A hyperconserved protein in *Prochlorococcus* and marine *Synechococcus*

Olga Zhaxybayeva*[1], J Peter Gogarten[2] and W. Ford Doolittle[1]

[1]Department of Biochemistry and Molecular Biology, Dalhousie University, 5850 College Street, Halifax, NS, B3H 1X5, Canada; olgazh@dal.ca, ford@dal.ca

[2]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA; gogarten@uconn.edu

*corresponding author

The definitive version of this manuscript is available at www.blackwell-synergy.com. The full citation for the article is *FEMS Microbiol Lett*, 2007, **274**: 30-34

Abstract

We describe an open reading frame (encoding a hyperconserved protein, or HCP) of unknown function in *Prochlorococcus*/marine *Synechococcus* genomes that is 100% conserved on the amino acid level but lacks homologs outside this group. Such conservation of an uncharacterized group-specific gene is unusual, and unexpected given the extensive genomic divergence of this group at other loci. Comparative analyses indicate that HCP is under stabilizing selection and has resided in these genomes since the last common ancestor of the group.

Introduction

Prochlorococcus and marine *Synechococcus* are abundant unicellular marine cyanobacteria (Waterbury *et al.*, 1986; Chisholm *et al.*, 1992), for which 13 completed genomes are now available (Table 1). Although members of the two genera are more than 96% identical in their 16S rRNA sequences, their genomes are surprisingly divergent, possibly due to high mutation rates and the lack of several DNA repair enzymes in some of the genomes (Rocap *et al.*, 2003). Pairwise Average Nucleotide Identity [ANI] values are as low as 64.5% and Average Amino acid Identity [AAI] as low as 57.4%. Such divergence is found even among the 933 “core” genes present in all 13 *Prochlorococcus* and marine *Synechococcus* genomes. On average these genes show amino acid identity across all 13 genomes at only 44% of their sites. It is noteworthy, then, that one open reading frame (here designated HCP for Hyper-Conserved Protein), of 62

amino acids, is conserved at all positions in all genomes. In this article, we characterize this protein *in silico*.

Materials and Methods

The ORF encoding the Hyper Conserved Protein (HCP) was used as query in BLAST (Altschul *et al.*, 1997) searches against the nr database, 705 microbial genomes and 21 environmental metagenomic databases that were available at NCBI BLAST web server on January 30, 2007. BLASTP, TBLASTN and PSI-BLAST searches, as well as CDD search with RPS-BLAST were performed. A TBLASTN search of the Global Ocean Sampling Expedition database was performed at the CAMERA version 0.7 website (http://camera.calit2.net/). Prediction of transmembrane helices, coiled-coils and helix-turn-helix motifs, and calculation of the predicted isoelectic point was performed using programs from the EMBOSS suite (Rice *et al.*, 2000). The core of 13 completely sequenced genomes was defined as a set of genes picking each other as top-scoring hits in a fully transitive BLASTP search (E-value < 10^{-4}). The core gene sets were aligned in ClustalW v. 1.83 (Thompson *et al.*, 1994). The alignment conservation was defined as proportion of alignment sites per gene that did not have amino acid substitutions (gaps were treated as missing data) and calculated using an in-house Perl script. Similarly, core genes and alignment conservation were calculated for three other groups of genomes with several completely sequenced genomes (genome lists are available upon request). The number of substitutions per site in the most conserved 20 genes in the 13 genomes was calculated as a

total branch length of a phylogenetic tree reconstructed from the concatenated alignment (the tree was reconstructed using NEIGHBOR from the PHYLIP package (Felsenstein, 1993) with distances calculated in TREE-PUZZLE (Schmidt *et al.*, 2002) under JTT+G substitution model). The probability of observing no amino acid substitutions was calculated under a Poisson distribution. Genome-wide ANI and AAI were calculated as described in (Konstantinidis & Tiedje, 2005a; Konstantinidis & Tiedje, 2005b). Values of GC3 were calculated using in-house Perl script. 10 other cyanobacterial genomes that were used in the genome context analyses are *Anabaena variabilis* ATCC 29413, *Gloeobacter violaceus* PCC 7421, *Nostoc* sp. PCC 7120, *Synechococcus elongatus* PCC 6301, *Synechococcus elongatus* PCC 7942, *Synechococcus* sp. JA-2-3B'a(2-13), *Synechococcus* sp. JA-3-3Ab, *Synechocystis* sp. PCC 6803, *Thermosynechococcus elongatus* BP-1, *Trichodesmium erythraeum* IMS101 (all available through the NCBI).

Results and Discussion

Remarkable conservation of the HCP

The HCP is conserved in all positions in all 13 *Prochlorococcus*/marine *Synechococcus* genomes. Given that the twenty next most conserved core genes have approximately 0.29 amino acid substitutions per site across all genomes, 100% conservation is surprising. Assuming this substitution rate the probability of observing a completely conserved protein with 62 amino acids is $2*10^{-8}$.

Misannotation of the HCP gene in three out of 13 genomes and on variation at the C-terminal

In the CC9902, MIT9515 and CC9311 genomes the HCP gene is annotated with an alternative start codon. This appears to be an artifact of automated genome annotations, since in the CC9902 genome this alternative start codon leaves only three nucleotides between the upstream Trp-tRNA gene and the HCP, while in the latter two genomes it overlaps with the Trp-tRNA gene by 29 nucleotides. There appears to be a length variant of the HCP gene, resulting from mutation of the stop codon TAA to CAA and hence causing two extra amino acids at the C terminal (QK) present in two out of 13 genomes (MIT9312 and AS9601), as well as in the Global Ocean Sampling Expedition database (see below).

HCP is a group-specific protein

This ORF is found in no other cyanobacterium (10 sequenced genomes), nor in any other genome or database, with the exception of the “Global Ocean Sampling Expedition” metagenome. Several samples of the Global Ocean Sampling Expedition database had significant BLAST hits to the HCP (total of 119 hits with E-value $<10^{-27}$). The samples that return BLAST hits are from the Sargasso Sea, Caribbean Sea, Eastern Tropical Pacific, and from the Pacific Ocean near the Galapagos Islands. Somewhat unexpectedly, no significant BLAST hits were found in samples from the North Pacific Subtropical Gyre (DeLong *et al.*, 2006), but this might be due to this metagenome being not

sampled as deeply as the ones from the Global Ocean Sampling Expedition. Also, no significant BLAST hits were found in the viral metagenomes that were available for the BLAST at the CAMERA website (see Materials and Methods).

HCP is not a recent transfer within the Prochlorococcus/marine Synechococcus group

At the nucleotide level there are 65 positions with substitutions within aligned HCP genes, all synonymous, indicating that the ORF does indeed encode protein, that it is under purifying selection, and that it has resided for some time in these genomes. Indeed, the 13 analyzed genomes have a very broad range of GC content, varying from 31% to 59% GC, and the GC content of the HCP gene follows that of the host genome (although it is in general higher due to the fact that 23 out its 62 amino acids require GC rich codons) (see Table 1). The correlation between GC content of the HCP gene and the host genome is more visible if only third codon positions are considered (see Fig. 1). A probable scenario is that the HCP gene was inserted between the two tRNA genes in the last common ancestor of the *Prochlorococcus*/marine *Synechococcus* group. Based on microarray experiments, HCP is expressed in *Prochlorococcus marinus* CCMP1986 (S.W. Chisholm, personal communication). The 5'-adjacent region is moderately conserved and the Trp-tRNA gene immediately upstream can be predicted to contain a promoter (Vogel *et al.*, 2003).

Genomic context of the HCP coding gene

Although dot plots (not shown) indicate that synteny is not widely conserved among the 13 genomes, the HCP gene occurs in the same genomic context in all, flanked by Glutamyl tRNA synthetase (gltX) and Asp-tRNA coding genes on its 3' end and by the ribosomal protein L19 (rpl19) and the Trp-tRNA genes on its 5' end (Fig. 2). In the 10 other completely sequenced cyanobacterial genomes, the gltX+Asp-tRNA and L19+Trp-tRNA gene clusters are usually found, but not in close proximity to each other. In several genomes, the L19 and Trp-tRNA genes are followed by a small ORF coding for SecE. However, no significant similarity was found between the SecE and the HCP genes.

HCP might be involved in DNA/RNA interactions

Computational predictions of transmembrane helices, coiled-coils and helix-turn-helix motifs did not return any significant results. However, the protein has a predicted isoelectric point of 11.3, and a high proportion of positively charged amino acids (16%), which suggests that it might be involved in DNA/RNA interactions. For the MIT9313 genome only 47 (2%) of the proteins are more alkaline and have a higher proportion of positively charged amino acids.

Conclusions

It is remarkable that HCP has been under such strong purifying selection that not a single amino acid change has occurred during the divergence of the *Prochlorococcus*/marine *Synechococcus* group (roughly estimated to be 100-200 millions years based on 16S rRNA divergence (Rappe & Giovannoni, 2003)) and

especially so since HCP is of unknown function. Examination of several other groups for which multiple genomes are available shows that the highly conserved core proteins (none were 100% conserved, but several families had 97-99% conserved alignment sites) are usually well-characterized functionally, with homologs outside the examined group (data not shown). The functional role of the HCP in the *Prochlorococcus*/marine *Synechococcus* group therefore invites further, experimental, investigation.

Acknowledgements

We would like to thank R. Thane Papke, Yuri Wolf and Kira Makarova for insightful discussions and suggestions. OZ is supported through a CIHR Postdoctoral Fellowship and is an honorary Killam Postdoctoral Fellow at Dalhousie University.

References:

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

Chisholm SW, Frankel SL, Goericke R, Olson RJ, Palenik B, Waterbury JB, West-Johnsrud L & Zettler ER (1992) *Prochlorococcus marinus* nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll a and b. *Archives of Microbiology* **157**: 297.

DeLong EF, Preston CM, Mincer T *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496-503.

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package). Distributed by the author. *Department of Genetics, University of Washington, Seattle*.

Konstantinidis KT & Tiedje JM (2005a) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* **102**: 2567-2572.

Konstantinidis KT & Tiedje JM (2005b) Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187**: 6258-6264.

Rappe MS & Giovannoni SJ (2003) The Uncultured Microbial Majority. *Annual Review of Microbiology* **57**: 369-394.

Rice P, Longden I & Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**: 276-277.

Rocap G, Larimer FW, Lamerdin J *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042.

Schmidt HA, Strimmer K, Vingron M & von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**: 502-504.

Thompson JD, Higgins DG & Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence

weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673-4680.

Vogel J, Axmann IM, Herzel H & Hess WR (2003) Experimental and computational analysis of transcriptional start sites in the cyanobacterium *Prochlorococcus* MED4. *Nucleic Acids Res* **31**: 2890-2899.

Waterbury J, Watson S, Valois F & Franks D (1986) Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Can Bull Fish Aquat Sci* **214**: 71-120.

Table 1. GI numbers for the HCP in completely sequenced genomes*.

Genome	GI number	GC content of HCP, %	GC content of the genome, %
Prochlorococcus marinus str. MIT 9312	78778859	42.6	31.2
Prochlorococcus marinus str. MIT 9313	33863575	55.0	50.7
Prochlorococcus marinus str. MIT 9303	124022390	55.0	50.0
Prochlorococcus marinus str. MIT 9515	123965773	45.5	30.8
Prochlorococcus marinus str. AS9601	123968066	43.1	31.3
Prochlorococcus marinus str. NATL1A	124025241	49.2	35.0
Prochlorococcus marinus str. NATL2A	72383643	48.7	35.1
Prochlorococcus marinus subsp. *marinus* str. CCMP1375	33239923	45.5	36.4
Prochlorococcus marinus subsp. *pastoris* str. CCMP1986	33861031	43.9	30.8
Synechococcus sp. WH 8102	33866357	60.8	59.4
Synechococcus sp. CC9605	78212192	58.7	59.2
Synechococcus sp. CC9902	78185284	58.7	54.2
Synechococcus sp. CC9311	113953642	52.4	52.4

*The HCP was also found in the following available unfinished genomes:

Synechococcus sp. WH7805, *Synechococcus* sp. WH5701, *Synechococcus* sp. RS9917, *Synechococcus* sp. RS9916, *Synechococcus* sp. BL107, *Prochlorococcus marinus* str. MIT9211.

Figure legends:

Fig. 1. Comparison of GC content at the third codon positions in the HCP and in each genome. GC3 values are given as a proportion of G+C nucleotides. The error bars correspond to the 95% confidence interval calculated assuming a binomial distribution for the GC3 values. The graph indicates that GC3 values of the HCP and the host genome are equal within the sampling error.

Fig. 2. A. Genomic context of the HCP gene in 13 *Prochlorococcus* and marine *Synechococcus* genomes. The unconserved region downstream from the HCP gene varies from 149 to 302 nucleotides in length, and in CC9311 (the genome with the largest, 302 nucleotide region) contains a hypothetical ORF that codes for a 54 a.a. protein (locus tag sync_2165). The latter ORF does not have matches anywhere in databases (including other marine *Synechococcus*/*Prochlorococcus* genomes) and perhaps is decayed in other genomes with smaller downstream regions. B. Genomic context of genes adjacent to the HCP gene in three selected cyanobacterial genomes. See text for more details.

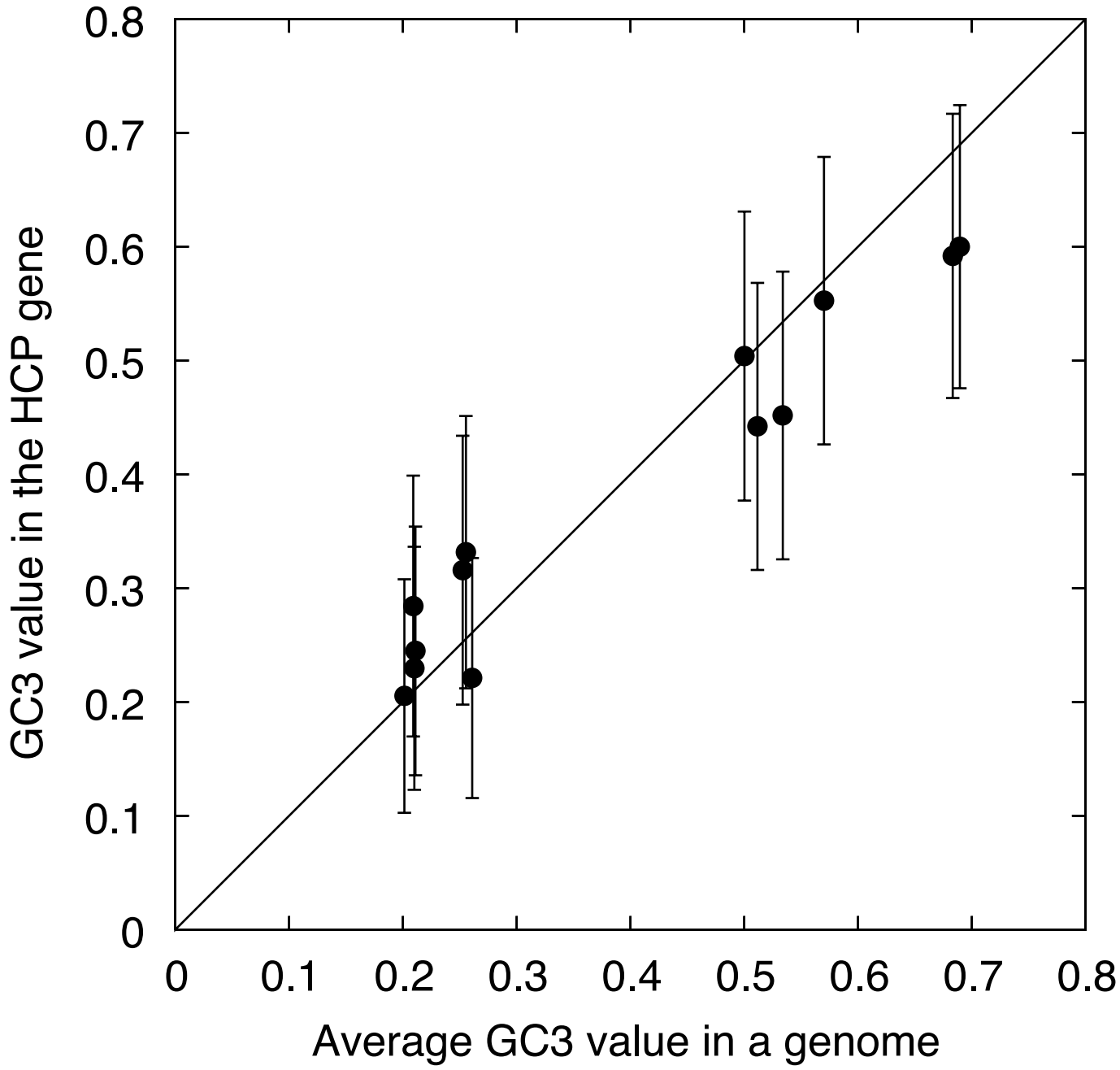
GC3 value in the HCP gene
Average GC3 value in a genome
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Figure 1

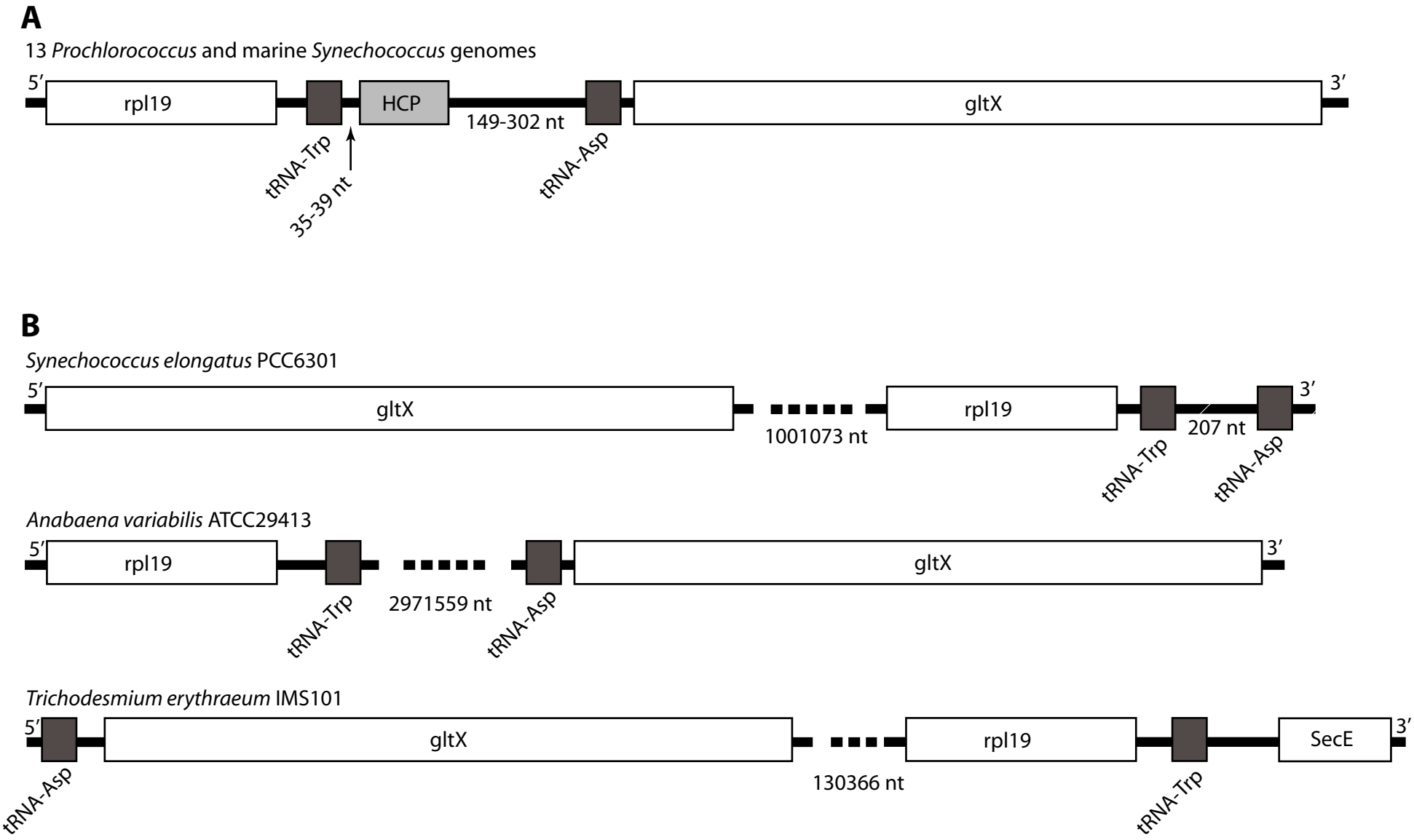
A
13 Prochlorococcus and marine Synechococcus genomes
5′
rpl19
HCP
149-302 nt
gltX
3′
tRNA-Trp
35-39 nt
tRNA-Asp
B
Synechococcus elongatus PCC6301
5′
gltX
1001073 nt
rpl19
207 nt
3′
tRNA-Trp
tRNA-Asp
Anabaena variabilis ATCC29413
5′
rpl19
2971559 nt
gltX
3′
tRNA-Trp
tRNA-Asp
Trichodesmium erythraeum IMS101
5′
gltX
130366 nt
rpl19
SecE
3′
tRNA-Asp
tRNA-Trp

Figure 2