A Catalog of the Diversity and Ubiquity of Metabolic Organelles in Bacteria

Markus Sutter[1,2], Matthew R. Melnicki[2], Frederik Schulz[3], Tanja Woyke[3], Cheryl A. Kerfeld[1,2,4]

[1] Environmental Genomics and Systems Biology and Molecular Biophysics and Integrative Bioimaging Divisions, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

[2] MSU-DOE Plant Research Laboratory, Michigan State University, East Lansing, MI 48824, USA.

[3] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

[4] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA.

Corresponding author: Cheryl A. Kerfeld; ckerfeld@lbl.gov

Abstract

Bacterial microcompartments (BMCs) are organelles that segregate segments of metabolic pathways that are incompatible with the surrounding metabolism. These metabolic modules consist entirely of protein, thus they are directly genetically encoded organelles. The BMC membrane is composed of families of proteins that oligomerize into pentagonal and hexagonal building blocks, typically perforated by pores, that tile into a polyhedral shell. The shell protein families are structurally homologous, with the function of the BMC determined by the encapsulated enzymes and the permeability properties of the constituent shell proteins. BMCs can be identified bioinformatically by locating genes encoding shell proteins, which are generally found proximal to those for the encapsulated enzymes. We here performed a large-scale sequence-based analysis of all shell proteins across the bacterial tree of life. With recent advances in genome-resolved metagenomics and the emphasis on "microbial dark matter", many new genome sequences from diverse and obscure bacterial species clades have become available. We find that locus-specific designations of shell proteins should be supplanted, because higher level patterns of co-occurrence are evident. Moreover, the number of identifiable BMC loci has increased twenty-fold since the last comprehensive census of 2014. While we can assign many to bioinformatically characterized loci, the addition of new types uncovered in this study doubles the number of distinct BMC types described. In addition, we predict several new functional types that expand the range of catalysis encapsulated in BMCs, an intriguing example is an organelle for the degradation of an aromatic substrate, compartmentalized in an unusually simple shell, with potential for bioremediation. Our comprehensive survey of bacterial metabolic organelles underscores that there is compartmentalized dark biochemistry yet to be discovered through genome sequencing. The finding of up to six distinct BMC loci in a single genome underscores the role of BMCs in conferring metabolic flexibility and provides new insights into how certain clades have adapted to or even dominate, as in dysbiosis, an environmental niche. Our catalog provides a rich substrate for downstream experimental characterization of these metabolic modules and broadens the foundation for the development of BMC-based nanoarchitectures for biomedical and bioengineering applications.

Main

Bacterial microcompartments (BMCs) are the functional analogue of the membrane-bound organelles of eukaryotic cells. A modular shell comprised of protein surrounds an enzymatically active core, with the shell functioning as a semipermeable membrane for substrates and products (Figure 1a). BMCs were discovered by electron microscopy as polyhedral structures in Cyanobacteria [1]. Carboxysomes enhance CO_2 fixation [2] in all cyanobacteria and some chemoautotrophs by encapsulating RuBisCO together with carbonic anhydrase (CA) to concentrate the substrate CO_2 (Figure 1b, top). Much later, similar structures were observed in heterotrophs, however only when grown in the presence of the substrate of those BMCs, ethanolamine or propanediol [3]. DNA sequencing confirmed that the shell proteins of those BMCs are similar to those of the carboxysomes, a fact that has enabled finding a multitude of BMCs with the advent of genomic sequencing [4,5]. The majority of BMCs are catabolic and are known as metabolosomes. Many diverse types share a common core chemistry of a signature enzyme, that generates an aldehyde, an ubiquitous pfam00171 aldehyde dehydrogenase (AldDh) to oxidize it as well as a phosphotransacylase (PTAC) to generate an acyl-phosphate and an alcohol dehydrogenase (AlcDh) for cofactor regeneration (Figure 1b, bottom) [6]. Targeting of enzymes into the lumen of many BMCs, including the beta-carboxysome [7], typically proceeds via encapsulation peptides (EPs), which consist of a ~15-20 amino acid amphipathic alpha-helix that is connected to the N- or C-termini of cargo proteins via a flexible linker [8].

BMCs have a polyhedral, often icosahedral shape and are about 40-200 nm in diameter [9]. Structures of model shells confirmed icosahedral symmetry with pentagons occupying 12 vertices and hexagons forming the facets [10-13] (Figure 1a). The pentagons (BMC-P) are formed by five subunits of the pfam03319 fold and have the shape of a truncated pyramid [14] (Supplementary Fig. 1a). BMC-H proteins contain the pfam00936 domain and form almost perfect hexagons and with a concave and convex side [15] (Supplementary Fig. 1b). A hexamer-forming subtype also exists, denoted BMC-H^p for permutation, because the order of secondary structure elements has been circularly permuted within the primary structure, with two elements translocated from the C- to the N-terminus (Supplementary Fig. 1e). The facets of shells also contain BMC-T proteins, which are genetic fusions of two pfam00936 domains that form trimers that in size and shape resemble the hexamers. In addition to the standard trimer form, BMC-T^s there is a permuted variant BMC-T^{sp}, in which the secondary structure elements each pfam00936 domain are reordered as in BMC-H^p, There another permuted variant in which two trimers dimerize across their concave faces to form an interior chamber (BMC-T^{dp}) (Supplementary Fig. 1c,d,f). The pores in BMC-T^{dp} trimers are relatively large (14 Å in diameter), gated by conformational changes in the surrounding sidechains and are predicted to the conduits for larger metabolites [16].

Both pfam03319 and pfam00936 protein folds are unique to BMCs and are often encoded together with their cargo proteins in chromosomal loci. This enables predicting and cataloguing of BMC loci within genomic databases. In this study, we have compiled a database of more than 7000 BMC loci that cluster into 68 BMC types or subtypes. In 2014, a comprehensive bioinformatic survey identified 23 types/subtypes of BMC loci across 23 phyla. Now, with the development of culture-independent methods for sampling, we detect BMC loci in 45 phyla across the bacterial Kingdom. We identify 29 new functional

types of BMCs. A phylogenomic analysis of these shell protein data reveals their supra-functional relationships. Collectively our results show that the known BMC functional diversity and distribution at the phylum level has essentially doubled in the last six years and foregrounds the widespread occurrence of bacterial organelles in microbial dark matter.

Results

BMC shell protein and locus analysis

We hypothesized that BMC diversity has greatly expanded since the previous survey (2014) due to expanded genome sequencing of microbial diversity. We compiled and curated an in-house dataset of all putative BMC loci based on the UniProt Knowledgebase (UniProtKB). After retrieving all available BMC shell protein sequences from data available as recently as March 2020, we compiled the sequences for neighboring genes encoded within 12 ORFs from any shell gene. Contiguous genes were classified as a "main locus" when they contained at least one BMC-H and one BMC-P gene, to distinguish from "satellite loci" [4]. Satellite loci were combined with the main locus to form a consolidated locus for genomes containing only a single main locus (Figure 1c).

Because the previous survey of BMC loci [4] relied on pfam co-occurrence, many of the uncharacterized BMC loci were difficult to categorize because they contained few type-specific proteins other than the shell components (unlike the well-characterized PDU and EUT loci with around 20 total gene products; Figure 2a). We therefore sought to improve BMC type clustering by subclassifying all pfam00936 and pfam03319 shell proteins using a phylogenomic approach [17]. Trees were built for representative sequences from each of the six shell protein types: BMC-H, BMC-P, BMC-T^{s}, BMC-T^{dp}, BMC-H^{p} and BMC-T^{sp} (Figure 3). Subclades for each shell type were identified visually, usually containing a long internal stem, and were each assigned a unique color name chosen from the xkcd color survey (https://xkcd.com/color/rgb/), using names from related color families for adjacent subclades within each major clade (for high resolution trees with full annotations, see Supplementary Data: High resolution shell protein trees). The sequences comprising each of the colored subclades were then used to calculate a profile Hidden Markov Model (HMM) [18] for each color group and combined with HMMs derived from proteins common to BMC loci that were clustered by protein pfam information (Figure 4a). Decoupling the identity of a shell protein from a specific BMC type allowed us to make unbiased observations of shell proteins that are similar, despite being components of functionally distinct BMCs. We find that for many BMC loci, the component shell proteins are drawn from across the tree, revealing functional and evolutionary relationships.

The combination of detailed shell protein, enzyme and accessory protein HMMs allowed us to cluster loci into distinct BMC types and subtypes (Figure 4a, see Materials and Methods for details). For example the sugar-phosphate utilization (SPU loci), while generally containing similar components can be separated into seven distinct subgroups (Figure 4b). For the naming of the loci we adopted and expanded the nomenclature from Axen et al [4]. New BMCs were named either by a distinguishing feature such as predominant occurrence in a certain taxon of organisms (e.g. ACI for Acidobacteria) or the organism from which a model system was derived (e.g. HO for *Haliangium ochraceum*), or putative class of BMC substrate

(e.g. ARO for aromatic substrate). In absence of any potentially defining feature, BMCs were classified under broad groupings such as Metabolosomes of unknown function (MUF), Metabolosomes with an incomplete core (MIC) or BMC of unknown function (BUF). MUF BMCs contain an AldDh and enzymes for cofactor recycling such as a phosphotransacylase and/or alcohol dehydrogenase. MIC types contain an AldDh but enzymes for most of the other elements are either absent or not detectable by sequence homology. BUF types lack an AldDh and therefore do not fit into the general metabolosome core chemistry scheme (Figure 1b). In this analysis we have added 29 new major BMC types or subtypes as well as 10 new subtypes for several established loci (Figure 2a). Representative locus diagrams and a short description of the common components for a total of 68 types are listed in Supplementary Data: BMC types.

Overview of BMC Distribution and Characteristics

A shell protein HMM search against 64,495 isolate genomes in IMG (img.jgi.doe.gov) found hits in 21% of the sequenced genomes (April 2020). BMCs occur widely across the Bacterial domain (Figure 5), they are found in 45 different phyla, compared to 23 phyla from an analysis in 2014 [4]. Some BMCs are confined to a single phylum (e.g. ACI, ARO, BUF2) but most are found across multiple phyla (Figure 5). Certain phyla are enriched in BMC functional diversity, such as Proteobacteria, Actinobacteria and Firmicutes (Figure 5), reflecting the importance of metabolic flexibility in diverse niches.

A core enzyme of many functionally distinct metabolosomes (Figure 1b) is the AldDh (Figure 2a). A phylogenetic tree of representative sequences from each BMC type reveals that the AldDh is specific to BMC functional type (Figure 6). AldDh typically have an encapsulation peptide on either the N- and C-termini; strikingly the two major branches of the tree also are distinct in the location of the EP extensions (Figure 6); type I have EP at their C-terminus, type II at the N-terminus. From what is known about experimentally characterized BMC loci we can conclude that AldDh from BMCs with similar substrates cluster on the tree. Accordingly, the AldDh is valid marker for BMC type, and can therefore be used to identify potential substrates for unknown BMCs by looking at the closest AldDh homologues.

In addition to enzymes, BMC loci also contain genes for functions that support the expression and activity of the organelle like transcriptional regulators and membrane transporters for the substrate. Like AldDh, these gene products also can provide clues as to the function of a BMC, by comparison to non-BMC related homologues. For example, for one of the new BMC types that we identified here involves an aromatic substrate (ARO), a SWISS-MODEL search [19] with the protein sequence of the regulator reveals the top characterized hit as a regulator of catechol degradation, which is consistent with the enzymes found in that BMC locus (Supplementary Data: BMC types).

When comparing the shell protein inventories across the BMC types, both unexpected differences and new patterns emerge. On average, BMC loci contain 1.7 BMC-P, 2.2 BMC-H, 0.4 BMC-Hp, 0.5 BMC-T^{s}, 0.4 BMC-Tsp and 0.3 BMC-T^{dp}. The distribution reveals that BMC-P most commonly occur singly or as three paralogs (Figure 2b). Correlating the BMC-P found in the loci that have three copies with their location on the BMC-P tree (Figure 3) we noticed that they follow a “BMC-P triplet” pattern: one member each from the grey and orange major clades and a third member from one of the other clades (Supplementary Fig.

2). Many BMC loci also contain multiples of BMC-H (Figure 2a, b) and for a specific BMC type, those do not necessarily cluster together on a phylogenetic tree. A reason for this could be that the different paralogs fulfill a function that is shared across BMC types. The wide variety of different compositions of the BMC shell highlights its modular construction from building blocks that have the same size and shape, but different permselectivities.

New and Expanded Variants of BMC loci

One of the most prominent expanded BMC type is the SPU (sugar phosphate utilization) BMC (for example locus diagrams of all types, see Supplementary Data: BMC types), first noted in Axen et al [4] as a type with only eight representatives. We have now found more than 280 members and our clustering has identified seven distinct subtypes (Figure 4b, Supplementary Data: BMC types), distributed across 26 bacterial phyla (Figure 5), establishing it as one of the most prevalent BMC types. All SPU loci encode two sugar phosphate processing enzymes (pfam01791, pfam02502). The SPU subtypes are distinguished by the order of genes and presence of additional shell proteins like BMC-T^s and BMC-T^{dp}. The DeoC-type pfam01791 aldolase converts 2-deoxy-D-ribose to glyceraldehyde-3-phosphate and acetaldehyde that can then be processed by the AldDh to acetate. The AldDh of SPU6 is close to EUT3 and the other SPU types are on the same major branch as EUT1, which both process acetaldehyde (Figure 6). The pfam02502 is of the RpiB type [20] that isomerizes ribose-5-phosphate. Collectively, these data suggest potential function of this BMC type is metabolizing the products of DNA degradation, presumably from the ubiquitous detritus available in diverse environments.

The HO BMC from *Haliangium ochraceum* has been the primary model system for structural studies of the shell [10,12,21]. We have identified similar loci in 40 other genomes but its function remains enigmatic; the HO AldDh is most closely related to those of SPU5 and SPU7 (Figure 6), they have the characteristic BMC-P triplet, and share similar types of BMC-T^{dp} (Figure 3d). Some genomes containing the HO loci do have a pfam01791 sugar processing enzyme in a different genomic location that sometimes contains an EP-like N-terminal extension so it could possibly be part of the BMC, suggesting functional similarity similar function of HO types to SPU BMCs.

Another functional type that has substantially increased in membership are the PVMs [22]; in 2014 we found 7 representatives, this has increased to 285 in this dataset (Figure 2). This reflects the increased attention to Verrucomicrobia species for their role in the global carbon cycle as degraders of complex algal and bacterial cell wall polysaccharides [23,24]. For example, Lentimonas species devote 4% of their proteome to the degradation of fucoidan, the major cell wall polysaccharide of brown algae into fucose, which is catabolized in a BMC [24]. The PVM-like BMC that shares the PVM aldolase signature enzyme that processes sugar derivatives also gained 50 new members. Those numbers will likely increase with increased attention on sequencing bacteria that can degrade complex polysaccharides.

Our analysis discovered several completely new BMC types. One, the ARO BMC locus for its predicted aromatic substrate, is found in the Micromonosporales and Pseudonocardiales orders of Actinobacteria. This locus contains two pfam02900 ring-opening oxygenases and a set of enzymes related to the degradation of aromatic aldehyde compounds (Supplementary Data: BMC types). A possible initial

substrate is 2-aminophenol based on the assignment of related pfam00171 AldDh as aminomuconate-semialdehyde dehydrogenase. The ARO AldDh is an outlier on the metabolosome AldDh tree (Figure 6). It does not contain a detectable encapsulation peptide, unlike most other AldDhs. The ARO shell protein composition is among the simplest observed: consisting of one BMC-P and two distinct types of BMC-H. The three ARO shell proteins are all found in late-branching subclades that are strongly divergent from other shell proteins (Figure 3a,b), suggesting that the predicted catalysis of aromatic compounds has imposed a distinctive function on these shell proteins that is under evolutionary constraint.

Another major new type with 125 members, ACI, is found exclusively in Acidobacteria, primarily in only two of the 26 Subdivisions, Subdivision 4 and Subdivision 6; these are predominantly soil organisms. The locus contains a pfam00596 class II aldolase with a C-terminal EP, a hydroxyacid dehydrogenase (pfam00389/02826), a triplet of BMC-P proteins, at least one BMC-H and several proteins of unknown function (Supplementary Data: BMC types). The same aldolase is also observed in the PVM and GRM5 types that process L-fuculose and L-rhamnulose phosphate so a similar function is possible. However, its AldDh is unique and phylogenetically similar only to MIC3, another uncharacterized BMC type (Figure 6). It contains a C-terminal EP and is always found on a satellite locus along with a BMC-H shell protein.

We discovered several enigmatic new BMC types that lack a defined locus organization. These FRAG BMCs are composed of shell protein genes encoded by as many as six different genomic locations. One common feature of FRAG BMCs is the presence of a BMC-P triplet (Figure 3a). Most of these genes are remote from genes for known enzymes, although a few have proximal AldDh (Figure 6, Supplementary Data: BMC types), that map to uncharacterized AldDhs in different parts of the tree (Figure 6). FRAG BMCs are found in diverse organisms, including a large number of candidate phyla (Figure 5).

In addition to those described above, we have also found several more distinct BMC types of unknown function: BUF2, BUF3, GRM6, MIC2-7, MUF2 and MUF3 (Supplementary Data: BMC types). There is a large number of new metabolosome loci that encode a pfam00171 AldDh yet do not have an obvious signature enzyme that generates that aldehyde (ACI, FRAG1-7, HO, MIC2-7, MUF2-3; Figure 2). It is possible that certain enzymes do not co-occur with shell proteins in loci and are therefore difficult to find. Such a case has been recently found for an AlcDh in the PDU BMC of *Acetobacterium woodii* [25] and possibly for the HO BMC mentioned above. Assignment of function for all of those will have to come from biochemical studies of the organisms containing those types, for example as done for the identification of the fucose/rhamnose substrate for PVM type BMCs [22].

BMC type co-occurrence and horizontal gene transfer

About 80% of BMC-containing genomes encode only one locus, but a substantial number encode two (20%) or more (2%) loci. The most frequent co-occurring pair are the PDU and EUT loci (Supplementary Table 1), providing catabolism for ethanolamine and 1,2-propanediol. In *Salmonella enterica*, this combination has been shown to be regulated mutually exclusively to prevent formation of mixed BMCs [26]. Most of the genomes encoding three or more BMC types are combinations of EUT, PDU and GRM (Supplementary Table 2). The most extreme examples of the potential to form multiple metabolic organelles is the genome of the Firmicute *Maledivibacter halophilus* that contains six loci: two EUT2B and

one each of PDU1D, GRM1A, BUF1 and BUF3. This organism is found in anoxic hypersaline sediment [27]. All of the organisms with three or more BMC types are either Proteobacteria or Firmicutes, predominantly from the Enterobacterales or Clostridiales orders. The ability to form multiple functionally distinct BMCs would confers metabolic potential and flexibility; for example, the prevalence of multiple BMCs in the Firmicutes (Clostridiaceae, Ruminococcaceae) likely contributes to their potential for breakdown complex polysaccharides [28].

BMC loci are genetic modules, a compact organization of the structural, regulatory and ancillary components necessary for BMC function. As such, they are an ideal target for horizontal gene transfer (HGT). One example of a possibly recent HGT event involves a EUT2I locus in *Oceanotoga teriensis*, the only BMC instance in the Thermotogae phylum. About 13% of the proteome of this organism shares >30% sequence identity with BLAST hits against Firmicutes (img.jgi.doe.gov) so the BMC locus is likely part of large horizontal gene transfer from a Firmicute. The closest relative to the BMC in this organism according to locus scoring is a EUT2I from *Clostridium scatologenes*, indicating a potential origin. In the case of RMM2 there are two phylum outliers, one in Epsilonproteobacteria and one in Candidatus Marinimicrobia (Supplementary Data: Locus diagrams). Both of them have phage integrase proteins (pfam13356 and pfam00589 domains) right next to the BMC locus, indicating a likely transfer via a phage vector.

Satellite loci

We define satellite loci (Figure 1c) as loci distal from the main locus and containing either shell proteins only, or a combination of shell proteins and enzymes. While some proteins found in satellite loci never occur in the main locus, others likely originated from a fission event of the main locus. Examples for the former are the carboxysomal CcmK3/4 proteins [29,30] and for the latter the satellite locus for type I HO BMCs that contains two BMC-P with an aldolase that are found in the main locus of type II HO BMCs (Supplementary Data: BMC types). We have found satellite loci in 18 different BMC types (with a threshold of >20% of the members that have satellites, Figure 2), highlighting their importance. The satellite locus of the β-carboxysome is one of the few to be experimentally characterized; the CcmK3/4 BMC-H proteins have been shown to form heterohexamers, and may serve as a means to regulate shell permeability under changing environmental conditions.

In HO, SPU3 and SPU6 BMC types the satellite loci resemble “EUT modules”, consisting of the ethanolamine degradation signature enzymes EutA/B/C and a EutL type BMC-T^{sp} shell protein (Supplementary Fig. 3). Since those BMC types are not expected to primarily process ethanolamine this could represent a functional extension of the main BMC to use ethanolamine as an alternate substrate, with the BMC-T^{sp} acting as a shell protein that facilitates entry and the EutA/B/C enzymes to process it. In the case of SPU6, there is even indication of integration of the satellite locus into the main locus, replacing the SPU type signature enzymes and the resulting locus is similar to EUT3 type loci (Supplementary Fig. 3). EUT3 is a new type of ethanolamine utilization BMC with an AldDh that is phylogenetically distinct from the ones from both EUT1 and EUT2 loci (Figure 6) and it contains a BMC-T^{dp} shell protein, unlike any other known EUT type BMC (Figure 3d). Phylogenetic trees validate the link between the two locus types; they are on the same major branch of the AldDh tree (Figure 6) and shell proteins like the BMC-T^{dp} are

also adjacent, with the members of the fused locus found at the base of the SPU6 part of the branch (Figure 3d), an example of how the expanded survey plausibly recounts evolutionary history.

Shell protein trees

From all the BMC loci we have collected more than 40,000 shell protein sequences. Almost 19,000 of them are unique, highlighting their diversity despite a common function to form BMC shells. Among them are 4900 BMC-P, 8000 BMC-H, 1550 BMC-H^p, 1700 BMC-T^s, 1600 BMC-T^{sp} and 1000 BMC-T^{dp}. A further reduced set of those was used to build phylogenetic trees that cover their diversity (Figure 3).

The BMC-P proteins resolve into five major clades: green, blue, purple, grey and orange (Figure 3a). Representatives of grey and orange clades always co-occur as two of three members of a BMC-P triplet. The third member of a given triplet varies across locus types, and is drawn from the green, blue or purple clades (Supplementary Fig. 2). The BMC-P proteins from these three clades frequently occur as the sole BMC-P gene in loci that lack BMC-P paralogs. The only exception is the alpha-carboxysomal CsoS4A and CsoS4B, BMC-P proteins that form a long stem of the purple clade; they always co-occur as pair. The presence of distinct BMC-P paralogs is unexpected because only 12 pentamers are needed to cap polyhedral structures and hints at additional functions of these. As opposed to the pfam00936 domains the BMC-P proteins are very homogeneous in length; there so far no observed instances of large extensions, domain duplication or permutation. Duplication and permutation are limited by a buried N-terminus and the most conserved region is the residues that have been observed to interact with BMC-H proteins, which is likely their primary function.

The defining, conserved primary structure of the BMC-H proteins, the pfam00936 domain is about 80 amino acids in length. However, extensions of up to 200 residues are observed, most frequently at the C-terminus. Each BMC locus type contains at least one BMC-H found close to the base of the tree (Figure 3b). We refer to these as “basal” BMC-H proteins, because they share sequence motifs including the highly conserved inter-hexamer interface residue motifs KAAN and (P/A)RPH of the facets [10]. In experimentally characterized shells, these basal hexamers are known to form the bulk of the BMC shell facets and likely conduct the largest portion of metabolites. We predict that they are interchangeable among functionally distinct BMCs, as has been shown for the PDU BMC [31] and for synthetic chimeric carboxysome shells [32]. The basal BMC-H sequences are also less likely to have C-terminal extensions, possibly reflecting constraints on their function as the major building block of the shell.

There are two major types of the permuted hexamer (BMC-H^p,) the GrpU type that is predicted to contain an Fe-S cluster in the pore [33] and the EutS/PduU type that forms an unusual beta-barrel protruding from the convex side [34] (Figure 3e). These, as well as all other EutS/PduU homologues always co-occur with a pfam10662 type EutP/PduV type protein, which have been shown to interact in the PDU system [35] and are likely involved in BMC positioning in the cell [36]. We observe these pairs of proteins frequently in satellite loci, suggesting that a positioning function is separately regulated from main BMC locus.

Characterization of BMC-Tdp proteins has long been limited to carboxysomal members CcmP and CsoS1D, as the two metabolosome model systems, EUT1 and PDU1, lack these proteins. In our analysis, we find BMC-T^{dp} proteins in a large variety of BMC types (Figure 2, Figure 3d). The BMC-T^{dp} tree can be divided

into four major clades. The yellow clade contains members from the RMM1 and RMM2 BMC types. An unusual outlier is found at the base of the tree with homologs found in PDU1D loci from Proteobacteria and a single Acidobacterium; no other PDU type BMCs contain BMC-T^{dp} shell proteins. The blue major clade contains a variety of uncharacterized BMCs (MIC12/BUF23) as well as EUT3 and SPU6 members that are in adjacent clades. The largest number of sequences are found in two clades that can be characterized by the presence of carboxysomal members. The purple clade contains the alpha-carboxysomal CsoS1D and a number of CsoS1D-like proteins from mainly HO and SPU3 and SPU5 type loci. The green major clade contains the beta-carboxysomal CcmP as well as CcmP-like proteins from HO, MIC3, SPU2 and SPU3 type loci. The proximity of the BMC-T^{dp} of SPU to carboxysome ones indicates parallels with regard to the molecules that enter or leave the shell through these proteins. Sugar phosphates seem a likely candidate that both BMC types have in common, however BMC-T^{dp} from the other two major clades do not all share that type of substrate. It is possible that those BMCs have adopted the gated shell proteins for other purposes, such as large cofactors. Universally conserved across all types are the residues for the gating mechanism, indicating that this is a crucial function of the BMC-T^{dp}.

The BMC-T^{s} type proteins form trimers that are structurally equivalent to hexamers in size and shape and likely evolved as tandem fusions of two BMC-H sequences. Phylogenetically there are three distinct groups, the beta-carboxysome exclusive CcmO, a cluster of sequences that appear to be recent fusions, and a large group of proteins related to PduT (Figure 3c). There is a second family of BMC-T^{s} proteins that have permuted secondary structural elements. These BMC-T^{sp} proteins group into two distinct clusters related to EutL or PduB (Figure 3f) and there is a strong correlation between the verified or most likely substrates of the BMC type and the position of its BMC-T^{sp} on the tree.

The fine classification of shell proteins allows us to find links between BMC types that share an evolutionary history, such as the close link between EUT3 and SPU6 replicated in the similarities of their BMC-T^{dp}. Another example can be found in the long known PDU1AB and GRM4 types. The “colors” of all shell proteins of the GRM4 type are also found in PDU1AB (Supplementary Data: Locus diagrams). Comparison of the locus diagrams reveals the same gene order, with the PDU signature enzymes replaced by the GRM4 signature enzyme. Such a transformation without changing the shell proteins is facilitated in this case by the fact that both BMCs have the same propanediol substrate. This highlights the importance of the proteins that form the shell and suggests that finding the closest homologues from known BMCs, and the combination of color types, may be useful in predicting function in the absence of any information about the encapsulated enzymes.

As opposed to the enzymes found in BMCs, which have homologues that are not BMC-related, the shell proteins are found exclusively associated with BMC loci. Scoring a proteome with the collection of BMC-type specific shell protein HMMs allows for quick assessment of both the presence of BMCs and initial identification due to their specificity. We have used our shell protein HMMs to score data from 26,948 metagenomes in IMG/M [37] and found hits in 15,604; the total shell protein gene count adding up to more than 1.7M. Using the assumption that the co-occurrence of the type specific BMC-H and BMC-P colors are a good indication of the presence of a certain BMC type (or types that use the same or closely

related substrates), we can make initial predictions of the kinds of encapsulated metabolism to be found in a metagenome (Supplementary Fig. 4). The distribution of BMC types, such as the prominent occurrence of PVM and SPU, provides a diagnostic marker of the specialized metabolism required in diverse nutritional landscapes of environmental samples.

Discussion

We present here an updated inventory of bacterial microcompartment loci. The classification is important to identify variations on already characterized BMCs as well as discovering new types of BMCs, only some for which we can propose functions. Classification of BMCs is important to recognize their distribution, especially when BMC types occur across diverse phyla as observed here. Common factors like environment can then give clues as to what potential substrate uncharacterized BMCs process. As bacteria from more diverse environments are sequenced, we will likely find more types of BMC, as exemplified by the large expansion of the SPU type loci that consisted of only a handful of members in 2014 and is now represented by more than 280 loci across 26 bacterial phyla.

Our approach of a functional agnostic classification of shell proteins revealed interesting correlations across BMC types. A striking example is the presence of the three distinct BMC-P in many, mostly new, loci. For capping of the icosahedral shell, only 12 BMC-P pentamers are needed so having three different ones indicates an additional function for at least some of the paralogs. Possible functions could be during shell assembly or, if it is possible that pentamers are dynamically dissociating-reassociating from mature shells as has been suggested [38], they could have some sort of gating function. BMC-P proteins are also surprisingly specific to BMC type, which would not be expected from a pure structural function of capping polyhedral structures. Another shell protein type that has seen an expansion is the BMC-T^{dp} type that is found in a variety of BMC types and will be an interesting target for engineering of BMCs due to its gating functionality. In the context of their color combinations, mosaicity of many BMC shells indicates a general compatibility of BMC shell proteins that allows combining shell protein parts from different BMC types for engineering of designed BMCs. The newly discovered BMC types expand the types of chemistry performed by BMCs and indicate that there are possibly more, yet undiscovered types. The wide distribution of BMCs we observe here indicates that they play an important role for bacteria that allows them to succeed in otherwise inaccessible environments.

Figures

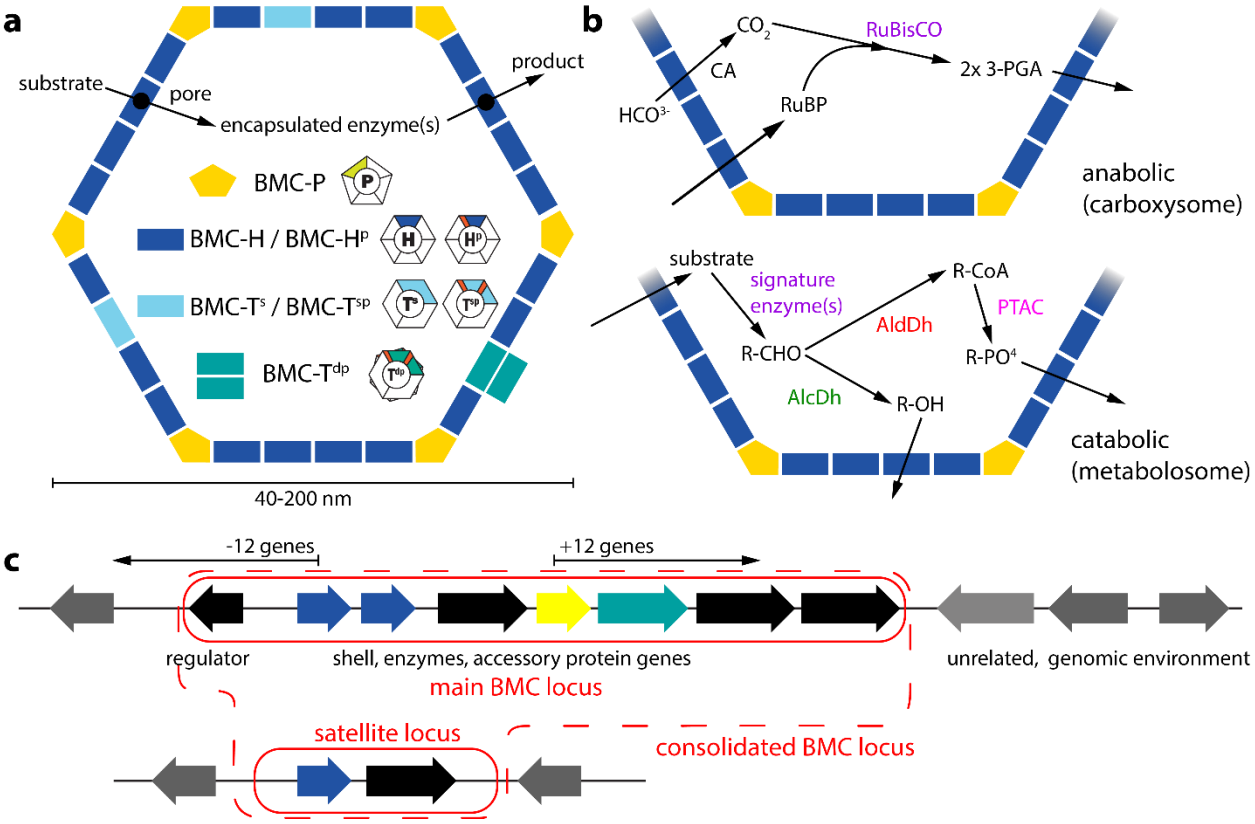

Figure 1. Generalized BMC structure, function and chromosomal organization of component genes. **a**, Overview of a BMC shell and types of protein components. **b**, Simplified reactions of anabolic and catabolic BMCs. CA: carbonic anhydrase, RuBP: ribulose 1,5-bisphosphate, 3-PGA: 3-phosphoglycerate, AlcDh: alcohol dehydrogenase, AldDh: aldehyde dehydrogenase, PTAC: phosphotransacylase, CoA: coenzyme A. **c**, Typical BMC locus consisting of genes for shell proteins, enzymes, regulators and ancillary proteins such as cell membrane transporters for substrates; collected genome fragments for analysis ranges from 12 genes before the first shell protein to 12 genes after the last shell protein encoded in the locus. The combination of main and satellite locus is termed a consolidated BMC locus.

a

BMC type name	number of loci	number in Axen et al	AldDh	satellites	average # of genes	# of phyla		BMC-H	BMC-Hp	BMC-Ts	BMC-Tsp	BMC-Tdp
ACI	125		40-80% of loci	>40% of loci	9	1	3	2				
ARO	17		>80% of loci		12	1	1	2				
BUF1	46	4			13	4	1	3				
BUF1B	55				13	1	1	3				
BUF2	8				5	1	1	2				1
BUF3	10				13	2	1	2				1
CsomeACh	214	20		20-40% of loci	10	6	2	2				1
CsomeACy	125	20			10	1	2	2				1
CsomeBCy	393	46		>40% of loci	11	1	1	4		1		1
ETU	3	2	>80% of loci	N/A	31	1	1	4	1	4	1	
EUT1	1377	72	>80% of loci		19	1	1	2	1		1	
EUT2A	169	12	>80% of loci		19	4	1	2	1		1	
EUT2B	145	9	>80% of loci		20	5	1	2	1	1	1	
EUT2C	88	2	>80% of loci		17	1	1	2	1		1	
EUT2D	28	2	>80% of loci		18	1	1	2	1	1	1	
EUT2E	75		>80% of loci		14	1	1	2	1		1	
EUT2F	59		>80% of loci		18	1	1	2	1		1	
EUT2G	28		>80% of loci		14	1	1	1			1	
EUT2H	38		>80% of loci		17	1	1	2	1		1	
EUT2I	159		>80% of loci		19	3	1	2	1	1	1	
EUT2K*	4	2			8	1	1	1			1	
EUT2x	23		40-80% of loci		10	1	1	3				
EUT3*	17		>80% of loci		14	3	1	1			1	1
FRAG1	56			N/A	7	5	3	2		1		
FRAG2	33		40-80% of loci	N/A	10	7	3	2				
FRAG3	8			N/A	6	3	3	2				
FRAG4	4			N/A	6	1	3	1				
FRAG5	3		>80% of loci	N/A	10	1	2	2		1		2
FRAG6	5			N/A	8	2	3	2				
FRAG7	2		>80% of loci	N/A	14	1	3	1		1		
FRAG8	7		40-80% of loci	N/A	8	2	2	1				
GRM1A	293	35	>80% of loci		19	5	1	4	2			
GRM1B	25		>80% of loci		20	2	1	3	1	3		
GRM2	252	16	>80% of loci		14	1	1	4				
GRM3A	84	19	>80% of loci		20	1	1	3	1	1	1	
GRM3B	92		>80% of loci		19	3	1	2	1		1	
GRM3C	24		>80% of loci		18	1	1	2	1		1	
GRM4	8	2	>80% of loci		14	1	1	3			1	
GRM5	202	5	>80% of loci	>40% of loci	24	1	1	5	2	1	1	
GRM5like	44		>80% of loci	>40% of loci	17	1	1	5	1	1		
GRM6	19		>80% of loci		15	1	1	2			1	
GRMguf	18		>80% of loci		15	2	1	3		1		

BMC type name	number of loci	number in Axen et al	AldDh	satellites	average # of genes	# of phyla	BMC-P	BMC-H	BMC-Hp	BMC-Ts	BMC-Tsp	BMC-Tdp
HO	41	1	>80% of loci	>40% of loci	10	2	3	1		1		1
MIC1	45	3	>80% of loci		13	3	1	2				1
MIC2	22		>80% of loci		7	4	1	1				1
MIC3	11		>80% of loci	>40% of loci	14	1	5	2		1		1
MIC4	15		>80% of loci	20-40% of loci	13	5	4	2		1		
MIC6	3		>80% of loci		13	1	1	1	1	1		2
MIC7	8		>80% of loci	>40% of loci	9	1	3	2		1		1
MUF1	7	2	>80% of loci		12	1	1	3		1	1	
MUF2	19		>80% of loci	>40% of loci	14	2	3	4		1		
MUF3	9		>80% of loci		12	1	1	3		1		
PDU1AB	1292	49	>80% of loci		21	1	1	3	1	1	1	
PDU1C	89	3	>80% of loci		20	1	1	3	1		1	
PDU1D	181	2	>80% of loci	20-40% of loci	22	10	1	3	1	1	1	
PDU1E	43		>80% of loci		22	1	1	3	1		1	
PDU1F	44		>80% of loci		19	1	1	2			1	
PVM	285	7	>80% of loci	20-40% of loci	13	5	3	2				
PVMlike	56	6	>80% of loci		9	1	1	4				
RMM1	140	4	>80% of loci		12	4	1	1				2
RMM2	32	4	>80% of loci		9	5	1	1				2
SPU1	62		>80% of loci	20-40% of loci	13	1	3	2		1		
SPU2	53		>80% of loci		11	6	4	2		1		1
SPU3	70		>80% of loci	20-40% of loci	10	14	3	1		1		1
SPU4	33	8	>80% of loci		14	3	3	2				
SPU5	8		>80% of loci		8	4	3	1		1		1
SPU6	46		>80% of loci	>40% of loci	12	2	3	1				1
SPU7	8		>80% of loci	>40% of loci	12	4	3	1	1	1		

AldDh: >80% of loci; 40-80% of loci
satellites: >40% of loci; 20-40% of loci

b

	P	H	H^p	T^s	T^p	T^sp	any T
BMC types with 0	0	0	45	40	43	49	15
BMC types with 1	44	15	21	26	25	15	31
BMC types with 2	4	32	2	0	0	4	18
BMC types with 3	17	13	0	1	0	0	3
BMC types with 4	2	6	0	1	0	0	0
BMC types with 5	1	2	0	0	0	0	1

Figure 2. Overview of BMC types and shell protein content. **a**, Table of number of loci for each BMC type in this study and Axen et al, AldDh occurrence, prevalence of satellite loci, average number of genes in the locus, number of observed phyla and number of each type of shell protein. Major new BMC types or subtypes identified in this study in bold. The asterisk denotes the name change of EUT3 in Axen et al. to EUT2K because this analysis indicates EUT2K is not a major new type. **b**, Numbers of each type of shell protein across BMC types.

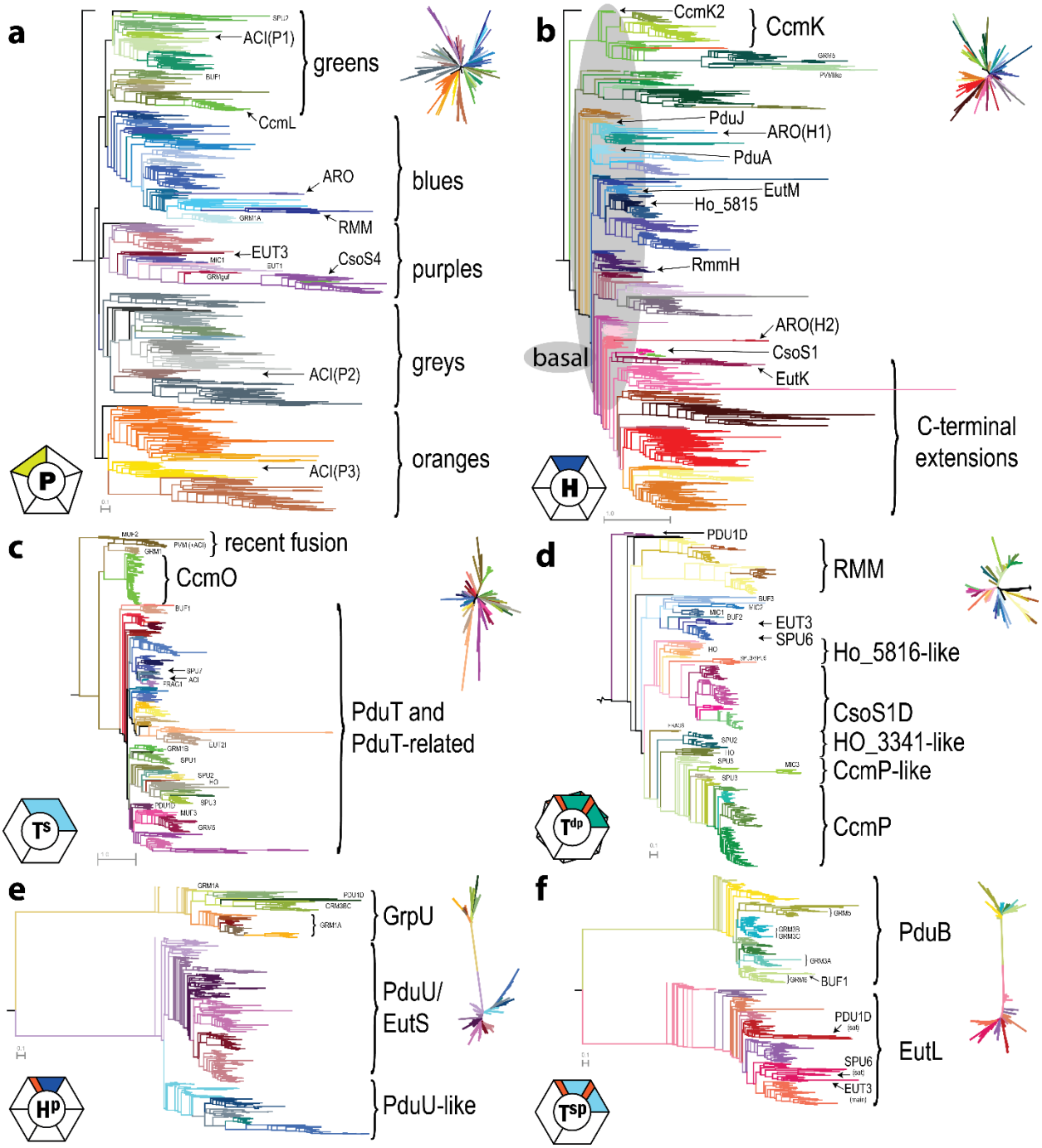

Figure 3. Phylogenetic maximum likelihood trees for the six types of shell protein: **a**, BMC-P, **b**, BMC-H with region of basal hexamers encompassed by gray shading, **c**, BMC-T^{s}, **d**, BMC-T^{dp}, **e**, BMC-H^{p} and **f**, BMC-T^{sp}. Representative sequences were selected by removing redundancy. Full hi-resolution images of each tree with annotation labels for all terminal nodes as well as bootstrap values are available as Supplementary Data: High resolution shell protein trees. All caps labels refer to the predominantly associated locus with a certain BMC type, mixed case labels denote a specific, previously characterized protein found in that clade. Unrooted versions of the trees are shown in the top right corner of each panel.

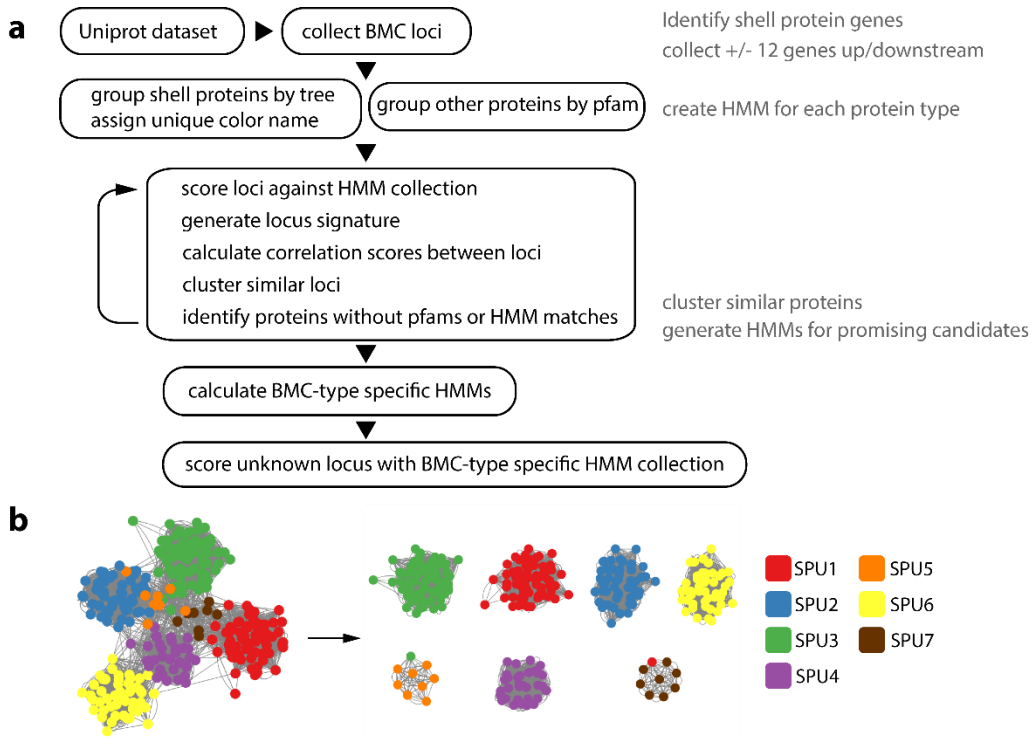

Figure 4. BMC classification methodology and example. **a**, Flowchart of the methodology of BMC locus classification. **b**, Example of clustering by locus scoring, shown for all SPU types: Force directed layout (left) and MCL clustering result (right) shows consistency of the BMC type clusters.

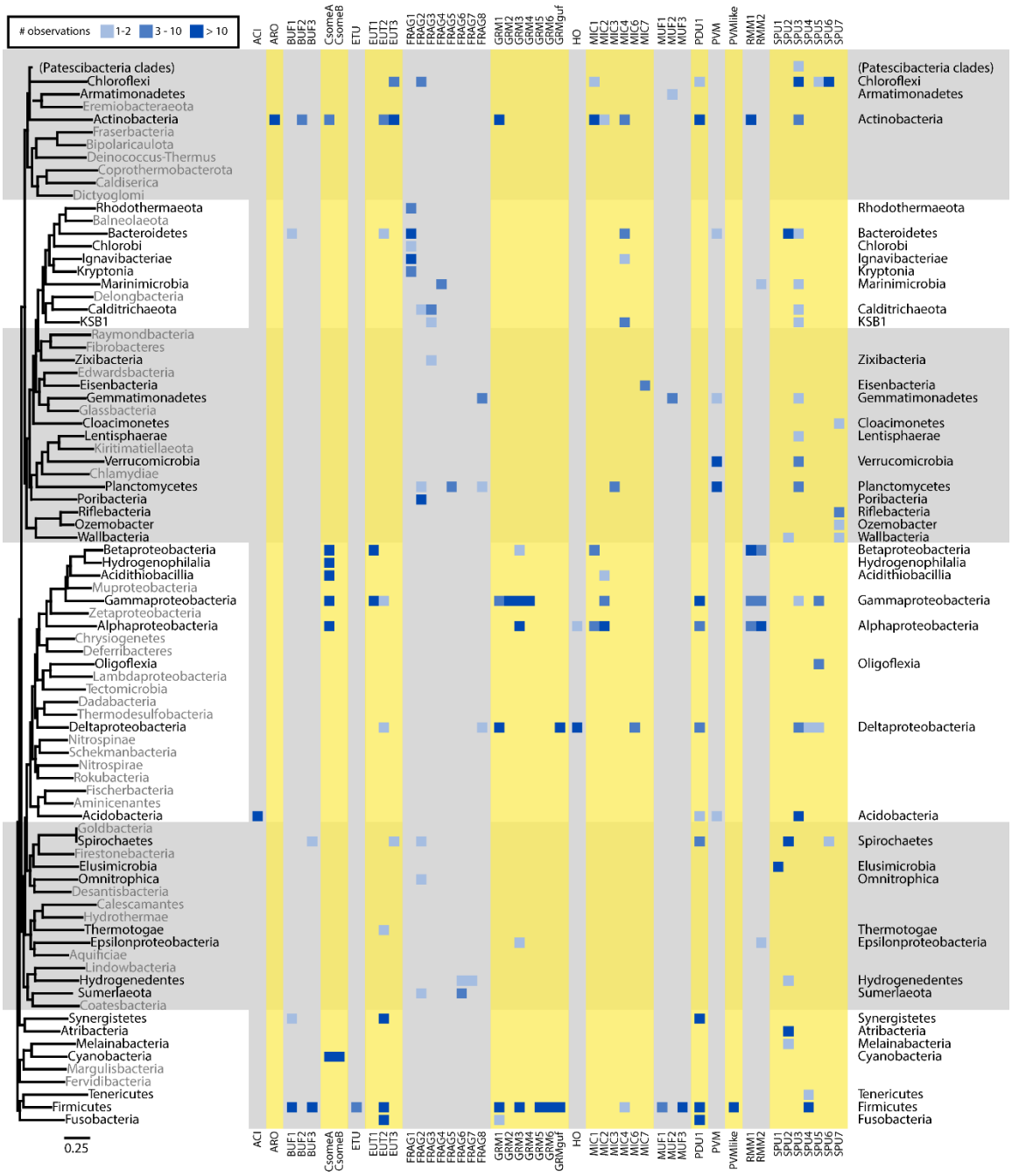

Figure 5. Distribution of BMC types in 46 phyla throughout the Bacteria Kingdom. Phylogenetic tree representing all known and candidate taxonomic phyla was generated by aligning a set of 56 different marker proteins. Phyla lacking BMC loci shown in grey and different numbers of BMCs in blue.

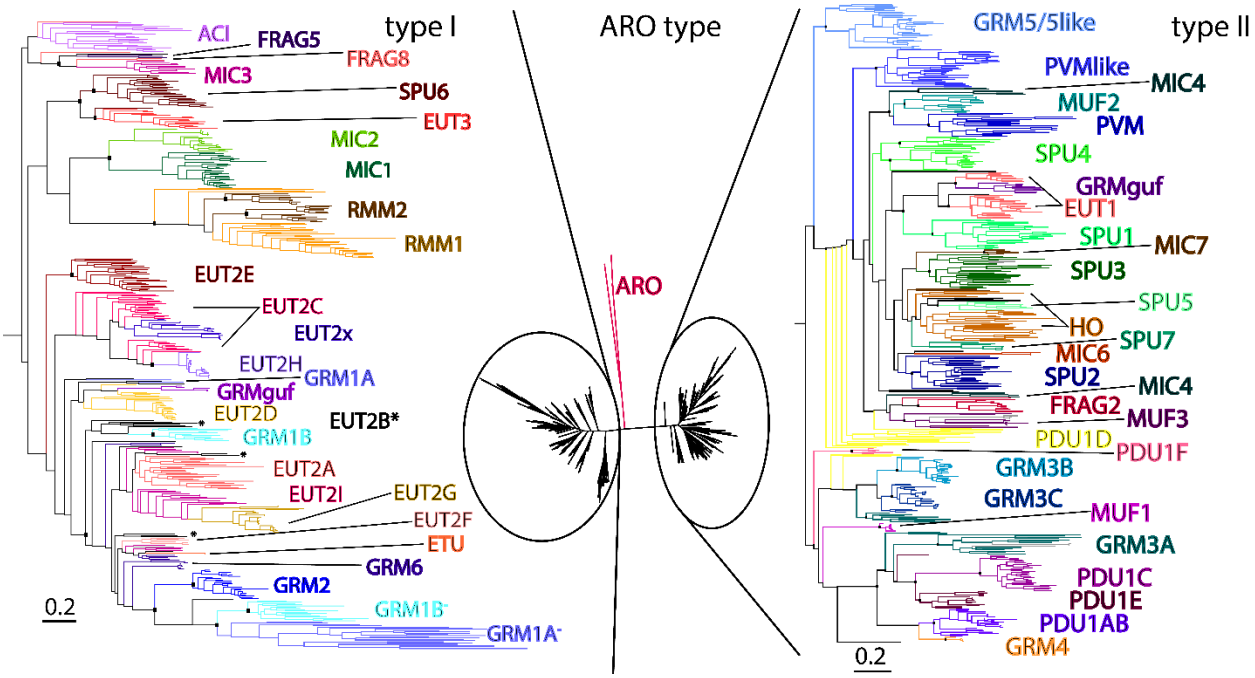

Figure 6. Phylogenetic tree of pfam00171 aldehyde dehydrogenases from BMC loci. Sequence redundancy was reduced by limiting to 30 sequences per BMC type (based on percent identity). Three major groups were identified: type I, which is dominated by EUT2, belongs to IPR013357, and contains a C-terminal EP (left); the ARO type outlier clade; and type II, which belongs to IPR012408, contains an N-terminal EP, and contains a larger diversity of BMC types, including all PDUs. Branch lengths are scaled by the number of substitutions per site. Bootstrap values for important nodes are represented as black squares (above 50%). GRM1A and GRM1B loci contain a putatively inactive second copy (marked with a - superscript). EUT2B AldDh marked with an asterisk are found in different basal locations.

Materials and Methods

Locus identification, clustering and HMM generation

The complete Uniprot database was searched with shell protein pfams (PF03319 and PF00936), InterPro domain identifiers (IPR000249, IPR004992, IPR009193, IPR009307, IPR013501, IPR014076, IPR014077, IPR020808, IPR030983, IPR030984), PROSITE identifier PS01139, and SMART ID SM0877, accessed as recently as March 2020. Adjacent proteins based on the numerical suffix were then obtained by downloading gene entries within 12 loci of the extracted gene identifier. Pfam tags were extracted from the uniprot annotations for initial HMM generation from proteins of the same pfam. HMMs were generated by aligning the sequences with clustalw 2.1 [39], trimming with trimAl 1.2rev59 with parameters -gt 0.6 -cons 30 -w 3 [40] and HMMs built with hmmbuild from the HMMER package version 3.1b2 [41]. Analogously aligned and trimmed sequences were used to construct the maximum likelihood AldDh tree with 30 redundancy reduced sequences for each BMC type using RAxML-NG with the WAG substitution model (v0.6.0, [42]). HMMs were calculated analogously for distinct shell proteins (see shell protein section below) and we then scored every locus protein against this combined HMM library, allowing us to represent each consolidated BMC locus as a string of identifier elements derived from the best-scoring HMM for each protein. To cluster the BMC loci and identify BMC types, we calculated a pairwise correlation scores across all loci. Pairwise locus-locus scores were calculated amongst all loci using a python script based on the sum of two values: the total number of HMMs found in both loci; and the length of the longest sequence of consecutive HMM matches, multiplied by a weighting factor of 10. Data was then imported in Cytoscape 3.7.2 [43] to visualize the locus clustering. See Supplementary Fig. 5 for an all-vs-all BMC types visualization. Clusters were manually matched to known BMC types and unknown types were assigned new identifiers. In a second round the proteins that were not assigned a pfam in Uniprot and did not match another protein in the HMM library were collected and the whole set was then clustered with MMseqs2 (mmseqs easy-cluster [44]). The clusters were then manually inspected and promising candidates were selected based on occurrence within a specific BMC type, distance to shell proteins and direction of translation. The clusters were then analyzed and HMMs were generated from those proteins to identify them in the locus analysis (for HMMs see Supplementary Data: HMM library).

Locus visualization

The HMM library was then used to score each BMC type separately to generate type-specific HMMs using only the sequence from one type. This type-specific HMM was then used to score loci with hmmsearch and locus data was visualized using a python script (Supplementary Data: Locus diagrams). The directionality of the gene coding for the proteins is shown as an arrow. While this data is not present in Uniprot files it can be determined by extracting the EMBL database identifier and parsing a corresponding DNA file downloaded through the European Nucleotide Archive (ENA). To determine the presence of encapsulation peptides (EPs) we collected sequences of known proteins with EPs and used the EP portion to generate HMMs. A separate HMM was generated for each BMC locus type and protein family. The combined HMM library was then used to identify potential encapsulation peptides. Despite the low sequence conservation and short length of EPs, this method is quite sensitive; however, manual inspection of the results is still necessary.

Shell protein phylogenies

An initial set of 6408 BMC-P was made non-redundant to 70% identity using the usearch –cluster_fast algorithm (v.11.0,[45]), resulting in a set of 1183 unique sequences. An initial alignment with muscle [46] was manually edited upon visual inspection with Jalview ([47] to prune fragments and problematic sequences likely arising from genome sequencing or gene modeling errors. Sequences were then realigned using MAFFT-linsi [48] and uninformative columns were removed with BMGE (-h 0.8 –g 0.05) [49]. ModelTest-NG [50] was used to determine the best-scoring substitution model, LG4M, which was then used to construct a maximum likelihood tree using RAxML-NG (v0.6.0, [42]). A similar approach was used to collapse the redundancy of BMC-H proteins (95% identity), BMC-H^p: (95%), BMC-T^s (90%), BMC-T^{dp} (90%), BMC-T^{sp} (90%). Trees were examined using Archaeopteryx (www.phylosoft.org/archaeopteryx) and significant clades were manually identified based on a general criterion of having a long internal stem, and were assigned unique color names from the XKCD color survey (https://xkcd.com/color/rgb/) and colored with their corresponding RGB hexcodes, selecting similar colors for nearby clades. The sequences from each color-based clade were then subdivided by the initial BMC locus type assignments and used to generate subtype-specific HMMs for scoring the entire BMC dataset. The phylogenies were later re-examined with later iterations of locus type assignments and annotated for functional correspondences to the clades or subclans (Figure 3). Vector-quality images of the six phylogenies are provided as supplementary material with sequence identifiers and color assignments provided, for legibility upon manual zoom in a PDF reader (Supplementary Data: High resolution shell protein trees).

Tree of bacterial phyla

A set of 56 universal single-copy marker proteins [51,52] was used to build a phylogenetic tree for the domain Bacteria based on a representative dataset that included one genome for each bacterial order present in the IMG/M database [37,53]; accessed March 2020). Marker proteins were identified with hmmsearch (version 3.1b2) using a specific HMM for each marker. Genomes lacking a substantial proportion of marker proteins (more than 26) or which had additional copies of more than five single-copy markers were removed from the dataset. For each marker, proteins were extracted, alignments built with MAFFT-linsi (v7.294b, [54]) and subsequently trimmed with BMGE (v1.12, [49]) using BLOSUM30. Single protein alignments were then concatenated resulting in an alignment of 10,755 sites. Maximum likelihood phylogenies were inferred with FastTree2 [55] using the options: -spr 4 -mlacc 2 -slownni -wag. The phylogenetic tree was pruned to keep only one representative genome for each phylum and then visualized using the ete3 package [56].

Acknowledgements

This work was supported by the National Institutes of Health, National Institute of Allergy and Infectious Diseases (NIAID) grant 1R01AI114975-01 and the U.S. Department of Energy, Basic Energy Sciences, Contract DE-FG02-91ER20021. We thank Jan Zarzycki and Stephanie A. Eichorst for helpful discussions, and Emiley Eloe-Fadrosh and Neha Varghese for custom HMM searches against the IMG/M database. Parts of this study were performed by the US Department of Energy Joint Genome Institute, a DOE Office

of Science User Facility, under Contract No. DE-AC02–05CH11231 and made use of resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02–05CH11231.

References

1. Drews G, Niklowitz W: **Beiträge zur Cytologie der Blaualgen. II. Zentroplasma und granuläre Einschlüsse von Phormidium uncinatum**. *Archiv für Mikrobiologie* 1956, **24**:147-162.
2. Shively JM, Ball F, Brown DH, Saunders RE: **Functional organelles in prokaryotes: polyhedral inclusions (carboxysomes) of *Thiobacillus neapolitanus***. *Science* 1973, **182**:584-586.
3. Shively JM, Bradburne CE, Aldrich HC, Bobik TA, Mehlman JL, Jin S, Baker SH: **Sequence homologs of the carboxysomal polypeptide CsoS1 of the thiobacilli are present in cyanobacteria and enteric bacteria that form carboxysomes - polyhedral bodies**. *Canadian Journal of Botany-Revue Canadienne De Botanique* 1998, **76**:906-916.
4. Axen SD, Erbilgin O, Kerfeld CA: **A taxonomy of bacterial microcompartment loci constructed by a novel scoring method**. *PLoS Comput Biol* 2014, **10**:e1003898.
5. Jorda J, Lopez D, Wheatley NM, Yeates TO: **Using comparative genomics to uncover new kinds of protein-based metabolic organelles in bacteria**. *Protein Science* 2013, **22**:179-195.
6. Kerfeld CA, Erbilgin O: **Bacterial microcompartments and the modular construction of microbial metabolism**. *Trends Microbiol* 2015, **23**:22-34.
7. Kinney JN, Salmeen A, Cai F, Kerfeld CA: **Elucidating essential role of conserved carboxysomal protein CcmN reveals common feature of bacterial microcompartment assembly**. *J Biol Chem* 2012, **287**:17729-17736.
8. Aussignargues C, Paasch BC, Gonzalez-Esquer R, Erbilgin O, Kerfeld CA: **Bacterial microcompartment assembly: The key role of encapsulation peptides**. *Commun Integr Biol* 2015, **8**:e1039755.
9. Kerfeld CA, Aussignargues C, Zarzycki J, Cai F, Sutter M: **Bacterial microcompartments**. *Nat Rev Microbiol* 2018, **16**:277-290.
10. Sutter M, Greber B, Aussignargues C, Kerfeld CA: **Assembly principles and structure of a 6.5-MDa bacterial microcompartment shell**. *Science* 2017, **356**:1293-1297.
11. Kalnins G, Cesle EE, Jansons J, Liepins J, Filimonenko A, Tars K: **Encapsulation mechanisms and structural studies of GRM2 bacterial microcompartment particles**. *Nat Commun* 2020, **11**:388.
12. Greber BJ, Sutter M, Kerfeld CA: **The Plasticity of Molecular Interactions Governs Bacterial Microcompartment Shell Assembly**. *Structure* 2019, **27**:749-763 e744.
13. Sutter M, Laughlin TG, Sloan NB, Serwas D, Davies KM, Kerfeld CA: **Structure of a Synthetic beta-Carboxysome Shell**. *Plant Physiol* 2019, **181**:1050-1058.
14. Tanaka S, Kerfeld CA, Sawaya MR, Cai F, Heinhorst S, Cannon GC, Yeates TO: **Atomic-level models of the bacterial carboxysome shell**. *Science* 2008, **319**:1083-1086.
15. Kerfeld CA, Sawaya MR, Tanaka S, Nguyen CV, Phillips M, Beeby M, Yeates TO: **Protein structures forming the shell of primitive bacterial organelles**. *Science* 2005, **309**:936-938.
16. Klein MG, Zwart P, Bagby SC, Cai F, Chisholm SW, Heinhorst S, Cannon GC, Kerfeld CA: **Identification and structural analysis of a novel carboxysome shell protein with implications for metabolite transport**. *Journal of Molecular Biology* 2009, **392**:319-333.
17. Sjolander K: **Phylogenomic inference of protein molecular function: advances and challenges**. *Bioinformatics* 2004, **20**:170-179.

18. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling**. *J Mol Biol* 1994, **235**:1501-1531.
19. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, et al.: **SWISS-MODEL: homology modelling of protein structures and complexes**. *Nucleic Acids Res* 2018, **46**:W296-W303.
20. Zhang RG, Andersson CE, Skarina T, Evdokimova E, Edwards AM, Joachimiak A, Savchenko A, Mowbray SL: **The 2.2 A resolution structure of RpiB/AlsB from Escherichia coli illustrates a new approach to the ribose-5-phosphate isomerase reaction**. *J Mol Biol* 2003, **332**:1083-1094.
21. Aussignargues C, Pandelia ME, Sutter M, Plegaria JS, Zarzycki J, Turmo A, Huang J, Ducat DC, Hegg EL, Gibney BR, et al.: **Structure and Function of a Bacterial Microcompartment Shell Protein Engineered to Bind a [4Fe-4S] Cluster**. *Journal of the American Chemical Society* 2016, **138**:5262-5270.
22. Erbilgin O, McDonald KL, Kerfeld CA: **Characterization of a planctomycetal organelle: a novel bacterial microcompartment for the aerobic degradation of plant saccharides**. *Appl Environ Microbiol* 2014, **80**:2193-2205.
23. Sizikov S, Burgsdorf I, Handley KM, Lahyani M, Haber M, Steindler L: **Characterization of sponge-associated Verrucomicrobia: microcompartment-based sugar utilization and enhanced toxin-antitoxin modules as features of host-associated Opitutales**. *Environ Microbiol* 2020, **22**:4669-4688.
24. Sichert A, Corzett CH, Schechter MS, Unfried F, Markert S, Becher D, Fernandez-Guerra A, Liebeke M, Schweder T, Polz MF, et al.: **Verrucomicrobia use hundreds of enzymes to digest the algal polysaccharide fucoidan**. *Nature Microbiology* 2020, **5**:1026-+.
25. Chowdhury NP, Alberti L, Linder M, Muller V: **Exploring Bacterial Microcompartments in the Acetogenic Bacterium Acetobacterium woodii**. *Front Microbiol* 2020, **11**:593467.
26. Sturms R, Streauslin NA, Cheng S, Bobik TA: **In *Salmonella enterica*, Ethanolamine Utilization Is Repressed by 1,2-Propanediol To Prevent Detrimental Mixing of Components of Two Different Bacterial Microcompartments**. *J Bacteriol* 2015, **197**:2412-2421.
27. Fendrich C, Hippe H, Gottschalk G: **Clostridium-Halophilium Sp-Nov and Clostridium-Litorale Sp-Nov, an Obligate Halophilic and a Marine Species Degrading Betaine in the Stickland Reaction**. *Archives of Microbiology* 1990, **154**:127-132.
28. Flint HJ, Scott KP, Duncan SH, Louis P, Forano E: **Microbial degradation of complex carbohydrates in the gut**. *Gut Microbes* 2012, **3**:289-306.
29. Long BM, Price GD, Badger MR: **Proteomic assessment of an established technique for carboxysome enrichment from Synechococcus PCC7942**. *Canadian Journal of Botany-Revue Canadienne De Botanique* 2005, **83**:746-757.
30. Sommer M, Cai F, Melnicki M, Kerfeld CA: **beta-Carboxysome bioinformatics: identification and evolution of new bacterial microcompartment protein gene classes and core locus constraints**. *J Exp Bot* 2017, **68**:3841-3855.
31. Slininger Lee MF, Jakobson CM, Tullman-Ercek D: **Evidence for Improved Encapsulated Pathway Behavior in a Bacterial Microcompartment through Shell Protein Engineering**. *ACS Synthetic Biology* 2017.
32. Cai F, Sutter M, Bernstein SL, Kinney JN, Kerfeld CA: **Engineering Bacterial Microcompartment Shells: Chimeric Shell Proteins and Chimeric Carboxysome Shells**. *ACS Synthetic Biology* 2015, **4**:444-453.
33. Thompson MC, Wheatley NM, Jorda J, Sawaya MR, Gidaniyan SD, Ahmed H, Yang Z, McCarty KN, Whitelegge JP, Yeates TO: **Identification of a unique Fe-S cluster binding site in a glycyl-radical type microcompartment shell protein**. *J Mol Biol* 2014, **426**:3287-3304.

34. Crowley CS, Sawaya MR, Bobik TA, Yeates TO: **Structure of the PduU shell protein from the Pdu microcompartment of Salmonella**. *Structure* 2008, **16**:1324-1332.
35. Jorda J, Liu Y, Bobik TA, Yeates TO: **Exploring bacterial organelle interactomes: a model of the protein-protein interaction network in the Pdu microcompartment**. *PLoS Comput Biol* 2015, **11**:e1004067.
36. Parsons JB, Frank S, Bhella D, Liang M, Prentice MB, Mulvihill DP, Warren MJ: **Synthesis of empty bacterial microcompartments, directed organelle protein incorporation, and evidence of filament-associated organelle movement**. *Mol Cell* 2010, **38**:305-315.
37. Chen IA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, Hajek P, Ritter S, Varghese N, Seshadri R, et al.: **The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities**. *Nucleic Acids Res* 2021, **49**:D751-D763.
38. Sun YQ, Wollman AJM, Huang F, Leake MC, Liu LN: **Single-Organelle Quantification Reveals Stoichiometric and Structural Variability of Carboxysomes Dependent on the Environment**. *Plant Cell* 2019, **31**:1648-1664.
39. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al.: **Clustal W and clustal X version 2.0**. *Bioinformatics* 2007, **23**:2947-2948.
40. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T: **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses**. *Bioinformatics* 2009, **25**:1972-1973.
41. Eddy SR: **Accelerated Profile HMM Searches**. *Plos Computational Biology* 2011, **7**.
42. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A: **RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference**. *Bioinformatics* 2019, **35**:4453-4455.
43. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A software environment for integrated models of biomolecular interaction networks**. *Genome Research* 2003, **13**:2498-2504.
44. Steinegger M, Soding J: **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets**. *Nature Biotechnology* 2017, **35**:1026-1028.
45. Edgar RC: **Search and clustering orders of magnitude faster than BLAST**. *Bioinformatics* 2010, **26**:2460-2461.
46. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Research* 2004, **32**:1792-1797.
47. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ: **Jalview Version 2-a multiple sequence alignment editor and analysis workbench**. *Bioinformatics* 2009, **25**:1189-1191.
48. Yamada KD, Tomii K, Katoh K: **Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees**. *Bioinformatics* 2016, **32**:3246-3251.
49. Criscuolo A, Gribaldo S: **BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments**. *BMC Evol Biol* 2010, **10**:210.
50. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T: **ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models**. *Molecular Biology and Evolution* 2020, **37**:291-294.
51. Yu FB, Blainey PC, Schulz F, Woyke T, Horowitz MA, Quake SR: **Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples**. *Elife* 2017, **6**.
52. Eloe-Fadrosh EA, Paez-Espino D, Jarett J, Dunfield PF, Hedlund BP, Dekas AE, Grasby SE, Brady AL, Dong H, Briggs BR, et al.: **Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs**. *Nat Commun* 2016, **7**:10476.

53. Chen IA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R, et al.: **IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes**. *Nucleic Acids Res* 2019, **47**:D666-D677.
54. Katoh K, Standley DM: **A simple method to control over-alignment in the MAFFT multiple sequence alignment program**. *Bioinformatics* 2016, **32**:1933-1942.
55. Price MN, Dehal PS, Arkin AP: **FastTree 2--approximately maximum-likelihood trees for large alignments**. *PLoS One* 2010, **5**:e9490.
56. Huerta-Cepas J, Serra F, Bork P: **ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data**. *Mol Biol Evol* 2016, **33**:1635-1638.

Supplementary Files

Supplementary Figures and Tables

Supplementary Data: BMC types

Supplementary Data: Locus diagrams

Supplementary Data: High resolution shell protein trees

Supplementary Data: HMM library

Supplementary Data: HMM names table