Journal of Phylogenetics & Evolutionary Biology

Abdul-Rahman et al., J Phylogen Evolution Biol 2013, 1:4
DOI: 10.4172/2329-9002.1000118

Research Article Open Access

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

Farah Abdul-Rahman[1], Elsa Petit[1] and Jeffrey L Blanchard[2,3,4]*

[1]Department of Microbiology; University of Massachusetts Amherst, Amherst, MA 01003 United States of America
[2]Organismal and Evolutionary Biology Graduate Program; University of Massachusetts Amherst, Amherst, MA 01003 United States of America
[3]Molecular and Cellular Biology Graduate Program, University of Massachusetts Amherst, Amherst, MA 01003 United States of America
[4]Biology 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 types 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 sequences. 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 transfer of the BMC loci is a common occurrence.

Keywords: Operon reassembly; Microcompartments; Alignment; Horizontal gene transfer; Phylogenetic tree

Introduction

Polyhedral protein microcompartments were first identified in Cyanobacteria in 1961 [1]. In Cyanobacteria and chemoautotrophs 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 layer of shell proteins similar in appearance to bacterial virus coats [5]. In *Synechococcus* WH8102 there are approximately 250 individual RuBisCO complexes per carboxysome, organized into three to four concentric layers [6]. The carboxysomes appear to be necessary to concentrate carbon dioxide [7] in close proximity to the RuBisCO 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(s) 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 bacterialmicrocompartment (BMCs) although they have also been called polyhedral microcompartment, 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 a propanediol dehydratase or an ethanolamine ammonia lyase. In addition to sharing shell structural proteins and a requirement for Ado-B12 each BMC operon also contains acetaldehyde and alcohol dehydrogenases, which lead to the production either of propionate and propanol from 1,2-propanediol (pdu-type BMCs) or of acetate and ethanol from ethanolamine (eut-type BMCs). Thus, the BMC pathways can be used as hydrogen sinks when the alcohols are produced or to provide carbon and adenosine triphosphate (ATP) through acetyl/propionyl-coenzyme A and ultimately 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 the 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 propanediol utilization-type (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

***Corresponding author:** Jeffrey L Blanchard, Biology Department, 221 Morrill Science Center South, 611 North Pleasant Street, University of Massachusetts Amherst, Amherst, USA, Tel: 413-577-2130; Fax: 413-545-3243; E-mail: jeffb@bio.umass.edu

Received July 27, 2013; **Accepted** September 12, 2013; **Published** September 14, 2013

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

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 genomic 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 analyses. These studies showed that a BMC locus present in *Rhodococcus erythropolis* and *Mycobacterium sp.* MCS is involved in aminoalcohol or aminoketone metabolism [29,30].

The polyhedral shape of BMCs suggests similarities to the 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 BMC loci using the vertex protein, then mapping the distribution unto a phylogenetic tree built from the 16S rRNA gene. 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 coding for the same function have evolved independently several times. Here, we discuss four main phyla that include the most sequenced species, Actinobacteria, Cyanobacteria, Firmicutes and Proteobacteria, to give us a more accurate view of the dispersal of BMCs than in phyla where less sampling exists.

Materials and Methods

Identifying vertex proteins in genome sequences

Predicted proteins from complete bacterial genomes were downloaded from the GenBank FTP (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) site on May 9, 2013. We searched the Pfam representation of the vertex protein (pfam03319) against this database using RPS-BLAST [32]. A cutoff for determining significant hits to pfam03319 was based on E-value (lower than 10^{-5}) and an amino acid length of ~100. All of these proteins had a member of the shell face protein family (pfam00936) within 20 genes on the chromosome, whereas none of the hits immediately above this cutoff contained proximal genes with the pfam00936 shell face protein domain.

Phylogenetic analysis

The vertex protein family sequences obtained above were aligned using ClustalW2 [33]. phylML [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/) and using TreeCollapseCL4 to collapse tree branches based on a bootstrap cutoff of 50% (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, eut, 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://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) respectively). For the eut-type BMC, we used pfam06277, pfam06751 and pfam05985 (activating enzyme (eutA), ethanolamine ammonia-lyase large subunit (eutB) and ethanolamine ammonia-lyase small subunit (eutC) respectively). For the pdu-type BMC, we used pfam02286 (B12-dependent propanediol dehydratase large subunit (pduC)). For the grp-type BMC, we used pfam01228, pfam04055 and pfam13247 (Glycyl radical domain (Gly_radical), the radical S-adenosylmethionine domain (radical_SAM) and pyruvate formate lyase (PFL) respectively).

Results

Taxonomic distribution of BMCs

The BMC vertex protein was chosen for identifying BMC loci and

Phylum	Genomes	Genomes with BMCs	% Genomes with 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%
Deferribacteres	4	0	0.0%
Nitrospira	4	0	0.0%
Elusimicrobia	3	0	0.0%
Dictyoglomi	2	0	0.0%
Fibrobacteres	2	0	0.0%
Thermodesulfobacteria	2	0	0.0%
Caldiserica	1	0	0.0%
Chrysiogenetes	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.

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

Several bacterial 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 supported the loss of all shell components of the carboxysome in *Cyanobacterium* UCYN-A (results not shown).

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 contained 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 *Meliobacteria* genome contained 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

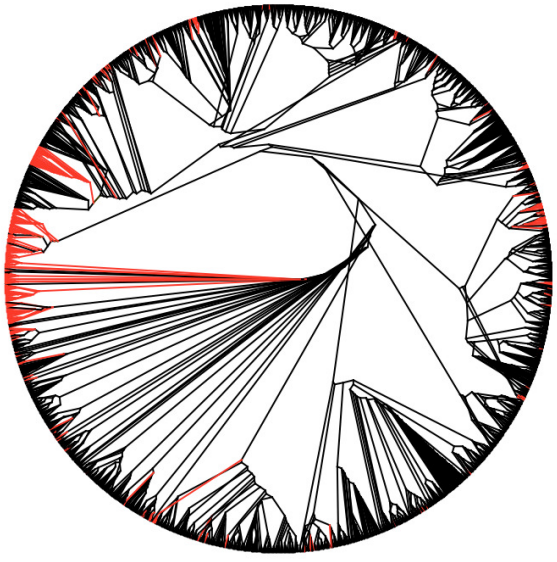

Figure 1: Radial 16S rRNA phylogenetic tree from all bacteria in the NCBI complete prokaryotic genomes database. The sequences from genomes that contain at least one BMC vertex protein are highlighted in red. A more detailed version of the tree including the taxon labels will be supplied upon request.

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
Vibrionales	33	2
Thiotrichales	24	2
Aeromonadales	5	2
Pseudomonadales	58	0
Pasteurellales	32	0
Xanthomonadales	27	0
Oceanospirillales	13	0
Legionellales	11	0
Methylococcales	4	0
Gammaproteobacteria_incertae_sedis	3	0
Cardiobacteriales	1	0
Family within Enterobacteriales		
Enterobacteriaceae	213	109
Genus within Enterobacteriaceae		
Escherichia/Shigella	62	57
Salmonella	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
Photorhabdus	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.

Classes within Firmicutes	Genomes	Genomes with BMCs
Clostridia	132	54
Bacilli	344	52
Erysipelotrichia	2	0
Negativicutes	5	0
Orders with Bacilli		
Bacillales	171	39
Lactobacillales	173	13
Families within Bacillales		
Listeriaceae	31	31
Bacillaceae 1	76	4
Paenibacillaceae 1	11	1
Planococcaceae	2	1
Sporolactobacillaceae	2	1
Pasteuriaceae	1	1
Staphylococcaceae	39	0
Bacillaceae 2	3	0
Alicyclobacillaceae	3	0
Bacillales_Incertae Sedis XII	3	0
Families with Lactobacillales		
Lactobacillaceae	45	5
Enterococcaceae	14	5
Streptococcaceae	98	2
Carnobacteriaceae	5	1
Leuconostocaceae	10	0
Aerococcaceae	1	0

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.

per genome. Forty percent of the genomes with BMC vertex proteins contained 2 or more separate BMC loci (Table 5).

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 boot strap support for some groupings was weak. However, many robust patterns are evident. Cyanobacteria were present in 2 major groups corresponding to their RuBisCO type. The multiple vertex proteins present in the marine *Prochlorococcus* and *Synechococcus* and some chemoautotrophs branched 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 HGT.

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 the 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, Chloroflexii, Synergistetes and Actinobacteria. In the pdu-type (Figure 4E), Proteobacteria (48%), composed of mostly

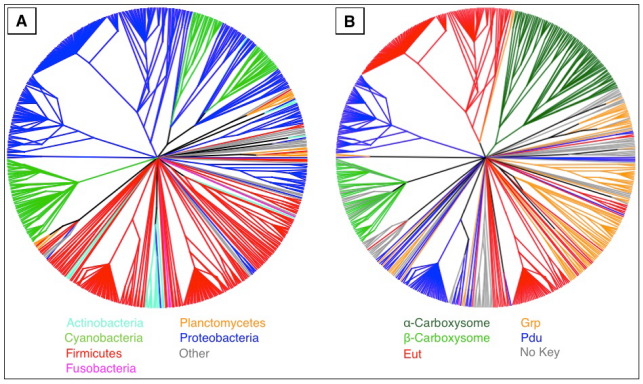

Figure 2. Phylogenetic trees constructed from the BMC vertex protein sequences.
(A) The phylum level taxonomy of the vertex proteins are color-coded according to the key. (B) The predicted functions of the BMC loci are color-coded according to the key. A more detailed version of the vertex protein tree including the taxon labels will be supplied upon request.

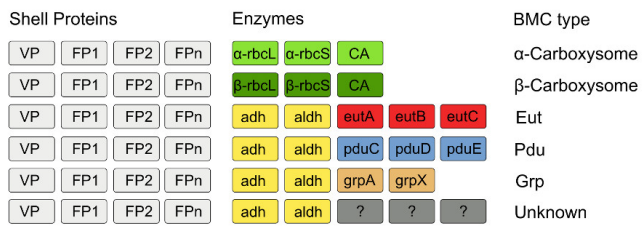

Figure 3: A simple representation of BMC types.
All BMCs contain at least one copy of the vertex protein (VP) and 3 or more shell face proteins (FP). They are functionally distinguished by the presence of different enzymes as shown. There are often other enzymatic (e.g. alcohol dehydrogenase (adh), aldehyde dehydrogenase (aldh) proteins), transcriptional regulatory factors and other genes associated with the loci, but these are not diagnostic.

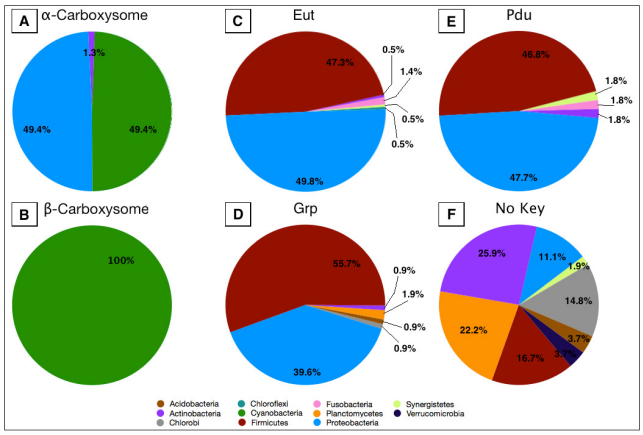

Figure 4: Taxonomic Distribution with each BMC type.
The taxonomic distribution of α-carboxysome (A), β-carboxysome (B), eut (C), pdu (E), grp (D) and unknown (F) were color-coded according to the key with the total BMC loci representing each being 79, 53, 207, 111, 113, and 53 respectively.

Gammaproteobacteria and some Deltaproteobacteria, along with Firmicutes (47%) formed the largest distributions. There were also minor taxa including Actinobacteria, Fusobacteria and Synergistetes. For the grp-type BMC (Figure 4D), Firmicutes (56%) formed the largest distribution, closely followed by Proteobacteria (40%), mostly composed of Gammaproteobacteria and some Alphaproteobacteria and Deltaproteobacteria. The minor taxa included Acidobacteria, Actinobacteria, Chlorobi and Planctomycetes.

Clades with no BMC type enzymes

Since we were looking for specific proteins as key enzymes in a 40 gene window, some BMC types eluded us and were designated as an unknown BMC type. The taxonomic distribution of the BMC loci with no adjacent eut, pdu or grp BMC types was very different and included Planctomycetes, Chlorobi, Chloroflexi, Actinobacteria and a Verrucomicrobium (Figure 4F). Most of these loci included an alcohol dehydrogenase and/or an aldehyde dehydrogenase indicative of the non-carboxysomal type BMCs. However, many of these loci had 2 or 3 vertex proteins, a feature common in α-carboxysomes.

We searched the entire genomes of taxa with unknown BMC type for type enzymes that might be part of a separate operon (Supplementary Table S1). We found that all Cyanobacteria had RuBisCO subunit sequences somewhere in their genomes (results not shown).

Mycobacteria have been proposed to have a different BMC type [30]. Our results showed that all Mycobacteria that had BMC genomic potential also had pdu-type enzymes elsewhere in their genome. Finally, there were three different clades consisting of four Planctomycetes, a Firmicute and a Verrucomicrobium that have an unknown BMC type. Most taxa that had an unknown BMC type had grp-related enzymes somewhere in their genomes which might be an overestimation since radical SAM and pyruvate formate lyase domains are found on other enzymes.

Discussion

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-carboxysomes 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-carboxysome (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 chemoautophic 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 in *Paulinella chromatophora,* a freshwater amoeba with photosynthetic endosymbionts of cyanobacterial origin [42]. This endosymbiotic event is not related to the origins of chloroplasts 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 prior to the symbiotic origin of chloroplasts when the atmospheric CO_2 levels were higher. 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 from the eut, pdu and grp-type BMCs suggests 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 *Escherichia. coli* contain each of the eut, pdu and grp-type BMCs. *C. 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 *Salmonella typhimurium*'s eut-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, β-carboxysome genes, which have been strongly selected for in Cyanobacteria, are more dispersed on physically disparate regions of the genome than other BMC loci which is consistent with the Selfish Operon Theory. However, the question remains of why α-carboxysome genes, 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 do not form monophyletic groups. This phylogenetic topology suggests that the reassembly of BMC loci of similar function has 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 pruniforme*. 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 (epsilon 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 cryotomography. 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 176: 5474-5482.
9. Shively JM, Bradburne CE, Aldrich HC, Bobik TA, Mehlman JL, et al. (1998) Sequence homologs of the carboxysomal polypeptide CsoS1 of the *thiobacilli* are present in cyanobacteria and enteric bacteria that form carboxysomes - polyhedral bodies. Can J Bot 76: 906–916.
10. Stojiljkovic I, Bäumler 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.
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, 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.
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: 1253-1261.
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, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211-222.
22. Wheatley NM, Gidaniyan 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 GntR 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, Garrity 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 phylum using diversity-driven genome sequencing. Proc Natl Acad Sci U S A 110: 1053-1058.

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

43. Marin 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 EC, 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.