The Distribution of Polyhedral Bacterial Microcompartments Suggests Frequent Horizontal Transfer and Operon Reassembly Proteins

Farah Abdul-Rahman1, Elsa Petit1 and Jeffrey L Blanchard2,3,4*

1Department of Microbiology; University of Massachusetts Amherst, Amherst, MA 01003 United States of America
2Organismal and Evolutionary Biology Graduate Program; University of Massachusetts Amherst, Amherst, MA 01003 United States of America
3Molecular and Cellular Biology Graduate Program; University of Massachusetts Amherst, Amherst, MA 01003 United States of America
4Biology Department, University of Massachusetts Amherst, Amherst, MA 01003 United States of America

Abstract

Bacterial microcompartments (BMCs) are proteinaceous organelles that carry out specific metabolic reactions. Using domain representations of the BMC shell proteins, we identified BMCs in genomes of 358 bacterial species including human gut microbes, bioremediation agents, cellulosic ethanol producers, and pathogens. Multiple BMCs of different metabolic type are present in 40% of the BMC-containing genomes. BMC genes frequently clustered at a single locus that includes enzymes related to the compartment’s metabolic function. The distribution of BMC-containing species was mapped onto a phylogenetic tree constructed from 16S rRNA. The presence of BMCs was sporadically distributed across the phylogenetic tree. All bacterial families that contained species with BMCs also had species without them. Even within a species, BMC number varied, indicative of frequent horizontal transfer and gene loss. Similarly, phylogenetic trees constructed from individual BMC genes indicated that horizontal gene of the BMC loci is a common occurrence.

Keywords: Operon reassembly; Microcompartments; Alignment

Introduction

Polyhedral protein microcompartments were first identified in Cyanobacteria in 1961 [1]. In Cyanobacteria and chemosynthetic bacteria these microcompartments contain the central enzyme involved in the fixation of carbon dioxide, ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) [2,3], and were hence named carboxysomes. Also associated with the carboxysome is carbonic anhydrase, which converts bicarbonate to carbon dioxide [4]. Carboxysomes are not bounded by a lipid bilayer, but instead are composed of a crystalline complex. Molecular transport across the shell is believed to occur via small pores at the center of each hexameric face of the shell structure [5], but the mechanism(s) which govern transport are not clear.

Genetic analyses resulted in the identification of orthologs of the carboxysome shell proteins in enteric bacteria and subsequently other bacterial families [8–10]. We will refer to these structures as bacterial BMCs (BMCs) although they have also been called polyhedral microcompartments, metabolosomes, polyhedral bodies, and protein microcompartments. BMCs are relatively large macromolecular complexes, 100 to 150 nm in cross section, and contain metabolic enzymes both within and, possibly as integral parts of, the polyhedral shell [9,11].

In Salmonella the shell protein genes are localized within two unusually large operons, comprised of 21 and 17 genes, involved in the catabolism of 1,2-propanediol (pdu-type BMC) and ethanolamine (eut-type BMC) [11,12]. Each BMC contains an adenosylcobalamin (Ado-B12) cofactor requiring enzyme, either propanediol dehydratase or ethanolamine ammonia lyase. In addition to sharing shell structural proteins and a requirement for Ado-B12 each BMC operon also contains acetaldehyde and alcohol dehydrogenase, which lead to the production either of propionate and propanol from 1,2-propanediol (pdu BMCs) or of acetate and ethanol from ethanolamine (eut BMCs). Thus, the BMC pathways can be used as a hydrogen sink when the alcohols are produced or to provide carbon and adenosine triphosphate (ATP) through acetyl/propionyl-coenzyme A and ultimately to pyruvate. BMCs may play a role in protecting the cell from toxic aldehydes [13–16], conserving volatile metabolic intermediates [7] and/or creating separate sub-pools of larger cofactors NAD and coenzyme A (CoA) [17].

In the human gut microbe, Roseburia inulinivorans, microarray analysis led to the identification of a novel BMC-associated B12-independent propanediol dehydratase [18]. The R. inulinivorans BMC functions in the metabolism of the animal host-derived fucose. A similar BMC has also been shown to exist in the forest soil-derived Clostridium phytofermentans, where it functions in the metabolism of fucose and rhamnose [19]. This type of BMC has been coined the glycyl radical prosthetic group-based (grp) BMC [20].

The carboxysome and BMC (eut, pdu and grp) shell proteins share two protein domains. One protein domain, represented by the Protein family (Pfam) database [21] model, (Pfam03319) is the EutN/PduN/CcmL/CsoS4/GrpN family. Proteins having these domains form pentamers which act as the vertices of the polyhedral shell structure and have thus been recently named vertex proteins [22]. The second domain is present in proteins which form cyclical hexamers that are believed to come together and form the faces of the shell [23–25]. This protein domain is represented by the Pfam database model (Pfam00936) and is present in multiple paralogs which number between 3 and 7 genes per BMC locus [25]. For this paper we will refer to proteins with this domain as shell face proteins, although the location and orientation of all family members has not been experimentally determined. Using these two protein domains or representative proteins, new
**BMC** shell protein genes are continuously being identified as more bacterial genomes are sequenced [20,25–28]. This comparative genome approach has led to the discovery of new types of BMCs and orphan BMCs whose metabolic role is still unknown. An additional metabolic type of BMC has been proposed based on genetic, biochemical and comparative genomic analysis that a BMC locus present in *Rhodococcus erythropolis* and *Myxococcus* sp. MCS is involved in aminoketone or aminoketone metabolism [29,30].

The polyhedral shape of BMCs suggests similarities to phage capsid, but there is no evidence for sequence or structural homology. Thus, the origin of the BMC shell proteins is an open mystery. Several BMC loci have been noted to undergo horizontal gene transfer (HGT) [23,25,31]. In this study we conduct comprehensive analysis of the available genomes by first identifying BMCloci using the vertex protein, then mapping the distribution unto a phylogenetic tree built from the 16S rRNA sequences. The diverse functions and taxonomic distribution suggest a complex evolutionary history. By looking at the distribution of BMC type and dispersal patterns across phyla, we also find that operons for coding for the same function have evolved independently several times. Here, we discuss four main phyla that include the most sequenced type and dispersal patterns across phyla, we also find that operons for coding for the same function have evolved independently several times. The BMC vertex protein family sequences obtained above were aligned using ClustalW2 [33], phyML [34] was used to construct our phylogenetic trees using the Maximum Likelihood method with 100 bootstrap replicates. The phylogenetic trees were visualized and manipulated in Dendroscope, [35] Figtree ([http://tree.bio.ed.ac.uk/ software/figtree/](http://tree.bio.ed.ac.uk/software/figtree/)) and using TreeCollapseCL to collapse tree branches based on bootstrap cutoff ([http://emmahodcroft.com/TreeCollapseCL.html](http://emmahodcroft.com/TreeCollapseCL.html)). 16S rRNA sequences were extracted from the NCBI complete genome data set using custom Perl scripts. Twenty-three genomes lacked 16S rRNA sequences and 15 16S rRNA sequences contained stretches of unknown nucleotides (N) or were truncated. These genomes were removed from the analysis. The remaining 16S rRNA sequences from NCBI were aligned to the master 16S rRNA alignment at the Ribosomal Database Project (RDP) website [36]. The taxonomy of each 16S rRNA gene was determined using RDP classifier [37]. Phylogenetic trees were constructed and visualized as described above.

### Determining BMC type based on neighboring enzymes

To identify enzymes that can be used to determine the BMC type (e.g., carboxysome, etu, pdu, grp) protein sequences 20 positions to the left and 20 positions to the right of each vertex protein were retrieved from the PTT files available for each genome at GenBank FTP site ([ftp.ncbi.nlm.nih.gov/genomes/Bacteria/](ftp.ncbi.nlm.nih.gov/genomes/Bacteria/)) using Perl scripts. The protein sequences were then annotated using the COG (Clustered Orthologous Genes) and Pfam databases implemented in the Conserved Domain Database (CDD) [38]. To identify the BMC type enzymes associated with this 40 gene region, we queried using RPS-BLAST for Pfams belonging to diagnostic enzymes. For the carboxysome, we used pfam02788 and pfam00101 (RubisCO large subunit (rbcL) and RubisCO small subunit (rbcS)). For the etu-type BMC, we used pfam06277, pfam06751 and pfam05985 (activating enzyme (eutA), ethanolamine ammonia-lyase large subunit (eutB) and ethanolamine ammonia-lyase small subunit (eutC)). For the pdu-type BMC, we used pfam02286 (B12-dependent propanediol dehydratase large subunit (pduD)). For the grp-type BMC, we used and pfam01228, pfam04055 and pfam13247 (Glycyl radical domain (Gly_radical), the radical S-adenosylmethionine domain (radical_SAM) and pyruvate formate lyase (PFL)).

### Results

#### Taxonomic distribution of BMCs

The BMC vertex protein was chosen for identifying BMC loci and constructing the phylogenetic relationship among BMCs, because it is usually present as a single copy, whereas the shell face proteins are frequently present as multiple paralogs. BMC vertex proteins were identified in 358 of the 2458 genomes in the NCBI database (Table 1). None of the 152 genomes sequenced to date from Archaea phyla

| Phylum | Genomes | Genomes with % Genomes |
|--------|---------|-----------------------|
|        | with BMCs | BMCs    |
| Cyanobacteria | 73 | 72 | 98.6% |
| Planctomycetes | 6 | 5 | 83.3% |
| Synergistetes | 4 | 3 | 75.0% |
| Fusobacteria | 6 | 2 | 33.3% |
| Verrucomicrobia | 4 | 1 | 25.0% |
| Firmicutes | 488 | 106 | 21.7% |
| Chlorobi | 13 | 2 | 15.4% |
| Proteobacteria | 1076 | 147 | 13.7% |
| Acidobacteria | 8 | 1 | 12.5% |
| Actinobacteria | 249 | 18 | 7.2% |
| Spirochaetes | 53 | 1 | 1.9% |
| Euryarchaeota | 102 | 0 | 0.0% |
| Bacteroidetes | 95 | 0 | 0.0% |
| Chlamydiae | 84 | 0 | 0.0% |
| Tenericutes | 59 | 0 | 0.0% |
| Crenarchaeota | 48 | 0 | 0.0% |
| Deinococcus-Thermus | 22 | 0 | 0.0% |
| Chloroflexi | 20 | 0 | 0.0% |
| Thermotogae | 14 | 0 | 0.0% |
| Aquificae | 12 | 0 | 0.0% |
| Deferribacteria | 4 | 0 | 0.0% |
| Nitrospirae | 4 | 0 | 0.0% |
| Elusimicrobia | 3 | 0 | 0.0% |
| Dictyoglomi | 2 | 0 | 0.0% |
| Fibrobacteres | 2 | 0 | 0.0% |
| Thermodesulfobacteria | 2 | 0 | 0.0% |
| Caldiversa | 1 | 0 | 0.0% |
| Chrysoygenetes | 1 | 0 | 0.0% |
| Gemmatimonadetes | 1 | 0 | 0.0% |
| Korarchaeota | 1 | 0 | 0.0% |
| Nanoarchaeota | 1 | 0 | 0.0% |
| **Total** | 2458 | 358 | 14.6% |

**Table 1:** Taxonomic distribution of prokaryotic genomes containing BMCs.
Citation: Abdul-Rahman F, Petit E, Blanchard JL (2013) The Distribution of Polyhedral Bacterial Microcompartments Suggests Frequent Horizontal Transfer and Operon Reassembly. J Phylogen Evolution Biol 1: 118. doi:10.4172/2329-9002.1000118

(Euryarchaeota, Crenarchaeota, Korarchaeota and Nanoarchaeota) contain BMCs.

Several phyla including the Bacteroidetes, Chlamydiae and Tenericutes have over 50 genomes represented in the database, but also completely lacked the signature protein for BMCs.

No phylum had vertex proteins in 100% of the genomes sequenced. In the Cyanobacteria, only *Cyanobacterium UCYN-A*, lacks a shell vertex protein.

*Cyanobacterium UCYN-A* has a reduced genome and is lacking many genes related to photosynthesis and the Calvin cycle, including RuBisCo [39]. A RPS-BLAST search for BMC shell proteins using pfam00936 domain supports the loss of all shell components of the carboxysome in *Cyanobacterium UCYN-A*.

The broad sporadic distribution of BMCs in many phyla is illustrated in Figure 1 using the context of a phylogenetic tree based on the 16S rRNA gene (which serves as a proxy for an organismal tree). The broad phylogenetic distribution of BMCs is unusual.

With the exception of Cyanobacteria, all other taxonomic groups contain BMCs in only a fraction of the members even at the family and genus levels (Tables 2,3).

**Many bacterial genomes contain more than one BMC vertex protein**

Many genomes contain more than one vertex protein (Table 4) which is indicative of either multiple BMC loci or paralogs within a locus. A *Melobacteria* genome contains 6 instances of the vertex protein distributed across 4 separate BMC loci. While most BMC loci contain only 1 copy of the vertex protein, there are exceptions, most notably that nearly all α-carboxysomes contain 2 copies. In Cyanobacteria and other carboxysome containing taxa there was only the 1 BMC locus, while many other phyla contained multiple BMCs per genome. Forty percent of the genomes with BMC vertex proteins contained 2 or more separate BMC loci (Table 5).

![Figure 1: Radial 16S rRNA phylogenetic tree from all bacteria in the NCBI complete prokaryotic genomes database. The sequences from genomes that contain at least one BMC vertex protein are highlighted in red. A more detailed version of the tree including the taxon labels will be supplied upon request.](Image)

| Class within Proteobacteria | Genomes | Genomes with BMCs |
|-----------------------------|---------|------------------|
| Gammaproteobacteria         | 488     | 128              |
| Deltaproteobacteria         | 57      | 7                |
| Alphaproteobacteria         | 265     | 6                |
| Betaproteobacteria          | 184     | 4                |
| Epsilonproteobacteria       | 82      | 0                |

| Order within Gammaproteobacteria |
|---------------------------------|
| Enterobacteriales               | 213     | 109              |
| Chromatiales                    | 12      | 6                |
| Acidithiobacillales             | 4       | 4                |
| Alteromonadales                 | 48      | 3                |
| Vibionales                      | 33      | 2                |
| Thiocriciales                   | 24      | 2                |
| Aeromonadales                   | 5       | 2                |
| Pseudomonadales                 | 58      | 0                |
| Pasteurellales                  | 32      | 0                |
| Xanthomonadales                 | 27      | 0                |
| Oceanospirillales               | 13      | 0                |
| Legionellales                   | 11      | 0                |
| Methylococcaceae                | 4       | 0                |
| Gammaproteobacteria_incertae_sedis | 3 | 0 |
| Cardiobacteriales               | 1       | 0                |

| Family within Enterobacteriales |
|---------------------------------|
| Enterobacteriaceae              | 213     | 109              |

| Genus within Enterobacteriaceae |
|---------------------------------|
| Escherichia/Shigella             | 62      | 57               |
| Salmonellæ                      | 31      | 29               |
| Klebsiella                      | 6       | 6                |
| Enterobacter                     | 12      | 5                |
| Yersinia                        | 19      | 3                |
| Pectobacterium                  | 4       | 2                |
| Sodalis                         | 10      | 1                |
| Serratia                        | 7       | 1                |
| Shimwellia                      | 1       | 1                |
| Raoultella                      | 1       | 1                |
| Proteus                         | 1       | 1                |
| Morganella                      | 1       | 1                |
| Citrobacter                     | 1       | 1                |
| Obesumbacterium                 | 12      | 0                |
| Erwinia                         | 8       | 0                |
| Thorsellia                      | 6       | 0                |
| Pantoea                         | 5       | 0                |
| Edwardsiella                    | 4       | 0                |
| Dickeya                         | 4       | 0                |
| Cronobacter                     | 4       | 0                |
| Rahnella                        | 3       | 0                |
| Xenorhabdus                     | 2       | 0                |
| Phototrophis                     | 2       | 0                |
| Budvicia                        | 2       | 0                |
| Trabulsiella                    | 1       | 0                |
| Samsonia                        | 1       | 0                |
| Providencia                     | 1       | 0                |
| Hafnia                          | 1       | 0                |
| Arsenophonus                    | 1       | 0                |

Table 2: Taxonomic distribution of BMCs at different phylogenetic levels in the Proteobacteria.
Phylogenetic analysis of the vertex protein

A phylogenetic tree was constructed from all vertex proteins (Figure 2). The vertex protein is short (~100 amino acids) and the bootstrap support for some groupings is weak. However, many robust patterns are evident. Cyanobacteria are present in 2 major groups corresponding to their RuBisCo type. The multiple vertex proteins present in the marine Prochlorococcus and Synechococcus and some chemosautotrophs branch deeply in the tree, suggesting ancient gene duplication. More frequently, vertex proteins belonging to the same phylum did not cluster together (Figure 2A) suggesting separate evolutionary trajectories resulting from horizontal gene transfer.

Distribution of BMC types

BMC types are distinguished by the different enzymes encapsulated within or associated with the compartment (Figure 3). The 5 previously identified types include the α-carboxysome, β-carboxysome, eut, pdu, and grp types. Because the type enzymes are nearly always associated with the vertex protein in the same operon/locus, we searched the 40 genes surrounding the vertex protein for the BMC type enzymes. The results are shown in the context of the vertex protein tree (Figure 2B) and with the taxonomic distribution (Figure 4). The α-carboxysomes and β-carboxysomes formed monophyletic groups. The β-carboxysomes consisted of only Cyanobacteria (Figure 4B). The α-carboxysomes contained equal numbers of Cyanobacteria and Proteobacteria (Alphaproteobacteria, Betaproteobacteria and Gammaproteobacteria) and an Actinobacteria (Figure 4A).

In the other BMC types the vertex protein phylogeny was inconsistent with the distribution of the type enzymes and none of these BMC types formed monophyletic groups (Figure 2B). This indicates multiple origins of each of the other BMC types since vertex proteins belonging to loci of similar function were not most closely related to each other. In the eut-type (Figure 4C), Proteobacteria (50%), composed of mostly Gammaproteobacteria, formed the largest distribution, closely followed by Firmicutes (47%). There were also minor players (less than 5%) including Fusobacteria, Chlororollexi, Synergistetes and Actinobacteria. In the pdu-type (Figure 4E), Proteobacteria (48%), composed of mostly Gammaproteobacteria and some Deltaproteobacteria, along with Firmicutes (47%) formed the largest distributions. There were also minor taxa including Actinobacteria, Fusobacteria and Synergistetes.

| Classes within Firmicutes | Genomes | Genomes |
|--------------------------|---------|---------|
| Clostridia               | 132     | 56      |
| Bacilli                  | 344     | 52      |
| Erysipelotrichia         | 2       | 0       |
| Negativicutes            | 5       | 0       |
| **Orders with Bacilli**  |         |         |
| Bacillales               | 171     | 39      |
| Lactobacillales          | 173     | 13      |

| Families within Bacillales | Genomes |
|---------------------------|---------|
| Listeriae                 | 36      |
| Bacillales 1              | 4       |
| Paenibacillales 1         | 11      |
| Planococaceae             | 2       |
| Sporolactobacillales      | 2       |
| Pasteuriaceae             | 1       |
| Staphylococaceae          | 39      |
| Bacillales 2              | 3       |
| Alicyclobacillales        | 3       |
| Bacillus[Incertae Sedis XII] | 3    |

| Families with Lactobacillales | Genomes |
|------------------------------|---------|
| Lactobacillaceae             | 45      |
| Enterococaceae               | 14      |
| Streptococaceae              | 98      |
| Carnobacteriaceae            | 5       |
| Leuconostocaceae             | 10      |
| Aerococaceae                 | 1       |

Table 3: Taxonomic distribution of BMCs at different phylogenetic levels in the Firmicutes.

| Number of vertex proteins per genome | Genomes |
|-------------------------------------|---------|
| 6                                   | 1       |
| 3                                   | 21      |
| 2                                   | 117     |
| 1                                   | 219     |

Table 4: Genomes with multiple vertex proteins per genome.

| Number of BMC loci per genome | Genomes |
|------------------------------|---------|
| 4                            | 1       |
| 3                            | 15      |
| 2                            | 86      |
| 1                            | 256     |

Table 5: Genomes with multiple BMCs.
We demonstrated that (1) BMCs are present in 11 different bacterial phyla, (2) with the exception of Cyanobacteria, BMC distribution is sporadic even at the family and genus levels, (3) Cyanobacteria and other carboxysome containing taxa only have 1 BMC type per genome, (4) multiple BMC types are present in 40% of the BMC containing genomes, (5) monophyletic clades of α-carboxysome and β-carboxysomes are evident on the vertex protein tree, and (6) there is an incongruence between BMC type and vertex protein evolution for eut, pdu and grp-type BMCs. Overall, our results suggest that, with the exception of carboxysomal loci in Cyanobacteria, BMC loci have undergone considerable HGT. The fraction of BMCs in our study (15%) is similar to an earlier report of 17%, which examined roughly half of the current complete genomes [30]. While the NCBI complete prokaryotic genome database contains over 2500 genomes, sequencing is concentrated in just a few phyla, whereas other phyla have not been extensively sampled for taxonomic breadth. Many phyla in our study contained less than 50 sampled genomes and BMCs may yet still be discovered in those groups.

Ancient origin of carboxysomes

No genomes from Archaea contain BMCs, suggesting that BMCs arose after the divergence of Bacteria from Archaea and that horizontal transfer of BMCs has not occurred between these domains. Where and when did BMCs originate? It is evident that carboxysomes have conferred a great advantage specifically for Cyanobacteria since virtually all of the species belonging to this group have the genomic potential to express these structures (Table 1). This suggests an origin of BMCs that dates back to the origin of Cyanobacteria nearly over three billion years ago [40].

Our phylogenetic analysis of the vertex protein was unable to provide strong support for a sister relationship between the α-carboxysome and β-carboxysomes (Figure 2B). A recent analysis of cyanobacterial genomes indicates that the marine Prochlorococcus and Synechococcus originated within the Cyanobacteria and are not ancestral to all other Cyanobacteria [41]. While it is possible that the α-carboxysome originated within chemoautrophic Proteobacteria, the sporadic distribution of the carboxysome in these bacteria does not provide evidence of an older lineage.

It is interesting that while no chloroplast genomes contain carboxysomes, a carboxysome operon is present Paulinella chromatophora, a freshwater amoeba with photosynthetic endosymbionts of Cyanobacterial origin [42]. This endosymbiotic event is not related to the origins of chloroplast and occurred more recently, about 60 mya, as the result of endosymbiosis of a member of the marine Prochlorococcus/Synechococcus group [43–45]. Why then have no chloroplast genomes retained the carboxysome shell proteins? One possible explanation is that Cyanobacteria did not require this structure early on when the atmospheric CO2 levels were higher and prior to the symbiotic origin of chloroplast. In either scenario of possible origin, the carboxysome is undoubtedly old in evolutionary sense, likely going back at least 2.5 billion years ago to the Great Oxidation event [46].

Recent origin of Eut, Pdu and Grp-type BMCs

The taxonomic segregation of the carboxysomes and the eut, pdu and grp-type BMCs suggest that ecological differences may play important roles in selecting for BMC types. The eut, pdu and grp-type BMCs are important in heterotrophic lifestyles such as those belonging to gut and soil microbes in the Proteobacteria and Firmicutes. There are 1654 genomes sequenced from these two phyla, representing 64% of the total NCBI complete genome database. Thus, it is not surprising that they are the most abundant taxa for each of these 3 BMC types. However, in these phyla the eut, pdu and grp-type BMCs do not seem to be subject to consistent strong selection as they are present in some taxa but absent in close relatives that appear in a similar environment. The limited phylogenetic depth of these BMC types in any taxonomic group suggests that they are of recent evolutionary origin and undergo frequent gene transfer and loss (Figure 2B).
Many of the bacteria with multiple BMC loci (Table 5) have different types of BMCs. It has long been known that *Salmonella* and *Listeria* strains contain both the eut and pdu-type BMCs [23,27,47]. Some strains of *E. coli* contain 1 each of the eut, pdu and grp-type BMCs. *Clostridium phytofermentans* contains 1 eut type and 2 grp types. However, the 2 grp-type BMCs are differentially expressed [19].

**BMC operon evolution**

While there were some loci with no key enzymes in the 40 gene region surrounding the vertex protein, for the most part BMC-related genes clustered closely together on the chromosome. One explanation could be offered by the Selfish Operon theory, which when first suggested, used the carboxysome (euta) type BMC locus as an example supporting the hypothesis [31]. The theory suggests that genes having a weakly selected function are more likely to persist in a population by clustering close to each other. This is because when these genes are located in a single locus they are more likely to be horizontally transferred together and maintain their function. In this manner they are less likely to be lost to genetic drift. The theory goes on to predict that genes having a weakly selected function, are more likely to cluster together than those that are strongly selected for. Interestingly, α-carboxysomal genes are more dispersed on physically disparate regions of the genome than other BMC loci and carboxysomes have been strongly selected for in the Cyanobacteria which is consistent with the Selfish Operon Theory. However, the question remains of why α-carboxysomes, if also strongly selected for, still cluster together.

The distribution of the diagnostic enzymes associated with the eut, pdu and grp-type BMCs is not consistent with the vertex protein phylogeny since vertex proteins belonging to BMC loci of similar function have occurred several times. Many bacteria contain other enzymes in the genome that share sequence similarity with these enzymes, so it is not clear why or when these enzymes need to be encapsulated in a BMC. It is possible that the “free standing” enzymes may replace the BMC type enzymes in the BMC locus leading to different BMC functions. Other enzymes associated with BMC loci also do not have a consistent phylogenetic pattern suggesting frequent association or loss from the BMC locus (results not shown). Detailed phylogenetic analysis of these genes is needed to determine rates of gene gain, replacement and loss from BMC loci to further evaluate the Selfish Operon Theory and other aspects of BMC evolution.

**Acknowledgements**

This research was supported by Cooperative State Research, Extension, Education Service, U.S. Department of Agriculture, Massachusetts Agricultural Experiment Station Project No. MAS009582 to J. B., a Sponsored Research Agreement between Qteros and J.B., and UMass College of Natural Sciences and Mathematics Excellence Initiatives funding to E.P.

**References**

1. Jensen T, Bowen C (1961) Organization of the centroplasm in Nostoc punctiforme. Proc Iowa Acad Sci 68: 89–96.
2. Codd GA, Stewart WDP (1976) Polyhedral bodies and ribulose 1,5-diphosphate carboxylase of the blue-green alga *Anabaena cylindrica*. Planta 130: 323–326.
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. So AK, Espie GS, Williams EB, Shively JM, Heinhorst S, et al. (2004) A novel evolutionary lineage of carbonic anhydrase (epilithon class) is a component of the carboxysome shell. *J Bacteriol* 186: 623-630.
5. Kerfeld CA, Sawaya MR, Tanaka S, Nguyen CV, Phillips M, et al. (2005) Protein structures forming the shell of primitive bacterial organelles. *Science* 309: 936-938.
6. Iancu CV, Ding HJ, Morris DM, Dias DP, Gonzales AD, et al. (2007) The structure of isolated Synechococcus strain WH8102 carboxysomes as revealed by electron cryomicroscopy. *J Mol Biol* 372: 764-773.
7. Penrod JT, Roth JR (2006) Conserving a volatile metabolite: a role for carboxysome-like organelles in *Salmonella enterica*. *J Bacteriol* 188: 2865-2874.
8. Chen P, Andersson DI, Roth JR (1994) The control region of the pdu/cob regulon in *Salmonella typhimurium*. *J Bacteriol* 174: 5474-5482.
9. Shively JM, Bradburne CE, Aldrich HC, Bobik TA, Mehlan JM, et al. (1998) Sequence homologs of the carboxysomal polypeptide CoS1 of the *Thiobacilli* are present in cyanobacteria and enteric bacteria that form carboxysomes - polyhedral bodies. *Can J Bot* 76: 906–916.
10. Stojilkovic I, Bäumler AJ, Heffron F (1995) Ethanolamine utilization in *Salmonella typhimurium*: nucleotide sequence, protein expression, and mutational analysis of the *cchA cohB eutE eutU eutG eutH* gene cluster. *J Bacteriol* 177: 1357-1366.
11. 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.
12. Kofoid E, Rappleye C, Stojilkovic 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.
13. Rondon MR, Kazmierczak R, Escalante-Semerena JC (1995) Glutathione is required for maximal transcription of the cobalamin biosynthetic and 1,2-propanediol utilization (cob/pdu) regulon and for the catabolism of ethanolamine, 1,2-propanediol, and propionate in *Salmonella typhimurium* LT2. *J Bacteriol* 177: 5434–5439.
14. 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.
15. Havemann GD, Sampson EM, Bobik TA (2002) PduA is a shell protein of polyhedral organelles involved in coenzyme B(12)-dependent degradation of 1,2-propanediol in *Salmonella enterica* serovar *Typhimurium* LT2. *J Bacteriol* 184: 1263-1261.
16. 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.
17. Huseby DL, Roth JR (2013) Evidence that a metabolic microcompartment contains and recycles private cofactor pools. *J Bacteriol* 195: 2864-2879.
18. 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.
19. Petit E, LaTouf WG, Coppi MV, Warnick TA, Currie D, et al. (2013) Involvement of a bacterial microcompartment in the metabolism of fucose and rhamnose by *Clostridium phytofermentans*. *PLoS One* 8: e54337.
20. Jorda J, Lopez D, Wheatley NM, Yeates TO (2013) Using comparative genomics to uncover new kinds of protein-based metabolic organelles in bacteria. *Protein Sci* 22: 179-195.
21. Finn RD, Mistry J, Tate J, Cogill P, Heger A, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38: D211-222.
22. Wheatley NM, Gidianyan SD, Liu Y, Cascio D, Yeates TO (2013) Bacterial microcompartment shells of diverse functional types possess pentameric vertex proteins. *Protein Sci* 22: 660-665.
23. Kerfeld CA, Heinhorst S, Cannon GC (2010) Bacterial microcompartments. *Annu Rev Microbiol* 64: 391-408.
24. Yeates TO, Thompson MC, Bobik TA (2011) The protein shells of bacterial microcompartment organelles. *Curr Opin Struct Biol* 21: 223-231.
25. Yeates TO, Crowley CS, Tanaka S (2010) Bacterial microcompartment organelles: protein shell structure and evolution. *Annu Rev Biophys* 39: 185-205.
26. Cannon GC, Bradburne CE, Aldrich HC, Baker SH, Heinhorst S, et al. (2001) Microcompartments in prokaryotes: carboxysomes and related polyhedra. *Appl Environ Microbiol* 67: 5351-5361.
27. Cheng S, Liu Y, Crowley CS, Yeates TO, Bobik TA (2008) Bacterial microcompartments: their properties and paradoxes. Bioessays 30: 1084-1095.

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

29. Urano N, Kataoka M, Ishige T, Kita S, Sakamoto K, et al. (2011) Genetic analysis around aminoalcohol dehydrogenase gene of Rhodococcus erythropolis MAK154: a putative GlnR transcription factor in transcriptional regulation. Appl Microbiol Biotechnol 89: 739–746.

30. Jorda J, Lopez D, Wheatley NM, Yeates TO (2013) Using comparative genomics to uncover new kinds of protein-based metabolic organelles in bacteria. Protein Sci 22: 179-195.

31. Lawrence JG, Roth JR (1996) Evolution of coenzyme B12 synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex. Genetics 142: 11-24.

32. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, et al. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30: 281-283.

33. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948.

34. Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.

35. Huson DH, Scornavacca C (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol 61: 1061-1067.

36. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37: D141-145.

37. Wang Q, Gavrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73: 5261-5267.

38. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, et al. (2013) CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res 41: D348-352.

39. Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, et al. (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. Nature 464: 90-94.

40. Olson JM (2006) Photosynthesis in the Archean era. Photosynth Res 88: 109-117.

41. Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, et al. (2013) Improving the coverage of the cyanobacterial phyllum using diversity-driven genome sequencing. Proc Natl Acad Sci U S A 110: 1053-1058.

42. Marín B, Nowack EC, Melkonian M (2005) A plastid in the making: evidence for a second primary endosymbiosis. Protist 156: 425-432.

43. Marín B, Nowack, EC, Glöckner G, Melkonian M (2007) The ancestor of the Paulinella chromatophore obtained a carboxysomal operon by horizontal gene transfer from a Nitrococcus-like gamma-proteobacterium. BMC Evol Biol 7: 85.

44. Reyes-Prieto A, Yoon HS, Moustafa A, Yang EC, Andersen RA, et al. (2010) Differential gene retention in plastids of common recent origin. Mol Biol Evol 27: 1530-1537.

45. Nowack E, Melkonian, M, Glöckner G (2008) Chromatophore genome sequence of Paulinella sheds light on acquisition of photosynthesis by eukaryotes. Curr Biol 18: 410-418.

46. Flannery DT, Walter MR (2012) Archean tufted microbial mats and the Great Oxidation Event: new insights into an ancient problem. Aust J Earth Sci 59: 1–11.

47. Bobik TA (2006) Polyhedral organelles compartmenting bacterial metabolic processes. Appl Microbiol Biotechnol 70: 517-525.