A new gene family diagnostic for intracellular biomineralization of amorphous

Karim Benzerara, Elodie Duprat, Tristan Bitard-Feildel, Géraldine Caumes, Corinne Cassier-Chauvat, Franck Chauvat, Manuela Dezi, Issa Diop Seydina, Geoffroy Gaschignard, Sigrid Görgen, et al.

A new gene family diagnostic for intracellular biomineralization of amorphous Ca-carbonates by cyanobacteria

Benzerara Karim[1,2,3,*], Duprat Elodie[1,2], Bitard-Feildel Tristan[1], Caumes Géraldine[1], Cassier-Chauvat Corinne[4], Chauvat Franck[4], Dezi Manuela[1], Diop Seydina Issa[1,5], Gaschignard Geoffroy[1], Görgen Sigrid[1,4], Gugger Muriel[6], López-García Purificación[7], Millet Maxime[1], Skouri-Panet Fériel[1], Moreira David[7, 8], Callebaut Isabelle[1, 8,*]

[1]Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590. Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC), 4 Place Jussieu, 75005 Paris, France.

[2]equal contributions

[3]lead contact

[4]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

[5]Present address: Department of Systematic and Evolutionary Botany & Zurich-Basel Plant Science Center, University of Zurich, Zollikerstrasse 107, 8008, Zurich, Switzerland.

[6]Institut Pasteur, Université de Paris, Collection of Cyanobacteria, F-75015 Paris, France

[7]Unité d'Ecologie Systématique et Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Orsay, France

[8]equal contributions

*corresponding authors: Karim Benzerara (karim.benzerara@sorbonne-universite.fr) & Isabelle Callebaut (isabelle.callebaut@sorbonne-universite.fr)

ABSTRACT

Cyanobacteria have massively contributed to carbonate deposition over the geological history. They are traditionally thought to biomineralize $CaCO_3$ extracellularly as an indirect byproduct of photosynthesis. However, the recent discovery of freshwater cyanobacteria forming intracellular amorphous calcium carbonates (iACC) challenges this view. Despite the geochemical interest of such a biomineralization process, its molecular mechanisms and evolutionary history remain elusive. Here, using comparative genomics, we identify a new gene (*ccyA*) and protein family (calcyanin) possibly associated with cyanobacterial iACC biomineralization. Proteins of the calcyanin family are composed of a conserved C-terminal domain, which likely adopts an original fold, and a variable N-terminal domain whose structure allows differentiating 4 major types among the 35 known calcyanin homologs. Calcyanin lacks detectable full-length homologs with known function. The overexpression of *ccyA* in iACC-lacking cyanobacteria resulted in an increased intracellular Ca content. Moreover, *ccyA* presence was correlated and/or co-localized with genes involved in Ca or HCO_3^- transport and homeostasis, supporting the hypothesis of a functional role of calcyanin in iACC biomineralization. Whatever its function, *ccyA* appears as diagnostic of intracellular calcification in cyanobacteria. By searching for *ccyA* in publicly available genomes, we identified 13 additional cyanobacterial strains forming iACC, as confirmed by microscopy. This extends our knowledge about the phylogenetic and environmental distribution of cyanobacterial iACC biomineralization, especially with the detection of multicellular genera as well as a marine species. Moreover, *ccyA* was probably present in ancient cyanobacteria, with independent losses in various lineages that resulted in a broad but patchy distribution across modern cyanobacteria.

Keywords:

Biomineralization, amorphous calcium carbonates, cyanobacteria, protein structure prediction, phylogeny, glycine zipper motifs

Significance statement

Few freshwater species of Cyanobacteria have been known to mineralize amorphous $CaCO_3$ (ACC) intracellularly. Despite the geochemical interest of this biomineralization, its evolutionary history and molecular mechanism remain poorly known. Here, we report the discovery of a new gene family that has no homolog with known function, which proves to be a good diagnostic marker of this process. Using this marker gene, we find new cyanobacteria forming ACC in several genera and environments such as seawater, where ACC biomineralization had not been reported before. Moreover, this gene is ancient and was independently lost in various lineages, resulting in a broad and patchy phylogenetic distribution in modern cyanobacteria.

INTRODUCTION

The formation of mineral phases by living organisms is widespread in both eukaryotes and prokaryotes (Weiner and Dove 2003). While many cases of biomineralization in eukaryotes involve specific genes (Marron et al. 2016; Wang et al. 2021; Yarra et al. 2021), there is presently only one documented case of genetically controlled biomineralization in bacteria: the intracellular magnetite formation by magnetotactic bacteria (Lefevre and Bazylinski 2013). The formation of Ca-carbonates by cyanobacteria has been studied for several decades and cyanobacteria are thought to have been among the main calcifiers at the Earth surface since their appearance several billion years ago (Altermann et al. 2006). However, it is only recently that a genetic control of iACC biomineralization by some species of cyanobacteria has been hypothesized (Benzerara et al. 2014), but not yet proven. Interestingly, the involvement of ACC has been widely documented and studied in the formation of eukaryotic skeletons (Blue et al.

2017). By contrast and although a growing number of bacterial occurrences are described (Monteil et al. 2020), the determinants of ACC formation in prokaryotes remain poorly understood.

The iACC-biomineralizing cyanobacteria are geographically widespread in freshwater, hotspring or karstic terrestrial systems (Ragon et al. 2014) and sometimes locally abundant (Bradley et al. 2017). They received particular attention since they challenge the usual paradigm that cyanobacteria biomineralize $CaCO_3$ extracellularly as an indirect byproduct of photosynthesis only (Altermann et al. 2006). Moreover, the geological history of iACC biomineralization remains mysterious since the fossilization potential of these bacteria appears uncertain (Couradeau et al. 2012; Riding 2012). They can form iACC even under thermodynamically unfavorable conditions, indicating that they consume energy to perform this process, possibly in relation with active sequestration of alkaline earth elements (Cam et al. 2018). An envelope of undetermined composition, either a lipid monolayer and/or proteins, surrounds the iACC granules (Blondeau, Sachse, et al. 2018) and it has been suggested that compartmentation is instrumental for the achievement of local Ca concentrations that are high enough for the formation of iACC (Cam et al. 2018). Furthermore, some iACC-forming species require higher Ca amounts for optimal growth than iACC-lacking ones, indicating that they possess an unusual Ca homeostasis (De Wever et al. 2019). Interestingly, by forming iACC granules, these cyanobacteria accumulate very high Ca amounts, as well as other alkaline earth elements such as strontium (Sr) and barium (Ba) (Cam et al. 2016; Blondeau, Benzerara, et al. 2018) and may impact the geochemical cycles of these trace elements (Blondeau, Benzerara, et al. 2018). Indeed, by normalizing the uptake to their cell mass, they are among the highest Sr and Ba-scavenging organisms known (Cam et al. 2016). Moreover, they can efficiently sequester radioisotopes such as ^{90}Sr or radium (Ra) isotopes, a capability that may be used for bioremediation (Cam et al. 2016; Blondeau, Benzerara, et al. 2018; Mehta et al. 2019).

All the members of some clades of cyanobacteria, such as the *Cyanothece-Synechococcus-Thermosynechococcus* clade, share this capability to form iACC, suggesting the genetic heritability of this trait in this specific group (Benzerara et al. 2014). Yet, despite the geochemical relevance of this process, the genetic control of iACC formation has not been identified. Moreover, whether the presently known iACC-forming cyanobacteria share ancestral genetic traits related to this biomineralization process or they convergently developed this capability to form iACC during cyanobacterial evolution remains unknown. In the absence of a fossil record, investigating the genetic basis of this biomineralization process appears as the only way to track its geological history.

Results and Discussion

Detection of a gene family diagnostic of iACC biomineralization

We applied comparative genomics to search for putative genes exclusively shared by iACC-forming cyanobacteria, and therefore absent in iACC-lacking species. We analyzed the genomes of 56 cyanobacterial strains (supplementary table 1), in which the presence or absence of iACC was previously determined by electron microscopy (EM) (Benzerara et al. 2014). Fifty strains were lacking iACC and 6 were shown to form iACC: *Synechococcus* sp. PCC 6312, *Synechococcus calcipolaris* PCC 11701, *Thermosynechococcus elongatus* BP-1, *Cyanothece* sp. PCC 7425, *Chroococcidiopsis thermalis* PCC 7203, and *Gloeomargarita lithophora* D10. Among the 523 680 translated coding sequences (CDSs) contained in the 56 genomes, only one group of orthologous sequences (among the 27 230 groups comprising at least 2 sequences) was shared by all six iACC-forming strains and absent in all 50 iACC-lacking strains. The corresponding gene was named *ccyA*. Its predicted protein product was named calcyanin (CcyA as a protein symbol). Conversely, we found no group of orthologous sequences shared by all 50 iACC-lacking strains and absent in all 6 iACC-forming strains. No functional annotation of

calcyanin could be achieved using profiles of known protein domain families.

We first investigated the architecture of calcyanin by Hydrophobic Cluster Analysis (HCA), an approach that has already been largely applied to the detection of novel domain families (Callebaut et al. 2017; Bitard-Feildel et al. 2018). The HCA two-dimensional representation of the protein sequence provides structural information based on the distribution of strong hydrophobic amino acids in clusters (representative of regular secondary structures) and their relative arrangement. This last feature allows to appreciate the segmentation of the protein into domains and their intrinsic nature (folded, disordered, *etc*.), as well as to detect repeated motifs and their overall conservation between sequences. The HCA approach revealed that calcyanin is composed of two domains (fig. 1). The C-terminal domain is composed of three long repeats of a periodic pattern (called GlyZip), including glycine (or small amino acids – indicated in yellow in fig. 1) and hydrophobic amino acids (green) every four residues (long, horizontal clusters). The pattern was clearly distinct for the N-terminal domains, possessing smaller hydrophobic clusters, usually encountered in current globular domains. While the C-terminal domain was highly conserved in the six different calcyanin sequences, the N-terminal domain appeared to be conserved in five sequences only, and exhibited significant differences in *G. lithophora*. Therefore, we used the conserved C-terminal domain to search for additional homologs in a comprehensive set of 594 cyanobacterial genomes available in public databases. We found additional *ccyA* homologs in 27 strains (supplementary table 2; supplementary fig. 1). Among them, we inspected 17 strains available to us, by EM coupled with energy dispersive x-ray spectrometry (EDXS), which allowed submicrometer-scale chemical mapping of several elements, including Ca and P. As shown by Benzerara et al. (2014) and Li et al. (2016), iACC can be recognized by the fact that they contain Ca but little to no P, in contrast with polyphosphate inclusions, which show a major P EDXS peak with Mg and K and, sometimes, Ca. We detected iACC in 13 of the 17 inspected strains (fig. 2; supplementary fig. 2), thereby

increasing the number of known iACC-forming cyanobacterial species from six to 19. Moreover, we detected *ccyA* in the two recently sequenced genomes of *Synechococcus* sp. PCC 6716 and PCC 6717 that were previously shown to form iACC (Benzerara et al. 2014) (supplementary table 2).

In some strains (e.g., *Fischerella* sp. NIES-4106, *Neosynechococcus sphagnicola* sy1), most of the cells exhibited abundant iACC granules. By contrast, for strains such as *Microcystis aeruginosa* PCC 7806, cells contained none or only few iACC granules. In other strains (e.g., *Chlorogloeopsis fritschii* PCC 9212), the cells contained few iACC granules and many Ca-rich polyphosphate inclusions that could be morphologically confused with iACC by EM alone but not chemically, hence requiring the use of EDXS (fig. 2, supplementary fig. 2). The four strains possessing *ccyA* but lacking iACC (*C. fritschii* PCC 6912; *Fischerella* sp. NIES-3754; *M. aeruginosa* PCC 9432 and PCC 9717; fig. 3) were phylogenetically very close to iACC-forming relatives. For example, *C. fritschii* PCC 9212 (iACC-forming) and PCC 6912 (no observed iACC) had only few differences in their gene repertoires (supplementary fig. 3; supplementary table 3) and the nucleotide sequences of the genomic regions containing *ccyA* in these two strains (corresponding to contigs of 97 542 bp and 97 528 bp in length, respectively) shared 100% identity over 97 528 bp. However, 57 genes of *C. fritschii* PCC 9212 had no homolog in *C. fritschii* PCC 6912. Their functional categories were annotated using the NCBI-curated clusters of orthologous groups (COG) protein classification resource. They mostly corresponded to unknown functions (46 without COG hit, 2 genes with COG category X indicating an unknown function) or inorganic ion transport (4 genes, COG category P; supplementary table 3). Moreover, although we did not observe iACC in *C. fritschii* PCC 6912 and *Fischerella* sp. NIES-3754 cells, they both showed Ca- and P-rich inclusions morphologically similar to iACC, suggesting that they may have some but not all the capabilities required to produce iACC (fig. 3). Benzerara et al (2014) and Cam et al. (2017)

previously concluded that iACC-forming strains tend to show iACC inclusions when cultured in different growth media and/or sampled at different stages of their growth. Moreover, we conducted observations on multiple cultures sampled at different times for the four strains, supporting the idea that iACC do not appear transiently in these cultures. However, whether these strains are genetically unable to form iACC or this capability may depend on specific conditions will need to be assessed by future studies. At any rate, the search for *ccyA* in available cyanobacterial genome sequences allowed the detection of 13 additional iACC-forming strains among the 17 strains whose genomes contained *ccyA*, largely extending and optimizing the initial detection of 8 iACC-forming strains (i.e. six strains whose genomes were used for comparative genomics plus *Synechococcus* sp. PCC 6716 and PCC 6717 whose genomes were recently sequenced) among 58 randomly selected, phylogenetically diverse cyanobacteria (Benzerara et al. 2014). Therefore, the search for *ccyA* occurrence significantly increased the probability of success in finding iACC-forming strains (binomial exact test, p=9.0e-09), indicating that *ccyA* can be used as diagnostic marker of intracellular biomineralization.

Thanks to this approach, we expanded considerably the phylogenetic diversity of known iACC-forming cyanobacteria (fig. 4a). So far, iACC biomineralization had been reported in unicellular cyanobacteria only (Benzerara et al. 2014). Here, we find iACC in several multicellular genera belonging to the most complex morphotypes of the cyanobacterial phylum with cellular differentiation and ramifications (*Chlorogloeopsis* and *Fischerella*). Moreover, we also discovered iACC in *Microcystis aeruginosa*, one of the most common, worldwide-distributed bloom-forming cyanobacteria (Humbert et al. 2013). *Microcystis* shows a life cycle with a benthic phase in winter and a planktonic phase in warmer seasons when cells produce gas vesicles to float in the water column (Reynolds and Rogers 1976; Latour et al. 2007). Considering the high density of ACC relative to cells, a controlled production of dense iACC granules might favor a shift to benthic life. Interestingly, *ccyA* is present in the genome of some

closely related *M. aeruginosa* strains but absent from others. This finding may be consistent with the high genome plasticity detected in this species, reflecting frequent horizontal gene transfers (Frangeul et al. 2008; Humbert et al. 2013).

The *ccyA* gene and iACC biomineralization were also found in four *Synechococcus*-like strains previously not known to produce iACC (*Neosynechococcus sphagnicola* sy1, *Synechococcus* sp. RS9917, *Synechococcus lividus* PCC 6715, and *Thermosynechococcus* sp. NK55a). *Synechococcus* is a polyphyletic genus, grouping strains isolated from very different environments (Komarek et al. 2020). We previously reported thermophilic and mesophilic freshwater iACC-biomineralizing *Synechococcus* representatives (Benzerara et al. 2014). Here, we significantly expanded this environmental distribution especially with the inclusion of the first marine (*Synechococcus* sp. RS9917) iACC-forming strain.

Sequence-based analysis of the calcyanin structure

With the exception of the *Thermosynechococcus* sp. NK55a calcyanin, fused with a polypeptide containing a PIN-TRAM domain, the other 34 calcyanin family homologs contained 264 to 375 amino acids (average 336±25; supplementary table 2). All showed the already mentioned two-domain modularity: a variable N-terminal domain and a conserved C-terminal domain.

The N-terminal domain was composed of hydrophobic clusters with lengths and shapes typical of regular secondary structures found in globular domains (Lamiable et al. 2019). According to their N-terminal domain, we classified the 35 calcyanin homologs into four groups: W, X, Y, and Z (fig. 4b). There was only one calcyanin in the X group. Amino acid identities between the N-terminal domains of calcyanin homologs were higher than 18, 84 and 82 % within the W, Y and Z groups, respectively. The Y-type N-terminal domain consisted of a duplicated small domain (measuring 66 amino acids in length, with a mean identity between

the repeated domains in a same protein of 35.6 %; supplementary fig. 4), which was predicted to contain five regular secondary structures (labeled a to e in fig. 4). As for X- and Z-type N-terminal domains, they were distinct from known protein domains, as inferred from the absence of significant similarities when searching sequence and domain databases. By contrast, significant sequence similarities were detected between the W-type N-terminal domain and three known domain families (fig. 5): 1) YAM domains, found in the cytosolic C-terminus of *Escherichia coli* Major Facilitator Superfamily transporter YajR (Jiang et al. 2013; Jiang et al. 2014); 2) Heavy-Metal Associated (HMA) domains (also called Metal Binding Domains) present in various proteins (e.g., P-type ATPases and metallochaperones), generally involved in metal transport and detoxification pathways (Bull and Cox 1994); and 3) integrated HMA (iHMA) domains detected in plant immune receptors, where they are involved in fungal effector recognition (De la Concepcion et al. 2018). Similarly to these three domains, the W-type N-terminal domain showed a repeated β–α–β motif corresponding to a ferredoxin-like fold, characteristic of the HMA superfamily (fig. 5). However, while most HMA domains possess two conserved cysteine residues directly involved in binding heavy metals, YAM, iHMA and W-type calcyanin N-terminal domains do not conserve these amino acids (fig. 5). Moreover, the W-type domain showed a specific signature consisting of several basic amino acids distributed in strands β1 and β2 and a histidine located upstream of strand β1, in a region appearing as a calcyanin-specific extension of the HMA core (strand β0 in fig. 5). Therefore, we named this novel domain family CoBaHMA, after *domain with Conserved Basic residues in the HMA superfamily*. A model of the CoBaHMA 3D structure was built using the experimental 3D structures of the HMA, iHMA and YAM as templates in Modeller 9.23 (Webb and Sali 2016). The position of strand β0 was moreover putatively assigned with reference to the 3D structure of KipI (pdb 2KWA), based on the results of HH-PRED searches and subsequent superimposition of the 3D corresponding 3D structures (pdb 2RU9 and 2KWA, root

mean square value of 2.1 Å on 55 Cα superimposed positions). The AlphaFold2 model (pLDDT scores above 85 from aa 7 to 81, with most of the values above 90) agreed with the first proposed model, in particular on the position of strand β0 relative to the β1-β3 core, but also led to propose a model for strand β4 as well as to refine the position of amino acids within strand β0 (fig. 5B). Although the calcyanin sequence of *Synechococcus* sp. Lanier also contained the specific signature of W-type domains with several basic amino acids, it clearly differed from the rest of the W-type N-terminal domains (fig. 5A), suggesting that calcyanin has deeply diverged in this species. Future studies should assess whether these different N-terminal domains can be found in other cyanobacterial proteins.

The C-terminal domain of the different calcyanin types consisted in three repetitions of a ~50 amino acid motif, which was largely apolar and displayed a constant periodicity in hydrophobic and small (glycine/alanine) amino acids (supplementary fig. 5). We called this motif "GlyZip" in reference to the name proposed by (Kim et al. 2005) to describe recurrent, short Gly-X(3)-Gly-X(3)-Gly motifs allowing tight packing of transmembrane helices (Senes et al. 2004). However, the calcyanin GlyZip motifs were much longer (12 basic Gly-X(3)-Gly units, interrupted in their middle by a central, highly conserved Gly-Pro dipeptide) than those already known at the 3D level, which generally contained no more than three such units (Leonov and Arkin 2005). Moreover, they did not share any obvious sequence similarity with known domains, suggesting that these repeated motifs form a novel architecture. The repeated presence of glycine and hydrophobic amino acids every four amino acid residues over a large sequence length, with an unusual persistence of this periodic motif across the different cyanobacterial lineages (especially for the first repeat) suggests that it may form compact and highly constrained assemblages of helices compatible with a membrane-embedded structure. These assemblages might resemble homo-oligomeric structures formed by short subunits, such as the c-rings of sodium-translocating ATP synthases (Kuehlbrandt 2019), which share similar, albeit

smaller, glycine zippers. Analysis of multiple sequence alignments (supplementary fig. 5) allowed discriminating each of the three GlyZip calcyanin motifs based on specific signatures, including the presence of aromatic and polar amino acids, outside the repeated patterns. In particular, a tryptophan and a glutamic acid were strictly conserved in the third GlyZip motif in all calcyanin sequences. The second GlyZip motif of several calcyanin sequences matched part of a family model called PdsO (sortase-associated OmpA-like protein), found in, e.g., *Shewanella oneidensis* (see NCBI Conserved Domain Database (CDD) annotations in supplementary table 2b). The matching region, located before the OpmA-like C terminal domain, shows the typical features of a GlyZip unit (supplementary fig. 6) and is present as a single copy in PdsO, suggesting that this basic unit evolved within calcyanin by triplication and enrichment in polar amino acids (see below). Last, among Y-type calcyanins, only that of *Fischerella* sp. NIES-4106 possessed all three GlyZip motifs. By contrast, all other Y-type calcyanins, including those found in iACC-forming strains, contained only the first and third GlyZip motifs. This suggests that calcyanins with only two GlyZip motifs remain functional (supplementary fig. 5). Interestingly, although it did not match the characteristic GlyZip profile, the duplicated domain found in the N-terminal region of Y-type calcyanins was also largely apolar and rich in small amino acids so as in GlyZip motifs.

Calcyanin may be involved in Ca homeostasis

In *C. fritschii* PCC 9212 and PCC 6912, the genes located directly upstream and downstream of *ccyA* were annotated as encoding a Ca(2+)/H(+) antiporter and a Na(+)-dependent bicarbonate transporter BicA, respectively (supplementary table 4). This is particularly interesting since bicarbonate and calcium are obvious crucial ingredients for the synthesis of $CaCO_3$. Moreover, these two transporter genes are located on the same DNA strand as *ccyA* and may therefore be transcribed simultaneously with *ccyA* in a single mRNA, although this

will have to be tested by future studies. By searching homologs of these two transporters in our complete dataset of 602 cyanobacterial genomes (i.e. the genomes of the 8 iACC-forming strains described by Benzerara et al. 2014 plus the 594 genomes in which we searched for new *ccyA* homologs), we observed that their combined presence was significantly associated with that of *ccyA* (chi2 test, p-value=1.4e-08; supplementary table 5). Indeed, all 35 genomes harboring *ccyA* had at least one copy of both genes, except *Synechococcus* sp. Lanier, which lacked BicA. The latter strain was also deviant from other *ccyA*-harboring strains based on very atypical N-terminal and C-terminal calcyanin sequences. Since this strain was not available for EM analysis, we could not test if it contained iACC or not. By contrast, among the 567 genomes lacking *ccyA*, only 293 contained both transporter genes. Interestingly, in *Fischerella* sp. NIES-4106 megaplasmid, *ccyA* was located downstream a calcium/proton exchanger (sharing 92.9% identity with the above-mentioned antiporter of *C. fritschii* PCC 9212 and PCC 6912), in a region containing several additional genes potentially involved in biomineralization, such as two cation-transporting ATPases and a carbonic anhydrase (supplementary fig. 7). Overall, the correlation and/or co-localization of *ccyA* and genes involved in Ca or HCO_3^- transport and homeostasis supports the hypothesis of a functional role of calcyanin in Ca-carbonate biomineralization.

Attempts to obtain *ccyA* deletion mutants in the iACC-forming strains *Synechococcus* sp. PCC 6312 were unsuccessful but there is no certainty at this point that the employed technique can generate deletion mutants in this strain. Some possibilities to be further explored in the future are that *ccyA* deletion is lethal and/or increase the sensitivity to toxicity by calcium, suggesting that this gene may carry out an essential function in these cyanobacteria. In the absence of a direct loss-of-function genetic analysis, we overexpressed the *ccyA* genes of the two evolutionary distant cyanobacteria *Synechococcus* sp. PCC 6312 and *G. lithophora* in the non-iACC-forming, but genetically manipulable host, *Synechococcus elongatus* PCC 7942, which

does not originally contain *ccyA*. Investigation by EM-associated elemental chemical analyses of *S. elongatus* PCC 7942 cells overexpressing these *ccyA* genes did not show the presence of typical iACC (i.e. inclusions with Ca only and little to no P), whereas polyphosphate inclusions were found in cells of all mutants (fig. 6 and supplementary fig. 8). However, the comparison of Ca chemical maps of *S. elongatus* PCC 7942 mutants harboring the empty plasmid (pC) and mutants harboring its derivative expressing the *ccyA* genes (pC-$ccyA_{\text{Gloeo}}$ and pC-$ccyA_{\text{S6312}}$) showed differences. No Ca hotspot was observed in cells with the empty plasmid (pC) sampled at two different growth stages, over a total of 135 counted polyphosphate inclusions. By contrast, 23 (pC-$ccyA_{\text{Gloeo}}$) and 10 (pC-$ccyA_{\text{S6312}}$) Ca hotspots were detected over a total of 117 and 90 polyphosphate inclusions observed in the pC-$ccyA_{\text{Gloeo}}$ and pC-$ccyA_{\text{S6312}}$ mutants, respectively (fig. 6). The Ca detection limit of SEM-EDXS is not precisely known and we likely overlook some Ca. Future studies using more sensitive spatially resolved techniques will be required to have more quantitative assessment of the Ca enrichment in these cells. However, these results suggest that higher amounts of Ca were sequestered within polyphosphate inclusions when *ccyA* was present and that this gene may be functionally involved in Ca homeostasis, via a molecular process that remains to be fully elucidated.

Phylogenetic distribution and evolution of calcyanin

Whatever the function of this diagnostic gene family, constructing its phylogeny allows to infer the possible evolutionary history of iACC biomineralization. We placed the species containing the *ccyA* gene on a general phylogeny of cyanobacteria constructed using 58 conserved proteins (supplementary table 6). The four calcyanin types were found in various lineages widely dispersed across this cyanobacterial tree (fig. 4a). Whereas the X, Y and Z types showed a distribution restricted to some particular clades (*Gloeomargarita*, *Fischerella* and closely related genera, and *Microcystis*, respectively), the CoBaHMA domain (i.e. W-type) was

found in several distantly related branches (fig. 4a). Similarly, *ccyA* was detected in all the species of some clades (e.g., the *Cyanothece-Synechococcus-Thermosynechococcus* clade), suggesting that it already existed in the genome of their last common ancestor, whereas it is missing in some species of other clades such as the *Chlorogloeopsis-Fischerella* one, suggesting several independent losses and/or horizontal gene transfer (HGT) events. To better characterize these evolutionary processes, we reconstructed the phylogeny of calcyanin using the conserved GlyZip domain sequences and compared it with the corresponding cyanobacterial species tree (fig. 4c). Despite a weaker resolution of the deep branches, reflecting the higher sequence variability of calcyanin, we retrieved the monophyly of most of the groups as found in the species tree (fig. 4c), supporting the idea that *ccyA* was ancestral in these groups, and that the *ccyA*-lacking species, most likely lost it secondarily. To further compare the two trees, we carried out an approximately unbiased (AU) test (Shimodaira 2002). Whereas the species tree topology was not rejected by the GlyZip dataset (p-value = 0.64), the GlyZip topology was strongly rejected by the dataset of conserved proteins used to build the species tree (p-value = 0.00198). This strongly suggests that the differences between both trees were due to the smaller amount of phylogenetic signal contained in the GlyZip sequences compared with the set of conserved proteins and that the GlyZip sequences have evolved following the species evolution.

The overall congruence between the two trees, both retrieving the monophyly of several large cyanobacterial clades (fig. 4c), supports a very ancient origin of *ccyA* in cyanobacteria, with independent losses in various lineages. The alternative scenario of a more recent origin of *ccyA* in one group followed by its transfer to the rest by HGT was unlikely given the congruence of both trees and the extreme divergence of the N-terminal domains among the different types of calcyanin (fig. 4b). Because of its larger phylogenetic distribution, the CoBaHMA-type seemed to be the most ancient calcyanin version, while the Y- and Z-types have likely evolved in cyanobacterial groups that diverged more recently. The situation is less clear for the X-type

due to its exclusive presence in *G. lithophora*, the so-far single representative species of the poorly known Gloeomargaritales. As mentioned above, the N-terminal domains of these four types of calcyanin did not share any apparent sequence similarity (fig. 4b). This could reflect either an extreme divergence from a common ancestral domain, potentially following the adaptation of the species to their habitat needs, or the independent recruitment of non-homologous domains generating the different calcyanins by their fusion to the conserved GlyZip C-terminal domain.

To investigate if calcyanin might have originated before the diversification of cyanobacteria, we used our HMM profile to search for the GlyZip domain in other sequences present in the NCBI non-redundant database. We found homologs with a complete C-terminal domain in only five non-cyanobacterial species: an uncultured candidate phyla radiation (CPR) *Gracilibacteria* genome and four gammaproteobacteria of the Methylococcales order. We included these new sequences in a phylogenetic analysis of the GlyZip domain and found that they did not form a monophyletic group but branched intermixed with the cyanobacterial sequences (supplementary fig. 9). On the one hand, the *Gracilibacteria* sequence was very close to the *Fischerella*-*Chlorogloeopsis* group and, in agreement with this similarity of the GlyZip domain, it also contained the typical Y-type N-terminal domain found in these cyanobacteria. On the other hand, the Methylococcales sequences branched close to the *Microcystis* group and, in fact, their N-terminal domains showed some similarity with the Z-type domains of the *Microcystis* sequences. This phylogeny and the extremely sparse distribution of *ccyA* outside the Cyanobacteria phylum suggest that these few non-cyanobacterial species acquired their *ccyA* genes by HGT from *Fischerella*- and *Microcystis*-like donors, respectively. It will be interesting in future work to investigate the possible presence of iACC inclusions in these bacteria.

Conclusions

Here we show that the newly identified *ccyA* gene family, belonging to the genomic 'dark matter' (i.e. unclassified or poorly understood genetic material) of cyanobacteria, can be used as a diagnostic iACC biomineralization marker. The *ccyA*-encoded calcyanin protein has a unique architecture composed of highly divergent N-terminal domains fused with a novel, much more conserved GlyZip-containing C-terminal domain, which may adopt an original, not yet described fold. Among the diverse N-terminal domains of calcyanin that we have identified here, the domain family that we named CoBaHMA is found in the most widespread, and likely most ancient, calcyanin version. This domain family likely supports an as-yet undisclosed function within the HMA superfamily, associated with a patch of conserved basic amino acids. By tracking this gene in available genome databases, we uncovered a diversity of *ccyA*-bearing cyanobacteria capable of iACC biomineralization that is phylogenetically and environmentally much broader than previously thought, supporting a potential environmental significance. Moreover, the distribution and phylogeny of *ccyA* suggest that iACC biomineralization is ancient, with independent losses in various lineages. Additional genes are likely involved in iACC formation but, unlike *ccyA*, they may not be specific to this function and/or they are not shared by all iACC-forming cyanobacteria. The specific distribution of *ccyA* in iACC-forming cyanobacteria, its correlated presence with bicarbonate and calcium transporters, and genetic analyses, all support a pivotal role of *ccyA* in iACC biomineralization. Further investigations are required to determine whether this function may involve the conserved glutamic acid residues of the C-terminal domain, reminding Glu-rich proteins involved in ACC biomineralization (Aizenberg et al. 2002), or the basic amino acids in the N-terminal domain, which may stabilize dense liquid phases of $CaCO_3$ and delay the formation of ACC (Finney et al. 2020). Alternatively, calcyanin may have a more indirect role in iACC biomineralization serving as a cation transporter or a signaling molecule. In any case, iACC biomineralization

clearly appears as an original case of controlled biomineralization in bacteria.

Materials and Methods

Identification of candidate iACC-specific orthologous groups

In a first step, the 56 genomic assemblies used to identify groups of orthologous genes specific to iACC-forming cyanobacteria (supplementary table 1) were retrieved from the NCBI database. The 523 680 translated coding sequences derived from these genomes were processed using OrthoMCL with default settings (Li et al. 2003). This analysis included an all-vs-all blastp routine (E-value < 1e-05) and a clustering procedure into orthologous groups using the MCL algorithm.

Iterative search for homologs of calcyanin in cyanobacterial genomes

Homologs of calcyanin were searched based on similarities of the conserved C-terminal domain in 594 available genomes of cyanobacteria using an iterative process. This search dataset corresponded to the NCBI genome assemblies assigned to Cyanobacteria, published online before December 1st, 2017 (except the 6 identified in the first step, see above). For each genome assembly, we iteratively searched for homologs of calcyanin in the first set of amino acid sequences available in the following ordered list (supplementary Data 2): (i) translated CDS or (ii) proteins in RefSeq annotation records, (iii) translated CDS or (iv) proteins in GenBank annotation records.

A multiple sequence alignment of the conserved C-terminal domain was built for the 6 calcyanin sequences identified in the first step (see above), using MAFFT (Katoh and Standley 2013). A HMM-profile was generated based on this alignment with the program hmmbuild from the HMMER package (version 3.3) (Eddy 2011). The options wblossum with wid 0.62 were used to downweight closely related sequences and upweight distantly related ones. To

avoid biases towards glycine-rich unrelated proteins, we artificially reduced glycine weight by 20% in the profiles. The profile *vs* sequence similarity search was done with the program hmmsearch (E-value < 1.0e-70). The hits matching 100% of the profile length and corresponding to newly identified sequences were added to the new calcyanin dataset. The multiple sequence alignment and the HMM-profile of this dataset were then updated. These steps (alignment, building of HMM-profile, similarity search) were repeated until no new sequence was detected. In order to detect remote homologs of calcyanin, seven iterations of the entire process were done as described in supplementary table 7, with a progressive decrease of the stringency of the similarity search. In the beginning, we set a very low E-value and high cover to the profile. As the iterations proceeded, we increased the E-value and decreased the cover to the profile down to 70%. This cover threshold higher than 66% was designed to avoid (Gly)2 (instead of (Gly)3) to be matched. At the end of the whole process, we used the final HMM profile (as provided in supplementary Data 3) to search for similarity in the GenBank records of the processed genomic assemblies.

Last, *ccyA* was searched in the newly sequenced genomes of *Synechococcus* sp. PCC 6716 and PCC 6717 using tBLASTn with all previously identified *ccyA* sequences as queries. The CDS boundaries of the best BLAST hits were further assessed using Prodigal (Hyatt et al. 2010).

Comparative genomics of C. fritschii PCC 6912 (no iACC observed) and C. fritschii PCC 9212 (with iACC)

The search of homologous genes shared by *C. fritschii* PCC 6912 and PCC 9212 genomes was achieved based on unidirectional BLASTp best hits as implemented in the PATRIC proteome comparison tool (Gillespie et al. 2011) (E-value < 1.0e-05, sequence coverage > 30%). For each genome assembly, we used the set of translated CDS as provided in the RefSeq annotation record. Gene functional categories were searched in COG database (v1) using CD-search

(Marchler-Bauer and Bryant 2004) (E-value < 1.0e-05). The nucleotide sequences of the *ccyA*-containing contigs of *C. fritschii* PCC 6912 and PCC 9212 (NCBI accessions NZ_AJLN01000033.1 and NZ_AJLM01000017.1, 97 542 bp and 97 528 bp, respectively) were compared using BLASTn.

Search for homologs of the Ca(2+)/H(+) antiporter and BicA Na(+)-dependent bicarbonate transporter in cyanobacterial genomes

Homologs of the Ca(2+)/H(+) antiporter and the Na(+)-dependent bicarbonate transporter BicA, encoded in *C. fritschii* PCC 6912 and PCC 9212 by the genes located upstream and downstream of *ccyA*, respectively, were searched using BLASTp (E-value < 1.0e-10) in our complete dataset of 602 cyanobacterial genomes (composed of the 8 iACC-forming strains described by (Benzerara et al. 2014) in which we initially detected *ccyA* and by the 594 genomes in which we iteratively searched for new *ccyA* homologs in a second step). Owing to the incompleteness of the *ccyA*-upstream gene in the genomic sequence of these two strains, we used the most similar full-length sequence as Ca(2+)/H(+) antiporter query (96% identity; accession WP_016868870.1, *Fischerella muscicola* PCC 7414). BicA homologs were identified using the protein sequence from *C. fritschii* PCC 6912 and PCC 9212 as query (accession WP_016872894.1).

Calcyanin functional annotation and structure prediction

The structural features of calcyanin were explored based on the information provided by amino acid sequences using Hydrophobic Cluster Analysis (HCA) (Callebaut et al. 1997; Bitard-Feildel et al. 2018). HCA provides a global view of the protein texture, with insights into the structural features of foldable regions (Bitard-Feildel et al. 2018). Similarities between domains composing calcyanin and known domains/3D structures were searched against different databases (NCBI nr sequence database, NCBI Conserved Domain Database (CDD) (Yang et

al. 2020) and the Protein Data Bank (PDB)) using tools for profile-sequence and profile-profile comparison such as PSI-BLAST (Altschul et al. 1997) and HH-PRED (Zimmermann et al. 2018), respectively.

3D structure modelling was performed using Modeller 9.23 (Webb and Sali 2016). This modeling was refined afterwards using AlphaFold (Jumper et al. 2021), through the notebook Alphafold2_advanced from Colabfold (Mirdita et al. 2021) (ngithub/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb). The full sequence of *Synechococcus* sp. RS9917 CcyA and the multiple sequence alignment of the 15 reported CoBaHMA-bearing CcyA were used as input. 3D structures were visualized using UCSF Chimera (Pettersen et al. 2004). Multiple sequence alignment handling and rendering were made using SeaView (Gouy et al. 2010) and EsPript (Robert and Gouet 2014), respectively.

Molecular phylogenetic analyses

Phylogeny of cyanobacteria using different sets of species was reconstructed using 58 conserved proteins (Moreira et al. 2017; supplementary table 6). Each individual protein was aligned using MAFFT with the accurate L-INS-I option (Katoh and Standley 2013) and poorly aligned regions were removed with trimAl –automated1 (Capella-Gutiérrez et al. 2009). Trimmed alignments were concatenated to produce a supermatrix and maximum likelihood phylogenetic trees were reconstructed with the program IQ-Tree using the mixture model LG+C60+F+G (Nguyen et al. 2015). Statistical support was estimated using 1000 bootstrap replicates. The phylogeny of calcyanin was studied using the manually curated alignment of the conserved GlyZip C-terminal domain. A maximum likelihood tree was constructed with the program IQ-Tree using the mixture model LG+C20+F+G (Nguyen et al. 2015). Statistical support was estimated using 1000 bootstrap replicates.

Tree topologies based in the set of 58 conserved proteins (species tree) and in the GlyZip C-terminal domain of calcyanin were compared using the AU test (Shimodaira 2002)

implemented in IQ-TREE with the options -n 0 -zb 10000 -au -zw (Nguyen et al. 2015). The sequence evolution models used were, as before, LG+C60+F+G for the conserved protein dataset and LG+C20+F+G for the GlyZip dataset.

Electron microscopy analyses of iACC

Strains recovered from culture collections were analyzed by scanning transmission electron microscopy (STEM) for iACC search. As previously shown by Benzerara et al (2014), Li et al. (2016) and De Wever et al. (2019), iACC can be recognized based on the fact that they mostly contain Ca with little to no P, whereas polyphosphate inclusions show a major P peak with Mg, K and/or Ca. For that purpose, we used a field emission gun JEOL-2100F microscope operating at 200 kV, equipped with a JEOL detector with an ultrathin window allowing detection of light elements. STEM allowed Z-contrast imaging in the HAADF mode. EDXS analyses rely on the detection of x-rays emitted by samples excited by the electron beam. Their energy is characteristic of the atoms and their intensity depends on the atomic content. Compositional maps of Ca, P, and C were acquired by performing EDXS analysis in the STEM high angle annular dark field (HAADF) mode. These EDXS analyses provide hyperspectral data, i.e. an image with EDXS spectra for each pixel of the image. For these analyses, a total of 0.5 mL of cultures aged between 5 and 30 days was centrifuged at 8,000 × g for 10 min. Pellets were rinsed three times in Milli-Q (mQ) water (Millipore). After the final centrifugation, pellets were suspended in 200 µL of mQ water. A drop of 5 µL was deposited on a glow discharged carbon-coated 200-mesh copper grid and let dry at ambient temperature.

For iACC-forming strains, we systematically measured several replicates by STEM and/or EDXS associated with scanning electron microscopy. Even more effort was invested in the analysis of strains harboring *ccyA* but not showing iACC. Indeed, while showing the presence of iACC in a strain only requires one single positive observation, concluding about the absence

of iACC is difficult, if not impossible. For *Fischerella* sp. NIES-3754, we performed 7 different SEM or STEM sessions over four different cultures, including two on the same culture with a 15 days interval and three on a second culture with three and nine days interval. For *Chlorogloeopsis fritschii* PCC 6912, we performed eight different SEM or STEM sessions over five different cultures, including three on the same culture with a four and six days interval and two on another culture with a 15 days interval. For *Microcystis aeruginosa* PCC 9432, we performed 7 different SEM or STEM sessions over four different cultures, including two on the same culture with a three days interval and three on another culture with three and eight days interval; For *M. aeruginosa* PCC 9717, we performed six different SEM or STEM sessions over four different cultures, including two on the same culture with a three days interval and two on another culture with 25 days interval.

Mutant strains of *Synechococcus elongatus* PCC 7942 harboring the pC, pC-$ccyA_{Gloe}$ or $ccyA_{S6312}$ were analyzed by scanning electron microscopy (SEM) in the backscattered electron mode, coupled with EDXS analyses to search for Ca enrichment. Analyses were replicated twice on at least three and up to six areas. Ca hotspots were identified each time and the signal in the Ca energy range was higher than the background by 1σ. One example of EDXS spectrum is provided per type of mutant in supplementary figure 8.

Genetics

The pC-$ccyA_{Gloeo}$ and pC-$ccyA_{S6312}$ plasmids were derivatives of the RSF1010-derived pC vector (Veaudor et al. 2018) replicating in *Escherichia coli* (supplementary table 8 and supplementary figure 10). Chenebault et al. (2020) showed that this expression plasmid allowed strong gene expression in cyanobacteria. The pC-$ccyA_{Gloeo}$ and pC-$ccyA_{S6312}$ plasmids were transferred to *Synechococcus elongatus* PCC 7942 by trans-conjugation (Mermet-Bouvier and Chauvat 1994), using the improved triparental-mating protocol that follows. Overnight-grown cultures of the *E. coli* strains CM404, which propagates the self-transferable mobilization vector

pRK2013, and TOP10, which propagates either pC, pC-*ccyA*$_{Gloeo}$ or pC-*ccyA*$_{S6312}$, were washed twice and resuspended in LB medium (1×10^9 cells.mL^{-1}). Meanwhile, *S. elongatus* PCC 7942 mid-log phase cultures grown in mineral growth medium (MM, a version of BG-11 supplemented with 3.78 mM Na_2CO_3) were centrifuged and concentrated five times (about 1×10^8 cells/mL) in fresh MM. Then, 100 µL of *S. elongatus* PCC 7942 cells were mixed with 30 µL of CM404 cells and 30 µL of TOP10 cells harboring either pC, pC-*ccyA*$_{Gloeo}$ or pC-*ccyA*$_{S6312}$. 30 µL aliquots of this mixture were spotted onto MM solidified with 1% agar (Difco), and incubated for 48 h under standard temperature (30°C) and light (2500 lux, i.e. 31 µE.m^{-2}.s^{-1}) conditions. Then, cells were collected from each plate and resuspended into 50 µL of liquid MM, prior to plating onto MM containing 5 µg.mL^{-1} of each the streptomycin (Sm) and spectinomycin (Sp) selective antibiotics. After about 10 days of incubation under standard light and temperature conditions, Sm^{R}Sp^R resistant conjugant clones were collected and re-streaked onto selective plates, prior to analyzing their plasmid content by PCR and DNA sequencing (Eurofins Genomics) using specific primers (supplementary Data 1c, 1d and supplementary figure 10).

Data availability statement

Further information and requests for resources, codes and reagents should be directed to and will be fulfilled by the lead contact, Karim Benzerara (karim.benzerara@upmc.fr).

• Plasmids and mutant strains generated in this study are available upon request to the lead contact.

• The genomic assemblies are available at GenBank as follows:

Synechococcus calcipolaris PCC 11701 - BioProject PRJNA800269

Synechococcus sp. PCC 6716 - BioProject PRJNA801107

Synechococcus sp. PCC 6717 - BioProject PRJNA801158 .

Accession numbers are listed in the key resources table.

• TEM-EDXS and SEM-EDXS data and the structure of the CoBaHMA domain have been uploaded to Zenodo: 10.5281/zenodo.5964253. DOIs will be listed in the key resources table.

Acknowledgments

We thank two anonymous reviewers for their constructive comments which improved the overall quality of the manuscript. We thank Alexis De Wever and Marine Blondeau for helping acquiring some TEM data. We thank Mélanie Poinsot for helping in the preparation of some samples for transmission electron microscopy analyses. We thank Eva Jahodarova for shipping a culture of *Neosynehcococcus. sphagnicola*. This work was supported by the Agence Nationale de la Recherche (ANR Harley, ANR-19-CE44-0017-01; ANR PHOSTORE, ANR-19-CE01-0005) and the European Research Council under the European Union's Seven Framework Program: ERC grants Calcyan (PI: K. Benzerara, Grant Agreement no. 307110) and PlastEvol (PI: D. Moreira, Grant Agreement no. 787904). Sigrid Görgen PhD grant was funded by the Sorbonne Université doctoral program Interfaces pour le Vivant.

Author Contributions:

KBE, EDU, CCC, FCH, DMO, PLG, ICA conceived and designed the work. KBE, EDU, TBF, GCA, CCC, MDE, IDI, GGA, MGU, SGO, FSP, DMO, ICA acquired, analysed and/or interpreted data. KBE, EDU, CCC, FCH, MGU, PLG, DMO, ICA drafted the work or substantively revised it.

References

Aizenberg J, Lambert G, Weiner S, Addadi L. 2002. Factors involved in the formation of amorphous and crystalline calcium carbonate: a study of an ascidian skeleton. *J Am Chem Soc* 124:32–39.

Altermann W, Kazmierczak J, Oren A, Wright DT. 2006. Cyanobacterial calcification and its rock-building potential during 3.5 billion years of Earth history. *Geobiology* 4:147–166.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.

Benzerara K, Skouri-Panet F, Li J, Férard C, Gugger M, Laurent T, Couradeau E, Ragon M, Cosmidis J, Menguy N, et al. 2014. Intracellular Ca-carbonate biomineralization is widespread in cyanobacteria. *Proc Natl Acad Sci U S A* 111:10933–10938.

Bitard-Feildel T, Lamiable A, Mornon J-P, Callebaut I. 2018. Order in disorder as observed by the "Hydrophobic Cluster Analysis" of protein sequences. *Proteomics* 18:e1800054.

Blondeau M, Benzerara K, Ferard C, Guigner J-M, Poinsot M, Coutaud M, Tharaud M, Cordier L, Skouri-Panet F. 2018. Impact of the cyanobacterium *Gloeomargarita lithophora* on the geochemical cycles of Sr and Ba. *Chemical Geology* 483:88–97.

Blondeau M, Sachse M, Boulogne C, Gillet C, Guigner J-M, Skouri-Panet F, Poinsot M, Ferard C, Miot J, Benzerara K. 2018. Amorphous Calcium Carbonate Granules Form Within an Intracellular Compartment in Calcifying Cyanobacteria. *Front. Microbiol.* 9:1768.

Blue CR, Giuffre A, Mergelsberg S, Han N, De Yoreo JJ, Dove PM. 2017. Chemical and physical controls on the transformation of amorphous calcium carbonate into crystalline CaCO3 polymorphs. *Geochim. Cosmochim. Acta* 196:179–196.

Bradley JA, Daille LK, Trivedi CB, Bojanowski CL, Stamps BW, Stevenson BS, Nunn HS, Johnson HA, Loyd SJ, Berelson WM, et al. 2017. Carbonate-rich dendrolitic cones: insights into a modern analog for incipient microbialite formation, Little Hot Creek, Long Valley Caldera, California. *npj Biofilms Microbomes* 3:32.

Bull PC, Cox DW. 1994. Wilson disease and Menkes disease: new handles on heavy-metal transport. *Trends Genet* 10:246–252.

Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP. 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci* 53:621–645.

Cam N, Benzerara K, Georgelin T, Jaber M, Lambert J-F, Poinsot M, Skouri-Panet F, Cordier L. 2016. Selective uptake of alkaline earth metals by cyanobacteria forming intracellular carbonates. *Environ. Sci. Technol.* 50:11654–11662.

Cam N, Benzerara K, Georgelin T, Jaber M, Lambert J-F, Poinsot M, Skouri-Panet F, Moreira D, López-García P, Raimbault E, et al. 2018. Cyanobacterial formation of intracellular Ca-carbonates in undersaturated solutions. *Geobiology* 16:49–61.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.Couradeau E, Benzerara K, Gerard E, Moreira D, Bernard S, Brown GE, Lopez-Garcia P. 2012. An Early-Branching Microbialite Cyanobacterium Forms Intracellular Carbonates. *Science* 336:459–462.

De la Concepcion JC, Franceschetti M, Maqbool A, Saitoh H, Terauchi R, Kamoun S, Banfield MJ. 2018. Polymorphic residues in rice NLRs expand binding and response to effectors of the blast pathogen. *Nature Plants* 4:576–585.

De Wever A, Benzerara K, Coutaud M, Caumes G, Poinsot M, Skouri-Panet F, Laurent T, Duprat E, Gugger M. 2019. Evidence of high Ca uptake by cyanobacteria forming intracellular $CaCO_3$ and impact on their growth. *Geobiology* 17:676–690.

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLOS Computational Biology* 7:e1002195.

Finney AR, Innocenti Malini R, Freeman CL, Harding JH. 2020. Amino acid and oligopeptide effects on calcium carbonate solutions. *Crystal Growth & Design* 20:3077–3092.

Frangeul L, Quillardet P, Castets A-M, Humbert J-F, Matthijs HCP, Cortez D, Tolonen A, Zhang C-C, Gribaldo S, Kehr J-C, et al. 2008. Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* 9:274.

Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, Hix D, Mane SP, Mao C, et al. 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun* 79:4286–4298.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 27:221–224.

Humbert J-F, Barbe V, Latifi A, Gugger M, Calteau A, Coursin T, Lajus A, Castelli V, Oztas S, Samson G, et al. 2013. A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium *Microcystis aeruginosa*. *PLoS One* 8:e70747.

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.

Jiang D, Zhao Y, Fan J, Liu X, Wu Y, Feng W, Zhang XC. 2014. Atomic resolution structure of the E. coli YajR transporter YAM domain. *Biochem Biophys Res Commun* 450:929–935.

Jiang D, Zhao Y, Wang X, Fan J, Heng J, Liu X, Feng W, Kang X, Huang B, Liu J, et al. 2013. Structure of the YajR transporter suggests a transport mechanism based on the conserved motif A. *Proc Natl Acad Sci U S A* 110:14664–14669.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy

E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583-589.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 30:772–780.

Kim S, Jeon T-J, Oberai A, Yang D, Schmidt JJ, Bowie JU. 2005. Transmembrane glycine zippers: physiological and pathological roles in membrane proteins. *Proc Natl Acad Sci U S A* 102:14278–14283.

Komarek J, Johansen JR, Smarda J, Strunecky O. 2020. Phylogeny and taxonomy of Synechococcus-like cyanobacteria. *Fottea* 20:171–191.

Kuehlbrandt W. 2019. Structure and Mechanisms of F-Type ATP Synthases. In: Kornberg RD, editor. Annual Review of Biochemistry. Vol. 88. Palo Alto: Annual Reviews. p. 515–549.

Lamiable A, Bitard-Feildel T, Rebehmed J, Quintus F, Schoentgen F, Mornon J-P, Callebaut I. 2019. A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. *Biochimie* 167:68–80.

Latour D, Salençon M-J, Reyss J-L, Giraudet H. 2007. Sedimentary Imprint of *Microcystis aeruginosa* (cyanobacteria) Blooms in Grangent Reservoir (Loire, France). *Journal of Phycology* 43:417–425.

Lefevre CT, Bazylinski DA. 2013. Ecology, diversity, and evolution of magnetotactic bacteria. *Microbiol. Mol. Biol. Rev.* 77:497–526.

Leonov H, Arkin IT. 2005. A periodicity analysis of transmembrane helices. *Bioinformatics* 21:2604–2610.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* 13:2178–2189.

Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* 32:W327-331.

Marron AO, Ratcliffe S, Wheeler GL, Goldstein RE, King N, Not F, de Vargas C, Richter DJ. 2016. The Evolution of Silicon Transport in Eukaryotes. *Molecular Biology and Evolution* 33:3226–3248.

Mehta N, Benzerara K, Kocar BD, Chapon V. 2019. Sequestration of radionuclides radium-226 and strontium-90 by cyanobacteria forming intracellular calcium carbonates. *Environ. Sci. Technol.* 53:12639–12647.

Mermet-Bouvier P, Chauvat F. 1994. A conditional expression vector for the cyanobacteria *Synechocystis* sp. strains PCC6803 and PCC6714 or *Synechococcus* sp. strains PCC7942 and PCC6301. *Curr Microbiol* 28:145–148.

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2021. ColabFold - Making protein folding accessible to all. bioRxiv, DOI: 10.1101/2021.08.15.456425.

Monteil CL, Benzerara K, Menguy N, Bidaud CC, Michot-Achdjian E, Bolzoni R, Mathon FP, Coutaud M, Alonso B, Garau C, et al. 2020. Intracellular amorphous Ca-carbonate and magnetite biomineralization by a magnetotactic bacterium affiliated to the Alphaproteobacteria. *The ISME Journal*:1–18.

Moreira D, Tavera R, Benzerara K, Skouri-Panet F, Couradeau E, Gerard E, Fonta CL, Novelo E, Zivanovic Y, Lopez-Garcia P. 2017. Description of Gloeomargarita lithophora gen. nov., sp nov., a thylakoid-bearing, basal-branching cyanobacterium with intracellular carbonates, and proposal for Gloeomargaritales ord. nov. *Int. J. Syst. Evol. Microbiol.* 67:653–658.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32:268–274.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera-a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612.

Ragon M, Benzerara K, Moreira D, Tavera R, Lopez-Garcia P. 2014. 16S rDNA-based analysis reveals cosmopolitan occurrence but limited diversity of two cyanobacterial lineages with contrasted patterns of intracellular carbonate mineralization. *Front. Microbiol.* [Internet] 5. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2014.00331/abstract

Reynolds CS, Rogers DA. 1976. Seasonal variations in the vertical distribution and buoyancy of *Microcystis aeruginosa* Kütz. emend. Elenkin in Rostherne Mere, England. *Hydrobiologia* 48:17–23.

Riding R. 2012. A Hard Life for Cyanobacteria. *Science* 336:427–428.

Robert X, Gouet P. 2014. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res* 42:W320-324.

Senes A, Engel DE, DeGrado WF. 2004. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr. Opin. Struct. Biol.* 14:465–479.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol*. 51:492–508.

Veaudor T, Ortega-Ramos M, Jittawuttipoka T, Bottin H, Cassier-Chauvat C, Chauvat F. 2018. Overproduction of the cyanobacterial hydrogenase and selection of a mutant thriving on urea, as a possible step towards the future production of hydrogen coupled with water treatment. *PLoS One* 13:e0198836.

Wang X, Zoccola D, Liew YJ, Tambutte E, Cui G, Allemand D, Tambutte S, Aranda M. 2021. The evolution of calcification in reef-building corals. *Mol Biol Evol*:msab103.

Webb B, Sali A. 2016. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci* 86:2.9.1-2.9.37.

Weiner S, Dove PM. 2003. An overview of biomineralization processes and the problem of the vital effect. In: Dove PM, DeYoreo JJ, Weiner S, editors. Biomineralization. Vol. 54. Chantilly: Mineralogical Soc Amer. p. 1–29.

Yang M, Derbyshire MK, Yamashita RA, Marchler-Bauer A. 2020. NCBI's conserved domain database and tools for protein domain analysis. *Curr Protoc Bioinformatics* 69:e90.

Yarra T, Blaxter M, Clark MS. 2021. A bivalve biomineralization toolbox. *Mol Biol Evol*:msab153.

Zhao L, Song Y, Li L, Gan N, Brand JJ, Song L. 2018. The highly heterogeneous methylated genomes and diverse restriction-modification systems of bloom-forming Microcystis. *Harmful Algae* 75:87–93.

Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J Mol Biol* 430:2237–2243.

Figure legends

Figure 1: **Domain architecture of calcyanins, as viewed by Hydrophobic Cluster Analysis. HCA plots of the calcyanin sequences of *Synechococcus calcipolaris* PCC 11701 and *Gloeomargarita lithophora* D10.** The protein amino acid sequences (one-letter code) are displayed on a duplicated alpha-helical net, on which the strong hydrophobic amino acids (V, I, L, F, M, Y, W) are contoured. The latter form clusters, which mainly correspond to the internal faces of regular secondary structures (α-helices and β-strands). The way to read the primary (1D) and secondary (2D) structures is shown with arrows (one amino acid or one hydrophobic cluster after another, respectively), whereas special symbols used for four amino acids with specific structural properties (P, G, S, T) are described in the inset, together with the color code used to highlight conserved amino acids within the periodic patterns of the two calcyanin sequences. The two distinct CcyA folded domains (~1/3 strong hydrophobic amino acids) are boxed.

Figure 2. Electron microscopy detection of iACC in 13 calcyanin-bearing cyanobacterial

strains not previously known to biomineralize carbonates. STEM-HAADF images of the 13 newly-identified iACC-forming strains and overlays of C (blue), Ca (green) and P (red) chemical maps as obtained by energy dispersive x-ray spectroscopy. The name of the strains is provided on the STEM-HAADF image. Numbers in parenthesis correspond to replicate numbers of SEM-EDXS, STEM-EDXS or both analyses. A and B: *Chlorogloeopsis fritschii* PCC 9212 (13); C and D: *Fischerella muscicola* PCC 7414 (4); E and F: *Fischerella* sp. NIES-4106 (5); G and H: *Microcystis aeruginosa* PCC 7806 (9); I and J: *M. aeruginosa* PCC 7941 (7); K and L: *M. aeruginosa* PCC 9443 (3); M and N: *M. aeruginosa* PCC 9806 (4); O and P: *M. aeruginosa* PCC 9807 (3); Q and R: *M. aeruginosa* PCC 9808 (4); S and T: *Neosynechococcus sphagnicola* sy1 (4); U and V: *Synechococcus lividus* PCC 6715 (3); W and X: *Synechococcus* sp. RS9917 (4); Y and Z: *Thermosynechococcus* sp. NK55 (6).

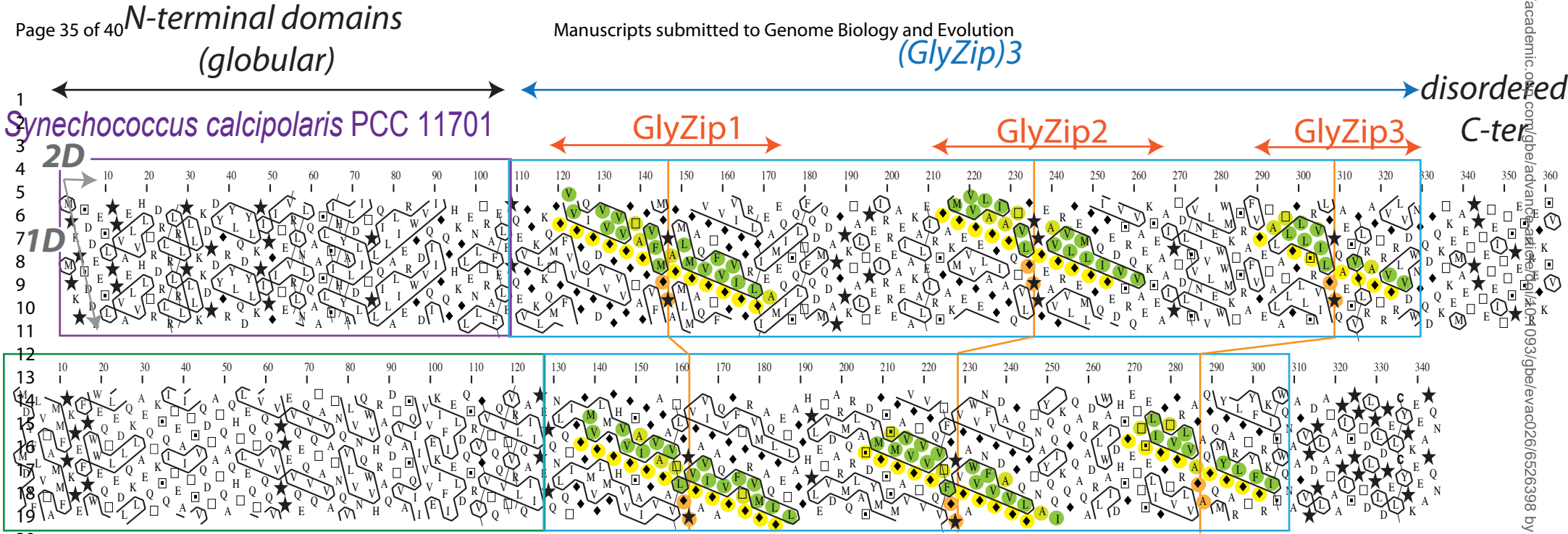
Figure 3: TEM analyses of the four *ccyA*-harbouring strains not forming iACC. Each row corresponds to one strain. The first column shows STEM-HAADF images. The second column shows overlays of C, Ca and P EDXS maps. The third column shows EDXS spectra of inclusions detected in the cells. A, B and C: *Fischerella* sp. NIES-3754. EDXS spectrum is extracted from the area indicated in A) by a dashed line; D, E and F: *Chlorogloeopsis fritschii* PCC 6912. G, H and I: *Microcystis aeruginosa* PCC 9432; J, K and L: *M. aeruginosa* PCC 9717.

Figure 4: Phylogenetic analysis and domain architecture of the calcyanin protein family. **a**, Maximum likelihood phylogenetic tree of Cyanobacteria based on 58 conserved proteins; the strains containing the *ccyA* gene are highlighted in bold and color. **b**, HCA plots of representative calcyanin sequences (see fig. 1 for details of the HCA representation). The positions of the domains are indicated, with red boxes corresponding to the duplicated

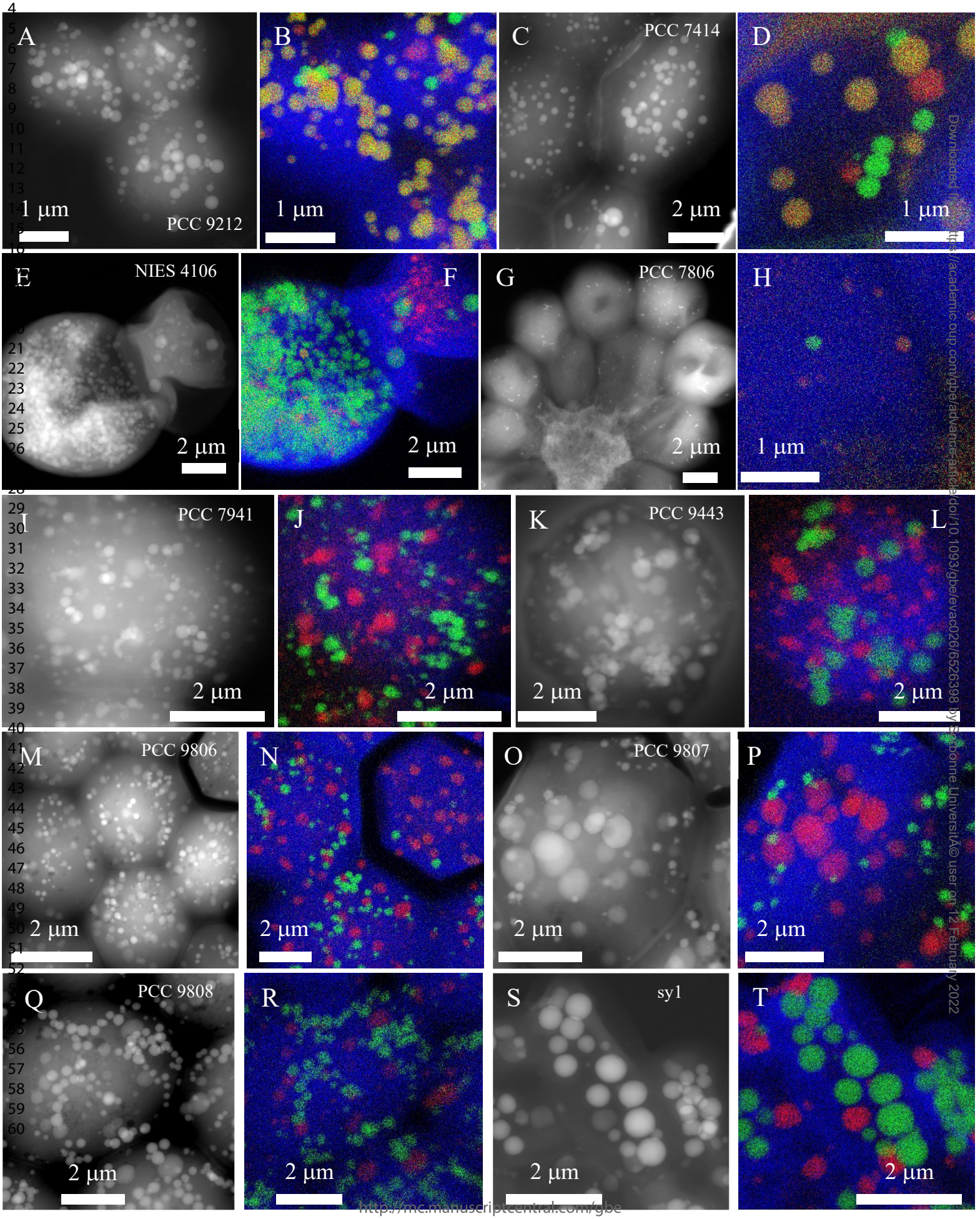
subdomain composing domain Y (labels *a* to *e* refer to equivalent hydrophobic clusters). The periodic patterns, made of glycine (or small amino acids – yellow) and hydrophobic amino acids (green) are highlighted for each GlyZip, with conserved signatures specific of each GlyZip shown with other colors. GlyZip2, which is present in only one species in the Y family, is indicated with a dotted box. **c**, Unrooted maximum likelihood phylogenetic tree of the GlyZip domain of calcyanin (left) compared with the species tree based on 58 conserved proteins (right). Numbers on branches indicate bootstrap support (BS, only values >50% are shown), BS of 100% are indicated by black circles. The species names and HCA profiles are color-coded according to the type of N-terminal domain of calcyanins (the code is shown at the bottom of the figure).

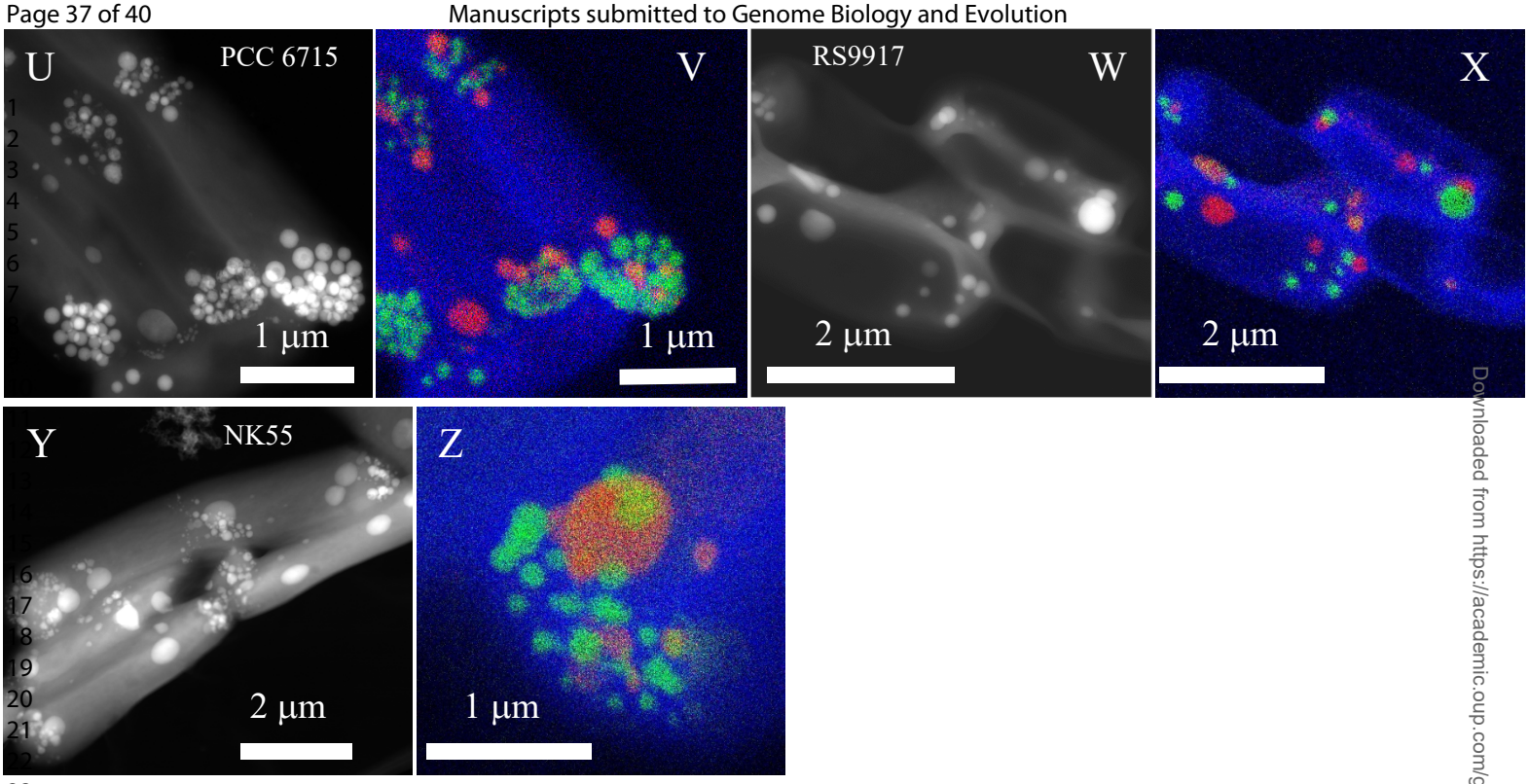
Figure 5: The CoBaHMA domain. A. Multiple sequence alignment of calcyanins and members of the HMA superfamily with known 3D structures. Identical amino acids are shown white on a black background, similarities are colored according to amino acid properties (inset). Sequences of proteins of the HMA superfamily, whose 3D structures are known and with which the CoBaHMA sequences can be aligned, are shown on top. PDB identifiers are provided. Observed 2D structures are boxed. The two cysteines of the CXXC motif specific of the HMA family are boxed in red. Green dots highlight the positions in which the hydrophobic character is strongly conserved, corresponding to amino acids participating in the hydrophobic core of the ferredoxin fold. An additional β-strand, named β0, is predicted in the CoBaHMA sequences, including a strictly conserved histidine. B**.** Model of the CoBaHMA 3D structure, illustrated here with the *Synechococcus* sp. RS9917 sequence. The HMA common core is colored in beige, whereas specific secondary structures of the CoBaHMA family are in blue. The four conserved basic amino acids are shown with atomic details.

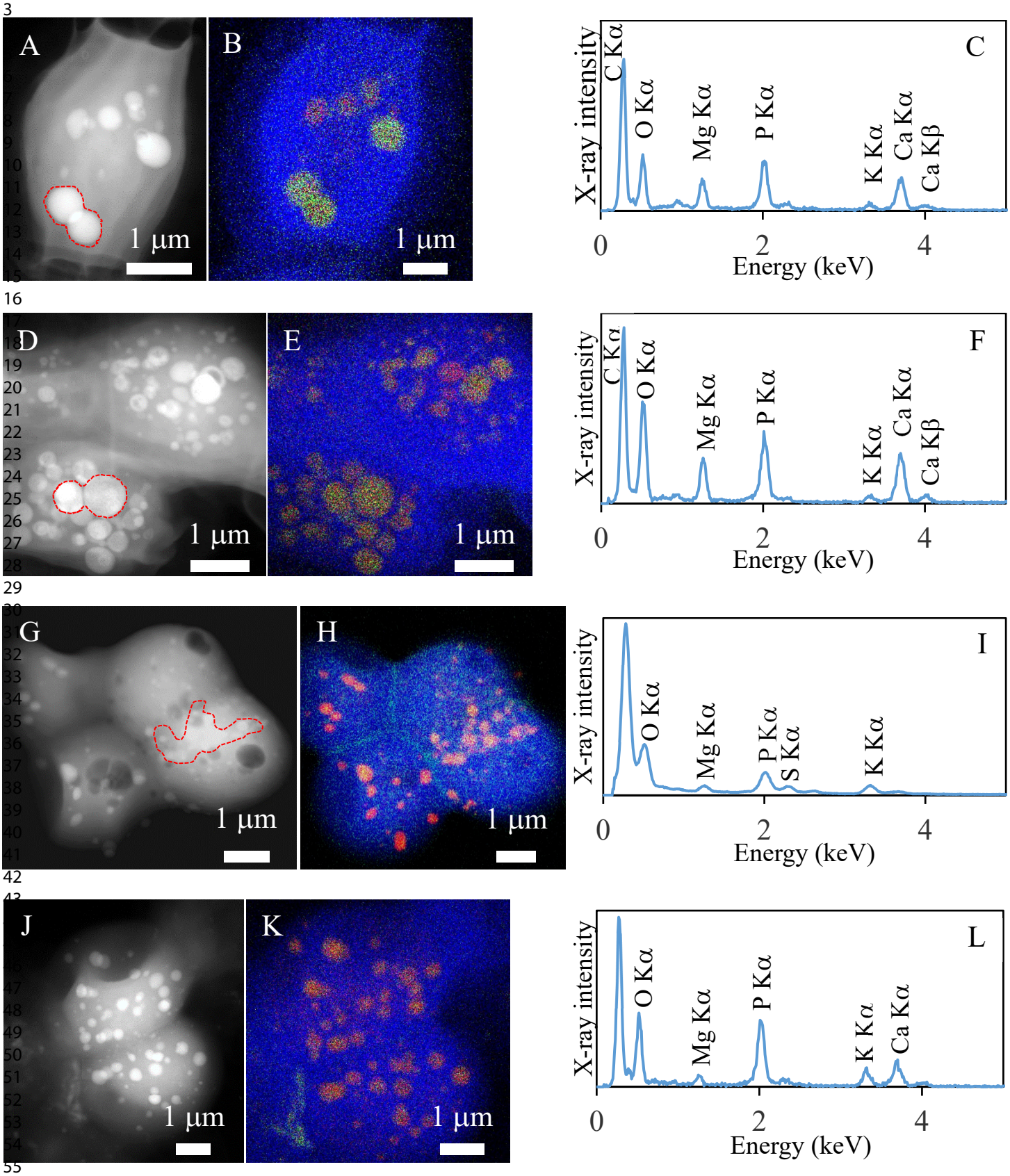
Figure 6. **SEM analyses of mutants overexpressing *ccyA*.** SEM-EDXS images (in backscattered electron (BSE) mode), P (green) and Ca (red) maps of *Synechochoccus elongatus* PCC 7942 mutants. The scale bar provided on the BSE images is the same for the corresponding P and Ca maps on each row. The 0.2 μm pores of the filters appear as dark disks in the BSE images. At the accelerating voltage used for these analyses, *S. elongatus* cells appear as relatively transparent, packed rods. Polyphosphate inclusions appear as brighter dots. The first three rows show cells of a *S. elongatus* PCC 7942 mutant harboring the empty pC plasmid. No Ca-rich inclusions are observed in these cells as shown by the homogeneous background in the Ca maps. In contrast, Ca-rich inclusions (polyphosphates) are observed in cells of *S. elongatus* PCC 7942 mutants harboring the plasmids pC-$ccyA_{\mathrm{Gloeo}}$ (fourth row) or pC-$ccyA_{\mathrm{S6312}}$ (fifth and sixth rows), appearing as hotspots in Ca maps. See supplementary data 3 for details concerning the plasmid and strains.

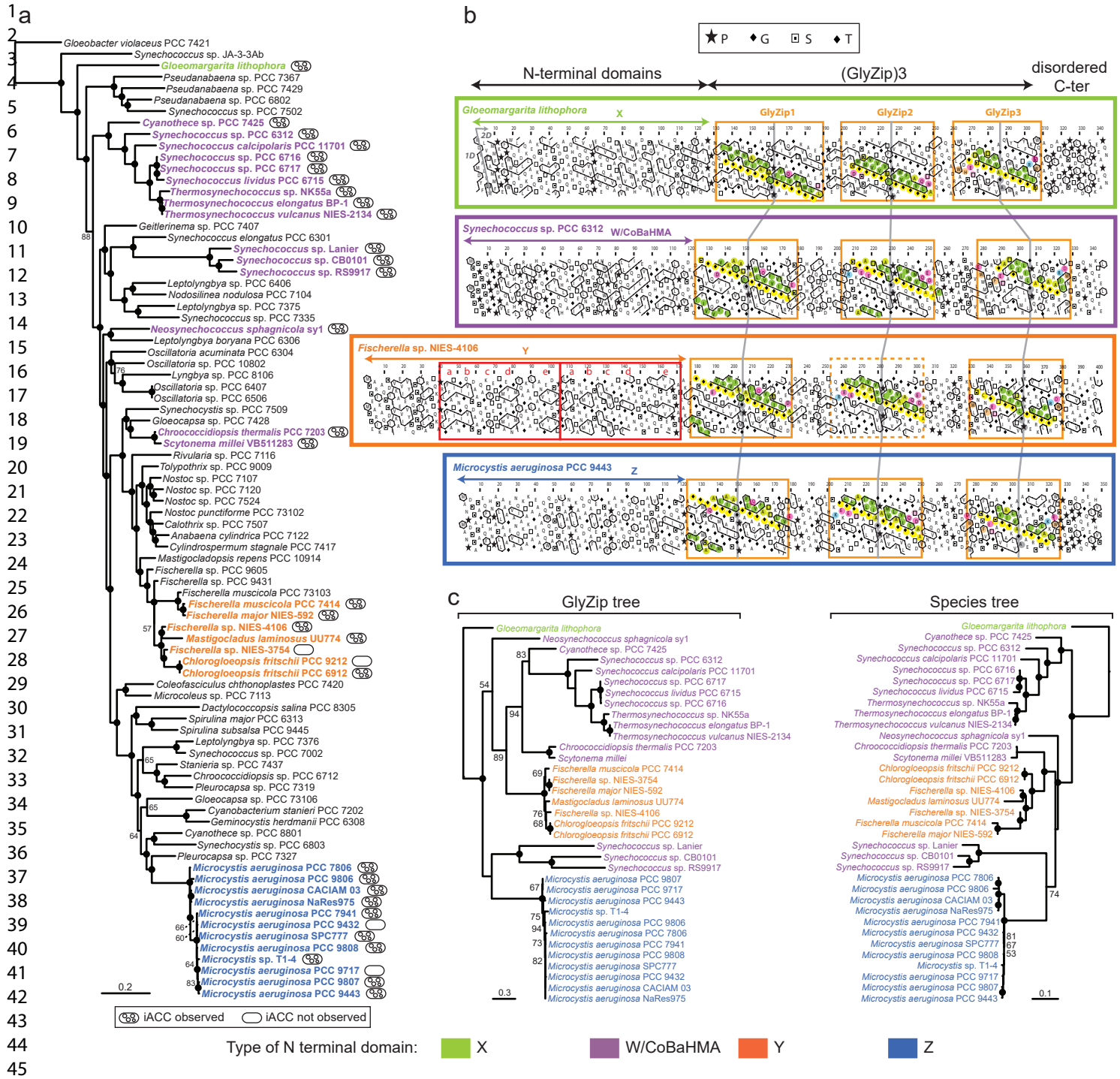


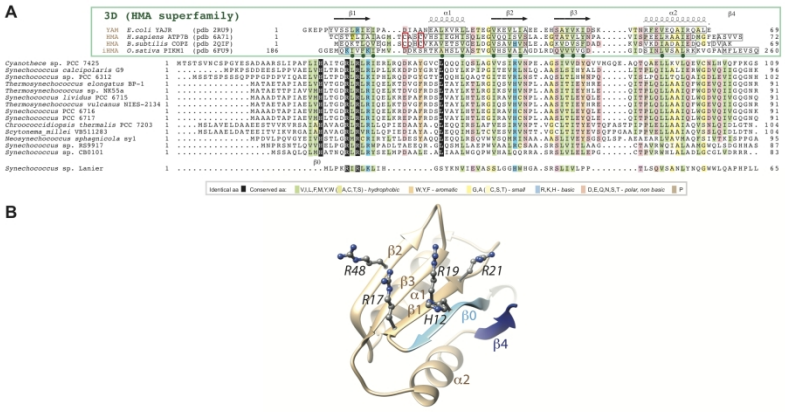
Symbols	Conserved amino acids in the GlyZip periodic pattern:
★ P ▣ S ◆ G □ T	hydrophobic (V,I,L,M,F,Y,W/A,T) small (G/A,T,S) central G-P/G-A dipeptide











874x1237mm (72 x 72 DPI)