

Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome

Alexis Dufresne*, Marcel Salanoubat†, Frédéric Partensky**, François Artiguenave†, Ilka M. Axmann[§], Valérie Barbe†, Simone Duprat†, Michael Y. Galperin[¶], Eugene V. Koonin[¶], Florence Le Gall*, Kira S. Makarova[¶], Martin Ostrowski^{||}, Sophie Oztas†, Catherine Robert†, Igor B. Rogozin[¶], David J. Scanlan^{||}, Nicole Tandeau de Marsac**, Jean Weissenbach†, Patrick Wincker†, Yuri I. Wolf[¶], and Wolfgang R. Hess^{§††}

*Station Biologique, Unité Mixte de Recherche 7127, Centre National de la Recherche Scientifique et Université Paris 6, BP74, 29682 Roscoff Cedex, France; †Genoscope et Unité Mixte de Recherche 8030, Centre National de la Recherche Scientifique, CP 5706, 91057 Evry Cedex, France; [§]Department of Biology, Humboldt University of Berlin, Chausseestrasse 117, 10115 Berlin, Germany; [¶]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; ^{||}Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, United Kingdom; and **Unité des Cyanobactéries, Unité de Recherche Associée 2172, Centre National de la Recherche Scientifique, Institut Pasteur, 28 Rue Dr. Roux, 75724 Paris Cedex 15, France

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved July 1, 2003 (received for review May 28, 2003)

Prochlorococcus marinus, the dominant photosynthetic organism in the ocean, is found in two main ecological forms: high-light-adapted genotypes in the upper part of the water column and low-light-adapted genotypes at the bottom of the illuminated layer. *P. marinus* SS120, the complete genome sequence reported here, is an extremely low-light-adapted form. The genome of *P. marinus* SS120 is composed of a single circular chromosome of 1,751,080 bp with an average G+C content of 36.4%. It contains 1,884 predicted protein-coding genes with an average size of 825 bp, a single rRNA operon, and 40 tRNA genes. Together with the 1.66-Mbp genome of *P. marinus* MED4, the genome of *P. marinus* SS120 is one of the two smallest genomes of a photosynthetic organism known to date. It lacks many genes that are involved in photosynthesis, DNA repair, solute uptake, intermediary metabolism, motility, phototaxis, and other functions that are conserved among other cyanobacteria. Systems of signal transduction and environmental stress response show a particularly drastic reduction in the number of components, even taking into account the small size of the SS120 genome. In contrast, housekeeping genes, which encode enzymes of amino acid, nucleotide, cofactor, and cell wall biosynthesis, are all present. Because of its remarkable compactness, the genome of *P. marinus* SS120 might approximate the minimal gene complement of a photosynthetic organism.

Marine cyanobacteria of the genus *Prochlorococcus* (1) dominate phytoplankton communities in most tropical and temperate open ocean ecosystems (2). Their tiny cell sizes (0.5–0.7 μm) make *Prochlorococcus* spp. the smallest photosynthetic organisms known to date. Their major pigments are divinyl derivatives of chlorophyll *a* and *b* (Chl *a*₂ and *b*₂), which are unique to this genus (3). *Prochlorococcus* lacks phycobilisomes, large extrinsic multisubunit light-harvesting complexes found in typical cyanobacteria. These complexes are replaced by Chl *a*₂/*b*₂-binding proteins called Pcb, which are analogous in function but are structurally and phylogenetically distinct from the light-harvesting complexes of higher plants (4).

In the ocean, *Prochlorococcus* cells face a number of natural constraints, including strong inverse vertical gradients of irradiance and nutrients. A key factor of the adaptation to such variable conditions seems to be the existence of several physiologically and genetically distinct genotypes growing in different ecological niches (5). High-light-adapted genotypes (or “ecotypes”) occupy the upper, well illuminated but nutrient-poor 100-m layer of the water column, whereas low-light-adapted genotypes preferentially thrive at the bottom of the euphotic zone (80–200 m) at dimmer light but in a nutrient-rich environment.

In this article we describe the complete genome sequence of the low-light-adapted *Prochlorococcus marinus* type strain SS120 (also known as CCMP1375). Genomes of a high-light-adapted strain, *P. marinus* MED4, and another low-light-adapted strain, *P. marinus* MIT9313, have also recently been sequenced, and their genomes have been compared (6). In terms of photophysiology, *P. marinus* SS120 represents an extreme within the *Prochlorococcus* genus because of its ability to grow at very low light levels (5). Analyses of the genomic information of *P. marinus* SS120 and comparisons with other cyanobacterial genomes available to date (7–9) allowed us to delineate a putative minimal gene set of an oxyphotoautotrophic bacterium.

Materials and Methods

P. marinus SS120 could not be cultured axenically, leading to an $\approx 5\%$ contamination of the genomic DNA. To construct a plasmid library with a low level of contaminants, we produced five pilot libraries with insert size ranging from 3 to 10 kb. Approximately 100 clones were end-sequenced, and the AT content of each read was calculated. The G+C content of SS120 and the contaminant were $\approx 40\%$ and $\approx 60\%$, respectively, allowing the assignment of low G+C sequences to *Prochlorococcus*. The contamination level was the lowest, with inserts of 7 and 10 kb. Two shotgun genomic libraries were made by mechanical shearing of the DNA, size selection of the fragments, ligation into a low-copy plasmid (pCNS), and electroporation into DH10b cells (Invitrogen, Cergy-Pontoise, France). Plasmid DNAs (11,944 and 26,821 for the 7- and 10-kb insert libraries, respectively) were purified and end-sequenced as described (10) by using dye-primer and dye-terminator chemistries (50/50) on Licor 4200L and ABI3700 sequencers. Data were assembled with PHRAP (www.phrap.org), taking all sequences into account. An additional 789 directed reactions were performed to close the gaps and raise the quality of the sequence to finished standards. The integrity of the assembly was verified by comparing the theoretical lengths of bands obtained with two restriction enzymes with those determined experimentally (11).

ORFs were identified by using Glimmer, GeneMarks, and Critica (12–14). Transfer RNAs were predicted by tRNAscan-SE (15). In addition, transcription start sites were predicted by using the

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: ABC, ATP-binding cassette; Pro, *Prochlorococcus*; PS, photosystem; Chl, chlorophyll.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AE017126).

[†]To whom correspondence should be addressed. E-mail: partensky@sb-roscoff.fr.

^{††}Present address: Ocean Genome Legacy, Beverly, MA 01915.

Table 1. Properties of the genomes of autotrophic microorganisms

| Organism | Genome size, kb | Protein-coding genes | rRNA genes | tRNA genes* | Paralog cluster sizes [†] |
|--|-----------------|----------------------|------------|-------------|------------------------------------|
| <i>A. aeolicus</i> | 1,551 | 1,522 | 6 | 44 | -5.39 |
| <i>Chlorobium tepidum</i> TLS Cyanobacteria | 2,155 | 2,252 | 6 | 50 | ND |
| <i>P. marinus</i> SS120 | 1,751 | 1,884 | 3 | 40 | -5.01 |
| <i>Thermosynechococcus elongatus</i> BP-1 | 2,594 | 2,475 | 3 | 42 | -3.98 |
| <i>Synechocystis</i> sp. PCC 6803 | 3,573 | 3,169 | 6 | 41 | -4.44 |
| <i>Anabaena</i> sp. (<i>Nostoc</i>) PCC 7120 | 6,414 | 6,129 | 12 | 48 + 19 | -3.22 |
| Archaea | | | | | |
| <i>Methanococcus jannaschii</i> | 1,665 | 1,715 | 6 | 37 | -4.39 |
| <i>Aeropyrum pernix</i> K1 | 1,670 | ≈1,720 | 5 | 47 | -5.25 |
| <i>Methanopyrus kandleri</i> AV19 | 1,695 | 1,691 | 3 | 35 | ND |
| <i>Methanobacterium thermoautotrophicum</i> | 1,751 | 1,869 | 6 | 39 | -4.11 |

The data are from original articles and from the latest variants of the genome sequences at the NCBI Genomes (www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html) and Genomic tRNA Database (<http://rna.wustl.edu/GtRDB>) web sites. Only those archaeal autotrophs with genome sizes <2 Mb are listed.

*The data for *Anabaena* sp. represent the sum of the chromosomal and plasmid-encoded tRNA genes.

[†]To assess the propensity of these different prokaryotes for gene duplication, the distribution of sizes of species-specific paralog clusters in each genome was approximated by the power law (20). The steeper the slope of the curve, the fewer large clusters are observed. ND, not determined.

raster-score filter method recently developed for *P. marinus* MED4 (16) with respect to the higher G+C content of 34% within upstream regions of *P. marinus* SS120 for the calculation of the scoring matrix.

Genome annotation was performed by comparing the protein sequences with the Cluster of Orthologous Groups (COG) database (www.ncbi.nlm.nih.gov/COG) (17) and with the National Center for Biotechnology Information (NCBI) protein database (www.ncbi.nlm.nih.gov) by using BLAST and PSI-BLAST (18) with manual verification, as described (19). The number of copies of each particular gene in cyanobacterial genomes was either taken from the COG database (17) or estimated by BLAST searches against cyanobacterial protein databases (7–9). Species-specific expansions of paralogous gene families were determined as described (20). The SS120 genome sequence can be blasted at www.sb-roscoff.fr/Phyto/ProSS120.

Results and Discussion

General Features. The genome of *P. marinus* SS120 is composed of a single circular chromosome of 1,751,080 bp with an average G+C content of 36.4% (Table 1).

The origin of replication (*oriC*) was mapped between the *dnaN* and *thrC* genes on the basis of GC- and AT-skew analyses. The intergenic region between these two genes is AT-rich (73%) and contains six possible DnaA boxes with the consensus sequence 5'-[AT]TTCCACA-3'. The gene arrangement around *oriC* differs from the conserved arrangement found in many bacteria but is similar to the one in *Synechococcus* sp. PCC 7942 (21).

The genome contains 1,884 predicted ORFs with an average size of 825 bp. These ORFs represent 88.5% of the chromosome sequence. There is no significant asymmetry in the distribution of ORFs between the leading strand (50.6% of ORFs) and the lagging strand (49.4%). Biological roles were assigned to 1,254 (66.6%) of these ORFs. Among ORFs with unknown function, 399 (21.2%) have detectable homologs in the National Center for Biotechnology Information nonredundant database, and 231 (12.2%) have no detectable homolog. Three hundred ninety ORFs code for polypeptides of ≤100 aa. Although some had known function, such as the small subunits of photosystem II (PSII; e.g., PsbI, M, T or X) or high-light-inducible proteins (22), the function of many of them is not assigned yet. The SS120 genome contains a single rRNA operon (16S-23S-5S), 40 tRNA genes (which include cognates for all amino acids), and genes for three other RNAs.

Transcription of the genome is regulated by a reduced set of RNA polymerase sigma factors. In addition to SigA, SS120 has four genes encoding putative group 2 (23) sigma factors. A total of 3,130 transcription start sites were predicted by using the raster-score-filter method (16) for 1,289 noncoding upstream regions of the 1,930 predicted protein-coding or RNA genes. Our analyses excluded 641 regions <49 bp. The complete prediction can be downloaded from www.biologie.hu-berlin.de/~genetics/hess/hessproj.html. By analogy to MED4 (16), ≈40% of these sites were estimated to be functional. An experimental validation using eight randomly chosen genes (data not shown) showed that promoter elements and transcriptional start sites were correctly predicted for *pcbA*, *psbA*, *psbD*, *petH*, *kaiB*, *cpeB*, *ftsZ*, and *ntcA*.

***P. marinus* SS120 as a Minimal Genome of an Oxyphototrophic Organism.** Although larger than the “minimal” genomes of *Mycoplasma* spp. (24) and other obligate parasites, *P. marinus* SS120 has the smallest genome of all cyanobacteria sequenced to date (7–9), with the sole exception of the closely related 1.66-Mbp *P. marinus* MED4 genome (6). The *P. marinus* SS120 genome is comparable in size to the genomes of *Aquifex aeolicus* and archaeal autotrophs (Table 1). However, as a free-living autotroph, *P. marinus* SS120 has to encode enzymes for complete biosynthetic pathways for amino acids, nucleotides, cell wall carbohydrates, and cofactors, not to mention numerous components of the photosynthetic machinery (Fig. 1). A comparison of the four complete cyanobacterial genomes available to date gives some clues as to which genes are missing or underrepresented in *P. marinus* SS120. These include a few photosynthetic genes, genes involved in DNA repair, solute uptake, intermediary metabolism, and many other systems (see supporting information on the PNAS web site, www.pnas.org). Systems of signal transduction and environmental stress response (e.g., two-component systems) that are widely represented in the genomes of *Synechocystis* sp. PCC 6803 and *Anabaena* (*Nostoc*) sp. PCC 7120 (25–27) show a particularly drastic reduction. The number of encoded components is much larger than expected from the simple difference in size between *P. marinus* and the other cyanobacterial genomes (Table 2). There are no genes encoding light receptors such as bacteriophytochromes, cryptochromes, or rhodopsin. Many classes of signaling proteins (including hybrid histidine kinases, adenylate cyclases, diguanylate cyclases, phosphodiesterases, serine/threonine kinases, and phosphatases) that are found in most microbial genomes (28) are also lacking in SS120.

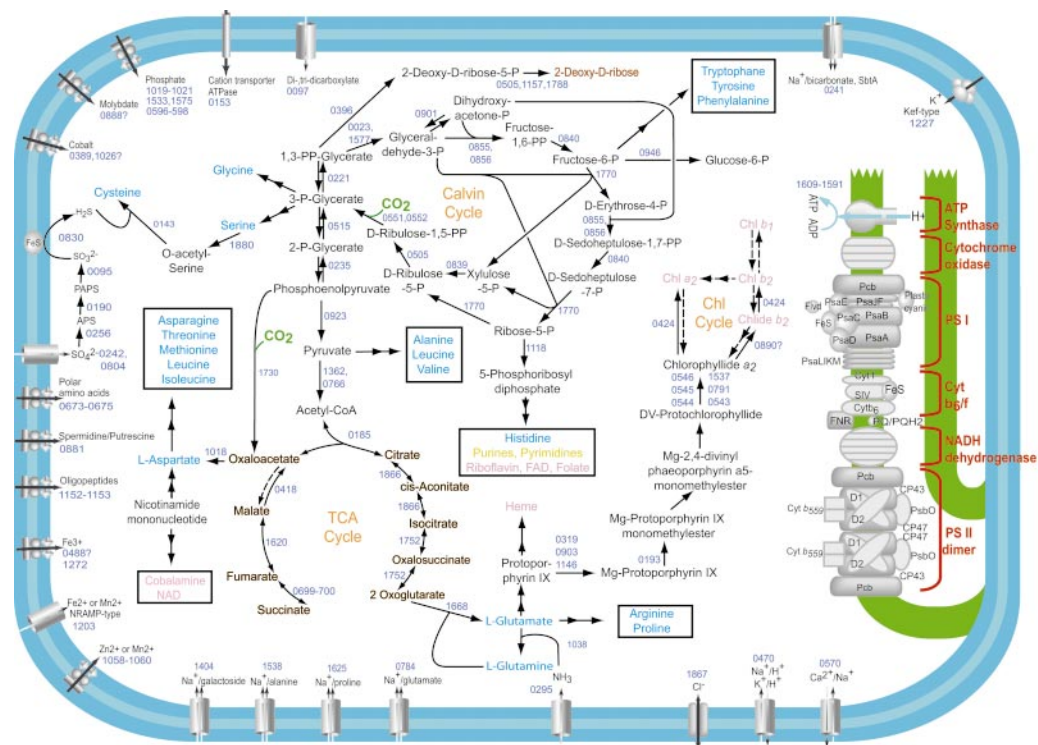


Fig. 1. Schematic organization of a *P. marinus* SS120 cell showing the main metabolic pathways and transporters. Gene identifiers are shown by numbers in blue. Genes with uncertain annotation are shown with a question mark. Reactions for which no candidate enzyme was confidently predicted are indicated by dashed arrows. Pathways that involve multiple reactions are shown by double arrows. Final biosynthetic products are indicated as follows: light blue, amino acids; dark yellow, nucleotides; brown, sugars; pink, cofactors. Chl cycle is based on ref. 37. Cyt, cytochrome; Flvd, flavodoxin; FNR, ferredoxin:NADP⁺ oxidoreductase; PQ, plastoquinone.

In contrast to almost every other microbial genome, of the five sensor histidine kinases and six response regulators encoded by *P. marinus* SS120 only one pair of genes forms an operon. The absence of diguanylate cyclases and phosphodiesterases that are often involved in regulation of extracellular polysaccharide production and biofilm formation (28) is in agreement with the lack of close association between *P. marinus* cells in their natural habitat.

A low number of lineage-specific duplications in the SS120 genome compared with other cyanobacteria was confirmed by the smaller size of paralog clusters (Table 1). Thus, in agreement with the trend noted previously for other small-genome organisms (29), SS120 has relatively few paralogous genes and encodes few analogous enzymes.

DNA Repair and Genome Rearrangements. Although *P. marinus* SS120 is a low-light-adapted strain, its DNA repair genes are similar in their number and diversity to those from other cyanobacterial genomes, with the exception of a few missing genes (see supporting information). It is so far the only cyanobacterium that encodes a complete exonuclease V/recombinase complex (RecBCD). Although SS120 lacks the gene for deoxyribodipyrimidine photolyase, which is present in other cyanobacteria, it encodes instead a pyrimidine dimer-specific glycosylase, Pro1489, which is absent in other cyanobacteria and might have been horizontally transferred from a viral/phage genome. Another possible example of horizontally transferred genes is the type-I restriction/modification system (Pro0628–0630, Pro1274), which might be responsible for specific DNA methylation. In contrast to freshwater cyanobacteria, which have numerous transposase genes, SS120 does not encode any transposases (see supporting information). Together with the absence of site-specific recombinases (homologs of DNA invertase Pin) and XerC-like integrases, this observation suggests

that SS120 is less prone to genome rearrangements than are other cyanobacteria.

Photosynthesis. Photosynthetic apparatus. The genome of *P. marinus* SS120 contains the whole set of PSII genes, with the exception of *psbU*, encoding the 12-kDa extrinsic subunit of PSII, which acts in the stabilization and protection of the oxygen-evolving complex against inactivation by heat, and *psbV*, encoding a low-potential cytochrome *c* associated with the luminal surface of the PSII reaction center complex.

All PSII genes in SS120 but one (*psbF*) are single-copy, in contrast to other cyanobacterial genomes, which contain multiple copies of *psbA*, *psbC*, and *psbD* genes (see supporting information).

SS120 has 11 genes coding for PSI proteins and 6 genes encoding the subunits of the cytochrome *b*₆/*f* complex. All these genes are single-copy. Ferredoxin:NADP⁺ oxidoreductase (PetH, encoded by Pro1123), which catalyzes electron transfer from photochemically reduced ferredoxin to NADP⁺, has an N-terminal domain with a predicted transmembrane helix that could anchor PetH to the thylakoid membrane. This contrasts with PetH from freshwater cyanobacteria (30, 31), whose N-terminal part is similar to a phycocyanin-associated linker polypeptide, allowing specific binding of this protein to phycocyanin. This observation is consistent with the different light-harvesting systems found in *P. marinus* and other cyanobacteria (4).

The *pcb* gene family, which encodes the proteins of the light-harvesting antenna complex in SS120, is a rare example of gene amplification in this otherwise compact genome. Compared with only one *pcb* gene in MED4 and all other high-light-adapted strains checked to date and two *pcb* copies in MIT9313 (6, 32), SS120 encodes eight different *pcb* genes, one more copy than previously reported (33).

SS120 has a small set of genes to synthesize the α , β , and γ

Table 2. Signaling and regulatory domains or genes that are absent in *P. marinus* SS120 (Pma) or in lower copy number than in the freshwater cyanobacteria *Thermosynechococcus elongatus* (Tel), *Synechocystis* sp. PCC6803 (Syn), and *Anabaena* sp. PCC7120 (Ana)

| Signaling system component or domain | Gene name | COG no. | No. of genes/genome | | | |
|--|--------------|----------------|---------------------|-----|-----|-----|
| | | | Pma | Tel | Syn | Ana |
| Signal transduction mechanisms | | | | | | |
| Histidine kinases | — | 0642 | 5 | 16 | 42 | 124 |
| Response regulators | — | 2197 | 6 | 26 | 41 | 84 |
| PAS domains | — | 2202 | 1 | 13 | 23 | 57 |
| GAF domains | — | 2203 | 0 | 15 | 28 | 62 |
| HPT domains | — | 2198 | 0 | 4 | 7 | 9 |
| GGDEF domains | — | 2199 | 0 | 10 | 23 | 14 |
| EAL domains | — | 2200 | 0 | 5 | 13 | 7 |
| HD-GYP domains | — | 2206 | 0 | 1 | 2 | 2 |
| Ser/Thr protein kinase | — | 0515 | 0 | 10 | 9 | 35 |
| Ser/Thr protein | — | 0631 | 0 | 3 | 3 | 4 |
| Phosphatase 1 | — | — | — | — | — | — |
| Ser/Thr protein | — | 0639 | 0 | 1 | 1 | 2 |
| Phosphatase 2 | — | — | — | — | — | — |
| Adenylate cyclase | <i>cyaA</i> | 2114 | 0 | 1 | 2 | 7 |
| cAMP-binding domains | <i>crp</i> | 0664 | 2 | 3 | 12 | 14 |
| Phototaxis* | | | | | | |
| Hybrid histidine kinase | <i>taxAY</i> | 0643/2198/0784 | 0 | 3 | 3 | 3 |
| Photoreceptor for positive phototaxis | <i>taxD</i> | 2203/0840 | 0 | 3 | 3 | 3 |
| Putative phototaxis protein | <i>taxP</i> | 0784 | 0 | 3 | 3 | 3 |
| CheW-like signal transducer | <i>taxW</i> | 0835 | 0 | 3 | 3 | 3 |
| CheY-like signal transducer | <i>taxY</i> | 0784 | 0 | 3 | 3 | 3 |
| Regulation of RNA polymerase | | | | | | |
| Alternative σ 28 | <i>fliA</i> | 1191 | 0 | 1 | 1 | 2 |
| Alternative σ 24 | <i>rpoE</i> | 1592 | 0 | 3 | 3 | 2 |
| Anti-sigma factor | <i>rsbW</i> | 2172 | 1 | 2 | 2 | 5 |
| Serine phosphatase regulator of σ | <i>rsbU</i> | 2208 | 1 | 2 | 5 | 5 |
| Others | | | | | | |
| Phycobilisome degradation proteins | <i>nbIA</i> | — | 0 | 1 | 2 | 2 |
| | <i>nbIB</i> | — | 0 | 2 | 1 | 1 |
| Universal stress protein | <i>uspA</i> | 0589 | 0 | 4 | 4 | 7 |
| Drought-induced stress protein | CDS P34 | — | 0 | 1 | 2 | 2 |
| Ankyrin repeats | — | 0666 | 0 | 1 | 1 | 2 |
| Forkhead domains | — | 1716 | 0 | 2 | 2 | 2 |

COG, clusters of orthologous groups.

*Gene designations after ref. 54.

subunits of phycoerythrin type III (PEIII) as the sole phycobiliprotein (32, 34). This also includes three bilin reductases, PcyA, PebA, and PebB (Pro0819, Pro1748–1749), for the biosynthesis of phycobiliprotein chromophores phycoerythrobilin and 3Z-phycoerythrobilin (35). Phycoerythrobilin is bound by PEIII, whereas 3Z-phycoerythrobilin, which is generated by the activity of PcyA, as shown for *Prochlorococcus* MED4 (35), serves as the chromophore for phycocyanin and allophycocyanin. Thus, questions arise as to which cognate polypeptide binds phycocyanobilin and why PEIII has been conserved in this genus given that the amount of PEIII per cell is very low (34) and light harvesting rests on Pcb proteins.

Chl biosynthesis. The set of Chl biosynthesis genes found in *P. marinus* SS120 is fairly complete (36). It has only one copy of *hemN* (Pro1385) encoding the oxygen-independent coproporphyrinogen III oxidase versus two in freshwater cyanobacteria (see supporting information). Similarly, SS120 has only one copy (compared with several copies in other cyanobacteria) of the *acsF/crdI* gene encoding an aerobic Mg-protoporphyrin IX monomethyl ester oxidative cyclase. Like *P. marinus* spp. MED4 and MIT9313 (32), SS120 has a gene (Pro0890) encoding a non-heme oxygenase with binding domains for a [2Fe-2S] Rieske center and for a mononu-

clear iron. These properties make it the most plausible candidate for chlorophyllide *a* oxygenase (Cao), an enzyme needed for the biosynthesis of chlorophyllide *b*₂ and therefore Chl *b*₂ (37) (see Fig. 1). However, Pro0890 is only distantly related to Cao previously identified in other Chl *b*-containing oxyphotobacteria, green algae and plants (38). Low-light-adapted strains such as SS120 or NATL1 can synthesize monovinyl Chl *b* (Chl *b*₁) when grown under high-light conditions (39), and Chl *b*₁ was assumed to derive from Chl *b*₂ (40). Thus, SS120 is likely to encode a 4-vinyl reductase, but no such enzyme has been found in the genome. Therefore, further studies are needed to elucidate the enzymes for the final phases of the biosynthesis of the specific divinyl-Chls of SS120.

Autotrophic Metabolism. Carbon assimilation. Despite the key role of carbon dioxide in autotrophic metabolism, the *P. marinus* SS120 genome does not encode any of the three principal CO₂/HCO₃⁻ uptake systems found in other cyanobacteria (41, 42): the ATP-binding cassette (ABC)-type bicarbonate transporter CmpABCD, the constitutive CO₂ uptake system NdhD4/NdhF4/CupB, or the low-CO₂ induced high-affinity system NdhD3/NdhF3/CupA. The only CO₂ uptake system found in SS120 is the ΔμNa⁺-dependent transporter SbtA (Pro0241). The *ntpJ* gene, whose product is

reportedly required for the SbtA-catalyzed bicarbonate uptake (41), is also present in the SS120 genome (Pro0098). SS120 does not encode any known carbonic anhydrases, so their function is probably fulfilled by an as-yet-unidentified protein. *P. marinus* SS120 assimilates CO₂ via the Calvin cycle and has the complete set of enzymes of this pathway (Fig. 1). However, Rubisco (large and small subunits), phosphoribulokinase, and pentose-5-phosphate epimerase of SS120 all have closer homologs in proteobacteria, such as *Thiobacillus* species, than in freshwater cyanobacteria, a feature shared with other marine picocyanobacteria (32, 42).

Whereas all freshwater cyanobacteria encode two analogous fructose-1,6-bisphosphatases, related, respectively, to *Escherichia coli* *fbp* and *glpX* genes, *P. marinus* SS120 has only the latter one. This form (F-1) has been shown to hydrolyze both fructose-1,6-bisphosphate and sedoheptulose-1,7-bisphosphate (43). This is yet another case of gene economy in which a gene encoding a monofunctional enzyme (*fbp*, active only with the former substrate) could have been eliminated because a bifunctional enzyme was present. **Nitrogen assimilation.** The *P. marinus* SS120 genome does not contain genes for transport systems for nitrate, nitrite, cyanate, and urea, which are present in freshwater cyanobacteria, nor that coding for a nitrate/nitrite permease recently discovered in a marine *Synechococcus* (44). Accordingly, it does not encode nitrate/nitrite reductases or urease. These results indicate that this strain relies for growth on reduced nitrogen compounds, such as NH₄⁺ and amino acids. Indeed, SS120 encodes an ammonia transporter (Fig. 1). Field studies recently showed that *Prochlorococcus* can import amino acids (45). There are four unassigned ABC-type transport systems and several Na⁺/amino acid symporters in SS120 that could provide that capability.

P. marinus SS120 does not possess enzymes for the synthesis or hydrolysis of cyanophycin (poly-L-arginyl-L-aspartate), which serves as nitrogen reserve in some cyanobacteria (see supporting information); however, it can conserve nitrogen by forming spermidine. Indeed, it possesses the *speE* gene encoding spermidine synthase (Pro1848), which is missing in the freshwater strains *Synechocystis* sp. PCC 6803 and *Anabaena* sp. PCC 7120.

Phosphorus assimilation. *P. marinus* SS120 encodes a typical ATP-dependent *PstCAB* system (Pro0598–Pro0596) for transporting phosphate, as well as an ABC-type transporter for potential use of phosphonate (Pro1019–1021). The regulatory component PhoU is missing, as are PhoR and PhoB proteins, which form a two-component system responsible for phosphate sensing and regulation in a variety of bacteria. This apparent gene loss goes a step further than even in the *Prochlorococcus* MIT9313 genome, which contains an intact *phoB* gene and a frameshifted *phoR* gene (46).

Sulfur assimilation. The pathway of sulfate uptake and reduction in *P. marinus* SS120 is similar to that in other cyanobacteria, except that the ABC-type sulfate transporter is replaced by two sulfate permeases of the major facilitator superfamily (Pro0242 and Pro0804). SS120 encodes two copies of cysteine synthase, Pro0143 and Pro0403.

Intermediary Metabolism. Tricarboxylic acid cycle (TCA) and related reactions. Like many other bacteria, *P. marinus* SS120 encodes an incomplete citric acid cycle (47). The *sucA* and *sucB* genes coding for the E1 (dehydrogenase) and E2 (dihydrolipoamide succinyl transferase) components of the 2-oxoglutarate dehydrogenase complex are missing, as are genes for both subunits of succinyl-CoA synthetase. NAD-dependent malate dehydrogenase is also lacking in SS120; its function might be taken over by malate:quinone oxidoreductase (Pro0418), an enzyme not found in other cyanobacteria but one that is highly similar to the TCA enzyme from *Helicobacter pylori* (47). However, the *H. pylori* enzyme could not catalyze the reverse reaction, reduction of oxaloacetate, which is required for the functioning of the reductive branch of the incom-

plete TCA cycle in SS120. Therefore, it remains to be determined whether Pro0418 actually catalyzes this reaction in SS120.

Amino acid biosynthesis. *P. marinus* SS120 encodes the complete set of enzymes for biosynthesis of all amino acids except for lysine. Lysine biosynthesis in SS120 must occur anyway and possibly proceeds via the diaminopimelate pathway, although the mechanism of conversion of tetrahydrodipicolinate into diaminopimelate remains unclear. The pathway of methionine biosynthesis in SS120 includes three enzymes that are found in *E. coli* and other bacteria but seem to be missing in other cyanobacteria. Conversion of homoserine into homocysteine, catalyzed by these enzymes [homoserine transsuccinylase (MetA; Pro0801), cystathionine γ -synthase (MetB; Pro0405), and β -cystathionase (MetC; Pro0404)], can also be catalyzed by a combination of the homoserine *O*-acetyltransferase and *O*-acetylhomoserine-sulfhydrylase (Pro0800).

In yet another manifestation of gene economy, a dedicated tyrosine aminotransferase is missing, and the last step in the biosynthesis of Phe, Tyr, and Trp is apparently catalyzed by the nonspecific aromatic acid aminotransferase HisC.

Nucleotide biosynthesis. *P. marinus* SS120 encodes a complete set of enzymes for *de novo* purine and pyrimidine biosynthesis. In a rare deviation from its “minimal” gene content, SS120 encodes two variants of phosphoribosylglycinamide formyltransferase, folate-dependent PurN and formate-dependent PurT. Likewise, it encodes two versions of dihydroorotase that have been previously described in *E. coli* and *Bacillus subtilis*, respectively. The glutamine amidotransferase and pyrophosphatase domains of GMP synthase (GuaA), which form a single polypeptide chain in all bacteria, are encoded in SS120 by two separate genes, as is the case in most archaea. Like other cyanobacteria, SS120 encodes the catalytic subunit of aspartate carbamoyltransferase (PyrB) but not the regulatory subunit (PyrI).

In contrast to *de novo* biosynthesis pathways, purine and pyrimidine salvage pathways are poorly represented in SS120. Like other cyanobacteria, SS120 lacks classical thymidylate synthase ThyA and instead encodes the recently described alternative enzyme ThyX (48). Similar to some other cyanobacteria, SS120 does not encode either anaerobic ribonucleoside triphosphate reductase (NrdDG) or ribonucleoside diphosphate reductase (NrdAF). Instead, SS120 encodes the B₁₂-dependent (class-II) ribonucleotide reductase that is also found in *Anabaena* sp. PCC 7120 (49).

Cell wall biosynthesis. As a typical bacterium, SS120 encodes a complete set of enzymes for peptidoglycan synthesis, including MurABCDEFGFI, alanine racemase, and D-ala-D-ala ligase.

Cofactor biosynthesis. SS120 encodes the full set of enzymes of the biosynthetic pathways for NAD, FAD, heme, B₁₂, biotin, folate, tetrahydrobiopterin, and phyloquinone. As in other cyanobacteria, the thiamine biosynthetic pathway lacks one enzyme, hydroxymethylpyrimidine/phosphomethylpyrimidine kinase (ThiD). Similarly, the set of *ubi* genes (which in cyanobacteria are likely involved in plastoquinone, not ubiquinone biosynthesis) is also incomplete, lacking *ubiB*. Thus, the reactions catalyzed in bacteria by ThiD and UbiB are likely performed in cyanobacteria by alternative enzymes yet to be characterized. The pathway of CoA biosynthesis lacks two enzymes, aspartate 1-decarboxylase (PanD) and pantothenate kinase (CoaA), whereas pyridoxine biosynthesis is represented only by pyridoxine synthase (PdxA and PdxJ subunits). The well known diversity of enzymes in these pathways (19) also suggests that these “missing” enzymes are encoded in the SS120 genome in alternative versions. In contrast, the genes for the biosynthesis of molybdenum cofactor are all missing, consistent with the absence of nitrate and nitrite reductases and other molybdopterin-containing enzymes.

Adaptation to the marine environment. *P. marinus* SS120 contains several systems that are likely critical in the adaptation to the marine

environment. Based on the sequence of the c subunit (AtpE) of its H⁺-ATP synthetase (50), it appears that, in addition to (or instead of) H⁺ ions, this enzyme can also transport Na⁺. Unlike many marine bacteria, *P. marinus* SS120 does not encode a primary Na⁺ pump. Nevertheless, it can extrude Na⁺ ions at the expense of the proton gradient, using an NhaP-type Na⁺/H⁺ antiporter (Pro0470). The resulting sodium gradient is apparently used for export of Ca²⁺ ions via the Ca²⁺/Na⁺ antiporter (Pro0570). Other Na⁺-dependent transporters in SS120 include Na⁺/bicarbonate symporter SbtA (Pro0241), Na⁺/proline symporter PutP (Pro1635), Na⁺/alanine symporter AlsT (Pro1538), Na⁺/glutamate symporter GltS (Pro0784), Na⁺/galactoside symporter MelB (Pro1404), and two SS120-specific predicted Na⁺-dependent permeases that belong, respectively, to the divalent anion:sodium symporter family (Pro0097) and the neurotransmitter:sodium symporter family (Pro1452). SS120 also encodes a homolog of a proteobacterial salt-induced outer membrane protein (Pro1529), which is absent in freshwater cyanobacteria.

Conclusions

We show here that *P. marinus* SS120, the type species of the dominant photosynthetic genus in the ocean, has a nearly minimal gene complement of an oxyphototrophic organism. The frugality of the gene repertoire of this organism is manifest at many levels. Signaling systems are either absent or represented by fewer domains in *P. marinus* SS120 than in other cyanobacteria, consistent with the fact that the oligotrophic marine environment where it preferentially thrives is much more stable than fresh waters. This argument is strengthened by the fact that, in the field, *P. marinus* SS120-like cells are restricted to the bottom part of the illuminated layer of oceans (5, 39). It appears likely that the compact genome of *P. marinus* SS120 is maintained by selection and is connected to the small cell volume of this organism ($\approx 0.1 \mu\text{m}^3$), which is the theoretical lower limit for an oxyphototroph (51). Small cell size has at least two distinct advantages for a phytoplanktonic organism: (i)

to increase the *in vivo* absorption coefficient by reducing the package effect (52) and (ii) to increase the cell surface to volume ratio and thereby improve nutrient uptake. Thus, *P. marinus* SS120 seems to be a rare case of a free-living organism for which a direct connection between fundamental characteristics of the genome and organism ecology is apparent.

Whether such a reduced genome is a derived state resulting from progressive gene loss or is an ancestral state is as yet unclear. Phylogenies based on 16S rRNA genes (see, for example, refs. 46 and 53) show that, within the *Prochlorococcus* radiation, SS120 is found in a "low-light clade" at an intermediate position between the "high-light clade," represented by MED4 (1) and another "low-light clade" containing MIT9313 (5) that is located near the base of the radiation. The MED4 strain has an even more compact genome than SS120 (1.66 vs. 1.75 Mbp, respectively), whereas that of MIT9313 is larger (2.41 Mbp) (6). The lower diversity within the high-light clade suggests that it has appeared more recently than the more highly divergent low-light clades (46, 53). Thus, evolution in the genus *Prochlorococcus* would have tended toward genome reduction. However, this phenomenon would certainly not be enough to account for the large differences in genome sizes and complexity between marine *P. marinus* SS120 and the freshwater cyanobacteria strains used as references in this paper. Specific genome amplification and diversification also must have taken place during adaptation of the latter to their specific environments. Confirmation of these hypotheses still awaits phylogenetic analysis of large gene regions, but these will be reliable only when more complete cyanobacterial genomes become available.

We thank D. Bhaya for helpful hints about motility and phototaxis genes. This work was supported by the European Union program MARGENES (QLRT-2001-01226) and Genomer (Région Bretagne). A.D. is supported by a doctoral fellowship from Région Bretagne, W.R.H. is supported by Deutsche Forschungsgemeinschaft Grant SFB 429-TPA4, and D.J.S. is a Royal Society University research fellow.

- Chisholm, S. W., Frankel, S. L., Goericke, R., Olson, R. J., Palenik, B., Waterbury, J. B., West-Johnsrud, L. & Zettler, E. R. (1992) *Arch. Microbiol.* **157**, 297–300.
- Partensky, F., Hess, W. R. & Vaulot, D. (1999) *Microbiol. Mol. Biol. Rev.* **63**, 106–127.
- Goericke, R. & Repeta, D. J. (1992) *Limnol. Oceanogr.* **37**, 425–433.
- LaRoche, J., van der Staay, G. W., Partensky, F., Ducret, A., Aebersold, R., Li, R., Golden, S. S., Hiller, R. G., Wrench, P. M., Larkum, A. W. et al. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 15244–15248.
- Moore, L. R. & Chisholm, S. W. (1999) *Limnol. Oceanogr.* **44**, 628–638.
- Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., et al. (2003) *Nature*, 10.1038/nature01947.
- Kaneko, T., Nakamura, Y., Wolk, C. P., Kuritz, T., Sasamoto, S., Watanabe, A., Iriguchi, M., Ishikawa, A., Kawashima, K., Kimura, T. et al. (2001) *DNA Res.* **8**, 205–213.
- Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirose, M., Sugiura, M., Sasamoto, S. et al. (1996) *DNA Res.* **3**, 109–136.
- Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., Watanabe, A., Iriguchi, M., Kawashima, K., Kimura, T. et al. (2002) *DNA Res.* **9**, 123–130.
- Artiguenave, F., Wincker, P., Brottier, P., Duprat, S., Jovelin, F., Scarpelli, C., Verdier, J., Vico, V., Weissenbach, J. & Saurin, W. (2000) *FEBS Lett.* **487**, 13–16.
- Strehl, B., Holtzendorff, J., Partensky, F. & Hess, W. R. (1999) *FEMS Microbiol. Lett.* **181**, 261–266.
- Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999) *Nucleic Acids Res.* **27**, 4636–4641.
- Besemer, J., Lomsadze, A. & Borodovsky, M. (2001) *Nucleic Acids Res.* **29**, 2607–2618.
- Badger, J. H. & Olsen, G. J. (1999) *Mol. Biol. Evol.* **16**, 512–524.
- Lowe, T. M. & Eddy, S. R. (1997) *Nucleic Acids Res.* **25**, 955–964.
- Vogel, J., Axmann, I. M., Herzel, H. & Hess, W. R. (2003) *Nucleic Acids Res.* **31**, 2890–2899.
- Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. (2000) *Nucleic Acids Res.* **28**, 33–36.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zheng, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
- Koonin, E. V. & Galperin, M. Y. (2002) *Sequence–Evolution–Function: Computational Approaches in Comparative Genomics* (Kluwer, Boston).
- Lespinet, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. (2002) *Genome Res.* **12**, 1048–1059.
- Liu, Y. & Tsinoiremas, N. F. (1996) *Gene* **172**, 105–109.
- Bhaya, D., Dufresne, A., Vaulot, D. & Grossman, A. (2002) *FEMS Microbiol. Lett.* **215**, 209–219.
- Imamura, S., Yoshihara, S., Nakano, S., Shiozaki, N., Yamada, A., Tanaka, K., Takahashi, H., Asayama, M. & Shirai, M. (2003) *J. Mol. Biol.* **325**, 857–872.
- Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M. et al. (1995) *Science* **270**, 397–403.
- Mizuno, T., Kaneko, T. & Tabata, S. (1996) *DNA Res.* **3**, 407–414.
- Ohmori, M., Ikeuchi, M., Sato, N., Wolk, P., Kaneko, T., Ogawa, T., Kanehisa, M., Goto, S., Kawashima, S., Okamoto, S. et al. (2001) *DNA Res.* **8**, 271–284.
- Meeks, J. C., Campbell, E. L., Summers, M. L. & Wong, F. C. (2002) *Arch. Microbiol.* **178**, 395–403.
- Galperin, M. Y., Nikolskaya, A. N. & Koonin, E. V. (2001) *FEMS Microbiol. Lett.* **203**, 11–21.
- Galperin, M. Y., Walker, D. R. & Koonin, E. V. (1998) *Genome Res.* **8**, 779–790.
- Schluchter, W. M. & Bryant, D. A. (1992) *Biochemistry* **31**, 3092–3102.
- Fillat, M. F., Flores, E. & Gomez-Moreno, C. (1993) *Plant Mol. Biol.* **22**, 725–729.
- Hess, W. R., Rocap, G., Ting, C. S., Larimer, F., Stilwagen, S., Lamerdin, J. & Chisholm, S. W. (2001) *Photosynth. Res.* **70**, 53–71.
- Garczarek, L., Hess, W. R., Holtzendorff, J., van der Staay, G. W. & Partensky, F. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 4098–4101.
- Hess, W. R., Steglich, C., Lichtle, C. & Partensky, F. (1999) *Plant Mol. Biol.* **40**, 507–521.
- Frankenberg, N., Mukougawa, K., Kohchi, T. & Lagarias, J. C. (2001) *Plant Cell* **13**, 965–978.
- Suzuki, J. Y., Bollivar, D. W. & Bauer, C. E. (1997) *Annu. Rev. Genet.* **31**, 61–89.
- Oster, U., Tanaka, R., Tanaka, A. & Rudiger, W. (2000) *Plant J.* **21**, 305–310.
- Shibata, M., Ohkawa, H., Katoh, H., Shimoyama, M. & Ogawa, T. (2002) *Funct. Plant Biol.* **29**, 123–129.
- Badger, M. R. & Price, G. D. (2003) *J. Exp. Bot.* **54**, 609–622.
- Tamoi, M., Murakami, A., Takeda, T. & Shigeoka, S. (1998) *Biochim. Biophys. Acta* **1383**, 232–244.
- Sakamoto, T., Inoue-Sakamoto, K. & Bryant, D. A. (1999) *J. Bacteriol.* **181**, 7363–7372.
- Zubkov, M. V., Fuchs, B. M., Tarran, G. A., Burkill, P. H. & Amann, R. (2003) *Appl. Environ. Microbiol.* **69**, 1299–1304.
- Scanlan, D. J. & West, N. J. (2002) *FEMS Microbiol. Ecol.* **40**, 1–12.
- Kather, B., Stingl, K., van der Rest, M. E., Altendorf, K. & Molenaar, D. (2000) *J. Bacteriol.* **182**, 3204–3209.
- Mylykallio, H., Lipowski, G., Leduc, D., Filee, J., Forterre, P. & Liebl, U. (2002) *Science* **297**, 105–107.
- Gleason, F. K. & Olszewski, N. E. (2002) *J. Bacteriol.* **184**, 6544–6550.
- Dzioba, J., Hase, C. C., Gosink, K., Galperin, M. Y. & Dibrov, P. (2003) *J. Bacteriol.* **185**, 674–678.
- Raven, J. A. (1994) *J. Plankton Res.* **16**, 565–580.
- Morel, A., Ahn, Y.-W., Partensky, F., Vaulot, D. & Claustre, H. (1993) *J. Mar. Res.* **51**, 617–649.
- Urbach, E., Scanlan, D. J., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. (1998) *J. Mol. Evol.* **46**, 188–201.
- Bhaya, D., Takahashi, A. & Grossman, A. R. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 7540–7545.