PNAS

Supporting Information for

Chitin utilization by marine picocyanobacteria and the evolution of a planktonic lifestyle

Giovanna Capovilla†*[1], Rogier Braakman†*[2], Gregory Fournier[2], Thomas Hackl[1,3], Julia Schwartzman[1,4], Xinda Lu[1,5], Alexis Yelton[1,6], Krista Longnecker[7], Melissa Kido Soule[7], Elaina Thomas[1,8], Gretchen Swarr[7], Alessandro Mongera[9,10], Jack Payette[2], Kurt G. Castro[1], Jacob Waldbauer[11], Elizabeth B. Kujawinski[7], Otto X. Cordero[1], Sallie W. Chisholm*[1,12]

[1]Department of Civil & Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[2] Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[3] Groningen Institute of Evolutionary Life Sciences, University of Groningen, Groningen, Netherlands
[4] present address: Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA
[5] present address: Foundation Medicine, Cambridge, MA, United States
[6] present address: Genvid Holdings Inc., Seattle, WA, United States
[7] Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA, USA
[8] present address: School of Oceanography, University of Washington, Seattle, WA, USA
[9] Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA
[10] Department of Genetics, Harvard Medical School, Boston, MA, USA
[11] Department of the Geophysical Sciences, University of Chicago, Chicago, IL, USA
[12] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

† Equal contributions

* Corresponding authors
Giovanna Capovilla (giocapo@mit.edu)
Rogier Braakman (braakman@mit.edu)
Sallie W. Chisholm (chisholm@mit.edu)

This PDF file includes:

Figures S1 to S8
Tables S1 to S2
Legends for Datasets S1 to S5
Supporting text including Figures S9 to S12
SI References

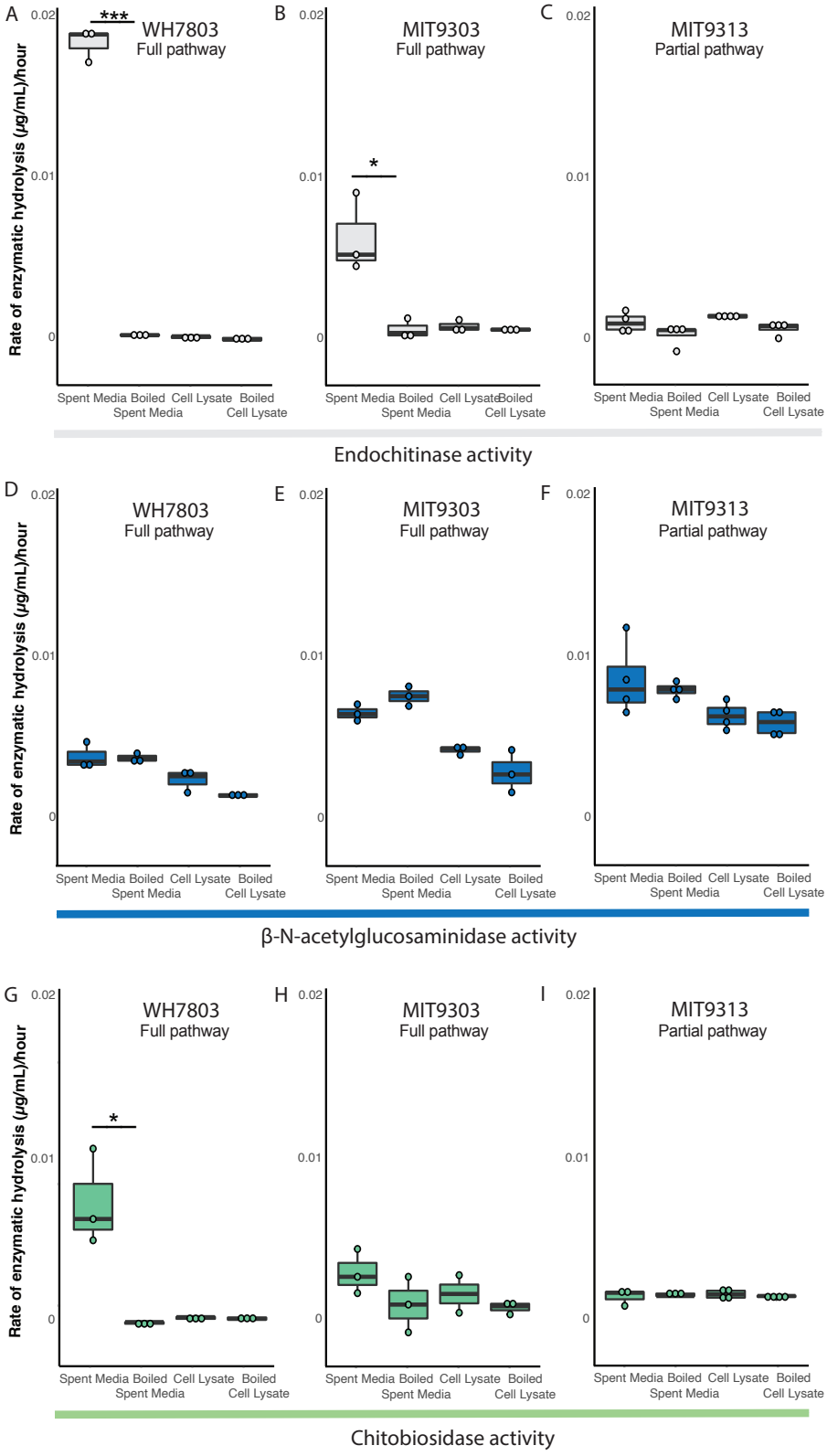

A
B
C
D
E
F
G
H
I
WH7803
Full pathway
MIT9303
Full pathway
MIT9313
Partial pathway
Rate of enzymatic hydrolysis (µg/mL)/hour
0.02
0.01
0
***
*
Spent Media
Boiled Spent Media
Cell Lysate
Boiled Cell Lysate
Endochitinase activity
β-N-acetylglucosaminidase activity
Chitobiosidase activity

FIGURE S1. Endochitinase and Exochitinase activities in marine picocyanobacteria. Chitinase activity in *Synechococcus* WH7803, *Prochlorococcus* MIT9303 and MIT9313 were measured in the spent media and cell lysates of cultures growing in Pro99 (Sea water media) amended with chitosan to a final concentration of 56 µg/ml. The degradation of chitin oligomers attached to the fluorophore methylumbelliferyl is estimated by the luminescence intensity. Three activities are shown. A-C) The endochitinase activity (ability to cleave glycosidic bonds at internal sites of the chitin polymer) shown in grey, D-F) β-N-acetylglucosaminidase activity (ability to cleave GlcNAc monomers from the nonreducing terminal of the chitin polymer) shown in blue G-I) Chitobiosidase activity (ability to cleave dimeric units of GlcNAc from the nonreducing terminal of the chitin polymer) shown in green. Statistical significance has been performed between each sample and its control with a Welch t-test. Data in (a), (b) and (c) are also shown in Figure 2.

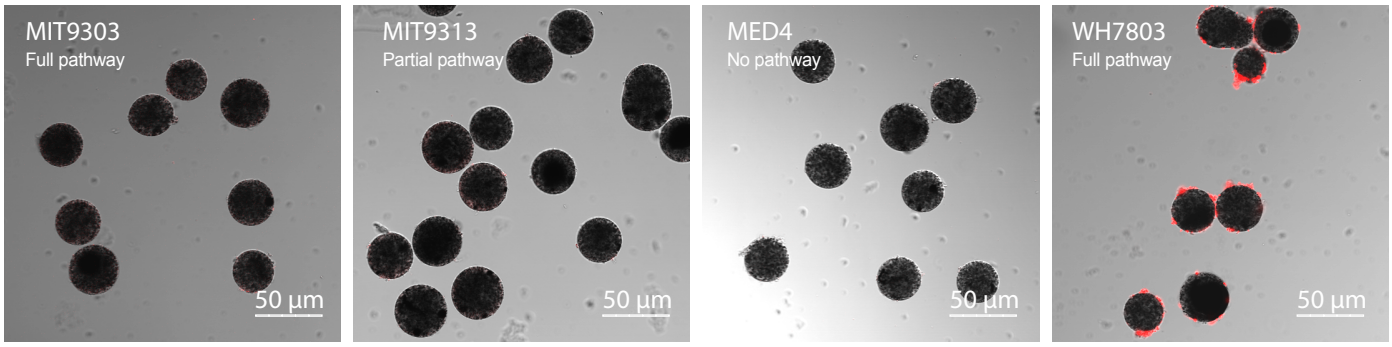


FIGURE S2. Colonization of agarose particles. Images of magnetic agarose beads analyzed in Figure 3 to illustrate consistency and uniformity across beads. Beads are shown in bright field mode. *Synechococcus* and *Prochlorococcus* are detected by their autofluorescence and highlighted in red. White bars are 50 µm long.

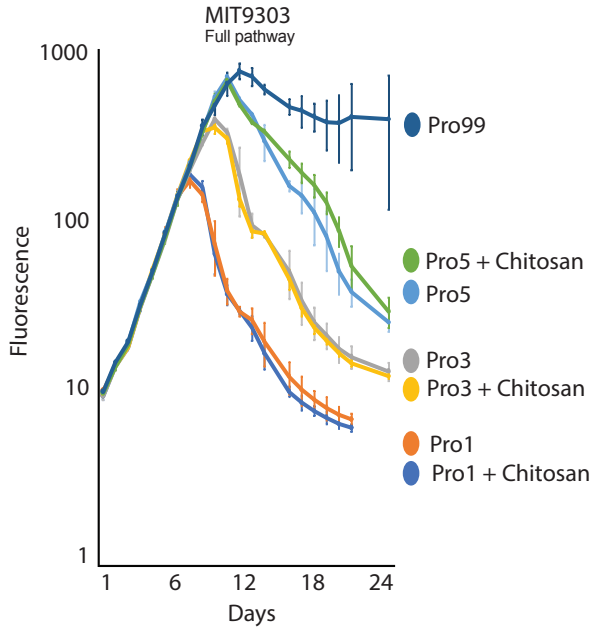


FIGURE S3. Response of *Prochlorococcus* to chitosan under nitrogen stress. Growth of *Prochlorococcus* MIT9303 in continuous light (at 12 µmol photons m^{-2} s^{-1}) in a variety of different media, monitored by relative bulk culture chlorophyll fluorescence. A. Different colors show the growth patterns of triplicate cultures grown in Pro99 media with different concentrations of NH_4^+, the sole N-source, with and without added chitosan (56 µg/ml). The concentrations of NH_4^+were 800 µM in Pro99, 250 µM in Pro5, 150 µM in Pro3, and 50 µM in Pro1.

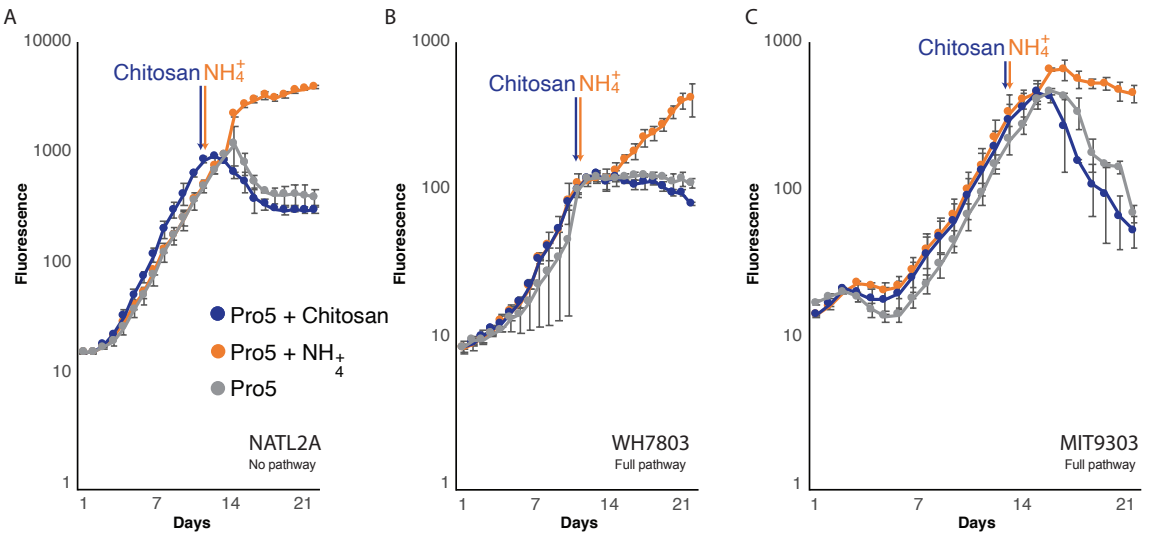


FIGURE S4. Comparison of the response to chitosan and ammonia under nitrogen stress. Growth of *Prochlorococcus* NATL2A, *Synechococcus* WH7803, and *Prochlorococcus* MIT9303 in N-depleted media (denoted as "Pro5" to indicate the lowered nitrogen-to-phosphorus ratio of N:P = 5) amended with NH_4^+ or chitosan when cells reached N-starved stationary phase. Different colors show the growth patterns in media Pro5 that has been amended with NH_4^+ (orange) at the time of the arrow, to a final concentration of 800 µM (N:P =16), or chitosan (blue) to a final concentration of 56 µg/ml, or non-amended (grey). Cultures were grown in continuous light (at 15 µmol photons m^{-2} s^{-1}) and growth was monitored by relative bulk culture chlorophyll fluorescence.

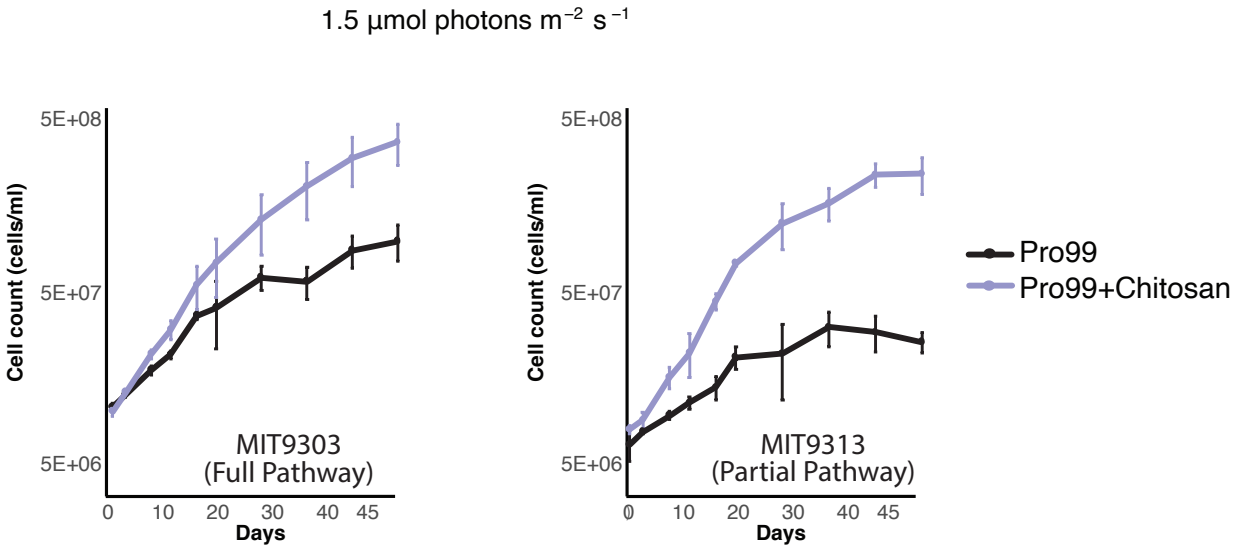


FIGURE S5. Effect of chitosan addition on growth of two Prochlorococcus strains in growth-limiting light level. Cultures of MIT9303 (primary chitin degrader) and MIT9313 (secondary degrader) were grown in Pro99 media with (purple) and without (black) added chitosan, at 1.5 µmol photons m^−2 s^−1. Growth was monitored by bulk chlorophyll fluorescence and reported in Fig.4. Cell counts measured using flow cytometry are shown here for each curve.

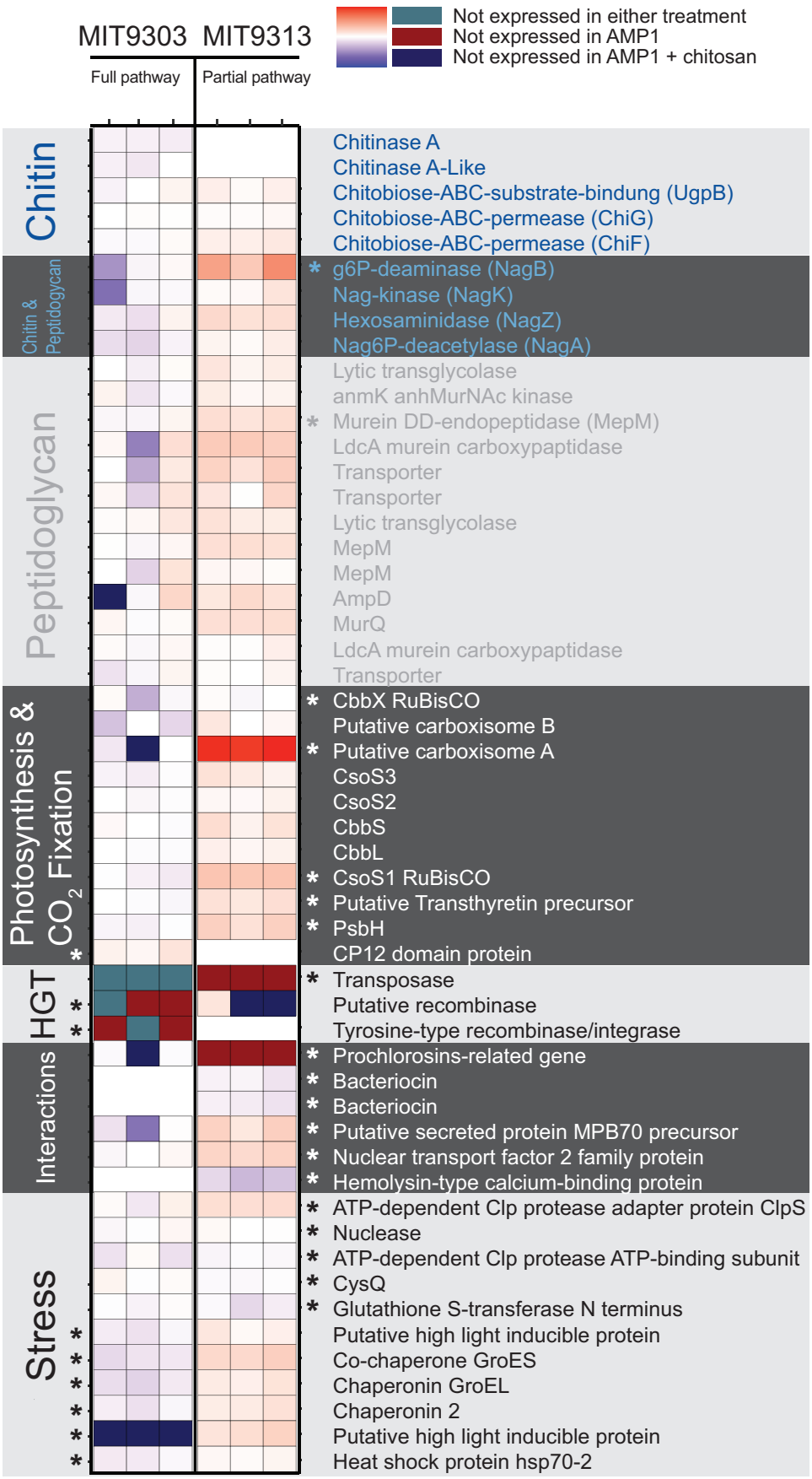

MIT9303
MIT9313
Full pathway
Partial pathway
Not expressed in either treatment
Not expressed in AMP1
Not expressed in AMP1 + chitosan
Chitin
Chitin & Peptidogycan
Peptidoglycan
Photosynthesis & CO_2 Fixation
HGT
Interactions
Stress
Chitinase A
Chitinase A-Like
Chitobiose-ABC-substrate-bindung (UgpB)
Chitobiose-ABC-permease (ChiG)
Chitobiose-ABC-permease (ChiF)
* g6P-deaminase (NagB)
Nag-kinase (NagK)
Hexosaminidase (NagZ)
Nag6P-deacetylase (NagA)
Lytic transglycolase
anmK anhMurNAc kinase
* Murein DD-endopeptidase (MepM)
LdcA murein carboxypaptidase
Transporter
Transporter
Lytic transglycolase
MepM
MepM
AmpD
MurQ
LdcA murein carboxypaptidase
Transporter
* CbbX RuBisCO
Putative carboxisome B
* Putative carboxisome A
CsoS3
CsoS2
CbbS
CbbL
* CsoS1 RuBisCO
* Putative Transthyretin precursor
* PsbH
CP12 domain protein
* Transposase
Putative recombinase
Tyrosine-type recombinase/integrase
* Prochlorosins-related gene
* Bacteriocin
* Bacteriocin
* Putative secreted protein MPB70 precursor
* Nuclear transport factor 2 family protein
* Hemolysin-type calcium-binding protein
* ATP-dependent Clp protease adapter protein ClpS
* Nuclease
* ATP-dependent Clp protease ATP-binding subunit
* CysQ
* Glutathione S-transferase N terminus
Putative high light inducible protein
Co-chaperone GroES
Chaperonin GroEL
Chaperonin 2
Putative high light inducible protein
Heat shock protein hsp70-2

FIGURE S6. Quantitative analysis of gene expression in *Prochlorococcus* in response to chitosan addition. Heat map representing the relative expression of genes in the two *Prochlorococcus* strains 24 hours after addition to chitosan, reported as a qualitative representation in figure 5B. Three biological replicates are shown for each strain. Transcriptional enrichment between the samples amended with chitosan and those that were not is indicated in shades of red, while transcriptional depletion is represented in shades of blue. Those genes which were statistically significantly differentially expressed are marked with a black or white star. All genes in the chitin degradation and peptidoglycan recycling pathways, as well as all genes of the CO_2-fixation, are shown. Genes are grouped by functional categories and represented in shades of gray.

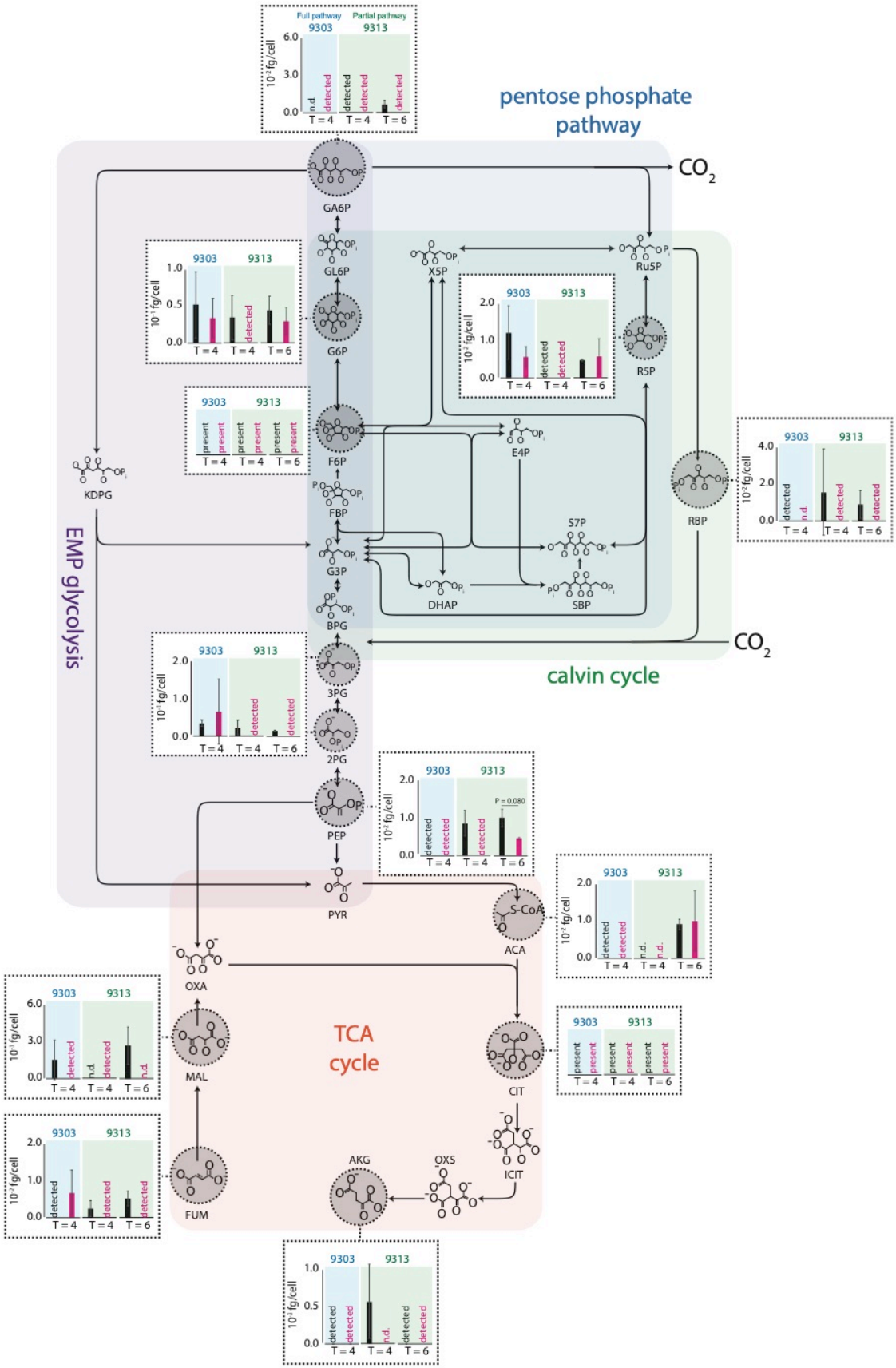
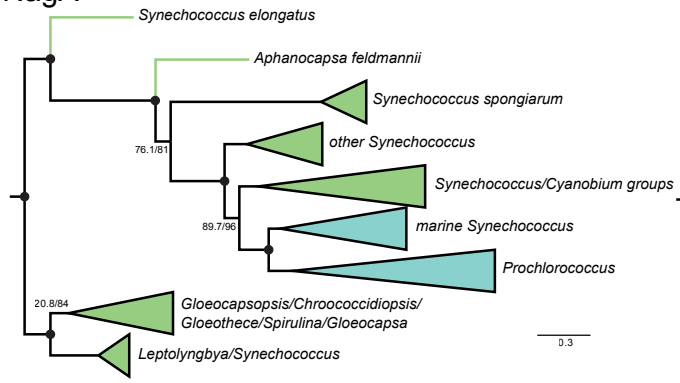
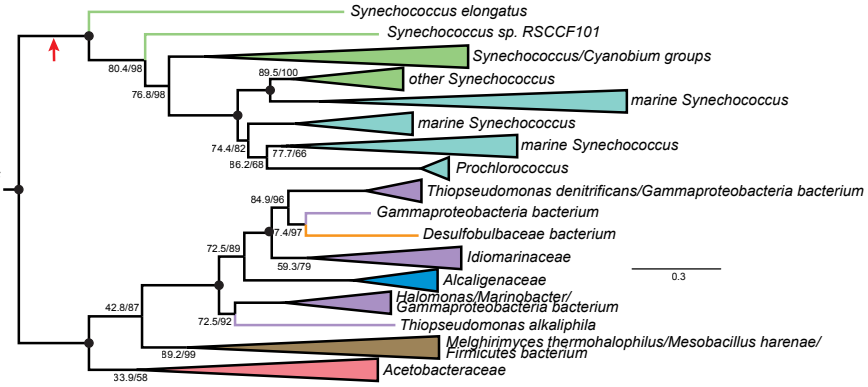

pentose phosphate pathway
EMP glycolysis
calvin cycle
TCA cycle
CO2
CO2
GA6P
GL6P
G6P
F6P
FBP
G3P
BPG
3PG
2PG
PEP
PYR
KDPG
X5P
Ru5P
R5P
E4P
S7P
DHAP
SBP
RBP
ACA
CIT
ICIT
AKG
OXS
OXA
MAL
FUM
Full pathway
Partial pathway
9303
9313
T = 4
T = 6
detected
n.d.
present
P = 0.080

FIGURE S7. Metabolomic analysis of core carbohydrate pathways in response to chitosan addition in *Prochlorococcus*. Concentrations, in femtograms per cell, of intermediates of core carbohydrate metabolic pathways occurring downstream of the chitin degradation pathway shown in Figure 5C, on days 4 and 6 after chitosan additions (pink) compared to unamended controls (black). Major pathways within the metabolic core are highlighted in different colors. Error bars show the standard deviation of 3 biological replicates of MIT9303 and MIT9313 on day 4, and of 2 replicates of MIT9313 on day 6. Samples in which metabolites were observed in only a single replicate are denoted 'detected', while samples, where metabolites were not observed in any replicates, are labeled 'n.d.' (not detected). Fructose-6P (F6P) and citrate (CIT) were present in most samples, but their levels could not be quantified due to interference by the organic carbon matrix. Abbreviations: GA6P - Gluconate-6-phosphate, GL6P - Gluconolactone-6-phosphate, G6P - Glucose-6-phosphate, F6P - Fructose-6-phosphate, FBP - Fructose bisphosphate, G3P - Glyceraldehyde-3-phosphate, DHAP - Dihydroxyacetone-phosphate, SBP - Sedoheptulose bisphosphate, S7P - Sedoheptulose-7-phosphate, E4P - Erythrose-4-phosphate, X5P - Xylulose-5-phosphate, R5P - Ribose-5-phosphate, Ru5P - Ribulose-5-phosphate, RBP - Ribulose bisphosphate, KDGP - 2-Keto-3-deoxy-6-phosphogluconate, BPG - bisphospho-glycerate, 3PG - 3-phosphoglycerate, 2PG - 2-phosphoglycerate, PEP - Phosphoenoylpyruvate, PYR - Pyruvate, ACA - Acetyl-CoA, OXA - Oxaloacetate, MAL - Malate, FUM - Fumarate, CIT - Citrate, ICIT - Isocitrate, OXS - Oxalosuccinate, AKG - alpha-ketoglutarate.

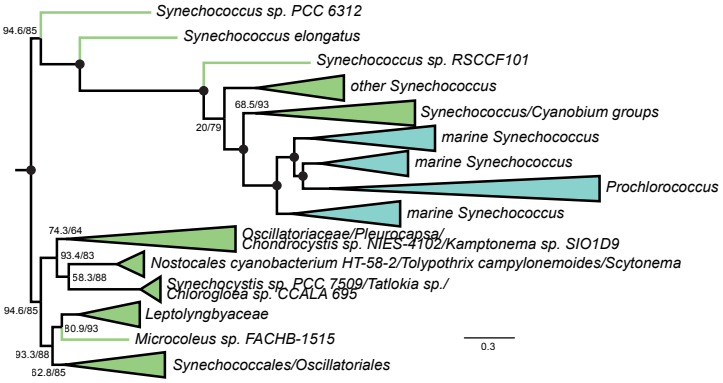
NagA



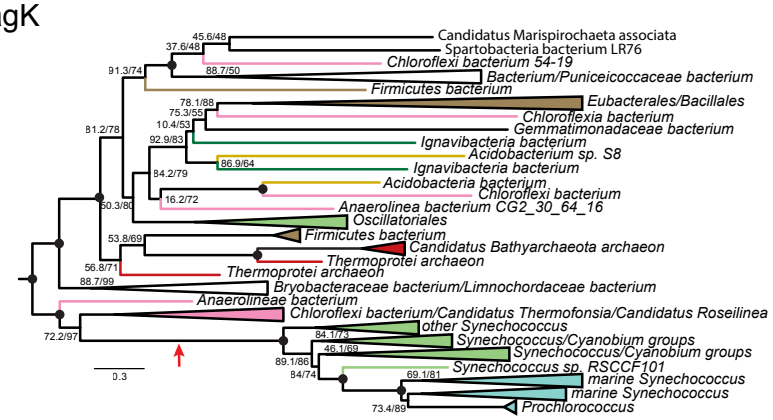
NagB



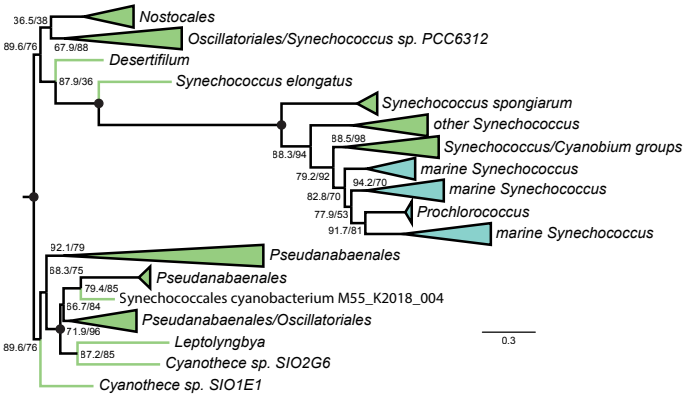
NagZ



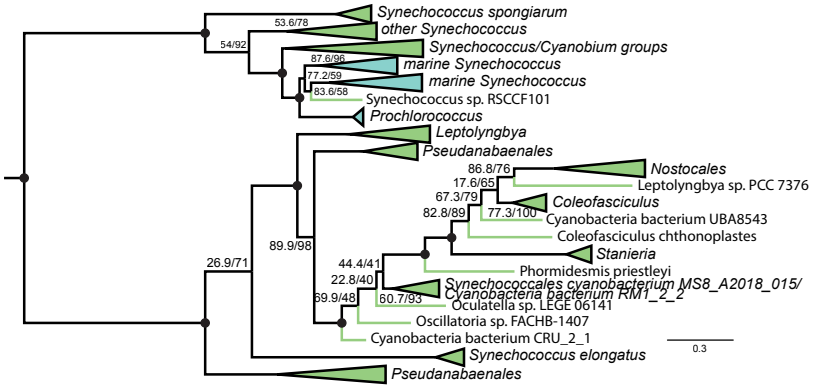
NagK



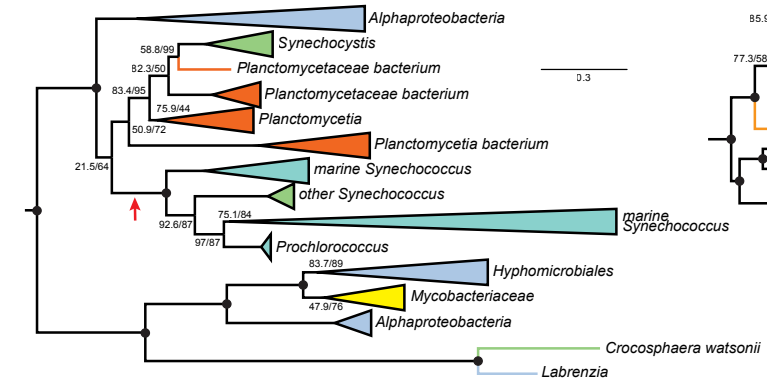
UgpA



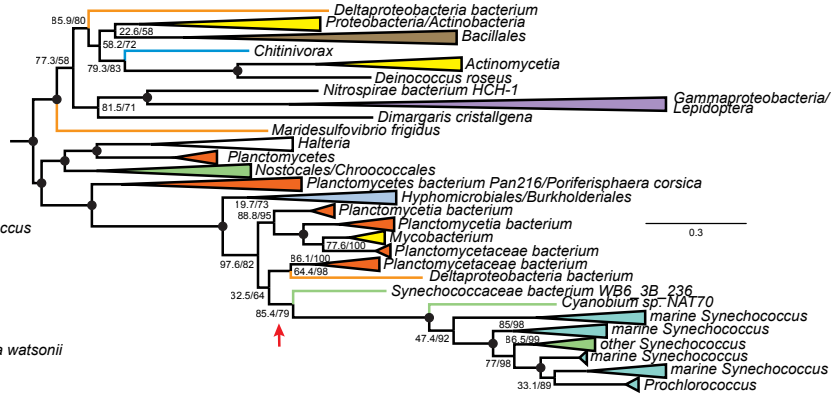
UgpE



ChiA domain N-terminal region



ChiA domain C-terminal region



- marine Synechococcus/Prochlrococcus
- other Cyanobacteria
- Planctomycetes
- Actinobacteria
- Firmicutes
- Chloroflexi
- Ignavibacteria
- Alphaproteobacteria
- Betaproteobacteria
- Gammaproteobacteria
- Deltaproteobacteria
- Acidobacteria
- Archaea
- other groups

FIGURE S8. Phylogenies of genes involved in chitin utilization in marine picocyanobacteria. Red arrows indicate inferred HGTs into SynPro ancestor lineages. Branch lengths are indicated by included scale bars (average substitutions/site). High-level taxonomic identities are indicated by color coding, with specific represented groups labeled for terminal and collapsed groups. In each case, trees are rooted at the branch closest to the midpoint that preserves the monophyly of Synechococcales groups including SynPro. Branch supports (approximate likelihood ratio test/bootstrap value) are included for bipartitions with less than 90% support for either metric, and black circles indicate >90% support. Phylogenies of the two ChiA domains and NagK are also shown in figure 6.

Table S1. List of primers used for the qPCR experiment

Gene	Forward primer	Reverse primer
chitinase A (ChiA)	GGTGCAAGACATCAATGTCG	GCAGCCCAAGAATCAAAAAG
chitinaseA-like (ChiA-like)	GTTCTCGTTGAGCAAAGCCAG	ACCCCACCAAAGATCACCAG
chitobiose-ABC-permease (UgpA)	CCAAAACAGGTCTCGACGTT	AACATCGGATCACCAGCAAG
chitobiose-ABC-permease (UgpE)	CCATTGCTTTGGCTGGTAAG	TCGGTTCGGCAGGTAGTAAG
chitobiose-ABC-substrate-binding (UgpB)	ACGTTGGCAGGATGTACCTT	GAATCGTCAGGGACCACAGT
g6P-deaminase (NagB)	GGCAGCTGATCGGCGCAGC	GAAGATGCAACTGCTGCGGG
hexosaminidase (NagZ)	CGACCCAGATCAGCCTTTAG	TTCCCGCGTAAGAAAAGTTG
nag6P-deacetylase (NagA)	GAGCGGGTGCTCAGTGTTC	AGCCCTCCATTGATTTGTAGA
nag-kinase (NagK)	GGAGCAAGCGGAATAGAACA	ATCCCCAGTTGCTAGGCATT
rnpB	TGAGGAGAGTGCCACAGAAA	ACCTCTCGATGCTGCTGGT

Table S2.

Complete set of metabolites measured using liquid chromatography-mass spectrometry (LC-MS). A novel method was developed to target this set of metabolites. The polarity (+ or -, for positive or negative) indicates the ionization mode used to analyze the metabolite. For select metabolites, concentrations could not be determined due to interference by the organic carbon matrix, these are marked as 'presence / absence' in the table.

compound	**KEGG**	**polarity**	**Detection in *Prochlorococcus***
2-keto-3-deoxy-6-phosphogluconic acid	C04442	-	not detected
6-phosphogluconic acid	C00345	-	yes
acetyl coenzyme A	C00024	-	yes
alpha-ketoglutaric acid	C00026	-	yes
chitobiose	C01674	+	yes
citric acid	C00158	-	presence / absence
3- and 2-phosphoglyceric acid	C00197 / C00631	-	yes, but cannot be analytically separated
D-glucosamine	C00329	+	yes
D-glucosamine 6-phosphate	C00352	-	yes
D-Ribose 5-phosphate	C00117	-	yes
D-ribulose 1,5-bisphosphate	C01182	-	yes
fructose 6-phosphate	C00085	-	presence / absence
fumaric acid	C00122	-	yes
glucose 6-phosphate	C00092	-	yes

malic acid	C00149	-	yes
N-Acetyl-D-glucosamine 6-phosphate	C00357	-	yes
phosphoenolpyruvate	C00074	-	yes

Dataset S1 (separate file). Gene annotations and gene frequency statistics

Dataset S2 (separate file). RNA-Seq data (also available at ENA #PRJEB53957)

Dataset S3 (separate file). Metabolites data (also available at Metabo-Lights #MTBLS5229)

Dataset S4 (separate file). Gene expression in MIT9313

Dataset S5 (separate file). Gene expression in MIT9303

Supporting Information Text

(also available at https://datadryad.org/stash/dataset/doi:10.5061/dryad.gxd2547pp)

Phylogenomic analysis of Chitinase Domains within SynPro proteins ChiA and ChiA-like

The majority of chitinase proteins homologous to ChiA and ChiA-like in *Prochlorococcus* genomes contain multiple conserved domains. These proteins are often annotated as "cellulose binding domain-containing protein" or "Chitinase A1 precursor". These proteins contain two cellulose binding (CBD II) domains (pfam00553) including the two key Trp residues involved in substrate binding, and a C-terminal glycosyl hydrolase family 18 (GH18) domain, also annotated as a type II chitinase domain (CDD database, cd06548/COG3325)(1).

Sequence retrieval and alignment
Using the *Prochlorococcus* MIT1303 WP_082824463.1 protein as a query, the top 500 BLAST hits were collected from genbank(2), including several homologs from other *Synechococcus/Prochlorococcus* (SynPro) genomes, and many other hits to protein sequences from other bacterial groups. These proteins were then aligned using the E-INS-i algorithm in MAFFT(3).

Linker region between inherited N-terminal and C-terminal chitinase regions for *Prochlorococcus* MIT1303:
AETLTINEVDLSMHHGTDHSDMDHSAIESDMPTGNGALVSNGPLSLEVSGSLYWGGMSGKLTLT
NSGNT
DLDGWSVSFVTPHTNFQSWAGDAQIESLADGTNRITLTPASWNQSIAIGQSIEVSFNAQSVGLPN
SGSLNSELFFADGQT
QMPSGGITVEADPMQPQEAETSSTATTTDFEPETGTNGDHNQNDMDHSAIESDMPTGDGALVS
NGPLSLEVSGSLYWGGM
SGKLTLTNSGNTDLDGWSVSFVTPHTNFQSWAGDAQIESLADGTNRITLTPASWNQSIAIGQSIE
VSFNAQSVGLPNSGS
LNSELFFADGQTQMPSGGITVEADPMQPQEAETSSTATTTDFEPQTGINDDAHPLEMSSTAIAD
GSKR

The resulting alignment showed a composite homology between SynPro proteins and other bacterial proteins, with the N-terminal region of the SynPro proteins being homologous to one group of bacterial proteins, and the C-terminal region being homologous to another group of bacterial proteins (Figure S9, SI file: ChiAChiAlikeFusion_BLAST500_EINSi_gappy.fasta). Furthermore, other bacterial proteins generally contained additional N-terminal and C-terminal protein regions absent in the SynPro homologs. These sequences also did not contain any sequences with regions that aligned well to the central region of the SynPro homologs containing the CBDII domains. These regions are variable, even within the SynPro homologs. **This alignment was consistent with the SynPro chitinases being the result of a fusion of at least two protein domains typically found in different protein sequences.**

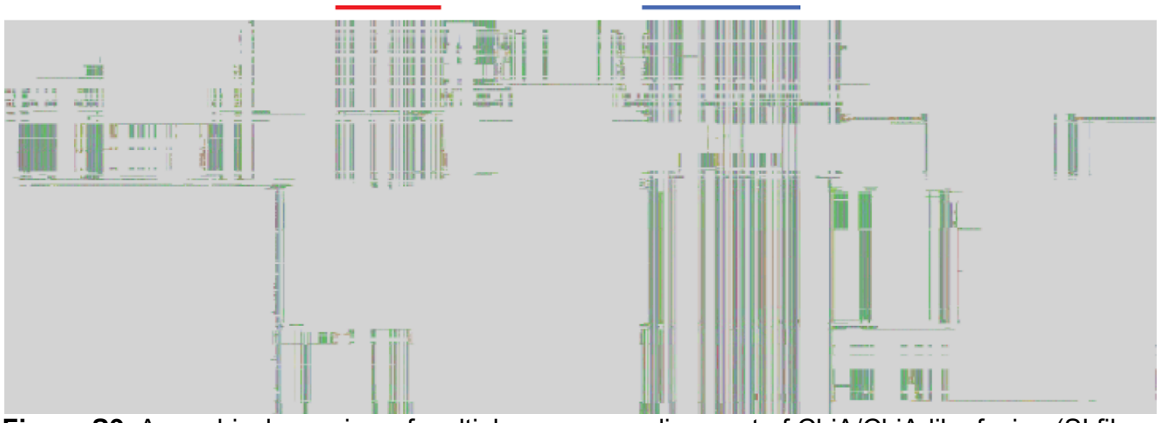


Figure S9. A graphical overview of multiple sequence alignment of ChiA/ChiA-like fusion (SI file: ChiAChiAlikeFusion_BLAST500_EINSi_gappy.fasta). SynPro sequences are in the top block (top rows); conserved N-terminal regions are in the block below the red line; conserved C-terminal regions are in the block below the blue line. Note that non-SynPro sequences containing one of the conserved regions almost never contain the second and SynPro sequences contain both N-terminal and C-terminal regions.

This observed alignment may have been impacted by BLAST searching using the full SynPro homolog, with top hits being driven by the most highly conserved regions of the query protein in one or another of the conserved domains. Therefore, an additional BLAST search was performed, this time independently using each of the three regions of the SynPro homolog: (1) conserved N-terminal region; (2) linker region containing CBDII domains; and (3) conserved C-terminal region containing the GH18 domain. If the chitinase protein within SynPro groups was the result of a gene fusion between different protein families, and potentially different Horizontal Gene Transfer (HGT) donors, the top hits outside of SynPro in each case may be different groups of proteins, within different bacterial taxa.

Region 1 query sequence:
LGGQTYAVNASGADITGFDPSRDRLDFGDISVHGLILGKLVDDTAVLVNPWQDSDYQRIL
DHNGNGISWNQLTLENFAPVGNEHLREDIGGVMSWELGIGPREADTVYIRSHEYGVHERVENFD
PQTQKLNFLYLGTRER
LSLTDTDEGLLISVDPSSQSLLLVGVKRTDLYAGNLEFHFDQVMEDNLEEPFGVAEDAVSLVSRE
LLLTPQSIGGATTDG
YQVRSG

Region 2 query sequence:
QLVQAAETLTINEVDLSMHHGTDHSDMDHSAIESDMPTGNGALVSNGPLSLEVSGSLYWGGMS
GKLTLTNSGNT
DLDGWSVSFVTPHTNFQSWAGDAQIESLADGTNRITLTPASWNQSIAIGQSIEVSFNAQSVGLPN
SGSLNSELFFADGQT
QMPSGGITVEADPMQPQEAETSSTATTTDFEPETGTNGDHNQNDMDHSAIESDMPTGDGALVS
NGPLSLEVSGSLYWGGM
SGKLTLTNSGNTDLDGWSVSFVTPHTNFQSWAGDAQIESLADGTNRITLTPASWNQSIAIGQSIE
VSFNAQSVGLPNSGS
LNSELFFADGQTQMPSGGITVEADPMQPQEAETSSTATTTDFEPQTG

Region 3 query sequence:
KRIVGYFEEWGIYS
RDFLVQDINVEDLTHINYSFFDVKANGDVNLFDSWAATDKRYSAEEQVSRTFSADEWAALDDSR
RSSYTSGSEFTTRTNG
NGSVSVSGVPVGWDVNGELAGNLRQFALLKQLNPDINLGLALGGWTLSDEFSLAFDDVAGRER
FTDNVISTLETYDFFNT
VDFDWEYPGGGGLSGNASSDQDGANFAATLKVLRQKMDLLETRTGEDYEISIATAGGQEKLANL
NLPAIDAYVDFYNVMT
YDFHGGWESVTGHQAAMTADAGGYDVVTAIQQFRNAGIAPEKVVLGAPTYTRAWGGVDSGEK
LGYGELGSANSAPGSYEA
GNYDQKDLVTGINNGSYDLAWDDDAKAAYLYNDQEQIWSSIETPSTIAGKAAYVDAAELGGMMF
WALSSDSSGEQSLIGA
ASDLL

For each region (1,2,3), the top 250 non-SynPro BLAST hits were retrieved from Genbank with E-values lower than 10^-5, with partial sequences removed (SI files: ChiA_Region1_BLAST_250_noSynPro.fasta, ChiA_Region2_BLAST_250_noSynPro.fasta, ChiA_Region3_BLAST_250_noSynPro.fasta). Region 3 hits initially had a large overrepresentation of sequences from *Paenibacillus*, *Brevibacillus*, and *Bacillus* genera; these genera were subsequently excluded and the search re-run, with single representatives from each genus subsequently added.

Interestingly, Region 2 had far fewer significant hits than Regions 1 or 3, including a large number of hits to proteins annotated as cellulose-binding proteins, but at below the E-value cutoff threshold.

Sequences retrieved from these three regional BLAST searches were added to the SynPro chitinase homolog sequences (SI file: SynProChiAChiAlike_1_2_3_EINSi_gappy.fasta). Duplicate sequences were removed. The 491 remaining sequences were then aligned in MAFFT(3) using two alignment strategies (G-INSi, leave gappy regions, ~0.7; E-INS-I, leave gappy regions, default). G-INSi with a high 0.7 unalign level produced an extremely long alignment with numerous long indels specific to only a few closely related sequences each; E-INSi leaving gappy regions, aligned much more of the sequences, and introduced fewer indels (Figure S10, SI file: SynProChiAChiAlike_1_2_3_EINSi_gappy_noDup.fasta). The E-INSi+gappy alignment was therefore used for subsequent analyses.

With this new alignment approach, the same pattern is clear; SynPro chitinase proteins contain an N-terminal region found in one set of bacterial homologs, and a C-terminal region found in another set of bacterial homologs.

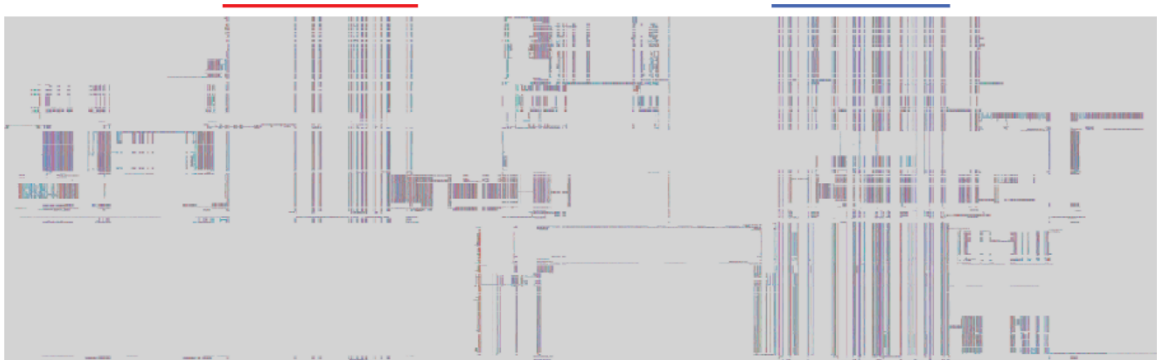


Figure S10. A graphical overview of multiple sequence alignment of chitinase homologs (SI file: SynProChiAChiAlike_1_2_3_EINSi_gappy_noDup.fasta). SynPro sequences are in the top block (top rows); conserved N-terminal regions are in the block below the red line; conserved C-

terminal regions are in the block below the blue line. Note that non-SynPro sequences containing one of the conserved regions almost never contain the second and SynPro sequences contain both N-terminal and C-terminal regions.

In order to test the fusion hypothesis, phylogenetic analyses were performed for the conserved N-terminal and C-terminal regions of the SynPro proteins, including the homologous regions of non-SynPro homologs. If these regions have disparate evolutionary histories outside of SynPro, but congruent evolutionary histories as part of the same protein within SynPro, this is strong evidence of a gene fusion event in the SynPro stem lineage.

Based on the alignment in Figure S10, the following sites were extracted for phylogenetic analysis:

N-terminal: 1745-3039 (218 sequences, 205 aligned sites)
C-terminal: 6077-7399 (313 sequences, 365 aligned sites)

Phylogenies were generated using IQTree(4), with 1000 rapid bootstraps and approximate likelihood ratio tests (SH-aLRT support %/Ultrafast bootstrap support %) used to evaluate bipartition supports.

IQTree determined via ModelFinder that the best-fitting evolutionary model for the N-terminal region was an LG substitution model with empirical frequencies and 5 distribution-free rate parameters: LG+F+R5, weighted BIC = 0.9898. The best-fitting model for the C-terminal region was LG+F+R6, with a weighted BIC = 0.9968. These generated treefiles, alignments, and IQtree logfiles are available in the SI.

N-terminal SI files:
SynProChiAChiAlike_EINSi_gappy_1745-3039_LG_F_R5_IQtree.tree
SynProChiAChiAlike123_EINSi_gappy_1745-3039_conserved.fasta
SynProChiAChiAlike123_EINSi_gappy_1745-3039_conserved.fasta.iqtree

C-terminal SI files:
SynProChiAChiAlike_EINSi_gappy_6077-7399_LG_F_R6_IQtree.tree
SynProChiAChiAlike123_EINSi_gappy_6077-7399_conserved.fasta
SynProChiAChiAlike123_EINSi_gappy_6077-7399_conserved.fasta.iqtree

Phylogenetic analysis
Both N-terminal and C-terminal region phylogenies (Figures S11 and S12, respectively) recovered a monophyletic group of SynPro sequences, most closely related to a group of sequences from marine and freshwater Planctomycetes metagenome sequences (see SI_Text1_Table1_PlanctomycetesStrainInfo.xlsx). These sequences from Planctomycetes were some of the few other proteins outside of SynPro to contain both of these N-terminal and C-terminal regions. However, these “fusion” chitinase-like proteins are sparsely distributed within the observed Planctomycetes sequences, and, for both N-terminal and C-terminal sequence regions, are not found as the closest relatives to the SynPro sequences. This suggests that these proteins have undergone multiple independent fusion events within Planctomycetes, including in the presumptive HGT donor to SynPro.

Support values
Support for the placement of SynPro as sister to the Planctomycetes sequences is relatively weak within the N-terminal sequence phylogeny (21.5/64). However, support for adjacent bipartitons constrains the alternative topology to group SynPro within Planctomycetes. The specific placement of SynPro within the C-terminal sequences is similarly weakly supported (32.5/64), but the bipartition grouping SynPro sequences within the broader diversity of Planctomycetes sequences is highly supported (97.6/82).

Additional HGTs
Another clear HGT from within Planctomycetes to a group of cyanobacteria is also found for sequences containing the N-terminal region, to a group of Synechocystis, including a member of Synechococcus that may represent a secondary transfer. This HGT is very strongly supported (58.8/99). The C-terminal region phylogeny recovered an additional HGT to cyanobacteria (92.5/100), to a group of Nostocales from Planctomycetes, including an additional likely secondary transfer to *Cyanobacterium aponium.* A group of Rhizobiales also acquired proteins containing the C-terminal region from Planctomycetes, and subsequently transferred the gene to a group of marine algae-associated Betaproteobacteria (*Alcaligenaceae* bacterium, *Algicoccus*, *Rhodocyclaceae* bacterium).

Proteins containing the C-terminal region show a set of striking HGTs to eukaryotes in more distant parts of the tree. These include HGTs from within Gammaproteobacteria to Ditrysia (a group of butterflies and moths) (97/91), HGT from within Planctomycetes to the planktonic ciliate *Halteria grandinella* (97.6/100), and a deep HGT to the fungal genus *Dimargaris* (81.5/71). This dataset did not recover closely related sequences to the fungal homologs, likely due to the sparse sampling at this phylogenetic depth. However, the HGT to *Halteria* further establishes Planctomycetes as a vector for the broad taxonomic distribution of these genes within marine environments.

Genomic Context
In both cases, the likely HGT donors to SynPro were members of Planctomycetes represented by metagenomic sequence data. Were the putative fusion protein constituents present in the same donor metagenomes? The sampling of Planctomycete metagenomes is too sparse to answer this question. However, it is telling that only one of the 36 genome/metagenomes returning proteins with significant hits to either the N-terminal or C-terminal fusion protein regions contains two distinct protein sequences homologous to both (metagenome DNLZ01000009.1 containing proteins HBB72892.1 and HBB72895.1). In other words, Planctomycetes genomes tend to contain proteins with homology to the N-terminal region of the fusion protein, or the C-terminal region of the fusion protein, but not both, except in the single case above, or in the case of the three detected fusion proteins within Planctomycetes (NBS32980.1, NDH93368.1, OUV71928.1). This suggests that these proteins may be frequently transferred between Planctomycetes genomes, but that they are rarely both retained within the same lineage. This makes it all the more notable that SynPro inherited a fusion gene product. This could have been the result of both genes being transferred into the same Planctomycetes genome and fusing to a single gene before being transferred to the SynPro stem lineage, or independent transfers into the SynPro stem lineage, with gene fusion subsequent to transfer.

Ecology context
The Planctomycete genomes and metagenomes sampled are found in both fresh and marine environments, without any clear indication of the environment of the HGT donor to SynPro. Therefore, from these phylogenies alone we cannot determine if the HGT(s) to SynPro occurred in a freshwater environment before a later adaptation to a planktonic environment, or within the planktonic environment. Other HGTs from within Planctomycetes to other groups of microbes include both marine and freshwater recipients, suggesting the lack of a strong congruence between ecological and phylogenetic signals for these genes.

Figure S11. Rooted phylogeny of the conserved N-terminal region identified in SynPro and other bacterial homologs (sites 1745-3039). Blue=SynPro; Red=Planctomycetes; Magenta=Gammaproteobacteria; Green=non-SynPro cyanobacteria; Dark Purple=Actinobacteria; Orange=Alphaproteobacteria. Tree is midpoint rooted. Detailed phylogeny files with sequence names and bootstrap supports are available in newick format (SI files: ChiA_ChiA_like_1745-3039_IQtree_Outfiles).

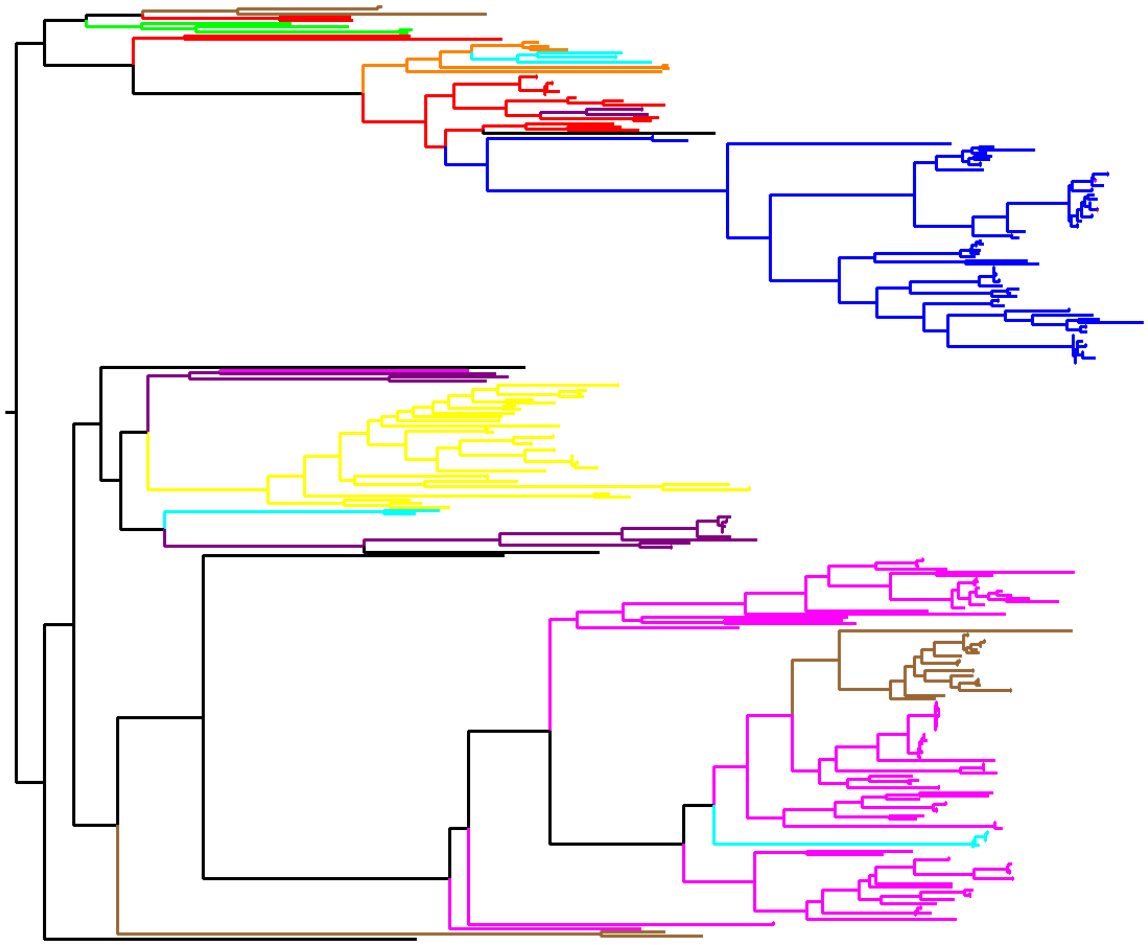


Figure S12. Rooted phylogeny of the conserved C-terminal region identified in SynPro and other bacterial homologs (sites 6077-7399). Blue=SynPro; Red=Planctomycetes; Magenta=Gammaproteobacteria; Green=non-SynPro cyanobacteria; Dark Purple=Actinobacteria, Orange=Alphaproteobacteria; Brown=Eukarya; Cyan=Betaproteobacteria; Yellow=Firmicutes, Black=other Bacteria. Tree is midpoint rooted. Detailed phylogeny files with sequence names and bootstrap supports are available in newick format (SI files: ChiA_ChiA_like_6077-7399_IQtree_Outfiles).

SI References

1. S. Lu, *et al.*, CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265–D268 (2020).

2. D. A. Benson, *et al.*, GenBank. *Nucleic Acids Res.* **46**, D41–D47 (2018).

3. K. Katoh, K.-I. Kuma, H. Toh, T. Miyata, MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).

4. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2014).