The ISME Journal (2016) 10, 2702–2714
© 2016 International Society for Microbial Ecology All rights reserved 1751-7362/16
www.nature.com/ismej

ORIGINAL ARTICLE

RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria

Kelly C Wrighton[1,7], Cindy J Castelle[2,7], Vanessa A Varaljay[1], Sriram Satagopan[1], Christopher T Brown[3], Michael J Wilkins[1,4], Brian C Thomas[2], Itai Sharon[2], Kenneth H Williams[5], F Robert Tabita[1] and Jillian F Banfield[2,6]

[1]*Department of Microbiology, The Ohio State University, Columbus, OH, USA;* [2]*Department of Earth and Planetary Science, UC Berkeley, Berkeley, CA, USA;* [3]*Department of Plant and Microbial Biology, UC Berkeley, Berkeley, CA, USA;* [4]*School of Earth Sciences, The Ohio State University, Columbus, OH, USA;* [5]*Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA and* [6]*Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA*

Metagenomic studies recently uncovered form II/III RubisCO genes, originally thought to only occur in archaea, from uncultivated bacteria of the candidate phyla radiation (CPR). There are no isolated CPR bacteria and these organisms are predicted to have limited metabolic capacities. Here we expand the known diversity of RubisCO from CPR lineages. We report a form of RubisCO, distantly similar to the archaeal form III RubisCO, in some CPR bacteria from the Parcubacteria (OD1), WS6 and Microgenomates (OP11) phyla. In addition, we significantly expand the Peregrinibacteria (PER) II/III RubisCO diversity and report the first II/III RubisCO sequences from the Microgenomates and WS6 phyla. To provide a metabolic context for these RubisCOs, we reconstructed near-complete (> 93%) PER genomes and the first closed genome for a WS6 bacterium, for which we propose the phylum name Dojkabacteria. Genomic and bioinformatic analyses suggest that the CPR RubisCOs function in a nucleoside pathway similar to that proposed in Archaea. Detection of form II/III RubisCO and nucleoside metabolism gene transcripts from a PER supports the operation of this pathway *in situ*. We demonstrate that the PER form II/III RubisCO is catalytically active, fixing CO_2 to physiologically complement phototrophic growth in a bacterial photoautotrophic RubisCO deletion strain. We propose that the identification of these RubisCOs across a radiation of obligately fermentative, small-celled organisms hints at a widespread, simple metabolic platform in which ribose may be a prominent currency.

The ISME Journal (2016) **10,** 2702–2714; doi:10.1038/ismej.2016.53; published online 3 May 2016

Introduction

The vast majority of the organisms in the environment have not been cultivated, obscuring our knowledge of their physiology. Recent metagenomic investigations have revealed the presence of a large diversity of uncultivated organisms in marine and terrestrial subsurface environments (Castelle *et al.*, 2013, 2015; Lloyd *et al.*, 2013; Baker *et al.*, 2015). For example, in a metagenomic study of a shallow alluvial aquifer, we determined that many of the uncultivated bacteria were associated with the candidate phyla radiation (CPR), a monophyletic group of at least 35 phyla that accounts for > 15% of all bacterial diversity (Brown *et al.*, 2015). Metagenomic sampling of almost 800 CPR bacteria suggested that members of this radiation consistently have small genomes with many metabolic limitations, including lack of an electron transport chain, no more than a partial tricarboxylic acid cycle and mostly incomplete nucleotide and amino-acid biosynthesis pathways. Using metabolic predictions from complete and near-complete genomes (Wrighton *et al.*, 2012), we inferred that members of the CPR are fermenters whose primary biogeochemical impact is primarily on subsurface organic carbon and hydrogen cycling (Wrighton *et al.*, 2014). However, many of the genes in these genomes remain poorly annotated and the metabolic platform of CPR bacteria remains uncertain. Further, no transcription and enzyme activities have been validated, hindering knowledge of the actual reactions catalyzed by these organisms.

Correspondence: JF Banfield, Department of Earth and Planetary Sciences, and Department of Environmental Science, Policy, and Man, University of California at Berkeley, 336 Hilgard Hall, Berkeley, CA 94720, USA.
E-mail: jbanfield@berkeley.edu

[7]These authors contributed equally to this work.

Received 25 August 2015; revised 29 February 2016; accepted 4 March 2016; published online 3 May 2016

The incorporation of CO_2 into organic carbon is a critical step in the carbon cycle. Ribulose-1,5-bisphophate carboxylase-oxygenase (RubisCO) is the most abundant enzyme on earth and is integral to the fixation of carbon dioxide. Currently four forms of RubisCO can fix atmospheric carbon dioxide (I, II, II/III and III) (Tabita *et al.*, 2007; 2008). Forms I and II are used by plants, algae and some chemoautotrophic and phototrophic bacteria for CO_2 fixation during primary production via the Calvin Benson Bassham (CBB) cycle. Forms III and II/III, found primarily in archaea, enable light-independent CO_2 incorporation into sugars derived from nucleotides like adenosine monophosphate (AMP) (Sato *et al.*, 2007; Tabita *et al.*, 2007). Form II/III RubisCOs, which were primarily known in methanogens, have greater overall similarity to bacterial Form II RubisCO but have specific residues, structural and catalytic features that more closely resemble archaeal form III (Alonso *et al.*, 2009).

Previously, we reported the first evidence of form II/III RubisCO genes in partial genomes from the Peregrinibacteria (PER) phylum in the CPR radiation (Wrighton *et al.*, 2012). Additionally, three bacterial form II/III RubisCOs were recovered from genomes of members of the SR1 phylum that also are inferred to be fermentative (Wrighton *et al.*, 2012; Campbell *et al.*, 2013; Kantor *et al.*, 2013). The presence of these annotated RubisCO genes in small genomes that house few ancillary genes suggests that RubisCO may have an important role in the metabolism of some of these bacteria.

To more comprehensively explore the diversity and potential role of CPR RubisCOs, we reproduced the prior experiment from which we first identified bacterial form II/III RubisCO sequences (Wrighton *et al.*, 2012). Size filtration was used to increase the relative abundance of the small-celled CPR organisms in samples used for sequencing (Luef *et al.*, 2015). Here we investigate CPR genomes that encode RubisCO genes, report a new RubisCO sequence type diverged from Archaeal form III and propose a broader metabolic context in which these genes occur. By detecting *in situ* transcripts and through directly cloning one CPR form II/III RubisCO sequence into an expression host, we suggest the potential functionality of this bacterial enzyme. This study represents the first enzymatic activity demonstrated from these broadly distributed but metabolically enigmatic candidate phyla bacteria.

Materials and methods

Groundwater field experiment and sample collection
The field experiment was carried out between 25 August and 12 December 2011 at the US Department of Energy Rifle Integrated Field Research Challenge site adjacent to the Colorado River, CO, USA as recently described (Brown *et al.*, 2015; Castelle *et al.*, 2015; Luef *et al.*, 2015). In brief, biostimulation experiments were performed by adding approximately 15 mM acetate to the groundwater over the course of 72 days to maintain an average *in situ* concentration of ~ 3 mM acetate during peak stimulation. Six microbial community samples (A–F) were collected over 93 days, including 3 days prior to acetate stimulation and after acetate addition ceased (amendment ceased on day 72). Ferrous iron and sulfide concentrations were analyzed immediately after sampling using the HACH phenanthroline assay and methylene blue sulfide reagent kit, respectively (HACH, Loveland, CO, USA). Acetate and sulfate concentrations were determined using a Dionex ICS-2100 ion chromatograph (Sunnyvale, CA, USA) equipped with an AS-18 guard and analytical column (Williams *et al.*, 2011). Microbial cells from pumped groundwater that passed through a 1.2-μm prefilter (Pall, Port Washington, NY, USA) were retained on 0.2- and 0.1-μm filters (Pall). Filters were flash-frozen in liquid nitrogen immediately upon collection.

Nucleic acid extraction and sequencing
Approximately a 1.5-g sample from each frozen filter was used for genomic DNA extraction. DNA was extracted using the PowerSoil DNA Isolation Kit (MoBio Laboratories, Inc., Carlsbad, CA, USA). Illumina HiSeq 2000 2 × 150 paired-end sequencing was conducted by the Joint Genome Institute on extracted DNA. Sequencing amounts for samples A–F from 0.1- and 0.2-μm filters ranged from 9.5 to 40.7 Gbp, as described in Castelle *et al.* (2015). RNA was extracted using Invitrogen TRIzol Reagent (Carlsbad, CA, USA) followed by genomic DNA removal and cleaning using Qiagen RNase-Free DNase Set Kit (Valencia, CA, USA) and Qiagen Mini RNeasy Kit (Valencia, CA, USA). An Agilent 2100 Bioanalyzer (Santa Clara, CA, USA) was used to assess the integrity of the RNA samples. Only RNA samples having RNA Integrity Number between 8 and 10 were used. The Applied Biosystems SOLiD Total RNA-Seq Kit (Invitrogen/LifeTechnologies, Carlsbad, CA, USA) generated the cDNA template library. The SOLiD EZ Bead system was used to perform emulsion clonal bead amplification to generate bead templates for SOLiD platform sequencing. Samples were sequenced on the 5500XL SOLiD platform (LifeTechnologies). The 75-bp sequences produced were mapped in color space using the SOLiD LifeScope software version 2.5 (Life Techno logies/Invitrogen) using the default parameters against the reference genome set of 2 302 715 separate gene FASTA entries. Corresponding GTF files were built to record the position of each gene on the set of artificial chromosomes.

Genome assembly, binning and functional annotations
Methods for assembly, binning and annotation are described from methods described in detail in prior publications (Wrighton *et al.*, 2012;

Castelle *et al*., 2015). Reads were quality trimmed using Sickle (https://github.com/najoshi/sickle) with default settings. Trimmed paired-end reads were assembled using IDBA_UD (Peng *et al*., 2012) with default parameters. Genes on scaffolds >5 kb in length were predicted using Prodigal (Hyatt *et al*., 2012) with the metagenome procedure (-p meta), and then USEARCH (–ublast) (Edgar, 2010) was used to search protein sequences against UniRef90 (Suzek *et al*., 2007), KEGG (Kanehisa and Goto, 2000; Kanehisa *et al*., 2012) and an in-house database composed of open reading frames from candidate phyla genomes. The in-house database includes previously published genomes (Wrighton *et al*., 2012; Castelle *et al*., 2013; Hug *et al*., 2013; Kantor *et al*., 2013; Wrighton *et al*., 2014) and genomes from ongoing work and primarily aiding in identifying putative CPR scaffolds. For each scaffold, we determined the GC content, coverage and profile of phylogenetic affiliation based on the best hit for each gene against the Uniref90 (Suzek *et al*., 2007) database. Scaffolds were binned to specific organisms by using coverage across the samples, phylogenetic identity and GC content, both automatically with the ABAWACA algorithm and manually using ggKbase (http://ggkbase.berkeley.edu/). ABAWACA is an algorithm that generates preliminary genome bins based on different characteristics of assembled scaffolds (https://github.com/CK7/abawaca). Six genome bins generated by ABAWACA were manually inspected within ggKbase. Binning purity was confirmed using an Emergent Self-Organizing Map (Dick *et al*., 2009). Our primary method for assessing genome completeness was based on the presence or absence of orthologous genes representing a core gene set that are widely conserved as single-copy genes among bacterial candidate phyla, but we also assessed global markers. All single copy and metabolic analyses reported here and summarized in the Supplementary Figures are reported in ggKbase (FASTA files can be accessed via links provided in the Supplementary Figures). Manual assembly curation improved some genome bins by extending and joining scaffolds, removing local scaffolding errors and, in some cases, bringing in small previously unbinned fragments. One genome was curated to closure and completion.

Inventory of the RubisCO genes

All putative RubisCO genes on genome fragments assembled from the 12 separate groundwater filter samples were recovered first by BLAST search (Altschul *et al*., 1990) against an in-house database that included candidate phyla sequences from prior studies (Wrighton *et al*., 2012; Campbell *et al*., 2013; Kantor *et al*., 2013; Castelle *et al*., 2015). The taxonomic affiliation of these putative RubisCO CPR scaffolds was confirmed based on phylogenetic analysis of genes included on the same genome fragment as the RubisCO or in the same genome bin. We then used these RubisCO sequences to recruit additional CPR sequences from currently unreported groundwater and sediment genomic data sets from the same field site.

16S rRNA gene recovery and phylogenetic analyses

16S rRNA gene sequences were identified in assembled scaffolds based on Hidden Markov Model searches using the cmsearch program from the Infernal package (cmsearch –hmmonly –acc –noali –T –1) (Nawrocki *et al*., 2009). This protocol is described in detail in Brown *et al*. (2015). Searches were conducted using the manually curated structural alignment of the 16S rRNA provided with SSU-Align (Nawrocki, 2009). Sequences corresponding with CPR genomes were aligned using SSU-Align along with the best hits of these sequences in the Silva database (v1.2.11) (Quast *et al*., 2013), a curated set of reference sequences associated with candidate phyla of interest, and a set of archaea sequences to serve as a phylogenetic root. Sequences with $\geqslant 800$ bp aligned in the 1582-bp alignment were used to infer a maximum-likelihood phylogeny using RAxML (Stamatakis, 2014) with the GTRCAT model of evolution selection via ProTest and 100 bootstrap re-samplings.

Protein phylogenetic analyses and modeling

For single protein trees (for example, RubisCO large subunit), the phylogenetic pipeline was performed as previously described (Wrighton *et al*., 2012). Briefly, protein sequences were aligned using MUSCLE version 3.8.31 (Edgar, 2004) with default settings. Problematic regions of the alignment were removed using very liberal curation standards with the program GBlocks (Talavera and Castresana, 2007). Alignments with and without GBlocks were confirmed manually. Best models of amino-acid substitution for each protein alignment were estimated using ProtTest3 (Abascal *et al*., 2005). Phylogenetic trees were generated using RAxML with the PROTCAT setting for the rate model and the best model of amino-acid substitution specified by ProtTest. Nodal support was estimated based on 100 bootstrap replications using the rapid bootstrapping option implemented in RAxML. Three-dimensional structure predictions were generated by SWISS-MODEL (an automated protein homology-modeling server) and i-tasser based on protein alignment and secondary structure prediction. The alignment mode was based on a user-defined target-template alignment. Conservation of key catalytic residues and the secondary structure for each model was confirmed by manual inspection.

Plasmids and cloning

Rifle metagenomic DNA was used as the template to amplify the exact coding regions for the RubisCO genes from PER genome PER_GWF2_38_29 (PER-2), the OD-1 genome GWE1_OD1_1_1_39 and WS6 genomes GWC1_1_418 and GWC1_33_22_1_93. Primer sequences used for RubisCO amplification are included (Supplementary Table S1). PCR was performed using the PrimeStar GXL (Clontech, Mountain view, CA, USA) high-fidelity polymerase. PCR products for all genes were cloned between NdeI and BamHI sites in plasmid pet11a (Novagen/EMD Millipore, Billerica, MA, USA), which does not incorporate an N-terminal hexa-histadine tag. For the PER, the PCR product was also cloned between KpnI and BamHI sites in plasmid p80, a high copy plasmid 3716-based vector (a derivative of pLO11 obtained from Oliver Lenz and Bärbel Friedrich, Berlin, Germany), which contained the *Rhodospirillum rubrum CbbR/cbbM* promoter region, for expression *in vivo*. This promoter has been used previously in the broad-host range vector pRPS-MCS3 to drive the expression of RubisCO genes to complement RubisCO deletion strain *Rhodobacter capsulatus* strain SB I/II (Smith and Tabita, 2003). Constructs for the coding regions of archaeal form III *Archaeoglobus fulgidus* RbcL2 (Kreel and Tabita, 2007), archaeal form II/III *Methanococcoides burtonii rbcL* and bacterial form II *R. rubrum cbbM* (Falcone and Tabita, 1993) were used for comparison and served as positive controls. The sequences of all constructs were verified at the Plant-Microbe Genomics Facility at The Ohio State University.

Growth and culture conditions

To test for *in vitro* activity, RubisCO proteins from pet11a constructs were expressed in *Escherichia coli* strain Rosetta 2 (DE3) pLysS (Novagen), which uses the pRARE plasmid to accommodate expression of rare codons such as those from candidate phyla and archaeal RubisCO sequences. Cells were grown in 50 ml lysogeny broth supplemented with 150 μg ml^{-1} ampicillin and 12.5 μg ml^{-1} chloramphenicol in 125-ml flasks. After reaching an optical density of 0.6–0.8 (at 600 nm), RubisCO gene expression was induced by adding IPTG to a final concentration of 1 mM. After 5–16 h of shaking at room temperature, cells were harvested, flash frozen and stored at − 80 °C prior to analysis. To ensure that candidate phyla RubisCOs were exposed to minimal amounts of oxygen, cell lysis and assay conditions were performed under strict anaerobic conditions. Using an anaerobic chamber, cell pellets were resuspended in 50 mM bicine-NaOH, 10 mM $MgCl_2$, at pH 8.0, supplemented with 2 mM $NaHCO_3$ and 1 mM dithiothreitol; crude protein lysates were obtained via bead beating with 0.1-mm beads for 3–4-min intervals in tightly capped tubes.

To test for *in vitro* activity from complemented autotrophic cultures, p80::PER, p80::*R. rubrum* and pRPS-MCS3::*A. fulgidus* RubisCO constructs were transformed into *E. coli* strain S17-1 (ATCC47055) and mobilized into *R. capsulatus* SB I/II$^-$ via biparental matings (Smith and Tabita, 2003; Satagopan *et al.*, 2009, 2014). Plasmid-complemented strain SB I/II$^-$ colonies were first selected on peptone yeast extract plates followed by streaking onto Ormerod's minimal medium (Ormerod *et al.*, 1961) agar plates with no added organic carbon. To test for photoautotrophic growth complementation, plates were incubated in illuminated anaerobic jars, which were periodically flushed with a gas mixture of 5% CO_2/H_2 as previously described (Satagopan *et al.*, 2009, 2014). Anaerobic CO_2-dependent photoautotrophic growth was obtained using Ormerod's minimal medium agar plates in a sealed jar flushed with a 5% CO_2/95% H_2 gas mixture. Colonies on minimal plates under these conditions were used to inoculate liquid photoautotrophic cultures bubbled continuously with 5% CO_2/H_2 for growth analysis; cells were harvested anaerobically for RubisCO activity assays, as described above.

In vitro RubisCO assays

RubisCO assays with activated enzyme preparations were performed according to Satagopan *et al.* (2014) with modifications to ensure anaerobiosis. All crude protein lysates, reagents and supplies were prepared in an anaerobic chamber and crimped in sealed glass vials prior to conducting the assays in a hood using gas-tight Hamilton syringes (Hamilton Robotics, Reno, NV, USA). Assay reaction mixtures contained 50 mM $NaHCO_3$ (~2 μCi of NaH ^{14}CO3), 10 mM $MgCl_2$ and 0.8 mM ribulose 1,5 bisphosphate (RuBP) in 50 mM bicine-NaOH buffer, pH 8.0. RubisCO activity is initiated by addition of its physiological substrate, RuBP. Reactions were initiated with the addition of crude protein lysate to its physiological substrate, RuBP, at 30 °C and incubated for 5–30 min. Negative control reactions consisted of no RuBP or no crude protein lysate added. Carboxylase activities were determined following the equimolar conversion of ^{14}C from $NaHCO_3$ into acid-stable 3-phosphoglycerate (3-PGA), as measured via radioactive scintillation counting. Protein was determined via the Bradford method and specific activities (nmol CO_2 fixed per min per mg protein) were calculated consistent with prior publications (Satagopan *et al.*, 2009).

Results and discussion

Recovery of CPR genomes that contain RubisCO

We reconstructed CPR genomes from sequence data sets for samples collected over 93 days (Brown *et al.*, 2015; Castelle *et al.*, 2015). As previously reported, bacteria from CPR phyla WS6, OD1 and OP11 were enriched (> 75% 16S rRNA relative abundance) on the 0.1-μm filter, and some members were shown to have ultra-small cells (median cell volume of 0.009 ± 0.002 μm^3) (Luef *et al.*, 2015). Multiple genomes belonging to the CPR phylum PER were recovered from all six 0.2-μm filter data sets, suggesting that these cells may be larger in size than other CPR bacteria (for example, OD1, WS6) enriched on the 0.1-µm filters.

Taxonomic identifications of the CPR organisms from which we recovered RubisCO genes are provided in Supplementary Figures S1 and S2. The emergent self-organizing map clustered genome fragments based on their tetranucleotide sequence composition (Supplementary Figure S3) and validated 16 unique CPR genomes that ranged in estimated completion from near-complete (>94%) to complete (Supplementary Figure S4). Here we report RubisCO sequences from five genomic bins assigned to the PER phylum, two to the Parcubacteria (OD1), five to the Microgenomates (OP11) and six to the WS6 phylum (Table 1). We also report RubisCO from additional scaffolds assigned to these CPR phyla.

The five WS6 genomes reported here span multiple taxonomic classes (Supplementary Figure S1 and Supplementary Table S2) (Brown *et al.*, 2015). The manually curated and closed WS6 genome is 0.896 Mbp in length and is the first complete genome reconstructed for a member of this lineage. The mapped read file confirming the assembly and the annotations can be accessed via ggKbase (see Methods section). Given the level of genome completion and the breadth of phylogenetic diversity sampled here (Konstantinidis and Rosselló-Móra, 2015), we propose the name *Candidatus* Dojkabacteria for the WS6 phylum to honor the memory of Michael A Dojka, who first reported this lineage based on 16S rRNA gene sequence data (Dojka *et al.*, 2000)

We recovered five distinct PER genomes that we estimated to be >93% complete (Table 1; Supplementary Figure S4). This data set augments the three partial PER genomes analyzed from samples collected from the same aquifer 5–7 days after acetate stimulation, two of which (ACD51, ACD65) encoded RubisCO genes (Wrighton *et al.*, 2012). Phylogenetic analyses that include these and other recently reported sequences (Rinke *et al.*, 2014; Brown *et al.*, 2015) support the identification of this lineage as a separate phylum (*Candidatus* Peregrinibacteria). Based on sequence divergence, the five Peregrinibacterial genomes reported here represent multiple taxonomic classes in this newly designated phylum.

CPR genomes contain novel and diverse RubisCO sequences

Fifty-one new RubisCO genes were identified from these CPR genomes. Some sequences, for example, from the WS6-1 and OD1-1 genomes, were independently reconstructed from multiple data sets. Ultimately, 18 representative unique sequences were manually confirmed by read mapping to verify their accuracy (Figure 1). These sequences were most similar to previously reported CPR or archaeal RubisCO sequences (and not bacterial RubiscCO sequences) (Supplementary Table S3) and the majority harbor the key active-site residues (Figure 2).

Phylogenetic analyses revealed that the seven bacterial CPR RubisCO sequences reported here are members of the previously defined archaeal intermediate form II/III RubisCOs (Figure 1) (Tabita *et al.*, 2007, 2008). Recent metagenomic studies reported the form II/III RubisCOs from PER and SR1 uncultivated bacterial lineages (Wrighton *et al.*, 2012; Campbell *et al.*, 2013; Kantor *et al.*, 2013). Although the overall sequence similarity is most closely related to type II, protein sequence alignment

Table 1 Summary of CPR genomes reported in this study

Genome ID	*ID ggKbase*	*Taxonomic affiliation*	*Completeness (%)*	*Number of scaffolds*	*Relative GC content (%)*	*Bin length (Mbp)*	*No. of protein-coding genes*	*RubisCo form*
PER-1	PER_GWA2_38_35	Peregrinibacteria	95	5	38.2	1.11	1020	II/III
PER-2*	PER_GWF2_38_29	Peregrinibacteria	93	23	38	1.23	1164	II/III
PER-3	PER_GWF2_39_17	Peregrinibacteria	100	15	38.8	1.31	1129	II/III
PER-4	PER_GWC2_39_14	Peregrinibacteria	98	21	38.5	1.32	1247	II/III
PER-5	PER_GWF2_43_17	Peregrinibacteria	100	16	43.1	1.20	1136	II/III
WS6-1	WS6_GWF2_39_15	Dojkabacteria	100 (closed)	1	38.7	0.896	891	II/III
WS6-2	WS6_GWC1_33_20	Dojkabacteria	98	23	32.7	0.65	666	III-like
WS6-3	GWF1_WS6_37_7	Dojkabacteria	100	18	36.7	1.22	1153	III-like
WS6-4	WS6_GWE2_33_157	Dojkabacteria	100	29	32.8	0.61	628	III-like
WS6-5	WS6_GWE1_33_547	Dojkabacteria	95	34	32.5	0.61	649	III-like
WS6-6	WS6_GWF1_33_233	Dojkabacteria	98	23	32.7	0.61	640	III-like
OD1-1	OD1_GWE2_42_8	Parcubacteria	93	34	42.5	0.743	781	III-like
OD1-2	OD1_RIFOXA1_OD1_43_6	Parcubacteria	84	183	31.7	0.67	850	III-like
OP11-1	OP11_GWA2_42_18	Microgenomates	95	53	42.3	1.24	1324	III-like
OP11-2	GWA2_OP11_43_14	Microgenomates	100	24	43.2	1.65	1699	III-like
OP11-3	RBG_16_OP11_37_8	Microgenomates	95	84	37.5	1.31	1506	III-like

Abbreviation: CPR, candidate phyla radiation. Bolded text includes manually curated and confirmed near-complete or complete genomes, while asterisk (*) denotes RubisCO that was cloned and functionally confirmed.

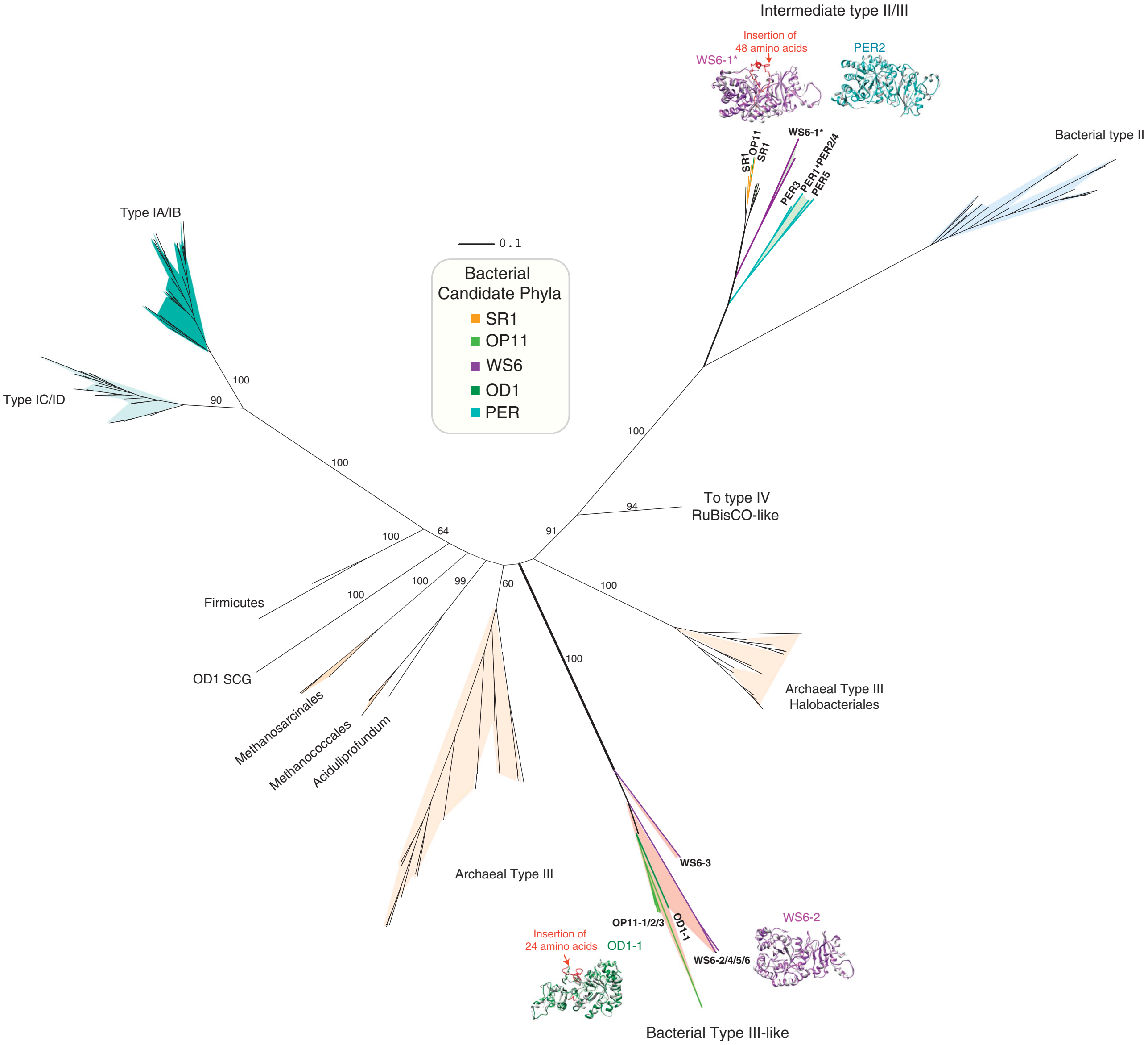
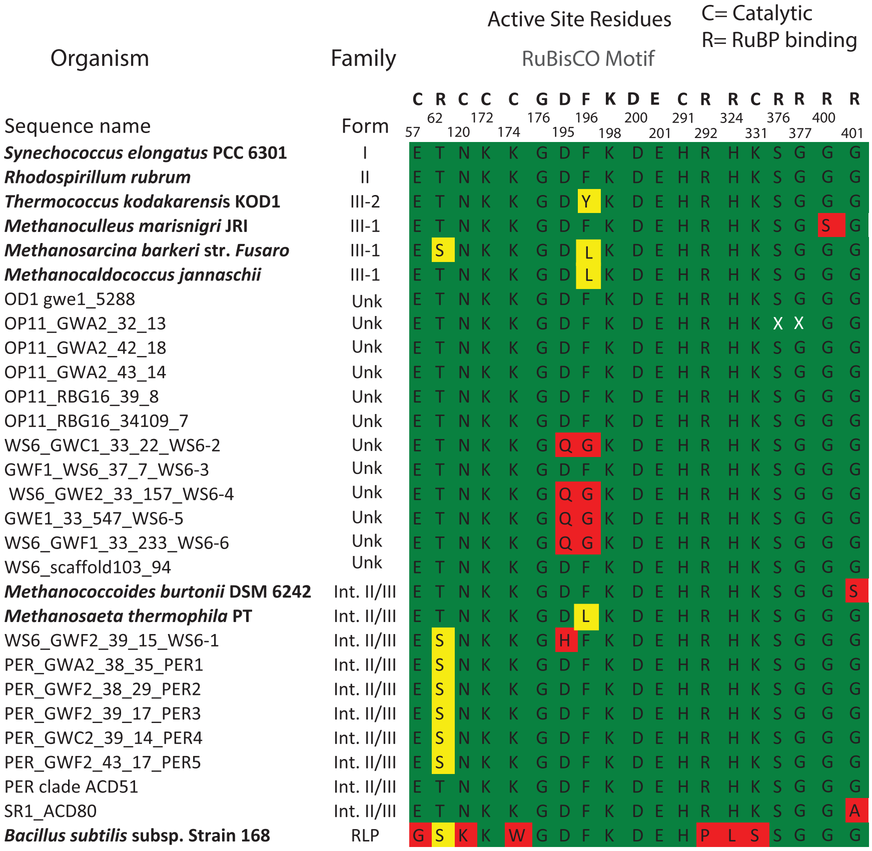


Figure 1 Maximum likelihood phylogenetic tree constructed for the RubisCO large subunit. Key bootstrap values >50 are shown based on 100 resamplings. Protein models of the PER and Dojkabacteria (WS6) type II/III RubisCOs and the Dojkabacteria and Parcubacterium (OD1) type III-like RubisCOs are included. The asterisk (*) refers to genomes that are complete or nearly complete and hand-curated. The additional inserts of the form II/III RubisCO from the closed Dojkabacteria genome (48 residues) or of the form III-like from the Parcubacterium genome (24 residues) are highlighted in red on the protein models (Expanded view shown in Supplementary Figure S6).

confirmed the presence of form III-specific residues (Supplementary Figure S5), validating the form II/III phylogenetic placement of these bacterial sequences. Our results expand the bacterial members of this form by four additional PER sequences, add the first three Dojkabacteria sequences and add a single-cell unpublished Microgenomates (OP11) sequence. Originally, two form II/III RubisCO subgroups were described based on the absence (PER) or the presence (SR1 and archaea) of a 29 amino-acid N-terminal insert sequence (Wrighton *et al.*, 2012; Campbell *et al.*, 2013; Supplementary Figures S5 and S6). Here our additional sampling of the PER RubisCO confirms this earlier discovery, as all PER genomes to date lack the insertion. The Microgenomates and two of the Dojkabacteria RubisCOs reported here have the 29 amino-acid insertion found in the SR1 and archaeal II/III form, while the closed WS6 genome (WS6-1) encodes a RubisCO with a novel 48 amino-acid insertion (Figure 1, Supplementary Figures S5 and S6). Both the 29 and 48 amino-acid insertion sequences are located in the same position as previously identified insertions. Similar to the methanogen sequences, the insertion is sandwiched between catalytic sites of the enzyme (Supplementary Figure S5), yet impact of these insertions on enzyme biochemical properties is currently unknown.



Active Site Residues

RuBisCO Motif

C= Catalytic
R= RuBP binding

Organism / Family

Sequence name	Form	C 57	R 62	C 120	C 172	C 174	G 176	D 195	F 196	K 198	D 200	E 201	C 291	R 292	R 324	C 331	R 376	R 377	R 400	R 401
***Synechococcus elongatus* PCC 6301**	I	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
Rhodospirillum rubrum	II	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
***Thermococcus kodakarensis* KOD1**	III-2	E	T	N	K	K	G	D	Y	K	D	E	H	R	H	K	S	G	G	G
***Methanoculleus marisnigri* JRI**	III-1	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	S	G
Methanosarcina barkeri* str. *Fusaro	III-1	E	S	N	K	K	G	D	L	K	D	E	H	R	H	K	S	G	G	G
Methanocaldococcus jannaschii	III-1	E	T	N	K	K	G	D	L	K	D	E	H	R	H	K	S	G	G	G
OD1 gwe1_5288	Unk	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
OP11_GWA2_32_13	Unk	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	X	X	G	G
OP11_GWA2_42_18	Unk	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
OP11_GWA2_43_14	Unk	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
OP11_RBG16_39_8	Unk	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
OP11_RBG16_34109_7	Unk	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
WS6_GWC1_33_22_WS6-2	Unk	E	T	N	K	K	G	Q	G	K	D	E	H	R	H	K	S	G	G	G
GWF1_WS6_37_7_WS6-3	Unk	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
WS6_GWE2_33_157_WS6-4	Unk	E	T	N	K	K	G	Q	G	K	D	E	H	R	H	K	S	G	G	G
GWE1_33_547_WS6-5	Unk	E	T	N	K	K	G	Q	G	K	D	E	H	R	H	K	S	G	G	G
WS6_GWF1_33_233_WS6-6	Unk	E	T	N	K	K	G	Q	G	K	D	E	H	R	H	K	S	G	G	G
WS6_scaffold103_94	Unk	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
***Methanococcoides burtonii* DSM 6242**	Int. II/III	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	S
***Methanosaeta thermophila* PT**	Int. II/III	E	T	N	K	K	G	D	L	K	D	E	H	R	H	K	S	G	G	G
WS6_GWF2_39_15_WS6-1	Int. II/III	E	S	N	K	K	G	H	F	K	D	E	H	R	H	K	S	G	G	G
PER_GWA2_38_35_PER1	Int. II/III	E	S	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
PER_GWF2_38_29_PER2	Int. II/III	E	S	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
PER_GWF2_39_17_PER3	Int. II/III	E	S	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
PER_GWC2_39_14_PER4	Int. II/III	E	S	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
PER_GWF2_43_17_PER5	Int. II/III	E	S	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
PER clade ACD51	Int. II/III	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	G
SR1_ACD80	Int. II/III	E	T	N	K	K	G	D	F	K	D	E	H	R	H	K	S	G	G	A
***Bacillus subtilis* subsp. Strain 168**	RLP	G	S	K	K	W	G	D	F	K	D	E	H	P	L	S	S	G	G	G

Figure 2 Conservation of RubisCO active-site residues. Alignment of active-site residues from representative RubisCO and RLP sequences as noted previously by Tabita *et al*., 2007, along with equivalent residues from the new sequences described in this study. Residues are noted in single-letter IUPAC code. Coloring scheme is based on the identities of residues relative to the *Synechococcus elongatus* PCC 6301 reference sequence. Amino-acid numbering according to *Synechococcus elongatus* PCC 6301 reference sequence have been indicated. Green shading refers to conserved residues, yellow indicates semi-conserved residues and red represents non-conserved residues. Catalytic (C) and RuBP-binding (R) residues are indicated on the top; X denotes missing residue. The RubisCO motif refers to a stretch of conserved residues that are proximal in the primary structure and is identified by consensus residue identities shaded gray. Residues K, D and E from this motif participate in catalysis.

In addition to the form II/III sequences, nine of our newly recovered bacterial sequences form a distinct and strongly supported monophyletic group that branches deeply from, yet is most similar to, the form III RubisCO sequences found in archaea. This current CPR clade includes RubisCO sequences from Dojkabacteria (3), Parcubacterium (1) and Microgenomates (5). Although the extent of sequence divergence between these bacterial sequences and the archaeal form III sequences is considerable, additional taxon sampling is needed to resolve this clade as completely distinct RubisCO form (Figure 1, Supplementary Figure S7). In our analyses, we have identified other bacterial genomes that contain homologs to archaeal form III sequences in public databases (for example, from a partial single-cell Parcubacteria genome and in two Firmicutes draft genomes; Supplementary Figure S7). Notably, these bacterial form III are not monophyletic with the CPR RubisCOs described here but clade with the canonical archaeal form III.

The parcubacterial sequence from the OD1-1 genome contains two, an 8 and 24 amino acid, insertions between alpha-helix 6 and beta-sheet 7 (Figure 1, Supplementary Figures S5 and S6). The WS6 genome containing divergent form III also contains the eight amino-acid insertion. To further confirm analyses indicating that these insertions were not due to sequencing or assembly error, we designed specific PCR primers for the gene encoding for the large RubisCO subunit and confirmed the insertions via cloning and sequencing. Prior sequence analyses demonstrated that this region was involved in interactions with the neighboring subunit (Satagopan *et al*., 2014); however, these insertions have not been previously identified in other RubisCO sequences.

CPR RubisCO is inferred to support a heterotrophic nucleoside pathway

Previously studied archaeal form III (Finn and Tabita, 2004; Sato *et al*., 2007; Estelmann *et al*., 2011) and form II/III (Tabita *et al*., 2008; Alonso *et al*., 2009) RubisCOs are not associated with the CBB pathway but instead function in a nucleosidesalvaging pathway. Consistent with the archaeal genomes that contain RubisCO, all bacterial CPR genomes containing RubisCOs sampled to date lack the gene encoding phosphoribulokinase, a key gene of the CBB pathway (Wrighton *et al.*, 2012; Campbell *et al.*, 2013; Kantor *et al.*, 2013). However, we note these genomes contain many genes that are too novel to be identified based on sequence similarity and thus we cannot eliminate the possibility that they contain new enzymes or divergent phosphoribulokinase homologs (Ashida *et al.*, 2008). Consistent with our hypothesis that these RubisCO function in archaeal nucleoside pathway, outside other CPR sequences, these RubisCO proteins have the closest homologs to similar archaeal and not bacterial RubisCOs (Supplementary Table S3). Thus we suggest RubisCOs studied here also function in a nucleoside/nucleotide metabolism pathway analogous to archaea.

Rather than using phosphoribulokinase, there are two proposed pathways to generate RuBP, the substrate for RubisCO in Archaea. In the first pathway, identified in *Thermococcus kodakarensis,* nucleosides are converted into 3-PGA. The first step in this short pathway dephosphorylates the nucleotide (AMP, CMP, UMP) to release the base (for example, adenine), yielding ribose-1,5-bisphosphate, which an isomerase converts to ribulose-1,5-bisphosphate (RuBP) (that is, AMP → ribose 1,5 bisphosphate → RuBP). From RuBP and CO_2, the type III or II/III archaeal RubisCO generate two molecules of 3-PGA, a central intermediate of core carbon metabolism (Aono *et al.*, 2012).

All CPR genomes with RubisCO reported here and studied previously possesses the two archaeal AMP pathway homologs (Figure 4), with the best hits to the AMP phosphorylase and the isomerase outside the CPR are largely from archaeal genomes (Supplementary Table S3). For both proteins in this pathway, the most similar sequences outside the CPR are to homologs in archaeal genomes, often times to methanogens (PER) or to uncultivated archaeal lineages such as the DPANN (Castelle *et al.*, 2015). We used an amino-acid alignment and protein modeling to confirm that the CPR isomerase has all residues for catalytic activity and is specific for ribose 1,5 bisphophate and not another substrate, for example, 5-methylthioribose-1-phosphate isomerase (Supplementary Figures S8 and S9). Based on the genomic evidence to date, we propose that the CPR bacteria likely use a RubisCO pathway analogous to that of *T. kodakarensis* (Figure 3, denoted in green).

In contrast to Thermococcus pathway, other thermophilic archaeal (*Methanocaldococcus jannaschii*, *Methanosarcina acetivorans* and *Archaeoglobus lithotrophicus)* use a second pathway. In this pathway, ribose-phosphate pyrophosphokinase (PRPP) is suggested to be desphophorylated to ribose 1,5 bisphosphate via an abiotic reaction that is facilitated by high temperatures. In archaea, the non-enzymatically produced ribose 1,5-bisphosphate is then converted to RuBP by an enzyme annotated to function in thiamine monophosphate formation (Figure 3, unique pathway denoted in yellow; Finn and Tabita, 2004) (that is, PRPP → ribose 1,5 bisphosphate → RuBP). Given that only one Peregrinibacterium genome contains a distant homolog of this RuBP-forming enzyme, and all of these CPR were recovered from an ~14 °C aquifer, this pathway seems less likely in most CPR bacteria whose genomes encode RubisCO. However, we note many PER genomes encode an enzyme that can convert PRPP to AMP. This enzyme, adenine phosphoribosyltransferase, may represent an alternative mechanism by which PER genomes can use PRPP via the AMP pathway and not the pathway previously reported in thermophilic archaea (that is, PRPP → AMP → ribose 1,5 bisphosphate → RuBP). Thus we cannot rule out that PRPP may also be used as a source for RuBP in some members of the CPR via a *T. kodakarensis*-like pathway (Figure 3).

Very recently, Aono *et al.* (2015) identified alternative enzymes involved in the nucleoside metabolism pathway in *T. kodakarensis*, linking nucleoside moieties to central carbon metabolism. Instead of AMP phosphorylase (step 1 *T. kodakarensis* pathway), the ribose moieties of adenosine, guanosine and uridine are converted to ribose-1-bisphosphate (R1P) by three nucleoside phosphorylases. R1P is converted to ribose-1,5-bisphosphate (R15P) via a nucleoside kinase (ADP-dependent ribose-1-phosphate kinase). R15P is likely directed to glycolysis via the R15P isomerase and RubisCO. In this metabolic context, using phosphate and ADP, the ribose moieties of adenosine, guanosine and uridine are converted to R15P via R1P by nucleoside phosphorylases and ADP-R1P kinase (Aono *et al.*, 2015) (Figure 3, blue dotted line). Finally, R15P can be directed to glycolysis via the functions of R15P isomerase and RubisCO (Aono *et al.*, 2015). We examined the CPR genomes for this alternative *T. kodakarensis* pathway and found some (albeit weak) support for this pathway in Microgenomates and most PER genomes. Thus it cannot be ruled out that some CPR bacteria could use an alternative route through which R15P is generated from the phosphorolysis of nucleosides (Figure 3).

CPR organisms that contain RubisCO are inferred to be obligate fermenters (Supplementary Figure S10), consistent with prior metabolic analyses (Wrighton *et al.*, 2012; Campbell *et al.*, 2013; Kantor *et al.*, 2013; Wrighton *et al.*, 2014). All of the genomes lack a complete tricarboxylic acid cycle, NADH dehydrogenase (complex I) and most other complexes from the oxidative electron transport phosphorylation chain (for example, complex II, complex III, complex IV and quinones). All CPR genomes included in this study, besides the Dojkabacteria (WS6), have complete or near-complete glycolysis and/or pentose phosphate pathway(s) and the capacity to produce acetate, lactate and/or hydrogen as byproducts of fermentation. Notably, the PER, in contrast to bacteria of the

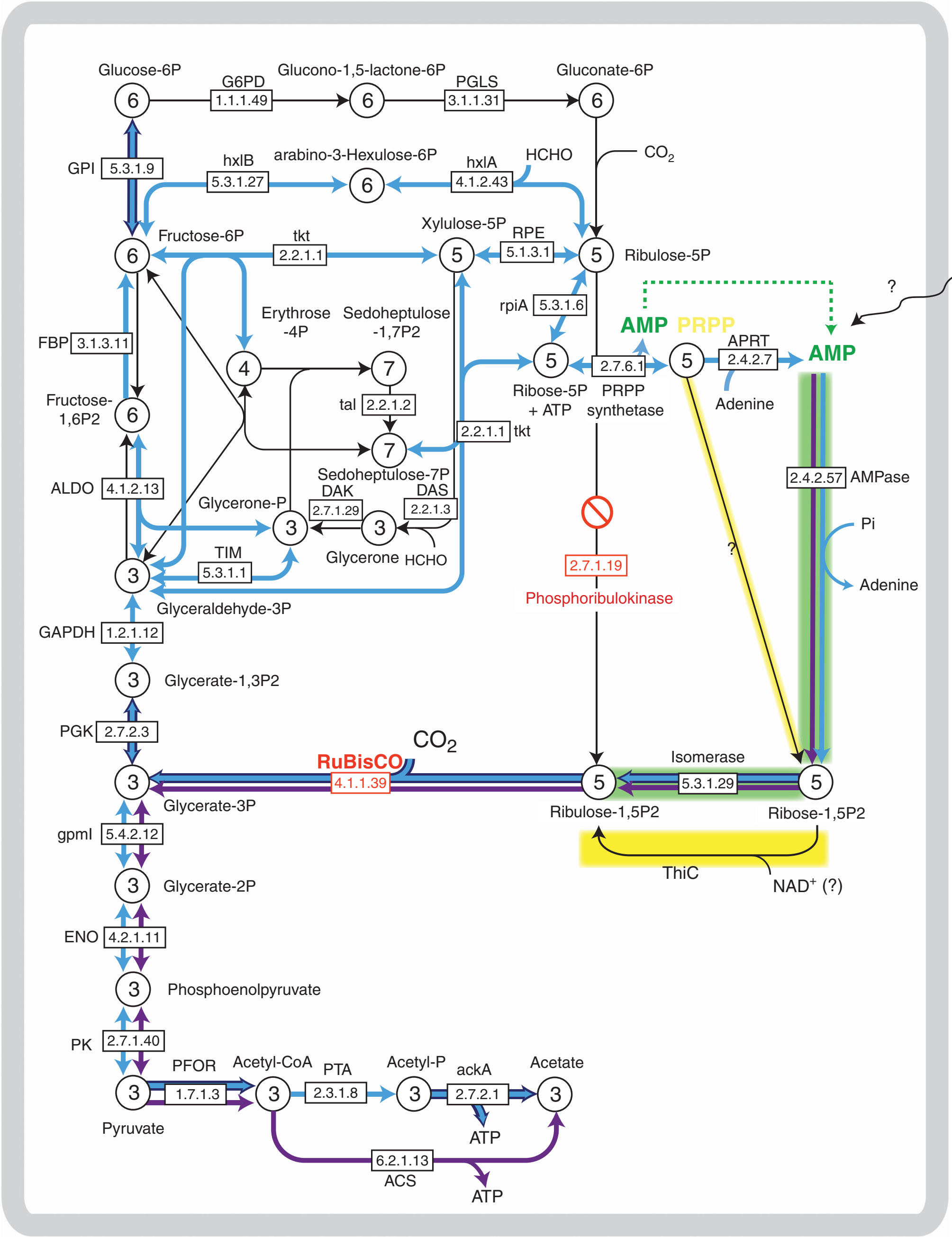


Figure 3 Proposed metabolic role of RubisCO, with genes identified in the PER-1 (blue line) and WS6-1 genomes (purple line). Black lines indicate genes not identified in either organism. Thick blue arrows represent genes identified by metatranscriptomics (Supplementary Table S4). Yellow lines denoted the proposed PRPP pathway indicated in the text. Abbreviations not indicated in the text: G6PD, glucose-6-phosphate dehydrogenase; PGLS, 6-phosphogluconolactonase; hxlB, 6-phospho-3-hexuloisomerase; hxlA, 3-hexulose-6-phosphate synthase Pgm, phosphoglucomutase; GPI, glucose-6-phosphate isomerase; FBP, fructose-1,6-bisphosphatase; Aldo, fructose-bisphosphate aldolase; GAPDH, glyceraldehyde 3-phosphate dehydrogenase, Pgk, phosphoglycerate kinase; TIM, triosephosphate isomerase; GpmI, 1,3-bisphosphoglycerate-independent phosphoglycerate mutase; Eno, enolase; PK, pyruvate kinase; PFOR, pyruvate/2-oxoacid-ferredoxin oxidoreductase; ACS, Acetyl-CoA synthetase; PTA, phosphotransacetylase; ackA, acetate kinase; RpiA, ribose 5-phosphate isomerase A; RPE, ribulose-phosphate 3-epimerase; Tkt, transkelotase; Tal, transaldolase; APRT, adenine phosphoribosyltransferase; PRPP synthetase, ribose-phosphate pyrophosphokinase; DAK, glycerone kinase; DAS, dihydroxyacetone synthase; AMPase, AMP phosphorylase; isomerase, Ribose 1,5-bisphosphate isomerase; NMP, nucleoside 5′-monophosphate.

other CPR genomes with RubisCO described here, seem to have the capacity to synthesize nucleotides and some amino acids (Supplementary Figure S11).

The metabolic information gleaned from the first complete WS6 genome is sparse compared with that for the other CPR genomes. Enzymes that were identified include those required for the last steps of glycolysis from PGA to pyruvate and ultimately fermentation to acetate and ATP. Notably, PGA is the product of RuBP conversion by RubisCO (Figure 3).

We failed to identify genes for upper glycolysis, gluconeogenesis and the pentose phosphate pathway in the WS6 genomes (Supplementary Figure S10). Gene and protein sequences discussed in this manuscript can be accessed through links in Supplementary Figures S10 and S11. This finding is similar to reports of bacteria of the SR1 phylum (placement within the CPR is uncertain) that also encode form II/III RubisCO and lack these pathways (Campbell *et al.*, 2013; Kantor *et al.*, 2013) (Supplementary Figure S10). Complete and closed SR1 and WS6 genomes also lack pathways for biosynthesis of nucleotides (purines and pyrimidines) and amino acids (Kantor *et al.*, 2013). Consistent with reports for SR1 and some small CPR genomes (Kantor *et al.*, 2013; Gong *et al.*, 2014; Brown *et al.*, 2015; He *et al.*, 2015; Luef *et al.*, 2015), we suggest that WS6 bacteria reported here have an obligatory symbiotic or parasitic lifestyle and are not free-living.

It is possible that WS6 and other CPR that encode RubisCO may be reliant on other members of the community for external ribose, which is utilized as carbon and energy sources. A bacterial or archaeal cell is on average 20% by weight RNA, and RNA is 40% by weight ribose, meaning that a cell is roughly 8% pure ribose. A recent paper used this information to suggest that ribose was likely an abundant sugar available on early earth for fermentation and that the archaeal type III RubisCO pathway of nucleoside monophosphate conversion to 3-PGA might be a relic of ancient heterotrophy (Schönheit *et al.*, 2015). In a similar manner, we suggest that in WS6 the ribose moiety can be funneled into RuBP via the two gene AMP pathway and converted to PGA via RubisCO to ultimately yield ATP and acetate via substrate-level phosphorylation. This use of ribose would minimize the energetic cost of upper glycolysis, maximizing energy yields for these organisms such as WS6 with tiny genomes.

CPR RubisCO transcripts detected in the environment

We examined the metatranscriptomic sequence data set collected from the 0.2-μm filter samples for reads that mapped uniquely to PER-1 genome (Supplementary Table S4), given that these organisms were enriched on this filter (there was insufficient biomass after metagenomics to investigate the 0.1-μm filters). In the sample collected prior to acetate amendment (sample A), we confirmed the expression of RubisCO (Scaffold 3_gene102) and the surrounding gene neighborhood (3_105, 3_107, 3_109) (Figure 4a). Notably, besides RubisCO, all of these transcribed genes were annotated as hypothetical. Also in sample A, we detected transcripts for step 1 of the AMP pathway (AMP phosphorylase (Scaffold 5_gene7)) and energy generation via acetate production (phosphotransacetylase (Scaffold 5_gene5)), as well as a poorly annotated phosphoesterase (Scaffold 5_gene4) (Figure 4b). The expression of RubisCO, a key gene in the AMP pathway, and genes for acetate production from the same genome prior to stimulation suggests a role for the RubisCO and likely the AMP pathway in bacteria under natural aquifer conditions.

The PER RubisCO can support autotrophic CO_2-dependent growth

We synthesized a recombinant protein in *E. coli* of the PER-1 form II/III RubisCO that was transcribed *in situ*. We demonstrated CO_2 fixation activity from the crude extract using RuBP as a substrate

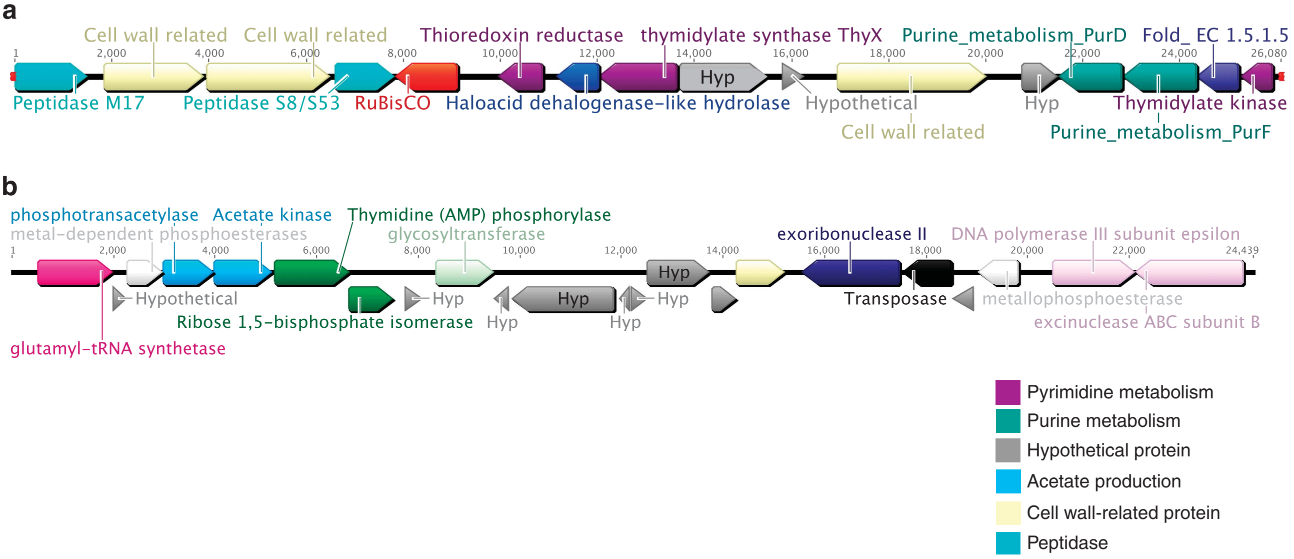


Figure 4 Gene neighborhoods near the form II/III RubisCO (**a**) as well as AMP phosphorylase and R15P isomerase (**b**) from the PER-1 genome. (**a**) RubisCO (red) clusters with genes for pyridmidine and purine metabolism. (**b**) Genes for producing acetate (blue) are near two homologs from the AMP pathway (dark green). Gene clusters were organized and visualized using Geneious v7.0.6.

(Kreel and Tabita, 2007; Supplementary Table S5). In contrast to the PER results, the OD1 and WS6 recombinant proteins failed to support comparable (OD1) or detectable (WS6) RuBP-dependent CO_2 fixation under these conditions. This latter WS6 finding is consistent with the missing catalytic residues and thus their annotation as RubisCOs remains uncertain (Figure 2).

More importantly, we demonstrated that the PER form II/III enzyme could complement a *Rhodobacter capsulatus* RubisCO-deletion host strain to support autotrophic CO_2-dependent growth, much like archaeal form III RubisCO (Finn and Tabita, 2003) (Figure 5). However, while under the conditions tested RubisCO fixed extracellular CO_2, we cannot rule out the possibility that *in situ* this RubisCO may scavenge intracellular CO_2 pools to support heterotrophic metabolism as previously reported (Joshi and Tabita, 1996; Tachi and Tabita, 2001; Kresge *et al.*, 2005; McKinlay and Harwood, 2010; Aono *et al.*, 2015).

Conclusion

Given that RubisCO is exceptionally well studied owing to its central role in CO_2 fixation in the Earth's biosphere, it is significant that this analysis of subsurface sediments has yielded many new sequence types. We expand the phylogenetic membership of CPR genomes known to contain RubisCO sequences from the PER to include members of the Dojkabacteria (WS6), Parcubacteria (OD1) and Microgenomates (OP11). We provide the first demonstration that genes of this pathway are transcribed *in situ* in bacteria. We also verified the first physiological complementation of a novel RubisCO that can support CO_2 fixation from a metagenomically derived sequence from the bacterial candidate phylum radiation. Future research is needed to identify the currently unknown but co-expressed flanking genes to better understand the regulatory as well as complete metabolic role for RubisCO in these small, apparently obligatory fermentative bacteria.

The presence of form II/III and a currently novel, undefined form of RubisCO in bacteria with a similar metabolic context (for example, anaerobic, non-CBB functionality) to RubisCO in archaea suggests a deep ancestry of this non-CBB cycle enzyme. Of particular note is that these bacterial RubisCOs occur in organisms with strikingly limited metabolic potential and no detected respiratory capacity. This metabolism, where ribose is suggested to have a key role in carbon compound transformations and energy generation, may have been important in organisms that existed before the emergence of respiration (Schönheit *et al.*, 2015).

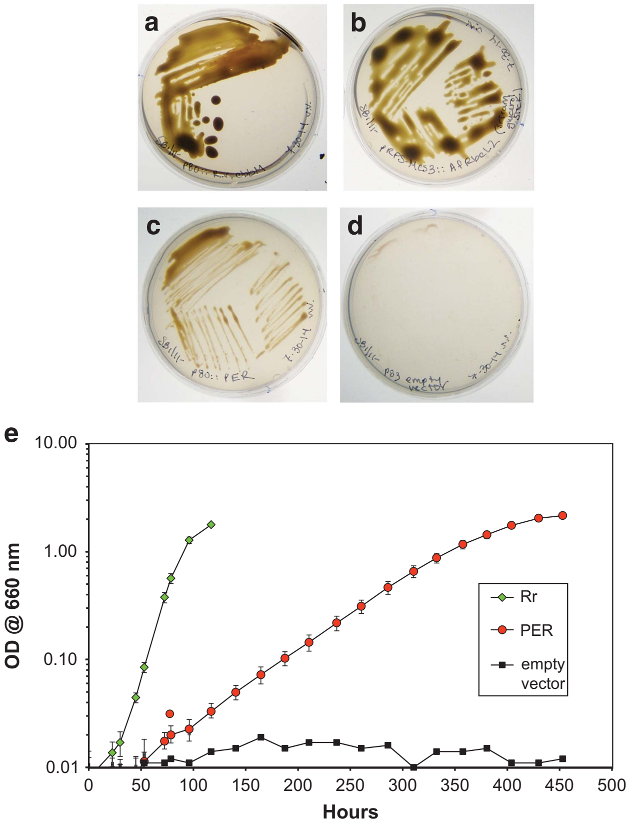


Figure 5 Photoautotrophic growth complemented by the PER form II/III RubisCO gene in *Rhodobacter capsulatus* strain SB I/II⁻. (**a**) p80::*Rhodospirillum rubrum cbbM*; (**b**) pRPS-MCS3::*Archaeoglobus fulgidus rbcL*; (**c**) p80::*Perigrinibacteria rbcL*; (**d**) empty vector. Anaerobic CO_2-dependent photoautotrophic growth was obtained using Ormerod's minimal medium agar plates in a sealed jar flushed with a 5% CO_2/95% H_2 gas mixture. (**e**) Anaerobic CO_2-dependent growth was performed in sealed tubes continuously bubbled with a 5% CO_2/95% H_2 gas mixture. Each data point represent the average ± s.d. of triplicate cultures.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work is partially based upon work supported through the Lawrence Berkeley National Laboratory's (LBNL) Sustainable Systems Scientific Focus Area. The US Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research funded the work under contract DE-AC02-05CH11231 (LBNL; operated by the University of California) and DE-SC0004918 to JB and NIH grant R01GM095742 to FRT. DNA sequencing was conducted by the DOE Joint Genome Institute through the Community Science Program. RNA sequencing was performed at the DOE-supported Environmental Molecular Sciences Laboratory at PNNL as part of an award to JFB, KCW, MJW and KHW. We thank Lye Meng Markillie and Ronald C Taylor for transcriptomic sequencing. We thank Professor B Alber at the Ohio State University for discussions on microbial metabolism. Genomes and raw reads in this manuscript

are deposited in NCBI under the accession numbers (pending) OP11_GWA2_42_18: LCDD00000000; OP11 _GWA2_43_14: LCFP00000000; PER_GWA2_38_35: LBUV00000000; PER_GWF2_38_29: LBUS00000000; PE R_GWF2_39_17: LBWM00000000; WS6_GWF2_39_15: LBWK00000000; WS6_GWC1_33_20: LBOV00000000 (BioProject: PRJNA273161 and BioSample: SAMN 03319638). RNA sequences have been deposited in the NCBI Sequence Read Archive under accession number SRP050083. The genomes and relevant fasta files for all mentioned proteins are available via the website: http://ggkbase.berkeley.edu/, under the project name 'RubiscO' under the link entitled 'Rifle groundwater metagenome'.

References

Abascal F, Zardoya R, Posada D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**: 2104–2105.

Alonso H, Blayney MJ, Beck JL, Whitney S. (2009). Substrate-induced assembly of *Methanococcoides burtonii* D-ribulose-1,5-bisphosphate carboxylase/oxygenase dimers into decamers. *J Biol Chem* **244**: 33876–33882.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Aono R, Sato T, Imanaka T, Atomi H. (2015). A pentose bisphosphate pathway for nucleoside degradation in Archaea. *Nat Chem Biol* **11**: 355–360.

Aono R, Sato T, Yano A, Yoshida S, Nishitani Y, Miki K *et al.* (2012). Enzymatic characterization of AMP phosphorylase and ribose-1,5-bisphosphate isomerase functioning in an archaeal AMP metabolic pathway. *J Bacteriol* **194**: 6847–6855.

Ashida H, Saito Y, Nakano T, Tandeau de Marsac N, Sekowska A, Yokota A. (2008). RubisCO-like proteins as the enolase enzyme in the methionine salvage pathway: functional and evolutionary relationships between RubisCO-like proteins and photosynthetic RubisCO. *J Exp Bot* **59**: 1543–1554.

Baker BJ, Lazar CS, Teske AP, Dick GJ. (2015). Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* **3**: 14.

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A *et al.* (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**: 208–211.

Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T *et al.* (2013). UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc Natl Acad Sci USA* **110**: 5540–5545.

Castelle CJ, Hug LA, Wrighton KC, Thomas BC, Williams KH, Wu D *et al.* (2013). Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nat Commun* **4**: 2120.

Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ *et al.* (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr Biol* **25**: 690–701.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP *et al.* (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**: R85.

Dojka MA, Harris JK, Pace NR. (2000). Expanding the known diversity and environmental distribution of an uncultured phylogenetic division of bacteria. *Appl Environ Microbiol* **66**: 1617–1621.

Edgar RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Estelmann S, Ramos-Vera WH, Gad'on N, Huber H, Berg IA, Fuchs G. (2011). Carbon dioxide fixation in 'Archaeoglobus lithotrophicus': are there multiple autotrophic pathways? *FEMS Microbiol Lett* **319**: 65–72.

Falcone DL, Tabita FR. (1993). Complementation analysis and regulation of CO_2 fixation gene expression in a ribulose 1,5-bisphosphate carboxylase-oxygenase deletion strain of *Rhodospirillum rubrum*. *J Bacteriol* **175**: 5066–5077.

Finn MW, Tabita FR. (2003). Synthesis of catalytically active form III ribulose 1,5-bisphosphate carboxylase/oxygenase in archaea. *J Bacteriol* **185**: 3049–3059.

Finn MW, Tabita FR. (2004). Modified pathway to synthesize ribulose 1,5-bisphosphate in methanogenic archaea. *J Bacteriol* **186**: 6360–6366.

Gong J, Qing Y, Guo X, Warren A. (2014). 'Candidatus *Sonnebornia yantaiensis*', a member of candidate division OD1, as intracellular bacteria of the ciliated protist *Paramecium bursaria* (Ciliophora, Oligohymenophorea). *Syst Appl Microbiol* **37**: 35–41.

He X, McLean JS, Edlund A, Yooseph S, Hall AP, Liu S-Y *et al.* (2015). Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc Natl Acad Sci USA* **112**: 244–249.

Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR *et al.* (2013). Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* **1**: 22.

Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**: 2223–2230.

Joshi HM, Tabita FR. (1996). A global two component signal transduction system that integrates the control of photosynthesis, carbon dioxide assimilation, and nitrogen fixation. *PNAS* **93**: 14515–14520.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**: D109–D114.

Kanehisa M, Goto S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**: 27–30.

Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ *et al.* (2013). Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *MBio* **4**: e00708–e00713.

Konstantinidis KT, Rosselló-Móra R. (2015). Classifying the uncultivated microbial majority: a place for metagenomic data in the Candidatus proposal. *Syst Appl Microbiol* **38**: 223–230.

Kreel NE, Tabita FR. (2007). Substitutions at methionine 295 of *Archaeoglobus fulgidus* ribulose-1,5-6bisphosphate carboxylase/oxygenase affect oxygen binding and CO2/O2 specificity. *J Biol Chem* **282**: 1341–1351.

Kresge N, Simoni RD, Hill RL. (2005). The discovery of heterotrophic carbon dioxide fixation by Harland G. Wood. *J Biol Chem* **280**: e15.

Lloyd KG, Schreiber L, Petersen DG, Kjeldsen KU, Lever MA, Steen AD *et al.* (2013). Predominant archaea in marine sediments degrade detrital proteins. *Nature* **496**: 215–218.

Luef B, Frischkorn KR, Wrighton KC, Holman H-YN, Birarda G, Thomas BC *et al.* (2015). Diverse uncultivated ultra-small bacterial cells in groundwater. *Nat Commun* **6**: 6372.

McKinlay JB, Harwood CS. (2010). Carbon dioxide fixation as a central redox cofactor recycling mechanism in bacteria. *Proc Natl Acad Sci* **107**: 11669–11675.

Nawrocki E. (2009). Structural RNA homology search and alignment using covariance models. All Theses and Dissertations (ETDs). Paper 256. http://openscholarship.wustl.edu/etd/256.

Nawrocki EP, Kolbe DL, Eddy SR. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**: 1335–1337.

Ormerod JG, Ormerod KS, Gest H. (1961). Light-dependent utilization of organic compounds and photoproduction of molecular hydrogen by photosynthetic bacteria; relationships with nitrogen metabolism. *Arch Biochem Biophys* **94**: 449–463.

Peng Y, Leung HCM, Yiu SM, Chin FYL. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–1428.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P *et al.* (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590–D596.

Rinke C, Lee J, Nath N, Goudeau D, Thompson B, Poulton N *et al.* (2014). Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat Protoc* **9**: 1038–1048.

Satagopan S, Chan S, Perry LJ, Tabita FR. (2014). Structure-function studies with the unique hexameric form II ribulose-1,5-bisphosphate carboxylase/oxygenase (RubisCO) from *Rhodopseudomonas palustris*. *J Biol Chem* **289**: 21433–21450.

Satagopan S, Scott SS, Smith TG, Tabita FR. (2009). A RubisCO mutant that confers growth under a normally 'inhibitory' oxygen concentration. *Biochemistry* **48**: 9076–9083.

Sato T, Atomi H, Imanaka T. (2007). Archaeal type III RubisCOs function in a pathway for AMP metabolism. *Science* **315**: 1003–1006.

Schönheit P, Wolfgang B, Martin WF. (2015). On the origin of heterotrophy. *Trends Microbiol* **24**: 12–25.

Smith SA, Tabita FR. (2003). Positive and negative selection of mutant forms of prokaryotic (cyanobacterial) ribulose-1,5-bisphosphate carboxylase/oxygenase. *J Mol Biol* **331**: 557–569.

Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**: 1282–1288.

Tabita FR, Hanson TE, Li H, Satagopan S, Singh J, Chan S. (2007). Function, structure, and evolution of the RubisCO-like proteins and their RubisCO homologs. *Microbiol Mol Biol Rev* **71**: 576–599.

Tabita FR, Satagopan S, Hanson TE, Kreel NE, Scott SS. (2008). Distinct form I, II, III, and IV RubisCO proteins from the three kingdoms of life provide clues about RubisCO evolution and structure/function relationships. *J Exp Bot* **59**: 1515–1524.

Talavera G, Castresana J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**: 564–577.

Tichi MA, Tabita FR. (2001). Interactive control of *Rhodobacter capsulatus* redox balancing systems during phototrophic metabolism. *J Bacteriol* **183**: 6344–6354.

Williams KH, Long PE, Davis JA, Wilkins MJ, N'Guessan AL, Steefel CI *et al.* (2011). Acetate availability and its influence on sustainable bioremediation of uranium-contaminated groundwater. *Geomicrobiol J* **28**: 519–539.

Wrighton KC, Castelle CJ, Wilkins MJ, Hug LA, Sharon I, Thomas BC *et al.* (2014). Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. *ISME J* **8**: 1452–1463.

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC *et al.* (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**: 1661–1665.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)