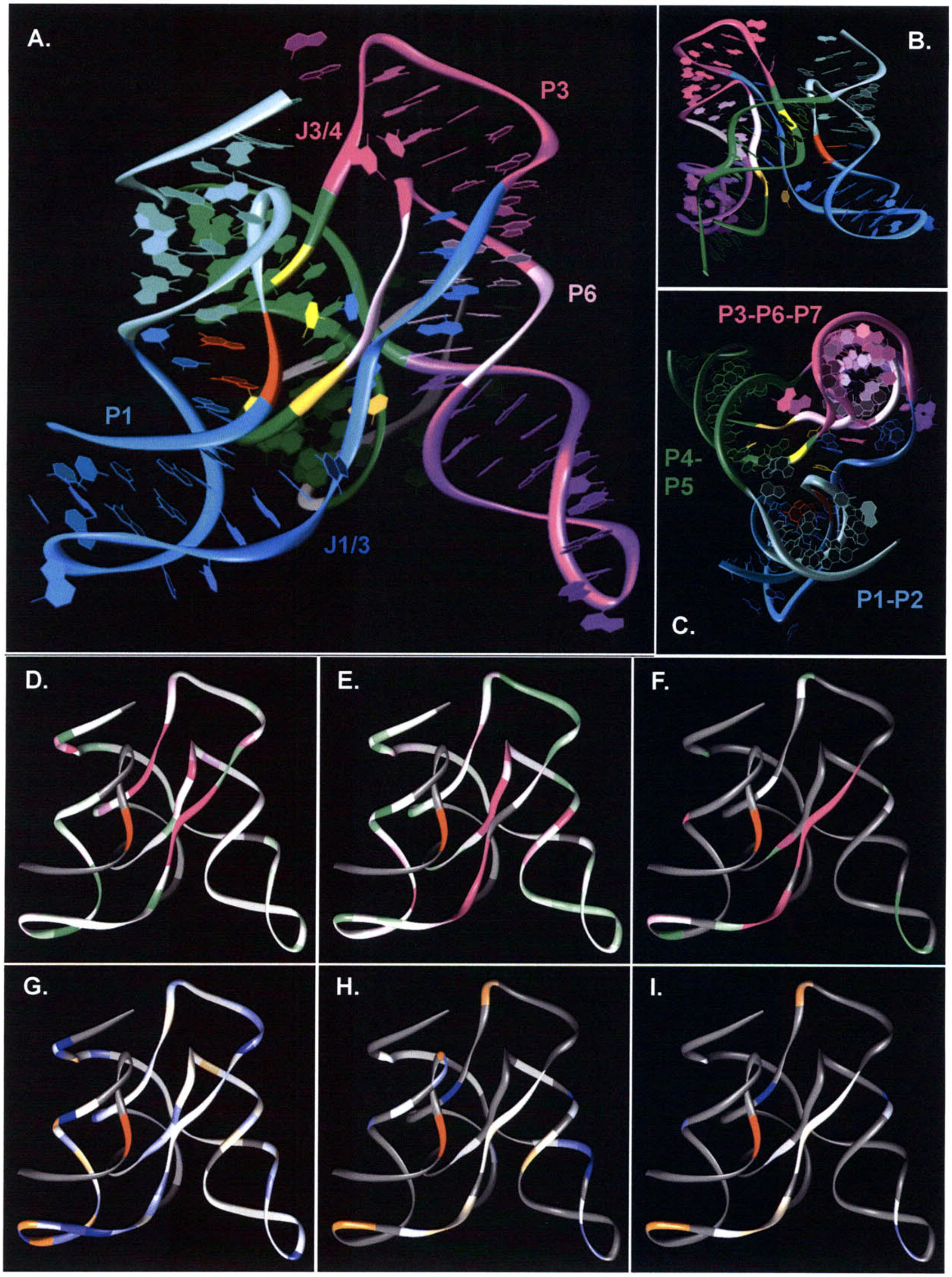
The idea that a slight but important realignment has taken place here is echoed in the differences in the relative importance of the other 2'-OH groups in the P6 helices of the two ribozymes. For instance, the improved ligase showed nearly 2-fold higher 2'-deoxy interference than did the parent ligase at A73, the complement to U84 (parent G71, C82), while contacts made by the 2'-hydroxyl groups of G74 and G75 (parent G72, G73) appeared less important in the improved ligase than in the parent. Continuing this trend, DMS-based disruptions of helix P6 and the proximal basepair of P7 do not affect the two ribozymes in quite the same way. At *C85, DMS modification gives only a 3-fold interference effect in the improved ligase, despite presumably disrupting the Watson-Crick basepair that caps P6. At C82 (parent C80), DMS modification has the expected very strong (5.5-fold) effect in the parent ligase, while tending, if anything, towards enhancement in the improved ligase. The tables are turned at C86, the proximal basepair of P7, where the 5.8-fold effect in the improved ligase is just over twice the interference seen in the parent. If indeed both ribozymes do make dense but distinct networks of tertiary contacts in this region of the ribozyme, it is possible that these networks provide an energetic buffer against disruption of the helices, but that the different networks can best accommodate disruption at different positions.

A hotspot of ribozyme evolution?

The DMS interference map also adds to the hints from other probes that continued selection has changed the role of the 5' end of J1/3 (Fig. 4–6). Selection favored non-parental residues at positions 19 and 21–23; C4' atoms at positions 19–20 and 22–23 are slightly more exposed than average in the parent and far more protected than average in the improved ligase (1, 2); there are large differences between the parent and improved ligase phosphorothioate interference effects at positions 19 and 21, and between the parent and improved ligase 2'-deoxy interference effects at positions 20 and 22; and now we find that positions 21–22 show DMS effects that are not only significantly different, and not only different by more than a factor of two, but also an enhancement in the parent and an interference in the improved ligase (Fig. 4). One clear possibility is that the difference in DMS effect is a function of nucleobase size. In the context of the parent cytosine residues here, an additional methyl group might represent a little more surface area that can profitably be buried, enhancing activity, while the larger adenosines in the improved ligase might already comfortably occupy the available cavity, so that an additional methyl group cannot crowd in too and produces a hydrophobic penalty rather than a van der Waals stabilization.

Whatever the nucleobase contacts involved, it is clear that the change in sequence between position 19 and position 23, from the parental [5']UACCG[3'] to the improved ligase [5']AUAA–[3'], drives the re-alignment of the backbone that is reflected in the phosophorothioate and 2'-deoxy interference maps and the solvent accessibility map. Moreover, the benefits of mutation to adenosine at residues in this region are not confined to the ligase ribozyme. Continued evolution of the polymerase ribozyme by Zaher and Unrau (36) led to a variant with markedly improved primer extension, as well as improved fidelity. The key difference—indeed, almost the only difference—between this variant and the polymerase of Johnston *et al.* (37) is the acquisition of three additional adenosine residues at positions analogous to ligase nucleotides 20, 21, and 23.

Figure 6. The improved class I ligase crystal structure with projections of the results of interference mapping. **(A, B, C)** Three views of the ligase, color coded as in the secondary-structure cartoons of previous figures, with the ligation junction highlighted in red. In (A), the long single-stranded region J1/3 wraps around the front of the ligase, its 5' end packing against P1 and its 3' end against P6. The view in (B) is rotated approximately 120° about the vertical with respect to the view in (A). (C) highlights the three coaxially stacked domains, and, at right, the transit of J1/3 from the P1-P2 domain to the P3-P6-P7 domain. **(D, E, F)** Projections of (D) phosphorothioate, (E) 2'-deoxy, and (F) DMS interference mapping results onto the view of the ligase shown in (A). As above, the ligation junction is highlighted in red. Residue colors are scaled from bright green for strong enhancement, through white for no effect, to bright purple for strong interference. Positions at which interference could not be quantified are colored dark gray. For clarity, the projection in (F) does not include positions at which DMS interference arises from Watson-Crick pairing in known helices. Note the concentration of effects in J1/3. **(G, H, I)** Projections of the extent of evolutionary change seen in the (G) phosphorothioate, (H) 2'-deoxy, and (I) DMS interference maps of the parent and improved ligase ribozymes. The mean effect in the improved ligase was divided by the mean effect in the parent ribozyme and the results scaled from blue (stronger enhancement or weaker interference—i.e., less reliance on the unperturbed residue—in the improved ligase) through white (no change) to orange (weaker enhancement or stronger interference—i.e., more reliance on the unperturbed residue—in the improved ligase). The strongest evolutionary change appears to be clustered towards the 5' end of J1/3 and in J3/4, with smaller differences in the stem regions. Note the predominance of blue-shaded positions in (G), consistent with the shallower Mg^{2+} dependence of the improved ligase.


A.
J3/4
P3
P6
P1
J1/3
B.
P3-P6-P7
P4-
P5
P1-P2
C.
D.
E.
F.
G.
H.
I.

CONCLUSIONS

Early in our understanding of catalytic RNAs, there were murmurs that having just four distinct building blocks would limit the catalytic repertoire of ribozymes. That speculation has turned out to be wildly off the mark, with the classical phosphoryltransferase and peptidyltransferase ribozymes now joined by an *in vitro*-evolved Diels-Alderase (38) and alcohol dehydrogenase (39), among others. As RNA crystallography improves and we learn more about the interactions that sustain structured RNAs, it becomes ever clearer that, though RNA's alphabet is small, its lexicon of interactions is large. Targeted structural probes like those performed here allow us to interrogate defined subsets of these possible interactions, often revealing congruent results when we separately disrupt the two sides of a proposed interaction. In general, the probe results we report here are in good agreement not only with each other but also with the recently solved crystal structure of the improved class I ligase ribozyme. Together, these methods offer evidence for catalytic metals bound at the ligase active site and reveal a diversity of adenosine-mediated interactions that belies the sequence-level homogeneity of the long single-stranded region J1/3. Our improved understanding of ligase structure and function, and our insights into the regions of the ligase ribozyme that are still actively evolving, will usefully inform future efforts to convert the RNA polymerase ribozyme into a true RNA replicase, the proof of principle the RNA world hypothesis requires.

MATERIALS AND METHODS

Nucleotide analog interference mapping

To incorporate nucleotide analogs into the ligase RNAs used for NAIM, linearized DNA encoding ribozyme RNAs at a final concentration of 0.5 mg/mL was mixed with 1 mM ATP, 1 mM GTP, 1 mM CTP, 1 mM UTP; the α-phosphorothioate nucleotide analog of interest (Glen Research) at the concentration recommended by the supplier; and 1X T7 Y639F buffer (40 mM Tris-HCl pH 8, 4 mM spermidine, 10 mM DTT, 15 mM $MgCl_2$, 0.05% Triton X-100). Transcriptions of the parent ligase also included 2.5 μM T7 promoter oligonucleotide. The reaction was mixed by vortexing prior to addition of the Y639F mutant of T7 RNA polymerase (16, 18), and incubated for 60–75 minutes at 37°C, when a cloudy precipitate had appeared in all reactions. Products were separated by denaturing gel electrophoresis and visualized by UV shadowing. Excised product bands were purified by overnight elution at 4°C in 300 mM NaCl and 1 mM EDTA, following by ethanol precipitation and spectrophotometric quantification.

Levels of nucleotide analog incorporation at each position were measured by non-selectively labeling the transcribed RNAs and subjecting them to I_2 cleavage. Briefly, 20 pmol of each transcript was treated with alkaline phosphatase (Roche) according to the manufacturer's instructions, then purified by phenol-chloroform extraction. Transcripts were then mixed with γ-[^{32}P]-labeled ATP and phosphorylated with polynucleotide kinase. Labeled transcripts were separated by denaturing gel electrophoresis and visualized by phosphorimaging using a Fujifilm BAS-2500. Transcript bands were excised and the labeled transcripts eluted as described above. Following ethanol precipitation, pellets were resuspended in a 1:2 mixture of RNase-free water and 2X denaturing gel loading buffer (8 M urea, 25 mM EDTA). Solutions were split into two aliquots; to the first was added 0.1 volumes freshly prepared 100 mM I_2 in EtOH, and to the second, to control for background degradation, 0.1 volumes EtOH. These solutions were heated to 50°C for 10 minutes and aliquots from each were loaded onto two sequencing gels, a 15% and a 6% polyacrylamide TBE-urea gel, to permit resolution of the 5' and 3' regions of the ribozyme, respectively. Banding patterns were visualized by phosphorimaging (Fujifilm BAS-2500) and quantified with the Fujifilm program ImageGauge. Two replicates of this calibration were performed for each nucleotide analog transcription.

For interference reactions, nucleotide analog-bearing ribozyme transcripts were mixed in a 2:1 molar ratio with γ-[^{32}P]-labeled substrate (5'-dAdAdACCAGUC-3', parent ligase substrate; 5'-UCCAGUA-3', improved ligase substrate), then denatured by heating to 80°C for 2 minutes and annealed by cooling to 22°C for 5 minutes. Ribozyme-substrate mixtures were spun down briefly and reactions were initiated by addition of 0.5 volumes 3X stringent ligase buffer (final concentrations 50 mM MES pH 6, 1 mM $MgCl_2$). Parent ligase reactions were incubated at 22°C for 100 minutes in an MJ Research PTC-100 thermocycler; improved ligase reactions, for 1 minute. Reactions were stopped by the addition of 2 volumes 2X denaturing gel loading buffer. Reactions were then split into two aliquots for treatment with freshly prepared dilute I_2 or mock-treatment with EtOH and subjected to gel electrophoresis and quantification as described above. The results reported are the average of three replicates of this interference experiment for each nucleotide analog.

2'-deoxy interference effects were calculated from background-subtracted peak intensities as follows. First, for each replicate of the interference experiment, the raw interference effect at each modified residue was calculated as

$$I_{raw} = \frac{\left(\frac{N}{dN}\right)_{selective}}{\left(\frac{\Sigma N}{\Sigma dN}\right)_{selective}}$$

where N is the background-subtracted peak intensity for the α-phosphorothioate ribonucleotide analog, dN is the background-subtracted peak intensity for the α-phosphorothioate deoxyribonucleotide analog, and the denominator is a gel exposure normalization factor (total background-subtracted intensity of the α-phosphorothioate ribonucleotide analog lane divided by total background-subtracted intensity of the α-phosphorothioate deoxyribonucleotide analog lane) to permit comparison of replicates run on different gels. Similarly, the relative levels of nucleotide analog incorporation for each ribo- and deoxyribonucleotide pair were calculated at each position as

$$R = \frac{\left(\frac{N}{dN}\right)_{non\text{-}selective}}{\left(\frac{\Sigma N}{\Sigma dN}\right)_{non\text{-}selective}}$$

from background-subtracted band intensities on the non-selectively-labeled calibration gels. The raw interference values from each replicate were then corrected for any differences in α-phosphorothioate ribonucleotide and deoxyribonucleotide analog incorporation at each position:

$$I_{cal} = \frac{I_{raw}}{<R>}$$

where $<R>$ is the average value obtained from two calibration replicates. Calculation of phosphorothioate effects is similar but uses only the α-phosphorothioate ribonucleotide analogs:

$$\Theta_{cal} = \frac{\left(\dfrac{<N_{non\text{-}selective}>}{N_{selective}}\right)}{\left(\dfrac{<\Sigma N_{non\text{-}selective}>}{\Sigma N_{selective}}\right)}$$

Because quantification of band intensities becomes less accurate as bands approach background levels, measured interference values become unreliable at very strong effect levels. We applied a cutoff at the six-fold effect level, truncating I_{cal} values at 6 and 0.167; any values that were measured outside this range are reported at the extremes of the range in Fig. 4. After application of this cutoff, the I_{cal} values measured at each position were averaged across the three interference replicates run for each nucleotide analog pair.

We considered effects to be significant if the 95% confidence interval about this average did not span 1. Effects were considered both significant and interesting if in addition the average is ≥ 2 or ≤ 0.5. At a given position, interference effects in the two ligase ribozymes were compared using the two-tailed t-test for two populations of equal sizes assuming equal variance. Positions at which average interference effects differed at the $p < 0.05$ level and at which the average parent and improved ligase interference effects were separated by at least a factor of 2 are highlighted in Fig. 4.

Dimethyl sulfate interference mapping

For DMS interference mapping experiments, we used ligase constructs bearing the 18-nt 3' tail that had been used during selection (5'-CCGGCCAGACGCCAGCGC-3'). We made this change because the readout of the DMS experiment is a primer extension assay, in which we wanted to ensure the primer would have unobstructed access to its binding site. That primer extension is the readout also means that ribozyme self-ligation to labeled substrate does not suffice as the selective step; instead, physical separation of reacted from unreacted ribozymes is required. Initial experiments with straightforward gel purification indicated that the resolution of these two populations (of lengths 137 and 146 nts in the 3'-extended parent ligase, and lengths 139 and 146 nts in the 3'-extended improved ligase) was inadequate on a standard gel; ribozyme reactions were performed under stringent conditions and stopped when only the most able 5–10% of the population had reacted, such that the shoulders of the very large unreacted-ribozyme band were broad enough to contaminate the reacted band. Again we turned to a technique from selection, using a 5'-thiophosphorylated substrate and separating the reaction products on an N-acryloyl aminophenylmercuric acetate (APM). So that the reacted and unreacted bands could be detected on these gels, we used ribozymes body-labeled with ^{32}P.

Ribozyme transcripts for these reactions were prepared by mixing 0.25 mg/mL DNA template for the 3'-extended ribozymes with 1X NTPs, T7 RNA polymerase buffer, 5 mM DTT, 0.1 μCi α-[^{32}P]-GTP per mol unlabeled GTP, and wild-type T7 RNA polymerase. Transcriptions ran for 50 minutes at 37°C and transcripts were purified by standard denaturing gel purification as described above. For DMS modification, 44 pmol transcript was mixed with 44 μg tRNA in buffer prepared without magnesium (50 mM Tris-HCl pH 7.5, 1 mM EDTA). Modification was initiated by addition of freshly prepared 5% DMS in 95% EtOH and proceeded at 37°C for 4 minutes. Modification was stopped with 1 volume freshly prepared DMS stop solution (1.5 M NaOAc pH 8.5, 7% β-mercaptoethanol, 0.1 mM EDTA). Reaction products were ethanol-precipitated twice to remove contaminants from the modification reaction and then resuspended in RNase-free water. This stock of modified RNA was split into two, with one aliquot to serve as the unselected control and the other to be put through the ligase reaction and APM gel-purification. In parallel, mock modifications were performed, using 95% EtOH alone in place of the 5% DMS in 95% EtOH, as a control for degradation during the DMS treatment.

Because the tRNA required in the DMS reaction inhibits the ligase reaction (data not shown), the aliquot of modified RNA that was to undergo selection was gel-purified prior to reaction with the 5'-thiolated substrate. The body-labeled, purified, modified ligase ribozymes were then mixed in a 2:1 molar ratio with substrates bearing a 5'-thiophosphate group and (to limit misfolding in the presence of the 3' primer binding site) a 1:1 molar ratio with unlabeled primer (5'-GCGCTGGCGTCTGGCCGG-3'), denatured by heating 2' at 80°C, and annealed by cooling to 22°C for 5 minutes. Ligation was initiated by addition of 0.5 volumes 3X stringent ligase buffer. Because ligation with the modified substrate and the 3' extension was slow, reaction times were extended to 11 h and 22 minutes for the parent and improved ligases, respectively. Reactions were then stopped by addition of one volume of denaturing gel loading buffer and loaded onto a 6% polyacrylamide TBE-urea gel prepared with 40 μM N-acryloyl aminophenylmercuric acetate and a non-APM stacking gel. Gels were imaged using the Fujifilm BAS-2500; reacted ribozymes were retained at the boundary between the stacking gel and the APM gel, separated by several cm from the unreacted band. Reacted-ribozyme bands were excised and the ligase products eluted and ethanol precipitated.

These selected ribozymes were compared to DMS-modified, unselected ribozymes and to mock-DMS-modified ribozymes using primer extension. Primer extension was performed largely as described in (25). Briefly, each ribozyme population was mixed with labeled primer and hybridization buffer (final concentrations: 50 mM K-HEPES pH 7.0, 100 mM KCl) and subjected to denaturation and slow cooling (85°C for 48 s followed by a decrease to 40°C at a rate of –2.5°C per minute) in an MJ Research PTC-100 thermocycler. Primer extension master mix (final concentrations: 125 mM Tris-HCl pH 8.4, 200 µM each dATP, dGTP, and dTTP, 300 µM dCTP, 6.4 mM $MgCl_2$, 9 mM DTT, 5% glycerol, and 0.027 units/µl AMV-RT [Seikagaku]) was added and extension allowed to proceed for 30 minutes at 42°C. Extension was stopped by addition of a solution of 1 part 300 mM NaOAc pH 8.5 and 1 mM EDTA to 3 parts 95% EtOH. After ethanol precipitation, pellets were resuspended in denaturing gel loading buffer and treated with 0.33 M NaOH at 5' for 90°C to remove RNA from the RT product. RT products were then separated on a sequencing-thickness 8% polyacrylamide TBE-urea gel with a buffer gradient running from 0.5X TBE in the upper buffer chamber to 1X TBE with 0.75M NaOAc in the lower chamber. To permit identification of bands, RT products were run in parallel with dideoxy sequencing ladders. Banding patterns were visualized by phosphorimaging and quantified with ImageGauge.

We quantified interference with respect to the extent of methylation at a given position as

$$I_{Me} = \frac{\left(\dfrac{N_{Me,\,non\text{-}selective}}{N_{Me,\,selective}}\right)}{\left(\dfrac{\Sigma N_{Me,\,non\text{-}selective}}{\Sigma N_{Me,\,selective}}\right)}$$

Reported interference effects are the average of three replicates, each with its own non-selective control. Significance was determined as for the phosphorothioate and 2'-deoxy interference maps. As noted above, not all positions in the ligase ribozymes could be resolved. The cDNAs of the improved ligase consistently showed a gel compression spanning nucleotides 34–41, and all primer extension reactions performed on the parent ligase, regardless of the presence or absence of DMS modification, showed intense bands at the base of P6, where the reverse transcriptase encounters four stacked G:C pairs.

WORKS CITED

1. DM Shechner, RA Grant, SC Bagby, and DP Bartel (2009). Crystal structure of the catalytic core of an RNA polymerase ribozyme. *Submitted manuscript.*
2. NH Bergman, NC Lau, V Lehnert, E Westhof, and DP Bartel (2004). The three-dimensional architecture of the class I ligase ribozyme. *RNA* **10:**176-84.
3. L Conway and M Wickens (1987). Analysis of mRNA 3'-end formation by modification interference: The only modifications which prevent processing lie in AAUAAA and the poly (A) site. *EMBO J* **6:**4177.
4. BC Rymond and M Rosbash (1988). A chemical modification/interference study of yeast pre-mRNA spliceosome assembly and splicing. *Genes Dev* **2:**428-39.
5. NH Bergman, WK Johnston, and DP Bartel (2000). Kinetic framework for ligation by an efficient RNA ligase ribozyme. *Biochemistry* **39:**3115-23.
6. F Conrad, A Hanne, RK Gaur, and G Krupp (1995). Enzymatic synthesis of 2'-modified nucleic acids: Identification of important phosphate and ribose moieties in RNase P substrates. *Nucleic Acids Res* **23:**1845.
7. SA Strobel and K Shetty (1997). Defining the chemical groups essential for *Tetrahymena* group I intron function by nucleotide analog interference mapping. *Proc Natl Acad Sci U S A* **94:**2903-2908.
8. SP Ryder and SA Strobel (1999). Nucleotide analog interference mapping. *Methods* **18:**38-50.
9. G Gish and F Eckstein (1988). DNA and RNA sequence determination based on phosphorothioate chemistry. *Science* **240:**1520-1522.

10. BM Chowrira and JM Burke (1992). Extensive phosphorothioate substitution yields highly active and nuclease-resistant hairpin ribozymes. *Nucleic Acids Res* **20:**2835-40.

11. F Eckstein (1985). Nucleoside phosphorothioates. *Annu Rev Biochem* **54:**367-402.

12. AD Griffiths, BV Potter, and IC Eperon (1987). Stereospecificity of nucleases towards phosphorothioate-substituted RNA: Stereochemistry of transcription by T7 RNA polymerase. *Nucleic Acids Res* **15:**4145-62.

13. VL Pecoraro, JD Hermes, and WW Cleland (1984). Stability constants of Mg^{2+} and Cd^{2+} complexes of adenine nucleotides and thionucleotides and rate constants for formation and dissociation of Mg-ATP and Mg-ADP. *Biochemistry* **23:**5262-71.

14. EL Christian and M Yarus (1992). Analysis of the role of phosphate oxygens in the group I intron from *Tetrahymena. J Mol Biol* **228:**743.

15. EL Christian and M Yarus (1993). Metal coordination sites that contribute to structure and catalysis in the group I intron from *Tetrahymena. Biochemistry* **32:**4475-4480.

16. R Sousa and R Padilla (1995). A mutant T7 RNA polymerase as a DNA polymerase. *EMBO J* **14:**4609-21.

17. Y Huang, F Eckstein, R Padilla, and R Sousa (1997). Mechanism of ribose 2-group discrimination by an RNA polymerase. *Biochemistry* **36:**8231-8242.

18. L Ortoleva-Donnelly, AA Szewczak, RR Gutell, and SA Strobel (1998). The chemical basis of adenosine conservation throughout the *Tetrahymena* ribozyme. *RNA* **4:**498-519.

19. R Padilla and R Sousa (1999). Efficient synthesis of nucleic acids heavily modified with non-canonical ribose 2'-groups using a mutant T7 RNA polymerase (RNAP). *Nucleic Acids Res* **27:**1561-3.

20. R Padilla and R Sousa (2002). A Y639F/H784A T7 RNA polymerase double mutant displays superior properties for synthesizing RNAs with non-canonical NTPs. *Nucleic Acids Res* **30:**e138.

21. AL Feig, OC Uhlenbeck (1999). The role of metal ions in RNA biochemistry. In *The RNA World.* RF Gesteland, TR Cech, and JF Atkins, eds. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

22. M Lindqvist, M Sarkar, A Winqvist, E Rozners, R Stromberg, and A Graslund (2000). Optical spectroscopic study of the effects of a single deoxyribose substitution in a ribose backbone: Implications in RNA-RNA interaction. *Biochemistry* **39:**1693-1701.

23. DA Peattie and W Gilbert (1980). Chemical probes for higher-order structure in RNA. *Proc Natl Acad Sci U S A* **77:**4679-4682.

24. DA Peattie and W Herr (1981). Chemical probing of the tRNA–ribosome complex. *Proc Natl Acad Sci U S A* **78:**2273-7.

25. D Moazed, S Stern, and HF Noller (1986). Rapid chemical probing of conformation in 16 S ribosomal RNA and 30 S ribosomal subunits using primer extension. *J Mol Biol* **187:**399-416.

26. S Stern, D Moazed, and HF Noller (1988). Structural analysis of RNA using chemical and enzymatic probing monitored by primer extension. *Methods Enzymol* **164:**481-9.

27. DA Peattie (1979). Direct chemical method for sequencing RNA. *Proc Natl Acad Sci U S A* **76:**1760-1764.

28. JA Piccirilli, JS Vyle, MH Caruthers, and TR Cech (1993). Metal ion catalysis in the *Tetrahymena* ribozyme reaction. *Nature* **361:**85-88.

29. ME Glasner, NH Bergman, and DP Bartel (2002). Metal ion requirements for structure and catalysis of an RNA ligase ribozyme. *Biochemistry* **41:**8103-12.

30. S Basu and SA Strobel (1999). Thiophilic metal ion rescue of phosphorothioate interference within the *Tetrahymena* ribozyme P4-P6 domain. *RNA* **5:**1399-407.

31. S-O Shan and D Herschlag (2000). An unconventional origin of metal-ion rescue and inhibition in the *Tetrahymena* group I ribozyme reaction. *RNA* **6:**795-813.

32. UF Müller and DP Bartel (2003). Substrate 2'-hydroxyl groups required for ribozyme-catalyzed polymerization. *Chem Biol* **10:**799-806.

33. EH Ekland and DP Bartel (1995). The secondary structure and sequence optimization of an RNA ligase ribozyme. *Nucleic Acids Res* **23:**3231.

34. P Nissen, JA Ippolito, N Ban, PB Moore, and TA Steitz (2001). RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proc Natl Acad Sci U S A* **98:**4899.

35. EH Ekland and DP Bartel, unpublished results.

36. HS Zaher and PJ Unrau (2007). Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *RNA* **13:**1017-26.

37. WK Johnston, PJ Unrau, MS Lawrence, ME Glasner, and DP Bartel (2001). RNA-catalyzed RNA polymerization: Accurate and general RNA-templated primer extension. *Science* **292:**1319-25.

38. B Seelig and A Jäschke (1999). A small catalytic RNA motif with Diels-Alderase activity. *Chem Biol* **6:**167-176.

39. S Tsukiji, SB Pattnaik, and H Suga (2003). An alcohol dehydrogenase ribozyme. *Nat Struct Biol* **10:**713-7.

40. PA Frey and RD Sammons (1985). Bond order and charge localization in nucleoside phosphorothioates. *Science* **228:**541.

41. S Verma and F Eckstein (1998). Modified oligonucleotides: Synthesis and strategy for users. *Annu Rev Biochem* **67:**99-134.

42. NH Bergman (2001). *The reaction kinetics and three-dimensional architecture of a catalytic RNA.* MIT Ph.D. Thesis.

43. EF Pettersen, TD Goddard, CC Huang, GS Couch, DM Greenblatt, EC Meng, and TE Ferrin (2004). UCSF chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **25:**1605-1612.

II.

In vivo

Physiological and Transcriptional Responses of *Prochlorococcus* to Carbon and Oxygen Limitation

The design of the experiments reported in this chapter benefited from discussions with members of the Chisholm lab. Their execution and analysis are entirely my own. I am indebted to Anne Thompson for training me in the use of the Influx flow cytometer for quantification of the growth curves, and to Matt Sullivan for comments on the manuscript. This work was supported by the Gordon and Betty Moore Foundation, the National Science Foundation, the Department of Energy, and the MIT Energy Initiative, in addition to my Howard Huges Medical Institute Pre-doctoral Fellowship.

ABSTRACT

The cyanobacterium *Prochlorococcus* is widespread in the open ocean, reaching high abundance in oligotrophic waters. Even given this tendency, the discovery by Goericke *et al.* that *Prochlorococcus* accounts for as much as 97% of photosynthetic pigments by mass in some extremely low-oxygen waters (1) came as a surprise. Because oxygen competes with carbon dioxide for binding to Rubisco, we wondered whether the apparent relative fitness of *Prochlorococcus* in these low-oxygen waters was connected to the observation that the selectivity of *Prochlorococcus* Rubisco is especially poor (2), and more generally how the metabolic ties between carbon dioxide and oxygen play out physiologically when *Prochlorococcus* is exposed to different mixtures of the two. I measured growth of the *Prochlorococcus* strain MED4 under a range of CO_2:O_2 ratios, finding no difference in growth when the supply of CO_2, O_2, or both was halved. A ten-fold reduction in CO_2 produced equivalent growth deficits regardless of O_2 level, while near-zero O_2 gave rise to smaller growth deficits when CO_2 was not limiting. This pattern of effects is clearly not consistent with a model in which CO_2 and O_2 physiology intersect only through a straightforward competition for Rubisco binding. I next used whole-genome microarrays to monitor the transcriptional profile of *Prochlorococcus* starved of CO_2, of O_2, or of both. Remarkably, cells deprived of exogenous O_2 showed no transcriptional difference from control cells in the 24-h period monitored. By contrast, at limiting CO_2, the presence or absence of O_2 did affect the transcriptome, most notably by allowing transcripts related to light harvesting and electron transfer to recover to basal levels after an initial downregulation. The results provide our first look at *Prochlorococcus* under these conditions and suggest that, at the irradiance studied, this strain of *Prochlorococcus*

has a carbon limitation-specific requirement for oxygen. I propose two pathways by which oxygen may act as a safety valve to protect *Prochlorococcus* cells from the light energy that carbon limitation prevents them from using.

ABBREVIATIONS

Ci, inorganic carbon; OMZ, oxygen minimum zone; CCM, carbon-concentrating mechanism; TPP, thiamine pyrophosphate; PTOX, plastoquinol terminal oxidase; SHMT, serine hydroxymethyltransferase; HEPES, N-(2-Hydroxyethyl)piperazine-N′-(2-ethanesulfonic acid); FCM, flow cytometry.

INTRODUCTION

Thirty years ago, the cyanobacterium *Synechococcus* was discovered to be widespread and abundant in the world's oceans, where it appeared in several structurally distinguishable subtypes (3, 4). A decade later, pigment analysis and shipboard flow cytometry indicated that organisms like the "type II" cells of Johnson and Sieburth in fact belonged to a novel genus, *Prochlorococcus*, remarkable for its sub-micron size and its unusual complement of photosynthetic pigments (5). *Prochlorococcus* has only grown more remarkable as our knowledge of it has deepened. We now know that *Prochlorococcus* outnumbers every other photosynthetic organism in the ocean (6); that, uniquely among phototrophs, it harvests light with divinyl chlorophyll antennae (7); that it encodes fewer genes than any other photosynthetic organism (8); and that it nonetheless has access to a seemingly endless *pan*-genome via horizontal gene transfer to and from its "islands" of niche-specific genes (9, 10).

The ocean is not one environment but many, defined by intersecting gradients of light, temperature, nutrients, oxygen, and salinity, such that marine microbes live not so much in a drop of water as (with apologies to Hutchinson (11)) in an *n*-dimensional hyperdrop, with perpetual flux across the boundaries. *Prochlorococcus* is thus not a creature of its environment but a creature of its environments, with distinct ecotypes adapted to life at high and low light levels, in warmer and cooler waters, with more or less nitrogen and phosphorus available in this or that form (9, 12-15). Although *Prochlorococcus* can live in relatively nutrient-rich coastal waters (16), in general its adaptations drive towards improving its relative fitness in oligotrophic waters: for instance, the streamlining of its genome (8) reduces its phosphorus demand, made lower still by its high baseline sulfolipid-to-phospholipid ratio (17, 18). Moreover, the very small size of *Prochlorococcus* cells gives them a higher surface-to-volume ratio, and thus greater ability to meet their reduced needs, than have their larger competitors. These and other factors combine to allow *Prochlorococcus* to keep dividing steadily in the open ocean, allowing it to reach high abundance in oligotrophic waters (6, 19) that greatly slow the growth of the less ascetic phototrophs that dominate more forgiving habitats.

Prochlorococcus at its most successful typically accounts for as much as half of net primary production at a given station (20). Thus the discovery of a "nearly monoalgal" population of *Prochlorococcus* in the Arabian Sea (1) was a noteworthy development, the more so because the habitat this population so thoroughly dominated was an oxygen minimum zone (OMZ), below the oxycline, with O_2 levels ~100-fold lower than in the surface mixed layer. Some suggested explanations for this population's success were extrinsic to *Prochlorococcus*, e.g., the hypothesis that the larger organisms that would normally keep *Prochlorococcus* populations in check by grazing cannot themselves survive sub-oxic waters , but the reason could also be a direct product of *Prochlorococcus* physiology (1, 21).

One candidate explanation is tied to the efficiency of carbon fixation by Rubisco. Rubisco can catalyze not only the on-pathway carboxylation of ribulose-1,5-bisphosphate but also its off-pathway oxygenation. The latter reaction yields one 3-phosphoglycerate and one 2-phosphoglycolate, not only costing a carbon-carbon bond (and, in the course of salvage, a phosphoester) but also producing a molecule whose accumulation is cytotoxic. But this pathway, called photorespiration, is not easy to avoid: from a structural perspective—the only perspective the enzyme has available—carbon dioxide is nearly indistinguishable from molecular oxygen, which is nearly 600-fold more abundant in the atmosphere. The absence of molecular oxygen from the terrestrial atmosphere when Rubisco first arose suggests that the oxygenation reaction would have posed little problem for the earliest carbon-fixing organisms. However, seelctive pressure to suppress the oxygenase reaction is likely to have begun to build well before the atmospheric oxygen concentration did: Raven *et al.* note that early oxygenic phototrophs growing in stromatolites would have encountered high O_2 concentrations because the stromatolite itself would greatly slow diffusive loss of photosynthetically evolved oxygen (22). But billions of years of selective pressure have neither ablated the oxygenase activity nor led to the widespread displacement of Rubisco by an alternative carbon-fixing enzyme. The persistence of the oxygenase activity testifies to the fundamental nature of the biophysical problem (23); the persistence of Rubisco and the Calvin cycle in the face of alternative carbon-fixing pathways (the reductive citric acid cycle, the reductive acetyl-CoA pathway, the 3-propionate cycle variants) may reflect either a vulnerability of enzymes in these pathways to oxidative damage or the difficulty of displacing an established multi-enzyme module so central to phototrophic metabolism (24, 25). Rather than replacing Rubisco and the Calvin cycle wholesale, many oxygenic phototrophs have instead evolved crutches for this inherently handicapped process; hence the widespread occurrence of the crassulacean acid metabolism, the C4 pathway (long known in land plants and recently discovered in a marine diatom (26)), and the cyanobacterial carbon-concentrating mechanism (CCM).

A key component of the cyanobacterial CCM is the sequestration of Rubisco in proteinaceous microcompartments called carboxysomes, which are thought to help increase the local concentration of carbon dioxide while excluding oxygen (27). We have little understanding of how well this strategy actually works, but recent work by Eisenhut et al. (28) suggests that cyanobacteria may photorespire constantly. Taking this result together with the fact that the only *Prochlorococcus* Rubisco yet characterized, from ecotype MIT9313, has unusually poor CO_2 specificity (2), we might hypothesize that living in an oxygen minimum releases *Prochlorococcus* from photorespiratory pressure, allowing even its very poor Rubisco to work efficiently because it encounters carbon dioxide far more often than it encounters oxygen. More generally, the carbon dioxide level that is limiting for *Prochlorococcus* growth, and the activity of the (unknown) pathways that signal carbon starvation (29), could be a function of the photorespiration rate. I explored this idea in two ways: by measurement of the growth physiology of *Prochlorococcus* strain MED4 under different carbon dioxide and oxygen partial pressures, and by whole-genome transcriptional profiling of the MED4 response to extreme carbon dioxide and oxygen conditions. Together, these experiments suggest that, at low levels of inorganic carbon, oxygen works not to starve the cell but to protect it, serving as a safety valve for excess light energy.

RESULTS AND DISCUSSION

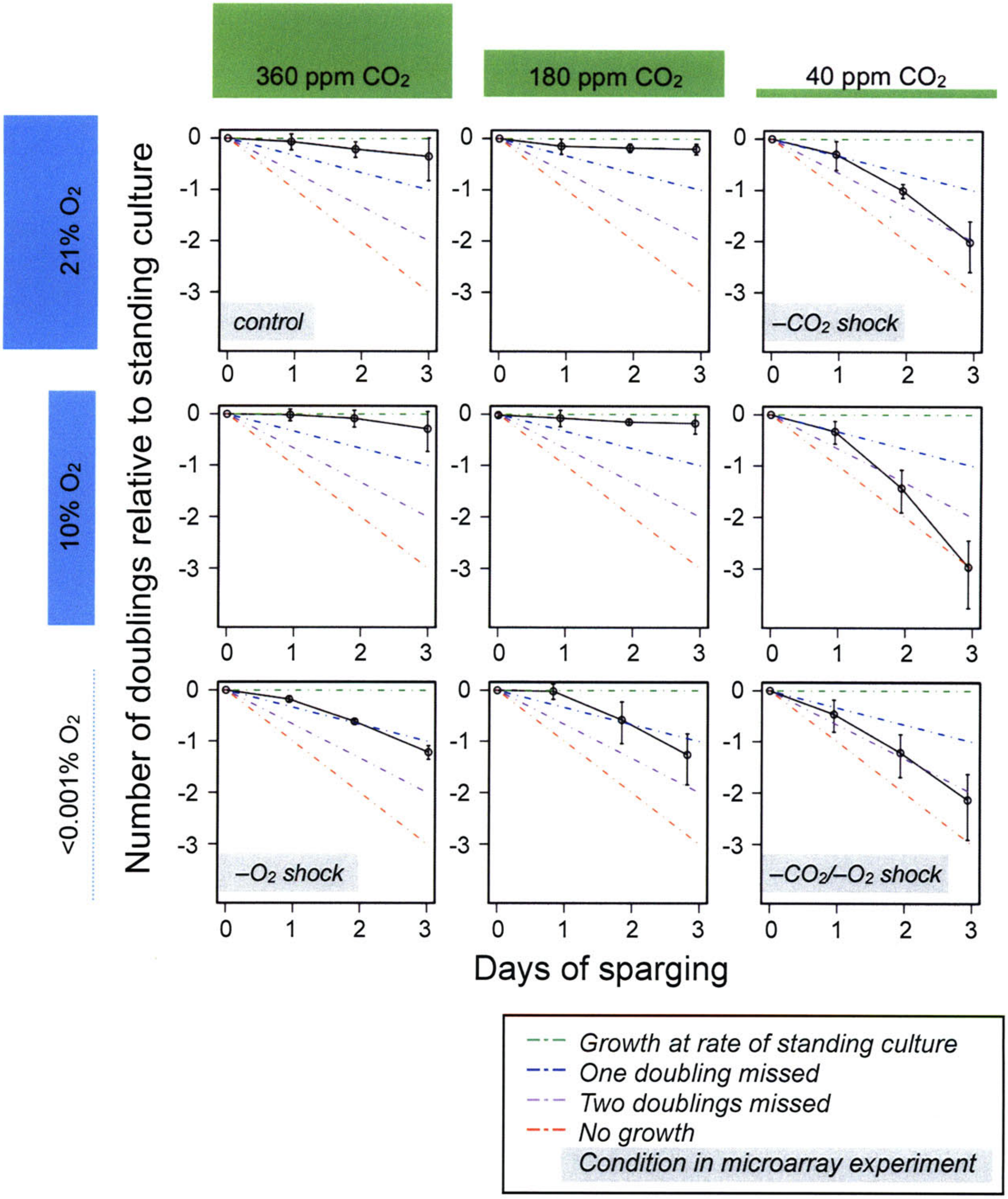
Growth rates under different steady-state CO_2 and O_2 regimes

We first asked how the growth of *Prochlorococcus* MED4 is affected by sparging with different partial pressures of CO_2 and O_2. We expected that competition between carbon fixation and photorespiration would cause abundant O_2 to heighten the effects of CO_2 limitation, leading to greater growth defects at low CO_2 partial pressure and normal levels of O_2 than at low levels of both gases. Three CO_2 concentrations (atmospheric = 360 ppm, 180 ppm, and 40 ppm) and three O_2 concentrations (atmospheric = 21%, 10%, and <0.001%) were examined in combination to provide nine growth conditions.[1] In the time required for control standing (un-bubbled) cultures to approach stationary phase, MED4 cells sparged with 180 ppm CO_2 + 21% O_2 kept pace with cells sparged with 180 ppm CO_2 + 10% O_2, cells sparged with 360 ppm CO_2 + 10% O_2, and cells sparged with air (~360 ppm CO_2 + 21% O_2; Fig. 1 and Appendix 1, Fig. 1). That 180 ppm CO_2 + 21% O_2 produced no growth deficit and 360 ppm CO_2 + 10% O_2 no advantage over air (in all, a four-fold change in the ratio between the gases) suggests either that MED4 has a very effective carbon-concentrating mechanism, such that Rubisco is fully sheltered from these changes, or that the cell can bear the energetic cost of a moderate amount of photorespiration.

More extreme gas conditions did elicit some differences in growth. After keeping pace initially, cells grown at <0.001% O_2 and 180 ppm or 360 ppm CO_2, fare worse than cells grown at 10% or

[1] The different units used for the two gases reflect the different scales of their atmospheric abundance: the atmospheric concentration of O_2 is 210000 ppm = 21%, while the current CO_2 level is ~360 ppm = 0.036%. The level of dissolved oxygen in OMZs is difficult to quantify, being below the range of commonly used detectors, but is expected to be roughly a hundredfold lower than the concentration in surface waters: very low, but at least an order of magnitude less extreme than our <0.001% O_2 conditions.

Figure 1. Growth of *Prochlorococcus* MED4 under different CO_2 and O_2 tensions. Cultures in the 360 ppm CO_2 + 21% O_2 condition (upper left) were sparged with room air; cultures in the remaining conditions were sparged with gas from pre-mixed cylinders in which N_2 was the carrier gas for the stated concentrations of CO_2 and O_2. Cells were inoculated from standing (unbubbled) cultures at –1 day and allowed to grow without sparging for 24 h. Cell counts were determined by flow cytometry. Results are plotted as the mean number of doublings under each condition (3–8 replicates) relative to paired standing (unbubbled) cultures (3 replicates; see Methods for details of calculations). A culture growing exactly as fast as the paired standing cultures, achieving roughly three doublings in three days, would lie on the green dashed line; a culture achieving two doublings, on the blue dashed line; one doubling, on the purple dashed line; and a culture achieving no increase in cell count, on the red dashed line. Error bars represent one standard deviation.


360 ppm CO2
180 ppm CO2
40 ppm CO2
21% O2
10% O2
<0.001% O2
Number of doublings relative to standing culture
0
-1
-2
-3
0
1
2
3
control
–CO2 shock
–O2 shock
–CO2/–O2 shock
Days of sparging
Growth at rate of standing culture
One doubling missed
Two doublings missed
No growth
Condition in microarray experiment

21% O_2 at these carbon dioxide levels. This is particularly interesting in that carbon-replete cells grown under constant light should have an active photosynthetic electron transport chain, with the water-splitting reaction at photosystem II providing a steady supply of O_2; either this supply is intrinsically inadequate to meet some cellular need, or sufficient oxygen is produced but then lost too rapidly to diffusion to be of use. Unlike growth at very low O_2 partial pressue, growth under 40 ppm CO_2 offers no initial respite, as cells grown at 40 ppm CO_2 and any level of O_2 had already fallen behind the more CO_2-replete cultures within 1 d. Cell division does continue even in these straitened circumstances, but at a roughly three-fold lower rate. Notably, growth was no more impaired at 40 ppm CO_2 + 21% O_2 than at 40 ppm + <0.001% O_2. That all three 40 ppm CO_2 conditions impose carbon limitation is clear; that O_2 tension does not interact with carbon limitation in these cultures in the hypothesized way seems equally so. Still, we cannot rule out the possibility that, in the gulf between unimpeded growth at 180 ppm CO_2 and carbon limitation at 40 ppm CO_2, and above the <0.001% O_2 that gives growth defects whatever the carbon supply, there is a CO_2 concentration range in which increasing oxygen increases the effective carbon limitation.

Global transcriptional responses to acute changes in CO_2 and O_2 tension

To examine more closely the effects of growth under limiting CO_2 and what appears to be limiting O_2, we used custom microarrays to follow the timecourse of gene expression in the first 24 h after mid-log-phase cultures grown with air bubbling (Fig. 2) were transferred to 40 ppm CO_2 + 21% O_2 (hereafter, –CO_2 shock), 40 ppm CO_2 + <0.001% O_2 (–CO_2/–O_2 shock), and 360 ppm CO_2 + <0.001% O_2 (–O_2 shock). Microarrays are a powerful tool, providing far more thorough data on the transcriptome than state-of-the-art proteomics and metabolomics methods can in their respective domains. But that thoroughness comes at the cost of an extra degree of remove from the real cellular economy: given the confounding factors of post-transcriptional and post-translational regulation, RNA turnover, and protein turnover, RNA levels cannot strictly be assumed to correlate with protein levels, nor enzyme levels with metabolic flux (30, 31). The connection between RNA levels and protein levels seems particularly pressing in a slow-growing organism: if *Prochlorococcus* essentially lives in slow motion, it would be particularly inappropriate to extrapolate from RNA levels to protein levels at a given timepoint. However, current data suggests that that extrapolation is not unusually hazardous: the median mRNA half-life in *Prochlorococcus* MED4 is even shorter than that in the much faster growing *E. coli* (32). The connections between enzyme levels and pathway fluxes are entirely unmapped in *Prochlorococcus*, and the possibility of nonlinear relationships between the two must be borne in mind in interpreting microarray data for enzyme-coding genes. In addition, it is extremely important to gate results stringently to keep the rate of false positives low. We set two conditions for assigning significant differential expression: a *q*-value (a measure of the false discovery rate) of ≤ 0.01, and at the same timepoint a ≥ 2-fold mean change in expression with respect to t = 0. Changes in expression that met one of these conditions but not the other were considered borderline significant.

The first result to emerge from our 24-h timecourse was as surprising as it was unmistakable: under $-O_2$ shock, not a single gene (and only one intergenic region[2]) showed significant

[2] Some intergenic regions of MED4 have been identified as genes for non-coding RNAs (33), but the sole $-O_2$ shock positive result, from IG1529f, is not among them.

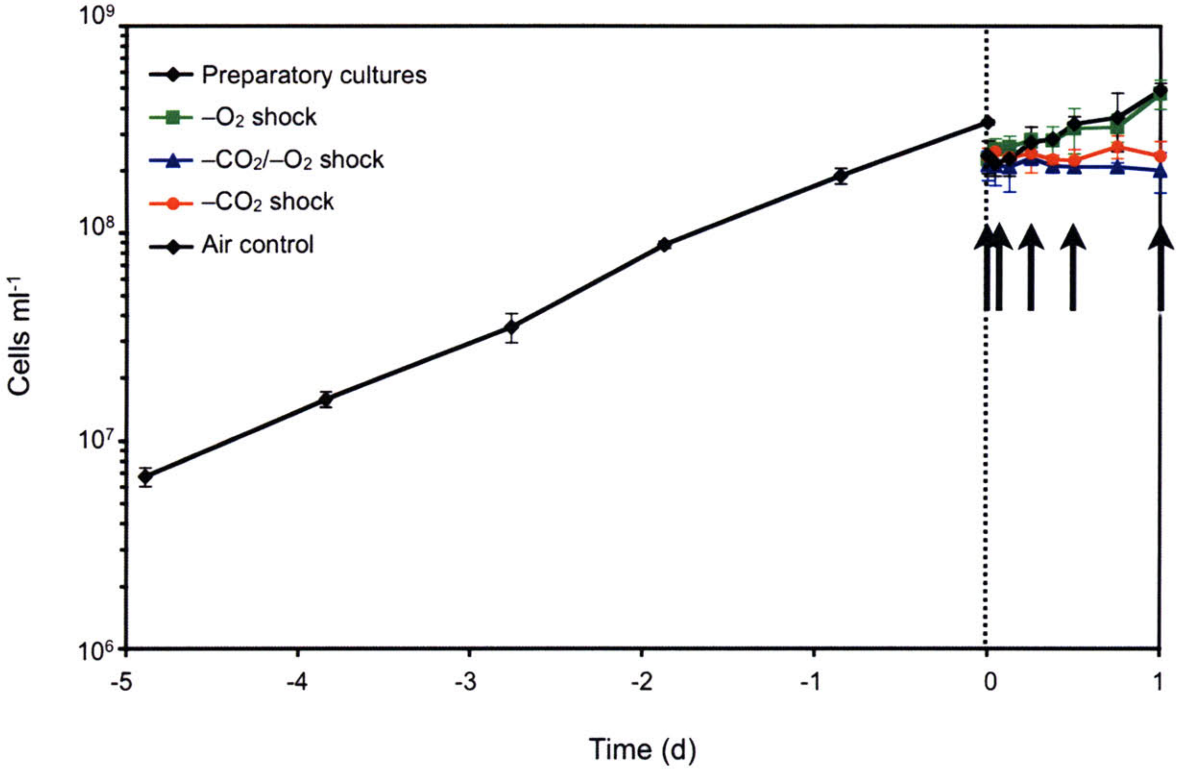
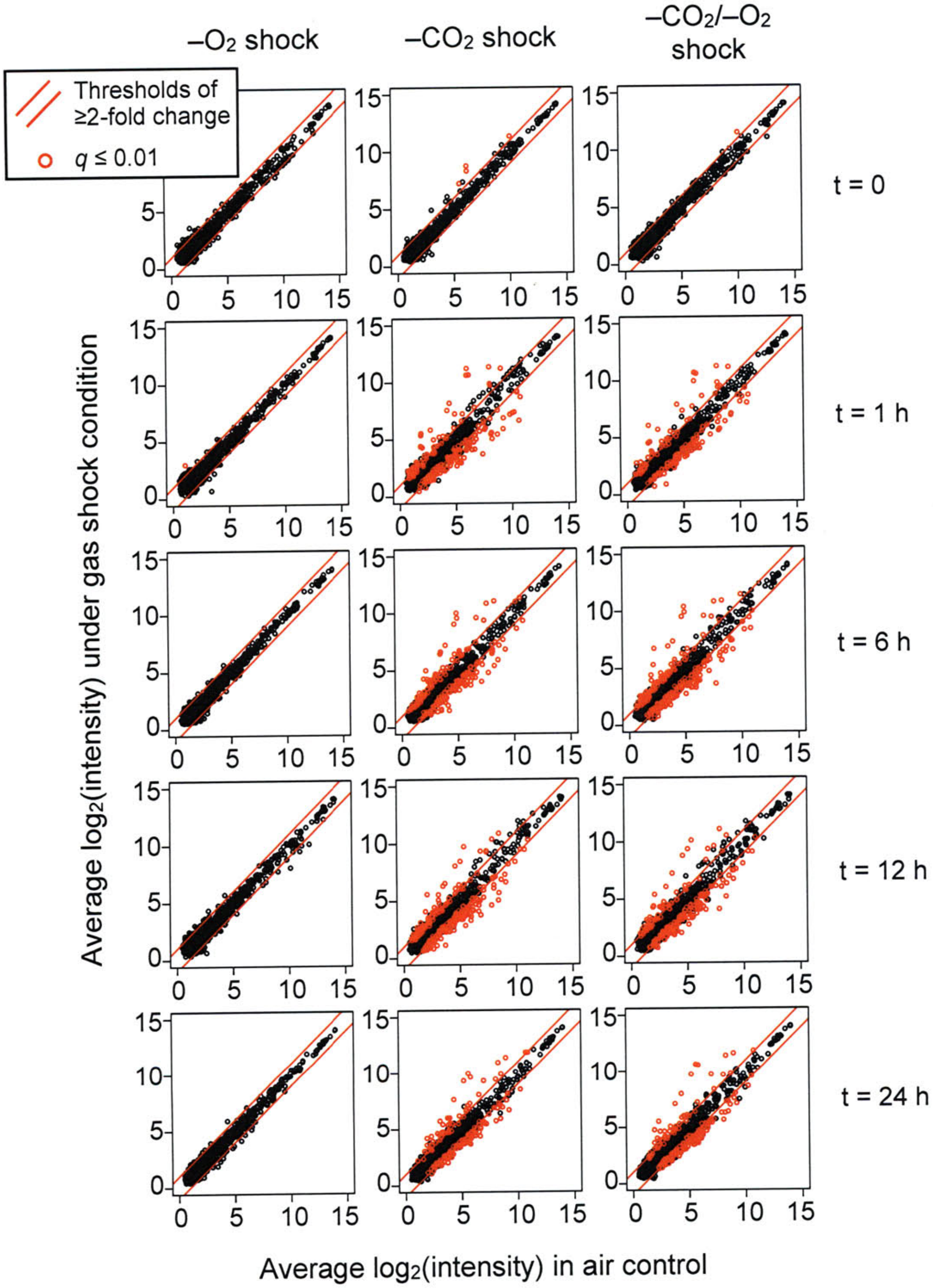


Figure 2. Growth of cultures used for gas shock microarrays. Three triplicate cultures were grown with air bubbling for 5 days prior to t = 0. At t = 0 (dotted line), cultures were collected by pelleting and resuspended in one of four conditions: control (air bubbling), $–O_2$ (360 ppm CO_2 + <0.001% O_2), $–CO_2$ (40 ppm CO_2 + 21% O_2), and $–CO_2/–O_2$ (40 ppm CO_2 + <0.001% O_2). Array hybridization was performed for samples taken at the timepoints indicated by black arrows (t = 0, 1 h, 6 h, 12 h, and 24 h).

differential expression with respect to the control, air-bubbled cells (Fig. 3), despite the growth deficit the cells show when subjected to the –O_2 shock condition for >24 h (Fig. 1). As noted above, cells under this condition should continue to see some molecular oxygen by virtue of continued photosynthesis, but the concentration gradient of oxygen across the cell membrane should keep intracellular oxygen levels relatively low. Lower than usual oxygen levels, in turn, might slow but not halt oxygen-dependent reactions, like the reactions of dioxygenases, so that pathways like tryptophan metabolism and porphyrin biosynthesis perform at levels that keep the cell from dividing at the usual pace but that do not signal a crisis. In this case, the transcriptional response might grow stronger with time, as gradually mounting metabolic imbalances begin to demand a new regulatory program; repeating this experiment with a 48 h or 72 h timepoint would allow us to test this conjecture. In addition, the growth curve should be repeated with more frequent sampling and flow cytometric cell-cycle analysis should be performed to fill in our picture of the time-dependent development of the physiological response to –O_2 shock. The alternative explanation that we failed to manipulate O_2 levels as desired is unlikely: the gas cylinder used had been individually assayed by the manufacturer and was certified to contain <0.001% O_2; the capacity of our experimental set-up to manipulate O_2 levels in culture is evident in the clear, and clearly oxygen-related, differences we observe between the –CO_2 and –CO_2/–O_2 shock conditions, discussed at length below; and we observed a drop in mean chlorophyll fluorescence per cell under –O_2 shock and not in the air control (Appendix 1, Fig. 3), indicating at the least that the conditions in these two sets of replicates were not the same. However, future work should incorporate a real-time measurement of oxygen levels in the experimental bottles.

Figure 3. Global comparison of microarray probeset intensities under –O_2 shock (360 ppm CO_2 + <0.001% O_2), –CO_2/–O_2 shock (40 ppm CO_2 + <0.001% O_2), and –CO_2 shock (40 ppm CO_2 + 21% O_2) vs. air. Red lines indicate two-fold change. Probesets with q-value < 0.01 (an indicator of the false discovery rate) are plotted in red. Genes required both ≥ 2-fold change and q-value < 0.01 to be called significantly differentially expressed. The t = 0 samples were collected immediately after air-grown cells were pelleted and resuspended in medium pre-equilibrated with the appropriate gas.


–O$_2$ shock
–CO$_2$ shock
–CO$_2$/–O$_2$ shock
Thresholds of ≥2-fold change
$q \leq 0.01$
t = 0
t = 1 h
t = 6 h
t = 12 h
t = 24 h
Average log$_2$(intensity) under gas shock condition
Average log$_2$(intensity) in air control
0
5
10
15

The extent of MED4's transcriptional response to the first 24 h of both carbon-limitation conditions is diametrically opposed to that of the carbon-replete cells that experienced $-O_2$ shock. Even with strict control of the false discovery rate, I find 165 genes to be significantly differentially expressed with respect to the air-bubbled cultures under $-CO_2$ shock, and 153 genes under $-CO_2/-O_2$ shock. The overlap between these two shock conditions is extensive, but it is far from complete: only 101 genes are significantly differentially expressed under both conditions, and often with intriguing differences in their expression pattern over time. In the sections that follow, I focus on the cellular schemes suggested by the results from these two shock conditions.

Rapid transcriptional regulation of the photosynthetic electron transport chain

The *psa* (photosystem I) and *psb* (photosystem II) genes are highly expressed in healthy *Prochlorococcus* cells, a reflection of the high rates of photodamage the photosystem proteins (most notably PsbA and PsbD) are subjected to (34). Under $-CO_2$ and $-CO_2/-O_2$ shock, MED4 cells rapidly suppress transcription of 8 of the 11 *psa* genes; a photosystem I assembly gene; the *pcb* gene, encoding the chlorophyll-binding antenna; and 5 of the 18 *psb* genes included on our microarray (Fig. 4a, b, g, h). The reaction-center genes *psbA* and *psbD* are not among the photosystem II proteins whose expression is significantly reduced, and expression of the reaction-center gene *psbN* shows an insignificant decline at 1 h and then actually increases significantly thereafter. Interestingly, the pattern of photosystem gene expression over time differs between cells under $-CO_2$ shock and those under $-CO_2/-O_2$ shock: to a very large extent, the former recover to basal or near-basal expression levels by t = 6 h. This rapid modulation of

RNA levels is likely to be evident at the protein level too; high rates of photodamage mean that photosynthesis proteins (most notably in the photosystem II reaction center) must be turned over rapidly (34), such that the cell must continually draw on these genes' transcripts to replenish the protein pool. By contrast to the widespread transcriptional recovery of photosystem genes under $-CO_2$ shock, recovery of photosystem expression under $-CO_2/-O_2$ shock is confined to a few genes, and the median expression level remains steady and low throughout the experiment (Fig. 4c, i). Nor is this pattern—recovery under $-CO_2$ shock, sustained suppression under $-CO_2/-O_2$ shock—confined to the photosystems themselves; it is also clearly in evidence in the genes encoding cytochrome b_6f, through which electrons flow en route from photosystem II to photosystem I (Fig. 4d–f). The significance of the difference between the time profiles of photosystem genes under these two conditions is confirmed by clustering analysis (Figs. 5–6). These consistent differences along the length of the photosynthetic electron transport chain strongly suggest that the $-CO_2$ cells, and not the $-CO_2$ /$-O_2$ cells, have access to some form of safety valve for dispersing the light energy or reducing equivalents they continue to gather.

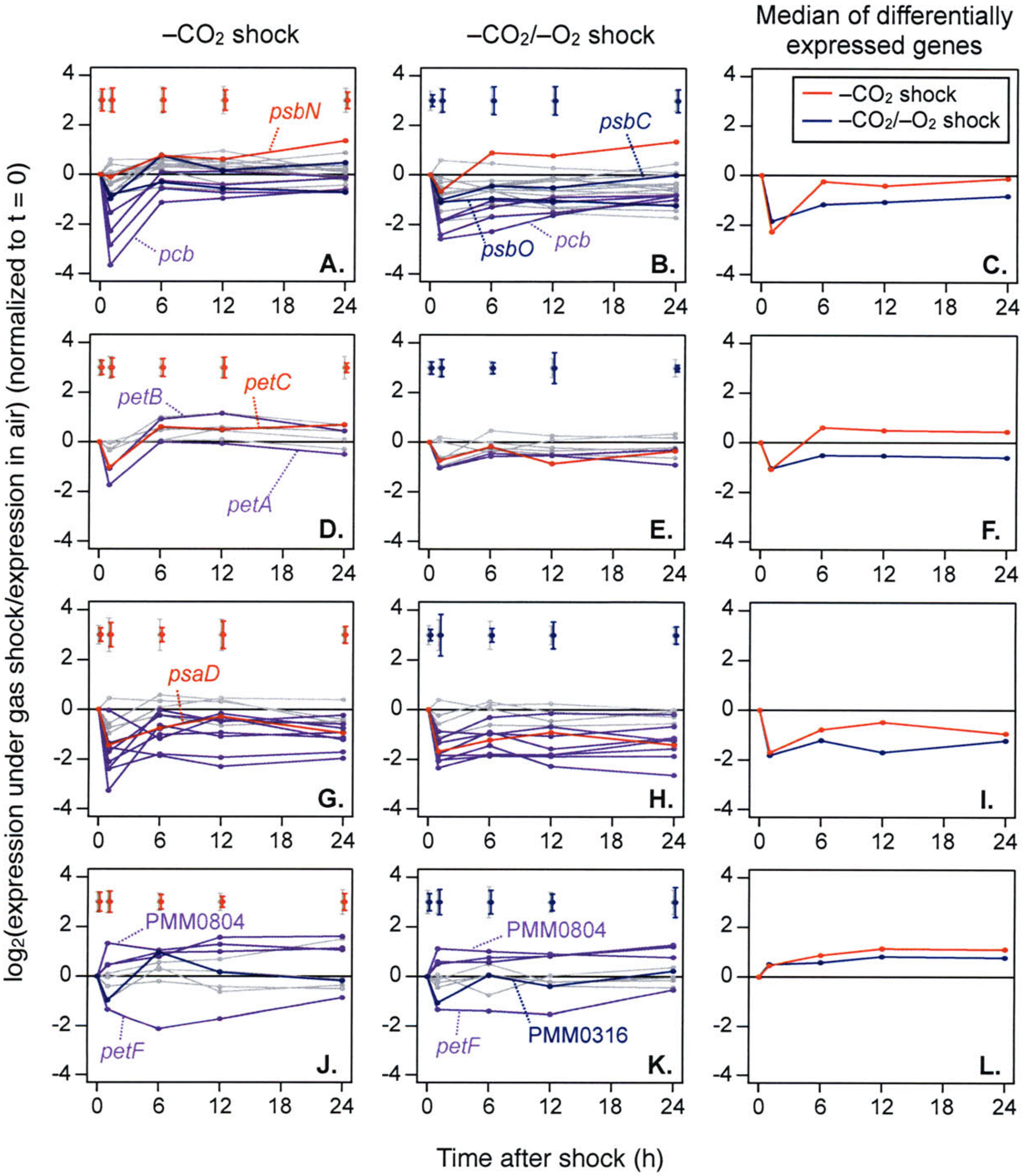
Iron-related genes

With the exception of the cytochrome b_6f genes discussed above, there is strong agreement in the patterns of iron-related gene expression under $-CO_2$ and $-CO_2/-O_2$ shock (Fig. 4j–l). In both conditions, expression of ferredoxin (PMM1449, *petF*) is strongly suppressed, while expression of the ferredoxin-NADP oxidoreductase (PMM1075, *petH*) is significantly upregulated. Interestingly, under both conditions we find significant upregulation of the iron-sequestration protein ferritin (PMM0804) and of one of MED4's two ferric uptake regulator genes (PMM0637, *fur*). The Fur protein is unusual among regulatory proteins in that it is typically present in the cell in high copy number; thus, though each subunit only binds one Fe(III) ion, the Fur population can contribute substantially to iron sequestration in its own right (35). The need for this additional iron-binding capacity is likely tied to the sharp initial downregulation of the photosystems: as damaged photosystems go unreplaced, the iron they release must be stored somewhere to prevent unchecked damage from the Fenton reaction. Under this hypothesis, however, it is not obvious why cells under –CO_2/–O_2 shock do not show a return to baseline ferritin and Fur expression levels once photosystem transcription has largely recovered.

A soft landing for the Calvin cycle

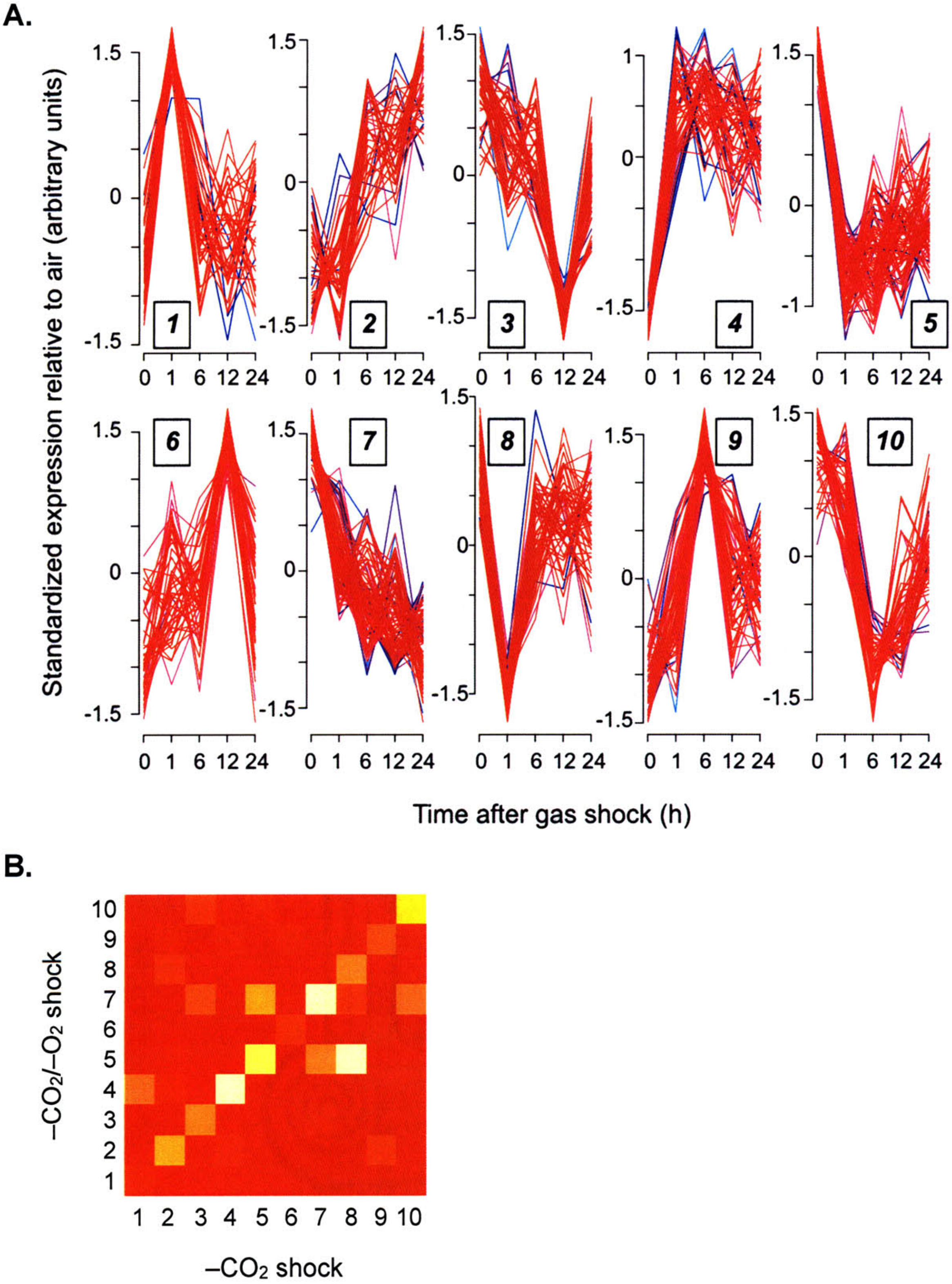
The downregulation of the Calvin cycle under inorganic carbon limitation is remarkably thorough. Twelve gene products catalyze the 14 reactions from the dehydration of bicarbonate to the regeneration of ribulose-1,5-bisphosphate. Under –CO_2/–O_2 shock, seven are significantly downregulated and the remaining five are borderline significant; similarly, –CO_2 shock yields significant downregulation of six genes and borderline significant downregulation of one more (Fig. 7a, b, e). The carboxysome shell gene PMM0552 (*csoS2*) is also significantly downregulated under –CO_2/–O_2 shock, and downregulation of the shell gene PMM0549 (*csoS1*) is borderline significant under both carbon-limitation conditions. Though thorough, these responses are not outstandingly swift or strong; the median expression levels of genes downregulated under both conditions bottom out at just over two-fold below the control. Expression of these genes develops over the timecourse with a remarkably unified profile, with the sharp drop and equally quick recovery of PMM0195 (*pgk*, encoding phosphoglycerate

Figure 4. Transcriptional responses of genes encoding **(A–C)** photosystem II proteins, **(D–F)** components of cytochrome b_6f, **(G–H)** photosystem I proteins, and **(J–K)** iron-related proteins. **(A, B, D, E, G, H, J, K)** Expression of individual genes in each functional category under $–CO_2$ shock (left) and $–CO_2/–O_2$ shock (middle). Genes in each functional class that were not differentially regulated under either shock are shown in gray; under $–CO_2$ shock only, in red; under $–CO_2/–O_2$ shock only, in blue; and under both, in purple. For clarity, error bars are not shown for every curve. The error bars shown in each panel are an estimate of error in each timepoint's measurements: in gray, the mean of the standard deviations of all genes in the functional category shown that are not differentially expressed under either shock; in red (left panels), the mean of the standard deviations of (red + purple) genes differentially expressed under $–CO_2$ shock; and in blue (middle panels), the mean of the standard deviations of (blue + purple) genes differentially expressed under $–CO_2/–O_2$ shock. Labeled genes are discussed in the text. **(C, F, I, L)** Median $\log_2$ fold change in expression levels of the (C) photosystem II, (F) cytochrome b_6f, (I) photosystem I, and (L) iron-related genes that are differentially regulated under each condition. Dashed red line, $–CO_2$ shock; dashed blue line, $–CO_2/–O_2$ shock. In (L), *petF* is excluded from the calculation of the median. Significant differential expression of *psbC, psbO*, and the possible ferredoxin gene PMM0316 was observed only under $–CO_2/–O_2$ shock; significant differential expression of *psbN, petC*, and *psaD*, only under $–CO_2$ shock. Both conditions gave significant differential expression of, in photosystem II, *psbB, psbJ*, and *psbT*; in cytochrome b_6f, *petA* and *petB*; in photosystem I, *psaA, psaB, psaF, psaI, psaJ, psaK, psaL*, assembly protein gene *ycf37*, and chlorophyll-binding antenna gene *pcb*; and, among iron-related genes, *petF, petH, fur*, and PMM0804 (ferritin).



–CO2 shock
–CO2/–O2 shock
Median of differentially expressed genes
–CO2 shock
–CO2/–O2 shock
psbN
pcb
A.
psbC
psbO
pcb
B.
C.
petB
petC
petA
D.
E.
F.
psaD
G.
H.
I.
PMM0804
petF
J.
PMM0804
petF
PMM0316
K.
L.
log2(expression under gas shock/expression in air) (normalized to t = 0)
Time after shock (h)
4
2
0
-2
-4
0
6
12
18
24

Figure 5. Clustering analysis of genes differentially expressed under –CO_2 and –CO_2/–O_2 shocks. **(A)** Gene expression profiles showing a significant response to –CO_2 shock were pooled with those showing a significant response to –CO_2/–O_2 shock; thus, the pooled set included two profiles for each gene that was differentially expressed under both conditions. The pooled set was subjected to fuzzy *c*-means clustering, revealing 10 distinct and well-represented patterns of gene expression. **(B)** For the genes that were differentially expressed under both conditions, the cluster membership of the –CO_2 shock profile was compared to that of the –CO_2/–O_2 shock profile. Off-diagonal peaks indicate sets of genes whose expression profiles under the two conditions differ. The strongest off-diagonal peak is at the intersection of clusters 8 and 5: genes that, under –CO_2 shock, were rapidly downregulated before returning to basal levels (cluster 8), and that, under –CO_2/–O_2 shock, were rapidly downregulated and remained suppressed throughout the timecourse (cluster 5).


A.
Standardized expression relative to air (arbitrary units)
1
2
3
4
5
6
7
8
9
10
1.5
0
-1.5
1
-1
0 1 6 12 24
Time after gas shock (h)
B.
–CO2/–O2 shock
1 2 3 4 5 6 7 8 9 10
–CO2 shock

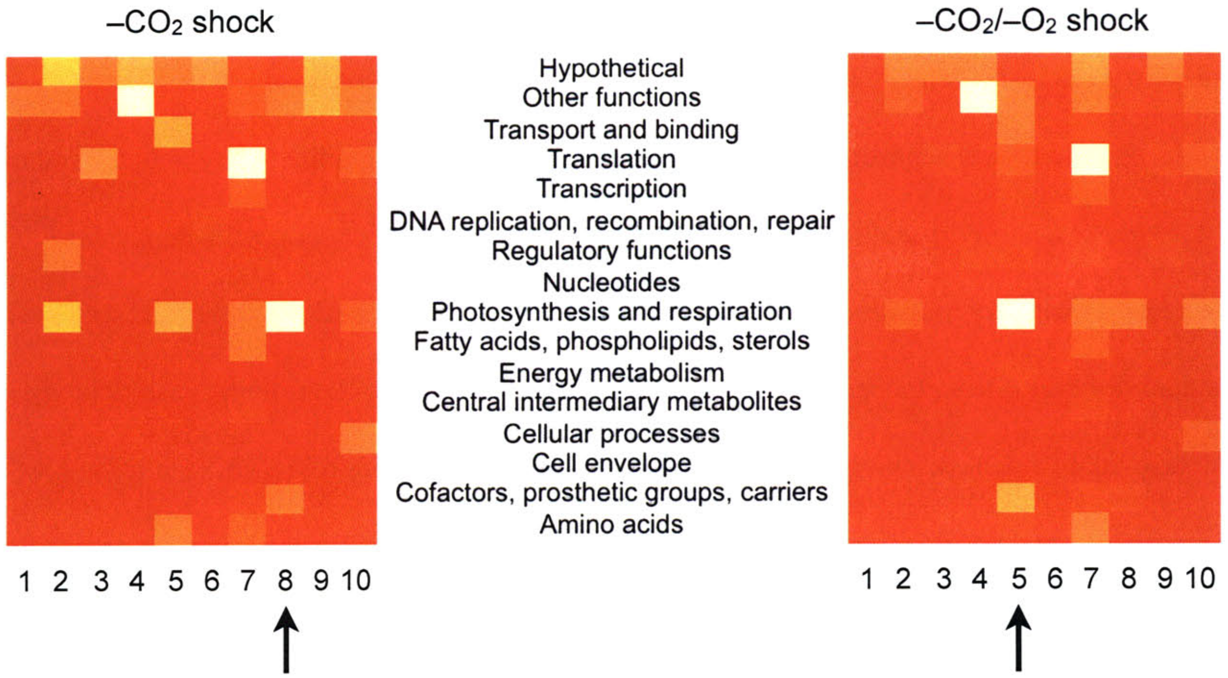
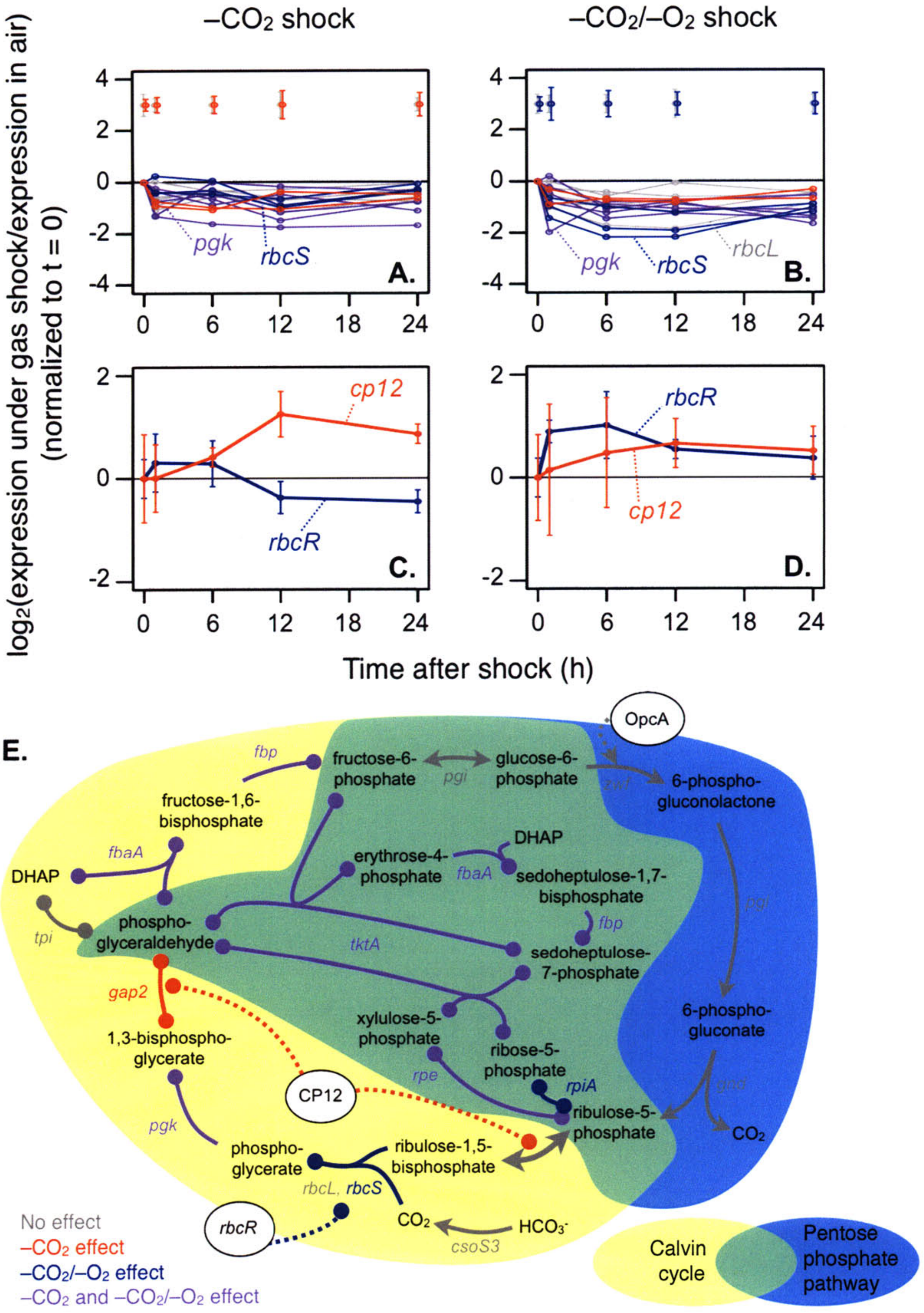


Figure 6. Breakdown of cluster membership by gene function as annotated in CyanoBase (http://genome.kazusa.or.jp/cyanobase/), showing that the major off-diagonal peak in Fig. 5b comprises genes involved in photosynthesis and respiration (Fig. 4a–i) and in cofactor (here, chlorophyll) metabolism (Fig. 11e–g).

Figure 7. Downregulation of the Calvin cycle under –CO_2/–O_2 and –CO_2 shock. **(A, B)** Transcriptional responses of Calvin-cycle enzymes, with colors and error bars as in Figure 4. **(D)** –CO_2 shock yields significant upregulation of the *gap2* and *prk* inhibitor *cp12* and not of the putative Rubisco transcriptional regulator *rbcR*. **(C)** –CO_2/–O_2 shock produces the opposite effect: significant upregulation of *rbcR* and not of *cp12.* Error bars in (C) and (D) represent one standard deviation (from the mean of two biological replicates). **(E)** The Calvin cycle (yellow) and the oxidative pentose phosphate pathway (light blue) share a series of enzymes and intermediates (green). Reaction arrows and gene names are colored according to the same scheme used in Figure 3: red if differentially expressed only under –CO_2 shock, dark blue if differentially expressed only under –CO_2/–O_2 shock, purple if differentially expressed under both, and gray if differentially expressed under neither. Circular arrowheads indicate downregulation/inhibition. Regulatory interactions are indicated by dashed lines. For clarity, the pentose-phosphate pathway reaction catalyzed by transaldolase (*tal*, borderline significant upregulation under –CO_2 shock) is not shown; it is the disproportionation that interconverts (fructose-6-phosphate + erythrose-4-phosphate) and (phosphoglyceraldehyde + sedoheptulose-7-phosphate). Both –CO_2 shock and –CO_2/–O_2 shock show widespread downregulation of the Calvin-cycle enzymes, while expression of the enzymes contributing exclusively to the oxidative pentose pathway is unchanged.


–CO2 shock
–CO2/–O2 shock
log2(expression under gas shock/expression in air)
(normalized to t = 0)
4
2
0
-2
-4
pgk
rbcS
A.
0
6
12
18
24
4
2
0
-2
-4
pgk
rbcS
rbcL
B.
0
6
12
18
24
2
0
-2
cp12
rbcR
C.
0
6
12
18
24
2
0
-2
rbcR
cp12
D.
0
6
12
18
24
Time after shock (h)
E.
OpcA
fbp
fructose-6-phosphate
pgi
glucose-6-phosphate
zwf
6-phospho-gluconolactone
fructose-1,6-bisphosphate
DHAP
erythrose-4-phosphate
fbaA
sedoheptulose-1,7-bisphosphate
fbaA
DHAP
pgl
fbp
tpi
phospho-glyceraldehyde
tktA
sedoheptulose-7-phosphate
gap2
xylulose-5-phosphate
6-phospho-gluconate
1,3-bisphospho-glycerate
ribose-5-phosphate
rpe
rpiA
gnd
CP12
ribulose-5-phosphate
pgk
CO2
phospho-glycerate
ribulose-1,5-bisphosphate
rbcL, rbcS
rbcR
CO2
HCO3-
csoS3
No effect
–CO2 effect
–CO2/–O2 effect
–CO2 and –CO2/–O2 effect
Calvin cycle
Pentose phosphate pathway

kinase) the only exception to the pattern of gentle decline (Fig. 7). The metabolic logic for this pattern of *pgk* expression is not clear.

The only genes associated with the Calvin cycle to show significant upregulation are two genes that encode regulators. PMM0147 encodes *rbcR,* a putative transcriptional regulator of Rubisco; it is significantly upregulated under $-CO_2/-O_2$ shock, and not under $-CO_2$ shock (Fig. 7c). In light of this difference, the annotated role of PMM0147 is clearly consistent with the observation that the small subunit of Rubisco (PMM0551, *rbcS*) is significantly downregulated, and the large subunit (PMM0550, *rbcL*) borderline significantly, under $-CO_2/-O_2$ shock, while neither subunit is affected by $-CO_2$ shock. Rubisco expression levels in cyanobacterial carbon-limitation experiments have rarely behaved as expected, sometimes staying steady or decreasing even when cellular carboxysome content increases (36). In our work, given the concordant downregulation of genes encoding the rest of the Calvin cycle, it is not surprising in itself that Rubisco expression should be repressed; the surprise is that the effect should be significantly stronger under $-CO_2/-O_2$ shock than it is under $-CO_2$ shock. By this metric, carbon starvation appears to be sensed more acutely by cells experiencing a $\geq$4:1 ratio of CO_2 to O_2 than by cells experiencing a ratio of 1:>5000. This result runs directly counter to the longstanding hypothesis that carbon starvation is sensed through the balance between Rubisco's on-pathway carboxylase reaction and its seemingly wasteful oxygenase reaction (29), and suggests that Rubisco may continue to have a productive role in $-CO_2$ cells. A possible role is discussed in the "Safety valves" section below.

The second regulatory gene of note is PMM0220, encoding the regulator CP12. CP12 inhibits two reactions in the Calvin cycle, the reduction of 1,3-bisphosphoglycerate to glyceraldehyde-3-phosphate and the phosphorylation of ribulose-5-phosphate to ribulose-1,5-bisphosphate. As discussed by Zinser *et al.* (37), the Calvin cycle and the oxidative pentose phosphate pathway seem to be hopelessly intertwined on the metabolic map: the two pathways share six reactions and five enzymes, but the Calvin cycle requires those reactions to run in one direction and the pentose phosphate pathway the other (Fig. 7e). Lacking the intrinsic redox control of the plant enzymes (38), cyanobacteria solve the problem in part by temporal compartmentalization and by coordinating two effectors, CP12 and the allosteric effector OpcA (39, 40), to shunt the pathways' common intermediates in the appropriate direction (37). Notably, the redox-regulated CP12 becomes an active inhibitor at night, when the cellular milieu is relatively oxidizing (37, 39). We find significant upregulation of CP12 expression by t = 12 h under $-CO_2$ shock, suggesting that MED4 under $-CO_2$ shock is attempting to inhabit a normal nighttime-like state (Fig. 7d). Just as the photosystem recovery was largely missing under the $-CO_2/-O_2$ condition, so too is the subsequent induction of CP12. Though $-CO_2/-O_2$ cells are not expected to fix much (if any) carbon, their not-fixing of carbon likely takes place in a much less oxidizing environment than is typical of cyanobacteria at night. In consequence, some of their usual habits of regulation, like CP12-mediated inhibition of the Calvin cycle, cannot be made to apply.

Downregulation of ATP synthase

The many genes contributing directly to ATP synthesis respond in concerted fashion to $-CO_2$ shock and $-CO_2/-O_2$ shock alike. The products of ten genes (PMM1438–1439 and PMM1450–

1457) are annotated as belonging to the F_oF_1 ATP synthase complex. All ten are significantly downregulated following $-CO_2/-O_2$ shock; eight of the ten are significantly downregulated following $-CO_2$ shock, and the remaining two (PMM1457 and PMM1443, encoding *atpF* and *atpI*) are borderline significant (Fig. 8c, d). The extent and development of this downregulation over time is similar between the two low-CO_2 conditions, in contrast to the differences seen between these conditions for the photosystem genes' response. The magnitude of this response is comparable to the fold change from the daytime peak to the evening trough of ATP synthase diel expression (37). Zinser *et al.* hypothesize that ATP synthase expression levels are tied to the maximum level of demand that carbon fixation will place on the system (rather than to the maximum potential ATP yield from aerobic respiration at night), hence the drop at the end of daylight hours. Although the cultures used here were asynchronous, the shock conditions applied should leave the cells in little doubt that, for the time being, they are through fixing carbon, and that ATP synthase abundance can be scaled back to the level that suffices for other, less energy-hungry metabolic pathways.

Downregulation of transport systems

Without inorganic carbon to fix, one straightforward way for the cell to conserve energy, conserve raw materials, and keep its nutrient levels relatively balanced is to cut off its imports. In line with this expectation, a number of genes annotated as encoding transport proteins are significantly downregulated under either or both low-carbon shocks, most often by t = 1 h. This includes genes encoding components of possible Mn^{2+} ABC transporters (PMM0603, PMM1029, PMM0710); *amt1* (PMM0263), encoding the ammonium transporter AmtB, one of

Figure 8. Transcriptional responses of **(A, B)** *hli* genes, **(C, D)** genes encoding ATP synthase components, **(E, F)** genes encoding known and suspected transporters, and **(G, H)** genes encoding RNA polymerase, ribosomal proteins, and translation elongation factors. Color coding and error bars are as described in Figure 4.

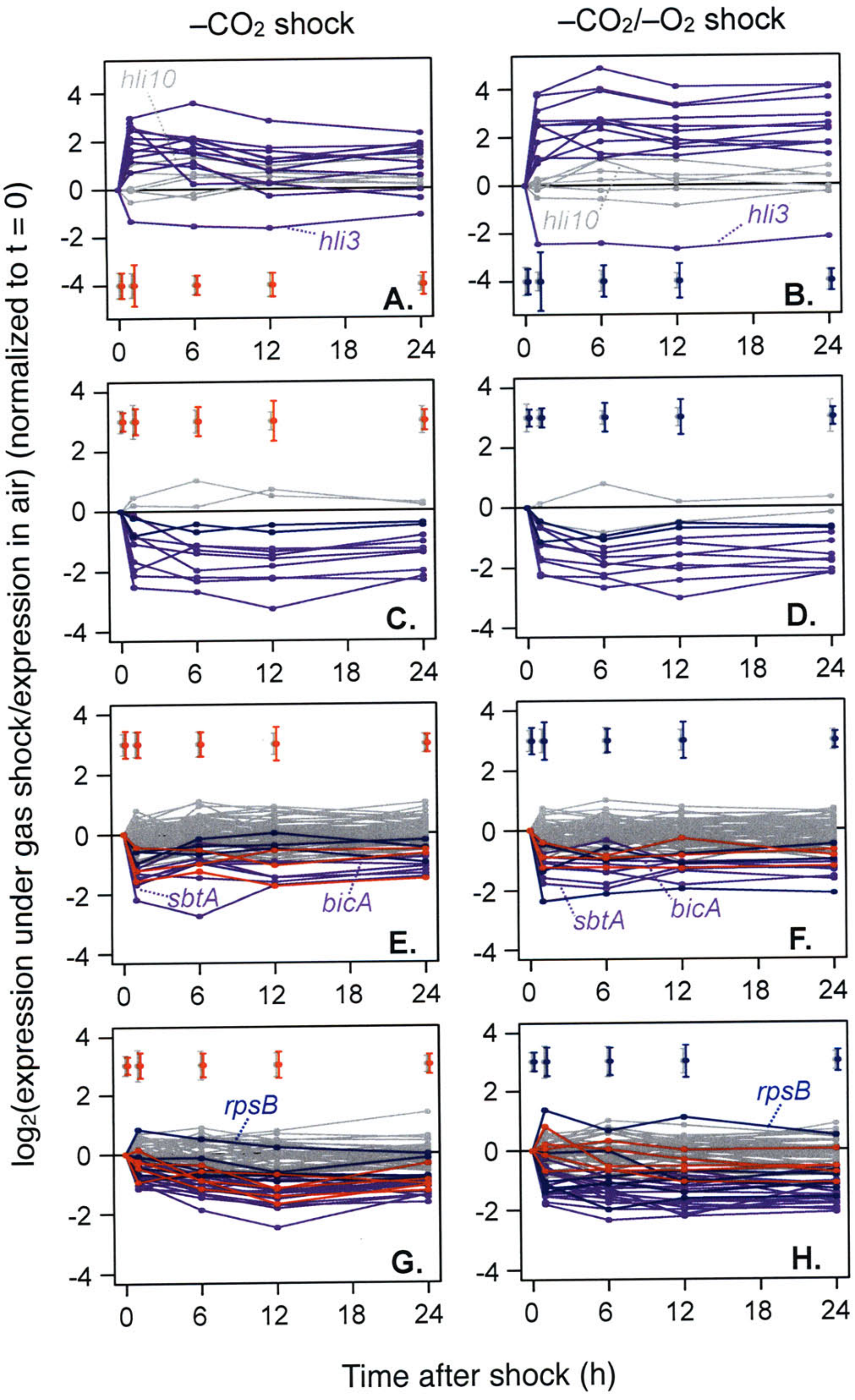

–CO2 shock
–CO2/–O2 shock
log2(expression under gas shock/expression in air) (normalized to t = 0)
hli10
hli3
A.
hli10
hli3
B.
C.
D.
sbtA
bicA
E.
sbtA
bicA
F.
rpsB
G.
rpsB
H.
4
2
0
-2
-4
0
6
12
18
24
Time after shock (h)

Table 1. Differentially expressed genes of unknown function. *, significant; ~, borderline significant.

$–CO_2/–O_2$	$–CO_2$	Locus	Annotation	COG
	*	PMM0015	domain of unknown function DUF25	COG229:Conserved domain frequently associated with peptide methionine sulfoxide reductase
*	~	PMM0051	conserved hypothetical protein	COG1192:ATPases involved in chromosome partitioning
~	*	PMM0086	conserved hypothetical protein	
*	~	PMM0121	conserved hypothetical protein	
*	~	PMM0212	conserved hypothetical protein	
*		PMM0265	conserved hypothetical protein	COG2259:Predicted membrane protein
*	~	PMM0307	hypothetical	
	*	PMM0313	conserved hypothetical protein	COG1327:Predicted transcriptional regulator, consists of Zn-ribbon and ATP-cone domains
	*	PMM0324	ctpA PDZ domain (DHR or GLGF)	COG793:Periplasmic protease
*	~	PMM0367	conserved hypothetical protein	Cellular processes:Toxin production and resistance
*		PMM0368	conserved hypothetical protein	
*	*	PMM0395	conserved hypothetical protein	
~	*	PMM0474	conserved hypothetical protein	
	*	PMM0482	phb Band 7 protein	COG330:Membrane protease subunits, stomatin/prohibitin homologs
	*	PMM0532	conserved hypothetical protein	
	*	PMM0751	conserved hypothetical protein	
	*	PMM0777	conserved hypothetical protein	Protein synthesis:Ribosomal proteins: synthesis and modification
	*	PMM0810	hypothetical	
*	*	PMM0812	hypothetical	
*	*	PMM0819	hypothetical	
	*	PMM0864	possible Fusion glycoprotein F0.	
*		PMM0875	conserved hypothetical protein	
	*	PMM0944	conserved hypothetical protein	
*		PMM0996	conserved hypothetical protein	
	*	PMM1015	conserved hypothetical protein	COG3686:Predicted membrane protein

$-CO_2/-O_2$	$-CO_2$	Locus	Annotation	COG
*	*	PMM1067	possible adenoviral fiber protein (repeat/shaft)	
*	~	PMM1071	conserved hypothetical protein	Protein synthesis:Ribosomal proteins: synthesis and modification
~	*	PMM1232	methyltransferase	
~	*	PMM1283	conserved hypothetical protein	COG670:Integral membrane protein, interacts with FtsH
	*	PMM1350	pentapeptide repeats	
*		PMM1355	conserved hypothetical protein	
	*	PMM1387	hypothetical	
*	*	PMM1400	possible Hemagglutinin-neuraminidase	
	*	PMM1416	conserved hypothetical protein	
~	*	PMM1435	conserved hypothetical protein	Protein fate:Protein and peptide secretion and trafficking
	*	PMM1478	conserved hypothetical protein	

the most highly expressed *Prochlorococcus* genes in the oligotrophic open ocean (41); and *urtABC* (PMM0970–0972), whose products together form the urea ABC transporter. This last is perhaps surprising, as urea could be imagined to serve the cell as a source of one-carbon units. However, it should be remembered that the urea cycle is a means of detoxification; with the downregulation of amino acid biosynthesis (see below) and protein translation, there is no good sink for nitrogen, and acquiring two ammonium ions for every carbon may simply be too high a price for the cell to pay for biomass.

One might also expect there to be another side of transport regulation under limiting carbon: inducing expression of transporters with high affinity for inorganic carbon could temporarily allow the cell to eke out a more normal living. Other cyanobacteria studied to date typically exhibit rapid induction of high Ci uptake affinity when external Ci levels drop (42-46). Although this physiological result is well established for many cyanobacteria and green algae, the mode (as CO_2 or HCO_3^-?) and mechanism of inorganic carbon acquisition remains a vexed topic (27, 47, 48). In *Prochlorococcus* and other marine cyanobacteria, the problem is the lack of candidate transporters. To all appearances, no sequenced *Prochlorococcus* isolate encodes a CO_2 uptake system, and prior to the identification of the BicA HCO_3^- transporter in *Synechococcus* PCC7942 (49) and recognition of its homolog PMM0214 in MED4, *Prochlorococcus* appeared to encode only the SbtA HCO_3^- transporter, and the gene was a weak *sbtA* homolog at that (50, 51). While there is some suggestion that *Prochlorococcus* strain PCC9511, closely related to MED4, has an inducible high-affinity bicarbonate uptake activity (52), the phenomenon is not well characterized and the protein (or proteins) responsible remain elusive.

We find that under both $-CO_2$ and $-CO_2/-O_2$ shock, MED4 rapidly downregulates both *bicA* (PMM0214) and *sbtA* (PMM0213), cutting off its two suspected routes for carbon acquisition (Fig. 8e, f). This is in direct opposition to repeated observations of extremely strong (30–70-fold) *sbtA* induction in carbon-limited *Synechocystis* spp. PCC6803 (36, 53). Moreover, we find repression of two putative Na^+/H^+ antiporters, PMM0472 and PMM1600. Such antiporters are hypothesized to promote inorganic carbon acquisition by Na^+/HCO_3^- symporters like BicA and SbtA (54). That these antiporters too should be downregulated under carbon limitation suggests that, if MED4 does have an as yet unrecognized high-affinity carbon uptake system, that system will not be a Na^+/HCO_3^- symporter. It has been suggested that an ABC-type cyanate transporter might be a cryptic Ci transporter (52, 55), but the candidate genes PMM0370–0372 show no significant change in expression under either $-CO_2$ or $-CO_2/-O_2$ shock. The results presented here do not rule out the induction of an unknown Ci transporter; the sets of genes upregulated under $-CO_2$ and $-CO_2/-O_2$ shock include dozens of genes of unknown function (Table 1). But *Prochlorococcus* is not expected to encounter carbon limitation often in the marine environment, so the absence of an inducible high-affinity Ci transporter would be consistent with the genome streamlining that is typical of *Prochlorococcus* (8, 9). If cells have a way of riding out the occasional carbon limitation, forgoing such an uptake pathway may be part of what seems to be the *Prochlorococcus* master plan for living where others starve: grow slowly, and abide.

Rapid differential regulation of high-light-inducible genes

As their name suggests, the high-light inducible proteins (HLIPs, encoded by *hli* genes) of cyanobacteria were originally identified in connection with light (56). HLIPs are small transmembrane proteins that resemble chlorophyll *a/b* binding proteins; their expression is induced by exposure to high light levels, and their presence is a key part of cyanobacterial survival of such light shocks (57). It is increasingly clear, however, that high light is only one of many stress conditions that evokes an *hli* response; in *Prochlorococcus* alone, *hli* induction has been observed in response to phage infection, N starvation, and to a lesser extent Fe starvation, in addition to the canonical high light response (58-61). The unifying feature of these disparate stresses may be their tendency to provoke damage to the photosystem, whether through unusually high levels of incoming excitation energy or through a reduced capacity to handle the usual photon flux. It is an open question whether HLIPs play their protective role through physical association with the photosystem (62) or mediation of pigment synthesis (63), or through some other means.

If there are many known ways to evoke *hli* gene expression in *Prochlorococcus*, there are even more *hli* genes to choose from in high-light-adapted ecotypes like MED4. MED4 encodes fully 22 *hli* genes, of which two sets of four (*hli6–9* and *hli16–19*) are, remarkably, nucleotide-perfect duplications (64). Though these are the only perfect duplications, they are not the only multi-copy *hli* genes; indeed, *hli1*, *hli2*, *hli3*, *hli13*, and *hli20* are the only single-copy high-light inducible genes in MED4. The multi-copy *hli* genes are similar to the *hli* genes found in some viruses of *Prochlorococcus* and are largely located in genomic islands, suggesting that they are part of the highly niche-dependent "flexible" genome (9, 10, 65); by contrast, the single-copy *hli* genes belong to the core genome. Intriguingly, while microarray experiments subjecting MED4 to phage infection, N starvation, altered N source, and altered light quality and quantity have all evoked strong upregulation of many or all multi-copy *hli* genes, none has produced differential expression of the single-copy *hli* genes (58, 59, 61, 66).

In general, $-CO_2$ shock and $-CO_2/-O_2$ shock yield patterns of differential *hli* expression (Fig. 8a, b) that are congruent to each other and to those seen in previous work. Both $-CO_2/-O_2$ shock and $-CO_2$ shock produce strong upregulation of all flexible *hli* genes save *hli10* (PMM1390), which is thought to be part of a specific N-stress response (59, 61), and *hli15* (PMM1128). Among the differentially expressed *hli* genes, the median fold change is smaller under $-CO_2$ shock than under $-CO_2/-O_2$ shock, perhaps indicating a less extreme stress in the former condition. Under N starvation, significant upregulation of *hli10,* the most N-responsive *hli* gene, peaks at 12 h with an 8.7-fold elevation above expression in the N-replete control. The *hli* response to C deprivation is both faster and more extreme: the most strongly upregulated *hli* gene under both carbon-limitation conditions, *hli5* (PMM1404), is immediately (t = 1 h) 14-fold more highly expressed under $-CO_2/-O_2$ shock than in the control and peaks at a remarkable 30-fold above control (t = 6 h). Even the weaker response under $-CO_2$ shock sees immediate 7.8-fold upregulation (t = 1 h) and peaks at 14-fold elevation (t = 6 h).

Strikingly, the one *hli* gene that was downregulated is part of the core genome (*sensu* (9). The downregulation of *hli3* (PMM1482; Fig. 8a, b) under carbon limitation is the first observation of significant differential expression of a core *hli* gene in *Prochlorococcus*. The drop in transcript abundance is immediate under both $-CO_2/-O_2$ and $-CO_2$ shock (t = 1 h, 5.4-fold and 2.7 fold decrease, respectively) and is sustained for the 24-h duration of the experiment. An obvious hypothesis for this strong suppression of *hli3* expression is that *hli3* transcription could be tied to that of its immediate upstream neighbor on the (-) strand, a gene from which it is separated by just 50 bp. That neighbor is PMM1483 (*rpoC2*), encoding the β' subunit of RNA polymerase; it is the third gene in an operon that also includes PMM1484 and PMM1485 (*rpoC1, rpoB*), encoding the β and γ subunits of RNA polymerase. As discussed below, RNA polymerase expression is rapidly suppressed under $-CO_2/-O_2$ and $-CO_2$ shock; if *hli3* is co-transcribed with the RNA polymerase genes, a decrease in *hli3* levels could be a byproduct of RNA polymerase transcriptional downregulation. Yet we know that *hli3* expression does not always change in lockstep with expression of the RNA polymerase; following phage infection of MED4, for instance, transcription of host RNA polymerase is strongly induced, while *hli3* transcript levels are not significantly elevated over the uninfected control (58).

Interestingly, *hli3* may have a functional parallel in *Synechocystis* spp. PCC6803. The *scp* genes of *Synechocystis* spp. PCC6803 are similar in sequence to *hli* genes and cluster with the core *hli* genes of *Prochlorococcus* rather than with the flexible *hli* genes (65). Just like the *Prochlorococcus* core *hli* genes, *scp* gene expression is not induced by high light (67). However, in *Synechocystis* spp. PCC6803 subjected to 24 h of carbon starvation, slr2542, one of the two *scp* genes whose sequences cluster most closely with MED4 *hli3,* shows significant downregulation (36). Expression of the other *scp* genes is unchanged.[3] Thus, it is tempting to

[3] It should be noted that a somewhat parallel experiment using a milder $-CO_2$ shock and a shorter timescale found upregulation of four *scp* genes (ssl 1633, ssl2542, ssr1789, and ssr2595) at one timepoint (53). However, I consider the design and analysis of this experiment to have been problematic in several respects, making its comparison with other gas-shock experiments inappropriate.

argue that the core *hli* gene *hli3* is not, as is usually thought, a stress-response gene per se, but rather encodes a protein whose normal function in photosystem maintenance or pigment maturation becomes unnecessary (or, conceivably, harmful) when carbon fixation is forcibly halted.

Other stress responses

Looking beyond the *hli* response, we find additional indications of cellular stress in response to carbon limitation. One such indication is the differential expression of genes whose products curate the proteome. Environmental stresses can damage proteins, forcing the cell to repair, refold, or degrade portions of the proteome (68); degradation can also be a complement to transcriptional changes, speeding the clearance of pre-existing proteins that have no place in the stress response. Thus, it is not surprising that we find upregulation of components of the ClpP protease (PMM0742 and PMM1313, both significantly upregulated under $-CO_2$ shock and borderline upregulated under $-CO_2/-O_2$ shock by t = 6 h); nor that $-CO_2$ shock produces upregulation of a trypsin-family serine protease (PMM1490) by t = 6 h (Fig. 9c, d). What is unexpected is the downregulation of Lon protease (PMM1506), an important part of the general cellular stress response (69), under both shocks. Two genes contributing to the chaperone GroEL (PMM1436, *groEL*; PMM0452, *groEL2*), another common part of the cellular stress response, are also downregulated under both shocks, as is the chaperone gene *dnaK2* (PMM1704). (Surprisingly we find significant upregulation under $-CO_2$ of PMM0698, a candidate gene for DnaJ, the DnaK binding partner.) The suppression of chaperone gene transcription (Fig. 9a, b) is consistent with the expectation that carbon limitation (unlike exposure to heat or heavy metals) should not lead to protein misfolding. Further, as described below, carbon limitation produces a widespread downregulation of genes whose products are involved in translation; with fewer newly translated proteins whose folding needs supervision, a reduced complement of chaperones should be adequate to the task.

One difference between the expression of protein-maintenance genes under $-CO_2$ and $-CO_2/-O_2$ shock bears mentioning. In many species, the enzyme thioredoxin peroxidase is induced by stress to repair the oxidative damage to proteins that is a common casualty of stress conditions (69). Expression levels of this enzyme (PMM0856, *tpx*) have been unaffected in all previous environmental-stress microarray experiments in *Prochlorococcus* MED4 (59-61, 70), suggesting that it should not be considered a general part of the *Prochlorococcus* stress response. Thus, it is not unexpected that *tpx* expression is unchanged by $-CO_2$ shock. We find the first instance of *tpx* differential expression under $-CO_2/-O_2$ shock (Fig. 9c), and the change is a mild but significant downregulation. The suppression of this redox-dependent enzyme may hint at a more reducing cellular environment in the absence of exogenous O_2.

The differential expression of stress-response genes is by no means confined to those involved in proteome maintenance. A clear sign of cellular distress is the strong upregulation of RecA (PMM1562) at $t = 6$ h under $-CO_2$ shock (borderline significant under $-CO_2/-O_2$ shock; Fig. 9e, f). When activated by binding to single-stranded DNA, RecA canonically catalyzes the cleavage of the LexA repressor and thereby initiates the SOS response to DNA damage (71). *Prochlorococcus* does not in general seem to rely heavily on the SOS response (72), and it is somewhat surprising that it should be provoked by carbon starvation, a stress unlikely to produce the DNA damage (e.g., double-strand breaks) that the downstream elements of the SOS pathway like UmuCD attempt to cope with. However, the SOS response has been observed in response to starvation in other bacteria, and this response appears to be essential for survival under carbon limitation for at least one cyanobacterium, *Synechocystis* spp. PCC6803 (73, 74).

Towards a Rosetta stone for *Prochlorococcus* regulatory responses

Overlapping regulatory regimes seem to be almost a necessary consequence of genome streamlining in *Prochlorococcus*, which in most sequenced strains appears to have maintained the diverse metabolic activities necessary to an autotroph at the cost of a large number of regulatory genes (8, 9). For instance, MED4 has only five sigma factors. One of these (PMM1289) is significantly induced under $-CO_2$ shock (borderline significant under $-CO_2/-O_2$ shock; Fig. 9f), indicating a broad switch from one transcriptional program to another, but the downstream targets will likely not be specific to carbon limitation, as induction of this sigma factor is also significant under N starvation (59). With so few regulatory genes at its disposal, *Prochlorococcus* seems likely to depend heavily on combinatorics to respond appropriately to the wide array of environmental challenges it faces in the ocean.

Until we are able to manipulate *Prochlorococcus* genetically, our toolbox for understanding its regulatory networks is arguably even more limited than the set of tools from which that network must be built. One of the few strategies currently available to us is scouring the corpus of *Prochlorococcus* array experiments to find differential responses among the few regulatory genes

MED4 has retained. The $-CO_2$ and $-CO_2/-O_2$ shock microarrays offer a few interesting contributions to this effort. We find that the two-component response regulator encoded by PMM1113 is much more strongly downregulated under $-CO_2/-O_2$ shock than under $-CO_2$ shock (Fig. 9e, f); this gene has previously been observed to be significantly induced by white light and borderline significantly by blue light (61). The putatively phosphorus-starvation inducible PhoH-like protein encoded by PMM1284 is also unresponsive to $-CO_2$ shock but repressed by $-CO_2/-O_2$ shock; it has shown no response in previous MED4 stress experiments, including the phosphorus starvation experiment (70).

Most strikingly among these genes, the Crp-family regulatory protein encoded by PMM0806 was borderline significantly downregulated under $-CO_2/-O_2$ shock and very strongly upregulated by $-CO_2$ shock (Fig. 9e, f). Phosphorus starvation and iron limitation also induce PMM0806 expression, while its expression is unchanged by nitrogen starvation, light intensity, and light color. Thus, microarray experiments with MED4 to date have defined a response spectrum for some unknown readout of cell state: some metabolic posture common to phosphorus-, iron-, andcarbon-starved but oxygen-replete cells demands one response; something common to nitrogen starvation and light shock makes cells indifferent to the issue; and something about anoxic carbon starvation makes the positive response worse than useless. Identification of the targets of PMM0806 would be an extremely useful next step in solving this riddle.

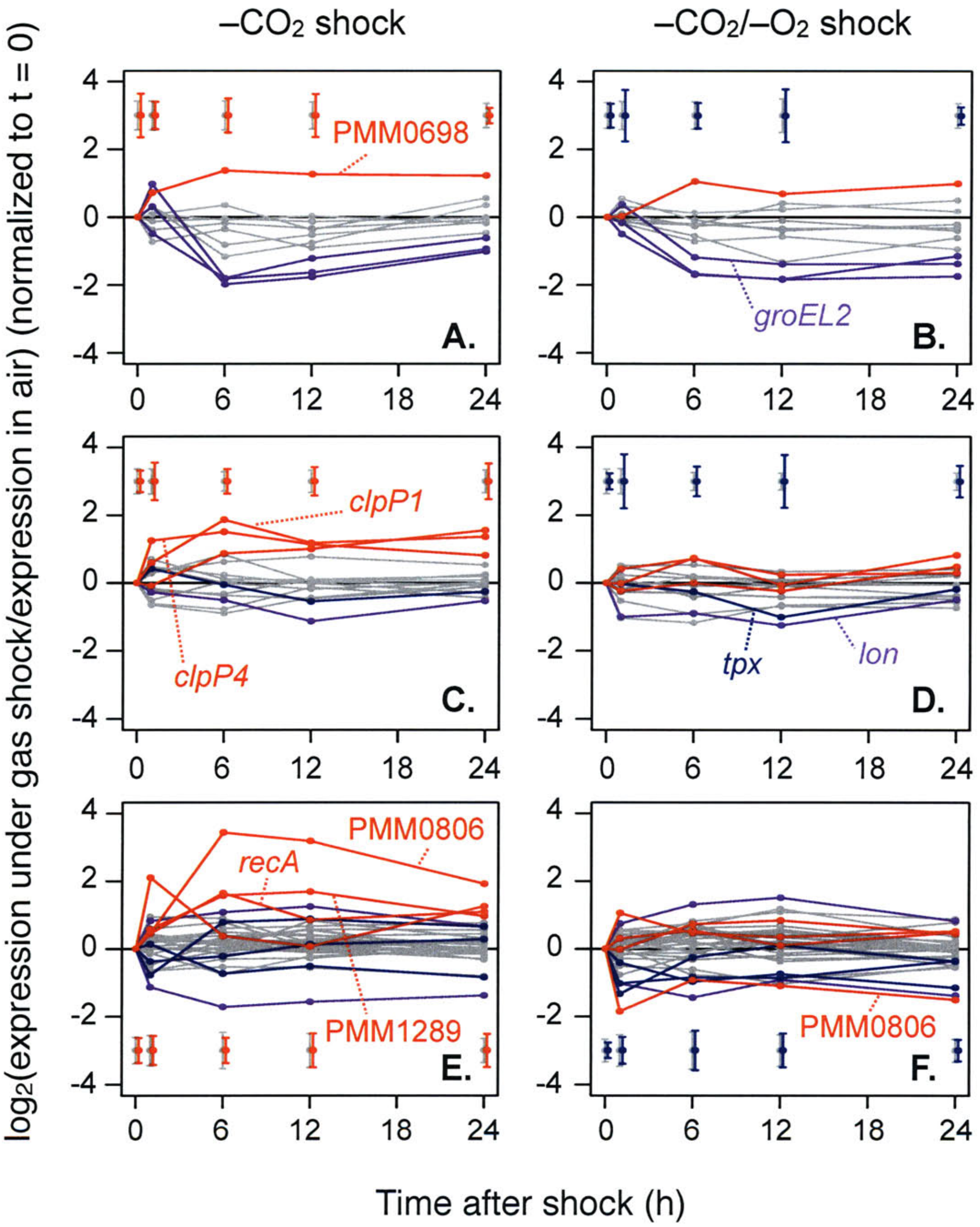


Figure 9. Transcriptional responses of genes annotated as **(A, B)** chaperones, **(C, D)** proteases, and **(E, F)** regulatory and stress-response genes. Color coding and error bars are as described in Figure 4.

Downregulation of transcription and translation under low-CO_2 stress

Under both –CO_2/–O_2 shock and –CO_2 shock, MED4 strongly clamps down on the production of new proteins, with downregulation of the gene expression machinery at several levels. The first indication of this came in purification of the RNA samples used for hybridization; total RNA yields from cultures subjected to both carbon-shock conditions were significantly lower for the t = 24 h timepoint than yields from equal volumes of cells harvested from the –O_2 shock condition and the control ($p < 1 \times 10^{-6}$). These lower yields are consistent with the sustained downregulation of RNA polymerase under low CO_2 (Fig. 8g, h): PMM1483 and PMM1535 (*rpoC2* and *rpoA*, respectively) are significantly downregulated under both –CO_2/–O_2 and –CO_2 shocks by t = 1 h; PMM1484 (*rpoC1*) is significantly downregulated under –CO_2/–O_2 shock (and almost significantly under –CO_2 shock) on the same timescale; and downregulation of the remaining RNA polymerase subunit, encoded by PMM 1485 (*rpoB*) is borderline significant under both –CO_2 and –CO_2/–O_2 shock.

Suppression of the machinery of gene expression extends under both these shocks to the elongation factors Tu (PMM1508, *tufA*) and G (PMM1509, *fusA*), and, under –CO_2/–O_2 shock, to the termination factor RF-1 (PMM1529, *prfA*), as well as to the ribosomes themselves. Of the 54 genes encoding MED4 ribosomal proteins, 23 are differentially expressed in at least one of the carbon-shock conditions; the large majority of these are downregulated (Fig. 8g, h). The suppression of translation-related genes under –CO_2/–O_2 shock is less universal but more rapid than that seen for MED4 under N starvation (59). Interestingly, the response of MED4 to –CO_2 shock proceeds at a more measured pace; just over half the genes that respond to –CO_2 shock show no significant effect until t = 12 h, whereas the $-CO_2/-O_2$ response is almost complete by t = 6 h. Despite the difference in timing, the overlap between the ribosomal genes affected by these two shocks is extensive, with 13 genes significantly differentially expressed under both conditions.

Breaking from the trend, PMM0753 (*rpsB*, encoding 30S ribosomal subunit protein S2) is upregulated ≥2-fold by t = 1 h under $-CO_2/-O_2$ shock and remains elevated through t = 12 h (Fig. 8g). Interestingly, *rpsB* transcription in *E. coli* and other γ-proteobacteria appears to be controlled by a particularly strong sigma-70 promoter, with finer control of S2 protein levels deriving primarily from translational repression mediated by S2 itself (75); if this regulatory scheme is conserved in *Prochlorococcus*, an increase in *rpsB* mRNA is far from predicting an increase in S2 protein. However, unless *rpsB* translation is coupled to mRNA degradation such that translational suppression increases transcript half-life, the continued elevation of *rpsB* transcripts over 12 h against a background of generally downregulated transcription does seem to suggest a role for S2 in the *Prochlorococcus* response to $-CO_2/-O_2$ shock. In this connection, it may be noteworthy that the PMM0312 transcript (*rpsA1*, encoding small-subunit ribosomal protein S1, homolog A), downregulated under $-CO_2$ shock, is elevated at borderline significant levels under $-CO_2/-O_2$ shock. In *E. coli*, S2 and S1 can be stably associated even outside the ribosome, with a role in transcriptional activation (76). Although this particular activity seems unlikely under $-CO_2/-O_2$ shock, given the broad transcriptional and translational downregulation observed (and MED4's lack of a homolog of Hfq, a key part of the *E. coli* S1-S2 complex), the potential for S2 and S1 to interact and mediate non-translational processes could underlie the divergence of *rpsB* and *rpsA1* transcript levels from the ribosomal norm.

Consistent with the general downregulation of the genes encoding the machinery of expression, we find moderate transcriptional downregulation of the enzymes that synthesize the raw materials, the nucleotides and amino acids. One enzyme that catalyzes the phosphorylation of nucleotides, phosphoribosyl phosphotransferase (PMM1122, *apt*), is downregulated, potentially slowing the re-charging of the nucleotide pool to the triphosphorylated, polymerization-competent state. Two genes encoding enzymes in the biosynthetic pathways of threonine (PMM0595, *thrB*) and methionine (PMM0642, *met17*) are significantly downregulated under both –CO_2 and –CO_2/–O_2 shock; three more, in the biosynthetic pathways of cysteine (PMM0227, *cysD*), lysine (PMM0832, *dapB*), and methionine (PMM0643, *metA*), are significantly downregulated under –CO_2 shock and borderline significantly under –CO_2/–O_2 shock. In addition, we observe downregulation of two amino acyl-tRNA synthetase genes (PMM0597, threonyl-tRNA synthetase, and PMM0473, glutamyl-tRNA synthetase).

Targeted regulation of cofactor biosynthesis

Because coenzymes and cofactors frequently participate in a much larger set of reactions than does a single enzyme, modulating the biosynthesis of cofactors may be a parsimonious way to regulate larger subsets of metabolism. In our experiments, MED4 offers several examples of the coordinated regulation of complex pathways of cofactor biosynthesis. In one such pathway, methionine is converted to S-adenosylmethionine, connected by two reversible reactions to Sadenosylhomocysteine and thence to homocysteine, whose methylation regenerates methionine (Fig. 10a). This loop is simple in itself, but each compound except homocysteine is a potential branchpoint. Under $-CO_2$ shock, MED4 significantly downregulates transcription of the gene encoding S-adenosylmethionine synthetase (PMM0311, *metK*) and concomitantly upregulates the gene encoding adenosylhomocysteinase (PMM1625, *ahcY/sahH*). These two responses develop with almost perfect symmetry over the 24 h sampled (Fig. 10b, c). (Under $-CO_2/-O_2$ shock, the *metK* response is borderline significant, but *ahcY* shows no sign of differential expression at all.) If flux through this pathway is under hierarchical control, the net effect should be to diminish the pool of S-adenosylmethionine, returning the carbon skeleton to the amino acid pool. The same strategy—downregulating PMM0311, upregulating PMM1625—appears to be in effect when MED4 is shocked with high light (61). High-light shock should parallel low-CO_2 shock insofar as each shock is manifested in carbon fixation's inability to keep pace with the reduction of the photosynthetic electron transport chain. Where $-CO_2$ shock and high-light shock together diverge from $-CO_2/-O_2$ shock, oxygen must somehow be involved, whether in terms of the redox state of the cell, the availability of molecular oxygen as a reactant, or some other readout.

In a second instance of cofactor targeting, both $-CO_2$ shock and $-CO_2/-O_2$ shock yield parallel increases in two cofactors that function in tandem. The NifS-like aminotransferase encoded by PMM0170 catalyzes the transfer of a sulfur from cysteine to an early intermediate in the synthesis of thiamine pyrophosphate (TPP), central to carbon-carbon bond chemistry in glycolysis, the citric acid cycle, and amino acid metabolism. A common TPP decarboxylation mechanism passes the substrate from cofactor to cofactor, using a series of covalent bonds—between substrate and TPP, then substrate and lipoic acid, and finally substrate and coenzyme A—to retain much of the bond energy that would otherwise be lost when the substrate carbon-carbon bond is broken. In such pathways, the regeneration of free TPP depends on the presence of lipoate, the next cofactor in line. Thus, the observation (Fig. 10d, e) that upregulation of PMM0170 under carbon limitation is accompanied by significant upregulation of lipoic acid synthetase (encoded by PMM1514, *lipA*) suggests that cells are attempting to facilitate flux through thiamine pyrophosphate/lipoic acid dependent pathways. It should be noted that the reaction catalyzed by lipoic acid synthase consumes S-adenosylmethionine, which, as noted above, should grow relatively scarce under $-CO_2$ shock; upregulating *lipA* transcription under this condition may thus be necessary to maintain normal levels of lipoate, rather than reflecting a need for increased lipoate accumulation. The broader caveat, of course, is that upregulation of PMM0170 and PMM1514 transcription is by no means a sufficient condition for upregulation of TPP and lipoate production, although the coordinated upregulation of both transcripts is suggestive. TPP and lipoate synthesis have not been targets of transcriptional regulation in MED4 under other environmental stresses (59-61, 70).

Finally, $-CO_2$ shock and $-CO_2/-O_2$ shock prompt MED4 to cut off access to the synthesis of cobalamin (vitamin B12), required for many amino acid isomerizations and, in MED4, for the reduction of ribonucleotides to deoxyribonucleotides. In the C5 pathway *Prochlorococcus* uses for cobalamin biosynthesis, glutamate is first charged with its cognate tRNA by glutamyl-tRNA synthetase, thereby becoming activated for conversion to glutamate-1-semialdehyde by

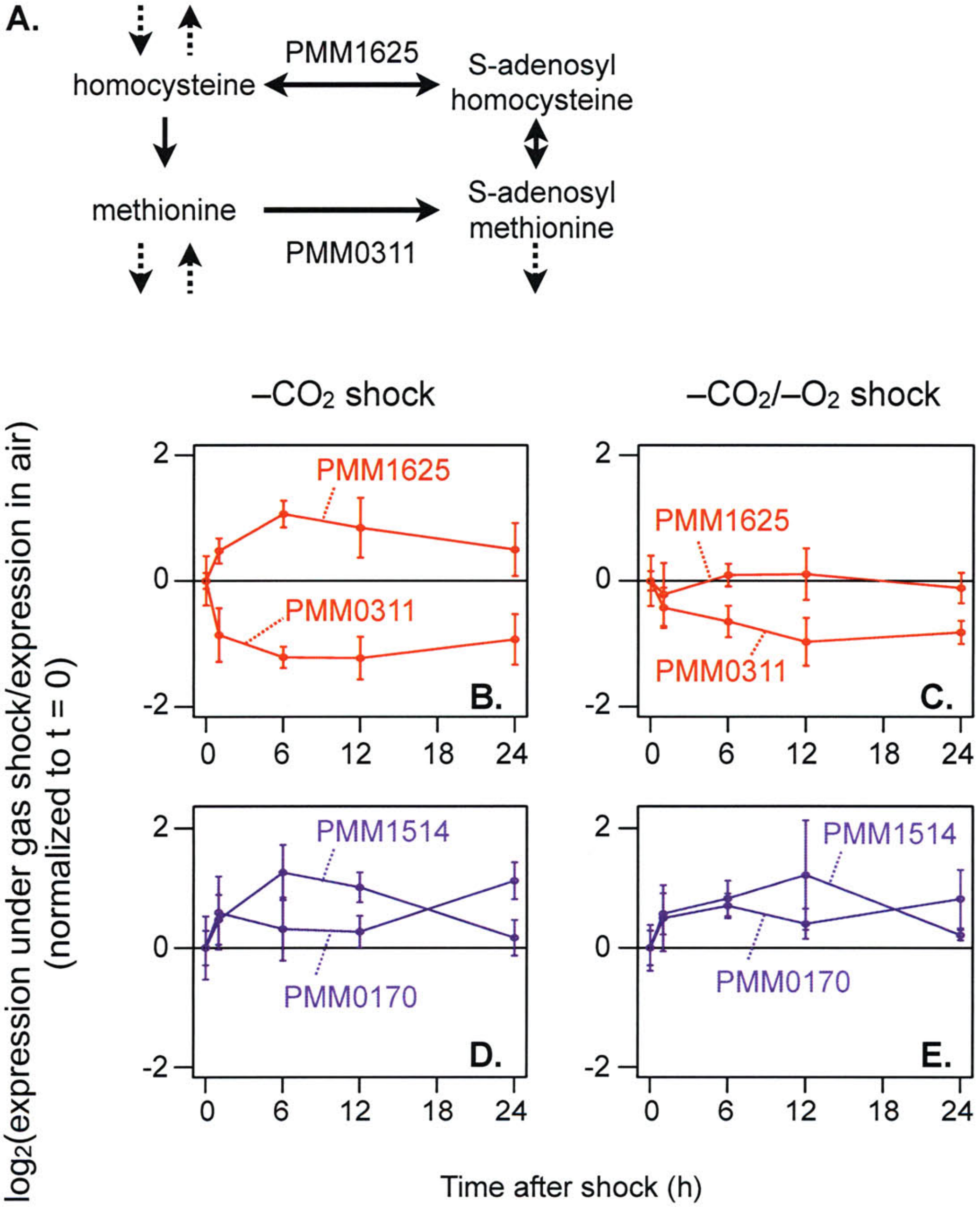


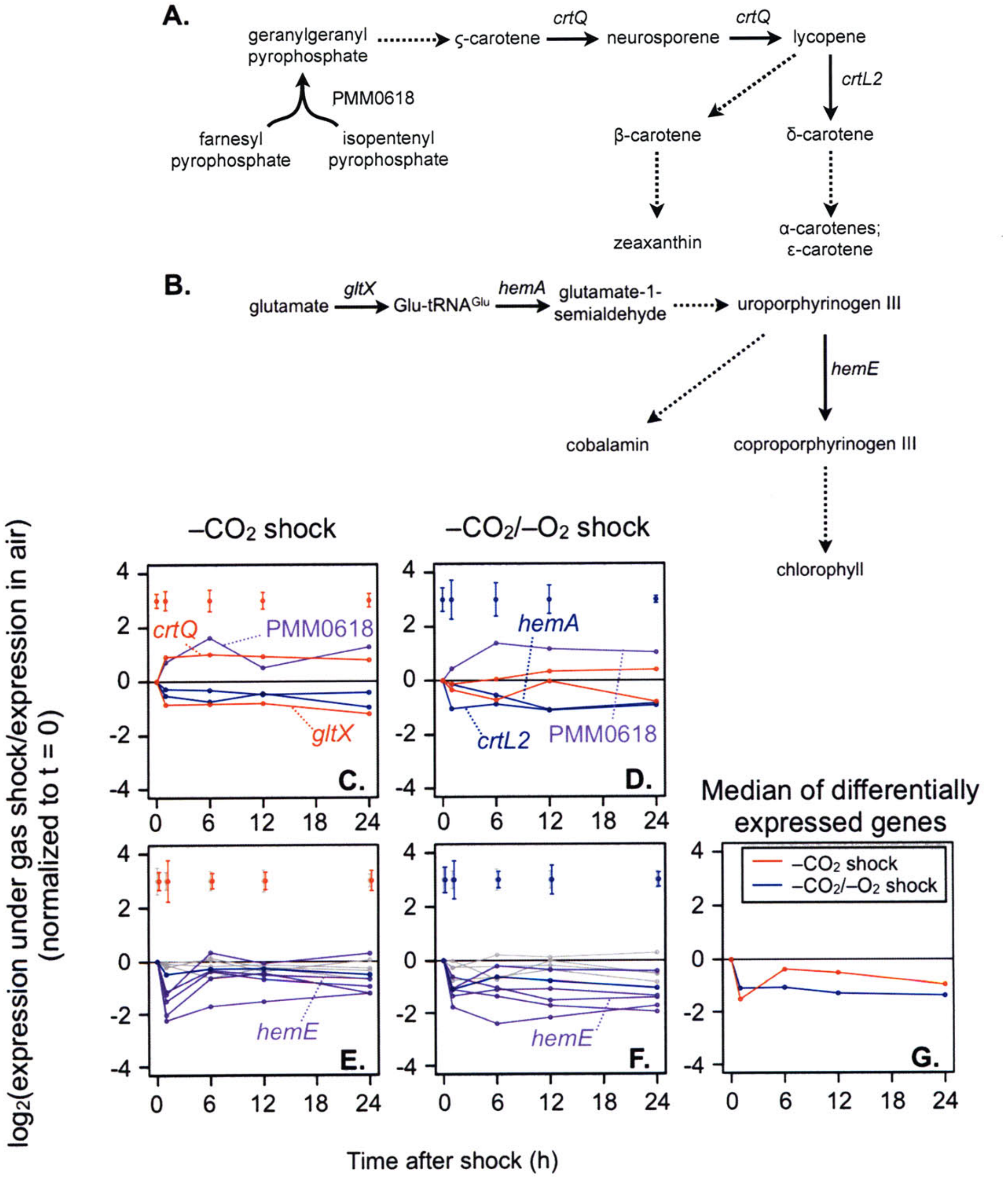
Figure 10. Regulation of biosynthetic pathways of **(A–C)** S-adenosyl methionine and **(D, E)** thiamine pyrophosphate and lipoate. Gene expression profiles are colored as in Figure 4; error bars represent one standard deviation (of the mean of two biological replicates). Dashed arrows in (A) indicate possible inputs to and from other pathways.

glutamyl-tRNA reductase (Fig. 11b). The genes encoding the enzymes that catalyze these steps —respectively, PMM0473 (*gltX*) and PMM0768 (*hemA*)—are downregulated under carbon limitation (Fig. 11c, d). As glutamyl-tRNA synthetase is one of only two amino acyl-tRNA synthetases to show significant differential expression under these conditions, it is likely that its differential expression is not merely an aspect of the general downregulation of translation but rather a means of specifically modulating a glutamyl-tRNA-dependent pathway. Furthermore, we find downregulation of the cobalamin-dependent enzyme ribonucleotide reductase (PMM0661, *nrdJ*). If the existing pools of GltX, HemA, and NrdJ are turned over on the timescale of hours, these transcriptional changes suggest that carbon-limited MED4 keeps its pool of nucleotides poised more for transcription than for genome replication.

Modulation of pigment biosynthesis

Intimately connected to the biosynthesis of cobalamin, the biosynthesis of pigments may also undergo a major shift under $-CO_2$ and $-CO_2/-O_2$ shocks. Chlorophyll biosynthesis branches from cobalamin biosynthesis at uroporphyrinogen III; in the chlorophyll arm, the next step is catalyzed by uroporphyrinogen decarboxylase (URO-D, encoded by PMM0583, *hemE*; Fig. 11b). Transcription of this gene is downregulated under both $-CO_2$ and $-CO_2/-O_2$ shock, but the timecourse expression profiles are distinctly different under the two conditions. $-CO_2/-O_2$ shock produces a slow decline, deepening steadily until t = 12 h and remaining low at t = 24 h; by contrast, $-CO_2$ shock evokes a stronger initial downregulation, and then a nearly full recovery of expression by t = 6 h (Fig. 11e–g). The same distinction applies to the patterns of downregulation of five other chlorophyll synthesis genes (PMM0543–0545, PMM0760, and

Figure 11. Regulation of **(A, C, D)** carotenoid and **(B, E–G)** cobalamin and chlorophyll biosynthesis. Gene expression profile coloring and error bars are as in Figure 4. Dashed arrows in (A, B) indicate intervening metabolic steps. In addition to *hemE*, the chlorophyll biosynthesis genes *chlLBNH* (purple curves, E and F) are differentially expressed under both –CO_2 and –CO_2/–O_2 shock, and *chlP* (blue curve, E and F) is differentially expressed under –CO_2/–O_2 shock.


A.
geranylgeranyl pyrophosphate
ς-carotene
crtQ
neurosporene
crtQ
lycopene
PMM0618
farnesyl pyrophosphate
isopentenyl pyrophosphate
crtL2
β-carotene
δ-carotene
zeaxanthin
α-carotenes; ε-carotene
B.
glutamate
gltX
Glu-tRNA^Glu
hemA
glutamate-1-semialdehyde
uroporphyrinogen III
hemE
cobalamin
coproporphyrinogen III
chlorophyll
–CO2 shock
–CO2/–O2 shock
log2(expression under gas shock/expression in air) (normalized to t = 0)
crtQ
PMM0618
gltX
C.
hemA
crtL2
PMM0618
D.
hemE
E.
hemE
F.
Median of differentially expressed genes
–CO2 shock
–CO2/–O2 shock
G.
4
2
0
-2
-4
0
6
12
18
24
Time after shock (h)

PMM0831; respectively, *chlLBNPH*) and bears a strong similarity to the pattern seen for photosystem I and II genes: sharp downregulation under both –CO_2 and –CO_2/–O_2 shocks, but much more extensive recovery by t = 6 h in simple –CO_2 shock than in –CO_2/–O_2 shock (Fig. 4).

It is difficult to say how chlorophyll levels are likely to change in these two scenarios; even if we assume that enzyme levels rapidly mirror RNA levels, such that the enzymatic capacity for chlorophyll biosynthesis is suppressed under –CO_2/–O_2 shock but suppressed and restored under –CO_2 shock, there remains the question of substrate. In a time of Ci starvation, the raw materials for biosynthetic pathways are likely to be limiting, so chlorophyll production may not rebound under –CO_2 shock despite the recovery of synthetic capacity. Further, if chlorophyll turnover is slow, then much of the cell's normal chlorophyll biosynthesis budget could be used for keeping pace with dilution as the cell grows; in that case, when Ci starvation precludes cell growth, chlorophyll per cell might stay fairly constant despite the vicissitudes in pathway transcript levels. On the other hand, rapid turnover of chlorophyll molecules could, by clearing damaged and potentially damaging pigments, contribute substantially to cellular health; in that case, we might imagine that other pathways will be starved of carbon in order to keep the chlorophyll pathway well stocked with substrate. In this latter hypothetical, chlorophyll levels would mirror URO-D transcript levels, recovering as soon as biosynthetic capacity expands. Whatever the change at the level of flux, –CO_2 cells appear from microarray analysis to deviate from and then rapidly return to the basal level of chlorophyll biosynthetic capacity, while –CO_2/–O_2 cells deviate and never recover.

Not all pigment-related genes follow this pattern, however. Both –CO_2 and –CO_2/–O_2 shock evoke significant upregulation of PMM0618, encoding polyprenyl synthetase. This enzyme is the gateway to the biosynthesis of carotenoids, catalyzing formation of the common precursor all-*trans* geranylgeranyl pyrophosphate (Fig. 11a). Carotenoids are a crucial element of the cyanobacterial defense against photodamage (77), with different mixtures of carotenoids conferring varying levels of protection (78). *Prochlorococcus* has an unusual complement of carotenoids, relying heavily on both α-carotene and zeaxanthin (7). The synthesis of these compounds diverges at lycopene; a β-cyclase activity (PMM1064) is required to produce α-carotene, while zeaxanthin requires an ε-cyclase. Possession of both activities is uncommon in cyanobacteria (79). Remarkably, *Prochlorococcus* not only has both activities but combines both into one enzyme, encoded by PMM0633 (80).

We find strong indications that MED4 attempts to modulate its carotenoid content under carbon limitation, and that the form of this modulation is dependent on the availability of oxygen. First, expression of the bifunctional cyclase gene PMM0633 is significantly downregulated under –CO_2/–O_2 shock and not under –CO_2 shock (Fig. 11c, d). That is, under –CO_2/–O_2 shock, MED4 closes access to the ε-cyclase activity, shunting the lycopene precursor exclusively towards production of the β-cyclized carotenoid zeaxanthin. By contrast, –CO_2 shock causes MED4 to upregulate PMM0115 (*crtQ*), encoding a ς-carotene desaturase activity upstream of the lycopene branchpoint, a regulatory strategy that, given an adequate supply of substrate, should permit the total rate of carotenoid biosynthesis to increase while leaving the balance of β- and ε-cyclic carotenoids unchanged.

This is a particularly interesting result in two respects. First, Goericke *et al.* observed that the carotenoid complement of a field population of *Prochlorococcus* growing in an oxygen minimum zone differed substantially from that of lab-grown cultures (1). The set of carotenoids observed in OMZ *Prochlorococcus*—plenty of α-carotene, but far less zeaxanthin than is typical in the lab, and far more of the zeaxanthin derivative parasiloxanthin—differs from the set that gene expression suggests $-CO_2/-O_2$ MED4 will be able to synthesize, whether because of the different *Prochlorococcus* strains involved or the very different levels of light. Nonetheless, our result is clearly concordant with that of Goericke *et al.* insofar as it argues that carotenoid composition in *Prochlorococcus* can be affected by oxygen partial pressure. Oxygen levels are known to play a role in the carotenoid composition of some photosynthetic bacteria (81, 82), and whether and how such modulations contribute to the successful adaptation of *Prochlorococcus* to growth in natural oxygen minima is a question of considerable interest.

The second reason this result stands out is the demonstration by Götz *et al.* that zeaxanthin has a stronger photoprotective effect than do other carotenoids (78). That the transcriptional response of cells under $-CO_2/-O_2$ shock appears to favor zeaxanthin biosynthesis, while $-CO_2$ cells simply increase their catalytic capacity for carotenoid biosynthesis across the board, is fully consistent with the argument that carbon-limited *Prochlorococcus* cells with access to oxygen face a smaller threat of photodamage than those without. Lacking the oxygen-dependent safety valves that I propose below, cells under $-CO_2/-O_2$ stress are left to cope with the constant barrage of photons by any means they can.

Opening safety valves under carbon limitation: PTOX and photorespiration

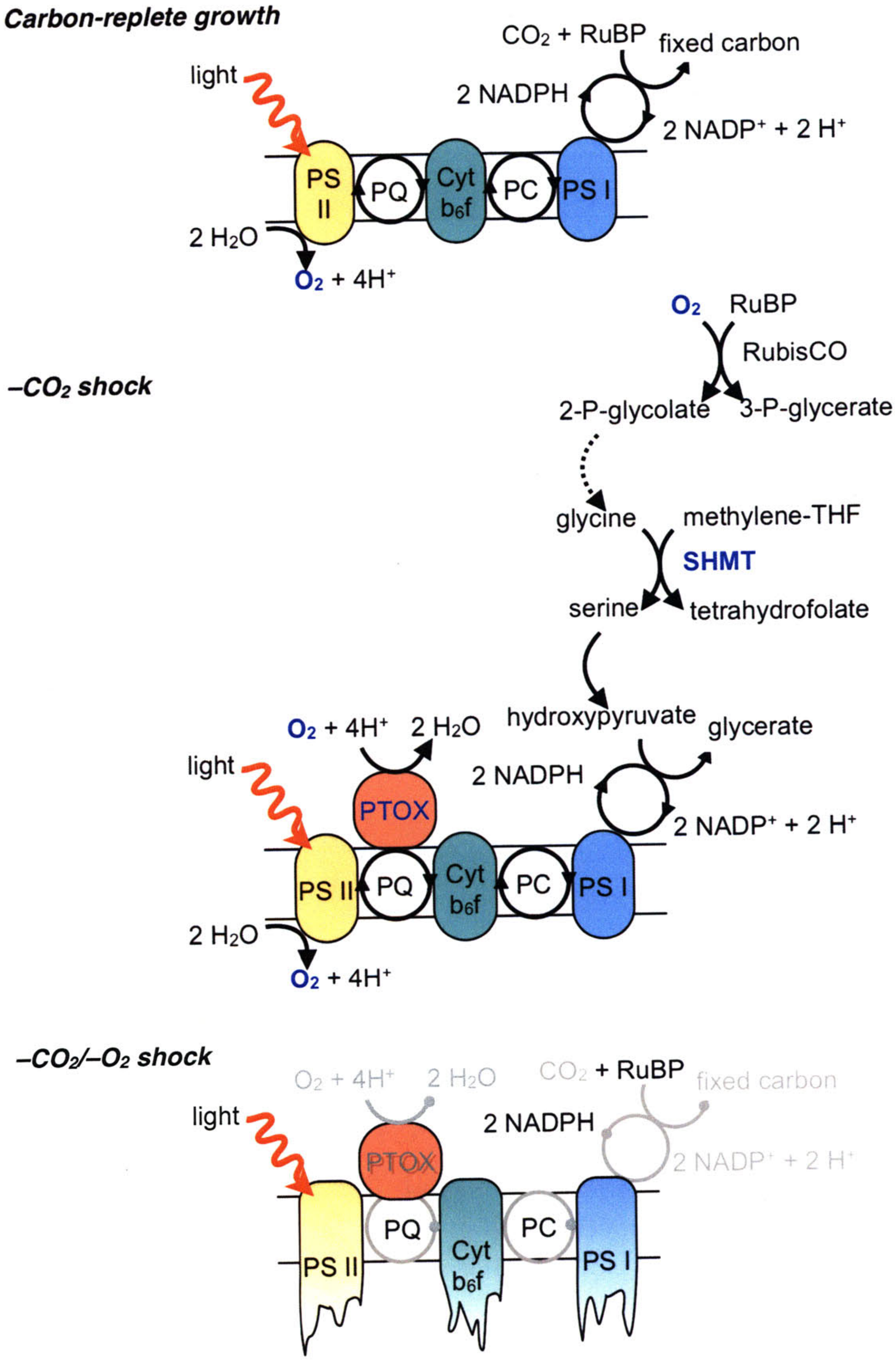
When the flow of energy from photons to reducing equivalents to carbon-carbon bonds is blocked by the absence of fixable carbon, the photosynthetic electron transport chain may stop but the flood of incoming photons does not. To survive, *Prochlorococcus* must find another outlet for this energy. MED4 appears to have two safety valves at its disposal, both of them dependent on oxygen (Fig. 12). First, it can decouple the photosynthetic electron transport chain from Rubisco by funneling excited electrons from photosystem II to an alternative oxidase rather than to cytochrome b_6f and photosystem I. Such a system, likely using the plastoquinol terminal oxidase (PTOX), was recently demonstrated in a laboratory population of *Synechococcus* WH8102 and observed in an open-ocean cyanobacterial population dominated by *Prochlorococcus* (83, 84). The populations studied should not be undergoing carbon limitation, but they produce lower levels of the iron-hungry complexes cytochrome b_6f and photosystem I, relative to photosystem II, than do coastal populations, suggesting a degree of iron limitation in the open-ocean populations (84). Just as insufficient inorganic carbon produces a bottleneck in energy flow at Rubisco, insufficient iron produces a bottleneck at photosystem I; so iron-limited cells too need a sink for reducing power upstream of photosystem I.

That reducing equivalents can flow to processes other than carbon fixation is suggested by the relationships Bailey *et al.* find between irradiance, carbon fixation, and photochemical efficiency in *Synechococcus* WH8102: substantial electron flow through photosystem II continues at irradiances more than an order of magnitude above the point where the rate of carbon fixation saturates, and far above the light level where electron flow through photosystem I is severely curtailed by photoinhibition. The maintenance of photosystem II's photochemical efficiency at high irradiance is dependent on oxygen, and even at moderate light levels the oxidase inhibitor propyl gallate causes the core of photosystem I to accumulate in the reduced state (83). PTOX is thought to be the unifying cause of these phenomena: on one side of the thylakoid membrane, photosystem II splits water to yield oxygen and protons; on the other, PTOX uses the excited electrons handed off from photosystem II to reduce oxygen to water once more. When oxygen is available, this alternative pathway allows photosystem II to remain active at high irradiance without risking over-reduction; without oxygen or without a functional oxidase, photosystem I is rapidly overwhelmed by electron flow.

We find that expression of the gene encoding PTOX (PMM0336) is dramatically induced by t = 1 h, with a more than 11-fold increase in expression under both $-CO_2$ and $-CO_2/-O_2$ shock. Expression remains strongly upregulated under both conditions throughout the timecourse (Fig. 13a, b). (The possible NADH-plastoquinone oxidoreductase subunit encoded nearby (PMM0335, on the opposite strand) also shows significant, albeit much milder, upregulation under $-CO_2/-O_2$ shock, and borderline significant upregulation under $-CO_2$ shock.) Previous array experiments support the idea that PTOX can serve in *Prochlorococcus* as an alternative electron acceptor downstream of photosystem II. Photosystem I is a voracious user of iron; iron limitation produces both downregulation of some photosystem I components and significant upregulation of PMM0336 (60). Two results from the study of MED4 response to light shocks are especially salient. First, following dark acclimation, treatment with high light, blue light, white light, and red light—each of which should produce an unwonted degree of reduction in photosystem II and the plastoquinone pool—all produced significant upregulation of PMM0336. By contrast, treatment with the photosystem II inhibitor DCMU gave significant PMM0336 repression (61). Notably, the specific inhibition of photosystem II by DCMU should prevent the reduction of the plastoquinone pool, so that there is neither need nor substrate for PTOX.

That the PTOX approach to photoprotection should be successful under $-CO_2$ shock, when the cells are still bathed in normal levels of oxygen, is highly likely. The obvious question, of course, is whether the expression of PTOX does any good at all for cells in the $-CO_2/-O_2$ condition. To judge from the results of Mackey *et al.* (84), the likelihood is low; in that work, *Prochloroccus*-dominated surface-water samples from the Atlantic under high light were able to maintain a large fraction of their photosystem II reaction centers in the oxidized state under oxic conditions but were unable to when oxygen was removed. The authors argue that the difference is the lack of the PTOX reaction in anoxia. As long as a cell continues to split water at photosystem II, it will see at least at trickle of molecular oxygen; the Mackey result suggests that, under anoxia, most of this newly evolved O_2 is unavailable to PTOX, lost to diffusion down the steep oxygen concentration gradient. The sustained upregulation of PTOX in *Prochlorococcus* MED4 under $-CO_2/-O_2$ shock does not shed much light on the subject; one can imagine either that PTOX expression is sustained because it is having the desired effect, or that it is sustained because the cell, continuing to need an outlet for excited electrons, deduces that yet more PTOX expression is required. Such questions are a reminder of the limitations of purely transcription-based measurements.

Figure 12. Proposed "safety valves" for carbon-limited *Prochlorococcus.* In carbon-replete cells (top), light energy drives the photosynthetic electron transport chain, transferring reducing power from photosystem II (PS II) via plastoquinone (PQ) to cytochrome b_6f (Cyt b_6f); from there via plastocyanin (PC) to photosystem I (PS I); and from there via the NADP(H) pool to CO_2, fixing a carbon atom. In the absence of CO_2, this final sink for reducing equivalents is unavailable, causing the electron transport chain to become over-reduced. However, so long as oxygen is still present (middle, $-CO_2$ shock), MED4 has two alternatives: electrons can flow from PQ to O_2 via the plastoquinone terminal oxidase (PTOX) (83, 84), or NADPH can be used to drive the oxygenation of ribulose-1,5-bisphosphate (RuBP) by Rubisco. In the latter case, the cytotoxic by-product 2-phosphoglycolate is cleared from the system through the action of (among other enzymes) serine hydroxymethyltransferase (SHMT). When cells are deprived of O_2 as well (bottom, $-CO_2/-O_2$ shock), these safety valves remain closed, and the continuing stream of photons causes damage to the photosynthetic chain.


Carbon-replete growth
light
CO_2 + RuBP
fixed carbon
2 NADPH
2 $NADP^+$ + 2 H^+
PS II
PQ
Cyt b_6f
PC
PS I
2 H_2O
O_2 + $4H^+$
–CO_2 shock
O_2
RuBP
RubisCO
2-P-glycolate
3-P-glycerate
glycine
methylene-THF
SHMT
serine
tetrahydrofolate
hydroxypyruvate
glycerate
O_2 + $4H^+$
2 H_2O
light
PTOX
2 NADPH
2 $NADP^+$ + 2 H^+
PS II
PQ
Cyt b_6f
PC
PS I
2 H_2O
O_2 + $4H^+$
–CO_2/–O_2 shock
O_2 + $4H^+$
2 H_2O
CO_2 + RuBP
fixed carbon
light
2 NADPH
2 $NADP^+$ + 2 H^+
PTOX
PQ
PC
PS II
Cyt b_6f
PS I

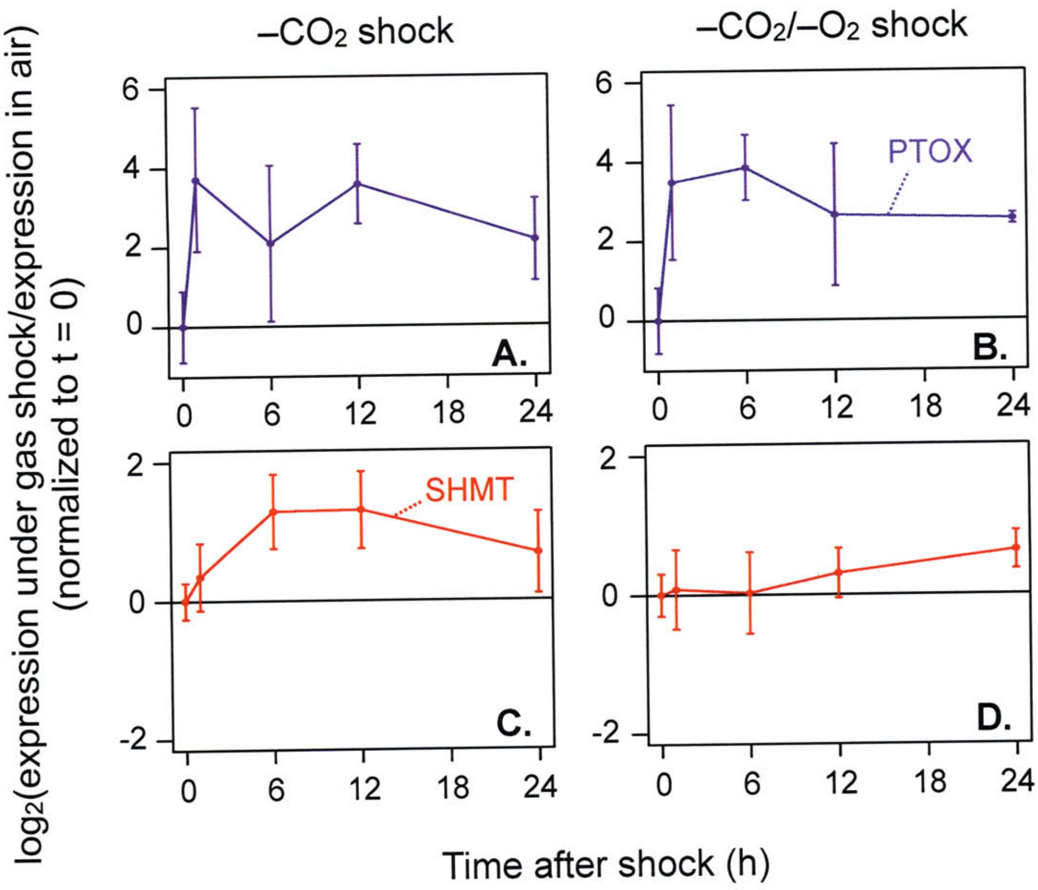


Figure 13. Differential expression of two possible "safety valve" systems. Expression of the **(A, B)** plastoquinol terminal oxidase (PTOX) gene and of **(C, D)** serine hydroxymethyltransferase (SHMT) under **(A, C)** –CO_2/–O_2 and **(B, D)** –CO_2 shocks. Gene expression profiles are colored as in Figure 4; error bars represent one standard deviation (of the mean of two biological replicates).

MED4's second option is to leave photosystem I coupled to photosystem II, but to decouple the re-oxidation of photosystem I's NADP(H) pool from carbon fixation. Here, the means of decoupling is a process that we are accustomed to thinking of as wasteful: photorespiration, the Rubisco oxygenation reaction. By cleaving ribulose-1,5-bisphosphate without fixing a molecule of carbon dioxide, the cell immediately loses a carbon-carbon bond, and in the course of salvaging the waste product 2-phosphoglycolate, a phosphoester bond is lost as well. But when the problem is a surfeit of energy, the solution is waste. This process has long been hypothesized to be a potentially useful energy sink for land plants, for just this reason (85). One key benefit of photorespiration is that it offers a safe way to waste energy: whereas photons wreak molecular havoc, the energy that is released in the controlled bond cleavages of photorespiration and 2-phosphoglycolate salvage is lost as waste heat.

The limitation of photorespiration is that it requires oxygen to reach the Rubisco active site, cloistered in cyanobacterial carboxysomes which function in part to limit the access of oxygen to the enzyme. While the oxygen produced by photosystem II stands a chance of reaching PTOX for re-reduction to water, the likelihood that any will reach Rubisco is poor. Thus, this safety valve should be firmly closed under $-CO_2/-O_2$ conditions, while the cell should be able to open it under $-CO_2$ shock. Consistent with this expectation, we observe significant upregulation of serine hydroxymethyltransferase (SHMT, encoded by PMM0258) under $-CO_2$ shock; at no time is expression under $-CO_2/-O_2$ shock even borderline significant (Fig. 13c, d). SHMT belongs to the so-called C2 cycle, in which 2-phosphoglycolate is broken down via glyoxylate to glycine and then built back up to glycerate via serine (Fig. 12). The reaction catalyzed by SHMT requires methylene-tetrahydrofolate; as noted above, while cells under $-CO_2/-O_2$ shock appear to shunt coenzyme biosynthesis away from folate production, cells under $-CO_2$ shock do not.

Two other aspects of the global transcriptional response under $-CO_2$ and $-CO_2$ /$-O_2$ shocks are suggestive of a role for photorespiration in the former and not the latter. First, as already noted, $-CO_2/-O_2$ shock leads to elevated *rbcR* transcription and suppressed *rbcLS* transcription (albeit borderline significant for *rbcL*); $-CO_2$ shock downregulates the rest of the Calvin cycle at the RNA level and probably, via CP12, the protein level, but it leaves Rubisco expression levels untouched, consistent with a photorespiratory need for Rubisco. Second, the distinct t = 6 h recovery we observe for photosystem I and cytochrome b_6f expression under $-CO_2$ shock (Fig. 4) makes sense only in the context of an outlet for reducing equivalents that is *downstream* of both complexes. Examination of previous MED4 microarray experiments uncovers no other instances of SHMT differential expression, even when there is evidence of photosystem stress and upregulation of PTOX expression. It is possible that MED4's carbon-concentrating mechanism is efficient enough that photorespiration is only available as a photoprotection strategy when carbon dioxide is scarce enough for oxygen to gain the upper hand in the competition for Rubisco.[4]

[4] Speculating wildly, one might imagine a cyanobacterium that, lacking a PTOX-type pathway, makes the photorespiration strategy more widely applicable by maintaining a second copy of Rubisco that is induced by light stress and tagged for cytoplasmic (rather than carboxysomal) localization.

In light of this picture of oxygen as a key player in photoprotection, the growth curves (Fig. 1) demand a second look. As noted upon initial examination of this data, 40 ppm CO_2 is clearly a limiting supply. Equally clearly, CO_2 supplied at $\geq$180 ppm is not limiting. What, then, is the reason for the growth impairment that appears at 180 ppm CO_2 + <0.001% O_2 and at 360 ppm CO_2 + <0.001% O_2? In constant light, MED4 grows optimally at ~100 $\mu E\ m^{-2}\ s^{-1}$; the irradiance used in this study (~65 $\mu E\ m^{-2}\ s^{-1}$) permits growth at ~90% of the maximal rate, and is well shy of the >200 $\mu E\ m^{-2}\ s^{-1}$ irradiance level normally needed to provoke photoinhibition of growth (13). Photoinhibition should arise when light is so intense that cells accumulate reducing equivalents faster than they can be used—a threshold that should be a function of the cell's access to the processes of carbon fixation, O_2 reduction, and photorespiration. I suggest that the oxygen-dependent processes are already in play at the irradiance used in my growth experiments. That is, in the 180 ppm CO_2 + <0.001% O_2 and 360 ppm CO_2 + <0.001% O_2 conditions, with the PTOX and photorespiration safety valves effectively stoppered by anoxia, the limiting factor may be not carbon supply but flux through the carbon-fixation pathway: without oxygen, cultures grown at moderately high light may not be able to fix even abundant carbon fast enough to match the needs of the photosynthetic electron transport chain for re-oxidation. If this is true, it is to be expected that the observed growth deficit at <0.001% O_2 should weaken, and the transcriptional profiles of $-CO_2$ and $-CO_2/-O_2$-shocked cells converge, as light levels are reduced and the need for electron sinks is accordingly diminished. Consistent with this hypothesis, the marine oxygen minima where *Prochlorococcus* thrive are at or below the base of the euphotic zone, where light is a limiting resource rather than a constant danger.

A second aspect of the growth data also requires consideration: if indeed oxygen serves as a photoprotectant, permitting photosynthesis-related genes to return to roughly basal expression levels in $-CO_2$ cells, how is it that the growth under the $-CO_2$ condition (40 ppm CO_2 + 21% O_2) gives no apparent advantage over growth under the $-CO_2/-O_2$ condition (40 ppm CO_2 + <0.001% O_2)? I propose that the real benefit of the oxygen-dependent safety valves will not become clear until cells are released from carbon starvation. Unharnessed light energy leads to the production of reactive oxygen species, leaving the cell exposed to crippling oxidative damage that will accumulate over the course of starvation. In some heterotrophs, the ability of starving cells to recover once nutrient levels are restored depends on the cells' ability to defend against reactive oxygen species (86); although the defense mechanism differs, the effect may be the same in *Prochlorococcus*. In this case, repeating the growth experiments with an additional recovery stage, by switching to air sparging at t ~ 2 d, would reveal a much larger viable population of cells after 40 ppm CO_2 + 21% O_2 growth than after 40 ppm CO_2 + <0.001% O_2.

CONCLUSIONS

The observation that *Prochlorococcus* dominates some suboxic waters to the near exclusion of other phototrophs (1) prompts the question of what adaptations permit it to live in these waters while others fail, and more generally of how levels of dissolved gases affect *Prochlorococcus* physiology. We find that the high-light-adapted *Prochlorococcus* strain MED4 shows no sign of distress when first transferred to a carbon-replete, anoxic environment, but neither does it benefit from the extremely high ratio of CO_2 to O_2, instead gradually falling behind cells given normal levels of both carbon dioxide and oxygen. High CO_2:O_2 ratios are no more useful for growth when the absolute level of CO_2 is limiting. Indeed, global transcriptional profiling suggests that oxygen offers cells some measure of protection at limiting carbon, permitting a cells to recover approximately basal photosystem expression levels after the initial shock of starvation. Whereas carbon-replete cells were completely indifferent at the transcriptional level to the early hours of anoxia, carbon limitation brought the relevance of oxygen into sharp relief, with differences in the stress the cell perceives under $-CO_2$ and $-CO_2/-O_2$ shocks indicated by the diverging expression patterns of regulatory genes; differences in cellular redox poise, by the expression patterns of redox-sensitive proteins; and differences in the need for photoprotection, by the expression patterns of carotenoid biosynthesis genes. I propose that the underlying difference between the state of cells in these two conditions is the availability or absence of functional alternative oxidase and photorespiration pathways, both oxygen-dependent ways to jettison the light energy that cannot now be usefully consumed by fixing carbon. Further, I hypothesize that, by analogy with oxidative stress defenses in heterotrophs, the ability of a constant-light culture to recover from carbon limitation will depend on its access to the "safety valve" processes during starvation. However, the utility of the safety valves need not be confined to starvation recovery; consistent with the first reports of PTOX use in marine cyanobacteria (83, 84), the evidence of growth defects in carbon-replete but anoxic conditions suggest that these oxygen-dependent processes contribute to the health of the cell at moderate irradiances even when carbon fixation is unimpaired.

The data presented here suggest a number of avenues for future inquiry. First, if we extend the growth experiments to include a recovery phase in which cultures are sparged with air, is it in fact the case that $-CO_2$ cells, having had the benefit of both PTOX and photorespiration, are able to recover more successfully than $-CO_2/-O_2$ cells? If so, does the greater success of $-CO_2$ cells appear as a larger viable fraction of the population or as a greater growth rate among a small subset of cells? What can this difference, and the transcriptional differences between these two recoveries, tell us about oxidative stress in *Prochlorococcus*? Second, if the expression timecourse is allowed to run past 24 h, at what point do cells under the $-O_2$ condition begin to diverge transcriptionally from cells bubbled with air, and what differences do we see? The growth curve from the $-O_2$ condition (360 ppm CO_2 + <0.001% O_2) suggests that these cells are stressed by t = 48 h. If indeed the PTOX and photorespiratory processes play a maintenance role in the photosynthetic electron transport chain, expression of these genes—and perhaps even of the carotenoid biosynthesis genes—in the $-O_2$ condition should eventually come to resemble the pattern seen more rapidly in $-CO_2/-O_2$ cells. Proteomic and metabolomic experiments would also be of significant interest in probing the $-O_2$ condition further; is the transcriptional similarity of the $-O_2$ and air conditions, so striking in the first 24 h, belied by subtler realignments at the protein and small-molecule level? Indeed, proteomic and metabolomic probes, especially probes targeting the redox state of the stressed cells, would greatly improve our understanding of the response of *Prochlorococcus* to these gas conditions; as discussed above, transcriptional profiling yields datasets whose richness is counterbalanced by their distance from metabolic fluxes and protein pools. In particular, a measurement of the NAD(P)/NAD(P)H balance in our experimental conditions over time would allow us to see how access to exogenous oxygen affects the redox state of *Prochlorococcus*.

This work began from the suspicion that the role of oxygen in *Prochlorococcus* could not be investigated properly without considering the role of carbon dioxide alongside it. It was an easy step from Rubisco's readiness to perform both carboxylation and oxygenation reactions to the notion that the stress of carbon limitation would be heightened, and the sensing of carbon limitation made more acute, by the presence of abundant oxygen able to compete with carbon dioxide for the enzyme's active site. But in making this step, we think like heterotrophs. We have to eat fast enough to keep up with our energy demand; phototrophs must eat fast enough to keep up with the supply. With that realization in mind, it will be essential in carrying this work forward to consider light as a third variable alongside carbon dioxide and oxygen levels. Photophysiology experiments should be used to ask whether the dependence of fluorescence dynamics on light, oxygen, and carbon dioxide availability in MED4 is consistent with the model presented here: is there evidence for an oxygen-dependent photochemical quenching process? Does the irradiance at which this process comes into play depend on the availability of inorganic carbon? It will be essential to remember the central role light has had in *Prochlorococcus* evolution: the number and nature of photoprotective strategies are likely to differ widely between high-light-adapted strains, like MED4, and low-light-adapted strains. A BLAST search for *Prochlorococcus* homologs of the *Synechococcus* WH8102-encoded PTOX revealed candidate homologs to be widespread in high-light *Prochlorococcus* strains (like MED4) and almost entirely absent from low-light *Prochlorococcus* strains (data not shown), but the machinery of photorespiration appears to have been maintained in low-light *Prochlorococcus* (28). How does access to a more limited range of photoprotective strategies affect the fitness of low-light *Prochlorococcus* in the space defined by CO_2 and O_2 concentrations? Is a full recovery of the photosynthetic electron transport chain impossible for low-light –CO_2 cells, or can they upregulate their photorespiratory capacity enough to compensate for the absence of PTOX? When irradiance and oxygen are both held to very low levels, can we recapitulate OMZ-like sustained, healthy growth of low-light *Prochlorococcus* in the lab, or is there another dimension, still unexplored, to the success of *Prochlorococcus* in the OMZ? In allowing us to ask these questions and many more, an extension of the present study to lower growth irradiances and to low-light strains would be of significant utility in moving us towards an understanding of the fitness landscape *Prochlorococcus* navigates in the oxygen minimum zones, and beyond.

MATERIALS AND METHODS

Growth curves

Axenic *Prochlorococcus* MED4 was grown in Sargasso seawater-based Pro99 medium (87) amended with 20 mM Na-HEPES pH 8.1. The addition of HEPES at this concentration did not affect growth rate (data not shown). Standing (un-bubbled) cultures were maintained in 25 mm tubes at ~60 μE m^2 s^{-1} and 21°C for >30 generations prior to inoculation of experimental cultures. After inoculation of experimental cultures and parallel standing controls, tubes were capped loosely and allowed to grow without bubbling for ~24 h, as pilot experiments indicated that beginning to sparge with any gas immediately after transfer led to cell death. Tubes were then sampled, fitted with sterile cap-and-frit assemblies and connected to gas lines. Gases (Airgas) were bubbled through Milli-Q water (to reduce evaporative loss from the cultures), then passed through a needle valve, an 0.2 μm filter and, in the culture tube, a fritted glass tube (Ace Glass porosity C). Cultures were sparged at a flow rate (~2 mL min^{-1}, estimated by eye) that gave 3–4 steady streams of very small bubbles, fast enough that a few dozen bubbles were always present at the air-water interface. Sparging rates were maintained as steadily as possible given the limitations of standard gas-cylinder regulators, but transient excursions to lower and higher bubbling rates did occur. Sparged cultures were found to crash more frequently than normal axenic standing cultures of MED4; the frequency of these crashes was independent of the gas used for sparging (data not shown). The apparent randomness of the crashes suggests that they may have been caused by mechanical stress during particularly high fluctuations of the gas flow rate. These crashed cultures were excluded as outliers.

Cultures were sampled daily by withdrawing 0.3 mL with a sterile transfer pipet and were immediately resealed. Sparged cultures were grown 9 at a time (3 replicates of each of 3 gases) in parallel with 3 standing cultures as a control for variations in medium, incubator temperature, etc. Once removed from the culture tubes and transferred to Eppendorf tubes, samples were protected from light and rapidly subsampled for flow cytometry and bulk fluorescence (measured with a BioTek Synergy 2 plate reader for daily monitoring; data are not reported here). For flow cytometry samples, 10 or 100 μL cultures was diluted into filtered seawater to a final volume of 1 mL. 5 μL 25% glutaraldehyde was added and samples were mixed, then incubated under foil for ~10 min. Fixed samples were flash-frozen in liquid nitrogen. Cells were counted using the blue laser on an Influx flow cytometer (Cytopeia) and a 2 μm fluorescent bead standard. Quantification of flow cytometric results was performed in the software package FlowJo.

The set of all replicates for a given gas mixture typically spanned more than one round of inoculations. Daily cell counts for all replicates within a given round of inoculations were normalized to the t = 0 count and the mean daily cell count for the paired standing cultures in that round. Normalized cell counts were then averaged among all replicates for each gas mixture and $\log_2$-transformed to give the mean number of doublings accomplished by the experimental cultures at each timepoint relative to control standing cultures. For example, if the replicates for a given gas condition were measured in three batches, A (3 experimental replicates, 3 unbubbled cultures), B (2 experimental replicates, 3 unbubbled cultures), and C (3 experimental replicates, 3 unbubbled cultures), the experimental (E) and standing (S) cultures' cell counts from each time t were treated as follows:

$$E_{batch\ A,\ tube\ 1,\ t} = \frac{\left(\frac{cells}{ml}\right)_{batch\ A,\ tube\ 1,\ t}}{\left(\frac{cells}{ml}\right)_{batch\ A,\ tube\ 1,\ t\,=\,0}}$$

$$S_{batch\ A,\ unbubbled\ tube\ 1,\ t} = \frac{\left(\frac{cells}{ml}\right)_{batch\ A,\ unbubbled\ tube\ 1,\ t}}{\left(\frac{cells}{ml}\right)_{batch\ A,\ unbubbled\ tube\ 2,\ t}}$$

$$S_{batch\ A\ ave,\ t} = \frac{1}{3}(S_{batch\ A,\ unbubbled\ tube\ 1,\ t} + S_{batch\ A,\ unbubbled\ tube\ 2,\ t} + S_{batch\ A,\ unbubbled\ tube\ 3,\ t})$$

$$E_{norm,\ tube\ 1,\ t} = \frac{E_{batch\ A,\ tube\ 1,\ t}}{S_{batch\ A\ ave,\ t}}$$

$$E_{norm,\ ave,\ t} = \frac{1}{8}(E_{norm,\ tube\ 1,\ t} + \cdots + E_{norm,\ tube\ 8,\ t})$$

$$E_{doublings\ missed,\ t} = log_2(E_{norm,\ ave,\ t})$$

Raw growth rates for control cultures were ~0.7 d^{-1}. Because of the need for normalization, growth curves are reported only through the point at which the control cultures entered stationary phase. Growth in healthy experimental cultures typically continued for 2–3 days past this point, consistent with previous observations that MED4 cultured without bubbling in Pro99 enters stationary phase due to carbon limitation (D. Lindell and S. Drakare, unpublished data).

Transcriptional profiling .

Axenic *Prochlorococcus* MED4 cultures were grown as described above ~60 µE m^2 s^{-1} and 21°C for >30 generations prior to inoculation of cultures for the shock experiment. For the experiment, three parallel 1.2-L cultures were grown in clear 2-L bottles sparged with air at ~10 mL min^{-1}. Cultures were grown to mid-log phase. One full day before the start of the experiment, 400 mL fresh medium was transferred to each of 12 clear 500-mL bottles for preequilibration with the experimental gases. Gases (Airgas) were 360 ppm CO_2 + <0.001% O_2 (balance N_2) for –O_2 shock; 40 ppm CO_2 + <0.001% O_2 (balance N_2) for –CO_2/–O_2 shock; and 40 ppm CO_2 + 21% O_2 (balance N_2) for –CO_2 shock. Three bottles were sparged vigorously with each experimental gas, and three control bottles with air, for 24 h.

The start of the experiment was staggered, with one hour separating the bottles' t = 0 timepoints. To start the experiment, each 1.2-L culture was split into six and harvested by pelleting in sterile acid-washed 250-mL centrifuge bottles in a JA-14 rotor, spinning at 8000 rpm for 10 minutes. Supernatants were immediately poured off each pellet resuspended in 3 mL fresh medium. Cell suspensions were pooled and split evenly among four 500-mL bottles, each pre-equilibrated with a different gas. Bottles were swirled to mix, 21-mL samples were immediately (t = 0) withdrawn into 30-mL centrifuge tubes, and sparging continued. Working rapidly, a 1-mL subsample was removed from each sample for bulk fluorescence and flow cytometry and protected from light; the remainder was pelleted by spinning at 12500 rpm and 4°C for 10 minutes in a JA-25.50 rotor. Meanwhile, for flow cytometry, 10 µL was transferred from each 1-mL subsample to a cryovial prepared with 990 µL filtered seawater and processed as described for growth curves. While FCM samples were undergoing fixation, a further 200 µL was removed from each 1-mL subsample to monitor bulk fluorescence in the Biotek plate-reader. Supernatants were immediately discarded and each pellet resuspended in 500 µL resuspension buffer (200 mM sucrose, 10 mM NaOAc pH 5.2, 5 mM EDTA). Cell suspensions were transferred to 1.5 mL tubes and flash-frozen in liquid nitrogen. Sampling for each timepoint was completed within 20 minutes (from withdrawal of samples to flash-freezing). Samples were taken at t = 0, 70 minutes (the 1 h timepoint), 3 h, 6 h, 9 h, 12 h, 18 h, and 24 h.

After the timecourse, RNA samples were thawed in small batches and RNA was extracted using the Zymo Research Mini RNA Isolation II kit. The manufacturer's protocol was followed except that cell lysis was allowed to continue for 20 minutes on ice. DNA was removed by treating 1 μg total nucleic acids with 2 μl Turbo DNAfree (Ambion) in a 50-μl reaction for 1 h. DNase was inactivated using the manager's reagent according to instructions. RNA was concentrated by ethanol precipitation and yields were quantified with Ribogreen (Molecular Probes). RNA amplification (using the MessageAmp II - Bacteria kit (Ambion) with 200 ng input), labeling, and hybridization to the custom Affymetrix microarray MD4-9313 was performed by the MIT BioMicro Center. Amplification, labeling, and hybridization were performed for two biological replicates.

Microarray data analysis was performed in R version 2.7.2 (88). Chip images were processed using the package Harshlight (89) prior to RMA normalization with the package affy (90). Empirical Bayes fitting of linear models was performed (comparing each condition at time *t* to the air control at time *t*) and *q*-values (roughly, false discovery rates) calculated using the package limma (91). To be said to show significant differential expression, a gene was required to have both $q \leq 0.01$ and mean fold change ≥ 2 (normalized to air at time *t* and to t = 0) at the same timepoint. "Borderline" significant genes met one of these conditions but not both. In addition, the function mas5calls (92) in the package affy was used as a coarse filter for genes with unreliably low signal intensities. Some genes with low baseline signal intensities are interesting: these are the genes that are "absent" at all timepoints in the air control and at t = 0 under an experimental condition, but are then upregulated at subsequent timepoints under that experimental condition; such genes would have 12 "absent" calls among the 20 arrays (2 biological replicates of 5 timepoints of the air control and of the given experimental condition). In order to avoid discarding these genes inappropriately, the threshold for passing the Present/ Marginal/Absent test was set at ≤60% "absent" calls.

Clustering of significantly differentially expressed genes was performed using the package Mfuzz (93). $-CO_2$ shock-responsive expression profiles were pooled with $-CO_2/-O_2$ shock-responsive expression profiles and the combined dataset was subjected to fuzzy *c*-means clustering. Clustering with c = 10 and m = 1.15 gave well-populated clusters with strong alpha cores. For analysis of differences in clustering behavior between genes responding to $-CO_2$ and $-CO_2/-O_2$ shocks, only those genes that were responsive to both conditions were considered. Gene functional categorizations are as annotated in CyanoBase (http://genome.kazusa.or.jp/ cyanobase/) as of May 2009.

WORKS CITED

1. R Goericke, RJ Olson, and A Shalapyonok (2000). A novel niche for *Prochlorococcus* sp. in low-light suboxic environments in the Arabian Sea and the Eastern Tropical North Pacific. *Deep Sea Res I: Oceanogr Res Papers* **47:**1183-1205.
2. KM Scott, M Henn-Sax, TL Harmer, DL Longo, CH Frame, and CM Cavanaugh (2007). Kinetic isotope effect and biochemical characterization of form IA Rubisco from the marine cyanobacterium *Prochlorococcus marinus* MIT9313. *Limnol Oceanogr* **52:**2199-2204.
3. JB Waterbury, SW Watson, RRL Guillard, and LE Brand (1979). Widespread occurrence of a unicellular, marine, planktonic cyanobacterium. *Nature* **277:**293-294.
4. PW Johnson and JMcN Sieburth (1979). Chroococcoid cyanobacteria in the sea: A ubiquitous and diverse phototrophic biomass. *Limnol Oceanogr* **24:**928-935.
5. SW Chisholm, RJ Olson, ER Zettler, R Goericke, JB Waterbury, and NA Welschmeyer (1988). A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334:**340-343.
6. F Partensky, WR Hess, and D Vaulot (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* **63:**106-27.
7. R Goericke and DJ Repeta (1992). The pigments of *Prochlorococcus marinus*: The presence of divinyl chlorophyll *a* and *b* in a marine procaryote. *Limnol Oceanogr* **37:**425-433.
8. A Dufresne, M Salanoubat, F Partensky, F Artiguenave, IM Axmann, V Barbe, S Duprat, MY Galperin, EV Koonin, F Le Gall, KS Makarova, M Ostrowski, S Oztas, C Robert, IB Rogozin, DJ Scanlan, N Tandeau de Marsac, J Weissenbach, P Wincker, YI Wolf, and WR

Hess (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A* **100:**10020-5.

9. GC Kettler, AC Martiny, K Huang, J Zucker, ML Coleman, S Rodrigue, F Chen, A Lapidus, S Ferriera, J Johnson, C Steglich, GM Church, P Richardson, and SW Chisholm (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3:**e231.

10. ML Coleman, MB Sullivan, AC Martiny, C Steglich, K Barry, EF Delong, and SW Chisholm (2006). Genomic islands and the ecology and evolution of *Prochlorococcus. Science* **311:**1768-70.

11. GE Hutchinson (1957). Concluding remarks. *Cold Springs Harbor Symp Quant Biol* **22:**415-427.

12. LR Moore, R Goericke, and SW Chisholm (1995). Comparative physiology of *Synechococcus* and *Prochlorococcus*: Influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Mar Ecol Prog Ser* **116:**259-275.

13. LR Moore and SW Chisholm (1999). Photophysiology of the marine cyanobacterium *Prochlorococcus*: Ecotypic differences among cultured isolates. *Limnol Oceanogr* **44:**628-638.

14. LR Moore, AF Post, G Rocap, and SW Chisholm (2002). Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus. Limnol Oceanogr* **47:**989-996.

15. ZI Johnson, ER Zinser, A Coe, NP McNulty, EM Woodward, and SW Chisholm (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311:**1737-40.

16. A Shimada, M Nishijima, and T Maruyama (1995). Seasonal appearance of *Prochlorococcus* in Suruga Bay, Japan in 1992-1993. *J Oceanogr* **51:**289-300.

17. BA Van Mooy, G Rocap, HF Fredricks, CT Evans, and AH Devol (2006). Sulfolipids dramatically decrease phosphorus demand by picocyanobacteria in oligotrophic marine environments. *Proc Natl Acad Sci U S A* **103:**8607-12.

18. BAS Van Mooy, HF Fredricks, BE Pedler, ST Dyhrman, DM Karl, M Koblizek, MW Lomas, TJ Mincer, LR Moore, T Moutin, MS Rappe, and EA Webb (2009). Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. *Nature* **458:**69-72.

19. EA Dinsdale, O Pantos, S Smriga, RA Edwards, F Angly, L Wegley, M Hatay, D Hall, E Brown, M Haynes, L Krause, E Sala, SA Sandin, RV Thurber, BL Willis, F Azam, N Knowlton, and F Rohwer (2008). Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE* **3:**e1584.

20. D Vaulot, D Marie, RJ Olson, and SW Chisholm (1995). Growth of *Prochlorococcus*, a photosynthetic prokaryote, in the equatorial pacific ocean. *Science* **268:**1480-1482.

21. Z Johnson, ML Landry, RR Bidigare, SL Brown, L Campbell, J Gunderson, J Marra, and C Trees (1999). Energetics and growth kinetics of a deep *Prochlorococcus* spp. population in the Arabian Sea. *Deep Sea Res II: Topical Stud Oceanogr* **46:**1719-1743.

22. JA Raven, CS Cockell, and CL De La Rocha (2008). The evolution of inorganic carbon concentrating mechanisms in photosynthesis. *Philos Trans R Soc Lond B Biol Sci* **363:**2641-50.

23. S Gutteridge and J Pierce (2006). A unified theory for the basis of the limitations of the primary reaction of photosynthetic CO_2 fixation: Was Dr. Pangloss right? *Proc Natl Acad Sci U S A* **103:**7203-4.

24. IA Berg, D Kockelkorn, W Buckel, and G Fuchs (2007). A 3-hydroxypropionate/4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway in archaea. *Science* **318:**1782-6.

25. JA Raven, M Giordano, and J Beardall (2008). Insights into the evolution of CCMs from comparisons with other resource acquisition and assimilation processes. *Physiol Plant* **133:**4-14.

26. JR Reinfelder, AM Kraepiel, and FM Morel (2000). Unicellular C4 photosynthesis in a marine diatom. *Nature* **407:**996-9.

27. GD Price, MR Badger, FJ Woodger, and BM Long (2008). Advances in understanding the cyanobacterial CO_2-concentrating-mechanism (CCM): Functional components, Ci transporters, diversity, genetic regulation and prospects for engineering into plants. *J Exp Bot* **59:**1441-61.

28. M Eisenhut, W Ruth, M Haimovich, H Bauwe, A Kaplan, and M Hagemann (2008). The photorespiratory glycolate metabolism is essential for cyanobacteria and might have been conveyed endosymbiontically to plants. *Proc Natl Acad Sci U S A* **105:**17199-17204.

29. FJ Woodger, MR Badger, and GD Price (2005). Sensing of inorganic carbon limitation in *Synechococcus* PCC7942 is correlated with the size of the internal inorganic carbon pool and involves oxygen. *Plant Physiol* **139:**1959-69.

30. BH ter Kuile and HV Westerhoff (2001). Transcriptome meets metabolome: Hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett* **500:**169-171.

31. S Rossell, CC van der Weijden, A Lindenbergh, A van Tuijl, C Francke, BM Bakker, and HV Westerhoff (2006). Unraveling the complexity of flux regulation: A new method demonstrated for nutrient starvation in *Saccharomyces cerevisiae. Proc Natl Acad Sci U S A* **103:**2166-2171.

32. C Steglich, M Futschik, D Lindell, T Rector, R Steen, and SW Chisholm, manuscript in preparation.

33. C Steglich, ME Futschik, D Lindell, B Voss, SW Chisholm, and WR Hess (2008). The challenge of regulation in a minimal photoautotroph: Non-coding RNAs in *Prochlorococcus. PLoS Genet* **4:**e1000173.

34. N Adir, H Zer, S Shochat, and I Ohad (2003). Photoinhibition—A historical perspective. *Photosynth Res* **76:**343-370.

35. SC Andrews, AK Robinson, and F Rodriguez-Quinones (2003). Bacterial iron homeostasis. *FEMS Microbiol Rev* **27:**215-237.

36. M Eisenhut, EA von Wobeser, L Jonas, H Schubert, BW Ibelings, H Bauwe, HC Matthijs, and M Hagemann (2007). Long-term response toward inorganic carbon limitation in wild type and glycolate turnover mutants of the cyanobacterium *Synechocystis* sp. strain PCC 6803. *Plant Physiol* **144:**1946-59.

37. ER Zinser, D Lindell, ZI Johnson, ME Futschik, C Steglich, ML Coleman, MA Wright, T Rector, R Steen, N McNulty, LR Thompson, and SW Chisholm (2009). Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, prochlorococcus. *PLoS ONE* **4:**e5135.

38. M Tamoi, A Murakami, T Takeda, and S Shigeoka (1998). Lack of light/dark regulation of enzymes involved in the photosynthetic carbon reduction cycle in cyanobacteria, *Synechococcus* PCC 7942 and *Synechocystis* PCC 6803. *Biosci Biotechnol Biochem* **62:**374-376.

39. M Tamoi, T Miyazaki, T Fukamizo, and S Shigeoka (2005). The Calvin cycle in cyanobacteria is regulated by CP12 via the NAD(H)/NADP(H) ratio under light/dark conditions. *Plant J* **42:**504-513.

40. KD Hagen and JC Meeks (2001). The unique cyanobacterial protein OpcA is an allosteric effector of glucose-6-phosphate dehydrogenase in *Nostoc punctiforme* ATCC 29133. *J Biol Chem* **276:**11477-11486.

41. J Frias-Lopez, Y Shi, GW Tyson, ML Coleman, SC Schuster, SW Chisholm, and EF DeLong (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A* **105:**3805-10.

42. MR Badger, A Kaplan, and JA Berry (1980). Internal inorganic carbon pool of *Chlamydomonas reinhardtii*: Evidence for a carbon dioxide-concentrating mechanism. *Plant Physiol* **66:**407-413.

43. Y Marcus, D Zenvirth, E Harel, and A Kaplan (1982). Induction of HCO_3 transporting capability and high photosynthetic affinity to inorganic carbon by low concentration of CO(2) in *Anabaena variabilis*. *Plant Physiol* **69:**1008-1012.

44. M Shibata, H Ohkawa, T Kaneko, H Fukuzawa, S Tabata, A Kaplan, and T Ogawa (2001). Distinct constitutive and low-CO_2-induced CO_2 uptake systems in cyanobacteria: Genes involved and their phylogenetic relationship with homologous genes in other organisms. *Proc Natl Acad Sci U S A* **98:**11789-94.

45. D Sültemeyer, B Klughammer, MR Badger, and GD Price (1998). Fast induction of high-affinity HCO_3^- transport in cyanobacteria. *Plant Physiol* **116:**183.

46. FJ Woodger, MR Badger, and GD Price (2003). Inorganic carbon limitation induces transcripts encoding components of the CO_2-concentrating mechanism in *Synechococcus* sp. PCC7942 through a redox-independent pathway. *Plant Physiol* **133:**2069-80.

47. A Kaplan, R Schwarz, J Lieman-Hurwitz, M Ronen-Tarazi, and L Reinhold (1994). Physiological and molecular studies on the response of cyanobacteria to changes in the ambient inorganic carbon concentration. In *The Molecular Biology of Cyanobacteria.* DA Bryant, ed. Dordrecht: Kluwer Academic Publishers.

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

49. GD Price, FJ Woodger, MR Badger, SM Howitt, and L Tucker (2004). Identification of a SulP-type bicarbonate transporter in marine cyanobacteria. *Proc Natl Acad Sci U S A* **101:**18228-33.

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

51. MR Badger, GD Price, BM Long, and FJ Woodger (2006). The environmental plasticity and ecological genomics of the cyanobacterial CO_2 concentrating mechanism. *J Exp Bot* **57:**249-65.

52. KA Palinska, W Laloui, S Bedu, S Loiseaux-de Goer, AM Castets, R Rippka, and N Tandeau de Marsac (2002). The signal transducer PII and bicarbonate acquisition in *Prochlorococcus marinus* PCC 9511, a marine cyanobacterium naturally deficient in nitrate and nitrite assimilation. *Microbiology* **148:**2405-2412.

53. HL Wang, BL Postier, and RL Burnap (2004). Alterations in global patterns of gene expression in *Synechocystis* sp. PCC 6803 in response to inorganic carbon limitation and the inactivation of *ndhR*, a LysR family regulator. *J Biol Chem* **279:**5739-51.

54. GS Espie and RA Kandasamy (1994). Monensin inhibition of Na^+-dependent HCO_3^- transport distinguishes it from Na^+-independent HCO_3^- transport and provides evidence for Na^+/HCO_3^- symport in the cyanobacterium *Synechococcus* UTEX 625. *Plant Physiol* **104:**1419-1428.

55. Y Harano, I Suzuki, S Maeda, T Kaneko, S Tabata, and T Omata (1997). Identification and nitrogen regulation of the cyanase gene from the cyanobacteria *Synechocystis* sp. Strain PCC 6803 and *Synechococcus* sp. Strain PCC 7942. *J Bacteriol* **179:**5744-5750.

56. NAM Dolganov, D Bhaya, and AR Grossman (1995). Cyanobacterial protein with similarity to the chlorophyll *a/b* binding proteins of higher plants: Evolution and regulation. *Proc Natl Acad Sci U S A* **92:**636-640.

57. Q He, N Dolganov, O Bjorkman, and AR Grossman (2001). The high light-inducible polypeptides in *Synechocystis* PCC6803. Expression and function in high light. *J Biol Chem* **276:**306-14.

58. D Lindell, JD Jaffe, ML Coleman, ME Futschik, IM Axmann, T Rector, G Kettler, MB Sullivan, R Steen, WR Hess, GM Church, and SW Chisholm (2007). Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449:**83-86.

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

60. AW Thompson (2009). *Iron and Prochlorococcus.* MIT Ph.D. Thesis.

61. C Steglich, M Futschik, T Rector, R Steen, and SW Chisholm (2006). Genome-wide analysis of light sensing in *Prochlorococcus*. *J Bacteriol* **188:**7796-806.

62. Q Wang, S Jantaro, B Lu, W Majeed, M Bailey, and Q He (2008). The high light-inducible polypeptides stabilize trimeric photosystem I complex under high light conditions in *Synechocystis* PCC 6803. *Plant Physiol* **147:**1239-50.

63. H Xu, D Vavilin, C Funk, and W Vermaas (2004). Multiple deletions of small Cab-like proteins in the cyanobacterium *Synechocystis* sp. PCC 6803: Consequences for pigment biosynthesis and accumulation. *J Biol Chem* **279:**27971-9.

64. D Bhaya, A Dufresne, D Vaulot, and A Grossman (2002). Analysis of the *hli* gene family in marine and freshwater cyanobacteria. *FEMS Microbiol Lett* **215:**209-19.

65. D Lindell, MB Sullivan, ZI Johnson, AC Tolonen, F Rohwer, and SW Chisholm (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci U S A* **101:**11013-8.

66. GC Kettler, personal communication.

67. C Funk and W Vermaas (1999). A cyanobacterial gene family coding for single-helix proteins resembling part of the light-harvesting proteins from higher plants. *Biochemistry* **38:**9397-404.

68. DA Parsell and S Lindquist (1993). The function of heat-shock proteins in stress tolerance: Degradation and reactivation of damaged proteins. *Annu Rev Genet* **27:**437-496.

69. D Kültz (2005). Molecular and evolutionary basis of the cellular stress response. *Annu Rev Physiol* **67:**225-57.

70. AC Martiny, ML Coleman, and SW Chisholm (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proc Natl Acad Sci U S A* **103:**12552-7.

71. GC Walker (1984). Mutagenesis and inducible responses to deoxyribonucleic acid damage in *Escherichia coli. Microbiol Rev* **48:**60-93.

72. M Osburne, personal communication.

73. F Taddei, I Matic, and M Radman (1995). cAMP-dependent SOS induction and mutagenesis in resting bacterial populations. *Proc Natl Acad Sci U S A* **92:**11736-40.

74. F Domain, L Houot, F Chauvat, and C Cassier-Chauvat (2004). Function and regulation of the cyanobacterial genes *lexA*, *recA* and *ruvB*: LexA is critical to the survival of cells facing inorganic carbon starvation. *Mol Microbiol* **53:**65-80.

75. LV Aseev, AA Levandovskaya, LS Tchufistova, NV Scaptsova, and IV Boni (2008). A new regulatory circuit in ribosomal protein operons: S2-Mediated control of the *rpsB-tsf* expression *in vivo*. *RNA* **14:**1882-94.

76. MV Sukhodolets and S Garges (2003). Interaction of *Escherichia coli* RNA polymerase with the ribosomal protein S1 and the Sm-like ATPase Hfq. *Biochemistry* **42:**8022-34.

77. NI Krinsky (1979). Carotenoid protection against oxidation. *Pure Appl Chem* **51:**649-660.

78. T Götz, U Windhovel, P Boger, and G Sandmann (1999). Protection of photosynthesis against ultraviolet-B radiation by carotenoids in transformants of the cyanobacterium *Synechococcus* PCC7942. *Plant Physiol* **120:**599-604.

79. WR Hess, G Rocap, CS Ting, FW Larimer, S Stilwagen, J Lamerdin, and SW Chisholm (2001). The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. *Photosynth Res* **70:**53-71.

80. P Stickforth, S Steiger, WR Hess, and G Sandmann (2003). A novel type of lycopene epsilon-cyclase in the marine cyanobacterium *Prochlorococcus marinus* MED4. *Arch Microbiol* **179:**409-415.

81. AA Yeliseev, JM Eraso, and S Kaplan (1996). Differential carotenoid composition of the B875 and B800-850 photosynthetic antenna complexes in *Rhodobacter sphaeroides* 2.4.1: Involvement of spheroidene and spheroidenone in adaptation to changes in light intensity and oxygen availability. *J Bacteriol* **178:**5877-83.

82. R Prasanna, A Pabby, S Saxena, and PK Singh (2004). Modulation of pigment profiles of *Calothrix elenkenii* in response to environmental changes. *J Plant Physiol* **161:**1125-1132.

83. S Bailey, A Melis, KR Mackey, P Cardol, G Finazzi, G van Dijken, GM Berg, K Arrigo, J Shrager, and A Grossman (2008). Alternative photosynthetic electron flow to oxygen in marine *Synechococcus. Biochim Biophys Acta* **1777:**269-76.

84. KRM Mackey, A Paytan, AR Grossman, and S Bailey (2008). A photosynthetic strategy for coping in a high-light, low-nutrient environment. *Limnol Oceanogr* **53:**900-913.

85. CB Osmond (1981). Photorespiration and photoinhibition. Some implications for the energetics of photosynthesis. *Biochim Biophys Acta* **639:**77.

86. D McDougald, L Gong, S Srinivasan, E Hild, L Thompson, K Takayama, SA Rice, and S Kjelleberg (2002). Defences against oxidative stress during starvation in bacteria. *Antonie Van Leeuwenhoek* **81:**3-13.

87. LR Moore, A Coe, ER Zinser, MA Saito, MB Sullivan, D Lindell, K Frois-Moniz, J Waterbury, and SW Chisholm (2007). Culturing the marine cyanobacterium *Prochlorococcus. Limnol Oceanogr: Meth* **5:**353-362.

88. R Development Core Team (2008). R: A language and environment for statistical computing. <http://www.R-project.org>.

89. M Suárez-Fariñas, M Pellegrino, KM Wittkowski, and MO Magnasco (2005). Harshlight: A corrective make-up program for microarray chips. *BMC Bioinformatics* **6:**294.

90. BM Bolstad, RA Irizarry, M Astrand, and TP Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19:**185-93.

91. GK Smyth (2005). Limma: Linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* R Gentleman, V Carey, S Dudoit, R Irizarry, and W Huber, eds. New York: Springer.

92. WM Liu, R Mei, X Di, TB Ryder, E Hubbell, S Dee, TA Webster, CA Harrington, MH Ho, J Baid, and SP Smeekens (2002). Analysis of high density expression microarrays with signed-rank algorithms. *Bioinformatics* **18:**1593-1599.

93. ME Futschik and B Carlisle (2005). Noise-robust soft clustering of gene expression time-course data. *J Bioinform Comput Biol* **3:**965-88.

Appendix 1

Notes on the Growth and Chlorophyll Fluorescence of *Prochlorococcus* under Carbon and Oxygen Limitation

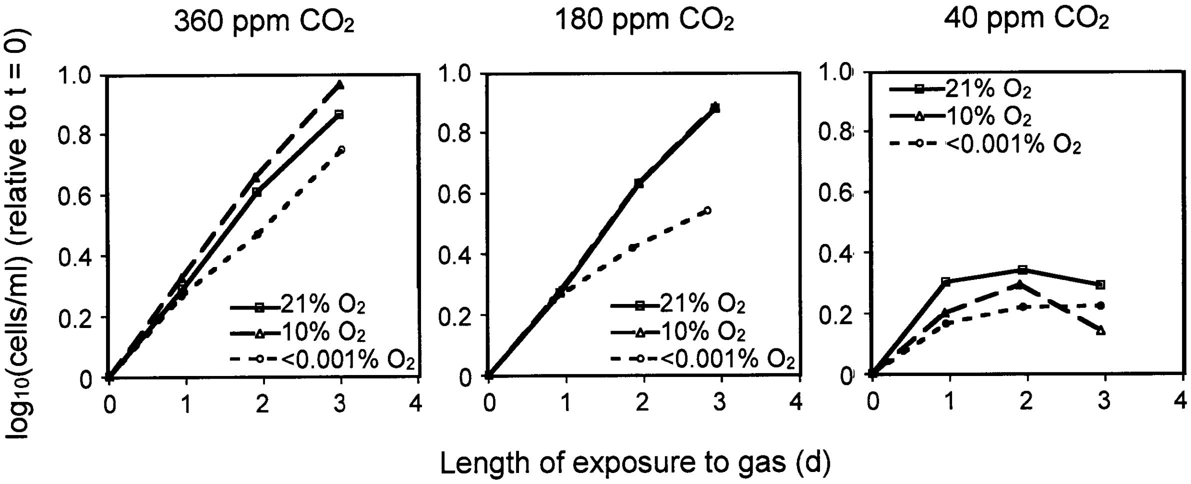


Figure 1. Representative raw growth curves from the nine gas conditions studied. Each curve presents data from one replicate of one condition. Curves are not normalized to the paired standing cultures. To facilitate comparison between conditions, data in each curve is presented relative to its own t = 0.

Note on gas-dependent changes in fluorescence signals

In the simplest case, bulk fluorescence (measured with a 96-well plate reader) scales linearly with the product of two parameters measured for single cells by flow cytometry (FCM): mean chlorophyll fluorescence per cell and the number of cells per unit volume. In the gas shock timecourse, the relationship between bulk fluorescence data and FCM data held roughly constant (within error) for the air control and the –O_2 shock condition, but not in the –CO_2/–O_2 shock and –CO_2 shock conditions, with the starkest difference in the latter. The plate reader interrogates a broader spectral region than does the blue laser used for FCM, and it is the plate-reader signal rather than the FCM signal that changes over the first 24 h (Fig. 2, 3); thus, one possible explanation for the divergence seen above is that the CO_2 starvation conditions lead to changes in accessory pigments that in turn change the cell's handling of light at wavelengths outside the narrow FCM band but still within the range used for excitation, or detected as emission, by the plate reader.

Also, note that, while cells under the –O_2 shock condition keep dividing as fast as cells in the air control (Part II, Fig. 2), the bulk fluorescence signal stays flat for the –O_2 shock condition, presumably because the mean chlorophyll fluorescence per cell decreases (Fig. 3). The decrease seen under –O_2 shock could indicate a dilution effect: these cells continue to divide over the 24-h timecourse (Part II, Fig. 2), but if chlorophyll synthesis does not keep pace with cell division, chlorophyll fluorescence per cell would decline. Such a dilution effect would not be in evidence in the –CO_2 and –CO_2/–O_2 shock conditions, in which cell division is minimal. Whatever the reason for the change under –O_2 shock, the clear difference between air and –O_2 shock does argue against the possibility that a failure to manipulate the experimental condition (as would have happened if, e.g., the Airgas-certified 360 ppm CO_2 + <0.001% O_2 gas cylinder were contaminated with oxygen) caused the lack of a transcriptional response to $-O_2$ shock.

Note that, in figs. 2 and 3, instrument problems with the plate reader prevented the measurement of bulk fluorescence at t = 24 h and caused the loss of bulk fluorescence data for two of the three t = 0 replicates and one of the three t = 1 h replicates in all conditions tested. In addition, the t = 18 h flow cytometry sample for one of the three air replicates was corrupted. With these exceptions, data points represent the mean of three replicates; error bars are ±1 standard deviation.

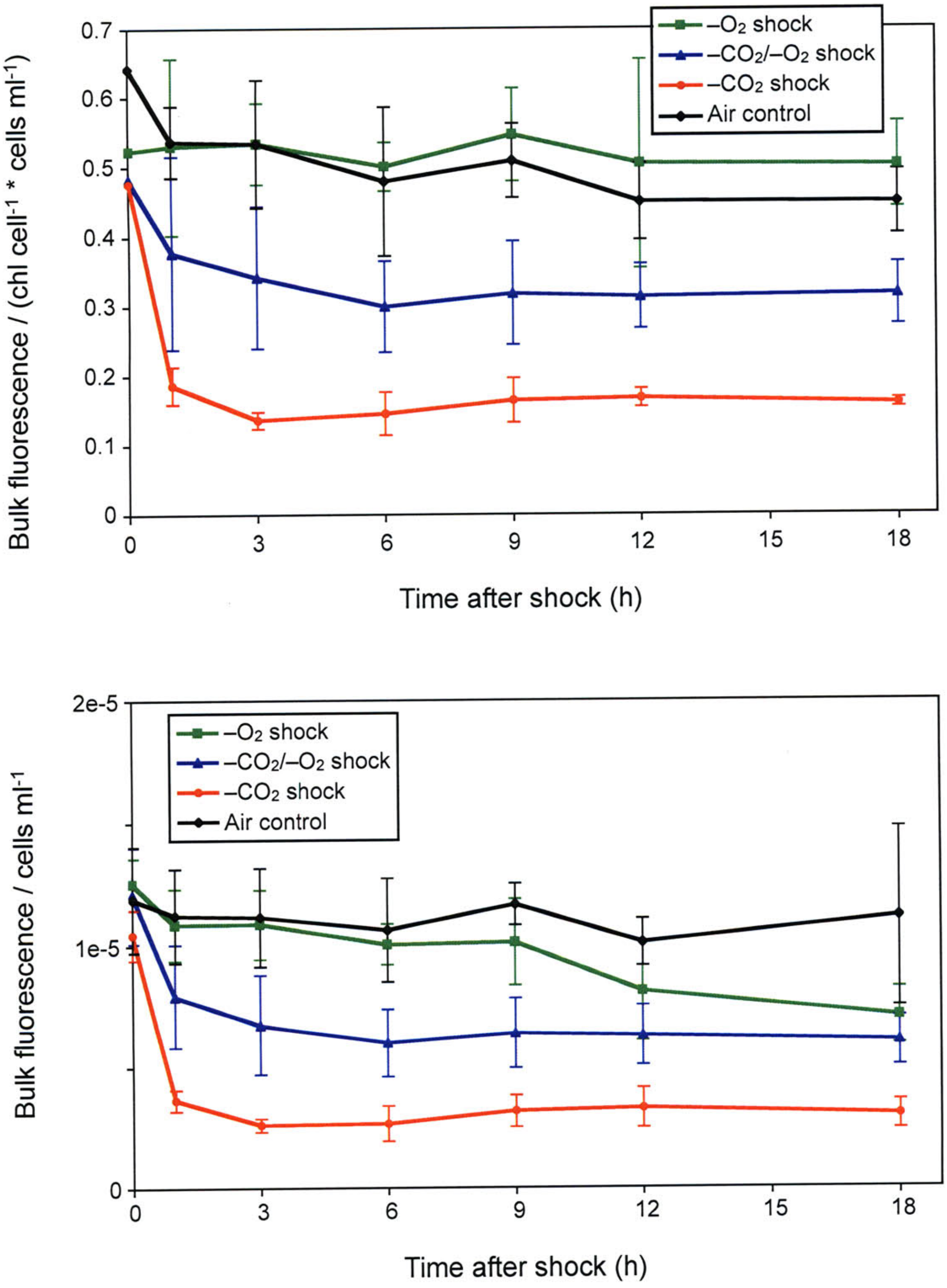


Figure 2. Changes in the relationship between bulk and single-cell measurements of fluorescence under CO_2 starvation.

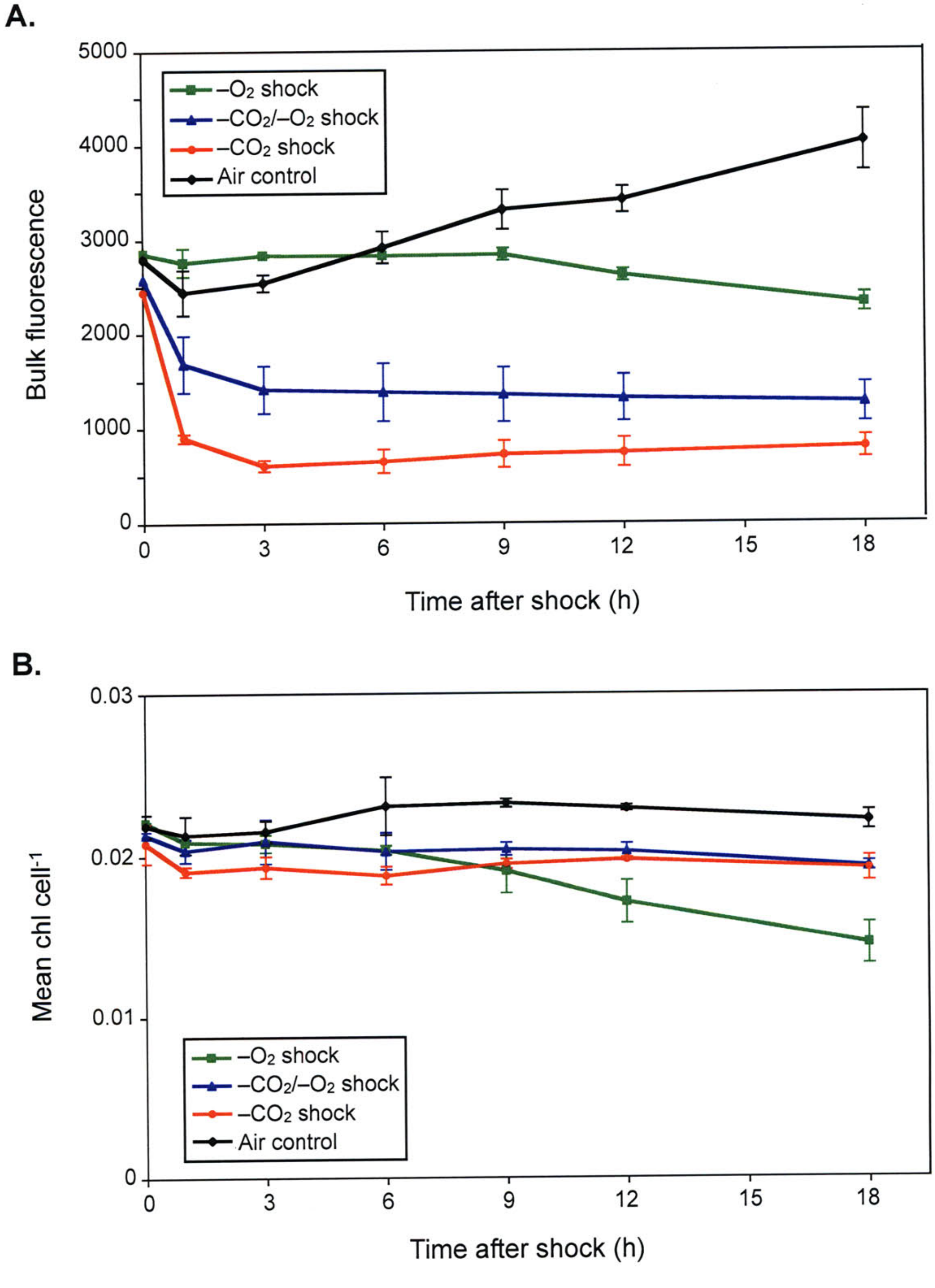


Figure 3. Gas-dependent changes in chlorophyll fluorescence per cell.

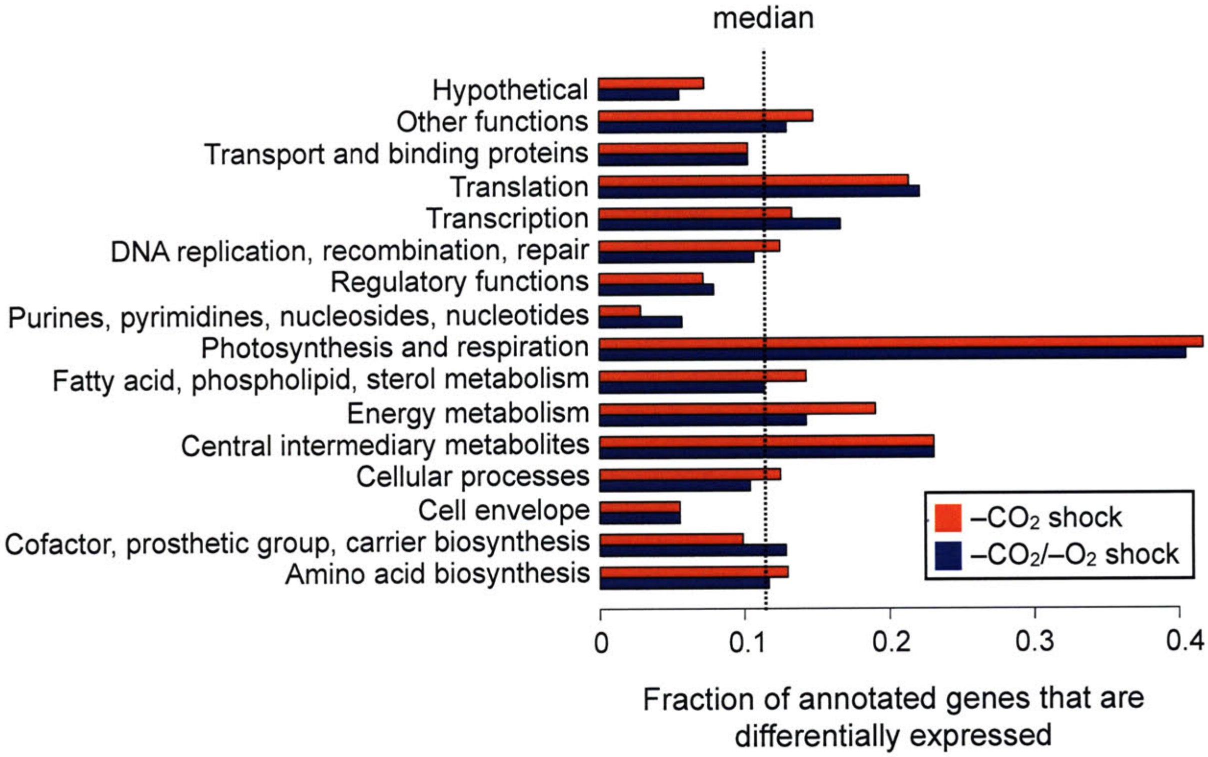


Figure 4. Functional categorizations of genes showing significant differential expression at any time sampled under $-CO_2$ shock (red) or under $-CO_2/-O_2$ shock (blue). When the data from the $-CO_2$ and $-CO_2/-O_2$ shocks is considered without reference to the development of transcriptional responses over the timecourse, the resemblance between the two is striking. As seen in Figs. 5 and 6, however, ignoring the time component masks the key difference between the two conditions.

Appendix 2

T Dammeyer, SC Bagby, MB Sullivan, SW Chisholm, and N Frankenberg-Dinkel (2008). Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. *Curr Biol* **18:**442-8.

Current Biology *18*, 442–448, March 25, 2008 ©2008 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2008.02.067

Report

Efficient Phage-Mediated Pigment Biosynthesis in Oceanic Cyanobacteria

Thorben Dammeyer,[1] Sarah C. Bagby,[2] Matthew B. Sullivan,[3]
Sallie W. Chisholm,[2,3] and Nicole Frankenberg-Dinkel[1,*]
[1]Physiology of Microorganisms
Ruhr-University Bochum
Universitaetsstr. 150
44780 Bochum
Germany
[2]Department of Biology
[3]Department of Civil and Environmental Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Summary

Although the oceanic cyanobacterium *Prochlorococcus* harvests light with a chlorophyll antenna [1–3] rather than with the phycobilisomes that are typical of cyanobacteria, some strains express genes that are remnants of the ancestral *Synechococcus* phycobilisomes [4]. Similarly, some *Prochlorococcus* cyanophages, which often harbor photosynthesis-related genes [5], also carry homologs of phycobilisome pigment biosynthesis genes [6, 7]. Here, we investigate four such genes in two cyanophages that both infect abundant *Prochlorococcus* strains [8]: homologs of heme oxygenase (*ho1*), 15,16-dihydrobiliverdin:ferredoxin oxidoreductase (*pebA*), ferredoxin (*petF*) in the myovirus P-SSM2, and a phycocyanobilin:ferredoxin oxidoreductase (*pcyA*) homolog in the myovirus P-SSM4. We demonstrate that the phage homologs mimic the respective host activities, with the exception of the divergent phage PebA homolog. In this case, the phage PebA single-handedly catalyzes a reaction for which uninfected host cells require two consecutive enzymes, PebA and PebB. We thus renamed the phage enzyme phycoerythrobilin synthase (PebS). This gene, and other pigment biosynthesis genes encoded by P-SSM2 (*petF* and *ho1*), are transcribed during infection, suggesting that they can improve phage fitness. Analyses of global ocean metagenomes show that PcyA and Ho1 occur in both cyanobacteria and their phages, whereas the novel PebS-encoding gene is exclusive to phages.

Results and Discussion

Although they do not have the typical cyanobacterial phycobilisome antennae [4], *Prochlorococcus* cells carry both the cellular machinery for the biosynthesis of the phycobiliprotein pigments phycocyanobilin (PCB) and phycoerythrobilin (PEB) (Figure 1, Table 1) and one of the three phycobilisome structural proteins [4]. Several lines of evidence suggest that these are playing some functional role [9–12]. The occurrence of phycobilisome-related genes in some marine cyanophage genomes (Table 1) is further evidence in support of a role for these genes in cell fitness. Although these types of genes are not found in cyanophage podoviruses [13–15], cyanophage myovirus genomes contain between one and three phycobilisome-related genes [14, 16, 17]. The two *Prochlorococcus* myovirus genomes available include some combination of the putative bilin reductase genes *pebA* and *pcyA*, heme oxygenase (*ho1*), and ferredoxin (*petF*). *Synechococcus* myoviruses, on the other hand, carry *cpeT* (a putative phycobiliprotein lyase) alone. The phage protein PetF is similar to a plant-type [2Fe-2S] ferredoxin [18], suggesting that it might serve as an electron donor for Ho1 and PebA (renamed PebS) in myovirus P-SSM2.

pebA from the Cyanophage P-SSM2 Encodes a Phycoerythrobilin Synthase

To better understand the role of these genes in cyanophages, we first investigated whether the phage *pebA*-encoded homolog, PebA_P-SSM2, whose sequence is highly divergent from cyanobacterial PebA, encodes a functional ferredoxin-dependent bilin reductase (FDBR). We found that recombinant PebA_P-SSM2 (Figure S1 available online) was highly active in vitro with biliverdin IXα (BV) as a substrate. However, instead of the expected two-electron-reduced product 15,16-dihydrobiliverdin (15,16-DHBV) (Figure 1), the PebA_P-SSM2-catalyzed reaction yielded the four-electron-reduced chromophore PEB (Figure 2 and Figure S2). Thus, the phage enzyme directly converts BV to PEB, whereas the host cells require the sequential action of the two enzymes PebA and PebB [7]. Because of this new FDBR activity, we renamed PebA_P-SSM2 as PebS, phycoerythrobilin synthase, by analogy with phytochromobilin synthase [6, 19]. PebS is only the second FDBR, after PcyA [20], to perform a formal four-electron reduction. The PebS-mediated reduction proceeds faster with PetF_P-SSM2 than with standard assay ferredoxin as a redox partner (data not shown), likely because of more efficient electron transfer among the phage proteins.

Phycoerythrobilin Synthase Converts BV via the Semireduced Intermediate 15,16-DHBV

By slowing down the in vitro reaction, we were able to observe the transient accumulation of the semireduced reaction intermediate 15,16-DHBV (Figure 2), as well as the appearance of the reaction product PEB. Hence, the sequence of reductions performed by the phage PebS is identical to that in the consecutive action of PebA and PebB in cyanobacterial and algal cells [21, 22]. Specifically, PebS and PebA alike catalyze the reduction of BV at the 15,16 double bond; PebS holds onto this 15,16-DHBV intermediate, whereas PebA passes it to PebB [22]. PebS and PebB then catalyze a reduction of the A-ring vinyl moiety of 15,16-DHBV, a formal 2,3 reduction most likely followed by isomerization to *3Z*-PEB.

PebS accepts the intermediate 15,16-DHBV as a substrate and completes its reduction to PEB, indicating that substrate recognition is not inhibited by a reduced 15,16 double bond. Nevertheless, the overall turnover rate of 15,16-DHBV to PEB in vitro was lower than that of the proper substrate BV (data not shown), possibly because of the instability of free 15,16-DHBV or because of an association rate constant (k_{on}) that is lower for free 15,16-DHBV's alternate conformation. We therefore propose that the intermediate is never released from the

[*]Correspondence: nicole.frankenberg@rub.de

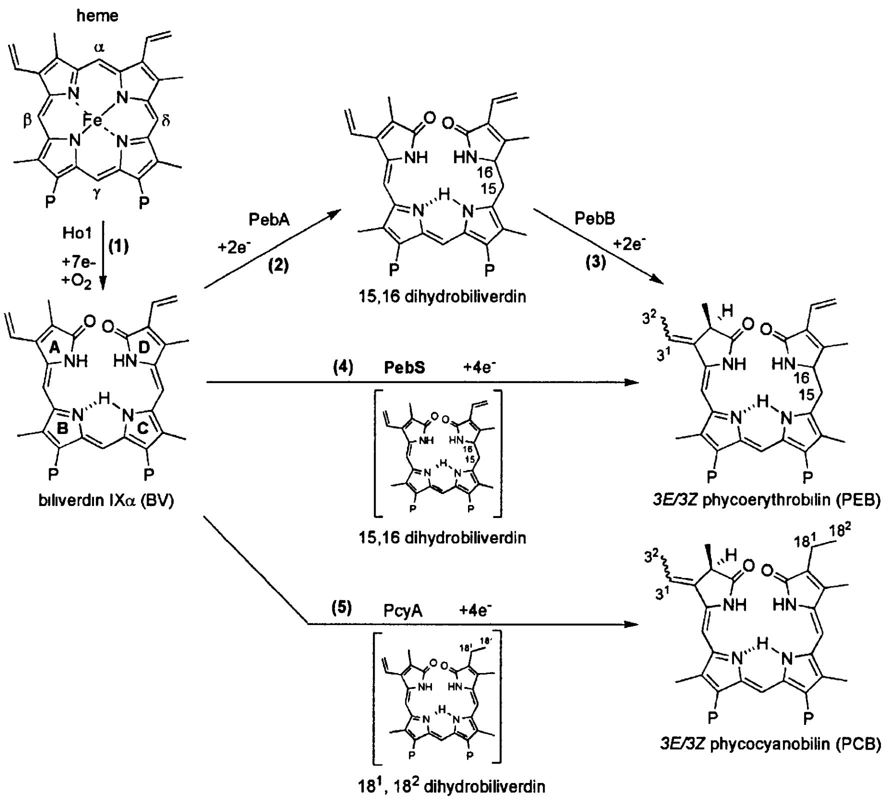


Figure 1. Biosynthesis of Open-Chain Tetrapyrroles in Cyanophages and Their Hosts

The first open-chain product biliverdin IXα (BV) is derived from heme by a heme oxygenase-catalyzed reaction (1). This open-chain product is the substrate for various enzymes of the FDBR family. Two sequential two-electron reductions catalyzed by PebA (2) and PebB (3) yield phycoerythrobilin (PEB). Phycoerythrobilin synthase (PebS) catalyzes a unique four-electron reduction of BV IXα to PEB (4). In a similar reaction, BV is reduced to phycocyanobilin (PCB) by the action of PcyA (5). All FDBRs obtain the required number of electrons from the redox cofactor ferredoxin. Enzymes catalyzing reactions 2 and 3 are found in cyanobacterial host cells. Enzymes catalyzing reactions 1 and 5 are found in both cyanobacterial host cells and cyanophages. The PebS enzyme (4) is found solely in cyanophage populations. The following abbreviations are used: propionate side chain (P), heme oxygenase (Ho1), PCB:ferredoxin oxidoreductase (PcyA), 15,16-DHBV:ferredoxin oxidoreductase (PebA), PEB:ferredoxin oxidoreductase (PebB), and PEB synthase (PebS).

enzyme but is directly converted to PEB. Similarly, cyanobacterial PebA and PebB have been postulated to associate transiently to perform metabolic channeling, the direct enzyme-to-enzyme transfer of the unstable intermediate 15,16-DHBV [22].

Reconstruction of Cyanophage Bilin Biosynthetic Pathways in *E. coli*

The functionality of the full-phage-encoded bilin biosynthetic pathway was further tested by reconstruction of the pathway with several expression vectors in *E. coli*, a useful system because *E. coli* can synthesize heme but cannot catabolize it. Extracted pigment from an *E. coli* culture expressing pTD*ho1* was indistinguishable from that of a BV IXα control (Figure S2), demonstrating that *ho1*_P-SSM2 encodes an active heme oxygenase that regiospecifically cleaves the heme macrocycle at the α-meso carbon.

Full reconstruction of PEB and PCB biosynthesis in *E. coli* was achieved with two more vectors, pTD*ho1pebS* and pTD*ho1pcyA*. Again, high-pressure liquid chromatography (HPLC) elution profiles of extracted expression culture pigments confirmed the expected products: The *pebS* construct yielded *3E*- and *3Z*-isomers of PEB, and the *pcyA* construct yielded *3E*- and *3Z*-isomers of PCB (Figure S2). Thus, expressed *ho1*_P-SSM2 and *pebS* can transform endogenous *E. coli* heme to PEB, and expressed *ho1*_P-SSM2 and *pcyA*_P-SSM4 can transform endogenous heme to PCB, in each case using only endogenous electron donors.

Finally, we wondered whether bilins produced by the cyanophage-encoded enzymes could be incorporated in phytochromes with the expected spectroscopic signatures. Here, the bilin biosynthesis vectors were used in coexpression experiments with bacterial BV-binding and cyanobacterial PCB- and PEB-binding apo-phytochromes, as previously described [23, 24]. Coexpression of pTD*ho1* with pASK_*bphP*, encoding the bacterial phytochrome of *Pseudomonas aeruginosa*, yielded a functional holophytochrome; red- and far-red-light difference spectroscopy revealed a typical phytochrome signature [25] (Figure S3). A similar result was obtained by coexpression of pTD*ho1pcyA* with the PCB-binding cyanobacterial phytochrome Cph1 of *Synechocystis* sp. PCC6803 (Figure S3). Notably, coexpression of pTD*ho1pebS* with *cph1* yields a highly fluorescent Cph1-PEB adduct (phytofluor)

Table 1. The Presence or Absence of Bilin Biosynthesis Genes in the Genomes of Fully Sequenced Cultured Marine Cyanobacteria and Cyanophages

Taxon	Strain	*pebS*	*pebA*	*pebB*	*pcyA*	*ho1*	*petF*	*cpeT*
Prochlorococcus podovirus	P-SSP7	–	–	–	–	–	–	–
Synechococcus podoviruses	Syn5	–	–	–	–	–	–	–
Synechococcus podoviruses	P60	–	–	–	–	–	–	–
Prochlorococcus myoviruses	P-SSM2	+	–	–	–	+	+	–
Prochlorococcus myoviruses	P-SSM4	–	–	–	+	–	–	–
Synechococcus myoviruses	S-PM2	–	–	–	–	–	–	+
Synechococcus myoviruses	Syn9	–	–	–	–	–	–	+
HL *Prochlorococcus*	6 strains[a]	–	+	+	+	+	+	–
LL *Prochlorococcus*	6 strains[b]	–	+	+	+	+	+	+
Synechococcus	6 strains[c]	–	+	+	+	+	+	+

+ indicates that the gene is present in the genome, and – indicates that the gene has not been identified in the genome. Each member of three cyanobacterial host cell groups (high-light [HL] adapted *Prochlorococcus*, low-light [LL] adapted *Prochlorococcus* and *Synechococcus*) contained the same bilin biosynthesis genes under investigation here; these data are grouped together under the appropriate taxon heading. Host genomes used here are strains as available as of October 15, 2007 at Microbes Online (http://www.microbesonline.org/).

[a] HL *Prochlorococcus* strains: MED4, MIT9215, MIT9301, MIT9312, MIT9515, and AS9601.

[b] LL *Prochlorococcus* strains: NATL1A, NATL2A, SS120, MIT9211, MIT9303, and MIT 9313.

[c] *Synechococcus* strains: CC9311, CC9605, CC9902, WH8102, JA-2-3B'a(2-13), and JA-3-3Ab.

[26], a useful biotechnology tool (Figure S4). Taken together, these experimental findings demonstrate that cyanophage enzymes are sufficient to efficiently produce the predicted bilin metabolites in a heterologous background.

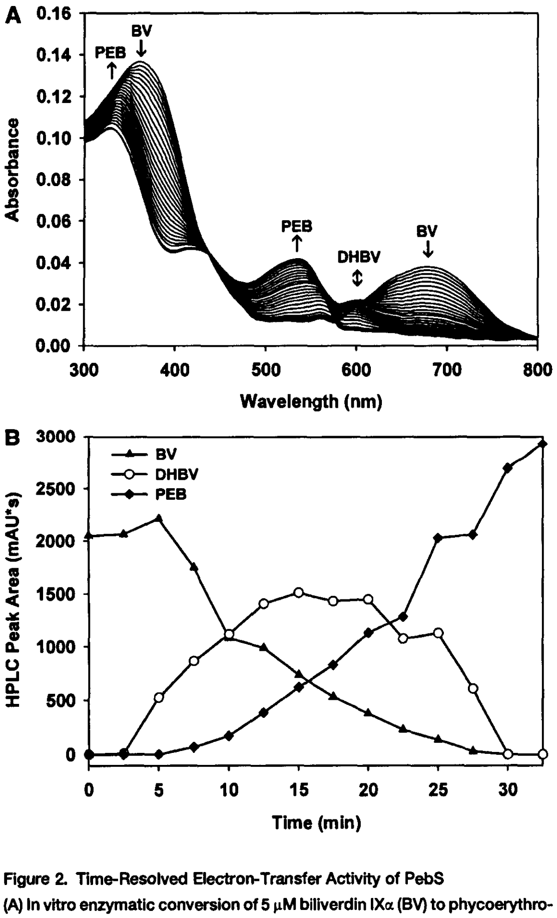


Figure 2. Time-Resolved Electron-Transfer Activity of PebS

(A) In vitro enzymatic conversion of 5 μM biliverdin IXα (BV) to phycoerythrobilin (PEB) was monitored by absorbance spectroscopy at 75 s intervals. Substrate-specific absorbance at 380 and 690 nm decreases, whereas the product-specific absorbance at 330 and 540 nm increases. Absorbance changes around 600 nm correspond to the appearance and disappearance of a partially reduced intermediate.

(B) Samples of the reaction were collected at 150 s intervals and subjected to HPLC analyses. HPLC peaks were identified by comparison to known standards. The integrated HPLC peak areas were plotted against sampling time. Triangles indicate BV IXα, circles indicate 15,16-dihydrobiliverdin (15,16-DHBV), and squares indicate *3E*- and *3Z*-PEB. The occurrence of 15,16-DHBV fits the approximately 600 nm absorbance change observed in (A), identifying 15,16-DHBV as the semireduced intermediate in this reaction. Detection wavelengths were 650 nm for BV IXα and 560 nm for 15,16-DHBV and PEB. Values for PEB correspond to the sum of both *3E*- and *3Z*-isomers.

Cyanophage-Encoded Bilin Biosynthesis Genes Are Expressed during Infection

Having shown that the *pebS*, *petF*, and *ho1* genes carried by the phage P-SSM2 encode proteins functional in vitro, we tested whether these three genes are expressed during infection of the low light (LL) *Prochlorococcus* strain NATL1A by P-SSM2. We used quantitative reverse transcriptase-polymerase chain reaction (qRT-PCR) to detect expression of *pebS*, *petF*_P-SSM2, and *ho1*_P-SSM2 at the messenger RNA (mRNA) level in infected and control cultures. In all infected cultures, all three phage transcripts were unambiguously detected as mRNA by 1 hr (Figure 3) after infection began ($p < 0.000003$ for each biological sample, binomial probability distribution; see the Supplemental Data). The early induction of these genes is consistent with the timing of phage gene induction of the core photosynthetic reaction center gene, *psbA*, in a *Prochlorococcus* podovirus phage-host system [27]. In light of the functionality of the recombinant proteins, the fact that all three genes in P-SSM2's PEB biosynthesis pathway are expressed during infection suggests that, rather than being excess genomic baggage en route to degradation, these genes likely play a functional role in this phage-host interaction.

Bilin Biosynthesis Genes in Host and Phage Genomes from Oceanic Communities

In the seven complete cultured cyanophage genomes, the *pebS* gene occurs only once, in P-SSM2 [13–17]. Furthermore, it is absent from the 18 *Prochlorococcus* and *Synechococcus* host genomes sequenced to date (S.C.B., M.B.S., and S.W.C., unpublished data, with data from [28] and unpublished

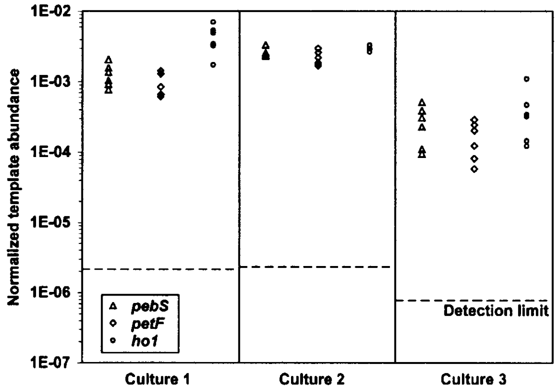


Figure 3. Bilin Biosynthesis Gene Expression during Infection of *Prochlorococcus* NATL1A by P-SSM2

Quantitative reverse transcriptase-polymerase chain reaction was used for the measurement of levels of cyanophage-encoded *pebS* (triangles), *petF* (diamonds), and *ho1* (squares) mRNA relative to host-encoded *mpB* mRNA 1 hr after infection of triplicate NATL1A cultures. Dotted lines indicate the theoretical detection limit (i.e., the starting abundance of template that would allow a reaction to reach the threshold fluorescence during the final cycle). At t = 1 hr, all P-SSM2 genes investigated were expressed as mRNA in all infected cultures, whereas the phage genes were undetectable in both the uninfected cultures (biological negative control; data not shown) and 17 of 18 replicates of *pebS* primed reactions with reverse transcriptase omitted (technical negative control; the 18th technical replicate, with template abundance 5.3 e-4, is taken to be an outlier as described in the Supplemental Data).

genomes available at http://www.microbesonline.org/). Conversely, *pebA* (like its pathway partner *pebB*) is found in all of these host genomes and none of the cyanophages (Table 1), suggesting that the canonical *pebA* and *pebB* pathway plays an essential role in *Prochlorococcus* and *Synechococcus* but that this pathway may be absent in their phages. To see whether the exclusive association of *pebS* with phage genomes and *pebA* with host genomes persists in the wild, we examined the Global Ocean Survey (GOS) metagenomics dataset [29–31], which consists of DNA fragments extracted from 0.1–0.8 μm diameter particles in surface waters of the Atlantic and Pacific Oceans and from specialized aquatic environments (e.g., lakes, hypersaline ponds, estuaries).

After curation of low-stringency recruits and length normalization (see the Supplemental Data and Table 2), PebA and PebS protein-sequence queries recruited 164 and 137 DNA fragments, respectively, from the GOS database (Table 2). The recruited read and the reverse sequence read (paired end) from each piece of cloned DNA were then used as BLAST queries for the comparison of the taxonomic assignments of the two queries' top hits. The taxonomic affiliation of the recruited and paired-end reads' best hits were strikingly similar (Figure 4), allowing us to classify each environmental sequence as originating from *Synechococcus* PebA (31 sequences), high light (HL)-adapted *Prochlorococcus* PebA (133 sequences), or phage PebS (137 sequences). We then examined the phylogenetic clustering of these PebA and PebS homologs (Figure 4). All of the recruited reads inferred by BLAST to be *Prochlorococcus* or *Synechococcus* clustered together with PebA from HL *Prochlorococcus* and *Synechococcus* isolates, respectively. Likewise, recruited reads already inferred by BLAST to be cyanophage clustered together with PebS from P-SSM2. Thus, it appears that environmental PebA-like sequences are found only in cyanobacterial host cells, whereas PebS-like sequences are found only in phages. Notably, the 137 "cyanophage" metagenomic PebS sequences originated from 24 sampling sites (Table 2, Figure 4) in varied environments, suggesting that this "viral" PebS bilin biosynthesis strategy is not geographically restricted.

We next examined the environmental distribution of other bilin pathway proteins, PcyA, Ho1, and PebB (PetF was not examined; see the Supplemental Data). We identified 159 PcyA, 264 Ho1, and 159 PebB homologs (Table 2). The origins of these PcyA and Ho1 sequences, as inferred from BLAST and clustering (Figures S5 and S6), are split between cyanobacteria and phage, with 46 PcyA and 91 Ho1 sequences of cyanophage origin. In contrast, BLAST analyses of the recruited and paired-end reads argue that the recruited PebB DNA fragments likely originated solely from cyanobacteria (Table 2;

Table 2. Occurrence of Phage and Host Bilin Biosynthesis Gene Homologs in the Global Ocean Survey Metagenomic Dataset

		Bit Score > 100, Any Size	Bit Score > 100 *and* Over 140 Amino Acids in Size				
Query Protein	Query Size (Amino Acids)	Total Recruited Sequence Reads	Total Recruited Sequence Reads	Number[a] of Putative Pro Sequences	Number[a] of Putative Syn Sequences	Number[a] of Putative Phage Sequences	Number of GOS Sites with Phage Sequences
PebA	234	166 (236)	115 (164)	93 (133)	22 (31)	0 (0)	0
PebB[b]	257	167 (216)	123 (159)	108 (140)	15 (19)	0 (0)	0
PebS	234	127 (181)	96 (137)	0 (0)	0 (0)	96 (137)	24
PcyA	237	124 (174)	113 (159)	58 (82)	22 (31)	33 (46)	15
Ho1	242	232 (319)	192[c] (264)	94 (130)	28 (39)	66 (91)	18

The "total" columns represent the sequence reads with bit score similarities greater than 100 to the query sequence, of any size or restricted to over 140 amino acids length as indicated. To allow for cross-query comparisons, we normalized the number of recruits to query size as follows: The normalized number of recruited reads is equal to the number of recruited reads divided by the protein query size times an average protein size of 333 amino acids; these data are presented in parentheses. These tabulated data represent the summary of inferences made with a two-tiered approach for the inference of the origin of each environmental DNA fragment: phylogenetic clustering and paired-end analysis (see the Supplemental Data, Figure 4, Figures S5 and S6, and Table S1).

[a] Taxon assignments were determined by BLAST similarity of the recruited and paired-end sequences, as well as phylogenetic clustering of the recruited sequence read (see the Supplemental Data, Figure 4, and Figures S5 and S6).

[b] No tree is presented for PebB because there are no phage sequences available to make phylogenetic clustering inferences meaningful—tabulated data are presented in Table S1.

[c] For Ho1, of the 192 recruited sequence reads, the original organism for four DNA fragments could not be confidently inferred with either phylogenetic clustering or paired-end analysis.

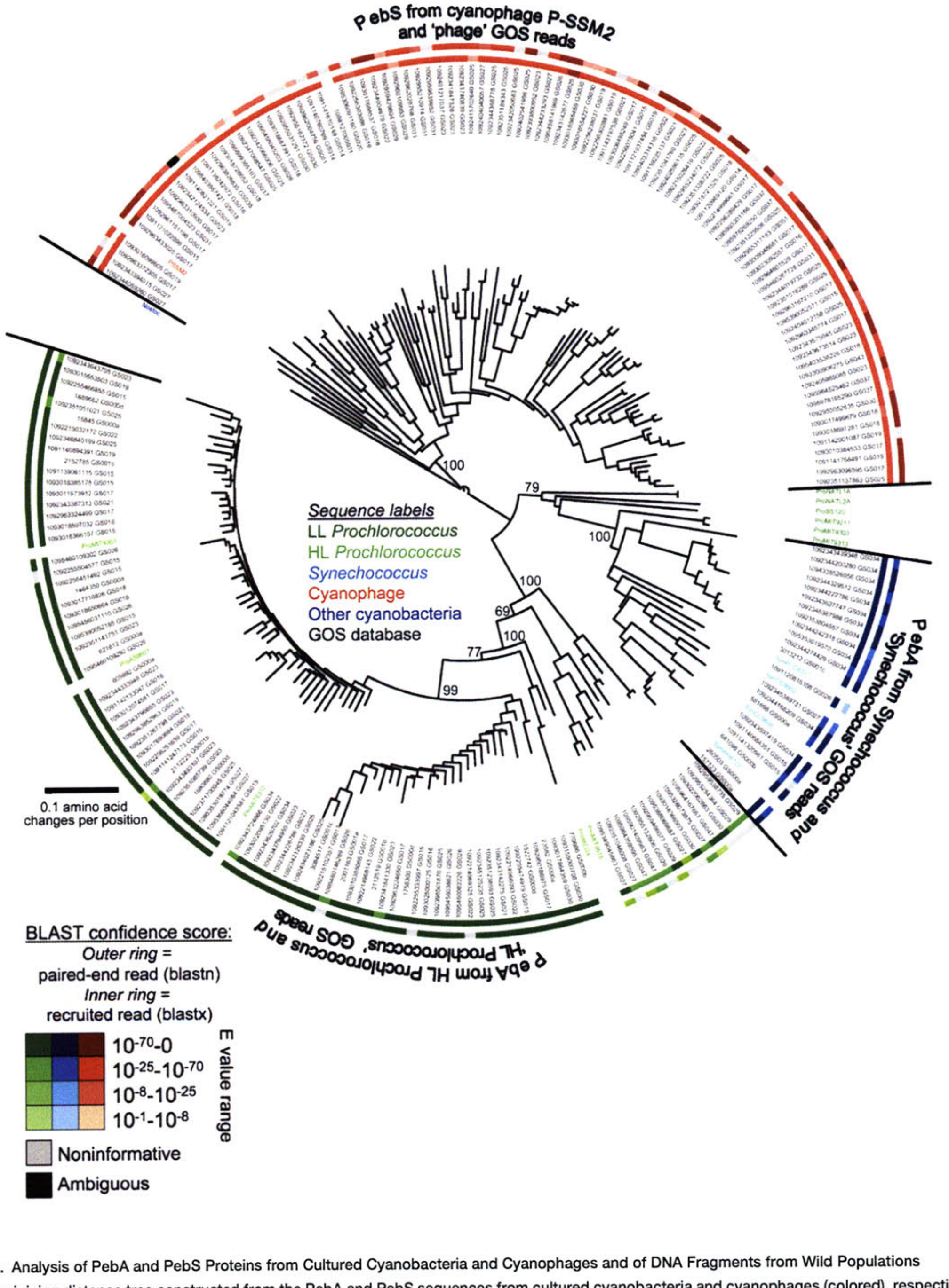

PebS from cyanophage P-SSM2 and 'phage' GOS reads
PebA from Synechococcus and 'Synechococcus' GOS reads
PebA from HL Prochlorococcus and 'HL Prochlorococcus' GOS reads
100
79
100
100
69
100
77
99
Sequence labels
LL Prochlorococcus
HL Prochlorococcus
Synechococcus
Cyanophage
Other cyanobacteria
GOS database
0.1 amino acid changes per position
BLAST confidence score:
Outer ring = paired-end read (blastn)
Inner ring = recruited read (blastx)
10^{-70}-0
10^{-25}-10^{-70}
10^{-8}-10^{-25}
10^{-1}-10^{-8}
E value range
Noninformative
Ambiguous

Figure 4. Analysis of PebA and PebS Proteins from Cultured Cyanobacteria and Cyanophages and of DNA Fragments from Wild Populations

Neighbor-joining distance tree constructed from the PebA and PebS sequences from cultured cyanobacteria and cyanophages (colored), respectively, as well as sequences recruited from the Global Ocean Survey (GOS) database (black) [29–31]. Statistical (neighbor-joining bootstraps) support is displayed only at critical nodes of the tree that help delineate cyanophage, *Prochlorococcus*, and *Synechococcus* sequence lineages. For environmental sequences, the read number is displayed, as well as the GOS site number (GS###) from which the DNA was originally obtained. The BLAST-based taxonomic assignments (see the Supplemental Data) are presented for each recruited and paired-end sequence from the wild DNA fragments as two color rings along the outside of the tree. The colors represent organismal identifications made from the taxon label of the top BLAST alignment of the recruited or paired-end reads against the GenBank nucleotide (NT) database; the intensity of the color reflects the confidence score of the BLAST hit. BLAST results were considered noninformative if there were no hits better than e-value 1e-1 from either cyanobacteria or cyanophages or, in the case of the paired-end reads, if the blast alignment was only to the query gene because the recruited read was already counted for that gene. BLAST results were considered ambiguous if the

Table S1). These findings suggest that cyanophages exploiting PEB synthesis may exclusively utilize the cyanophage PebS pathway rather than the "cyanobacterial" PebA and PebB pathway.

Phylogenetic analyses shed light on the evolutionary history of these genes. The monophyletic clusters observed for "phage" copies of these genes suggested that PebA and PebS and Ho1 were obtained from marine cyanobacteria only once, whereas PcyA phage sequences group into two clusters (Figure 4, Figures S5 and S6) and may have been obtained twice. In comparison, the genes encoding the core photosystem II reaction center proteins, PsbA and PsbD, are thought to have been obtained repeatedly (four and two times, respectively) by cyanophages from their cyanobacterial hosts [32].

Finally, although we recognize that the phage signal in these cell-fraction metagenomic data are predominantly intracellular phages unlikely to represent the entire free-phage community, we estimated the fraction of T4-like phages in the metagenome that contain phage bilin biosynthesis genes. By using normalized recruit frequencies that account for variable gene sizes, we observed 137 phage PebS, 46 phage PcyA, and 91 phage Ho1 DNA fragments (Table 2), as compared to 1018 DNA fragments (Document S2) recruited for the T4-like portal protein-encoding gene (gene 20), which is universal among T4-like phages, one of the most common phage types observed in marine metagenomes to date [33, 34]. Thus, as much as 13%, 5%, and 9% of the T4-like phages captured in these samples contained PebS, PcyA, and Ho1, respectively, suggesting that these biosynthetic pathways are an important component of wild T4-like phage populations. Notably, one *pcyA* sequence resides on a 12.7 kb phage genomic fragment (JCVI_SCAF_1096626959277) that also includes a *pebS* homolog; thus, some phage genomes may be capable of both PEB and PCB biosynthetic pathways.

Conclusions

We have demonstrated that the cyanophage protein PebS has acquired a novel activity, combining the functions of two separate enzymes in the host cell, PebA and PebB. Further, all evidence suggests that cyanophages that maintain bilin biosynthesis activity have replaced the canonical PebA and PebB pathway found in host cells with the single PebS enzyme. We hypothesize that the maintenance of two individual genes might allow for tighter regulation (e.g., feedback inhibition) in the cell or a metabolic branchpoint upstream of an alternative, beneficial function for 15,16-DHBV. Conversely, phage fitness may be influenced more by short-term efficiency rather than long-term flexibility; a one-enzyme system would allow a cyanophage to channel all PebS-bound BV toward a single metabolic fate. Further, phage genomes are likely under tighter size selection than microbial genomes because of headful packaging; again, the single enzyme would prove advantageous because PebS requires less than half of the genetic material as the PebA and PebB system.

More broadly, the auxiliary metabolic genes in oceanic viral genomes presumably reveal the evolutionary swapping of metabolic components critical to phage and host reproduction. The importance of these components to cyanophages is not always clear. In the case of the bilin biosynthesis genes studied here, their role is not even well understood in the host cells from which the phage genes were derived. Nonetheless, we find that cyanophage-encoded bilin biosynthesis genes are functional, that they are expressed during infection, and that they are represented in wild populations of phage. Moreover, PebS has a novel activity and a sequence found only in cyanophages. Together, these data contribute to our growing understanding of cyanophages as a laboratory for metabolic innovation.

Experimental Procedures

Standard methods for manipulation of nucleic acids and purification of recombinant proteins were used throughout. Full experimental procedures and associated references are in the Supplemental Data available online.

Supplemental Data

Experimental Procedures, six figures, one dataset, and one table are available at http://www.current-biology.com/cgi/content/full/18/6/442/DC1/.

Acknowledgments

We thank J.C. Lagarias for the gift of a Cph1 expression vector, F. Narberhaus, M. Osburne, J. Bragg, L. Thompson, M. Coleman, and J. Waldbauer for critically reading the manuscript, S. Kern and K. Huang for technical assistance, and D. Lindell and M. Coleman for helpful discussions. Financial support from the Deutsche Forschungsgemeinschaft, the Sonderforschungsbereich 480 (Teilprojekt C8), the Fonds der chemischen Industrie (to N.F.D.), and National Science Foundation (this is a contribution of C-MORE [NSF]), Department of Energy, Massachusetts Institute of Technology Center of Environmental Health Sciences, and the Gordon and Betty Moore Foundation (to S.W.C.) is gratefully acknowledged. S.C.B. is a Howard Hughes Medical Institute Predoctoral Fellow.

Received: December 18, 2007
Revised: February 18, 2008
Accepted: February 20, 2008
Published online: March 20, 2008

References

1. Goericke, R., and Repeta, D.J. (1992). The pigments of *Prochlorococcus marinus*: The presence of divinyl chlorophyll a and b in a marine procaryote. Limnol. Oceanogr. *37*, 425–433.
2. Lichtle, C., Thomas, J.C., Spilar, A., and Partensky, F. (1995). Immunological and ultrastructural characterization of the photosynthetic complexes of the prochlorophyte *Prochlorococcus* (oxychlorobacteria)1. J. Phycol. *31*, 934–941.
3. La Roche, J., Van Der Staay, G.W.M., Partensky, F., Ducret, A., Aebersold, R., Li, R., Golden, S.S., Hiller, R.G., Wrench, P.M., Larkum, A.W.D., and Green, B.R. (1996). Independent evolution of the prochlorophyte and green plant chlorophyll a/b light-harvesting proteins. Proc. Natl. Acad. Sci. USA *93*, 15244–15248.
4. Hess, W.R., Partensky, F., Van Der Staay, G.W.M., Garcia-Fernandez, J.M., Borner, T., and Vaulot, D. (1996). Coexistence of phycoerythrin and a chlorophyll a/b antenna in a marine prokaryote. Proc. Natl. Acad. Sci. USA *93*, 11126–11130.
5. Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc. Natl. Acad. Sci. USA *101*, 11013–11018.
6. Frankenberg, N., Mukougawa, K., Kohchi, T., and Lagarias, J.C. (2001). Functional genomic analysis of the HY2 family of ferredoxin-dependent bilin reductases from oxygenic photosynthetic organisms. Plant Cell *13*, 965–978.

top hits included both cyanobacteria and cyanophage. Sequences from *Prochlorococcus* MIT9303 and MIT9313 often cluster with homologs from *Synechococcus* [28], as observed here. Together these analyses identify the environmental PebA and PebS sequences as having originated either from cyanobacteria or cyanophages, respectively.

7. Dammeyer, T., Michaelsen, K., and Frankenberg-Dinkel, N. (2007). Biosynthesis of open-chain tetrapyrroles in *Prochlorococcus marinus*. FEMS Microbiol. Lett. *271*, 251–257.
8. Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003). Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. Nature *424*, 1047–1051.
9. Hess, W.R., Steglich, C., Lichtle, C., and Partensky, F. (1999). Phycoerythrins of the oxyphotobacterium *Prochlorococcus marinus* are associated to the thylakoid membrane and are encoded by a single large gene cluster. Plant Mol. Biol. *40*, 507–521.
10. Steglich, C., Behrenfeld, M., Koblizek, M., Claustre, H., Penno, S., Prasil, O., Partensky, F., and Hess, W.R. (2001). Nitrogen deprivation strongly affects photosystem II but not phycoerythrin level in the divinyl-chlorophyll b-containing cyanobacterium *Prochlorococcus marinus*. Biochim. Biophys. Acta *1503*, 341–349.
11. Steglich, C., Frankenberg-Dinkel, N., Penno, S., and Hess, W.R. (2005). A green light-absorbing phycoerythrin is present in the high-light-adapted marine cyanobacterium *Prochlorococcus* sp. MED4. Environ. Microbiol. *7*, 1611–1618.
12. Steglich, C., Mullineaux, C.W., Teuchner, K., Hess, W.R., and Lokstein, H. (2003). Photophysical properties of *Prochlorococcus marinus* SS120 divinyl chlorophylls and phycoerythrin in vitro and in vivo. FEBS Lett. *553*, 79–84.
13. Chen, F., and Lu, J. (2002). Genomic sequence and evolution of marine cyanophage P60: A new insight on lytic and lysogenic phages. Appl. Environ. Microbiol. *68*, 2589–2594.
14. Sullivan, M.B., Coleman, M.L., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005). Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. PLoS Biol. *3*, e144.
15. Pope, W.H., Weigele, P.R., Chang, J., Pedulla, M.L., Ford, M.E., Houtz, J.M., Jiang, W., Chiu, W., Hatfull, G.F., Hendrix, R.W., and King, J. (2007). Genome sequence, structural proteins, and capsid organization of the cyanophage syn5: A "horned" bacteriophage of marine synechococcus. J. Mol. Biol. *368*, 966–981.
16. Mann, N.H., Clokie, M.R.J., Millard, A., Cook, A., Wilson, W.H., Wheatley, P.J., Letarov, A., and Krisch, H.M. (2005). The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus* strains. J. Bacteriol. *187*, 3188–3200.
17. Weigele, P.R., Pope, W.H., Pedulla, M.L., Houtz, J.M., Smith, A.L., Conway, J.F., King, J., Hatfull, G.F., Lawrence, J.G., and Hendrix, R.W. (2007). Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. Environ. Microbiol. *9*, 1675–1695.
18. Fukuyama, K. (2004). Structure and function of plant-type ferredoxins. Photosynth. Res. *81*, 289–301.
19. McDowell, M.T., and Lagarias, J.C. (2001). Purification and biochemical properties of phytochromobilin synthase from etiolated oat seedlings. Plant Physiol. *126*, 1546–1554.
20. Frankenberg, N., and Lagarias, J.C. (2003). Phycocyanobilin:ferredoxin oxidoreductase of *Anabaena* sp. PCC 7120. Biochemical and spectroscopic. J. Biol. Chem. *278*, 9219–9226.
21. Beale, S.I., and Cornejo, J. (1991). Biosynthesis of phycobilins. 15,16-dihydrobiliverdin IXα is a partially reduced intermediate in the formation of phycobilins from biliverdin IXα. J. Biol. Chem. *266*, 22341–22345.
22. Dammeyer, T., and Frankenberg-Dinkel, N. (2006). Insights into phycoerythrobilin biosynthesis point toward metabolic channeling. J. Biol. Chem. *281*, 27081–27089.
23. Gambetta, G.A., and Lagarias, J.C. (2001). Genetic engineering of phytochrome biosynthesis in bacteria. Proc. Natl. Acad. Sci. USA *98*, 10566–10571.
24. Mukougawa, K., Kanamoto, H., Kobayashi, T., Yokota, A., and Kohchi, T. (2006). Metabolic engineering to produce phytochromes with phytochromobilin, phycocyanobilin, or phycoerythrobilin chromophore in *Escherichia coli*. FEBS Lett. *580*, 1333–1338.
25. Tasler, R., Moises, T., and Frankenberg-Dinkel, N. (2005). Biochemical and spectroscopic characterization of the bacterial phytochrome of *Pseudomonas aeruginosa*. FEBS J. *272*, 1927–1936.
26. Murphy, J.T., and Lagarias, J.C. (1997). The phytofluors: A new class of fluorescent protein probes. Curr. Biol. *7*, 870–876.
27. Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. Nature *438*, 86–89.
28. Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., et al. (2007). Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. PLoS Genet. *3*, e231.
29. Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., et al. (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. PLoS Biol. *5*, e77.
30. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science *304*, 66–74.
31. Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W., et al. (2007). The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. PLoS Biol. *5*, e16.
32. Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol. *4*, e234.
33. DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U., Martinez, A., Sullivan, M.B., Edwards, R., Brito, B.R., et al. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. Science *311*, 496–503.
34. Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., Kelley, S., Liu, H., et al. (2006). The marine viromes of four oceanic regions. PLoS Biol. *4*, e368.

Appendix 3

MG Klein, P Zwart, SC Bagby, F Cai, SW Chisholm, S Heinhorst, GC Cannon, and C Kerfeld (2009). Identification and structural analysis of a novel carboxysome shell protein with implications for metabolite transport. *J Mol Biol* in press.

Identification and Structural Analysis of a Novel Carboxysome Shell Protein with Implications for Metabolite Transport

Michael G. Klein,[1] Peter Zwart,[2] Sarah C. Bagby,[3] Fei Cai,[4] Sallie W. Chisholm,[3] Sabine Heinhorst,[4] Gordon C. Cannon[4] and Cheryl A. Kerfeld[1,5]

[1]US Department of Energy - Joint Genome Institute, Walnut Creek, CA 94598

[2]US Department of Energy – Lawrence Berkeley National Laboratory, Berkeley CA 94720 USA

[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

[4]Department of Chemistry and Biochemistry, The University of Southern Mississippi, Hattiesburg, MS 39406 USA

[5]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720 USA

Corresponding author: ckereld@lbl.gov Phone: (925) 296-5691 FAX: (925) 296-5752

Running Title: Structure of a Novel Carboxysome Shell Protein

Summary

Bacterial microcompartments are polyhedral bodies composed entirely of protein that function as organelles in bacteria; they promote subcellular processes by encapsulating and co-localizing targeted enzymes with their substrates. The best-characterized bacterial microcompartment is the carboxysome, a central part of the carbon concentrating mechanism that greatly enhances carbon fixation in cyanobacteria and some chemoautotrophs. Here we report the first structural insights into the carboxysome of *Prochlorococcus*, the numerically dominant cyanobacterium in the world's oligotrophic oceans. Bioinformatic methods, substantiated by analysis of gene expression data, were used to identify a new carboxysome shell component, CsoS1D, in the genome of *Prochlorococcus* strain MED4; orthologs were subsequently found in all cyanobacteria. Two independent crystal structures of *Prochlorococcus* MED4 CsoS1D reveal three features not seen in any BMC-domain protein structure solved to date. First, CsoS1D is composed of a fused pair of bacterial microcompartment domains. Second, this double-domain protein trimerizes to form a novel pseudohexameric building block for incorporation into the carboxysome shell, and the trimers further dimerize, forming a two-tiered shell building block. Third, and most strikingly, the large pore formed at the 3-fold axis of symmetry appears to be gated. Each dimer of trimers contains one trimer with an open pore and one whose pore is obstructed due to sidechain conformations of two residues that are invariant among all CsoS1D orthologs. This is the first evidence of the potential for gated transport across the carboxysome shell and reveals a new type of building block for bacterial microcompartment shells.

Keywords: Carboxysome, bacterial microcompartment, gated transport, *Prochlorococcus*, carbon concentrating mechanism

Introduction

The cyanobacterium *Prochlorococcus* is the numerically dominant photosynthetic organism in much of the world's oceans. As such it is responsible for a significant fraction of biological carbon fixation in these systems. Like all cyanobacteria, *Prochlorococcus* is presumed to rely on a carbon concentrating mechanism to facilitate CO_2 capture and fixation [1; 2; 3]. Central to this mechanism is a microcompartment in the cell, the carboxysome (Figure 1A) which is the site of CO_2 concentration and fixation.

Carboxysomes are the foremost example of the polyhedral subcellular inclusions that have been termed bacterial microcompartments (Figure 1A), self-assembling protein shells that encapsulate enzymes and other functionally related proteins. In addition to carboxysomes, two other types of bacterial microcompartments are relatively well characterized (reviewed in [4; 5; 6]); they function in propane-diol utilization (encoded by the pdu operon) and ethanolamine utilization (encoded by the eut operon) in heterotrophic bacteria.

Carboxysomes have been observed in all cyanobacteria and many chemoautotrophs. They contain the key enzyme of carbon fixation, ribulose1,5-bisphosphate carboxylase/oxygenase (RubisCO) [7], which has a relatively high K_M and as such is not a kinetically efficient enzyme. Its efficiency is further compromised because it catalyzes the apparently unproductive fixation of O_2 in competition with CO_2. Cyanobacteria compensate for these deficiencies by encapsulating RubisCO within the carboxysome. In *Prochlorococcus*, apparently all of the cellular RubisCO is located in the carboxysomes [8]; this is likely to be particularly important for this genus as the K_{CO2} value of its RubisCO (750 umol L^{-1}; [3]) is the highest known among cyanobacterial RubisCOs. Carbonic anhydrase (CA) activity, which converts the abundant intracellular bicarbonate to CO_2 is tightly associated with the carboxysome shell [9]. RubisCO derives a catalytic advantage not only from its proximity to CA activity in the carboxysome but also because the carboxysome shell impedes diffusive loss of HCO_3^- and CO_2 [10; 11].

There are two types of carboxysome, alpha and beta, distinguished by the form of RubisCO they encapsulate. Alpha (or cso-type) carboxysomes, containing Form 1A RubisCO, are found in the dominant oceanic picocyanobacteria, *Prochlorococcus* and *Synechococcus*, and in chemoautotrophs, such as *Halothiobacillus neapolitanus*; beta carboxysomes contain Form 1B RubisCO and are broadly distributed among cyanobacteria, occurring in *Trichodesmium, Thermosynechococcus, Synechocystis, Gloeobacter*, and *Crocosphaera*, and others [12; 13]. The two types of carboxysomes have very different genomic architecture; the alpha carboxysome is encoded by an operon (Figure 1B) that includes genes for the large and small subunits of RubisCO, the carbonic anhydrase CsoSCA (formerly CsoS3), CsoS2, a protein of unknown function and two small shell proteins that form the vertices of the carboxysome [14]. In contrast, the components of the beta carboxysome are encoded by several small, scattered gene clusters (the *ccm* genes;[4; 12; 15]).

This disparity at the operon level belies the similarity of the proteins that constitute alpha and beta carboxysome shells. The most common structural motif in carboxysome shell proteins is the bacterial microcompartment (BMC) domain (Pfam 00936), which occurs in multiple paralogous proteins in both alpha and beta carboxysomes. Among these paralogs, CcmK1, 2 and 4 from *Synechocystis* sp PCC 6803 and CsoS1A from *H. neapolitanus* have been structurally characterized [14; 16; 17], revealing that they form cyclic hexamers. Pores formed at the six-fold symmetry axis of individual hexamers are thought to permit metabolite flow across the carboxysome shell. Hexamers of CcmK1, CcmK2 and CsoS1A tend to tile into uniformly oriented layers that presumably constitute the facets of the carboxysome shell.

Interestingly, the alpha carboxysome operons of cyanobacteria and chemoautotrophs differ in the number and position of BMC domains each encodes (Figure 1B). Whereas all chemoautotrophs contain 2-3 highly homologous single-BMC polypeptides (CsoS1A, B and C), at the end of the operon, in the *Prochlorococcus* species these are absent; instead there is a single, ~100 amino acid BMC protein gene (*csoS1*) preceding the gene for the large subunit of RubisCO. Not only the difference in position, but the difference in number of BMC domain-containing proteins in the genome is notable, for in all other organisms expected to form carboxysomes or related bacterial microcompartments, there are always at least two BMC domains present (Kerfeld et al., unpublished); the reason for this redundancy is at present, unknown. However, closer inspection of the high light-adapted *Prochlorococcus* genomes reveals that this shortfall is superficial. Upstream of *csoS1* and separated from it by a gene encoded on the opposite strand, there is a gene for a 256 amino acid protein predicted to contain a BMC domain in its C-terminus. Both the sequence of this protein (PMM0547 in *Prochlorococcus* MED4; Figure 1B) and its genomic position relative to the other carboxysome genes are conserved in all sequenced *Prochlorococcus* strains, and indeed in all sequenced alpha carboxysome-containing cyanobacteria.

Here we report two independent crystal structures of PMM0547 at 2.2 and 2.3Å resolution, and a model for its interaction with CsoS1 in the *Prochlorococcus* carboxysome shell. The structural models and several complementary lines of evidence suggest that PMM0547 is a carboxysome shell protein, therefore we have provisionally dubbed it CsoS1D.

Results

Identification of a New Cyanobacterial Carboxysome Shell Protein

The gene for CsoS1D was detected by scanning the *Prochlorococcus* MED4 genome for bacterial microcompartment (BMC) domains. Subsequently, a search of the Pfam database with the *Prochlorococcus* MED4 CsoS1D amino acid sequence detected a single BMC domain in residues 190-256 and no other recognizeable domains. Putative orthologs (defined as bidirectional best BLASTP hits) to CsoS1D are found in all cyanobacteria, suggesting that this protein could be a part of both alpha and beta carboxysome shells. The primary structure is highly conserved (~80% identical) among the cyanobacteria that form alpha carboxysomes. Orthologs found in the genomes of cyanobacteria that form beta carboxysomes, such as the product of *slr0169* in *Synechocystis* sp PCC6803, lack the first 50 residues and are ~ 45% identical to residues 52-256 of CsoS1D. In addition, orthologs to CsoS1D are also found in several other organisms that do not form carboxysomes, but based on the presence of BMC-domain containing proteins, do appear to have the potential to form bacterial microcompartments of unknown function.

The alpha carboxysome genes are arranged similarly in the 22 genomes of marine *Synechococcus* and *Prochlorococcus* species sequenced to date, with *csoS1D* upstream of and separated from the carboxysome operon by one gene that is transcribed from the opposite strand (Figure 1B). The gene that separates *csoS1D* from the cluster of carboxysome genes, annotated as encoding a Ham1 protein (PMM0548 in *Prochlorococcus* MED4), is widespread among bacterial genomes but is of unknown function. To determine the likelihood that *csoS1D* is a component of the *Prochlorococcus* carboxysome, we analyzed the RNA expression pattern of *csoS1D* and its neighbors in axenic *Prochlorococcus* MED4 cells over the diel cycle, using data collected for a study of the entire *Prochlorococcus* transcriptome in synchronously dividing cells entrained to a 24-h light-dark cycle (Zinser *et al.*, manuscript submitted). Like the genes in the carboxysome operon (*csoS1-csoS4B*, PMM0549-PMM0553; Figure 1B), *csoS1D* (PMM0547) is expressed with 24-h periodicity with peak expression occurring at sunrise. We subjected all expressed, cycling genes (a total of 1405; Zinser *et al.*, manuscript submitted) to "fuzzy" clustering to examine more closely the association of *csoS1D* expression with expression of *csoS1-csoSCA* (PMM0549-PMM0553), known carboxysome-related genes. Unlike "hard" clustering, which assigns each gene to exactly one cluster, "fuzzy" clustering determines the fractional membership of each gene in each cluster, permitting relationships between clusters that share genes to be recognized [18]. Following the model of Zinser et al., (manuscript submitted), our initial analysis, used 16 clusters and "fuzzification" parameter $m = 1.25$. In 100 out of 100 runs, *csoS1, cbbL, cbbS,* and *csoS2* clustered together, each gene with very high fractional membership in this cluster (mean $\mu = 0.991 \pm 0.004$, 0.990 ± 0.006, 0.957 ± 0.031, 0.913 ± 0.058, respectively). To check whether the association of *csoS1D* with the carboxysome-containing cluster was robust, we repeated the clustering analysis with m ranging from 1.15 ("harder") to 1.35 ("fuzzier") and the number of clusters ranging from 10 to 18, performing 100 runs with each pair of input parameters. Cluster stability began to diminish at $m \geq 1.30$ (data not shown); nonetheless, across the entire range, *csoS1D* was more strongly clustered with *csoS1-csoS2* than with genes in any other cluster (Figure S1). Indeed, *csoS1D* was more stably clustered with *csoS1-csoS2* than were the known carboxysome components *csoSCA* and *csoS4A* and *B* (formerly known as *orfA* and *orfB*); at 16 clusters and $m = 1.25$, for instance, these shell proteins have mean carboxysome-cluster membership of just 0.262 ± 0.063 and 0.308 ± 0.095, respectively (Figures S1, S2)

We also examined the published expression patterns of the *csoS1D* ortholog *slr0169*, found in the genome of the beta carboxysome-forming cyanobacterium *Synechocystis* sp PCC6803. In these organisms, the carboxysome genes (*sll1028–1032* (*ccmK2, ccmK1, ccmL, ccmM, ccmN*) *slr0009*, *slr0011*, and *slr0012* (RubisCO-related genes), *slr0436* (*ccmO*), *slr1838–1839* (*ccmK3, ccmK4*), and *slr1347*(*ccaA*) are scattered throughout the genome. Eisenhut *et al.* [19] examined the whole-genome transcriptional response of *Synechocystis* sp. PCC 6803 to long-term inorganic carbon limitation, a stress condition that is expected to affect expression of carboxysome-related genes. In this experiment, 24 h after WT 6803 cells were shifted from 5% CO_2 to 0.03% CO_2, expression of *slr0169* was found to be significantly (>2-fold) downregulated, in tandem with several known carboxysome genes: *ccmK2*, *ccmK3*, *ccmK4*, and *ccmLMN*. Expression levels of the remaining carboxysome genes (*ccmK1, rbcLXS*, and *ccaA*) dropped by a factor of 1.2–1.8, or remained unchanged (*ccmO*). Under these conditions, then, expression of *slr0169* has more in common with expression of most genes encoding carboxysome shell proteins than have genes as central to carboxysome function as *rbcLS*. The same phenomenon is observed in the light-stress study of Singh *et al.* [20].

The Structure of CsoS1D: A Tandem BMC Domain Protein

CsoS1D was expressed and crystallized in two distinct crystal forms. The orthorhombic ($P2_12_12_1$) crystal form was solved by multiwavelength anomalous dispersion (MAD) methods using 2.3 Å diffraction data from a crystal prepared with selenomethionine substituted protein. This crystal form contains six molecules in the crystallographic asymmetric unit. The structure refined to final R and Rfree values of 18.5 and 21.6 %, respectively (Table S1, Figure S3). The second crystal form (R3, rhombohedral) was subsequently solved by molecular replacement using the monomer from the orthorhombic crystal form structure as a search model. Although twinned, the rhombohedral crystal form refined to 2.2 Å with 20.4% and 25.9% for R and Rfree values, respectively. The twin fraction for the rhombohedral form refined to 0.45. For both crystal forms, the amino acid sequence was readily modeled with the exception of the N-terminus (~50 residues) of the protein which was absent from the electron density maps in both crystal forms. The stereochemistry of the final refined model in both crystal forms is excellent with 98.7 % and 94.3% of the residues in the most favored regions of the Ramachandran plot for the orthorhombic and rhombohedral crystal forms, respectively and no residues in the disallowed regions.

Carboxysome shell proteins are characterized by a small (~80 amino acid) domain known as the BMC domain (Pfam00936). To date, structures have been determined for BMC-domain containing proteins ranging from 94 to 116 amino acids in length; each containing a single-BMC domain. Sequence analysis of CsoS1D predicted that its C-terminus (residues 190-256) contained a BMC domain, but the N-terminal 150 amino acids lack sequence homology to any known domains. The crystal structures revealed that residues 50-150 in CsoS1D also contain a BMC domain; thus CsoS1D is composed of a tandem pair of BMC domains (referred to as N-and C-BMC; Figure 2). The N- and C-BMC share less than 18% sequence identity (Figure 2B) yet their alpha-carbon backbones superimpose with an RMS deviation of 1.27 Å over 95 residues (Figure 2B).

Overall, the fold of the N-BMC and C-BMC of CsoS1D are very similar to those of the single-BMC domain proteins that have been structurally characterized [14; 16; 17; 21]. The fold consists of three alpha helices (designated A, B and C) and four beta strands (designated β1, β2, β3, β4; Figures 2B, 2C). The C-BMC overlaps with CsoS1A, a shell protein from the alpha carboxysome of the chemoautotroph *H. neapolitanus*, over 81 alpha carbon atoms with a RMS deviation of 1.43Å. The N-BMC superimposes on CsoS1A with a RMS deviation of 1.67Å over 79 alpha carbon atoms. This consistency in domain fold is despite a difference in connectivity; both BMC domains of CsoS1D contain the permutation in secondary structure recently noted in PduU, a single-BMC domain shell protein of the bacterial microcompartment involved in 1,2-propanediol utilization [21]. Like PduU, the N-terminal ~34 residues of each BMC domain in CsoS1D contribute a beta-strand and a short alpha helix (β1 and α–helix A), to the overall fold of the domain; in the typical single-domain BMC proteins these secondary structure elements are instead contributed by the C-terminus.

The CsoS1D Trimer Forms a Pseudohexamer that Contains a Gated Pore at the Three-fold Symmetry Axis

The asymmetric unit of the orthorhombic crystal form of CsoS1D contains six molecules tightly associated with one another, to form a dimer of trimers that constitutes a hexamer (Figures 2D, 2E). In the rhombohedral form, the two CsoS1D molecules in the asymmetric unit form a dimer, with the crystallographic three-fold axis generating a similar dimer of trimers. The interactions between the trimers to form a dimer of trimers/hexamer are discussed below.

Because each CsoS1D monomer is composed of a tandem pair of BMC domains, CsoS1D trimers form pseudo-hexamers (Figure 2A) that recapitulate the single-BMC domain carboxysome shell protein hexamers that are thought to constitute the facets of the carboxysome shell [14; 16; 17]. One face of the trimer appears slightly broader than the other and more closely resembles a regular hexagon (Figures 2, 3). The two sides of the trimer also have distinctive electrostatic surfaces with the broader, more hexagonal face being less polar (Figure 3).

In addition to these differences between the two faces of each trimer, we observed a striking difference between the two trimers in each CsoS1D hexamer. In the single-BMC domain carboxysome shell proteins that have been structurally characterized, the hexamer's six-fold axis of symmetry is lined by conserved, positively charged sidechains. These pores are proposed to be conduits for diffusion of metabolites such as HCO_3^-, ribulose-1,5-bisphosphate (RuBP) and 3-phosphoglycerate (PGA) across the carboxysome shell. In each trimer of CsoS1D, the sidechains of the invariant residue Arg121 converge at the three-fold axis of symmetry; however the sidechains of Arg121 and of Glu120, which is also absolutely conserved among CsoS1D orthologs, adopt dramatically different conformations in the two trimers (Figures 4A, 4B; Supplementary animation). In one trimer, Arg121 and Glu120 sidechains are oriented to leave a pore of ~14Å at the three-fold axis of symmetry (Figures 2A, 3C, 3D Figure 4); we refer to this trimer as the open trimer. In contrast, in the other "closed" trimer, each CsoS1D monomer's Glu120 sidechain forms a salt bridge with Arg121 of the adjacent monomer, effectively closing the pore

(Figure 4A). Strikingly, in both crystal forms, the CsoS1D hexamer is a dimer of one open and one closed trimer. The open and closed trimers in the orthorhombic crystal superimpose with an RMS deviation of 1.43 Å over 599 alpha carbon atoms.

To quantify the differences between the monomers in the open and closed trimers, the six monomers in the asymmetric unit of the orthorhombic crystal forms were compared to one another (Table S2). Two classes of conformation of the monomer emerged, each distinctive of either the open or closed trimer. The three monomers in the open trimer (class 1) share the same conformation, as they superimpose with a RMS deviation ranging from 0.14 to 0.21 Å. The three monomers in the closed trimer (class 2) also share essentially the same conformation, as they superimpose with RMS deviations ranging from 0.19 to 0.27 Å. When compared to each other, however, the two classes of monomers appear to be significantly different, as they superimpose with RMS deviations ranging from 0.95 to 1.1 Å over 204 alpha carbon atoms.

This distinct conformational difference between the monomers in the two types of trimers is also observed in the rhombohedral crystal form. The asymmetric unit of the rhombohedral crystal form contains only two molecules of CsoS1D; one from the open and one from the closed trimer. They superimpose with an RMS deviation of 1.03 Å over 201 alpha carbon atoms. Further analysis (data not shown), indicates that the open trimer in the R3 form contains monomer similar to the class 1 and the closed trimer contains monomers similar to class 2. Although the crystal forms are clearly distinct and grown from very different crystallization conditions, CsoS1D adopts the same two specific conformations that make up the open and closed trimers.

The structural differences between the monomers that compose the open and closed trimers are distributed across the polypeptide chain. The largest differences are localized to the loop that gates the pore (residues 120-123). The counterpart to the gating loop in the C-BMC (residues 224 to 227) also differs substantially between the monomers of the open and closed trimers. Likewise, in both the N- and C-BMC, the beta strand (β2) structurally adjacent to these loops is shifted concomitantly. Finally, part of the region of permutation in the primary structure (the A helix and the loop connecting the A helix to the β2 strand) in both the N- and C-BMC also differ substantially in conformation between the class 1 and class 2 monomers (rmsd between the class 1 and class 2 monomers alpha carbon backbones differ by ~0.68 Å and 0.36 Å in this region of the N- and C-BMC, respectively). The collective effect of these differences between the class 1 and class 2 monomers on the trimers can be viewed by toggling between the open and closed trimers in a superposition (Supplementary animation).

CsoS1D is a Dimer of an Open and a Closed Trimer

The CsoS1D structure from the orthorhombic crystal form contains six molecules arranged as a dimer of trimers in the crystallographic asymmetric unit (Figures 2D, 2E). Likewise, in the rhombohedral crystal form, a dimer of open and closed trimers is revealed by applying the crystallographic symmetry to the two molecules contained in the asymmetric unit. In both cases, the dimer interface is formed by interaction between the relatively non-polar and hexagonal (Figures 3B, 3D) surface of the two trimers.

The interface between the two trimers buries 6,573 Å^2 of surface area (all surface areas are reported as the total surface area buried). The two surfaces fit snugly together (Figures 2E, 4C) as reflected by their shape correlation statistic, Sc. Values for Sc range between 0 and 1; 1 being a perfect fit and 0 no interaction, [22]; the Sc for the trimer-trimer interface of CsoS1D is 0.70 and 0.68 for the orthorhombic and rhombohedral crystal form structures, respectively. The trimer-trimer interface residues are primarily found on the A helix and the loop that connects the A helix to beta strand 2 in both the N- and C-BMC domains (residues 67-84 and 170-188; Figure 2C); one of the most structurally variable regions between the class 1 and class 2 monomers. These are also the regions of secondary structure in the N-BMC and C-BMC domains that are permuted relative to known single-BMC domain proteins. A closer inspection of the architecture of the CsoS1D hexamer reveals that the open and closed trimers are offset by a rotation of ~60° degrees across the dimerization interface. Thus, each monomer from one trimer thus interacts with two monomers from the opposite trimer (Figure 2E).

CsoS1D Models into a Carboxysome Shell Layer

In previous studies of the single-BMC domain carboxysome shell proteins, the proteins formed hexamers that tend to assemble into uniformly oriented layers in the crystals, suggesting a model for the facets of the carboxysome shell [14; 16; 17]. The size and hexagonal shape of the CsoS1D pseudohexamer suggested that it could fit into such a model. To generate a model for a facet of the *Prochlorococcus* MED4 carboxysome shell, we constructed a layer of CsoS1 (a single-BMC domain protein) hexamers based on the layer structure of CsoS1A from *H. neapolitanus* [23]. Both *Prochlorococcus* MED4 and *H. neapolitanus* contain alpha carboxysomes and the sequence identity between CsoS1 of *Prochlorococcus* MED4 and *H. neapolitanus* CsoS1A is 80.4% overall. More than half of the sequence variation is confined to the C-terminal 10 residues of CsoS1, which are not part of the inter-hexamer interface in the CsoS1A layer; when these residues are omitted from the alignment the sequence identity is 93%.

Adjacent hexamers in the resulting CsoS1 model layer (Figure 6A) pack together with a Sc of 0.67 (compared to 0.68 for the adjacent CsoS1A hexamers observed in the crystal structure), burying 1,697Å^2 of surface area between adjacent hexamers (Table 1). We then substituted either an open or a closed CsoS1D trimer for a CsoS1 hexamer in the layer and examined the initial fit of each. The open trimer clashed somewhat with adjacent CsoS1 hexamers, and would require a slight adjustment to the mainchain in addition to some sidechain remodeling to obtain a realistic fit. In contrast, the closed trimer modeled readily into the CsoS1 layer (Figure 6) and required only rotamer adjustments in two amino acid sidechains to prevent any clashing. This CsoS1D-CsoS1 model was refined by energy minimization. The Sc and buried surface area in the resulting heterologous interface between the closed trimer of CsoS1D and the adjacent CsoS1 hexamers was calculated; they indicate that modeled interaction is comparable to the interfaces among shell proteins that have been observed or modeled to-date (Table 1).

DISCUSSION

The overall pseudohexameric structure of the CsoS1D trimer and its fit within a model layer of CsoS1 (Figure 6), suggest that it is part of the carboxysome shell of *Prochlorococcus*. The amount of surface area buried (Table 1) between CsoS1D and CsoS1 in the modeled interaction is comparable to that in experimentally observed the homo-oligomeric interactions between hexamers of single-BMC domain proteins. The shape complementarity of and the amount of surface area buried in the modeled CsoS1D-CsoS1 heterologous interfaces are comparable to the experimentally observed homo-hexameric interfaces and to the heterologous interfaces between the pentameric and hexameric shell proteins in the most plausible models for the carboxysome shell (Table 1).

The CsoS1D pseudohexamer exhibits features necessary for incorporation into proposed models for the facets of the carboxysome shell. The pseudohexamer edges are of appropriate length and each edge contains the most strongly conserved residue among all BMC domains, the lysine residue at position 23 in the canonical BMC domain (see Pfam00936 HMM Logo http://pfam.sanger.ac.uk/family?acc=PF00936). This amino acid is found at the two-fold symmetry axis between adjacent hexamers in the CcmK1, CcmK2 and CsoS1A layers (and the model layer of CsoS1, Figure 6); in each case its sidechain is positioned roughly parallel to the edge of the hexamer, held in that conformation by hydrogen bonds to the main chain carbonyl oxygen of the lysine in the opposite BMC domain and to the sidechain of either an Asp or Asn residue (conserved at position 19 in the Pfam00936 HMM Logo) within the same monomer. The conformation and the hydrogen bonding interactions between these residues in CsoS1D (Asp104 and Lys108, and Asn 208 and Lys212 in the N-BMC and C-BMC, respectively), are present in both crystals forms. Thus the conformation of the sidechains of the conserved lysine residues in the N-BMC and C-BMC are disposed to interact with adjacent hexamers at the two fold-axis of symmetry (Figure 6). Presumably, if it was not a carboxysome shell component, CsoS1D would be released from selective pressure to retain these residues (and their conformation) and the overall size and shape of the pseudohexamer.

In addition to the structural evidence for a role in the carboxysome shell, several other lines of evidence substantiate the proposed role for CsoS1D in the carboxysome. From genomic sequence analysis, all bacteria that appear to have the potential to form bacterial microcompartments (Kerfeld et al., unpublished) contain two or more BMC domains within their genomes; high light-adapted *Prochlorococcus* strains appeared to be the exception. The detection of the tandem BMC domain protein CsoS1D increases the known number of BMC domains in the genomes of *Prochlorococcus* MED4 and other high light adapted *Prochlorococcus* strains from one to three.

At the level of gene expression, the co-expression of *csoS1D* and the known carboxysome proteins in synchronized *Prochlorococcus* MED4 cells is consistent with a role in the carboxysome for CsoS1D. In unperturbed, synchronously cycling

Prochlorococcus MED4 cells, the *csoS1D* expression pattern clusters more tightly with that of RubisCO and the major carboxysome shell genes than does expression of the known shell genes *csoS4A* and *csoS4B.* A similar pattern for the *csoS1D* ortholog in *Synechocystis* PCC 6803, *slr0169*, emerges under low-carbon and high-light stresses [19; 20]. CsoS1D even offers a rationalization of one seemingly discordant expression pattern uncovered in an examination of the transcriptional effects of deleting *glcD* [19]. The enzyme GlcD is involved in detoxifying the byproduct of photorespiration, long proposed to be a signal of C_i starvation [24]. Growing the Δ*glcD* mutant grown at high CO_2 does partially phenocopy the metabolic state of wild-type cells grown at low CO_2 [25], but analysis of wild-type and Δ*glcD* mutant gene expression under high and low CO_2 appears to preclude the possibility that a single regulatory mechanism coordinates expression of all *Synechocystis* PCC 6803 carboxysome genes, *slr0169* included [19]. Eisenhut *et al.* propose that governing carboxysome gene expression by multiple mechanisms would allow the cell to tune the composition of the carboxysome shell according to the circumstances of the cell's growth. The identification of CsoS1D as a carboxysome shell protein gives this argument credence in two regards. First, all alpha carboxysome-containing cyanobacteria maintain a slight but consistent genomic separation of *csoS1D* from the alpha carboxysome operon; this conserved genomic context is consistent with there being a selective advantage to the disjoint regulation of different shell proteins. (Such an argument is difficult to advance in the beta carboxysome-containing *Synechocystis* sp. 6803, as its operons are in general poorly organized.) Second, the ability to tune carboxysome shell composition might be particularly useful if different shell proteins have different metabolite transport properties, as the novel pore structure in CsoS1D strongly suggests.

The presence of orthologs to *csoS1D* (defined as a reciprocal bi-directional BLAST hit) in all cyanobacterial genomes also is consistent with a structural role for this protein in the carboxysome. A *csoS1D* ortholog (PCC-0905) is also found in the genome of the chromatophore of the amoeba *Paulinella* [26]. The *Paulinella* chromatophore appears to have originated from a relatively recent endosymbiosis of a cyanobacterium closely related to *Synechococcus* WH5701, however the chromatophore genome has undergone a massive genome reduction (to 26% of the gene content of *Synechococcus* WH5701). Among the genes retained in the chromatophore genome is a complete set of photosynthesis genes [26], in which the cyanobacterial alpha carboxysome gene cluster (Figure 1B), including *csoS1D* and *ham1*, is conserved intact.

The pore formed at the three-fold axis of the CsoS1D pseudohexamer is unique among bacterial microcompartment shell proteins that have been structurally characterized to date. CsoS1D is the first BMC protein structure to reveal a potential for gating at the pore (Figures 3, 4). Because the sidechains of the conserved residues that line the pore in the CcmK1, 2 and 4 and CsoS1A hexamers are fully extended, these 4-7Å pores appear to be constitutively open [14; 16; 17]. In contrast, the CsoS1D pore is open or closed depending on the sidechain conformations of two residues, Glu120 and Arg121 (Figure 4), both absolutely conserved among all

CsoS1D orthologs. The open pore in CsoS1D is ~14Å in diameter, significantly larger than those of single-BMC domain carboxysome shell protein hexamers. A larger pore may be advantageous for the diffusion of the bulkier carboxysome metabolites (e.g. RuBP) which would be expected to pass through the CcmK1, CcmK2, CcmK4 and CsoS1A pores only with difficulty.

Because both of the CsoS1D structures determined here lack the N-terminal 50 amino acids, it is possible that there is an additional domain situated over the CsoS1D pore that may be involved in regulating transport across the trimer. This is an enigmatic region of primary structure as it is only found in the *Prochlorococcus* and *Synechococcus* CsoS1D orthologs and its primary sequence is relatively poorly conserved.

The structure of CsoS1D is the first model of a tandem BMC domain containing protein (Figure 2); the second BMC domain in the N-terminus of the protein was unexpected, since it was not predicted from the primary structure. The genomes of organisms that produce beta carboxysomes all contain a gene for a tandem-BMC domain protein, *ccmO*, but one had not been identified in cyanobacteria that form alpha carboxysomes. In CsoS1D, combining the two BMC domains and a trimeric oligomeric state results in a pseudohexamer, which presents two types of edges (Figures 2, 6) for interaction with neighboring (pseudo)hexamers within a model carboxysome shell layer. To date, only homohexamers of single-BMC domain carboxysome shell proteins have been structurally characterized; it is not yet known whether mixtures of single-BMC domain containing proteins such as

CcmK1-4 or CsoS1A,B,C form mixed hexamers under physiological conditions. Fusing two BMC domains has implications for their assembly into shell layers; fusion increases the effective concentration of protein domains with respect to each other [27] which could possibly enhance the rate of shell assembly. At the same time, the presence of two distinct edges to match reduces the number of degrees of freedom of fit. This would be expected to slow the assembly of heterohexamers into a layer.

Interestingly, there are no single-BMC domain homologs of the N-terminal BMC domain of CsoS1D in the sequence database. Because of this and the general observation that gene fusion is a later event in protein evolution, the most plausible scenario for the evolutionary history of CsoS1D includes a duplication of a permuted BMC domain, followed by fusion with subsequent divergence of the primary structure of the N-BMC domain. The underlying functional reason for the divergence in sequence of the N-BMC domain is presently unknown.

In both crystal forms of CsoS1D, the protein was arranged as a dimer of trimers (Figures 2D, 2E) formed by the dimerization of the broader, more hexagonal, faces of two trimers. Several lines of evidence suggest that a dimer of trimers is a physiologically relevant form of CsoS1D. In general, interfaces in obligate oligomers, such as most homodimers, are large and relatively hydrophobic [28]; in CsoS1D the dimerization buries a total of 6,573 Å^2. Moreover, the face of the CsoS1D trimer buried by the dimerization is less polar than the opposite face that is exposed to solvent (Figure 3); more than half of residues buried in dimerization are nonpolar (14 of 26 in each monomer). Furthermore, the fit between the two interfaces is snug; the shape complementarity is 0.70 for the orthorhombic crystal form and 0.68 for the rhombohedral crystal form (a typical value for an antibody: antigen complex is 0.64-0.68 [22]). Finally, the residues interacting across the trimer-trimer interface also are strongly conserved; 21 of the 26 residues are identical among all *Prochlorococcus* and *Synechococcus* CsoS1D orthologs.

The stability of interaction between the trimers is likely enhanced by staggering the interactions among domains; each monomer in a trimer interacts with two different monomers in the opposite trimer (Figure 2E). Notably, the 26 residues each monomer contributes to the trimer-trimer interface are all found in the permuted segments of the primary structure (between residues 50-84 of the N-BMC and 156-190 of the C-BMC) in CsoS1D. Likewise, in the structure of PduU, the only single-BMC domain protein containing this permutation that has been structurally characterized, there is a similar dimerization of hexamers, which is likewise mediated by the permuted regions of primary structure [21]. The elution behavior of CsoS1D in size exclusion chromatography is also consistent with hexameric and larger assemblies (data not shown). Collectively these observations suggest that the dimer of trimers configuration, found in the two crystal forms of CsoS1D, is physiologically relevant.

A structural consequence of the tight appression of two CsoS1D trimers is a central channel formed by the continuity between the pores of the two trimers (Figure 4D) that is open on one end and closed on the other. The Glu120 and Arg121 residues involved in gating the opening are closer to the outer, solvent exposed surface of each trimer. The channel is approximately 73 Å lengthwise and ranges in diameter between 0 and 18 Å (Figure 4D). The interior of the channel is mainly positively charged (Figure 4C). The volume of the channel is substantial, 13,613 Å^3, of sufficient size to be a microenvironment for ligand binding and catalysis. Accordingly, we screened CsoS1D for a range of enzymatic activities and attempted co-crystallization with metabolites that must cross the carboxysome shell (e.g. RuBP and bicarbonate) but these experiments did not yield conclusive results.

While the trimeric/pseudohexameric structure of CsoS1D is compatible with current models for the facets of the carboxysome shell, the dimerization of the CsoS1D trimers raises new questions about the details of shell architecture: for example, could the entire carboxysome shell be composed of a bilayer of BMC-domain containing proteins? In the structures of BMC proteins that form layers of uniformly oriented hexamers (CsoS1A, CcmK1 and CcmK2), adjacent layers in the crystals are aligned with respect to their pores, but the orientation of molecules between layers varies (i.e. in CsoS1A all of the layers are oriented similarly, in CcmK1 the convex surfaces of hexamers in adjacent layers appear to interact, in CcmK2 isologous interfaces are juxtaposed but any two adjacent layers were not as tightly appressed as in the CsoS1D dimer of trimers). Alternatively, it may be that only a subset of proteins making up the carboxysome shell are assembled in two tiers; this subset is perhaps restricted to those containing the secondary structure permutation observed in CsoS1D and PduU, which also forms dimers of hexamers [21]. Dyad shell building blocks may occur for a specific functional purpose and may be differentially regulated; CsoS1D's position outside of the known carboxysome operon and the expression profile of its *Synechocystis* ortholog *slr0169* are consistent with this hypothesis. Based on the CsoS1-CsoS1D model, the second trimer could occur as a protrusion from the surface of the carboxysome (where, for example, it could be a part of an adjacent carboxysome). Alternatively, the second trimer could face the interior of the carboxysome, where it could play a role in organizing the interior enzymes RubisCO and/or CsoSCA.

CONCLUSIONS

There is compelling evidence that PMM0547 in *Prochlorococcus* (provisionally dubbed CsoS1D) is a carboxysome shell protein. Here we report two independent crystal structures of this protein, and a model for its interaction with CsoS1 in the *Prochlorococcus* carboxysome shell. The structural models are complemented by genomic and transcriptomic evidence which together suggest that PMM0547 is a carboxysome shell protein. Strikingly, although bioinformatic methods predicted that CsoS1D would contain a single BMC domain, the crystal structure reveals it to have two tandem BMC domains, making it the first double BMC-domain protein to be structurally characterized. In addition to uncovering the second, cryptic, BMC domain, structural analysis of CsoS1D revealed several other unexpected features. All characterized single-BMC-domain proteins form a hexameric ring; CsoS1D trimerizes, but the tandem BMC domains make it a pseudohexamer that is compatible with the hexameric layers that are proposed to form the facets of the carboxysome shell. Moreover, CsoS1D trimers form tightly appressed dimers, sandwiching two rings together to form a novel two-tiered carboxysome shell building block. The structures of CsoS1D also reveal an unusually large pore that can adopt both open and closed conformations. No other carboxysome shell protein has been observed to possess a two-state pore; CsoS1D thus is the first shell protein to present a means for gating metabolite flow across the carboxysome shell. Collectively, these features are likely to have a significant impact on carboxysome ultrastructure and function in *Prochlorococcus*. Moreover, the universality of CsoS1D orthologs alongside canonical shell proteins in other cyanobacteria suggests that, should the environment demand it, the functions made possible by these novel carboxysome features may be readily available to all cyanobacteria

Materials and Methods

PMM0547 Cloning, Expression, and Purification

Prochlorococcus marinus subsp. MED4 genomic DNA was used as template in polymerase chain reaction using the oligonucleotide primers with the following sequences: CGACCATGGAACCAACTTCTAGC and CGAAGCTTAATAATTTGATATTTGATCAATTGC, which contain *NcoI* (C^CATGG) and *HindIII* (A^AGCTT) restricstion sites. The PCR product was digested with *NcoI* and *HindIII*, and ligated into the multiple cloning site of pProEX-HTb (Life Technologies, Inc.), which adds a His_6 to the N-terminus of the PMM0547 protein coding sequence. The sequence of the protein coding region of the recombinant plasmid was confirmed by DNA sequencing (The Univ. of Maine DNA Sequencing Facility).

For expression of recombinant protein, the constructs were transformed into *Escherichia coli* BL21 cells (Novagen). For Seleno-methionine labeling of the PMM0547 gene product, the plasmid was transformed into the *E. coli* methionine auxotroph B834 strain (Novagen). PMM0547 was purified from fresh BL21 cultures grown in LB (ALDRICH-SIGMA) to $O.D._{600}$=0.8 at 37 °C, and induced at 25 °C for 18 h with 0.4 mM IPTG (ALDRICH-SIGMA). Following centrifugation, the cell pellets were lysed by sonication (model W-220F, Branson) in buffer containing 500 mM NaCl, 50 mM Tris-HCl (pH 8.0), 10% glycerol and 2 mM Beta-ME. Poly-histidine tagged proteins were purified using the standard purification protocol recommended for use with Ni-TA resin (QIAGEN). To obtain higher purity preparations for use in SAXS, the recombinant PMM0547 protein was purified using TALON resin (CLONTECH). Proteins were eluted from these resins with 200 mM imidazole, 500 mM NaCl, 100 mM Tris-HCl (pH 8.0). The purified recombinant 0547 protein (5-10 mg/ml) was dialyzed overnight in 10 mM Tris-HCl (pH 8.0) at 4 °C. Purified protein was stored at 4 °C and diluted to 1 mg/ml with dH_2O prior to setting up crystallization trials.

Crystallization and Structure Determination

Orthorhombic crystals were grown at 18 °C by hanging drop vapor diffusion by mixing with a solution containing 12 % Tacsimate (HAMPTON RESEARCH) adjusted to pH 6.2 with HCl. The rhombohedral crystal form was grown at 18 °C by hanging drop vapor diffusion by mixing with 0.7M ammonium phosphate (HAMPTON RESEARCH), adjusted to pH 4.2 with 100 mM CHES (ALDRICH-SIGMA) (pH 9). Crystallization drops were prepared with either 4+1 or 2+1 μl protein-to-crystallization buffer drop ratios for the Tacsimate and ammonium phosphate conditions, respectively.

Native and selenomethionine (ALDRICH-SIGMA) derived crystals were soaked rapidly in crystallization solutions supplemented with 30% ethylene glycol (HAMPTON RESEARCH) prior to freezing. Tacsimate-grown crystals were plunged into liquid nitrogen and crystals from the ammonium phosphate condition were frozen by placing directly into the cryo-stream of nitrogen gas. Diffraction from the crystals prepared with Tacsimate was consistent with orthorombic space group $P2_12_12_1$ (unit cell dimensions: A=122.41, B=131.286, C=131.762, $\alpha,\beta,\gamma=90$) and the diffraction from crystals grown from the ammonium phosphate condition was Rhombohedral Space Group R3 (unit cell dimensions: A=117.163, B=117.163, C=101.463 $\alpha=\beta=90, \gamma=120$). Diffraction data were collected at Lawrence Berkeley's Advanced Light Source beamline 8.2.2 and processed with DENZO and SCALEPACK [29].

We prepared crystals of seleniomethionine substituted protein to assist in solving the phases for the orthorhombic crystal form. Forty heavy atom selenium positions were identified from a dataset collected at the peak wavelength for selenium using the program Phenix.hyss using diffraction data in resolution ranges 30 to 3.5 Å, where the anomalous signal appeared to be strong based on results from the program xtriage in the Phenix software suite [30]. Heavy atom refinement and phasing was performed using autoSHARP [31]. Density modification using SOLOMON [32] produced exceptionally high quality electron density map calculated without NCS averaging to 2.3 Å. Using the density modified phases from SOLOMON, the program ARP/wARP [33], implemented in CCP4 [34], constructed an initial model for the six monomers in the crystallographic asymmetric unit, including fitting the amino acid sequence into the correct registry. Model building was completed using the graphics program COOT [35] and a well defined NCS averaged electron density map (Figure S3) produced using the phases calculated in SHARP after applying solvent flattening and density averaging procedures in the program RESOLVE [36]. Positional and B factor refinements and water-picking were performed using Phenix.refine. The stereochemistry of the final refined model was analyzed with the program MolProbity [37]. The 6 chains that comprise the hexamer found in the asymmetric unit have connectivity throughout the polypeptide chain yet the residues absent at the N-and C-terminal are variable. Residues 50 to 256, 48 to 256, 50 to 256, 53 to 256, 52 to 252, and 52 to 256 are present in chains A-F of the final model, respectively.

Diffraction data from the rhombohedral crystal form was consistent with twinning as detected in xtriage from the phenix suite. Nevertheless, the structure was solved by molecular replacement methods using the monomeric form from the orthorhombic crystal form using the program Phaser [38] and refined to 2.2 resolution using phenix. The structure was refined using a twin fraction of 0.45. In R3 chain

B the loop residues 120 to 122 are refined with occupancies of zero, as the electron density map that corresponds to this region of the polypeptide chain was not well defined.

Analysis of structure surface complementarity and surface area was performed with CCP4 programs SC [39] and AREAIMOL [40], respectively. Least squares RMS deviation was calculated in CCP4 using program LSQKAB [41]. The model of CsoS1 was generated with SWISSMODEL [42]. The docking model of CsoS1D-CsoS1 was refined using CNS energy minimization program [43].

Clustering analysis of diel expression data

For details of sample collection and data processing, see Zinser *et al.* (submitted). In their RMA-averaged data, they identified 1405 genes in *Prochlorococcus* strain MED4 whose expression oscillated when cells were grown in light:dark (14:10) synchronized cells doubling once per day. The expression profiles of these genes were subjected to soft c-means clustering using the R package Mfuzz [18; 44]. Clustering was performed with 25 input parameter pairs, each combination of 5 cluster numbers *c* (10, 12, 14, 16, and 18) and 5 "fuzzification" paramters *m* (1.15, 1.20, 1.25, 1.30, and 1.35). This analysis was repeated 100 times with each set of input parameters as a check on the variability of the output fractional cluster membership (μ) values. Figure S1 was generated using the R package Heatplus, taking as input the output from one randomly selected clustering run for each pair of input parameters.

REFERENCES

1. Dobrinski, K. P., Longo, D. L. & Scott, K. M. (2005). The carbon-concentrating mechanism of the hydrothermal vent chemolithoautotroph *Thiomicrospira crunogena*. *J Bacteriol* **187**, 5761-5766.
2. Price, G. D., Sultemeyer, D., Klughammer, B., Ludwig, M. & Badger, M. R. (1998). The functioning of the CO_2 concentrating mechanism in several cyanobacterial strains: a review of general physiological characteristics, genes, proteins and recent advances. *Can J Bot* **76**, 973-1002.
3. Scott, K. M., Henn-Sax, M., Harmer, T. L., Longo, D. L., Frame, C. H. & Cavanaugh, C. M. (2007). Kinetic isotope effect and biochemical characterization of form IA RubisCO from the marine cyanobacterium *Prochlorococcus marinus* MIT9313. *Limnol Oceanogr* **52**, 2199-2204.
4. Yeates, T. O., Kerfeld, C. A., Heinhorst, S., Cannon, G. C. & Shively, J. M. (2008). Protein-based organelles in bacteria: carboxysomes and related microcompartments. *Nat Rev Microbiol.*
5. Bobik, T. A. (2006). Polyhedral organelles compartmenting bacterial metabolic processes. *Appl Environ Microbiol* **70**, 517-525.
6. Heinhorst, S., Cannon, G. C. & Shively, J. M. (2006). Carboxysomes and carboxysome-like inclusions. In *Complex Intracellular Structures in Prokaryotes* (Shively, J. M., ed.), Vol. 2, pp. 141-164. Springer, Berlin/Heidelberg.
7. Shively, J. M., Ball, F., Brown, D. H. & Saunders, R. E. (1973). Functional organelles in prokaryotes: Polyhedral inclusions (carboxysomes) in *Thiobacillus neapolitanus*. *Science* **182**, 584-586.
8. Lichtle, C., Thomas, J. C., Spilar, A. & Partensky, F. (1995). Immunological and ultrastructural characterization of the photosynthetic complexes of the prochlorophyte *Prochlorococcus* (Oxychlorobacteria). *J. Phycology* **31**, 934-941.
9. So, A. K., Espie, G. S., Williams, E. B., Shively, J. M., Heinhorst, S. & Cannon, G. C. (2004). A novel evolutionary lineage of carbonic anhydrase (epsilon class) is a component of the carboxysome shell. *J Bacteriol* **186**, 623-630.
10. Dou, Z., Heinhorst, S., Williams, E. B., Murin, C. D., Shively, J. M. & Cannon, G. C. (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.
11. Heinhorst, S., Williams, E. B., Cai, F., Murin, C. D., Shively, J. M. & Cannon, G. C. (2006). Characterization of the carboxysomal carbonic anhydrase CsoSCA from *Halothiobacillus neapolitanus*. *J Bacteriol* **188**, 8087-8094.
12. Badger, M. R. & Price, G. D. (2003). CO_2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. *J Exp Bot* **54**, 609-622.
13. Badger, M. R., Price, G. D., Long, B. M. & Woodger, F. J. (2006). The environmental plasticity and ecological genomics of the cyanobacterial CO_2 concentrating mechanism. *J Exp Bot* **57**, 249-265.
14. Tanaka, S., Kerfeld, C. A., Sawaya, M. R., Cai, F., Heinhorst, S., Cannon, G. C. & Yeates, T. O. (2008). Atomic-level models of the bacterial carboxysome shell. *Science* **319**, 1083-1086.

15. Cannon, G. C., Bradburne, C. E., Aldrich, H. C., Baker, S. H., Heinhorst, S. & Shively, J. M. (2001). Microcompartments in prokaryotes: carboxysomes and related polyhedra. *Appl Environ Microbiol* **67**, 5351-5361.

16. Tsai, Y., Sawaya, M. R., Cannon, G. C., Cai, F., Williams, E. B., Heinhorst, S., Kerfeld, C. A. & Yeates, T. O. (2007). Structural analysis of CsoS1A and the protein shell of the Halothiobacillus neapolitanus carboxysome. *PLoS Biol* **5**, e144.

17. Kerfeld, C. A., Sawaya, M. R., Tanaka, S., Nguyen, C. V., Phillips, M., Beeby, M. & Yeates, T. O. (2005). Protein structures forming the shell of primitive bacterial organelles. *Science* **309**, 936-8.

18. Futschik, M. E. & Carlisle, B. (2005). Noise-robust soft clustering of gene expression time-course data. *J Bioinform Comput Biol* **3**, 965-88.

19. Eisenhut, M., von Wobeser, E. A., Jonas, L., Schubert, H., Ibelings, B. W., Bauwe, H., Matthijs, H. C. P. & Hagemann, M. (2007). Long-term Response Towards Inorganic Carbon Limitation in Wild Type and Glycolate Turnover Mutants of the Cyanobacterium Synechocystis sp. Strain PCC 6803 10.1104/pp.107.103341. *Plant Physiol.*, pp.107.103341.

20. Singh, A. K., Elvitigala, T., Bhattacharyya-Pakrasi, M., Aurora, R., Ghosh, B. & Pakrasi, H. B. (2008). Integration of carbon and nitrogen metabolism with energy production is crucial to light acclimation in the cyanobacterium synechocystis. *Plant Physiol* **148**, 467-78.

21. Crowley, C. S., Sawaya, M. R., Bobik, T. A. & Yeates, T. O. (2008). Structure of the PduU Shell Protein from the Pdu Microcrocompartment of *Salmonella*. *Structure* **16**, 1324-1332.

22. Lawrence, M. C. & Colman, P. M. (1993). Shape Complementarity at Protein/Protein Interfaces. *J. Mol Biol.* **234**, 946-950.

23. Tsai, Y., Sawaya, M. R., Cannon, G. C., Cai, F., Williams, E. B., Heinhorst, S., Kerfeld, C. A. & Yeates, T. O. (2007). The structure of the shell protein CsoS1A from *Halothiobacillus neapolitanus* and its implications for carboxysome function. *PLoS Biology* **5**, 1345-1354.

24. Marcus. Y., Harel, E. & Kaplan, A. (1983). Adaptation of the cyanobacterium *Anabaena varaibilis* to low CO2 concentration in the environment. *Plant Physiology* **71**, 208-210.

25. Eisenhut, M., Huege, J., Schwarz, D., Bauwe, H., Kopka, J. & Hagemann, M. (2008). Metabolome phenotyping of inorganic carbon limitation in cells of the wild type and photorespiratory mutants of the cyanobacterium *Synechocystis* sp Strain PCC6803. *Plant Physiology* **in press**.

26. Nowack, E. C. M., M., M. & Glockner, G. (2008). Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Current Biology* **18**, 410-418.

27. Kuriyan, J. & Eisenberg, D. S. (2007). The origin of protein interactions and allostery in colocalization. *Nature* **450**, 983-990.

28. Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **93**, 13-20.

29. Otwinowski, Z., and Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. In *Methods in Enzymology: Macromolecular Crystallography, part A* (Carter, C. W., and Sweet, R.M., ed.), Vol. 276, pp. 307-326. Academic Press, New York.

30. Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). PHENIX: building new software for automated crystallographic structure determination. *Acta*

Crystallogr D Biol Crystallogr **58**, 1948-54.

31. Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. (2007). Automated structure solution with autoSHARP. *Methods Mol Biol* **364**, 215-30.

32. Abrahams, J. P. & Leslie, A. G. (1996). Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Crystallogr D Biol Crystallogr* **52**, 30-42.

33. Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). ARP/wARP's model-building algorithms. I. The main chain. *Acta Crystallogr D Biol Crystallogr* **58**, 968-75.

34. (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* **50**, 760-3.

35. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-32.

36. Terwilliger, T. C. (2000). Maximum-likelihood density modification. *Acta Crystallogr D Biol Crystallogr* **56**, 965-72.

37. Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., 3rd, Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* **35**, W375-83.

38. Read, R. J. (2001). Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr D Biol Crystallogr* **57**, 1373-82.

39. Lawrence, M. C. & Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. *J Mol Biol* **234**, 946-50.

40. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**, 379-400.

41. Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A* **A32**, 922-923.

42. Peitsch, M. C. (1996). ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem Soc Trans* **24**, 274-9.

43. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **54**, 905-21.

44. Team, R. D. C. (2008). A language and environment for statistical computing. http://www.R-project.org.

45. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876-4882.

46. DeLano, W. L. (2002). DeLano Scientific, Palo Alto, CA.

47. Petrey, D. & Honig, B. (2003). GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol* **374**, 492-509.

48. Smart, O. S., Goodfellow, J. M. & Wallace, B. A. (1993). The Pore Dimensions of Gramicidin A. *Biophysical Journal* **65**, 2455-2460.

ACKNOWLEDGEMENTS

We would like to thank Sandra Schwarte, Martin Hagemann, Hermann Bauwe Michael Sawaya and Todd Yeates for helpful discussions and Edwin Kim and Jay Kinney for figure preparation and technical assistance. SCB is a Howard Hughes Medical Institute Predoctoral Fellow. GRC and SH acknowledge support from the National Science Foundation (MCB-0818680 and DMR-0213883). The work of CAK is performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract number DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract number DE-AC52-07NA27344.

FIGURE LEGENDS

Figure 1. (A) Negatively stained transmission electron micrograph of *Prochlorococcus* MED4 showing a single carboxysome (arrow). (Electron Micrograph courtesy of Luke Thompson (MIT) and Nikki Watson, Keck Imaging Center, Whitehead Institute for Biomedical Research). (B) Genomic organization of the carboxysome gene clusters of the cyanobacterium *Prochlorococcus* MED4 and the chemoautotroph *Halothiobacillus neapolitanus* indicating the positions of the bacterial microcompartment (BMC) domains. Arrows representing genes are proportional to gene size. Panel 2C prepared with CLUSTAL_X [45]

Figure 2. (A) The CsoS1D trimer. (B) Superposition of the N-BMC (green) and C-BMC (blue) domain structures of CsoS1D. (C) Sequence alignment and secondary structure of the N-BMC and C-BMC domains of CsoS1D. Arrows correspond to beta strands and heavy blue lines represent alpha helices, colored as in (B). (D) The CsoS1D dimer of trimers in ribbons and (E) space-filling representation. The N-BMC and C-BMC are colored as in (B). Panels A, C-E and Figures 4-6 prepared with PyMOL [46].

Figure 3. Electrostatic surface representation of the (A and B) closed CsoS1D trimer and the open (C and D) CsoS1D trimer. The surfaces shown in A and C are exposed to solvent in the crystal; those in (B) and (D) are involved in the dimerization of the open and closed trimer. Figures are colored according to electrostatic potential with blue positive and red negative. This figure and figures 4A-C prepared with GRASP2 [47].

Figure 4. A close-up view (from the solvent exposed side of each trimer) of the gated pore formed at the three-fold axis of symmetry in its closed (A) and open (B) conformations. The electrostatic potential is colored as in Figure 3. The conserved amino acids (Glu120 and Arg121) that gate the pore are highlighted. (C) The central channel created by the dimerization of the open and closed CsoS1D trimers. (D)

Radius of the central channel in the dimer of trimers plotted as a function of vertical position along the channel (calculated with HOLE2, ^{48}).

Figure 5. Superposition of the open (gray) and closed (blue) trimers (A) and of (B) individual monomers distinctive of the open (gray) and closed (blue) trimers.

Figure 6. (A) Model of the interaction of the closed CsoS1D trimer (colored as in Figure 2A) and a layer of CsoS1 hexamers (blue). A single CsoS1D monomer interacts with three CsoS1 hexamers, labeled in reference to the interface analysis reported in Table 1. (B and C) Close-up views of the conserved Lys sidechains (circled) at the intermolecular interface of adjacent BMC domains between CsoS1D and CsoS1. (B) and between adjacent CsoS1 hexamers.

Table 1. Properties of protein-protein interfaces among bacterial microcompartment shell proteins

Data bank accession codes

The coordinates and structure factors for both crystal forms of CsoS1D have been deposited in the RCSB PDB (http://www.rcsb.org/pdb/home/home.do) under the accession numbers 3F56 (orthorhombic crystal form) and 3FCH (rhombohedral crystal form).

Supplementary Material

Figure S1. Heat map depicting the clustering patterns (see Methods) of gene expression from 25 time points over two diel cycles in synchronized *Prochlorococcus* MED4 cells for a range of clustering parameters (c, number of clusters; m, "fuzziness" of clustering). Clustering was run on the entire set of 1405 expressed, cycling genes; output is shown for only those genes in the immediate vicinity of the carboxysome operon. The heat map gives the output of one arbitrarily selected clustering run for each set of parameters. Genes *PMM0549-0555* constitute the known carboxysome operon in *Prochlorococcus* MED4. *PMM0547* (*csoS1D*) is separated from the main carboxysome operon by *PMM0548*, encoding a non-carboxysome-related Ham1 protein. Genes not labeled in the operon diagram encode hypothetical proteins. Coloring indicates the strength, μ, of each gene's association with each cluster under a given pair of clustering parameters.

Figure S2. Fractional membership, μ, in the cluster containing known carboxysome genes *CsoS1*, *CbbL, CbbS,* and *CsoS2* for the 20 genes surrounding *CsoS1D.* Genes not labeled in the operon diagram encode hypothetical proteins. Each panel plots the output μ values for these 20 genes from 100 soft c-means clustering runs using the given pair of input parameters and diel expression data for all 1405 *Prochlorococcus* MED4 genes that oscillate in synchronized cells grown on a 24-h light-dark cycle (Zinser et al, submitted). Run-to-run variability, seen as the vertical smearing of

each gene's cluster membership values, increases with cluster number and "fuzziness", but for each set of input parameters, the distribution of membership values for *CsoS1D* (highlighted in light blue) in the cluster containing the carboxysome operon (orange) is higher than that of shell proteins *CsoS4A* and *CsoS4B* (red).

Figure S3. Stereo image showing residues 55 to 65 from Chain A of the refined model inside the experimental electron density map prepared in RESOLVE with density modification using solvent flattening and NCS averaging.

Table S1. Data collection and refinement statistics.

Table S2. Summary of RMS deviations between CsoS1D monomers

Supplemental Movie Comparison of the open and closed CsoS1D trimers

Figure 1

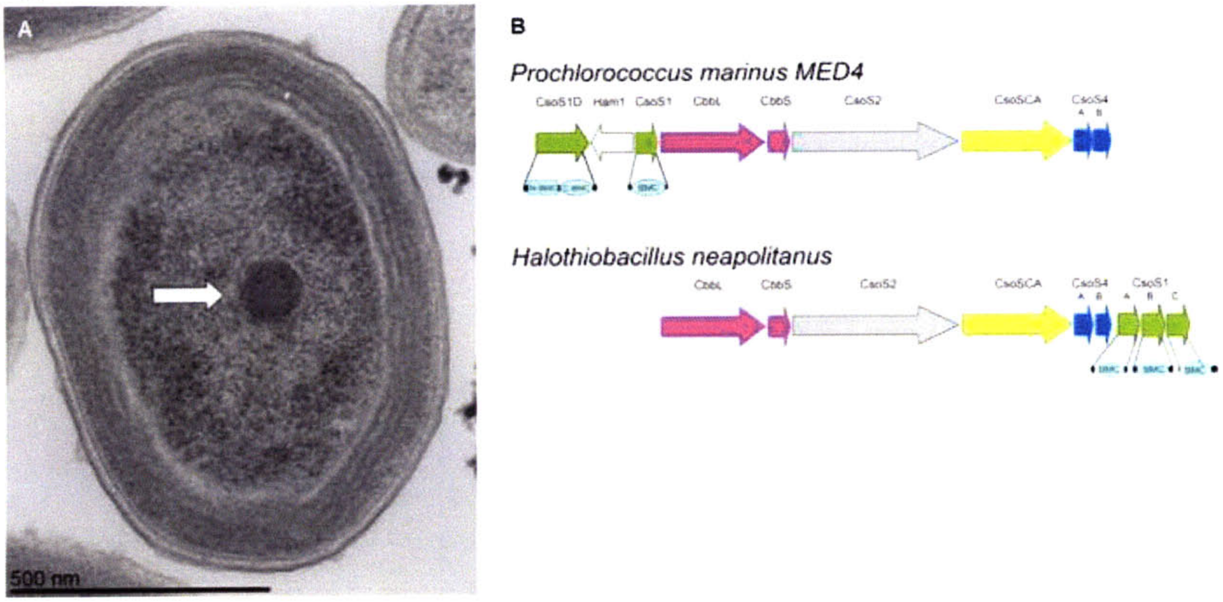

A
500 nm
B
Prochlorococcus marinus MED4
CsoS1D
Ham1
CsoS1
CbbL
CbbS
CsoS2
CsoSCA
CsoS4
A
B
BMC
BMC
BMC
Halothiobacillus neapolitanus
CbbL
CbbS
CsoS2
CsoSCA
CsoS4
CsoS1
A
B
A
B
C
BMC
BMC
BMC

Figure 2

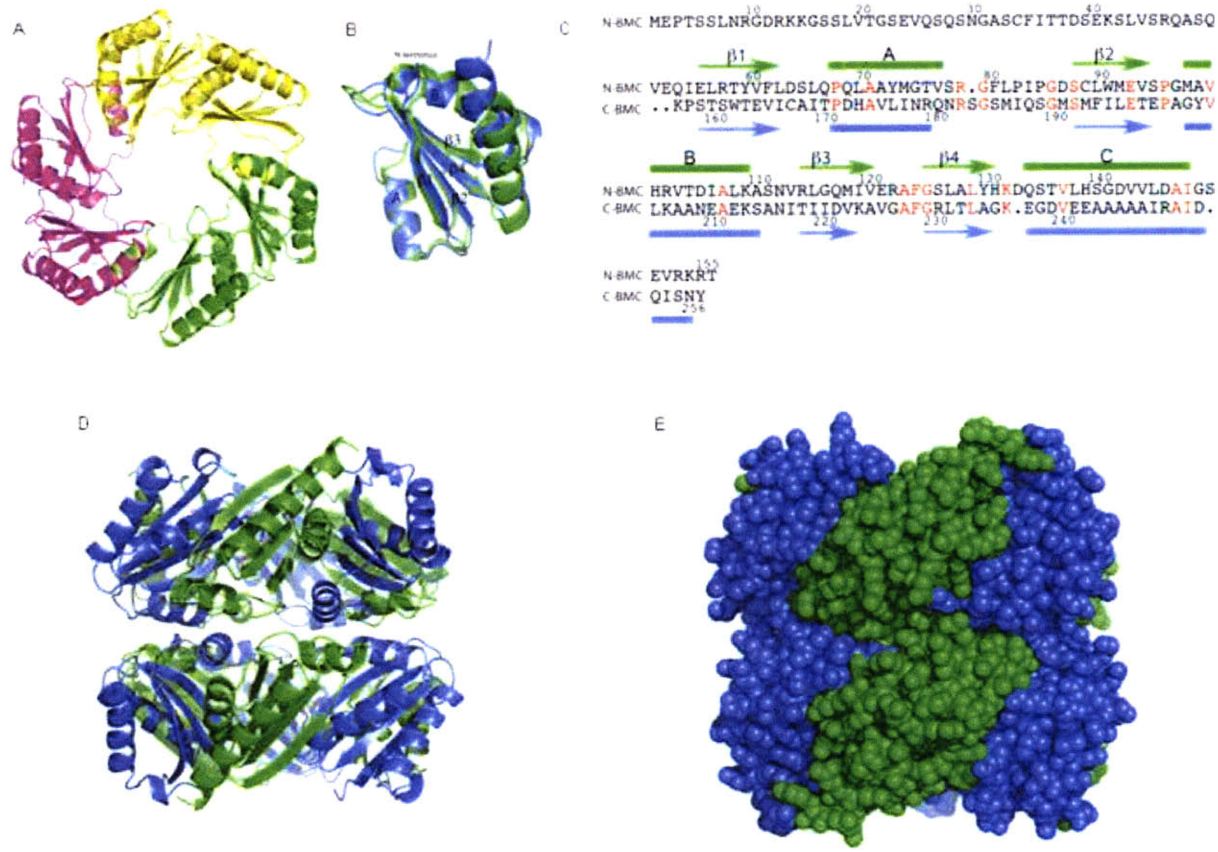

A
B
β3
β4
β2
C
N-BMC MEPTSSLNRGDRKKGSSLVTGSEVQSQSNGASCFITTDSEKSLVSRQASQ
10 20 30 40
β1 A β2
N-BMC VEQIELRTYVFLDSLQPQLAAYMGTVSR.GFLPIPGDSCLWMEVSPGMAV
60 70 80 90
C-BMC ..KPSTSWTEVICAITPDHAVLINRQNRSGSMIQSGMSMFILETEPAGYV
160 170 180 190
B β3 β4 C
N-BMC HRVTDIALKASNVRLGQMIVERAFGSLALYHKDQSTVLHSGDVVLDAIGS
110 120 130 140
C-BMC LKAANEAEKSANITIIDVKAVGAFGRLTLAGK.EGDVEEAAAAAIRAID.
210 220 230 240
N-BMC EVRKRT
155
C-BMC QISNY
256
D
E

Figure 3

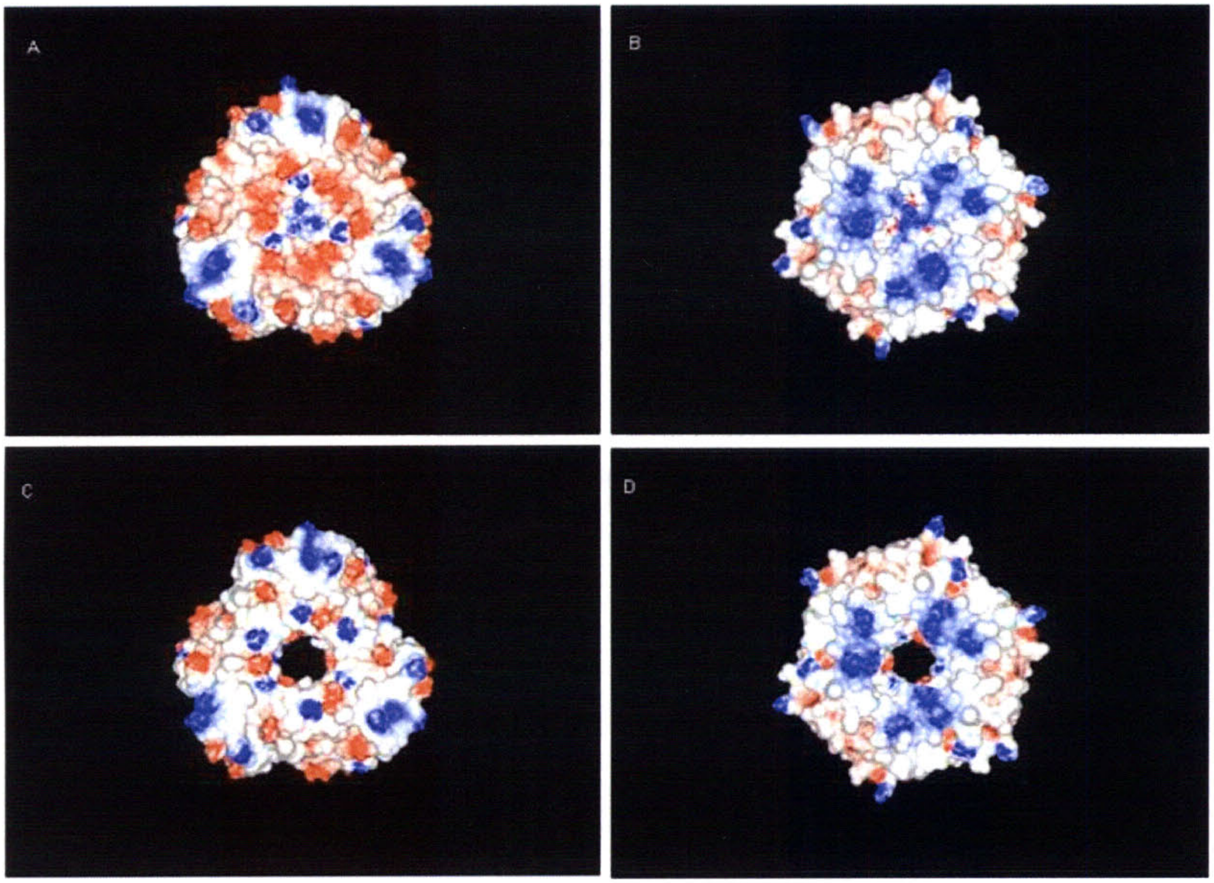
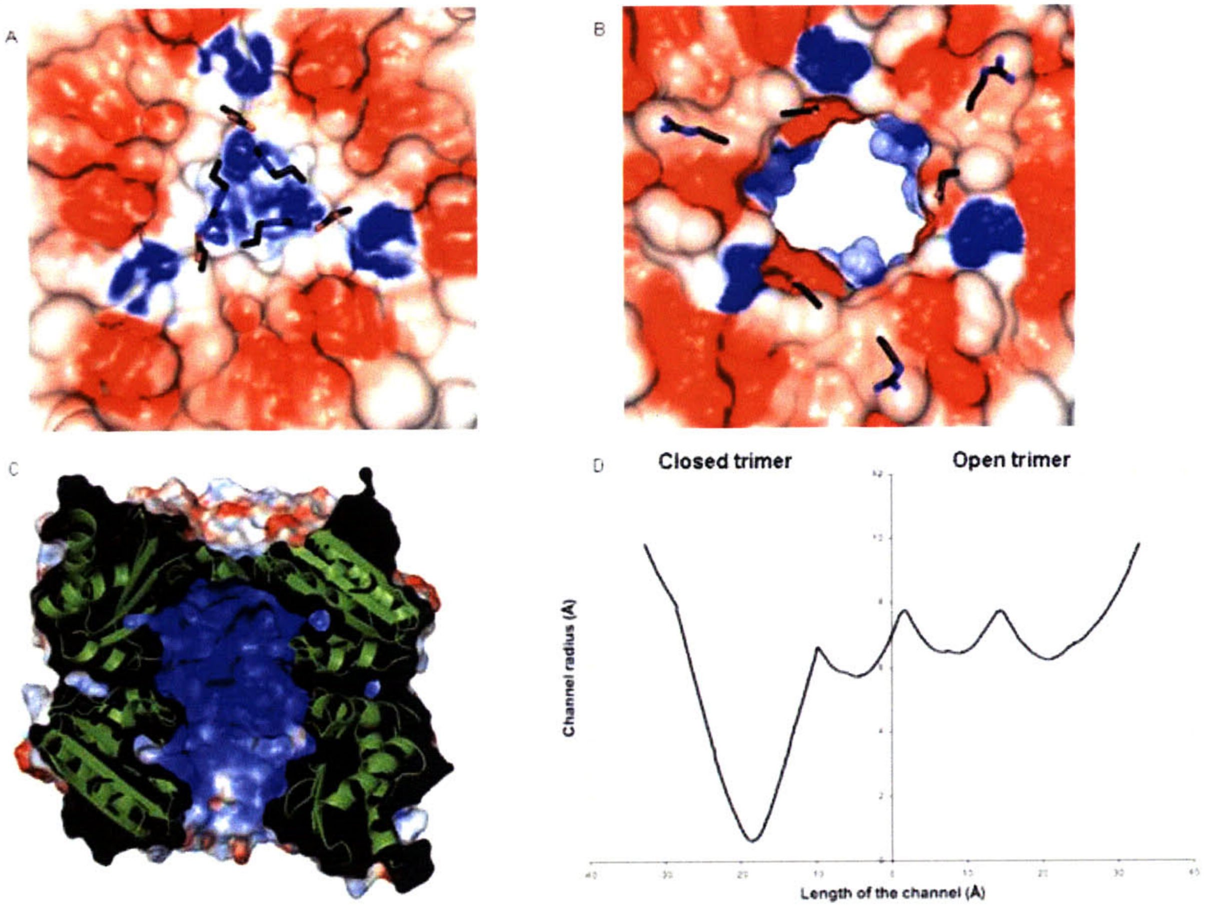


Figure 4



A
B
C
D
Closed trimer
Open trimer
12
10
8
6
4
2
0
Channel radius (Å)
40
30
20
10
0
10
20
30
40
Length of the channel (Å)

Figure 5

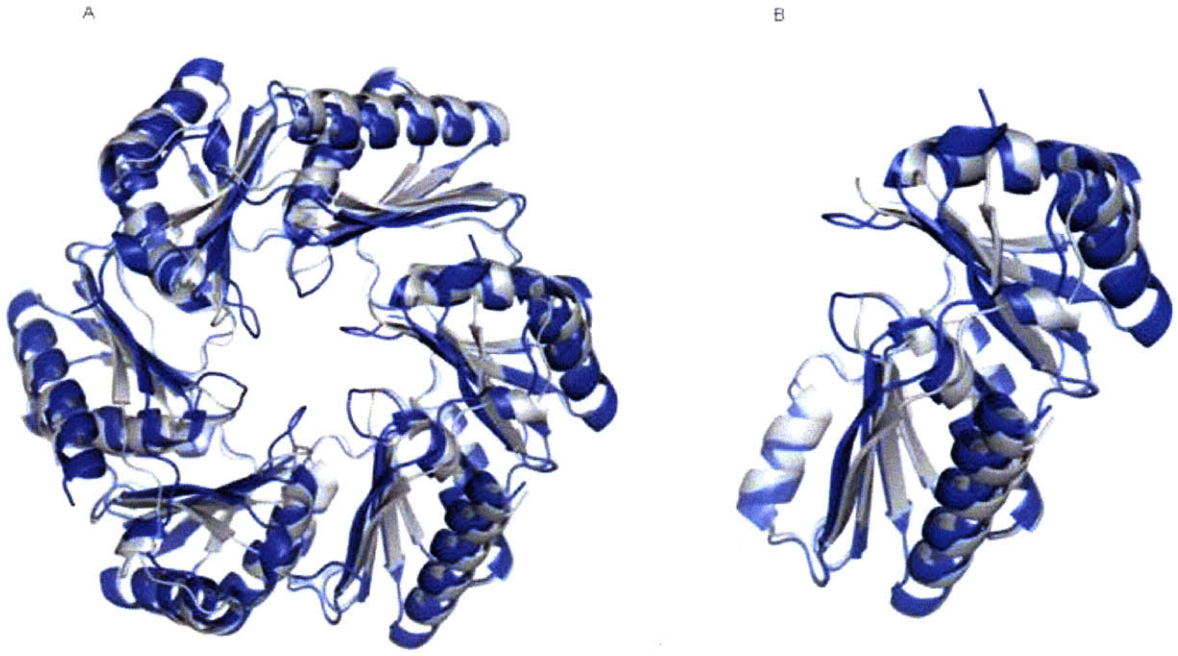


Figure 6

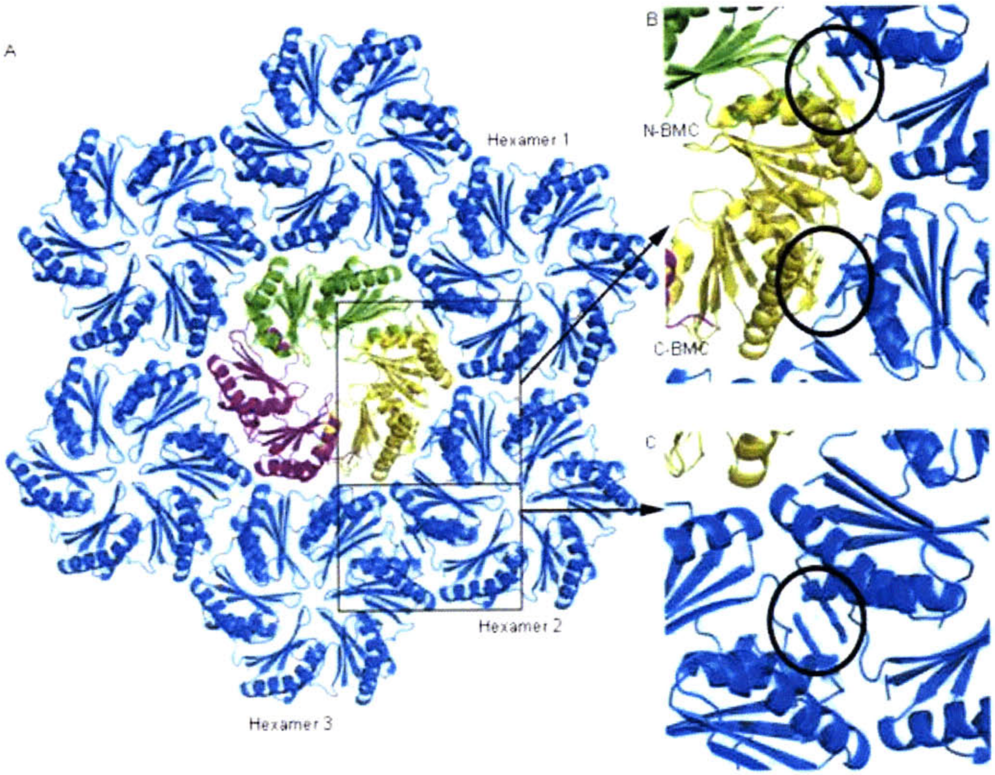


Table 1. Properties of protein-protein interfaces among bacterial microcompartment shell proteins

Interface	Total Buried Surface Area of the interface (Å^2)	Shape Complementarity (Sc)
CsoS1A-CsoS1A	1,450	0.68
CcmK2-CcmK2	x,y,z of chains A,B,C,D,E,F with: sym x,y,z+1 = 541 sym x+1,y,z+1 = 561 sym x+1,y,z = 488	x,y,z of chains A,B,C,D,E,F with: x,y,z+1 = 0.25 x+1,y,z+1 = 0.27 x+1,y,z = 0.41
CcmK4-CcmK4 (similar orientation)	878	0.58
CcmK4-CcmK4 (opposite orientation)	626	0.73
PduU-PduU	Chain A with sym chain B=397 Chain D with sym chain C=673	Chain A with sym chain B=0.60 Chain D with sym chain C=0.79
CsoS1D-CsoS1D (opposite orientation observed in R3 crystal form)	1,092	0.55
CsoS1D-CsoS1* (CsoS1 hexamers 1, 2 and 3 as shown in Figure 6)	N-BMC + CsoS1 hexamer1=728 N-BMC and C-BMC + CsoS1 hexamer2=1,559 C-BMC + CsoS1 hexamer3=606	N-BMC + CsoS1 hexamer1=0.58 N-BMC and C-BMC + CsoS1 hexamer2=0.48 C-BMC + CsoS1 hexamer3=0.45
CsoS1-CsoS1*	1,511	0.67
CcmK-CcmL*± (similar to CsoS1A-CsoS4A)	999 (model 1) 937 (model 2)	0.50 (model 1) 0.39 (model 2)

*Denotes computational model

±As reported in (11)

Buried surface area and surface complementary statistics were calculated in CCP4 using areaimol and SC, respectively. The pdb codes for the included structures are: 2ewh (CsoS1A), 2a1b (CcmK2), 2a18 (CcmK4) and 3cgi (PduU).

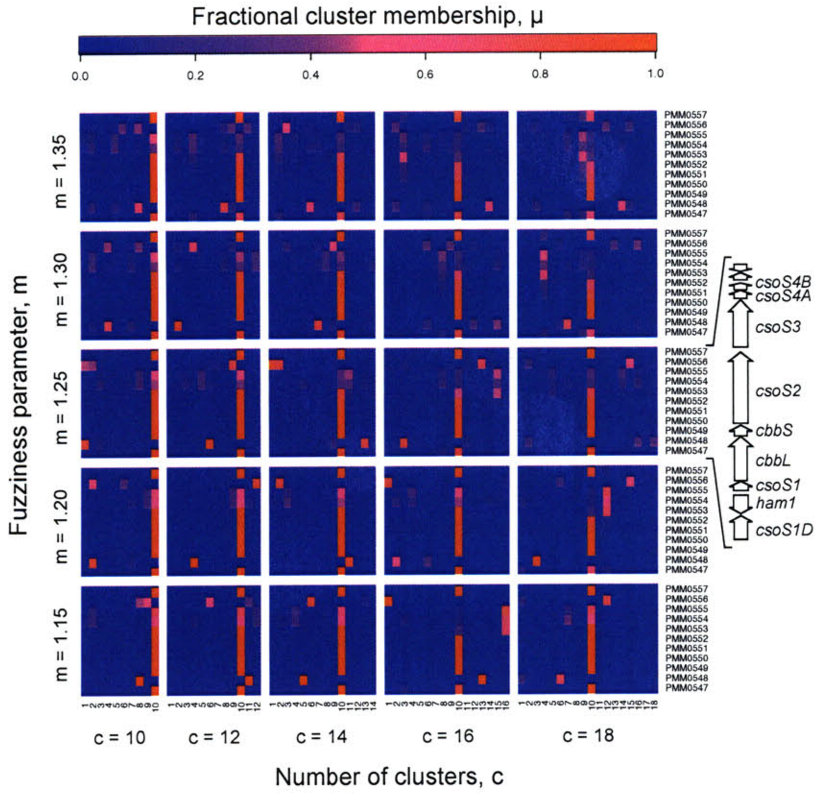


Figure S1. Heat map depicting the clustering patterns (see Methods) of gene expression from 25 time points over two diel cycles in synchronized *Prochlorococcus* MED4 cells for a range of clustering parameters (c, number of clusters; m, "fuzziness" of clustering). Clustering was run on the entire set of 1405 expressed, cycling genes; output is shown for only those genes in the immediate vicinity of the carboxysome operon. The heat map gives the output of one arbitrarily selected clustering run for each set of parameters. Genes *PMM0549-0555* constitute the known carboxysome operon in *Prochlorococcus* MED4. *PMM0547* (*csoS1D*) is separated from the main carboxysome operon by *PMM0548*, encoding a non-carboxysome-related Ham1 protein. Genes not labeled in the operon diagram encode hypothetical proteins. Coloring indicates the strength, μ, of each gene's association with each cluster under a given pair of clustering parameters.

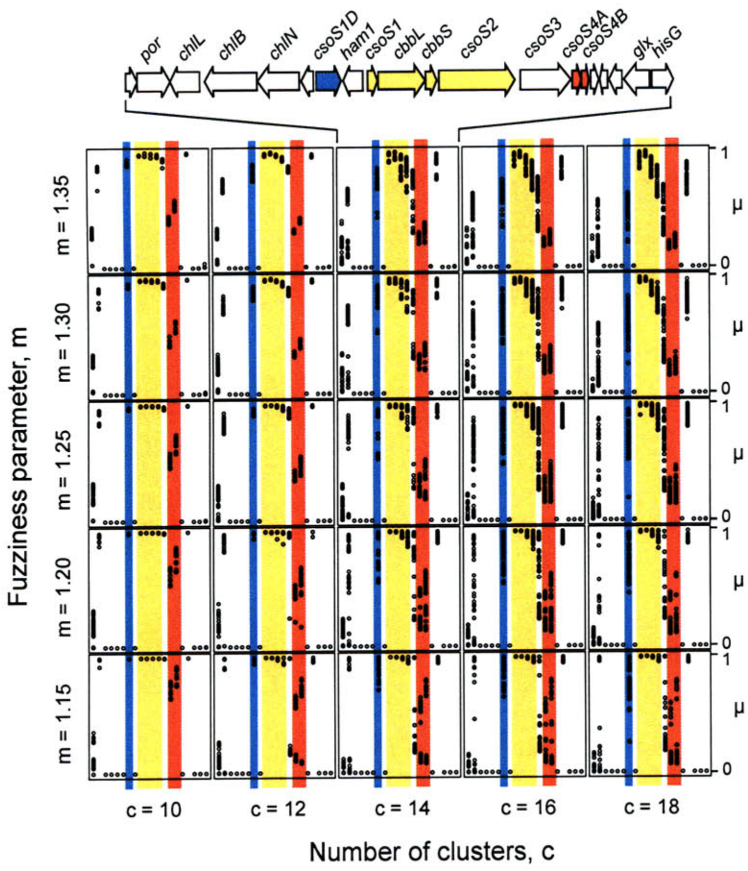


Figure S2. Fractional membership, µ, in the cluster containing known carboxysome genes *CsoS1*, *CbbL, CbbS,* and *CsoS2* for the 20 genes surrounding *CsoS1D*. Genes not labeled in the operon diagram encode hypothetical proteins. Each panel plots the output µ values for these 20 genes from 100 soft c-means clustering runs using the given pair of input parameters and diel expression data for all 1405 *Prochlorococcus* MED4 genes that oscillate in synchronized cells grown on a 24-h light-dark cycle (Zinser et al, submitted). Run-to-run variability, seen as the vertical smearing of each gene's cluster membership values, increases with cluster number and "fuzziness", but for each set of input parameters, the distribution of membership values for *CsoS1D* (highlighted in light blue) in the cluster containing the carboxysome operon (orange) is higher than that of shell proteins *CsoS4A* and *CsoS4B* (red).

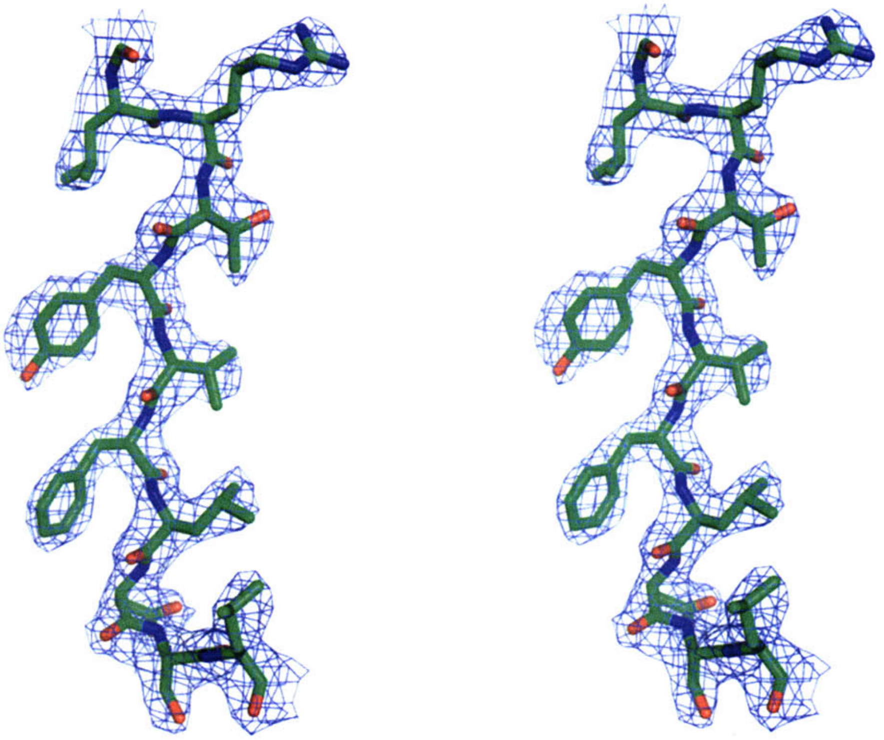


Figure S3. Stereo image showing residues 55 to 65 from Chain A of the refined model inside the experimental electron density map prepared in RESOLVE with density modification using solvent flattening and NCS averaging.

Table S1. Data collection and refinement statistics

		Crystal Form 1	Crystal Form 2
Data collection			
Space group		$P2_12_12_1$	R3:H
Cell dimensions			
a, *b*, *c* (Å)		122.410, 131.286, 121.762	117.163, 117.163, 101.464
α, β, γ (°)		90, 90, 90	90, 90, 120
	Peak	*Remote*	*Native*
Wavelength	0.9796	0.9768	1.0000
Resolution (Å)	50-2.5(2.59-2.50)	50-2.3(2.38-2.30)	50-2.2(2.28-2.20)
R_{merge}	11.0(53.3)	8.8(49.8)	5.2(66.2)
$I / \sigma I$	16.9(3.9)	16.7(2.9)	34.2(2.4)
Completeness (%)	100(100)	99.4(99.9)	99.7(99.7)
Redundancy	7.3(7.4)	4.4(4.4)	5.6(5.0)
Refinement			
Resolution (Å)		46.5-2.3	45.4-2.2
No. reflections		177663	25647
R_{work} / R_{free}		185/216	204/259
No. atoms			
Protein		9314	3082
Water		988	43
B-factors			
Protein		30.5	62.8
Water		37.0	40.0
R.m.s deviations			
Bond lengths (Å)		0.007	0.001
Bond angles (°)		1.048	0.454

*One crystal for each structure. *Values in parentheses are for highest-resolution shell.

Table S2. Summary of RMS deviations between CsoS1D monomer structures

	B	C	D	E	F	R3-A	R3-B
A	0.133	0.153	0.948	1.008	0.932	0.388	0.864
B		0.166	0.979	1.036	0.964	0.385	0.890
C			0.796	1.028	0.961	0.375	0.881
D				0.232	0.162	0.965	0.366
E					0.256	0.999	0.340
F						0.950	0.389
R3-A							0.871

All results are RMS deviation of alpha carbon atoms of 200 amino acids, results are reported in Å.