The ISME Journal (2014) 8, 2056–2068

www.nature.com/ismej

ORIGINAL ARTICLE

Comparative transcriptomics of two environmentally relevant cyanobacteria reveals unexpected transcriptome diversity

Karsten Voigt[1], Cynthia M Sharma[2], Jan Mitschke[1], S Joke Lambrecht[1], Björn Voß[1], Wolfgang R Hess[1] and Claudia Steglich[1]
[1]*Faculty of Biology, University of Freiburg, Freiburg, Germany and* [2]*Research Center for Infectious Diseases (ZINF), University of Würzburg, Würzburg, Germany*

Prochlorococcus **is a genus of abundant and ecologically important marine cyanobacteria. Here, we present a comprehensive comparison of the structure and composition of the transcriptomes of two** ***Prochlorococcus*** **strains, which, despite their similarities, have adapted their gene pool to specific environmental constraints. We present genome-wide maps of transcriptional start sites (TSS) for both organisms, which are representatives of the two most diverse clades within the two major ecotypes adapted to high- and low-light conditions, respectively. Our data suggest antisense transcription for three-quarters of all genes, which is substantially more than that observed in other bacteria. We discovered hundreds of TSS within genes, most notably within 16 of the 29 prochlorosin genes, in strain MIT9313. A direct comparison revealed very little conservation in the location of TSS and the nature of non-coding transcripts between both strains. We detected extremely short 5′ untranslated regions with a median length of only 27 and 29 nt for MED4 and MIT9313, respectively, and for 8% of all protein-coding genes the median distance to the start codon is only 10 nt or even shorter. These findings and the absence of an obvious Shine–Dalgarno motif suggest that leaderless translation and ribosomal protein S1-dependent translation constitute alternative mechanisms for translation initiation in** ***Prochlorococcus*****. We conclude that genome-wide antisense transcription is a major component of the transcriptional output from these relatively small genomes and that a hitherto unrecognized high degree of complexity and variability of gene expression exists in their transcriptional architecture.**
The ISME Journal (2014) **8,** 2056–2068; doi:10.1038/ismej.2014.57; published online 17 April 2014
Subject Category: Integrated genomics and post-genomics approaches in microbial ecology
Keywords: antisense RNA; cyanobacteria; dRNA-seq; *Prochlorococcus*; transcriptomics; prochlorosin

Introduction

Prochlorococcus is a marine and unicellular cyanobacterium that populates the oligotrophic open oceans between 40°N and 40°S (Partensky *et al.,* 1999). In these areas, *Prochlorococcus* numerically dominates the phytoplankton with up to 5×10^5 cells per ml, contributing a significant fraction of the photosynthetic biomass (Goericke and Welschmeyer, 1993;Vaulot *et al*., 1995). The oceanic ecosystem is strongly affected by the interplay between *Prochlorococcus*, cyanophage that infect *Prochlorococcus* and alphaproteobacteria of the SAR11 clade. Whereas the latter is dependent on organic matter produced by primary producers such as *Prochlorococcus*, cyanophage contribute to this interaction by lysing *Prochlorococcus*, hence making the organic matter available (Thompson *et al*., 2013). Two distinct ecotypes have been defined according to their adaptation to high-light (HL; for example, strain MED4) or low-light regimes (LL; for example, strain MIT9313) (Moore *et al.,* 1995). However, other environmental factors, such as temperature, nutrients and competitor abundance, also affect the distribution of ecotypes (Johnson *et al.,* 2006). The HL and LL ecotypes can be further divided into six distinct subclades, two within the high-light ecotype (HLI and HLII) and four within the low-light ecotype (LLI-LLIV) (Kettler *et al*., 2007). Within this classification, MED4 and MIT9313 represent the two most distantly related *Prochlorococcus* clades (HLI vs LLIV). Although their 16S rRNA is 97.5% identical, their protein-coding potential differs, with 1955 annotated protein-coding genes for MED4, 2843 for MIT9313 and a shared number of only 1447 genes (as determined by BLAST search using an e-value of 10^{-8}). The suite of strain-specific open reading

Correspondence: C Steglich, Faculty of Biology, University of Freiburg, Freiburg D-79104, Germany.
E-mail: claudia.steglich@biologie.uni-freiburg.de
Received 10 December 2013; revised 3 March 2014; accepted 5 March 2014; published online 17 April 2014

frames (ORFs) includes genes encoding photosynthetic proteins and also genes that are involved in nutrient uptake, assimilation and metabolic functions (Rocap *et al.*, 2003; Kettler *et al.*, 2007). This suggests that there could be ecotype-specific regulatory elements that specifically control these functions and therefore can be expected to differ between the two strains.

The genomes of 12 isolates of *Prochlorococcus* have been completely sequenced, revealing a trend toward compact and streamlined genomes (Dufresne *et al.,* 2003; Rocap *et al.,* 2003; Kettler *et al.,* 2007). Such genomic streamlining has been recognized to be important in many other open ocean marine bacteria, such as representatives of the SAR11 clade of alphaproteobacteria (Grote *et al.,* 2012), and has been interpreted as a typical adaptation of the marine bacterioplankton to oligotrophy (Swan *et al.*, 2013). Due to this streamlining, *Prochlorococcus* genomes are depleted in guanine and cytosine residues, are densely packed with small intergenic regions, and contain only a few genes encoding proteins involved in transcription, signal transduction and the regulation of gene expression (Dufresne *et al.,* 2003; Rocap *et al.,* 2003; Kettler *et al.,* 2007). Such genomic features have also been recognized for many other bacterioplankton species (Swan *et al.*, 2013).

Despite their densely packed genomes and relatively low number of transcriptional protein regulators (only five- and eight-sigma factors, six and seven response regulators as well as six and 10 histidine kinases in MED4 and MIT9313, respectively (Scanlan *et al.,* 2009)), *Prochlorococcus* is capable of adapting to environmental perturbations. The identification of relatively high numbers of non-coding RNAs (ncRNAs) in MED4 (17 small RNAs and 24 antisense RNAs (asRNAs)) suggested that regulatory RNAs may play an important role in the regulation of gene expression in *Prochlorococcus* (Steglich *et al.,* 2008). For the ncRNA Yfr1, a function in the control of major outer-membrane proteins PMM1119 and PMM121 was shown (Richter *et al.,* 2010). More recently, antisense transcripts were detected for 73% of all genes in MED4 at some point in time over the diel cycle (Waldbauer *et al.*, 2012). This percentage is substantially higher than the previously reported fraction of genes associated with asRNA molecules in any bacterium, which was at 46% the highest in *Helicobacter pylori* and less in other bacteria (Georg *et al.,* 2009). Moreover, the abundance of asRNAs was reported to be relatively high in MED4 with an average of 35% of the corresponding mRNA concentration (Waldbauer *et al.*, 2012). However, despite the wealth of available genome information, insight into the transcriptional architecture and the numbers and types of potentially regulatory RNA molecules remains largely fragmentary and limited to MED4.

On the basis of high-density microarray hybridizations, several studies have provided valuable transcriptomic information on genes of MED4 and to some extent on genes of MIT9313 that are differentially expressed under environmentally relevant perturbations such as light stress, phosphorus, nitrogen and iron starvation and phage infection (Martiny *et al.,* 2006; Steglich *et al.,* 2006; Tolonen *et al.,* 2006; Lindell *et al.,* 2007; Thompson *et al.,* 2011) and hence are involved in the adaptation to those conditions. However, except for the experimental determination of 25 transcriptional start sites (TSS) in MED4 (Vogel *et al.,* 2003) and the mapping of the TSS of *cpeB* (Steglich *et al.,* 2005), *psbA*, *psbC*, *psbD* and *pcbA* genes (Garczarek *et al.,* 2001), there is no information on promoters that are actually utilized in *Prochlorococcus.* The combination of our study with previous microarray analyses will promote the targeted search for new regulatory promoter elements in *Prochlorococcus*, which so far is restricted on the knowledge of the NtcA regulon (Tolonen *et al.,* 2006) and computational modeling studies of the cyanobacterial Pho, LexA and cAMP receptor protein regulons (Su *et al.*, 2007; Li *et al.*, 2010; Xu and Su, 2009).

Here, we present a comparative analysis of the transcriptomes of two *Prochlorococcus* strains representing the two most diverse clades within the two major ecotypes adapted to high- and low-light conditions, respectively. We used three different high-throughput methodologies (454 and Solexa sequencing platforms) in combination with a differential RNA-seq approach selective for the analysis of primary transcriptomes and global mapping of TSS (Sharma *et al.,* 2010), as well as Affymetrix high-density microarrays for the independent verification of antisense transcripts. We complemented these with computational and experimental validation of the selected genes. We present genome-wide TSS maps for both strains at a single nucleotide resolution that enables the interrogation of transcriptional activity in a comparative fashion. We found a high degree of antisense transcription and identified new ncRNAs. With the exception of photosynthesis-related genes, the TSS of protein-coding genes and asRNAs are not conserved, suggesting a high degree of variability in their transcriptional architecture.

Materials and methods

Culture growth conditions, RNA isolation and northern blot analysis

Cells were grown at 22 °C in AMP1 medium (Moore *et al.,* 2007) under 10–30 μmol quanta $m^{-2}s^{-1}$ continuous white cool light and harvested in an exponential growth phase. For microarray analysis and northern verifications, the cells were subjected to several stress conditions for 30 min: light stress (light shifts from 10 to 100 μE or darkness, respectively, or from 30 to 300 μE) and temperature stress (shifts from 22 to 12 or 32 °C, respectively). For nitrogen and iron starvation, the cells were washed twice in nitrogen- or iron-free medium and grown in

minus N or minus Fe medium for 2, 3 or 6 days. Alternatively, iron depletion was induced via the addition of 0.6 μM DFB for 24–48 h. Total RNA was extracted from the cells via filtration following the hot phenol method (Steglich *et al.,* 2006) or a modified protocol using PGTX buffer, which consists of 4.2 M phenol amended with 6.9% v/v glycerol, 5 mM 8-hydroxyquinoline, 15.6 mM Na-EDTA, 0.1 M sodium acetate, 0.8 M guanidine thiocyanate and 0.48 M guanidine hydrochloride (Pinto *et al.,* 2009). The MIT9313 cells were lysed prior to RNA extraction using a cell disruption device (Precellys, PeqLab, Erlangen, Germany) applying 6 cycles at power level of 6.5 for 20 s. Northern hybridizations were performed as described (Stazic *et al.,* 2011).

454 and Solexa dRNA sequencing and computational analysis

Details for the 454 differential RNA-seq approach protocol are described in Sharma *et al.* (2010) and Mitschke *et al.* (2011). For both strains, a (+) cDNA library synthesized from a primary RNA pool of exponentially grown cells (total RNA treated with Terminator 5′ P-dependent exonuclease, Epicenter, San Diego, CA, USA) and a (–) cDNA library synthesized from total RNA were generated. To produce the RNA 5′-monophosphates necessary for RNA linker ligation, the samples were treated with tobacco acid pyrophosphatase (after exonuclease treatment if applied). Each RNA sample was ligated to an RNA oligonucleotide containing a unique sequence tag (see Supplementary Table S1). cDNA libraries were sequenced on a Roche FLX sequencer (Basel, Switzerland) and the resulting data were analyzed as described (Sittka *et al.,* 2008). For MED4 (NCBI accession number BX548174.1), a total of 60 457 and 61 633 sequence reads and for MIT9313 (NCBI accession number BX548175.1) a total of 104 037 and 113 994 sequence reads were obtained for the (−) and (+) libraries, respectively. From these, 47 339 and 46 555 sequence reads for MED4 as well as 72 988 and 74 389 sequence reads for MIT913 were ⩾18 nt in length. After filtering out ribosomal RNA matches, 40 291 and 40 739 reads remained in the MED (−) and (+) libraries, respectively, and 53 287 and 59 761 reads remained in the MIT9313 (−) and (+) libraries, respectively.

TSS were determined from (+) libraries following the same work flow as described in Mitschke *et al.* (2011). An individual position-specific scoring matrix for the −10 element of MED4 or MIT9313 was derived, taking the putative TSS of all expressed protein-coding genes of the respective transcriptome into account (Supplementary Tables S2 and S3). Subsequently, a minimum −10 element score of 2.0 followed by at least two sequences was set for the TSS minimum threshold. TSS were classified into four groups: gTSS (start sites of annotated protein-coding genes within a range of 0 to 100 nt upstream of the ORF), iTSS (start sites within annotated genes), aTSS (start sites opposite to annotated genes or within 30 nt of its 5′ and 3′ untranslated regions (UTR) giving rise to asRNAs) or nTSS (all remaining TSS giving rise to potential ncRNAs) (Supplementary Tables S4 and S5). For classification, gTSS were prioritized over aTSS and iTSS.

cDNA libraries for Solexa sequencing were prepared in the same manner as for 454 sequencing with one modification: the RNA for the (−) cDNA library preparation was neither treated with terminator-dependent exonuclease nor treated with tobacco acid pyrophosphatase, resulting in a (−) cDNA library entirely depleted of the primary RNA pool (because triphosphate 5′-termini cannot be ligated to the RNA linker oligonucleotide). For cDNA synthesis and the amplification, Solexa-specific TrueSeq sequencing primers with a unique sequence tag for each library were used. The cDNA libraries were analyzed on an Illumina GA IIx sequencer (San Diego, CA, USA). Sequence data have been deposited in NCBI's Sequence Read Archive under accession number SRR1045147 for MED4 and SRR1045146 for MIT9313. Sequence lengths of 72 nt were obtained for both MED4 libraries and for the (−) MIT9313 library, and sequence lengths of 26 nt were obtained for the (+) MIT9313 library. For MED4 totals of 8 278 893 and 6 324 321 sequence reads, and for MIT9313 totals of 10 689 054 and 8 245 154 sequence reads, were obtained for the (−) and (+) libraries, respectively. The sequences were mapped to the respective genome with *segemehl* (Hoffmann *et al.,* 2009), resulting in 7 429 557 and 5 750 991 redundant read mappings for MED4 and 14 945 219 and 4 663 103 redundant read mappings for MIT9313. After filtering out ribosomal RNA matches, 6 389 217 and 6 389 217 reads remained in the MED4 (−) and (+) libraries and 4 251 797 and 3 423 681 reads remained in the MIT9313 (−) and (+) libraries. The (+) libraries of MED4 and MIT9313 were normalized to a read value per million bp and million Solexa reads. A minimum −10 element score of 2.0 followed by at least 27 reads for MED4 and 25 reads for MIT9313 starting within a window of five to seven nucleotides (three sequence reads each after normalization) were set as minimum thresholds for the definition of a TSS. The same criteria as above for the 454 data sets were applied to classify TSS into gTSS, iTSS, aTSS or nTSS. TSS with the same −10 element were clustered to one entry. The entries were subsequently clustered when −10 elements were located within a 6-nt window (Supplementary Tables S4 and S5).

Prediction of transcriptional units

Transcriptional units are genomic segments with uniform read coverage from the (−) cDNA library. Segmentation of the data into transcriptional units with RNASeg (to be published elsewhere) was performed based on the (−) cDNA library and the

previously defined gTSS, aTSS and nTSS. For this, primary read starts were normalized to 1 for starts and to 0 for non-starts to assure that only previously defined TSS were used. A maximum segment length of 7000 nt and 10 000 nt was defined for MED4 and MIT9313, respectively. If the distance between two subsequent TSS exceeded this length, this distance was used for the corresponding genomic region. The minimum segment length was set to 45 for both strains to cover small transcripts. For transcript segments, the maximum distance to a TSS was set to 10 for both strains and the mean secondary read coverage was set to 1 for MED4 and to 0.5 for MIT9313. Segments with lower coverage were defined as non-transcripts. If the transcript segment exceeded 70% coverage of an already classified region, the segment was assigned to the appropriate class.

Determination of conserved TSS and functional enrichment

TSS were considered as conserved between MED4 and MIT9313 if their distance from the respective start codon differed no more than 10 nt (gTSS) or if their relative distance with respect to a protein alignment was not more than 10 nt (iTSS and aTSS).

Conserved TSS between MED4 and MIT9313 of each TSS class were inspected for functional enrichment using the DAVID web interface (http://david.abcc.ncifcrf.gov/). Functional classes with an enrichment score above 1.3, corresponding to a *P*-value below 0.05, were considered as true enriched functional categories.

Microarray labeling, hybridization, normalization and segmentation

For the detection of asRNA expression signals on microarrays, RNA was first treated with TURBO DNase (Life Technologies, Carlsbad, CA, USA). In total, 6 units of DNase were added to the RNA samples and digestion of DNA was carried out in three consecutive incubation steps, each at 37 °C for 10 min. RNA was either directly labeled (without cDNA synthesis) with the Kreatech (Amsterdam, The Netherlands) 'ULS labeling kit for Affymetrix arrays' with Cy3 according to the manufacturer's protocol or previously depleted of rRNA using Ambion's (Carlsbad, CA, USA) MICROBExpress kit. Of note, probes for the Affymetrix gene expression arrays were designed for hybridization with cDNA. Therefore, direct hybridization with RNA results in signals for the respective asRNA strand. Fragmentation was performed at 70 °C for 15 min using fragmentation buffer (Ambion). Hybridization and scanning were performed according to Affymetrix protocols for *E. coli* (http://www.affymetrix.com/support/technical/manual/expression_manual.affx and Steglich *et al.,* 2006) using 2.5 μg of total RNA (or 1 μg rRNA depleted RNA) on an Affymetrix high density array MD4-9313, which contains probes for *Prochlorococcus* MED4 and MIT9313. The custom array covers all gene-coding regions with a probe pair (match and mismatch) every 80 bases and every 45 bases in the intergenic regions in both the sense and antisense orientations. For both strains, one microarray each was hybridized with RNA extracted from cells grown under standard conditions. A second microarray was hybridized with pooled RNA of cells subjected to different stress conditions (light and temperature stress, nitrogen and iron starvation). Microarray data have been deposited in NCBI's Gene Expression Omnibus under accession number GSE17075. Microarray expression data of single probes were quantile-normalized using the R package LIMMA (Smyth, 2005). The expression threshold value was individually evaluated for MED4 and MIT9313 and was set to 250 and 400, respectively. Consecutively expressed probes (in at least one condition, stress or standard) were combined into segments allowing one probe below the threshold within the segment if the following probe showed expression above the set threshold.

Verification of potential ncRNAs via secondary structure prediction using RANDfold

For the computational prediction of ncRNAs, we extracted candidate sequences based on nTSS reads and analyzed their thermodynamic structural stability with RANDfold (Bonnet *et al.*, 2004). We used two different approaches to identify promising ncRNA candidates: a sliding window approach and an expanding window approach. In the expanding window approach, all possible lengths between 30 and 200 nt were folded one by one; in the sliding approach, windows of 100 nt beginning at the start of the predicted nTSS were shuffled in a range of 0–300 nt in 10-nt increments. For both approaches, *P*-values below 0.05 were counted and summarized in a hit list (Supplementary Tables S9 and S10).

Results and discussion

The primary transcriptomes of two Prochlorococcus strains are highly diverse

To characterize the transcriptomes of the two *Prochlorococcus* strains, total RNA samples of MED4 and MIT9313 cultures grown under standard conditions were used to generate two strand-specific cDNA libraries that allow for a differential sequencing of primary transcripts and processed transcripts. In bacteria, most primary transcripts carry a triphosphate at their 5′ ends resulting from initiation of transcription, whereas processed or degraded RNA fragments possess a 5′ mono-phosphate or 5′ hydroxyl. These differences were employed here by synthesizing two cDNA libraries for each strain: one from the original, untreated RNA pool containing both primary and processed transcripts ((−) cDNA library), and one from RNA

that was enriched in primary transcripts by selective degradation of RNAs containing mono-phosphates using Terminator 5′ P-dependent exonuclease ((+) cDNA library) (Sharma *et al.*, 2010; Mitschke *et al.*, 2011). After sequencing of these differential libraries of both strains, reads were mapped to the genomes of MED4 and MIT9313 (Table 1), and two separate position-specific scoring matrixes (Supplementary Tables S2 and S3) were generated for the −10 elements of both strains. These position-specific scoring matrixes were based on the nucleotides at positions −7 to −12 of all mapped gTSS with a minimum of 11 reads (9927 5′ ends for MED4 and 13 626 for MIT9313). Subsequently, a position-specific scoring matrix minimum score of 2.0 for the −10 element served as a filter criterion for the identification of TSS. For the same sequence motif of the −10 element, different score values were obtained for MED4 and MIT9313 due to the difference in GC content (30.8% for the MED4 genome and 47.5% for MIT9313). Of the 26 previously characterized gTSS (TSS of protein-coding genes) (Vogel *et al.,* 2003; Steglich *et al.,* 2005), 18 were identified directly. Two of the eight missing gTSS had a score below 2.0 for the new −10 element, and four failed the read number criterion, indicating that there is an even higher number of active TSS in MED4. Consistent with previous results (Vogel *et al.,* 2003), the targeted search for the −35 promoter element 'TTGACA' yielded a very small subset of 3.3% (MED4) or 2.1% (MIT9313) TSS with a conserved *E. coli*-like −35 element.

In total, 4126 and 8587 TSS were defined for MED4 and MIT9313, resulting from Solexa and 454 sequencing data that were further classified into gTSS (protein-coding genes), aTSS (asRNAs), nTSS (potential ncRNAs) or iTSS (within genes) (for more details, see Material and methods) (Supplementary Tables S4 and S5 and Supplementary Files S1–S6). For approximately half of all protein-coding genes (MED4, 49% and MIT9313, 41%), a gTSS could be assigned. Surprisingly, there was no good correlation between highly expressed genes in MED4 and those in MIT9313 (Supplementary Table S6). Highly expressed TSS are dominated by gTSS, followed by nTSS, including transfer-messenger RNA— the only nTSS shared between both strains among the top 20—and the newly verified ncRNAs Yfr23 (MED4), Yfr102 and Yfr103 (MIT9313) (Supplementary Table S6). For more details, see the section on ncRNAs. The only gTSS in the top 20 list that occurs in both strains is driving transcription of the *psbEFLJ* operon, encoding the cytochrome b559 alpha and beta subunits and the L and J proteins of PSII (Supplementary Table S6). Other examples of the top 20 list provide a clear connection of gene expression to the physiology of *Prochlorococcus* – for example, as illustrated by the gTSS for genes such as *pcb* or *hli10* in MED4, which encode the single light harvesting protein and one of the high-light-inducible proteins. Another difference between both transcriptomes is the density of TSS. The median distance between two consecutive TSS is 353 nt and 563 nt for MIT9313 and MED4, respectively. The functional relevance of the higher TSS density in MIT9313 remains unclear.

Conservation of distinct TSS of photosynthesis-related genes

We searched for gTSS conserved between both strains (defined by an identical distance to the translational start site ±10 nt). Of the 1447 MED4-MIT9313 shared protein-coding genes, 436 were

Table 1 Summary of genome and transcriptome information. Sequencing numbers are given for primary libraries only

	MED4		*MIT9313*		*Reference*
Ecotype adapted to	High light		Low light		Moore *et al.,* 1995
Hfq	No		Yes		Axmann *et al.,* 2005
Genome size	1.66 Mbp		2.41 Mbp		Rocap *et al.,* 2003
Annotated protein-coding genes	1974		2.894		Kettler *et al.,* 2007
	454 sequencing	*Solexa sequencing*	*454 sequencing*	*Solexa sequencing*	
Total number of mapped reads	61 633	5 750 991	113 994	4 663 103	This work
Number of reads w/o ribosome	40 739	5 425 895	59 761	3 712 862	
Total number of TSS[a]	1643	3459	3402	7410	
gTSS	608	1059	801	1284	
iTSS	423	658	928	2256	
aTSS	520	1566	1372	3231	
nTSS	92	176	300	639	
aTSS with most reads	asPMED4_06291 (*clpB*, 52 reads)	asPMED4_04671 (*cons. hyp.,* 21 506 reads)	asP9313_16021 (*priA*, 184 reads)	asP9313_06941 (*recG*, 3542 reads)	
gTSS with most reads	PMED4_11782 (*cons. hyp.,* 482 reads)	PMED4_16461 (*groES,* 93 670 reads)	P9313_14931 (*psbD*, 263 reads)	P9313_20011 (*smc*, 124 649 reads)	
nTSS with most reads	Yfr11 (1700 reads)	Yfr16 (1 184 303 reads)	Yfr103 (413 reads)	Yfr2-5 (48 227 reads)	
Filter criteria for TSS mapping	⩾2 reads, score⩾2.0	⩾26 reads, score⩾2.0	⩾2 reads, score⩾2.0	⩾24 reads, score⩾2.0	

Abbreviation: w/o, without.
[a]Total number of TSS after clustering of neighboring TSS that start ±10 nt of each other or have a similar ±3 nt −10 element (as described above) and removal of potential processing sites (primary coverage is lower than secondary coverage).

associated with one or more gTSS. From these, 214 possess a conserved gTSS (Supplementary Table S7). A functional enrichment analysis revealed that photosynthesis-related genes were highly overrepresented in this group (enrichment score 7.54). Among the photosynthesis-related genes, those encoding proteins of both photosystem I (*psaC, psaD, psaE* and *psaF*) and photosystem II (*psbA, psbB, psbE, psbH, psbO* and *psb28*) are found. Other genes in this category encode electron transfer proteins (*petE, petH*) and cytochrome b6/f complex components (*petB*). As many of these genes are located in operons, an even higher number of photosynthesis-related genes belong to this class, illustrated by the conserved gTSS upstream of *psbE*, in fact driving the transcription of *psbEFLJ*. For other functional classes, the enrichment was not as pronounced as for the photosynthesis genes; however, this enrichment had good statistical support. These classes include histidine metabolism, aminoacyl-tRNA biosynthesis, genes encoding subunits of the NAD(P)H-quinone oxidoreductase, ribosomal proteins and terpenoid backbone biosynthesis, which reached enrichment scores between 3.5 and 1.4. Especially for the ribosomal proteins, the number of genes with similar transcript initiation is even higher because many of these genes are organized in huge operons. These data suggest that the gene regulation of many housekeeping processes underlies a conserved regulation.

5′UTRs in Prochlorococcus

The median distance from the gTSS to the start codon of protein-coding genes in *Prochlorococcus* turned out to be very short, with 26 and 28 nt for MED4 and MIT9313 (Figure 1), respectively, compared with 42 nt in *Synechocystis* PCC6803 (Mitschke *et al.,* 2011) or 43 nt in *H. pylori* (Sharma *et al*., 2010). Half of all MED4 (56%) and MIT9313 (53%) protein-coding genes have a 5′UTR length between 10 and 40 nt with a maximum frequency at 17 and 15 nt, respectively (Figure 1). We searched for the Shine–Dalgarno sequence 'RGGRGG' in all MED4 and MIT9313 5′UTRs excluding those shorter than 17 or 15 nt and allowing one mismatch. Of the 905 MED4 5′UTRs and the 1202 MIT9313 5′UTRs, 29 MED4 genes and 156 MIT9313 genes possess a possible Shine–Dalgarno sequence in the -4 to -23 region. This frequency equals that of a randomly chosen 6-nt-long motif. When searching for the complementary 'YCCYCC' sequence, we detected very similar numbers of 23 MED4 and 161 MIT9313 5′UTRs with that motif in the same search region. These findings point to a little if any functional role of a Shine–Dalgarno sequence in *Prochlorococcus*. In *E. coli,* the Shine–Dalgarno sequence is bound by the anti-Shine–Dalgarno sequence of the 16S rRNA, anchoring the 30S ribosomal subunit around the start codon to form an initiation complex that covers the -35 to $+19$ region (Hüttenhofer and Noller, 1994). Although the 16S rRNA sequence of *Prochlorococcus* carries the same anti-SD sequence as described for *E. coli*, the absence of a Shine–Dalgarno sequence and the high number of protein-coding genes with relatively short 5′UTRs suggest alternative modes of translation initiation preferentially used in *Prochlorococcus*. Other mechanisms for translation initiation in *E. coli* such as the binding of ribosomal protein S1 to the 5′UTR in mRNAs are also known (Boni *et al*., 1991) and could be the preferred mechanism for translation initiation used in *Prochlorococcus*. In many cyanobacteria, two homologous ribosomal S1 proteins Rps1a and Rps1b exist (Figure 2). Rps1a proteins are more closely related to their orthologs than to their paralogous Rps1b protein. Furthermore, orthologous Rps1b proteins seem to be more diverse than orthologous Rps1a proteins. The conservation of duplicated S1 proteins among cyanobacteria indicates an important role of Rps1b; however, whether it is also involved in translation remains enigmatic. The two homologs in MED4 and MIT9313 are predicted to be 40.8 and 45.1 kDa (34% sequence identity) and 40.5 and 45.2 kDa (29% sequence identity) in size. Mutsuda and Sugiura (Mutsuda and Sugiura, 2006) showed that the 38 kDa S1 protein (Rps1a) of *Synechococcus elongatus* PCC 6301 is involved in the efficient initiation of translation via the association with pyrimidine-rich sequences, regardless of the presence or absence of the Shine–Dalgarno sequence. *S. elongatus* Rps1a exhibits a sequence identity of 75% to the *Prochlorococcus* homologs (40.8 kDa in MED4 and 40.5 kDa in MIT9313), whereas the identity of *S. elongatus* Rps1b within the central-to-C terminal region is only 34% and 38%, respectively (45.1 kDa in MED4 and 45.2 kDa in MIT9313).

Another mechanism of translation initiation exists for leaderless mRNAs that lack a 5′UTR and directly bind a 70S ribosome complemented with *N*-formyl-methionyl-transfer RNA (Moll *et al.,* 2002). Here we found in both strains ~8% of all gTSS to be located within the first 10 nt to the initiation codon of translation, suggesting that leaderless translation occurs in *Prochlorococcus*. This is in agreement with computational predictions based on 953 bacterial and 72 archaeal genomes, which suggested that leaderless transcription is widespread among bacteria (207 of the 953 genomes) although not dominant (Zheng *et al*., 2011). Transcription starts for 30 MED4 ORFs and 41 MIT9313 ORFs on the first nucleotide of the start codon. For 21 MED4 and 34 MIT9313 genes, this is the only detectable TSS under standard growth conditions. Except for three homologs, leaderless transcripts are not conserved between MED4 and MIT9313 and are distributed evenly over the entire genome. The only exceptions are the *ruvB, degT* and the PMED4_05961/P9313_15281 homologs, encoding the Holliday junction DNA helicase RuvB, a DegT/DnrJ/EryC1/StrS

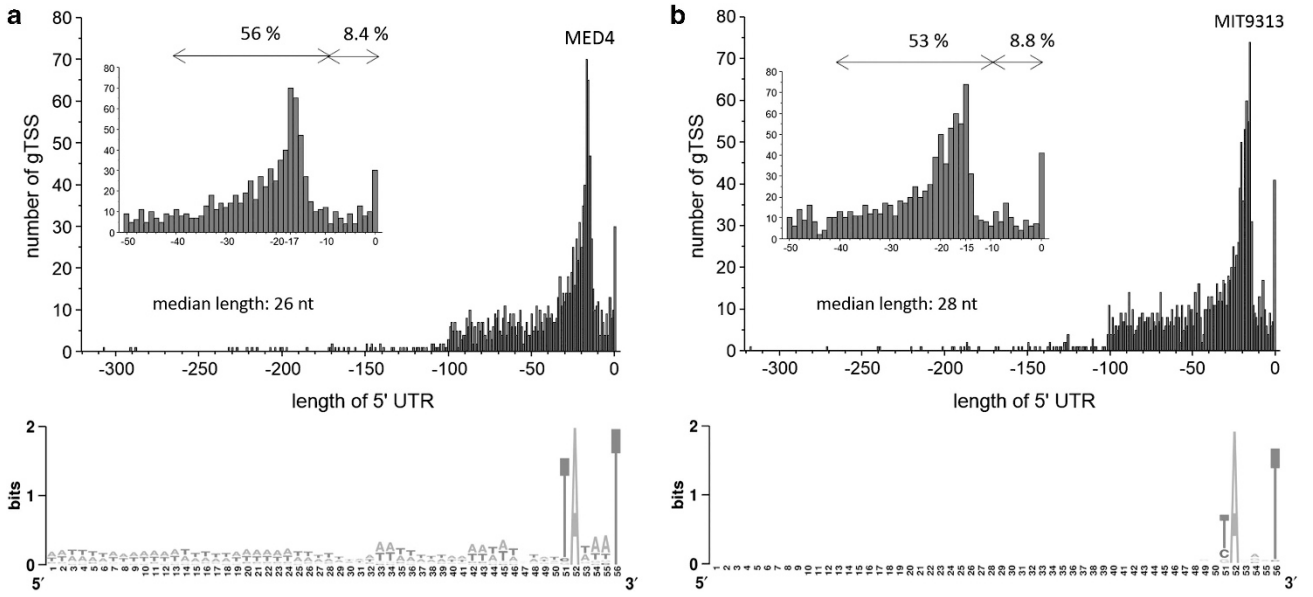

Figure 1 Promoter as well as 5′UTR characteristics of MED4 (**a**) and MIT9313 (**b**) gTSS. Based on 1182 MED4 and 1383 MIT9313 gTSS, a minus −10 element motif 6–8 nt upstream of the transcription initiation site was detected, followed by an upstream periodic AT-stretch signal in MED4. Extremely short 5′UTRs with a median length of 26 nt (MED4) and 28 nt (MIT9313) and the absence of a conserved Shine–Dalgarno sequence suggest alternative modes of action for translation initiation independent of 16S rRNA-mediated ribosome binding. The common 5′UTR length of MED4 is 17 nt and of MIT9313 is 15 nt. The majority of protein-coding genes have a 5′UTR length between 10 and 40 nt.

aminotransferase family protein and the previously undetected carboxysome shell protein CsoS1d (Klein *et al.*, 2009). All of these three proteins are highly conserved throughout the cyanobacterial phylum; therefore, their leaderless translation might be conserved beyond the two strains studied here. Together, our data indicate that S1 protein-mediated and leaderless translation constitute mechanisms for translation initiation, whereas Shine–Dalgarno sequence-dependent translation initiation plays a secondary role (if at all) in *Prochlorococcus*. Zheng *et al.* (2011) determined cyanobacterial-specific sequence signatures in the 5′UTR and suggested that other so far unknown mechanisms of translation initiation must exist, which is in total agreement with our results.

Alternative TSS in prochlorosin genes

Our data show a high number of TSS located within the coding region of protein-coding genes, of which only 14 iTSS (transcripts that initiate within a gene) are conserved among the 377 shared MED4-MIT9313 genes associated with an iTSS. Some of the iTSS are actually gTSS of incorrectly annotated start codons due to automatic ORF calling that usually defines the longest possible ORF as a CDS region. Of the 22 MED4 and 155 MIT9313 iTSS within the first 50 nt of an annotated CDS, which do not possess an additional gTSS, 3 MED4 and 61 MIT9313 iTSS were wrongly classified because of incorrect annotation of the start codon. We have corrected these and reclassified them as gTSS (Supplementary Tables S4 and S5). The iTSS located in the 3′ regions of ORFs that are in the vicinity of a downstream following ORF could give rise to UTRs of the downstream ORF of at least 101 nt. However, there is evidence for a subclass of iTSS that gives rise to shorter, very likely functionally relevant transcripts. Among those are TSS that are located within prochlorosin (*procA*) genes, of which 29 homologs exist in MIT9313 (Li *et al.*, 2010). Prochlorosins are lantipeptides, which are ribosomally translated and posttranslationally modified in the C-terminal core region of the precursor peptide and are distinguishable by the thioether amino acids lanthionine and methyllanthionine (Willey and Van der Donk, 2007). Maturation of the *procA* is completed by the proteolytic removal of the N-terminal leader sequence, which itself is not modified, resulting in an active metabolite (Willey and Van der Donk, 2007). All 29 homologs are expressed under standard growth conditions, although at a low level of expression (Supplementary Table S10). For 11 of the 29 *procA* genes, transcription starts exclusively upstream of the annotated gene, whereas 16 *procA* homologs possess a gTSS and at least one additional iTSS (Supplementary Table S8). Interestingly, the median distance of the iTSS to the C-terminal core peptide is 26 nt, which is very similar to the median 5′UTR length of 28 nt that was determined for all MIT9313 gTSS.

Northern hybridizations with specific probes that target either the 5′UTR plus the N-terminal region of *procA1.4* or the C-terminus including the iTSS gave distinct signals that could be assigned to the full-length transcripts of the precursor peptide and also to transcripts that start at the mapped iTSS and cover the entire core peptide (Figure 3). The functional

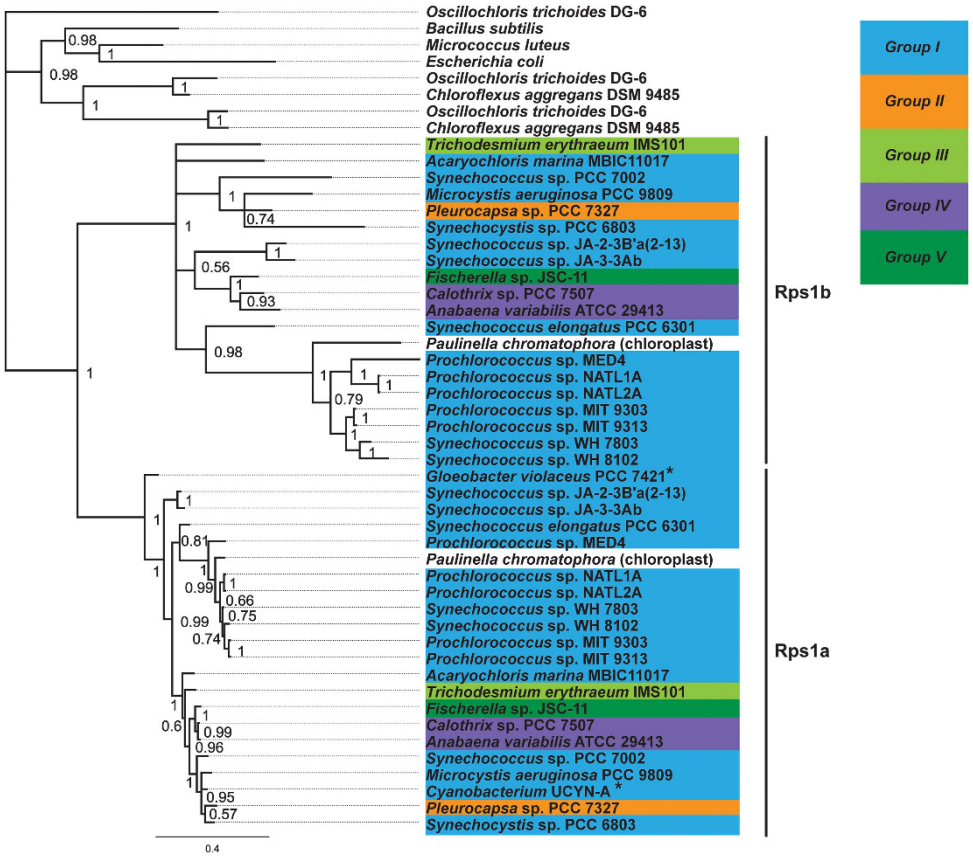

Figure 2 Bayesian inference of phylogeny (Huelsenbeck & Ronquist 2001) for ribosomal S1 proteins in cyanobacteria, gram-positive bacteria and gram-negative bacteria based on the alignment in Supplementary File S7. We used a poisson rate matrix, gamma distributed rate variation, *Oscillochloris trichoides* DG-6 as an outgroup and stopped after the average s.d. of split frequencies of 0.016 was reached. Numbers at the nodes represent the posterior probability. The grouping of cyanobacteria is according to Schirrmeister *et al.*, 2011. Cyanobacteria marked with an asterisk (*Gloeobacter violaceus* PCC 7421 and *Cyanobacterium* UCYN-A) are devoid of the Rps1b homolog.

relevance of these 3′ transcripts remains enigmatic as the core peptide does not start with a methionine or an alternative start codon. There are two *procA* genes—*procA2.3* and *procA4.1*—that are transcribed from an iTSS only. Intriguingly, the core peptide of *procA2.3* starts with a methionine and shows distinct signals of 70-, 90-, 200- and ~400-nt length when probing against the 3′UTR region of the core peptide (Figure 3). The 200-nt fragment is the most abundant of all detected and could encompass the entire core peptide plus a 3′UTR of 80 nt. In agreement with sequencing data, when probing against the 5′UTR region, signals were observed neither under standard growth conditions (Figure 3) nor during adaptation to various stress conditions (data not shown). The transcription of only the *procA2.3* core peptide raises the question of whether a translated core peptide without the leader peptide could be transformed into a bioactive form at all; the common maturation pathway starts with the posttranslational modification of the precursor peptide by the bifunctional lanthionine synthetase and requires at least certain parts of the leader peptide for substrate binding of the enzyme (Xie *et al.,* 2004), or if the core peptide cannot be modified, thus fulfilling another function. If the prochlorosin could be converted into a bioactive compound by an alternative pathway, a completely new tool set for the synthesis of lantipeptides would become available—a group of small peptides that are already routinely used by the food industry as a preservative and in the clinical field for the eradication of infections caused by multi-drug-resistant pathogens (Piper *et al.,* 2009).

High incidence of short cis-encoded asRNAs

The ubiquitous occurrence of *cis*-encoded asRNAs has been reported for various bacteria (reviewed in Georg and Hess, 2011), and every new transcriptome data set published reveals new examples of antisense transcription. However, we were surprised by the high number of genes that were associated with at least one asRNA. In total, 1117 and 1789 genes with an asRNA within the 454 data set as well as 1600 and 2280 genes within the Solexa data set were determined for MED4 and MIT9313, respectively. The shared number of genes with an asRNA in both the Solexa and 454 data sets was 1008 (54% of all genes) and 1625 (57% of all genes) for MED4 and MIT9313, respectively (Figure 3). We furthermore applied an alternative approach to the differential RNA-seq method and hybridized *Prochlorococcus*-specific

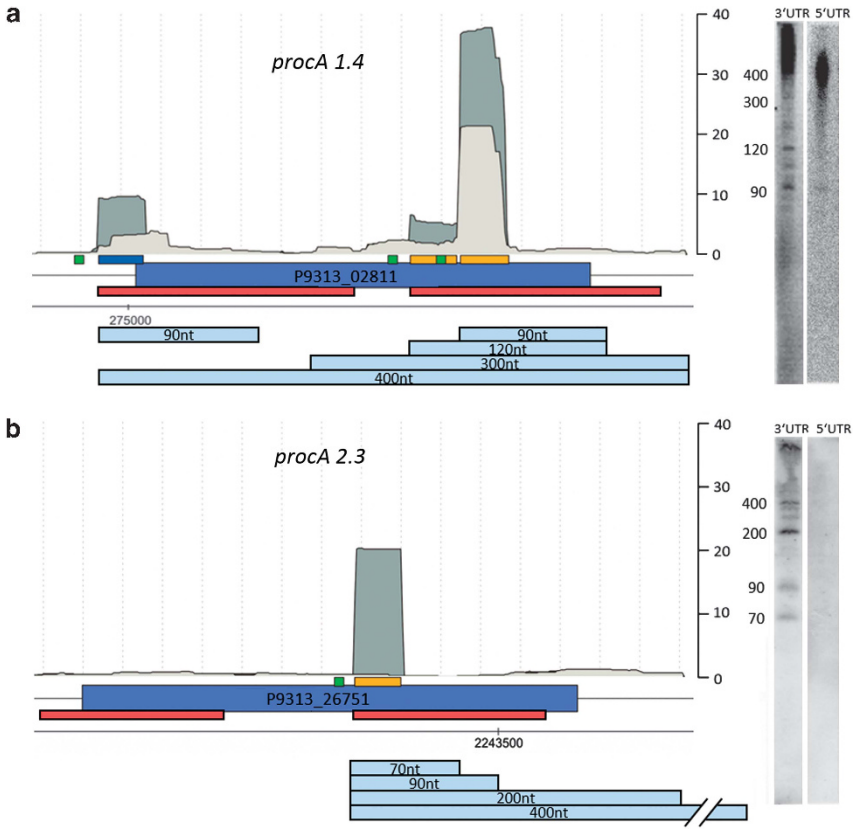

Figure 3 Overview on *procA* genes *procA1.4* (**a**) and *procA2.3* (**b**) read distribution based on mapped Solexa reads of the [+] (dark gray lines) and [−] (light gray lines) cDNA libraries and northern blot hybridizations with probes (red boxes) recognizing either the 5′ region or the 3′ region. Regions of mapped reads that give rise to a TSS are marked with blue (gTSS) and orange (iTSS) boxes. The −10 element is marked with a green box. Transcript lengths as determined by northern hybridization are given as light blue boxes. A total of 10 μg of total RNA was loaded on 10% PAA gels and blotted onto nitrocellulose membranes.

Affymetrix microarrays with directly Cy3-labeled RNA from cultures grown under standard conditions or several stress conditions (for more details see Material and Methods) (Supplementary Files S8 and S9). Because probes of these microarrays were initially designed for hybridization with cDNA, signal intensities above the threshold would provide information on the existence and expression of asRNAs when hybridizing with directly labeled RNA. In total, we found 1875 and 2613 genes with an associated asRNA for MED4 and MIT9313, respectively, confirming a potentially global antisense transcription in *Prochlorococcus*. A high degree of antisense transcription has been previously reported for MED4 (Waldbauer *et al.*, 2012) and appears to be a common feature not only for high-light-adapted ecotypes but also for low-light-adapted strains according to the presented data. Notably, the overlap between the three methods used for the identification of asRNAs was smaller than expected, which can be explained by distinct biases of the different technologies (Harismendy *et al.*, 2009; Raabe *et al.*, 2014). To ensure the validity of our analysis, we considered only those asRNAs that were confirmed by at least two methods (Supplementary Figure S1). As a result, for both MED4 and MIT9313, approximately three-quarters of all genes possess an asRNA.

To obtain a better understanding of the characteristics of asRNAs in *Prochlorococcus*, we determined the average transcript length and the relative abundance of asRNAs in comparison with mRNAs. For that analysis, aTSS and gTSS reads were assembled to longer transcripts using the segmentation algorithm of RNASeg (to be published elsewhere). The stoichiometric mean ratio of mRNA to asRNA reads was 19 and 15, and the average asRNA length was 242 and 408 nt for MIT9313 and MED4, respectively. Of 616 shared genes with an associated asRNA, only 52 asRNAs were conserved with respect to position. Because we noticed a very dense promoter occurrence in general for both MIT9313 (every 353 nt) and MED4 (every 563 nt), these short asRNAs could solely be stable non-functional by-products – for example, in the form of double-stranded RNA.

However, the sheer abundance of asRNAs spurred us to investigate the functional relevance of this class of molecules in more detail. We chose putative asRNAs of MED4 and MIT9313 and tested their possible differential expression under various stress conditions. All tested asRNAs showed an altered

expression behavior in one or more conditions compared with standard growth conditions (Figure 4). *CynS* (PMM0373) encodes a cyanase that converts cyanate to carbon dioxide and ammonia and appears to have a role in cyanate utilization rather than in detoxification in MED4 (Kamennaya and Post, 2011). Intriguingly, under nitrogen-limiting conditions, an asRNA of ~300 nt is induced that may have an important role in the regulation of *cynS* (Figure 4a). *YocE* (PMM1378, *desA*) encodes a fatty acid desaturase, type II, and its expression is induced during low temperature and high light in *Synechococcus* PCC7002 (Sakamoto *et al*., 1997, 1998). Our data suggest a potential *yocE* asRNA-mediated regulation of *yocE* during stationary phase and iron-limiting conditions (Figure 4a). The mRNA of the manganese-binding photosystem II protein Y, encoded by *psbY*, is covered by a 120-nt-long asRNA (Figure 4b). *PsbY* and the gene immediately downstream, *gidA* (PMT1049, containing a NAD-binding domain), appear to be organized in a dicistron (Figure 4b), and the asRNA of *psbY* may enable the decoupling of the gene expression regulation of *psbY* and *gidA*. An ~130-nt asRNA covers the 3′ region of P9313_10131—a gene with homology to the thiol oxidoreductase of *Synechococcus* CC9311 (Figure 4b). The specific decrease in the asRNA abundance under nitrogen starvation suggests its involvement in the gene expression regulation of P9313_10131.

These few examples point to a very complex regulatory network through asRNAs and emphasize the important role of functional RNAs in the gene regulation of *Prochlorococcus*.

Non-coding RNAs are light clade-specific in Prochlorococcus

Gene expression control through ncRNAs constitutes an important regulatory pathway in MED4 (Axmann *et al.,* 2005; Steglich *et al.,* 2008). However, until now, there has been little information on the ncRNA content in other *Prochlorococcus* strains. Of the 20 previously described ncRNAs in MED4 (Axmann *et al.,* 2005; Steglich *et al.,* 2008), 16 were also detected here using the differential RNA-seq approach approach. Additionally, we found six new

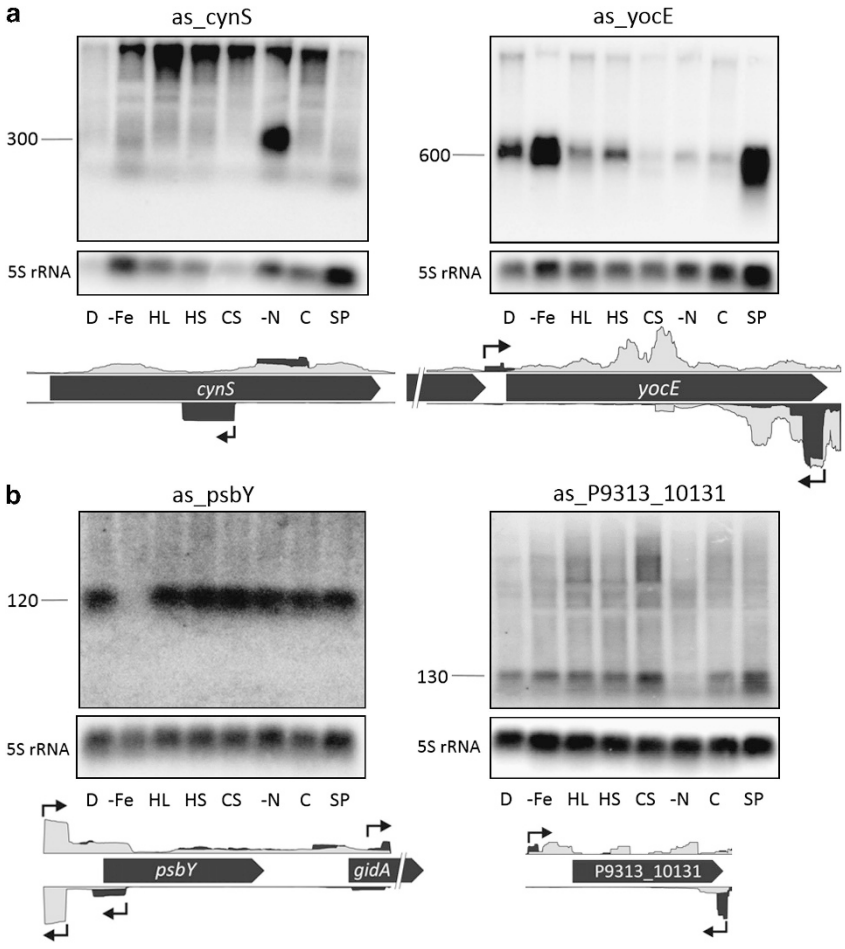

Figure 4 Verification of selected asRNAs of MED4 (**a**) and MIT9313 (**b**), and determination of expressional changes during several stress conditions (D—darkness, -Fe—iron depletion, HL—high light, HS—heat shock, CS—cold shock, -N—nitrogen deprivation, C—standard growth conditions and SP—stationary phase) by northern blots. 5S RNA was used as a loading control. The gene arrangement and coverage with reads of the primary (black) or secondary (gray) library are shown at the bottom. Arrows indicate positions of TSS.

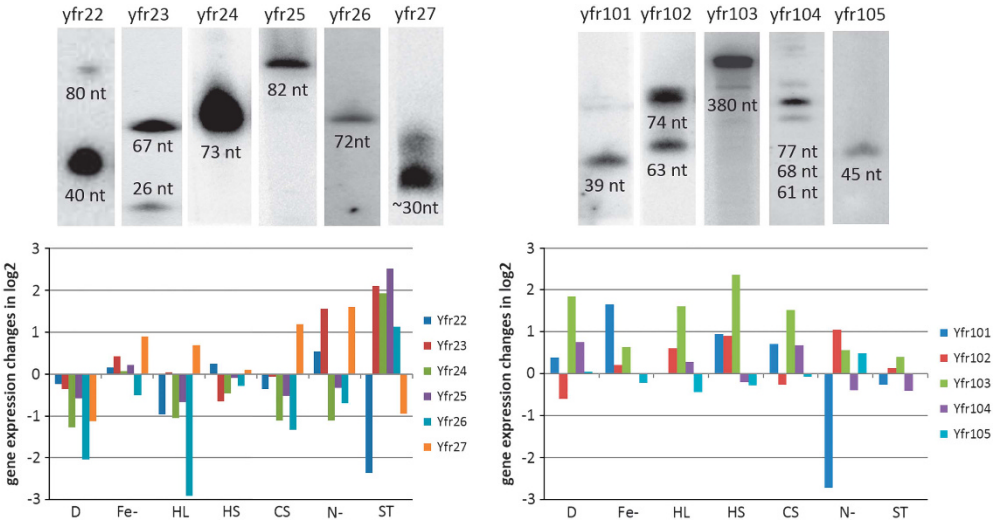

Figure 5 Verification of ncRNA candidates by northern blot hybridization. Different stress conditions have been applied for MED4 and MIT9313 and 5 μg of extracted total RNA was separated on PAA gels and blotted onto nitrocellulose membranes. The size of RNAs was estimated by RNA ladder interpolation and from sequencing data. Fold changes (normalized against the internal standard 5S rRNA) of seven different stresses in comparison with the control hybridization are given in log2.

ncRNAs either through manual inspection of nTSS or via computational prediction using RANDfold (Bonnet *et al.*, 2004) (Figure 5 and Supplementary Figure S2 and Supplementary Table S9). For MIT9313, only five ncRNAs that are all homologs of MED4 ncRNAs have been previously identified. Applying the same search criteria as for MED4, we detected five additional ncRNAs in MIT9313 and some potential ncRNA candidates, however with ambiguous hybridization results (Supplementary Figure S3). From sequencing results (300 nTSS in the 454 data set and 639 in the Solexa data) we can expect more ncRNAs to be detected in MIT9313. The five new identified ncRNAs in MIT9313 only occur (in contrast to the known ncRNAs) in low-light-adapted ecotypes (Figure 5 and Supplementary Figure S4 and Supplementary Table S10). The only exception is Yfr103, which also exists in *Synechococcus* strains CC9311, CC9605 and CC9902. In the same manner, homologs of the newly detected MED4 ncRNAs are restricted to the high-light clade or exclusively occur in MED4—for example, Yfr26 and Yfr27. From this, we conclude that most ncRNAs of *Prochlorococcus* are high-light- and low-light-clade specific, which indicates the functional importance of these regulators for the adaptation to clade-specific niches. It is well documented that ncRNAs are frequently coregulated with the environmental conditions in which they have a role. Therefore, it is consistent that the gross of tested ncRNAs responded during the adaptation to light, nutrient and temperature fluctuations (Figure 5). The strongest response of ncRNA expression was detectable during nitrogen deprivation and the stationary phase. The adaptation to nitrogen-limiting conditions and stationary phase through regulatory circuits involving ncRNAs has been previously reported for other bacteria (Jäger *et al.*, 2009; Fröhlich *et al.*, 2012).

In summary, our data support the previously suggested importance of ncRNAs in the regulatory network of MED4, a view that now can be expanded to the low-light-adapted strain MIT9313 and probably *Prochlorococcus* in general.

Conclusions and possible implications

We discovered a dense promoter activity in both strains and very diverse transcriptome architectures between MED4 and MIT9313. A major fraction of newly identified transcripts is represented by asRNAs, which infers that the transcript pool is nearly twice as complex as the annotated gene content in both MED4 and MIT9313. The accumulation of short asRNAs may be a factor contributing to the reported short global RNA half-life times (2.4 minutes) (Steglich *et al.,* 2010) despite *Prochlorococcus*' slow doubling times of usually one per day (Partensky *et al.,* 1999). Not only are these short asRNAs upon pairing to mRNA perfect substrates for ribonuclease III, which recognizes double-stranded RNA, but the partial overlap of mRNA – asRNA duplexes could also generate additional entry sites in single-stranded regions for ribonuclease E (Stazic *et al.,* 2011), which is the major ribonuclease in regulated mRNA ribolysis. Thus, a high substrate availability for ribonuclease might lead to enhanced RNA turnover rates. The high number of asRNAs compared protein regulators in *Prochlorococcus* could explain the short length of the 5′UTRs; these regions are essential for gene regulation through protein regulators (and trans-acting ncRNAs) but are less important for asRNA-mediated gene regulation.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank Stephanie Hein for providing RNA of stressed MED4 cultures for microarray analyses and Jörg Vogel for helping with 454 sequencing. The research was supported by the DFG (SPP 1258) to CS and WRH and the EU project MaCuMBA (grant agreement no: 311975) to WRH.

References

Axmann IM, Kensche P, Vogel J, Kohl S, Hess WR. (2005). Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol* **6**: R73.

Boni IV, Isaeva DM, Musychenko ML, Tzareva NV. (1991). Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1. *Nucleic Acids Res.* **19**: 155–162.

Bonnet E, Wuyts J, Rouzé P, Van de Peer Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20**: 2911–2917.

Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V *et al.* (2003). Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci USA* **100**: 10020–10025.

Fröhlich KS, Papenfort K, Berger AA, Vogel J. (2012). A conserved RpoS-dependent small RNA controls the synthesis of major porin OmpD. *Nucleic Acids Res* **40**: 3623–3640.

Garczarek L, Partensky F, Irlbacher H, Holtzendorff J, Babin M, Mary I *et al.* (2001). Differential expression of antenna and core genes in Prochlorococcus PCC 9511 (Oxyphotobacteria) grown under a modulated light-dark cycle. *Environ Microbiol* **3**: 168–175.

Georg J, Voss B, Scholz I, Mitschke J, Wilde A, Hess WR. (2009). Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. *Mol Syst Biol* **5**: 1–17.

Georg J, Hess WR. (2011). cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol Biol Rev* **75**: 286–300.

Goericke R, Welschmeyer NA. (1993). The marine prochlorophyte Prochlorococcus contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. *Deep Sea Res* **40**: 2283–2294.

Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ *et al.* (2012). Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *MBio* **3**: pii: e00252–12.

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY *et al.* (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10**: R32.

Hoffmann S, Otto C, Kurtz S, Sharma C, Khaitovich P, Vogel J *et al.* (2009). Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures. *PLoS Comput Biol* **5**: e1000502.

Huelsenbeck JP, Ronquist F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Hüttenhofer A, Noller HF. (1994). Footprinting mRNA-ribosome complexes with chemical probes. *EMBO J* **13**: 3892–3901.

Jäger D, Sharma CM, Thomsen J, Ehlers C, Vogel J, Schmitz RA. (2009). Deep sequencing analysis of the Methanosarcina mazei Gö1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci USA* **106**: 21878–21882.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. (2006). Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737–1740.

Kamennaya NA, Post AF. (2011). Characterization of cyanate metabolism in marine Synechococcus and Prochlorococcus spp. *Appl Environ Microbiol* **77**: 291–301.

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S *et al.* (2007). Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *PLoS Genet* **3**: e231.

Klein MG, Zwart P, Bagby SC, Cai F, Chisholm SW, Heinhorst S *et al.* (2009). Identification and structural analysis of a novel carboxysome shell protein with implications for metabolite transport. *J Mol Biol* **392**: 319–333.

Li B, Sher D, Kelly L, Shi Y, Huang K, Knerr PJ *et al.* (2010). Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proc Natl Acad Sci USA* **107**: 10430–10435.

Li S, Xu M, Su Z. (2010). Computational analysis of LexA regulons in Cyanobacteria. *BMC Genomics* **11**: 527.

Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, Rector T *et al.* (2007). Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449**: 83–86.

Martiny AC, Coleman ML, Chisholm SW. (2006). Phosphate acquisition genes in Prochlorococcus ecotypes: evidence for genome-wide adaptation. *Proc Natl Acad Sci USA* **103**: 12552–12557.

Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, Bantscheff J *et al.* (2011). An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803. *Proc Natl Acad Sci USA* **108**: 2124–2129.

Moll I, Grill S, Gualerzi CO, Bläsi U. (2002). Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Mol Microbiol* **43**: 239–246.

Moore LR, Coe A, Zinser ER, Saito MA, Sullivan MB, Lindell D *et al.* (2007). Culturing the marine cyanobacterium Prochlorococcus. *Limnol Oceanogr Meth* **5**: 353–362.

Moore LR, Goericke R, Chisholm SW. (1995). Comparative physiology of Synechococcus and Prochlorococcus: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Mar Ecol Prog Ser* **116**: 259–275.

Mutsuda M, Sugiura M. (2006). Translation initiation of cyanobacterial rbcS mRNAs requires the 38-kDa ribosomal protein S1 but not the Shine-Dalgarno sequence: development of a cyanobacterial in vitro translation system. *J Biol Chem* **281**: 38314–38321.

Partensky F, Hess WR, Vaulot D. (1999). Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* **63**: 106–127.

Pinto FL, Thapper A, Sontheim W, Lindblad P. (2009). Analysis of current and alternative phenol based RNA extraction methodologies for cyanobacteria. *BMC Mol Biol* **10**: 79.

Piper C, Cotter P, Ross R, Hill C. (2009). Discovery of medically significant lantibiotics. *Curr Drug Discovery Technol* **6**: 1–18.

Raabe CA, Tang T-H, Brosius J, Rozhdestvensky TS. (2014). Biases in small RNA deep sequencing data. *Nucleic Acids Res* **42**: 1414–1426.

Richter AS, Schleberger C, Backofen R, Steglich C. (2010). Seed-based INTARNA prediction combined with GFP-reporter system identifies mRNA targets of the small RNA Yfr1. *Bioinformatics* **26**: 1–5.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al.* (2003). Genome divergence in two Prochlorococcus ecotypes reflects Oceanic niche differentiation. *Nature* **424**: 1042–1047.

Sakamoto T, Higashi S, Wada H, Murata N, Bryant DA. (1997). Low-temperature-induced desaturation of fatty acids and expression of desaturase genes in the cyanobacterium Synechococcus sp. PCC 7002. *FEMS Microbiol Lett* **152**: 313–320.

Sakamoto T, Shen G, Higashi S, Murata N, Bryant DA. (1998). Alteration of low-temperature susceptibility of the cyanobacterium Synechococcus sp. PCC 7002 by genetic manipulation of membrane lipid unsaturation. *Arch Microbiol* **169**: 20–28.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR *et al.* (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73**: 249–299.

Schirrmeister BE, Antonelli A, Bagheri HC. (2011). The origin of multicellularity in cyanobacteria. *BMC Evol Biol* **11**: 45.

Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A *et al.* (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* **464**: 250–255.

Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT *et al.* (2008). Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet* **4**: e1000163.

Smyth G. (2005). limma: Linear models for microarray data. In Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S (eds *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Statistics for Biology and Health, Springer: New York, NY, USA, pp. 397–420.

Stazic D, Lindell D, Steglich C. (2011). Antisense RNA protects mRNA from RNase E degradation by RNA-RNA duplex formation during phage infection. *Nucleic Acids Res* **39**: 4890–4899.

Steglich C, Frankenberg-Dinkel N, Penno S, Hess WR. (2005). A green light-absorbing phycoerythrin is present in the high-light-adapted marine cyanobacterium Prochlorococcus sp. MED4. *Environ Microbiol* **7**: 1611–1618.

Steglich C, Futschik M, Rector T, Steen R, Chisholm SW. (2006). Genome-wide analysis of light sensing in Prochlorococcus. *J Bacteriol* **188**: 7796–7806.

Steglich C, Futschik ME, Lindell D, Voss B, Chisholm SW, Hess WR. (2008). The challenge of regulation in a minimal phototroph: Non-coding RNAs in Prochlorococcus. *PLoS Genet* **4**: e10000173.

Steglich C, Lindell D, Futschik M, Rector T, Steen R, Chisholm SW. (2010). Short RNA half-lives in the slow-growing marine cyanobacterium Prochlorococcus. *Genome Biol* **11**: R54.

Su Z, Olman V, Xu Y. (2007). Computational prediction of Pho regulons in cyanobacteria. *BMC Genomics* **8**: 156.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM *et al.* (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci USA* **110**: 11463–11468.

Thompson AW, Huang K, Saito MA, Chisholm SW. (2011). Transcriptome response of high- and low-light-adapted Prochlorococcus strains to changing iron availability. *ISME J* **5**: 1580–1594.

Thompson LR, Field C, Romanuk T, Ngugi D, Siam R, El Dorry H *et al.* (2013). Patterns of ecological specialization among microbial populations in the Red Sea and diverse oligotrophic marine environments. *Ecol Evol* **3**: 1780–1797.

Tolonen AC, Aach J, Lindell D, Johnson ZI, Rector T, Steen R *et al.* (2006). Global gene expression of Prochlorococcus ecotypes in response to changes in nitrogen availability. *Mol Syst Biol* **2**: 53.

Vaulot D, Marie D, Olson RJ, Chisholm SW. (1995). Growth of Prochlorococcus, a photosynthetic prokaryote, in the equatorial Pacific Ocean. *Science* **268**: 1480–1482.

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

Waldbauer JR, Rodrigue S, Coleman ML, Chisholm SW. (2012). Transcriptome and proteome dynamics of a light-dark synchronized bacterial cell cycle. *PLoS One* **7**: e43432.

Willey JM, Van der Donk WA. (2007). Lantibiotics: Peptides of diverse structure and function. *Ann Rev Microbiol* **61**: 477–501.

Xie L, Miller LM, Chatterjee C, Averin O, Kelleher NL, van der Donk WA. (2004). Lacticin 481: In Vitro reconstitution of lantibiotic synthetase activity. *Science* **303**: 679–681.

Xu M, Su Z. (2009). Computational prediction of cAMP receptor protein (CRP) binding sites in cyanobacterial genomes. *BMC Genomics* **10**: 23.

Zheng X, Hu G-Q, She Z-S, Zhu H. (2011). Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics* **12**: 361.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)