Using comparative genomics to uncover new kinds of protein-based metabolic organelles in bacteria

Julien Jorda,¹ David Lopez,¹ Nicole M. Wheatley,² and Todd O. Yeates^{1,2,3*}

¹UCLA-DOE Institute for Genomics and Proteomics, 611 Charles Young Dr. East, Los Angeles, California 90095

²UCLA Molecular Biology Institute, 611 Charles Young Dr. East, Los Angeles, California 90095

³Department of Chemistry and Biochemistry, University of California, Los Angeles, 611 Charles Young Dr. East, Los Angeles, California 90095

Received 12 September 2012; Revised 9 November 2012; Accepted 12 November 2012

DOI: 10.1002/pro.2196

Published online 27 November 2012 proteinscience.org

Abstract: Bacterial microcompartment (MCP) organelles are cytosolic, polyhedral structures consisting of a thin protein shell and a series of encapsulated, sequentially acting enzymes. To date, different microcompartments carrying out three distinct types of metabolic processes have been characterized experimentally in various bacteria. In the present work, we use comparative genomics to explore the existence of yet uncharacterized microcompartments encapsulating a broader set of metabolic pathways. A clustering approach was used to group together enzymes that show a strong tendency to be encoded in chromosomal proximity to each other while also being near genes for microcompartment shell proteins. The results uncover new types of putative microcompartments, including one that appears to encapsulate B₁₂-independent, glycyl radical-based degradation of 1,2-propanediol, and another potentially involved in amino alcohol metabolism in mycobacteria. Preliminary experiments show that an unusual shell protein encoded within the glycyl radical-based microcompartment binds an iron-sulfur cluster, hinting at complex mechanisms in this uncharacterized system. In addition, an examination of the computed microcompartment clusters suggests the existence of specific functional variations within certain types of MCPs, including the alpha carboxysome and the glycyl radical-based microcompartment. The findings lead to a deeper understanding of bacterial microcompartments and the pathways they sequester.

Keywords: microcompartment; carboxysome; bacterial organelle; metabolic pathways; glycyl radical enzymes

Abbreviations: BMC, bacterial microcompartment shell protein; Eut, ethanolamine utilization; Etu, ethanol utilization; Grp, glycyl radical-based 1,2-propanediol utilization; MCP, microcompartment; PCC, pairwise correlation coefficient; Pdu, propanediol utilization.

Additional Supporting Information may be found in the online version of this article.

Julien Jorda and David Lopez contributed equally to this work.

Grant sponsor: NIH; Grant number: R01AI081146; Grant sponsor: Ruth L. Kirschstein National Research Service; Grant number: GM007185.

*Correspondence to: Todd O. Yeates, Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095-1569. E-mail: yeates@mbi.ucla.edu

Introduction

Over the last few decades, the general view that bacteria have simple internal structures has changed. Electron microscopy investigations have demonstrated that bacteria produce a wide variety of intracellular inclusions.^{1,2} The discovery and subsequent isolation of one particular class of polyhedrally shaped bodies dates back almost 40 years.³ These giant proteinaceous bodies, called bacterial microcompartments (hereafter referred to as MCPs) are typically 80–150 nm in diameter and consist of a set of interior enzymes surrounded by a thin protein shell reminiscent of a viral capsid.^{4–8} MCPs have been proposed to serve diverse functional roles:

improving flux through key metabolic pathways,⁹ sequestering cytotoxic or volatile intermediates in a pathway,^{10,11} and protecting the encapsulating enzymes from exposure to competing or reactive molecules (e.g., O₂),¹² all while allowing passage of substrates and products across the shell. Biochemical and structural studies have revealed microcompartments to be mechanistically complex entities, warranting their classification as organelles (Fig. 1).

The enzymes and metabolic pathways encapsulated by microcompartments are diverse, allowing the delineation of a few distinct classes of MCPs.⁷ The founding member is the carboxysome, present in cyanobacteria and some chemoautotrophs.^{3,13} The carboxysome houses two enzymes: RuBisCO (a low efficiency enzyme essential to autotrophic fixation of carbon dioxide) and carbonic anhydrase [Fig. 1(B)]. The catalytic efficiency of RuBisCO is improved by having its CO₂ substrate produced by carbonic anhydrase inside the MCP, where its escape might be retarded by the shell.^{14,15} Two carboxysome subtypes (alpha and beta) are delineated by their partially distinct protein components; they are distributed along phylogenetic lines within chemoautotrophs (alpha only) and cyanobacteria (alpha or beta). Biochemical and genetic studies have been conducted on two other microcompartments: the Pdu microcompartment of enteropathic *Salmonella enterica*^{16–18} and the Eut microcompartment of *Salmonella* (also found in some strains of *Escherichia coli*) (refs. 11,19,20). These MCPs metabolize 1,2-propanediol and ethanolamine, respectively [Fig. 1(B)].

In contrast to the metabolic variations presented by different MCPs, the proteins that self-assemble to form the outer shell are homologous across the disparate functional types. MCP shells are composed mainly by proteins bearing one or sometimes two tandem bacterial microcompartment (or BMC) domains, identified first by Shively and coworkers.¹³ We refer to these major shell components as BMC shell proteins. Bacterial microcompartments themselves are sometimes referred to as BMCs, but in this paper we refer to microcompartments as MCPs to avoid confusing the compartment with its main structural proteins. In each MCP, a few (three to seven) different paralogs of the BMC shell protein assemble, from a few thousand copies in total, to form the shell (Supporting Information Fig. S1). Crystallographic studies have given insight into the significance of the conserved BMC domain and how it relates to microcompartment shell organization as a whole.^{4,5,8,21,22} In particular, structures of several BMC shell proteins have revealed that they generally assemble as cyclic homohexamers, which pack side-by-side to build a molecular layer^{4,5,21,23} (Supporting Information Fig. S1). The center of each hexamer is typically perforated by a narrow pore along the sixfold axis of symmetry; these pores are

Figure 1. Morphology of MCPs and models of their sequestered pathways. (A) Thin-section EM of a dividing cell of the cyanobacterium *Synechocystis* sp. PCC6803 (left) along with an enlargement of a single carboxysome (right, courtesy of Wim Vermaas). (B) Electron micrographs of a thin-section of *Salmonella enterica* serovar Typhimurium LT2 (left; Reproduced from Ref. ⁹⁶, with permission from Thomas Bobik) and purified Pdu microcompartments (right; Reproduced from Ref. ⁸, with permission from Thomas Bobik). (C) Models for CO₂ fixation and 1,2-propanediol and ethanolamine metabolism in the carboxysome, Pdu and Eut microcompartments, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

believed to provide the conduits for substrates and products to cross the shell.^{4,23–25} Another gene family (*ccmL/csoS4/eutn/pduN*), which is distinct from the BMC family, is typically present as well and is believed to code for minor vertex proteins in the shell.²³ Genomic analyses have shown that the encapsulated enzymes and the BMC shell proteins are frequently encoded within the same operon,

consistent with the idea that expression and assembly of the distinct MCP components occur in a coordinated fashion.⁷ Mechanisms underlying the targeting of enzymes to their correct destination in MCPs have been partially clarified. Experiments show that, for a few cases, short N-terminal sequence extensions are necessary and sufficient to target enzymes to the MCP interior or lumen.^{26–30} In another case, a C-terminal region has been shown to be important for targeting an interior protein.³¹

Genomic studies offer prospects for new discoveries related to MCPs. A current search of sequence databases for proteins bearing a BMC domain indicates that microcompartments are distributed across approximately 17% (265 out of 1568) of the fully sequenced bacterial genomes currently available. Several comparative genomics analyses have suggested the existence of other types of MCPs besides the three currently studied,^{4,8,21,22} but clear metabolic models have not been developed. The ever-expanding body of genomic sequences, combined with the tendency of BMC shell proteins to be encoded in proximity to the enzymes they encapsulate, suggests bioinformatics strategies for predicting the existence of novel metabolic schemes within MCPs. In this study, we sought to identify potentially novel MCP pathways by searching for the co-occurrence of groups of enzyme-encoding genes in MCP operons. The approach is based on the idea that groups of enzymes that tend to occur together within individual MCP operons are likely to represent encapsulated pathways. Here, we describe how our method recapitulates the metabolic pathways hosted in known MCPs, while also uncovering MCPs that are novel or that represent variations on previously studied types.

Results

MCP operons were examined across the 113 fully sequenced bacterial genomes where BMC shell proteins could be identified. Based on their identified PFAM domains and annotations from the KEGG Orthology system, proteins encoded by genes proximal to BMC shell genes were collapsed into distinct Protein Functional Groups intended to represent unique cellular functions. In order to cluster these into disjoint sets that might each represent one type of MCP, we evaluated in a full pairwise fashion the tendency of every pair of Protein Functional Groups to co-occur within individual MCP operons. Statistical tests were applied throughout the procedure to maximize the likelihood of producing biologically meaningful results (see Methods and Fig. 2).

Our genomic context-based approach produced 10 candidate metabolic clusters containing between two and 13 proteins and enzymes (Fig. 3). Protein Functional Groups are represented as nodes, with two nodes being connected by a line or edge when

Figure 2. Schematic for identifying pairs of proteins or enzyme families that tend to co-occur in the context of microcompartment (MCP) operons. As step 1, BMC-proximal genes and their encoded proteins are collected following our operational definition of a MCP operon (see Methods). In step 2, to gain clarity and statistical power, the BMC-proximal genes are assigned to Protein Functional Groups, grouping together similar protein sequences where possible. In step 3, the co-occurrence tendency is evaluated by a pairwise correlation coefficient (PCC) for every pair of Functional Groups. In the last step, after applying statistical confidence tests, strongly linked Protein Functional Groups are clustered together. Different clusters identify MCPs with distinct metabolic functions. The scheme shown is a simplification; application to real genomic data leads to more MCP types and more proteins per cluster.

their tendency to co-occur was judged to be statistically significant. Among the 10 clusters, four were consistent with well-characterized, canonical MCPs: carboxysomes of the alpha and beta type, along with the Pdu and Eut microcompartments. Most of the enzymes known to participate in MCP function were found to be effectively clustered, in some cases together with other unanticipated Protein Functional Groups. Of particular interest, a number of poorly

bacterioferritin, pterin-4a-carbinolamine dehydratase homologues, an ATP-binding protein homologous to CbbQ, along with its activation protein (CbbO), and cobyrinic acid a,c-diamide synthase (CobQ). The presence of these noncanonical genes appears to arise from their co-occurrence in the vicinity of the BMC shell gene *csoS1D*, which was recently confirmed to be a bona fide carboxysome shell protein.³⁵ The tendency of CsoS1D to be encoded near some of the proteins identified here was reported in an earlier analysis,³⁵ but the biological relevance of these BMC-proximal genes has not been clarified yet.

The CsoS1D protein²⁴ belongs to a group of BMC shell proteins that each contain two tandem BMC domains, and which assemble as pseudo-hexameric trimers.^{4,25,36,37} Tandem domain BMC shell proteins from different kinds of MCPs have been shown to support both open and closed pore conformations.^{4,25,36} In the case of the Eut MCP, it was proposed that the open conformation in the tandem BMC protein EutL is required for transporting bulky cofactors across the shell for their use by the encapsulated ethanolamine utilizing enzymes. The core enzymes of the carboxysome—carbonic anhydrase and RuBisCO—do not require cofactors, leaving the purpose of a large pore in CsoS1D unexplained. We propose here, based on the tight genomic association of CsoS1D with various complex enzymes, that this shell protein likely supports the transport of molecules—possibly including bulky cofactors—for as-yet undefined reactions that might occur within variant forms of the alpha carboxysome in diverse bacteria.

The identities of the above five noncanonical Protein Functional Groups that appear linked to the alpha carboxysome allow for speculation regarding possible functional variations within this type of MCP. Two of the associated Protein Functional Groups, CbbQ and its activation protein, modulate the conformation and activity of RuBisCO, inducing a twofold increase in its $V_{\max}$.³⁸ CbbQ relies on ATP for activity, raising the prospect of nucleotide transport across the alpha carboxysome shell. A third protein, bacterioferritin, is known to form molecular cages for iron storage. Its functional connection to the alpha carboxysome, if any, is unknown. The remaining two Protein Functional Groups identified here have functions likely related to cofactor synthesis. One is CobQ, which uses glutamine or ammonia as a substrate in a reaction for synthesis of cobalamin,³⁹ a cofactor used in other MCPs but not previously associated with the carboxysome. The involvement of ammonia is intriguing in view of the enzyme's relatively high K_m (26–200 μM) for this substrate,⁴⁰ along with the well-established confinement or channeling of ammonia between enzymes in other metabolic contexts.^{41,42} The possibility that ammonia could be used for a reaction inside a func-

tionally extended form of the alpha-carboxysome constitutes a hypothesis for future testing.

The final Protein Functional Group strongly co-occurring with the canonical proteins of the alpha-carboxysome appears by sequence analysis to be distantly related to pterin-4 alpha-carbinolamine dehydratase (PCD), though key catalytic residues believed to play a role in PCD activity⁴³ are not obviously preserved by sequence alignments to the enzymes co-occurring with the alpha-carboxysome genes. PCD is involved in tetrahydrobiopterin recycling, where it catalyzes the reversible conversion of 4a-hydroxytetrahydrobiopterin to dihydrobiopterin.⁴⁴ Only tentative connections can be suggested between these Protein Functional Groups. For example, we note that tetrahydrobiopterin is often seen as a cofactor for aromatic amino acid hydroxylases,⁴⁵ which also require non-heme iron for activity;⁴⁶ this would provide a potential link to bacterioferritin. Also intriguing is the observation that some ferritin-like proteins are encapsulated within a different kind of protein compartment, considerably smaller than an MCP, known as the encapsulin nanocompartment.⁴⁷ Although clear functional relationships between these additional proteins found to be associated with alpha carboxysome operons cannot be established at the present time, recent experimental studies confirm that bacterioferritin and the enzyme homologous to PCD are in fact upregulated in concert with the canonical carboxysome genes.⁴⁸

Pdu microcompartment. Cluster 3 represents the Pdu MCP (Fig. 3). It contains all the enzymes known to operate in the 1,2-propanediol utilization pathway (Fig. 1), with the exception of PduH and PduV. These enzymes share identical domains with other clustered enzymes (PduH with PduD and PduV with EutP respectively) and thus were missed by our approach due to the inability in each case to segregate the homologous enzymes into two distinct groups. The Protein Functional Group annotated as *hisG* was represented as a weak node due to its tendency to occur at the margins of the *pdu* operon.

Cluster 4 contains a few enzymes involved in synthesizing the cobalamin cofactor (Fig. 3). This cluster of *cob* genes was not automatically joined to the canonical Pdu MCP (cluster 3) by our analysis, but the two clusters appear to be functionally linked. Experimental studies show that the *pdu* and *cob* operons are both tightly regulated by the PocR protein, and that propanediol degradation is dependent on cobalamin, B₁₂.^{18,49} In most cases we observe that the *cob* genes are adjacent or peripheral to the *pdu* operon and not interspersed with the BMC shell proteins. Thus in our analysis the correlations between these distinct clusters of Protein Functional Groups were not significant enough to merge the Pdu and Cob pathways into a single cluster.

Likewise, there are no experimental data tying these particular cobalamin synthesis reactions directly to the Pdu MCP. Nonetheless, B₁₂ is a required cofactor for 1,2-propanediol degradation and there are a few bacterial species where the genomic arrangement is distinct, and suggestive of a closer relationship. The *cobU* *cobC* and *cobS* genes are used to synthesize the lower ligand of B₁₂, suggesting that lower ligand synthesis may be limiting for B₁₂ production in some environments. Similarly, the *PduX* gene often found near the end of the *pdu* operon in enteric bacteria is also used for lower ligand synthesis.⁵⁰

Eut microcompartment. Cluster 5 represents the ethanolamine utilization (Eut) MCP. The proteins typically encoded by that operon are clustered by our method. Some additional proteins, more weakly connected, are also identified, including two genes coding for a sensor histidine kinase and a response regulator. Indeed, it has been previously established that among the species associated to the Eut microcompartment, some of them present an extended version of the canonical operon and embed a two-component signal transduction system: a histidine kinase and its response regulator, referred to as EutW and EutV, respectively.¹⁹ In vitro assays showed that ethanolamine induces a 15-fold increase in the rate of autophosphorylation of EutW, followed by the activation of EutV through phosphotransfer.^{19,51} Reciprocally, a closer look at the 17 organisms featuring this variant of the *eut* operon showed that the *eutR* gene is absent. The latter is known to regulate the *eut* operon in response to ethanolamine and adenosylcobalamin (AdoCbl).⁵² The EutVW and EutR regulatory systems appear to exist in mutually exclusive species that use Eut MCPs. The observed dichotomy appears to be largely phylogenetic; EutV and EutW are found mainly in the Firmicutes while EutR is found only in the Enterobacteriaceae.

We note that the putative microcompartment for ethanol utilization (Etu) discussed by Heldt *et al.*⁵³ gets grouped with the Eut cluster by our automatic clustering. This is because the latter pathway includes just two Protein Functional Groups, and these are also found in the ethanolamine utilization pathway (namely EutG and EutE). The operation of this presumptive Etu MCP has not been clarified yet by experimental studies, though physiological considerations suggest that it may be involved in converting ethanol to acetyl-CoA.⁵⁴

Protein clusters representing new presumptive MCP types

Five additional small clusters, besides the clusters clearly related to the canonical MCPs discussed above, are identified in our analysis. They highlight systems that have not yet been characterized in detail (Fig. 3). Four of these clusters (6–9) are

weakly interconnected at a level not sufficient for them to be automatically joined by our approach—they appear to represent variations within a complex type of MCP. Finally, a 10th cluster appears to represent a distinct entity. Findings related to these two presumptive MCP types—clusters 6–9 and cluster 10—are discussed below.

A putative glycy radical-based MCP. Among several enzymes identified in cluster 6, one of particular note belongs to a diverse family of glycy radical enzymes. The identification (in *Vibrio furnissi* M1) of BMC shell genes interspersed with an enzyme from this family was discussed by Wackett *et al.*⁵⁵ Based on sequence similarity, this enzyme has been previously annotated as a pyruvate formate lyase. However, the sequence similarity is low, and other considerations discussed subsequently argue that this enzyme, and the MCP that harbors it, most likely utilizes 1,2-propanediol rather than pyruvate, in a B₁₂-independent pathway. We report here that these enzymes are encoded in a genomic pattern substantially conserved across more than 20 species examined, suggesting the existence of a broadly distributed class of microcompartment apart from the better-known MCPs.

Cluster 7 consists of two co-occurring Functional Groups: an enzyme similar to the C-terminal domain of PduO (an ATP/cobalamin adenosyltransferase involved in vitamin B₁₂ synthesis in the Pdu pathway), and a MIP family channel protein known to transport small neutral metabolites across the membrane.⁵⁶ Cluster 8 includes two drug resistance proteins along with two regulatory proteins and a phosphotransacylase (a PduL homolog). Strikingly, an analysis of the operons supporting clusters 7 and 8 showed that the enzymes identified in cluster 6 were also present, though our algorithmic approach did not automatically detect connections between cluster 6 and either cluster 7 or 8. The Protein Functional Groups represented by clusters 7 and 8 are only sometimes present in the larger set of operons that contain cluster 6 as the conserved core (Fig. 4). This is consistent with the failure of our automatic procedure to connect clusters 7 and 8 to cluster 6; they appear to represent specific compositional variations.

Clusters 7 and 8 contain a number of proteins or enzymes without obvious relationships to glycy radical-based metabolism. The MIP channel protein from cluster 7 could be responsible for importing the substrate or substrates of this MCP, but the role of the enzyme similar to the PduO C-terminus is unclear. Among the five organisms found to have these two genes, three belong to the set of more than 20 where the cluster 7 genes occur together with the core enzymes of cluster 6. When present, these genes are found dispersed between multiple BMC shell protein genes (Fig. 4). Cluster 8,

Figure 4. Phylogenetic profile of the BMC-proximal proteins and enzymes from the presumptive glycyl radical-based propanediol utilization (Grp) microcompartment operon across 23 bacterial species. (A) The 23 microorganisms are extracted from the analysis of cluster 6 (Fig. 3). Each colored block refers to the presence of a given protein (on the left) in the operon featured by a given organism (on the top). Cluster 6, corresponding to the core enzymes of the pathway, is depicted in light blue while supplemental enzymes from clusters 7, 8, and 9 are shown in red, purple and green respectively. A black block next to a protein name indicates the presence of a presumptive N-terminal targeting extension in this protein compared to homologues not involved in microcompartments, as analyzed following previously described methods.²⁷ The last row represents the profile of a divergent BMC shell protein, apparently specific to the Grp MCP, which was subjected to initial experimental characterization (see text). Next to each organism name, the number of tandem BMC proteins present in the operon is enclosed in red brackets. (B) Examples of gene organization in species featuring the Grp operon. BMC shell genes are colored in light gray, while the genes clustered by our approach are colored consistently with their corresponding clusters depicted in panel A. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

composed of four different Protein Functional Groups, was present in five different species (Fig. 4). These proteins appear to be associated with drug resistance mechanisms: two transcription regulators with the canonical helix-turn-helix DNA-binding motif found in antibiotic-resistance repressors like TetR,⁵⁷ two paralogs of a drug resistance protein,

and one yet uncharacterized protein conserved across the five species. Genes from this group were sometimes found flanking BMC shell genes at both upstream and downstream positions within a genome (Fig. 4). It is presently unclear what functional connections might relate the glycyl radical-based enzyme cluster (6) to clusters 7 and 8.

Cluster 9, whose genes also occasionally occur with those of cluster 6, involves two Protein Functional Groups similar to those found in the expanded version of the *eut* operon: a histidine kinase sensor and a response regulator receiver. These genes occur in the operon containing the glycyl radical-based degradation enzyme in about half of the species examined, but their frequent location at the upstream end of the operon, along with their absence from the other half of the species, caused them to cluster separately.

Diverse MCP operons encoding a glycyl radical enzyme have been noted in the literature, and a just-published survey (see Supporting Information in Ref. 31) suggests the potential for multiple types of MCP based on those operons. In the present study, the diverse MCP operons belonging to clusters 6–9 are seen to share a conserved set of enzymes (Fig. 4). We therefore group them for the present as a single distinct type of MCP, though functional variations are evident.

A presumptive MCP in mycobacteria. Cluster 10 identifies another distinct type of MCP operon present in four organisms (*Mycobacterium smegmatis*, *Mycobacterium* sp. MCS, *Mycobacterium gilvum*, and *Mycobacterium vanbaalenii*) and containing at least three Protein Functional Groups (Fig. 3): an amino-transferase, a short chain dehydrogenase similar to amino alcohol dehydrogenase, and a GntR family transcriptional regulator that Vindal *et al.*⁵⁸ have described as belonging to the FadR/HutC subfamily, whose members are known to bind ligands such as oxidized substrates related to amino acid metabolism or long chain fatty acids. Recently, genetic analysis of an operon coding for a similar MCP in *Rhodococcus erythropolis* MAK154 highlighted that amino alcohol dehydrogenase expression was repressed by a GntR transcriptional regulator.⁵⁹ That repression was relieved in the presence of 1-amino-2-propanol, which is the substrate of the amino alcohol dehydrogenase. After visual examination of the four operons, we found other enzymes missed by our automatic approach, which allowed for manual improvement of the functional predictions. These additional enzymes include an amino acid permease and an aldehyde dehydrogenase. Beyond a presumptive connection to amino alcohol metabolism, a more specific functional role for this putative MCP, which was also listed among prospective MCPs by Kinney *et al.*,³¹ cannot be offered at this time.

Genomic characterization of a putative glycyl radical-based propanediol utilization (Grp) MCP

As noted above, the enzymes of cluster 6 exhibit a strongly conserved pattern of co-occurrence with BMC shell proteins across many bacteria, covering 23 species from the Firmicutes and Proteobacteria

phyla.⁵⁵ The key enzyme from cluster 6 exhibits low but recognizable sequence similarity to pyruvate formate lyase from *Escherichia coli* (21% amino acid sequence identity), and glycerol dehydratase from *Clostridium butyricum* (33% identity). Both of those enzymes belong to a broad family of glycyl-radical enzymes.^{60–62} Experimental studies demonstrate that a pyruvate-formate lyase activase enzyme, which converts Gly734 (G-H) into a glycine radical (G*) in the active site of pyruvate formate lyase, is necessary for activating the latter, and for triggering the reaction in anaerobic conditions.⁶³ Glycyl radical enzymes generally require such activators to modify a critical residue in their active site.^{61,62,64} A glycyl-radical enzyme activase is indeed present in cluster 6; its occurrence is perfectly correlated with the presence of the glycyl-radical enzyme across the genomes in our analysis (Fig. 4). The activase is known to use S-adenosyl methionine as a substrate to generate the required radical,^{65,66} and a closer look at the genomic context showed the presence in some cases of a S-adenosylmethionine synthase gene (Supporting Information).

Pyruvate formate lyases contain two adjacent cysteine residues within their active sites, which are believed to be important in radical transfer.⁵⁶ Other types of glycyl radical enzymes, such as B₁₂ independent glycerol dehydratase, ribonucleotide reductases, and 1,2-propanediol dehydratases, contain only one of those cysteine residues. Sequence alignments indicate that the glycyl radical enzymes in our cluster 6 MCP contain only one cysteine. This, coupled with the closer similarity of the cluster 6 glycyl radical enzyme to a B₁₂-independent glycerol dehydratase, calls into question its annotation as a pyruvate formate lyase. Likewise, experimental studies highlighted that two of the organisms represented in cluster 6, *Proteus mirabilis* and *Escherichia fergusonii* are not able to ferment glycerol,⁶⁷ suggesting that glycerol is not the primary substrate for this type of MCP. Conversely, it has been shown that the enzyme annotated as glycerol dehydratase from *Clostridium butyricum* is also able to catalyze the dehydration of 1,2-propanediol to propionaldehyde.⁶⁸ Further evidence that 1,2-propanediol is the primary substrate in these systems comes from the upregulation of microcompartment genes when 1,2-propanediol is metabolized anaerobically in *Roseburia inulinivorans*.⁶⁹ The 1,2-propanediol arises from the anaerobic degradation of fucose, followed by conversion to propionaldehyde by a B₁₂-independent glycyl-radical enzyme in *R. inulinivorans* that is highly similar to the glycerol dehydratase in *Clostridium butyricum*.⁶⁸ Consistent with this theme, one of the cluster 6 microcompartments identified in *Clostridium phytofermentans* (Supporting Information) has been shown to be involved in fucose/rhamnose degradation (GSM333252 data from the Gene

Figure 5. Proposed model for metabolism in the Grp microcompartment. Similar to the Pdu microcompartment, the expected final products include propanol and propionyl-phosphate. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Expression Omnibus, submitted by J.L. Blanchard). Based on these observations, we surmise that the glycy radical enzymes in these MCPs act as B_{12} -independent 1,2-propanediol dehydratases. In line with previous naming schemes, we refer to these systems as glycy radical-based propanediol utilization or Grp MCPs.

Following the initial enzymatic step, similarities are evident between the Grp MCP and the B_{12} -dependent Pdu (and Eut) MCPs. Cluster 6 contains two Protein Functional Groups related to enzymes known to be targeted to the Eut or Pdu microcompartments: the aldehyde dehydrogenases EutE/PduP and EutJ, which plays a possible role of chaperone. Other enzymes believed to operate in the Eut or Pdu MCP are not seen in the glycy radical-based cluster 6. However, an enzyme closely related in sequence to the phosphotransacylase PduL (whose substrate is propionyl-CoA) is present,⁷⁰ further supporting the hypothesis of a pathway beginning with 1,2-propanediol as the initial MCP substrate. The PduL-like Functional Group is not automatically placed with cluster 6 in our approach, mainly because of its wide occurrence in other genomic contexts, including those represented by cluster 3 (Pdu). This situation highlights a limitation of our approach, which seeks to divide MCPs based on their use of distinct enzymes. A similar situation was seen for the Protein Functional Group representing alcohol dehydrogenases. The alcohol dehydrogenase functional group clusters automatically with the Eut MCP (Fig. 3, cluster 5, EutG), but genes for alcohol dehydrogenases also

co-occur systematically with the 1,2-propanediol metabolizing enzymes of cluster 6 (Fig. 4).

Modeling a glycy radical-based propanediol utilization pathway

Looking at the substrates and products of the enzymes identified computationally (cluster 6), and also by subsequent visual inspection of the corresponding operons, one can logically combine the reactions in a pathway that would use 1,2-propanediol as a main metabolic substrate. As a first step in modeling a reaction pathway, we used the KEGG webserver⁷¹ to identify candidate pathways that might involve a glycy radical-based propanediol dehydratase enzyme. The most likely candidate was the glycerolipid pathway (K00128), leading to two products: propionylphosphate and 1-propanol. By mapping the enzymes identified in our cluster 6 onto the glycerolipid metabolism pathway, we found that they could carry out a reaction sequence similar to the one found in Pdu. The putative 1,2-propanediol dehydratase in the Grp MCP would play a role analogous to the PduCDE enzyme in the B_{12} -dependent Pdu MCP; both produce propionaldehyde. This intermediate would then be converted to propionyl-CoA and propanol by the sequential actions of aldehyde dehydrogenase and alcohol dehydrogenase, both found in cluster 6. Based on similarities with the characterized Pdu pathway, the PduL homolog can be integrated into the reaction scheme to carry out the transacylation between propionyl-CoA and propionyl-phosphate. From these reactions, we can draw a

prospective pathway likely to occur within the Grp MCP (Fig. 5).

Detection of N-terminal extensions in the glycyl radical-based propanediol utilization enzymes

Recent experiments have demonstrated that some enzymes in the B₁₂-dependent Pdu system are targeted to the MCP interior via the presence of short N-terminal extensions in their sequences.^{27,30} Bioinformatics studies suggest that an equivalent mechanism might exist in other MCPs as well.^{27,28,31} Here, we asked whether special N-terminal extensions might be evident in the enzymes predicted to be associated with the proposed Grp MCP (Fig. 4). Consistent with our previous calculations, we found that the phosphotransacylase and the putative glycyl radical 1,2-propanediol dehydratase both present special N-terminal extensions.²⁷ Additionally, the enzyme similar to the PduO C-terminus in cluster 7 was also detected as having such an extension. These observations support a model that places those enzymes within (or physically bound to) the MCP.

The approach we used to identify potential targeting signals is based, not on recognition of any particular sequence motif, but on the presence of extended sequences at the termini.²⁷ In addition, following recent work,³¹ we attempted to identify potentially conserved sequence motifs in the enzymes of the Grp MCP operons that might match other established targeting sequences. In our examination we did not judge matches to potential targeting motifs as being statistically significant enough to make clear predictions about other targeting mechanisms in this MCP.

Shell proteins of the Grp MCP

In parallel with our analysis of Protein Functional Groups from cluster 6 and their genomic organization (Fig. 4), we examined the BMC genes that would presumably form the shell of the Grp MCP. In a typical bacterial genome, an MCP operon for glycyl radical-based propanediol utilization contains about four distinct BMC shell protein paralogs. According to current models for MCP architecture,^{5,6,25} a few thousand copies of these proteins would assemble into hexagonally-packed arrays in forming the surface of the shell, with pores allowing for molecular transport (Supporting Information Fig. S1). Also in keeping with other MCP operons, the operons for Grp MCPs code for a gene from the *ccmL/csoS4/pduN/eutN* family, which are presumed to code for minor (e.g., vertex) proteins in the shell.²³

Structural studies of shell proteins from different MCPs have revealed interesting variations that have been ascribed to different functional requirements of the distinct shells. The presumptive BMC shell proteins for the Grp MCP reveal additional variations. For instance, the BMC shell proteins

encoded by Sputw3181_0423 in *Shewanella putrefaciens* and Pecwa_4089 in *Pectobacterium wasabiae* exhibit long C-terminal tails extending beyond the conserved BMC domain; these segments are predicted by IUPRED⁷² to be disordered. Short, flexible extensions of about 10–20 residues have been noted in previous structural studies,^{21,23,73} but the extensions observed here are unusually long, ranging from 80 to 100 residues. Also unusual is the presence of a BMC shell protein (present in eight of the species examined) that is especially divergent from those previously characterized. The occurrence of this shell protein across various species has been integrated into Figure 4. Another interesting feature is the abundance of tandem BMC proteins in the Grp MCP operons of some bacteria, such as *Desulfovibrio salexigens* and *Desulfovibrio desulfuricans* G20, which exhibit up to three tandem BMC proteins (Fig. 4). The apparently novelty of the shell proteins in the Grp MCP is consistent with the expectation that its metabolic features will be different in substantial ways from those explored so far.

As an initial step in characterizing the protein components of a Grp MCP, we overexpressed the unusually divergent shell protein noted above, Pecwa_4094 from *Pectobacterium wasabiae*, in *E. coli* and purified it to homogeneity. As judged by gel filtration, the protein appears to assemble into a homooligomer consistent with a hexamer, by analogy to previously studied BMC shell proteins. Surprisingly, the purified protein was brown in color and an absorption spectrum showed broad peaks at 330 and 420 nm, consistent with a partially oxidized iron-sulfur cluster (Supporting Information Fig. S2). An iron-sulfur cluster has been proposed to occupy the pore of a previously characterized tandem BMC shell protein, PduT, in the Pdu MCP.^{25,74,75} The unusual shell protein from *Pectobacterium wasabiae* reported here is the first single-domain BMC to be identified as a metalloprotein. Further biochemical and structural studies on components of this system will be required to clarify how the shell proteins of the Grp MCP modulate its function.

Discussion

Since the discovery of the carboxysome and the Pdu and Eut MCPs, important clues about MCP composition and function have come from examining sequence data and genomic organization.^{12,76–79} In this study, we have taken those ideas further with an algorithmic approach aimed at classifying new and varied types of MCPs, relying on patterns across more than a hundred bacterial genomes in which BMC shell proteins can be found. Because each type of MCP houses a group of enzymes, a central element of our strategy was to try to automatically group together genes that co-occur strongly in the vicinity of BMC shell proteins. The resulting

clustering strategy was generally effective in detecting genomic patterns and organizing the data in a functionally relevant form. A minor challenge arose from the fact that different MCPs actually contain some overlapping enzyme activities (e.g., alcohol dehydrogenases), whereas the clustering approach naturally seeks to generate separate enzyme groupings. Nonetheless, with some manual analysis to mitigate such challenges, the method was able to classify two types of uncharacterized MCPs, while also illuminating potential variations within previously characterized types, most notably the alpha carboxysome.

Our analysis suggests a model for a microcompartment for B₁₂-independent, glycy radical-based propanediol utilization, for which we have introduced the name Grp. Multiple lines of reasoning support the proposed operation of such a microcompartment. An operon-like organization is evident, with genes for enzymes dispersed among multiple BMC shell genes. The enzymes can be sequentially connected in a pathway for 1,2-propanediol metabolism that resembles the one that occurs in the Pdu MCP, with the key distinction being the initial reaction, which involves a glycy radical enzyme and its activase in the Grp MCP, instead of a B₁₂ cofactor. Further experimental studies will be required to test whether MCPs of the Grp type might be able to metabolize a wider range of substrates, perhaps with different specificities in different bacteria. Finally, multiple enzymes encoded in the operon reveal special N-terminal extensions; this has been established as a key mechanism for targeting enzymes to MCPs.^{27–29}

The encapsulated enzymes and metabolic intermediates provide clues regarding biological advantages that could be offered by the Grp microcompartment. One of the pathway intermediates is the cytotoxic propionaldehyde; retaining such intermediates is recognized as a key role for MCPs.^{10,80} The reactivity of the key glycy radical enzyme offers another clue about potential roles for this MCP. The presence of a glycy radical in the activated state of this enzyme renders it sensitive to oxygen, via oxygen-mediated cleavage of the polypeptide backbone. This property is general to glycy radical enzymes, confining their existence to strictly anaerobic conditions.⁶² The suggestion that the Grp MCP could protect the glycy radical enzyme from destructive oxygen exposure would parallel similar ideas in other MCPs; enzymes in the Pdu, Eut, and carboxysome systems are all sensitive to either damage by or competition with molecular oxygen.⁶⁰

In addition to the core pathway illustrated for propanediol utilization in the Grp MCP, several genomic variations were found across the species examined. A few distinct groups of additional proteins were sometimes present, with different groups appearing in different bacteria (Fig. 4). The

existence of three types of extensions could be postulated from the gene clusters automatically identified. The first extension, involving a histidine kinase and a regulatory sensor, is reminiscent of an analogous pair observed in the Eut MCP (and also present in the proposed Eut system). It occurs in about half of the Grp operons identified here. The similarity suggests that this ubiquitous phosphorylation-based signal transduction mechanism also regulates the Grp system, perhaps in response to propanediol. Another variant of the Grp operon appears in several bacteria in the form of two transcriptional regulators homologous to the tetR family of repressors at the upstream end of the operon, and two genes coding for small multidrug resistance (SMR) family proteins at the downstream end. It seems unlikely that the drug resistance proteins are sequestered in the microcompartment lumen since their primary role is to export small molecules extracellularly.⁸¹ Moreover, their structural properties as transmembrane efflux proteins seem incompatible with the structural requirements of a microcompartment shell. Drug resistance elements are known to be prone to horizontal gene transfer, and their presence in this case may be explained by a necessity to be under the influence of the promoter controlling the expression of the Grp operon. This extended Grp operon occurs almost exclusively in enteropathic bacteria. In another variational form, we identified two proteins homologous to proteins usually found in the Pdu operon, PduF and PduO. PduF is a channel protein transporting small metabolites, potentially facilitating 1,2-propanediol diffusion in this case.⁸² A last variant, highlighted by J. L. Blanchard in *Roseburia inulinovorans*⁶⁹ (but not automatically identified in our bioinformatics analysis), appears to be used for anaerobic fucose and rhamnose degradation. In this variation, which eluded our automatic analysis owing to its relative rarity, the Grp operon is extended by the presence of an aldolase, which is required in a multistep conversion of 6-carbon sugars to 1,2-propanediol, most likely before entering the MCP. Expression profiling data showed that, in the presence of fucose or rhamnose in anaerobic conditions, the aldolase and 12 other genes matching our definition of the Grp operon are indeed found in the top 20 upregulated genes in *Roseburia*.

The genomic variations observed in the Grp system suggest that MCPs providing core metabolic functions can be used differently or modified in distinct ways in diverse bacteria. The existence of a Eut operon variant with a signal transduction system, and alpha carboxysome operons extended by various additional genes in the vicinity of the specialized CsoS1D shell gene,³⁵ are consistent with this general view. For the latter, we speculate that extended functions beyond the canonical alpha carboxysome activities (i.e., CO₂ fixation) would require

the transport of larger molecules, perhaps cofactors, and the specialized shell protein CsoS1D could serve those functions with its larger pore. What additional functions might extend the core CO₂ fixing reactions have not been articulated yet, but preliminary observations implicate bacterioferritin and a homolog of a pterin recycling enzyme in this type of MCP.

Finally, another type of presumptive microcompartment was identified, specific to *Mycobacterium* species. Genes for a group of about four proteins or enzymes are found interspersed with typically two BMC shell protein genes as well as a gene for the minor (vertex) shell protein. The enzymes suggest potential involvement in amino alcohol metabolism. Two of them have been shown to be involved in utilizing amino alcohols such as 1-amino-2-propanol, while other types of proteins occurring in these operons include a class III aminotransferase, an amino acid permease-associated protein, an aminoglycoside phosphotransferase, and a protein of unknown function. By analogy to other MCP systems, the amino alcohol dehydrogenase is likely to represent the key first reaction in some encapsulated pathway. The repression of this enzyme by the GntR transcriptional regulator would be lifted in the presence of the amino alcohol substrate, leading to the expression of structural shell proteins and the enzymes to be encapsulated. The permease, while unlikely to be part of the MCP structure itself, could facilitate uptake of an amino acid or amino alcohol substrate. The structural similarity of 1-amino-2-propanol to ethanolamine raises the possibility that the mycobacterial MCP discussed here could be similar to the Eut microcompartment. However, the presence of distinct groups of enzymes supports a separate classification for this presumptive MCP. Among the distinctive enzymes appearing to be associated with this MCP, the aminoglycoside phosphotransferases are bacterial antibiotic resistance proteins, conferring resistance to many aminoglycosides,⁸³ while aminotransferases have been presumed to play a role in aminoglycoside antibiotics biosynthesis.^{86,87} This suggests potentially interesting connections between this novel MCP and mycobacterial persistence, a dormant phase of host infection sometimes lasting decades,⁸⁸ though we note that an MCP of this type is not found in the *M. tuberculosis* genome. Finally, sequence comparisons (using reciprocal Blast searches) between the putative MCP operons across different mycobacterium species highlighted one well-conserved protein (MSMEG_0274 as in *Mycobacterium smegmatis*) whose function is presently unknown.

The automatic computational analysis of MCP types presented here does not successfully identify all the MCP types proposed in the recent literature. Owing to the statistical criteria applied, the computational approach overlooks potential MCP types that occur in only a few instances across the genomic data. The failure to automatically classify

the Eut MCP was discussed above. In addition, in a few select bacteria Kinney *et al.*³¹ describe potential MCP types based on fucose aldolase as a key encapsulated enzyme; the purpose of such an MCP could be to sequester the lactaldehyde intermediate. Indeed, we note that the fucose aldolase in these operons carries an extended N-terminal domain that could be involved in targeting it to the MCP. An alternate sequence feature in the C-terminal region of fucose aldolase has been indicated as a likely targeting signal by Kinney *et al.*³¹

In summary, our bioinformatics approach has allowed us to more clearly articulate the diversity of MCPs in up-to-date sequenced genomes, and to glean new insights from the organization of their underlying operons. Combining these findings with other recent analyses of MCP operons in the literature, we can assemble a census of all the microcompartment types and variants currently supported by genomic data (Fig. 6). As always with genomic/bioinformatics approaches, the lack of functional annotations for many genes leads to challenges and limitations. Exploring those uncharacterized proteins could be fruitful. Our initial experimental investigation of an unusual shell protein from the proposed Grp MCP supports that view. Meanwhile, further bioinformatics studies are likely to add additional discoveries, for example, using different algorithmic approaches, or using larger data sets as additional genomes are sequenced. Somewhat different approaches may be required to gain insights into systems like the beta carboxosome, where the genomic organization of MCP components is more fragmented. Bearing in mind that bioinformatics studies are only predictive, experimental investigations will be necessary to unravel the biological functions of the new microcompartment types and variations reported here, and to more fully understand the mechanisms by which they operate.

Material and Methods

Operational definition of an MCP operon and BMC-proximal genes

To circumvent the problem of defining true operons across hundreds of microbes, many without experimentally characterized regulatory signatures, we adopted a statistical view of MCP operons. We considered genes encoded within a certain number of open reading frames (upstream or downstream) of one or several paralogous BMC shell genes as potentially belonging to an MCP operon. The general idea of using conserved (bacterial) chromosomal proximity as an indicator of functional linkage has been explored widely and with good success in previous bioinformatics studies.^{87–91} In our application, proteins whose genes satisfy the chromosomal proximity requirement in multiple genomes are statistically likely to be part

Figure 6. An updated classification of MCPs including presumptive types identified computationally. Microcompartments are divided here into seven main classes according to the core enzymes and pathways confined in their lumen; the two carboxysome subtypes are separated here on the basis of their partially distinct compositions. For some of the MCPs, their core functions appear to be augmented by the presence of additional groups of proteins or enzymes; some of these may be directly involved with MCP function while others could be more peripheral (e.g., regulatory). These extensions suggest the existence of more complex or diverse MCP variants. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

of an encapsulated pathway. The strength of this assertion depends on the maximum number of ORFs allowed between a gene in question and a BMC gene; we elected not to consider actual physical intergenic distance or direction of transcription. We refer to this vicinity metric as v . In our analysis, setting v to a maximum value of five yielded the most consistent results in subsequent analyses.

Extracting BMC-proximal proteins

BMC genes were collected from fully sequenced bacterial genomes by scanning for the InterPro profile of the BMC domain (IPR000249)⁹² against the UniProt database (release of September 2011) and mapping the resulting hits onto their corresponding gene names. Of all the fully sequenced bacterial genomes present in Uniprot, 113 genomes were identified as carrying BMC genes; no BMC genes were identified in the archaea. Following our operational definition of an MCP operon, the BMC-proximal genes (i.e., those within v genes of a BMC shell gene) were retrieved from the EBI integr8 database and subsequently mapped to their UniProt ID.⁹⁵ These constituted our starting set of BMC-proximal proteins. For $v = 5$, the dataset gathered a total of 3120 BMC-proximal protein sequences.

Classifying BMC-proximal protein into functional groups

A first assignment of BMC-proximal protein sequences into internally homologous protein families (or groups) was based on Pfam HMM queries of our database using the *hmmer* package.^{94,95} Only those hits reporting an E -value lower than 10^{-4} were considered. Proteins exhibiting the same combination of domains were collapsed into the same group. Of the 3120 proteins, 2616 could be collapsed down to 759 protein groups. We judged that some of these groups contained somewhat divergent sequences, with possibly distinct metabolic functions. Therefore, to increase the sensitivity of our classification, a complementary scan of the dataset searching for KEGG Orthology annotations⁷¹ allowed a subdivision of some of these groups into smaller ones, leading to a total number of 807 groups, referred to as Protein Functional Groups in our analysis.

Pairwise correlation coefficients between protein functional groups

Each possible pair of Protein Functional Groups was examined to see if they tended strongly to occur together within individual MCP operons (Fig. 2). To obtain a correlation coefficient for each Protein

Functional Group pair (A,B) we filled a 2×2 contingency table according to the number of operons exhibiting the following combinations: containing both A and B, containing A without B, containing B without A, and lacking both A and B. A pairwise correlation coefficient (PCC) was obtained by assigning an ordered pair value [(A,B) = (1,1), (1,0), (0,1), or (0,0)] for each operon according to the presence or absence of A and B, and then calculating the Pearson correlation coefficient. This measure ranges from +1 to -1, with the extremes indicating a perfect correlation or a perfect anticorrelation respectively, and 0 indicating independence between the two Protein Functional Groups.

Clustering BMC-proximal functional groups based on pairwise correlations

BMC-proximal functional groups were clustered using a graph-based strategy. In this scheme, each Protein Functional Group is a node, and a correlation between a pair of Protein Functional Groups, if it satisfies our statistical tests, is an edge between those two nodes (Fig. 2). We applied multiple statistical criteria to reduce the likelihood of including spuriously identified proteins and linkages in the final analysis. From the list of the Protein Functional Group pairs, we filtered out those combinations having a PCC lower than 0.5, a threshold below which we considered the correlations likely to be functionally insignificant. We applied two additional statistical tests to assess the reliability of both the edges and nodes. For edge validation, we computed a double-tail Fischer's exact test to cope with unevenly distributed values in the contingency tables that could lead to spurious correlations. In other words, we calculated a *P*-value for an edge between Protein Functional Groups A and B, corresponding to the probability of having B occur by chance at least as many times as actually observed in a set of operons that already have A, under the null hypothesis that both groups occur independently. Edges that had a *P*-value higher than a given threshold—set to 10^{-4} in our study—were discarded from the list, along with any subsequently unconnected nodes. This filtering step had the effect of discarding from further consideration Protein Functional Groups that did not occur enough times to allow statistically significant inferences to be made. Out of a starting set of 807 Protein Functional Groups, 64 survived this filtering step. A final statistical analysis aimed to estimate the strength of the nodes upon variation of our operon definition. Indeed, the edges connected to a specific node can be affected through varying the parameter *v*, since the BMC proximal proteins dataset is built upon its value. For a given node, we analyzed the variation of the set of other nodes to which it was connected when setting *v* from 5 to 10 by a dependent paired

Student's *t*-test (confidence interval equal to 0.95), where the null hypothesis is true if both sample means are not significantly different. Nodes for which the test yielded a value more significant than 10^{-5} were designated as strong, while values inferior to that limit were synonymous for weak nodes. We used the remaining pairs as a basis for clustering.

The final list of Protein Functional Groups and their linkages were clustered using GraphViz(version 2.x) with the “neato” layout (Fig. 3). A node was named either by choosing a consensus string from genomic annotations, when available, or after one of the gene names associated with the node when a consensus could not be established.

Cloning, expression, and purification of the BMC shell protein Pecwa_4094

A codon-optimized version of the Pecwa_4094 gene was synthesized by assembly PCR to include an N-terminal hexa-histidine tag. The gene was cloned into pET22b+ vector via NdeI and XhoI restriction sites. Pecwa_4094 was expressed in BL21(DE3) cells in Luria-Broth at 37°C, shaking at 225 rpm, for 3 h. Cells were pelleted, frozen and stored at -20°C. The cells were resuspended in 50 mM Tris pH 7.6, 300 mM NaCl with protease inhibitors and sonicated until lysed. Lysate was centrifuged at 16,500 rpm to separate soluble and insoluble fractions for 30 min. The soluble fraction was filtered through a 0.22 μ M filter before being loaded onto a 5 mL HiTrap Ni Column at room temperature. Protein was eluted in one step with 50 mM Tris pH 7.6, 300 mM NaCl, 300 mM imidazole pH 8.

Acknowledgments

The authors thank Thomas Bobik, Rob Gunsalus, and Michael Thompson for helpful comments and for critical reading of the manuscript. We also thank Joan Valentine, Kevin Barnese, James Liao and Jenny Takasumi for assistance with anaerobic experiments, and Danny Gidaniyan for assistance in protein expression.

References

1. Gitai Z (2005) The new bacterial cell biology: moving parts and subcellular architecture. *Cell* 120:577–586.
2. Jensen GJ, Briegel A (2007) How electron cryotomography is opening a new window onto prokaryotic ultrastructure. *Curr Opin Struct Biol* 17:260–267.
3. Shively JM, Ball F, Brown DH, Saunders RE (1973) Functional Organelles in Prokaryotes - Polyhedral Inclusions (Carboxysomes) of *Thiobacillus Neapolitanus*. *Science* 182:584–586.
4. Kerfeld CA, Heinhorst S, Cannon GC (2010) Bacterial microcompartments. *Annu Rev Microbiol* 64:391–408.
5. Yeates TO, Thompson MC, Bobik TA (2011) The protein shells of bacterial microcompartment organelles. *Curr Opin Struct Biol* 21:223–231.

6. Yeates TO, Tsai Y, Tanaka S, Sawaya MR, Kerfeld CA (2007) Self-assembly in the carboxysome: a viral capsid-like protein shell in bacterial cells. *Biochem Soc Trans* 35:508–511.
7. Bobik TA (2006) Polyhedral organelles compartmenting bacterial metabolic processes. *Appl Microbiol Biotechnol* 70:517–525.
8. Yeates TO, Kerfeld CA, Heinhorst S, Cannon GC, Shively JM (2008) Protein-based organelles in bacteria: carboxysomes and related microcompartments. *Nat Rev Microbiol* 6:681–691.
9. Raushel FM, Thoden JB, Holden HM (2003) Enzymes with molecular tunnels. *Acc Chem Res* 36:539–548.
10. Sampson EM, Bobik TA (2008) Microcompartments for B12-dependent 1,2-propanediol degradation provide protection from DNA and cellular damage by a reactive metabolic intermediate. *J Bacteriol* 190:2966–2971.
11. Penrod JT, Roth JR (2006) Conserving a volatile metabolite: a role for carboxysome-like organelles in *Salmonella enterica*. *J Bacteriol* 188:2865–2874.
12. Cannon GC, Bradburne CE, Aldrich HC, Baker SH, Heinhorst S, Shively JM (2001) Microcompartments in prokaryotes: carboxysomes and related polyhedra. *Appl Environ Microbiol* 67:5351–5361.
13. English RS, Lorbach SC, Qin X, Shively JM (1994) Isolation and characterization of a carboxysome shell gene from *Thiobacillus neapolitanus*. *Mol Microbiol* 12:647–654.
14. Heinhorst S, Williams EB, Cai F, Murin CD, Shively JM, Cannon GC (2006) Characterization of the carboxysomal carbonic anhydrase CsoSCA from *Halothiobacillus neapolitanus*. *J Bacteriol* 188:8087–8094.
15. Cannon GC, Heinhorst S, Kerfeld CA (2010) Carboxysomal carbonic anhydrases: Structure and role in microbial CO₂ fixation. *Biochim Biophys Acta* 1804:382–392.
16. Bobik TA, Havemann GD, Busch RJ, Williams DS, Aldrich HC (1999) The propanediol utilization (pdu) operon of *Salmonella enterica* serovar Typhimurium LT2 includes genes necessary for formation of polyhedral organelles involved in coenzyme B(12)-dependent 1, 2-propanediol degradation. *J Bacteriol* 181:5967–5975.
17. Bobik TA, Xu Y, Jeter RM, Otto KE, Roth JR (1997) Propanediol utilization genes (pdu) of *Salmonella typhimurium*: three genes for the propanediol dehydratase. *J Bacteriol* 179:6633–6639.
18. Chen P, Ailion M, Bobik T, Stormo G, Roth J (1995) Five promoters integrate control of the cob/pdu regulon in *Salmonella typhimurium*. *J Bacteriol* 177:5401–5410.
19. Garsin DA (2010) Ethanolamine utilization in bacterial pathogens: roles and regulation. *Nat Rev Microbiol* 8:290–295.
20. Stojiljkovic I, Baumler AJ, Heffron F (1995) Ethanolamine utilization in *Salmonella typhimurium*: nucleotide sequence, protein expression, and mutational analysis of the cchA cchB eutE eutJ eutG eutH gene cluster. *J Bacteriol* 177:1357–1366.
21. Yeates TO, Crowley CS, Tanaka S (2010) Bacterial microcompartment organelles: protein shell structure and evolution. *Annu Rev Biophys* 39:185–205.
22. Cheng S, Liu Y, Crowley CS, Yeates TO, Bobik TA (2008) Bacterial microcompartments: their properties and paradoxes. *Bioessays* 30:1084–1095.
23. Tanaka S, Sawaya MR, Phillips M, Yeates TO (2009) Insights from multiple structures of the shell proteins from the beta-carboxysome. *Protein Sci* 18:108–120.
24. Klein MG, Zwart P, Bagby SC, Cai F, Chisholm SW, Heinhorst S, Cannon GC, Kerfeld CA (2009) Identification and structural analysis of a novel carboxysome shell protein with implications for metabolite transport. *J Mol Biol* 392:319–333.
25. Crowley CS, Cascio D, Sawaya MR, Kopstein JS, Bobik TA, Yeates TO (2010) Structural insight into the mechanisms of transport across the *Salmonella enterica* Pdu microcompartment shell. *J Biol Chem* 285:37838–37846.
26. Choudhary S, Quin MB, Sanders MA, Johnson ET, Schmidt-Dannert C Engineered protein nano-compartments for targeted enzyme localization. *PLoS One* 7:e33342.
27. Fan C, Cheng S, Liu Y, Escobar CM, Crowley CS, Jefferson RE, Yeates TO, Bobik TA (2010) Short N-terminal sequences package proteins into bacterial microcompartments. *Proc Natl Acad Sci U S A* 107:7509–7514.
28. Heinhorst S, Cannon GC (2010) Addressing microbial organelles: a short peptide directs enzymes to the interior. *Proc Natl Acad Sci U S A* 107:7627–7628.
29. Parsons JB, Frank S, Bhella D, Liang M, Prentice MB, Mulvihill DP, Warren MJ (2010) Synthesis of empty bacterial microcompartments, directed organelle protein incorporation, and evidence of filament-associated organelle movement. *Mol Cell* 38:305–315.
30. Fan C, Bobik TA (2011) The N-terminal region of the medium subunit (PduD) packages adenosylcobalamin-dependent diol dehydratase (PduCDE) into the Pdu microcompartment. *J Bacteriol* 193:5623–5628.
31. Kinney JN, Salmeen A, Cai F, Kerfeld CA (2012) Elucidating essential role of conserved carboxysomal protein CcmN reveals common feature of bacterial microcompartment assembly. *J Biol Chem* 287:17729–17736.
32. Cot SS, So AK, Espie GS (2008) A multiprotein bicarbonate dehydration complex essential to carboxysome function in cyanobacteria. *J Bacteriol* 190:936–945.
33. Pena KL, Castel SE, de Araujo C, Espie GS, Kimber MS (2010) Structural basis of the oxidative activation of the carboxysomal gamma-carbonic anhydrase, CcmM. *Proc Natl Acad Sci U S A* 107:2455–2460.
34. Baker SH, Lorbach SC, Rodriguez-Buey M, Williams DS, Aldrich HC, Shively JM (1999) The correlation of the gene csoS2 of the carboxysome operon with two polypeptides of the carboxysome in *thiobacillus neapolitanus*. *Arch Microbiol* 172:233–239.
35. Roberts EW, Cai F, Kerfeld CA, Cannon GC, Heinhorst S (2011) Isolation and characterization of the *Prochlorococcus* carboxysome reveal the presence of the novel shell protein CsoS1D. *J Bacteriol* 194:787–795.
36. Tanaka S, Kerfeld CA, Sawaya MR, Cai F, Heinhorst S, Cannon GC, Yeates TO (2008) Atomic-level models of the bacterial carboxysome shell. *Science* 319:1083–1086.
37. Sagermann M, Ohtaki A, Nikolakakis K (2009) Crystal structure of the EutL shell protein of the ethanolamine ammonia lyase microcompartment. *Proc Natl Acad Sci U S A* 106:8883–8887.
38. Hayashi NR, Arai H, Kodama T, Igarashi Y (1997) The novel genes, cbbQ and cbbO, located downstream from the RubisCO genes of *Pseudomonas hydrognothermophila*, affect the conformational states and activity of RubisCO. *Biochem Biophys Res Commun* 241:565–569.
39. Galperin MY, Grishin NV (2000) The synthetase domains of cobalamin biosynthesis amidotransferases cobB and cobQ belong to a new family of ATP-dependent amidoligases, related to dethiobiotin synthetase. *Proteins* 41:238–247.

40. Fresquet V, Williams L, Raushel FM (2004) Mechanism of cohydrin acid a,c-diamide synthetase from *Salmonella typhimurium* LT2. *Biochemistry* 43:10619–10627.
41. Floquet N, Mouilleron S, Daher R, Maigret B, Badet B, Badet-Denisot MA (2007) Ammonia channeling in bacterial glucosamine-6-phosphate synthase (Glns): molecular dynamics simulations and kinetic studies of protein mutants. *FEBS Lett* 581:2981–2987.
42. Holden HM, Thoden JB, Raushel FM (1999) Carbamoyl phosphate synthetase: an amazing biochemical odyssey from substrate to product. *Cell Mol Life Sci* 56: 507–522.
43. Naponelli V, Noiriell A, Ziemak MJ, Beverley SM, Lye LF, Plume AM, Botella JR, Loizeau K, Ravel S, Rebeille F, de Crecy-Lagard V, Hanson AD (2008) Phylogenomic and functional analysis of pterin-4a-carbinolamine dehydratase family (COG2154) proteins in plants and microorganisms. *Plant Physiol* 146: 1515–1527.
44. Hauer CR, Rebrin I, Thony B, Neuheiser F, Curtius HC, Hunziker P, Blau N, Ghisla S, Heizmann CW (1993) Phenylalanine hydroxylase-stimulating protein/pterin-4 alpha-carbinolamine dehydratase from rat and human liver. Purification, characterization, and complete amino acid sequence. *J Biol Chem* 268: 4828–4831.
45. Fitzpatrick PF (2000) The aromatic amino acid hydroxylases. *Adv Enzymol Relat Areas Mol Biol* 74:235–294.
46. Fitzpatrick PF (2003) Mechanism of aromatic amino acid hydroxylation. *Biochemistry* 42:14083–14091.
47. Sutter M, Boehringer D, Gutmann S, Gunther S, Prangishvili D, Loessner MJ, Stetter KO, Weber-Ban E, Ban N (2008) Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nat Struct Mol Biol* 15:939–947.
48. Dobrinski KP, Enkemann SA, Yoder SJ, Haller E, Scott KM (2012) Transcriptional response of the sulfur chemolithoautotroph *Thiomicrospira crunigena* to dissolved inorganic carbon limitation. *J Bacteriol* 194: 2074–2081.
49. Ailion M, Roth JR (1997) Repression of the cob operon of *Salmonella typhimurium* by adenosylcobalamin is influenced by mutations in the pdu operon. *J Bacteriol* 179:6084–6091.
50. Fan C, Bobik TA (2008) The PduX enzyme of *Salmonella enterica* is an L-threonine kinase used for coenzyme B12 synthesis. *J Biol Chem* 283:11322–11329.
51. Del Papa MF, Perego M (2008) Ethanolamine activates a sensor histidine kinase regulating its utilization in *Enterococcus faecalis*. *J Bacteriol* 190:7147–7156.
52. Roof DM, Roth JR (1992) Autogenous regulation of ethanolamine utilization by a transcriptional activator of the eut operon in *Salmonella typhimurium*. *J Bacteriol* 174:6634–6643.
53. Heldt D, Frank S, Seyedarabi A, Ladikis D, Parsons JB, Warren MJ, Pickersgill RW (2009) Structure of a trimeric bacterial microcompartment shell protein, EutB, associated with ethanol utilization in *Clostridium kluyveri*. *Biochem J* 423:199–207.
54. Seedorf H, Fricke WF, Veith B, Bruggemann H, Liesegang H, Strittmatter A, Miethke M, Buckel W, Hinderberger J, Li F, Hagemeyer C, Thauer RK, Gottschalk G (2008) The genome of *Clostridium kluyveri*, a strict anaerobe with unique metabolic features. *Proc Natl Acad Sci U S A* 105:2128–2133.
55. Wackett LP, Frias JA, Seffernick JL, Sukovich DJ, Cameron SM (2007) Genomic and biochemical studies demonstrating the absence of an alkane-producing phenotype in *Vibrio furnissii* M1. *Appl Environ Microbiol* 73:7192–7198.
56. Reizer J, Reizer A, Saier MH, Jr. (1993) The MIP family of integral membrane channel proteins: sequence comparisons, evolutionary relationships, reconstructed pathway of evolution, and proposed functional differentiation of the two repeated halves of the proteins. *Crit Rev Biochem Mol Biol* 28:235–257.
57. Ramos JL, Martinez-Bueno M, Molina-Henares AJ, Teran W, Watanabe K, Zhang X, Gallegos MT, Brennan R, Tobes R (2005) The TetR family of transcriptional repressors. *Microbiol Mol Biol Rev* 69:326–356.
58. Vindal V, Suma K, Ranjan A (2007) GntR family of regulators in *Mycobacterium smegmatis*: a sequence and structure based characterization. *BMC Genomics* 8: 289.
59. Urano N, Kataoka M, Ishige T, Kita S, Sakamoto K, Shimizu S (2010) Genetic analysis around aminoalcohol dehydrogenase gene of *Rhodococcus erythropolis* MAK154: a putative GntR transcription factor in transcriptional regulation. *Appl Microbiol Biotechnol* 89: 739–746.
60. Frey PA, Hegeman AD, Reed GH (2006) Free radical mechanisms in enzymology. *Chem Rev* 106:3302–3316.
61. Selmer T, Pierik AJ, Heider J (2005) New glycol radical enzymes catalysing key metabolic steps in anaerobic bacteria. *Biol Chem* 386:981–988.
62. Buckel W, Golding BT (2006) Radical enzymes in anaerobes. *Annu Rev Microbiol* 60:27–49.
63. Wagner AF, Frey M, Neugebauer FA, Schafer W, Knappe J (1992) The free radical in pyruvate formate-lyase is located on glycine-734. *Proc Natl Acad Sci U S A* 89:996–1000.
64. Vey JL, Yang J, Li M, Broderick WE, Broderick JB, Drennan CL (2008) Structural basis for glycol radical formation by pyruvate formate-lyase activating enzyme. *Proc Natl Acad Sci U S A* 105:16137–16141.
65. Lehtio L, Goldman A (2004) The pyruvate formate lyase family: sequences, structures and activation. *Protein Eng Des Sel* 17:545–552.
66. Vey JL, Drennan CL Structural insights into radical generation by the radical SAM superfamily. *Chem Rev* 111:2487–2506.
67. Bouvet OM, Lenormand P, Ageron E, Grimont PA (1995) Taxonomic diversity of anaerobic glycerol dissimilation in the Enterobacteriaceae. *Res Microbiol* 146:279–290.
68. O'Brien JR, Raynaud C, Croux C, Girbal L, Soucaille P, Lanzilotta WN (2004) Insight into the mechanism of the B12-independent glycerol dehydratase from *Clostridium butyricum*: preliminary biochemical and structural characterization. *Biochemistry* 43:4635–4645.
69. Scott KP, Martin JC, Campbell G, Mayer CD, Flint HJ (2006) Whole-genome transcription profiling reveals genes up-regulated by growth on fucose in the human gut bacterium “*Roseburia inulinivorans*”. *J Bacteriol* 188:4340–4349.
70. Liu Y, Leal NA, Sampson EM, Johnson CL, Havemann GD, Bobik TA (2007) PduL is an evolutionarily distinct phosphotransacylase involved in B12-dependent 1,2-propanediol degradation by *Salmonella enterica* serovar typhimurium LT2. *J Bacteriol* 189:1589–1596.
71. Mao X, Cai T, Olyarchuk JG, Wei L (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 21:3787–3793.
72. Dosztanyi Z, Csizmek V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically

- unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21:3433–3434.
73. Kerfeld CA, Sawaya MR, Tanaka S, Nguyen CV, Phillips M, Beeby M, Yeates TO (2005) Protein structures forming the shell of primitive bacterial organelles. *Science* 309:936–938.
 74. Pang A, Warren MJ, Pickersgill RW (2011) Structure of PduT, a trimeric bacterial microcompartment protein with a 4Fe-4S cluster-binding site. *Acta Crystallogr D Biol Crystallogr* 67:91–96.
 75. Parsons JB, Dinesh SD, Deery E, Leech HK, Brindley AA, Heldt D, Frank S, Smales CM, Lunsdorf H, Rambach A, Gass MH, Bleloch A, McClean KJ, Munro AW, Rigby SE, Warren MJ, Prentice MB (2008) Biochemical and structural insights into bacterial organelle form and biogenesis. *J Biol Chem* 283:14366–14375.
 76. Kofoed E, Rappleye C, Stojiljkovic I, Roth J (1999) The 17-gene ethanolamine (eut) operon of *Salmonella typhimurium* encodes five homologues of carboxysome shell proteins. *J Bacteriol* 181:5317–5329.
 77. Cannon GC, Baker SH, Soyer F, Johnson DR, Bradburne CE, Mehlman JL, Davies PS, Jiang QL, Heinrichs S, Shively JM (2003) Organization of carboxysome genes in the thiobacilli. *Curr Microbiol* 46:115–119.
 78. Badger MR, Price GD (2003) CO₂ concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. *J Exp Bot* 54:609–622.
 79. Lawrence JG, Roth JR (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 143:1843–1860.
 80. Brinsmade SR, Paldon T, Escalante-Semerena JC (2005) Minimal functions and physiological conditions required for growth of *salmonella enterica* on ethanolamine in the absence of the metabolosome. *J Bacteriol* 187:8039–8046.
 81. Bay DC, Rommens KL, Turner RJ (2008) Small multidrug resistance proteins: a multidrug transporter family that continues to grow. *Biochim Biophys Acta* 1778:1814–1838.
 82. Chen P, Andersson DI, Roth JR (1994) The control region of the pdu/cob regulon in *Salmonella typhimurium*. *J Bacteriol* 176:5474–5482.
 83. Kim C, Mobashery S (2005) Phosphoryl transfer by aminoglycoside 3'-phosphotransferases and manifestation of antibiotic resistance. *Bioorg Chem* 33:149–158.
 84. Aoki R, Nagaya A, Arakawa S, Kato C, Tamegai H (2008) Identification and diversity of putative aminoglycoside-biosynthetic aminotransferase genes from deep-sea environmental DNA. *Biosci Biotechnol Biochem* 72:1388–1393.
 85. Nagaya A, Takeyama S, Tamegai H (2005) Identification of aminotransferase genes for biosynthesis of aminoglycoside antibiotics from soil DNA. *Biosci Biotechnol Biochem* 69:1388–1393.
 86. Daniel J, Deb C, Dubey VS, Sirakova TD, Abomoelak B, Morbidoni HR, Kolattukudy PE (2004) Induction of a novel class of diacylglycerol acyltransferases and triacylglycerol accumulation in *Mycobacterium tuberculosis* as it goes into a dormancy-like state in culture. *J Bacteriol* 186:5017–5030.
 87. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* 96:2896–2901.
 88. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* 402:83–86.
 89. Kyrpides NC, Ouzounis CA, Iliopoulos I, Vonstein V, Overbeek R (2000) Analysis of the *Thermotoga maritima* genome combining a variety of sequence similarity and genome context tools. *Nucleic Acids Res* 28:4573–4576.
 90. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* 5:R35.
 91. Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* 28:3442–3444.
 92. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimmma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* 37:211–215.
 93. Mulder NJ, Kersey P, Pruess M, Apweiler R (2008) In silico characterization of proteins: UniProt, InterPro and Integr8. *Mol Biotechnol* 38:165–177.
 94. Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14:755–763.
 95. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39:29–37.
 96. Crowley CS, Sawaya MR, Bobik TA, Yeates TO (2008) Structure of the PduU shell protein from the Pdu microcompartment of *Salmonella*. *Structure* 16:1324–1332.