


# A New Gene Family Diagnostic for Intracellular Biomineralization of Amorphous Ca Carbonates by Cyanobacteria

Karim Benzerara <sup>1,\*†,‡</sup>, Elodie Duprat<sup>1,†</sup>, Tristan Bitard-Feildel<sup>1</sup>, Géraldine Caumes<sup>1</sup>, Corinne Cassier-Chauvat<sup>2</sup>, Franck Chauvat<sup>2</sup>, Manuela Dezi<sup>1</sup>, Seydina Issa Diop<sup>1,§</sup>, Geoffroy Gaschignard<sup>1</sup>, Sigrid Görden<sup>1,2</sup>, Muriel Gugger<sup>3</sup>, Purificación López-García<sup>4</sup>, Maxime Millet<sup>1</sup>, Fériel Skouri-Panet<sup>1</sup>, David Moreira<sup>4,†,‡</sup> and Isabelle Callebaut<sup>1,\*†</sup>

<sup>1</sup>Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590. Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC), Paris, France

<sup>2</sup>Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France

<sup>3</sup>Institut Pasteur, Université de Paris, Collection of Cyanobacteria, Paris, France

<sup>4</sup>Unité d'Ecologie Systématique et Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Orsay, France

<sup>†</sup>These authors contributed equally to this work.

<sup>\*</sup>Lead contact.

<sup>§</sup>Present address: Department of Systematic and Evolutionary Botany & Zurich-asel Plant Science Center, University of Zurich, Zollikerstrasse 107, Zurich, Switzerland.

\*Corresponding authors: E-mails: karim.benzerara@sorbonne-universite.fr; isabelle.callebaut@sorbonne-universite.fr.

Accepted: January 29, 2022

## Abstract

Cyanobacteria have massively contributed to carbonate deposition over the geological history. They are traditionally thought to biomineralize CaCO<sub>3</sub> 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 four 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 colocalized with genes involved in Ca or HCO<sub>3</sub><sup>-</sup> 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.

**Key words:** biomineralization, amorphous calcium carbonates, cyanobacteria, protein structure prediction, phylogeny, glycine zipper motifs, comparative genomics.

© The Author(s) 2022. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

## Significance

Few freshwater species of Cyanobacteria have been known to mineralize amorphous  $\text{CaCO}_3$  (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). Although 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 intracellular amorphous calcium carbonates (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. 2021), the determinants of iACC formation in prokaryotes remain poorly understood.

The iACC-biomineralizing cyanobacteria are geographically widespread in freshwater, hot spring, or karstic terrestrial systems (Ragon et al. 2014) and sometimes locally abundant (Bradley et al. 2017). They received particular attention because they challenge the usual paradigm that cyanobacteria biomineralize  $\text{CaCO}_3$  extracellularly as an indirect byproduct of photosynthesis only (Altermann et al. 2006). Moreover, the geological history of iACC biomineralization remains mysterious because 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  $^{90}\text{Sr}$  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

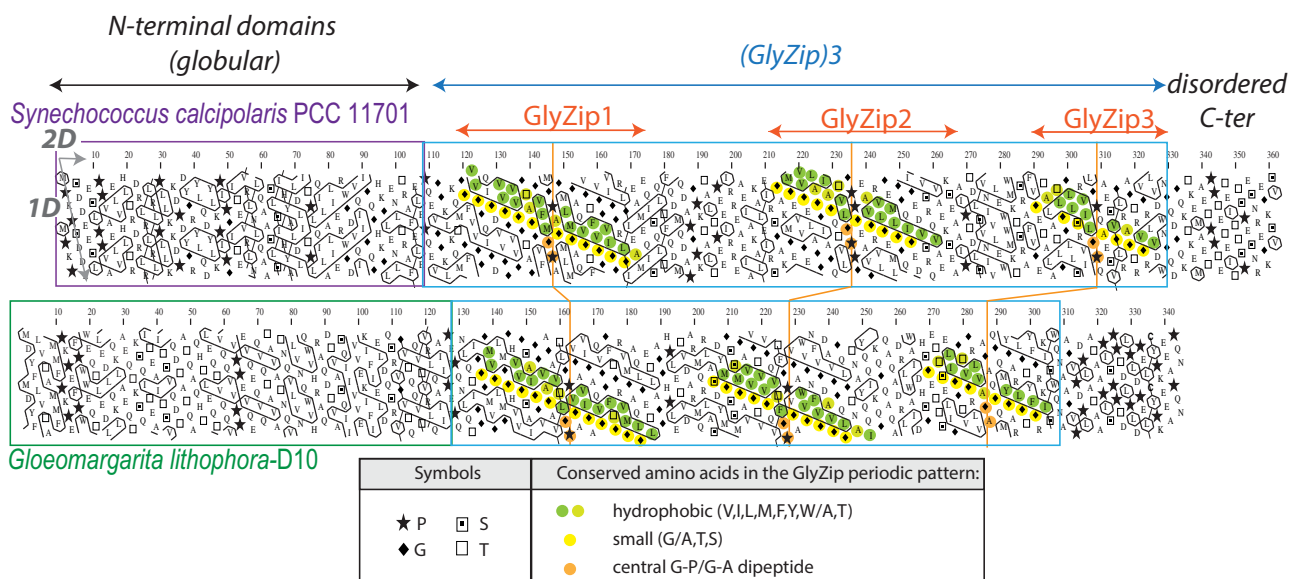
We applied comparative genomics to search for putative genes exclusively shared by iACC-forming cyanobacteria, and therefore absent in iACC-lacking species. We analyzed the genomes of 56 cyanobacterial strains (supplementary table 1, Supplementary Material online), in which the presence or absence of iACC was previously determined by electron microscopy (EM) (Benzerara et al. 2014). Fifty strains were lacking iACC and six were shown to form iACC: *Synechococcus* sp. PCC 6312, *Synechococcus calcipolaris* PCC 11701, *Thermosynechococcus elongatus* BP-1,

*Cyanothece* sp. PCC 7425, *Chroococciopsis 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 two 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. 1997; Bitard-Feildel et al. 2018). The two-dimensional HCA 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 (e.g., folded, disordered), 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. Although 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 and fig. 1, Supplementary Material online). 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 and supplementary fig. 2, Supplementary Material online), thereby increasing the number of known iACC-forming cyanobacterial species from 6 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, Supplementary Material online).

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 and supplementary fig. 2, Supplementary Material online). 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 and table 3, Supplementary Material online) and the nucleotide sequences of the genomic regions containing *ccyA* in these two strains (corresponding to contigs of 97,542 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, Supplementary Material online). 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. (2018) 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., 6 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



**FIG. 1.**—Domain architecture of calcyanins, as viewed by HCA. 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, and W) are contoured. The latter form clusters, which mainly correspond to the internal faces of regular secondary structures ( $\alpha$ -helices and  $\beta$ -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, and 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 ( $\sim 1/3$  strong hydrophobic amino acids) are boxed.

(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 *M. 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 (HGTs) (Frangeul et al. 2008; Humbert et al. 2013).

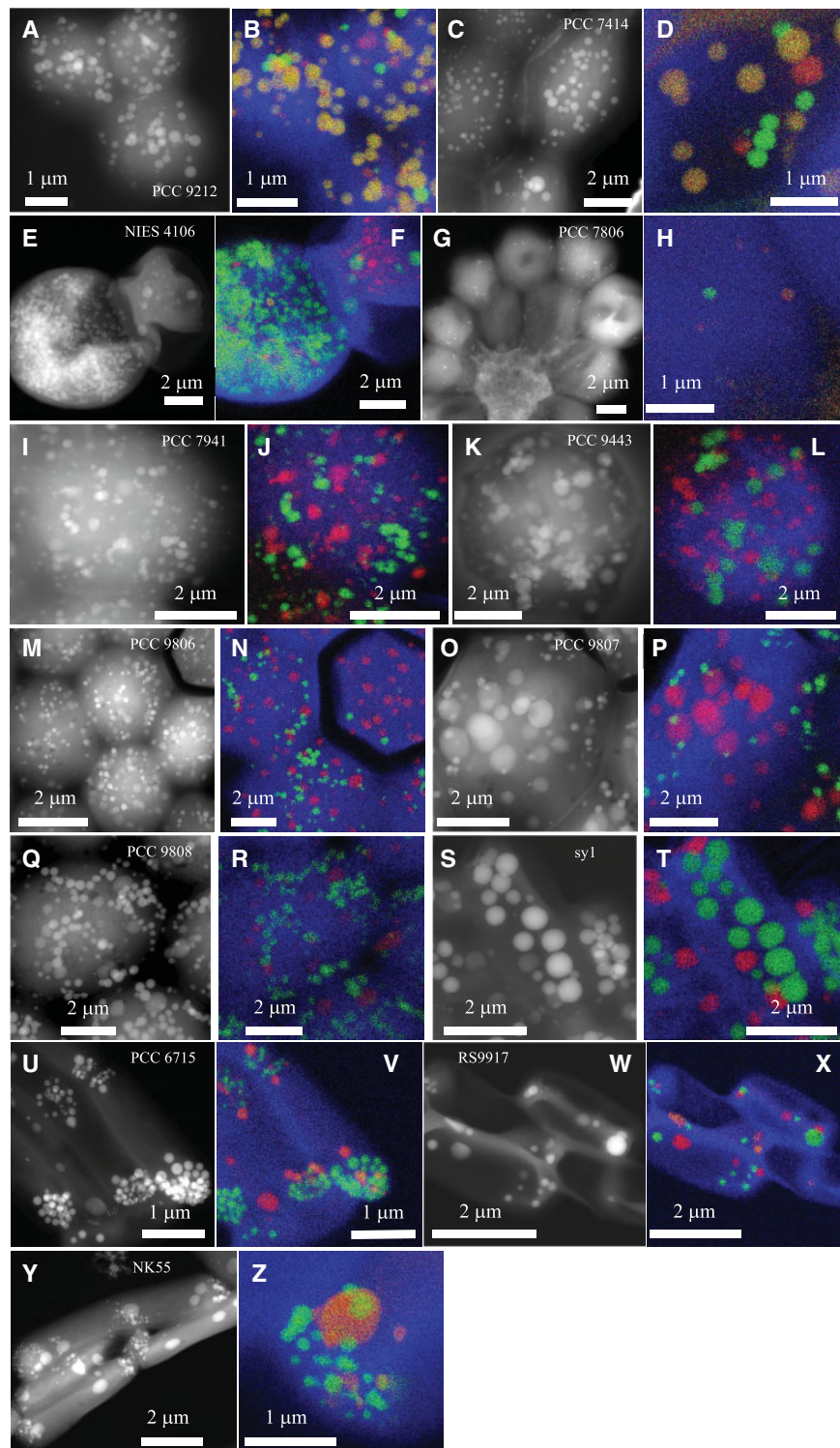
The *ccyA* gene and iACC biomineralization were also found in four *Synechococcus*-like strains previously not known to produce iACC (*Neosynechococcus sphagnicola* sy1, *Synechococcus* sp. RS9917, *Synechococcus lividus* PCC 6715, and *Thermosynechococcus* sp. NK55a). *Synechococcus* is a polyphyletic 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

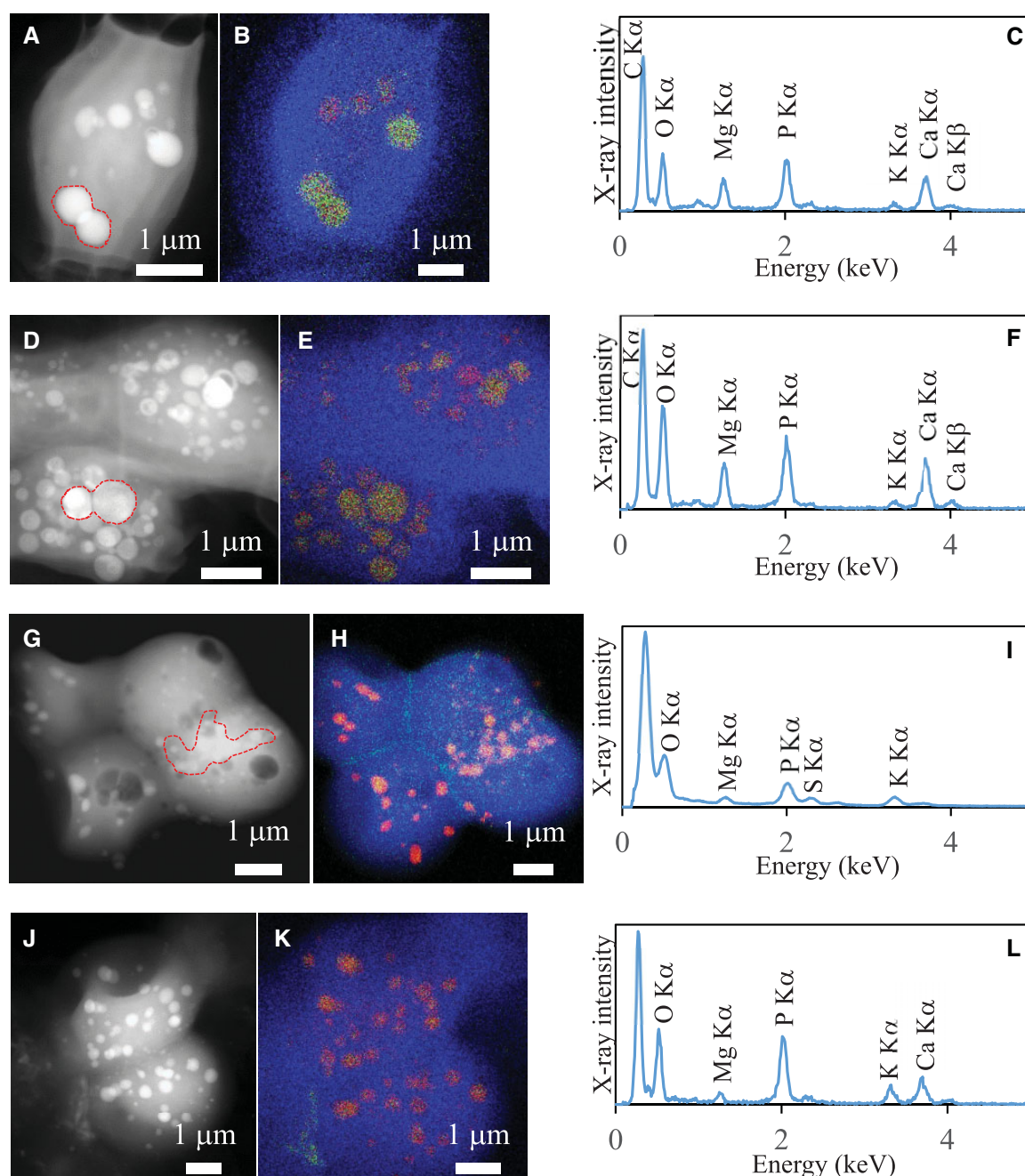
With the exception of the *Thermosynechococcus* sp. NK55a calcyanin, fused with a polypeptide containing a PIN-TRAM domain, the other 34 calcyanin family homologs contained 264–375 amino acids (average  $336 \pm 25$ ; supplementary table 2, Supplementary Material online). All showed the already mentioned 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 domains, 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





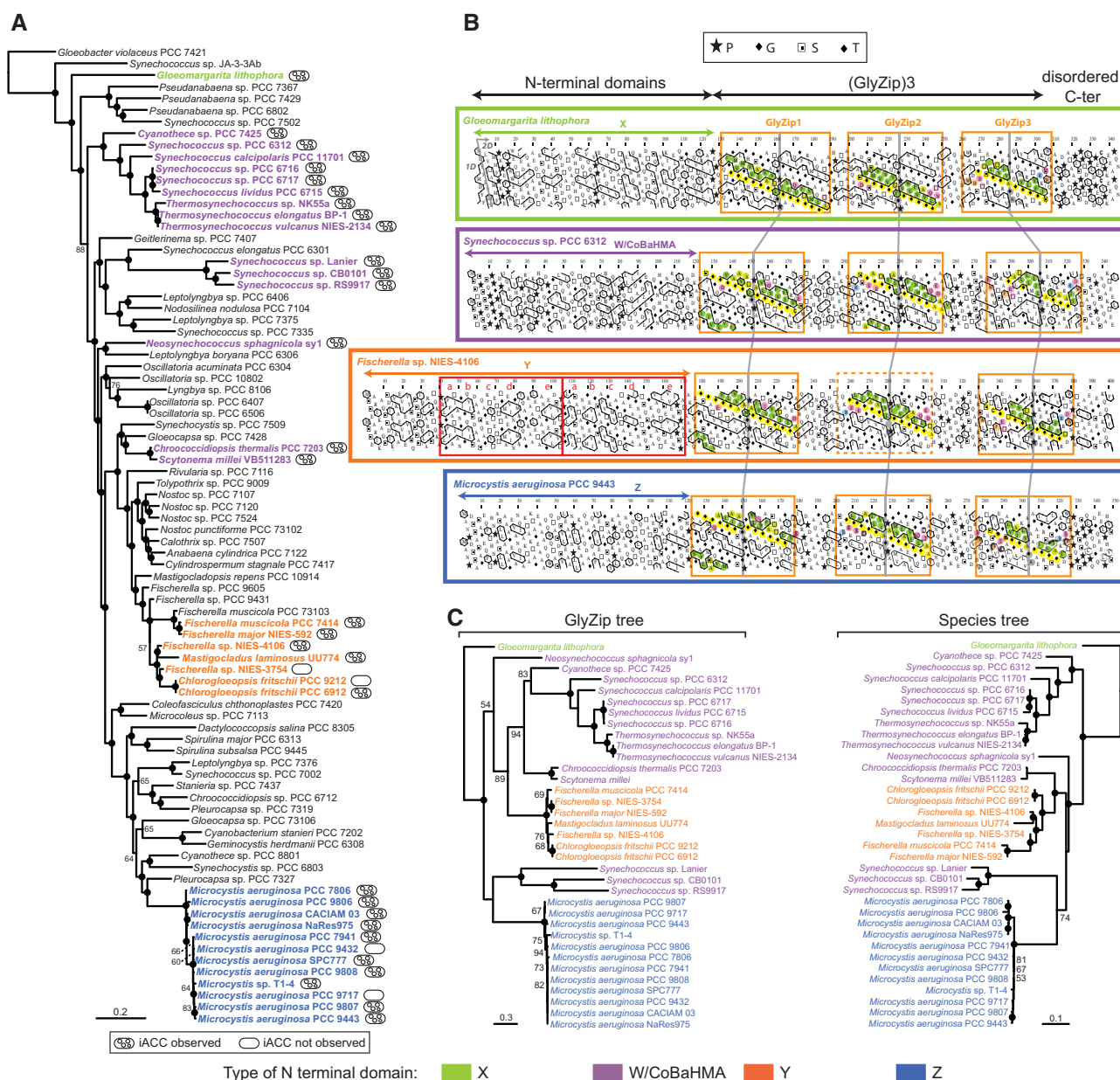
**Fig. 2.**—EM detection of iACC in 13 calycinin-bearing cyanobacterial strains not previously known to biomineralize carbonates. STEM-HAADF images of the 13 newly identified iACC-forming strains and overlays of C (blue), Ca (green), and P (red) chemical maps as obtained by EDXS. The name of the strains 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).



**Fig. 3.**—TEM analyses of the four *cyjA*-harboring 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.

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, Supplementary Material online](#)), which was predicted

to contain five regular secondary structures (labeled a–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

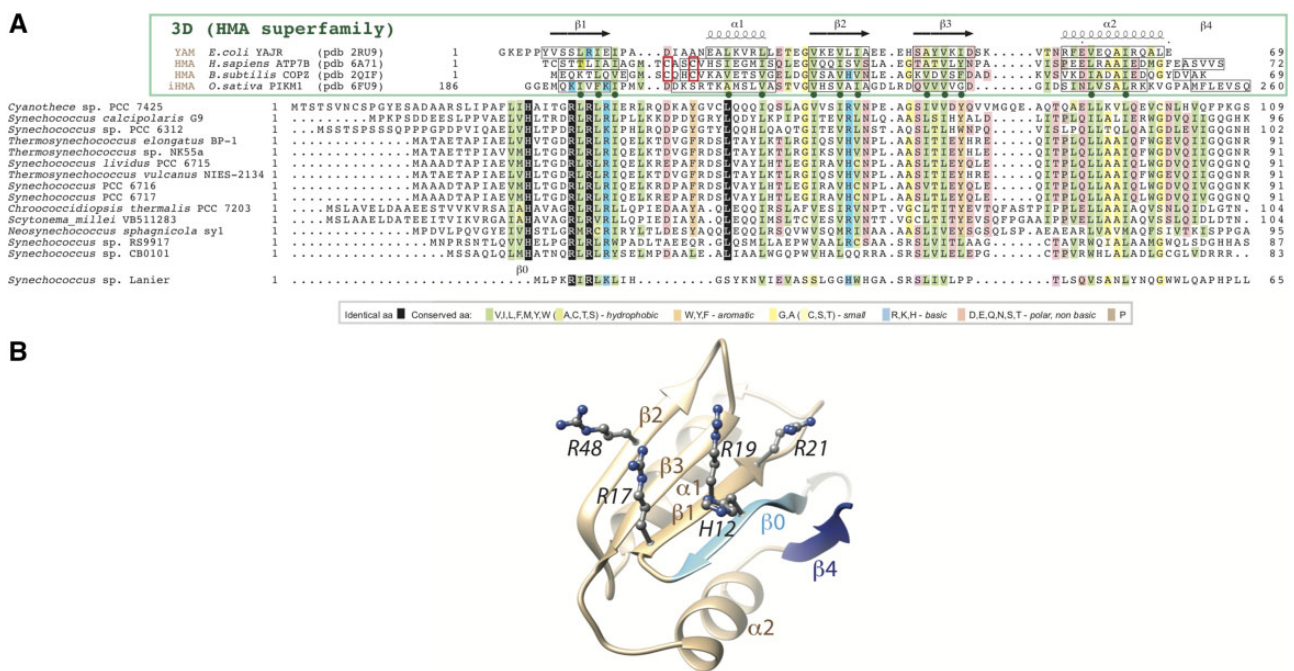


**Fig. 4.**—Phylogenetic analysis and domain architecture of the calcyanin protein family. (A) Maximum-likelihood phylogenetic tree of Cyanobacteria based on 58 conserved proteins; the strains containing the *cyA* gene are highlighted in bold and color. (B) HCA plots of representative calcyanin sequences (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–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).

domains, found in the cytosolic C-terminus of *Escherichia coli* Major Facilitator Superfamily transporter YajR (Jiang et al. 2013, 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). Similar to these three domains, the W-type N-terminal





**FIG. 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 in 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  $\beta$ -strand, named  $\beta_0$ , 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 highly conserved basic amino acids are shown with atomic details.

domain showed a repeated  $\beta - \alpha - \beta$  motif corresponding to a ferredoxin-like fold, characteristic of the HMA superfamily (fig. 5). However, although 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  $\beta_1$  and  $\beta_2$  and a histidine located upstream of strand  $\beta_1$ , in a region appearing as a calcyanin-specific extension of the HMA core (strand  $\beta_0$  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 (Webb and Sali 2016). The position of strand  $\beta_0$  was moreover putatively assigned with reference to the 3D structure of Kipl (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 C $\alpha$  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  $\beta_0$

relative to the  $\beta_1 - \beta_3$  core, but also led to propose a model for strand  $\beta_4$  as well as to refine the position of amino acids within strand  $\beta_0$  (fig. 5B). Although the calcyanin 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, Supplementary Material online). 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



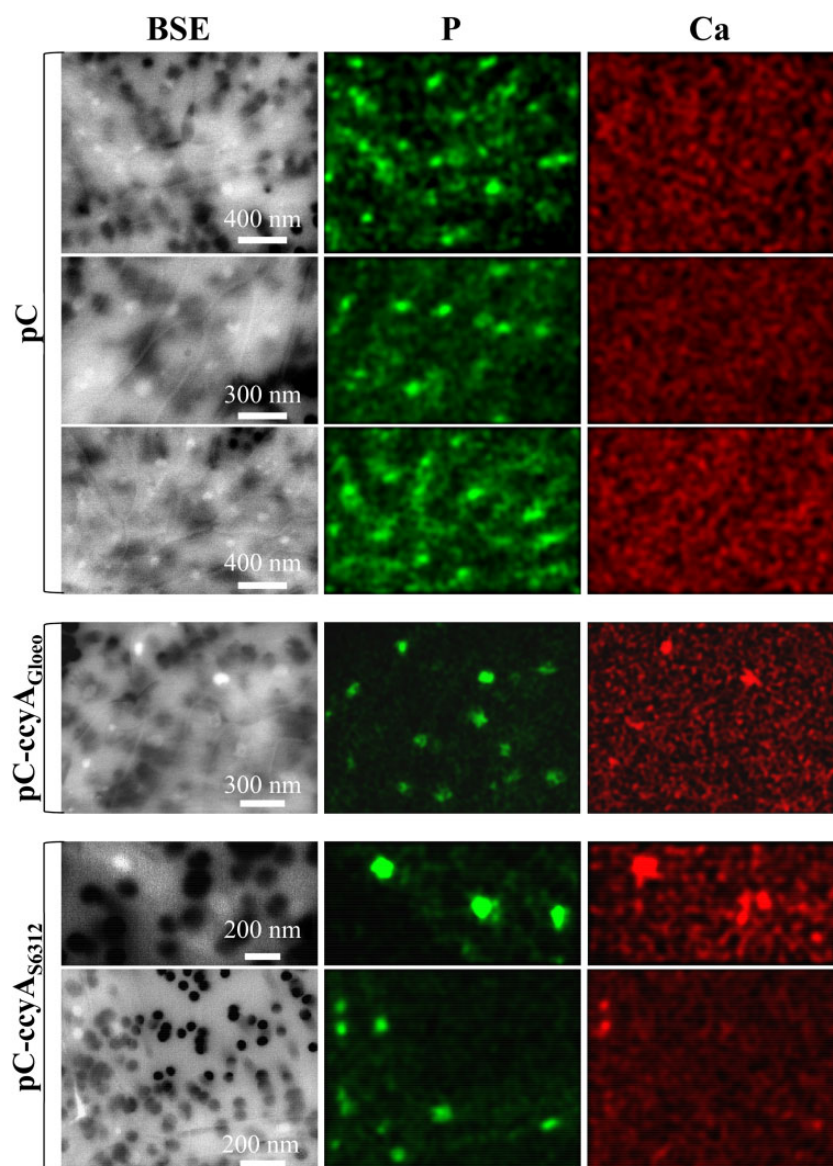
sequence similarity with known domains, suggesting that these repeated motifs form a novel architecture. The repeated presence of glycine and hydrophobic amino acids every four amino acid residues over a large sequence length, with an unusual persistence of this periodic motif across the different cyanobacterial 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, Supplementary Material online) 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, for example, *Shewanella oneidensis* (see NCBI Conserved Domain Database [CDD] annotations in supplementary table 2b, Supplementary Material online). The matching region, located before the OpmA-like C terminal domain, shows the typical features of a GlyZip unit (supplementary fig. 6, Supplementary Material online) 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, Supplementary Material online). 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, Supplementary Material online). This is particularly interesting because bicarbonate and calcium are obvious crucial ingredients for the synthesis of CaCO<sub>3</sub>. 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 data set 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* ( $\chi^2$  test,  $P$  value = 1.4e-08; supplementary table 5, Supplementary Material online). 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. Because 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, Supplementary Material online). Overall, the correlation and/or colocalization of *ccyA* and genes involved in Ca or HCO<sub>3</sub><sup>-</sup> 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, Supplementary Material online). 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*<sub>G10e0</sub> and pC-*ccyA*<sub>S6312</sub>) 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*<sub>G10e0</sub>) and 10 (pC-*ccyA*<sub>S6312</sub>) Ca hotspots were



**FIG. 6.**—SEM analyses of mutants overexpressing *ccyA*. SEM-EDXS images (in BSE mode), P (green), and Ca (red) maps of *Synechococcus 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  $\mu\text{m}$  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*<sub>Gloeo</sub> (fourth row) or pC-*ccyA*<sub>S6312</sub> (fifth and sixth rows), appearing as hotspots in Ca maps. See [supplementary data 3, Supplementary Material online](#) for details concerning the plasmid and strains.

detected over a total of 117 and 90 polyphosphate inclusions observed in the pC-*ccyA*<sub>Gloeo</sub> and pC-*ccyA*<sub>S6312</sub> mutants, respectively (fig. 6). The Ca detection limit of scanning electron microscopy (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, Supplementary Material online). The four calcanin 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 HGT events. To better characterize these evolutionary processes, we reconstructed the phylogeny of calcanin 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 calcanin, 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 data set ( $P$  value = 0.64), the GlyZip topology was strongly rejected by the data set 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 calcanin (fig. 4B). Because of its larger phylogenetic distribution, the CoBaHMA-type seemed to be the most ancient calcanin version, whereas 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 calcanin 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 nonhomologous domains generating the different calcanins by their fusion to the conserved GlyZip C-terminal domain.

To investigate if calcanin 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 nonredundant database. We found homologs with a complete C-terminal domain in only five noncyanobacterial species: an uncultured candidate phyla radiation *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, Supplementary Material online). 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, consistently, 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 noncyanobacterial 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 calcanin 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 calcanin that we have identified here, the domain family that we named CoBaHMA is found in the most widespread, and likely most ancient, calcanin 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  $\text{CaCO}_3$  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, Supplementary Material online) were retrieved from the NCBI database. The 523,680 translated CDSs derived from these genomes were processed using OrthoMCL with default settings (Li et al. 2003). This analysis included an all-versus-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 data set corresponded to the NCBI genome assemblies assigned to Cyanobacteria, published online before December 1, 2017 (except the six 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, Supplementary Material online): 1) translated CDS or 2) proteins in RefSeq annotation records, 3) translated CDS or 4) proteins in GenBank annotation records.

A multiple sequence alignment of the conserved C-terminal domain was built for the six 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 toward glycine-rich unrelated proteins, we artificially reduced glycine weight by 20% in the profiles. The profile versus 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 data set. The multiple sequence alignment and the HMM profile of this data set 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, Supplementary Material online, 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)<sub>2</sub> (instead of (Gly)<sub>3</sub>) to be matched. At the end of the whole process, we used the final HMM profile (as provided in supplementary data 3, Supplementary Material online) to search for similarities 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 and 97,528 bp, respectively) were compared using BLASTn.

### Search for Homologs of the $\text{Ca}(2+)/\text{H}(+)$ Antiporter and $\text{Na}(+)$ -Dependent Bicarbonate Transporter in Cyanobacterial Genomes

Homologs of the  $\text{Ca}(2+)/\text{H}(+)$  antiporter and the  $\text{Na}(+)$ -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 data set 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 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.

Three-dimensional structure modeling 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) (<https://colab.research.google.com/github/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. Three-dimensional 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, Supplementary Material online). 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 1,000 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 1,000 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 10,000 -au -zw (Nguyen et al. 2015). The sequence evolution models used were, as before, LG+C60+F+G for the conserved protein data set and LG+C20+F+G for the GlyZip data set.

### 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 high angle annular dark field (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 HAADF mode. These EDXS analyses provide hyperspectral data, that is, 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  $8,000 \times g$  for 10 min. Pellets were rinsed three times in Milli-Q (mQ) water (Millipore). After the final centrifugation, pellets were suspended in 200  $\mu$ L of mQ water. A drop of 5  $\mu$ L 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 SEM. Even more effort was invested in the analysis of strains harboring *ccyA* but not showing iACC. Indeed, although 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 seven 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 3 and 9 days interval. For *C. fritschii* PCC 6912, we performed eight different SEM or STEM sessions over five different cultures,

including three on the same culture with a 4 and 6 days interval and two on another culture with a 15 days interval. For *M. aeruginosa* PCC 9432, we performed seven different SEM or STEM sessions over four different cultures, including two on the same culture with a 3 days interval and three on another culture with 3 and 8 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 3 days interval and two on another culture with 25 days interval.

Mutant strains of *S. elongatus* PCC 7942 harboring the pC, pC-ccyA<sub>Gloeo</sub>, or ccyA<sub>S6312</sub> were analyzed by SEM in the back-scattered electron (BSE) 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 1 $\sigma$ . One example of EDXS spectrum is provided per type of mutant in [supplementary figure 8](#), [Supplementary Material online](#).

### Genetics

The pC-ccyA<sub>Gloeo</sub> and pC-ccyA<sub>S6312</sub> plasmids were derivatives of the RSF1010-derived pC vector (Veaudor et al. 2018) replicating in *E. coli* ([supplementary table 8 and fig. 10](#), [Supplementary Material online](#)). Chenebault et al. (2020) showed that this expression plasmid allowed strong gene expression in cyanobacteria. The pC-ccyA<sub>Gloeo</sub> and pC-ccyA<sub>S6312</sub> plasmids were transferred to *S. 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<sub>Gloeo</sub>, or pC-ccyA<sub>S6312</sub>, were washed twice and resuspended in LB medium ( $1 \times 10^9$  cells.mL<sup>-1</sup>). 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<sub>2</sub>CO<sub>3</sub>) were centrifuged and concentrated five times (about  $1 \times 10^8$  cells/mL) in fresh MM. Then, 100  $\mu$ L of *S. elongatus* PCC 7942 cells were mixed with 30  $\mu$ L of CM404 cells and 30  $\mu$ L of TOP10 cells harboring either pC, pC-ccyA<sub>Gloeo</sub>, or pC-ccyA<sub>S6312</sub>. A total of 30  $\mu$ L 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 (2,500 lux, i.e., 31  $\mu$ E.m<sup>-2</sup>.s<sup>-1</sup>) conditions. Then, cells were collected from each plate and resuspended into 50  $\mu$ L of liquid MM, prior to plating onto MM containing 5  $\mu$ g.mL<sup>-1</sup> of each the streptomycin (Sm) and spectinomycin (Sp) selective antibiotics. After about 10 days of incubation under standard light and temperature conditions, Sm<sup>R</sup>Sp<sup>R</sup>-resistant conjugant clones were collected and restreaked onto selective plates, prior to analyzing their plasmid content by PCR and DNA sequencing (Eurofins Genomics) using specific primers

([supplementary data 1c and 1d and fig. 10](#), [Supplementary Material online](#)).

### Supplementary Material

[Supplementary data](#) are available at *Genome Biology and Evolution* online.

### 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 Poinset for helping in the preparation of some samples for transmission electron microscopy analyses. We thank Eva Jahodarova for shipping a culture of *Neosynechococcus 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 Görden PhD grant was funded by the Sorbonne Université doctoral program Interfaces pour le Vivant.

### Author Contributions

K.B.E., E.D.U., C.C.C., F.C.H., D.M.O., P.L.G., and I.C.A. conceived and designed the work. K.B.E., E.D.U., T.B.F., G.C.A., C.C.C., M.D.E., I.D.I., G.G.A., M.G.U., S.G.O., F.S.P., D.M.O., and I.C.A. acquired, analyzed, and/or interpreted data. K.B.E., E.D.U., C.C.C., F.C.H., M.G.U., P.L.G., D.M.O., and I.C.A. drafted the work or substantively revised it.

### Data Availability

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.



## Literature Cited

- 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(1):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(3):147–166.
- Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389–3402.
- Benzerara K, et al. 2014. Intracellular Ca-carbonate biomineralization is widespread in cyanobacteria. *Proc Natl Acad Sci U S A.* 111(30):10933–10938.
- Bitard-Feidel T, Lamiable A, Mornon J-P, Callebaut I. 2018. Order in disorder as observed by the “hydrophobic cluster analysis” of protein sequences. *Proteomics* 18(21–22):e1800054.
- Blondeau M, Benzerara K, et al. 2018. Impact of the cyanobacterium *Gloeomargarita lithophora* on the geochemical cycles of Sr and Ba. *Chem Geol.* 483:88–97.
- Blondeau M, Sachse M, et al. 2018. Amorphous calcium carbonate granules form within an intracellular compartment in calcifying cyanobacteria. *Front Microbiol.* 9:1768.
- Blue CR, et al. 2017. Chemical and physical controls on the transformation of amorphous calcium carbonate into crystalline CaCO<sub>3</sub> polymorphs. *Geochim Cosmochim Acta.* 196:179–196.
- Bradley JA, 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 Microbiomes.* 3:32.
- Bull PC, Cox DW. 1994. Wilson disease and Menkes disease: new handles on heavy-metal transport. *Trends Genet.* 10(7):246–252.
- Callebaut I, et al. 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci.* 53(8):621–645.
- Cam N, et al. 2016. Selective uptake of alkaline earth metals by cyanobacteria forming intracellular carbonates. *Environ Sci Technol.* 50(21):11654–11662.
- Cam N, et al. 2018. Cyanobacterial formation of intracellular Ca-carbonates in undersaturated solutions. *Geobiology* 16(1):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(15):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, et al. 2018. Polymorphic residues in rice NLRs expand binding and response to effectors of the blast pathogen. *Nat Plants.* 4(8):576–585.
- De Wever A, et al. 2019. Evidence of high Ca uptake by cyanobacteria forming intracellular CaCO<sub>3</sub> and impact on their growth. *Geobiology* 17(6):676–690.
- Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol.* 7(10):e1002195.
- Finney AR, Innocenti Malini R, Freeman CL, Harding JH. 2020. Amino acid and oligopeptide effects on calcium carbonate solutions. *Cryst Growth Des.* 20(5):3077–3092.
- Frangeul L, et al. 2008. Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics.* 9(274):274.
- Gillespie JJ, et al. 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun.* 79(11):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(2):221–224.
- Humbert J-F, et al. 2013. A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium *Microcystis aeruginosa*. *PLoS One.* 8(8):e70747.
- Hyatt D, et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.* 11:119.
- Jiang D, 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(36):14664–14669.
- Jiang D, et al. 2014. Atomic resolution structure of the *E. coli* YajR transporter YAM domain. *Biochem Biophys Res Commun.* 450(2):929–935.
- Jumper J, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583–589.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.
- Kim S, et al. 2005. Transmembrane glycine zippers: physiological and pathological roles in membrane proteins. *Proc Natl Acad Sci U S A.* 102(40):14278–14283.
- Komarek J, Johansen JR, Smarda J, Strunecky O. 2020. Phylogeny and taxonomy of *Synechococcus*-like cyanobacteria. *Fottea* 20(2):171–191.
- Kuehlbrandt W. 2019. Structure and mechanisms of F-type ATP synthases. *Annu Rev Biochem.* 88: 515–549.
- Lamiable A, et al. 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). *J Phycol.* 43(3):417–425.
- Lefevre CT, Bazylinski DA. 2013. Ecology, diversity, and evolution of magnetotactic bacteria. *Microbiol Mol Biol Rev.* 77(3):497–526.
- Leonov H, Arkin IT. 2005. A periodicity analysis of transmembrane helices. *Bioinformatics* 21(11):2604–2610.
- Li J, et al. 2016. Biomineralization patterns of intracellular carbonatogenesis in cyanobacteria: molecular hypotheses. *Minerals* 6(1):10.
- Li L, Stoekert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13(9):2178–2189.
- Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* 32(Web Server Issue):W327–W331.
- Marron AO, et al. 2016. The evolution of silicon transport in eukaryotes. *Mol Biol Evol.* 33(12):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(21):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(3):145–148.
- Mirdita M, et al. 2021. ColabFold—making protein folding accessible to all. *bioRxiv.* doi:10.1101/2021.08.15.456425.
- Monteil CL, et al. 2021. Intracellular amorphous Ca-carbonate and magnetite biomineralization by a magnetotactic bacterium affiliated to the Alphaproteobacteria. *ISME J.* 15(1):1–18.
- Moreira D, et al. 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(3):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(1):268–274.
- Pettersen EF, et al. 2004. UCSF chimera—a visualization system for exploratory research and analysis. *J Comput Chem.* 25(13):1605–1612.

- Ragon M, Benzerara K, Moreira D, Tavera R, López-García 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.* 5:331.
- 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(1):17–23.
- Riding R. 2012. A hard life for cyanobacteria. *Science* 336(6080):427–428.
- Robert X, Gouet P. 2014. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* 42(Web Server Issue):W320–W324.
- 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(4):465–479.
- Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol.* 51(3):492–508.
- Veaudor T, et al. 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(6):e0198836.
- Wang X, et al. 2021. The evolution of calcification in reef-building corals. *Mol Biol Evol.* 38:3543–3555.
- 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. *Rev Mineral Biochem.* 54: 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(1):e90.
- Yarra T, Blaxter M, Clark MS. 2021. A Bivalve Biomineralization Toolbox. *Mol Biol Evol.* 38(9):4043–4055.
- Zimmermann L, et al. 2018. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J Mol Biol.* 430(15):2237–2243.

**Associate editor:** Tal Dagan