Corresponding Author: Dr Wolfgang R Hess,

Corresponding Author's Institution: Biologisches Institut/Genetik

First Author: Danny Ionescu

Order of Authors: Danny Ionescu; Björn Voss; Aharon Oren; Wolfgang R Hess; Alicia M Muro-Pastor

Abstract: In response to nitrogen deficiency some cyanobacteria develop heterocysts, a terminally differentiated cell type, specialized for the fixation of atmospheric nitrogen. In Nostocales this differentiation process is controlled by two major regulators, NtcA and HetR, but additional unknown factors are likely to be involved as well. In the context of a genome-wide search for potential non-coding RNAs, we identified an array of 12 tandem repeats that is transcribed in high amounts when cells enter conditions that trigger cell differentiation and switch to nitrogen fixation. The main accumulating transcript, which we suggest designating NsiR1, has properties similar to regulatory non-coding RNAs. In Anabaena sp. PCC 7120, it is about 60 nt in length, has a very distinct predicted secondary structure and is expressed very early and transiently after nitrogen step-down. Moreover, its expression requires HetR and NtcA, and is restricted to cells that are differentiating into heterocysts, clearly placing NsiR1 within the regulon that controls the switch to nitrogen fixation and heterocyst formation. The genomic arrangement of NsiR1, located upstream of hetF, a gene whose product is involved in heterocyst formation, is conserved in all five Nostocales whose genomes are completely sequenced. Additionally, we detected NsiR1 expression in 19 different heterocyst-forming cyanobacteria. Our data suggest that every repeat is a complete transcriptional unit furnished with a cell-type specific promoter and a Rho-independent terminator, that gives rise to a very high NsiR1 transcript level. NsiR1 is the first known bacterial non-coding RNA that is specifically upregulated in response to nitrogen step-down.

Heterocyst-specific transcription of NsiR1, a non-coding RNA encoded in a tandem array of direct repeats in cyanobacteria

Danny Ionescu[1,2,#], Björn Voss[3], Aharon Oren[1], Wolfgang R. Hess[3*] and Alicia M. Muro-Pastor[4]

[1]Department of Plant and Environmental Sciences, The Institute of Life Sciences, and the Moshe Shilo Minerva Center for Marine Biogeochemistry, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

[2]The School of Marine Sciences, The Ruppin Academic Center, Michmoret, Israel

[3]University of Freiburg, Faculty of Biology, Genetics and Experimental Bioinformatics, Schänzlestr. 1, D-79104 Freiburg, Germany

[4]Instituto de Bioquímica Vegetal y Fotosíntesis, Consejo Superior de Investigaciones Científicas, E-41092 Seville, Spain

[#]current address: Max-Planck-Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany

*Corresponding author: Tel: +49-761-203-2796; Fax: +49-761-203-2745; email: wolfgang.hess@biologie.uni-freiburg.de

Abbreviations: CRISPR, Clustered Regularly Interspaced Short Palindromic Repeats; TSS, transcription start site.

Abstract

In response to nitrogen deficiency some cyanobacteria develop heterocysts, a terminally differentiated cell type, specialized for the fixation of atmospheric nitrogen. In Nostocales this differentiation process is controlled by two major regulators, NtcA and HetR, but additional unknown factors are likely to be involved as well. In the context of a genome-wide search for potential non-coding RNAs, we identified an array of 12 tandem repeats that is transcribed in high amounts when cells enter conditions that trigger cell differentiation and switch to nitrogen fixation. The main accumulating transcript, which we suggest designating NsiR1, has properties similar to regulatory non-coding RNAs. In *Anabaena* sp. PCC 7120, it is about 60 nt in length, has a very distinct predicted secondary structure and is expressed very early and transiently after nitrogen step-down. Moreover, its expression requires HetR and NtcA, and is restricted to cells that are differentiating into heterocysts, clearly placing NsiR1 within the regulon that controls the switch to nitrogen fixation and heterocyst formation. The genomic arrangement of NsiR1, located upstream of *hetF*, a gene whose product is involved in heterocyst formation, is conserved in all five Nostocales whose genomes are completely sequenced. Additionally, we detected NsiR1 expression in 19 different heterocyst-forming cyanobacteria. Our data suggest that every repeat is a complete transcriptional unit furnished with a cell-type specific promoter and a Rho-independent terminator, that gives rise to a very high NsiR1 transcript level. NsiR1 is the first known bacterial non-coding RNA that is specifically upregulated in response to nitrogen step-down.

Keywords: cell differentiation, cyanobacteria, non-coding RNAs, heterocysts, nitrogen fixation

Introduction

Cyanobacteria are oxygen-producing photosynthetic organisms that are responsible for over half of the global nitrogen fixation in both aquatic and terrestrial environments.[1,2] Nitrogenase, the enzyme that transforms dinitrogen into ammonium, is extremely sensitive to oxygen.[3] Therefore, nitrogen-fixing cyanobacteria spatially or temporally separate nitrogen fixation from oxygenic photosynthesis.[2,4] In some filamentous cyanobacteria such as *Anabaena* sp. PCC 7120 (also known as *Nostoc* sp. PCC 7120; from here on *Anabaena* 7120), nitrogen fixation occurs in heterocysts, specialized cells that create a local microaerobic environment.[5] The developmental program that results in the differentiation of a vegetative cell into a heterocyst requires the sequential activation of a large number of genes involved in regulatory, structural, or enzymatic aspects of heterocyst differentiation and function (recent reviews[6-10]) and depends on two major regulators, NtcA, the global nitrogen control protein[11,12] and HetR[13,14], a positive-acting factor that exhibits protease[15,16] and DNA binding[17] activities *in vitro*. This process is further modulated by a plethora of additional regulators, one of them HetF, controlling the level of HetR,[18] and additional unknown factors are likely to participate in this process.

The genomes of heterocyst-forming cyanobacteria are unusual in the fact that up to 1.5% of their intergenic sequences (7.5% in *Nostoc punctiforme*[19]) comprise tandem repeats that are at least 20 nt in length.[19,20] Tandem repeats are very well studied and occur at a high frequency in eukaryotic non-coding regions,[21] with eukaryotic microsatellites being one of the best-characterized classes. In contrast, tandem repeats have seldom been reported in bacteria[22] and, except for the class of Clustered Regularly

Interspaced Short Palindromic Repeats (CRISPR),[23] have not been functionally characterized. CRISPR-derived non-coding RNAs guide antiviral defense in prokaryotes,[24] illustrating that some bacterial direct repeats are biologically highly relevant.

Processes leading to the differentiation of eukaryotic cells are frequently controlled by regulatory RNA molecules such as miRNAs,[25] whereas not a single such case has yet been reported for bacteria. However, regulatory RNAs in bacteria are well known to control adaptive responses to virtually all possible stress conditions.[26]

Here, we present a widely conserved tandem array of sequence repeats that become expressed upon nitrogen deprivation specifically in cells that are differentiating into heterocysts. These repeats are located upstream of *hetF*, whose product is required for the differentiation of heterocysts in *Anabaena* 7120[18,27] and *Nostoc punctiforme*[28], as well as for the establishment of a functional symbiosis between *Nostoc punctiforme* and the bryophyte *Anthoceros punctatus*.[29] The main transcript originating from the repeats has properties typically associated with regulatory non-coding RNAs and also showed up in a genome-wide search for such elements focusing on possible Rho-independent terminators in intergenic regions. We detected this transcript in 19 different cyanobacteria belonging to the classes Nostocales and Stigonematales, and characterized it in more detail in *Anabaena* 7120.

Results

A genome-wide prediction of possible Rho-independent terminators identifies a distinct set of conserved direct repeats upstream of *hetF*

4

To identify non-coding RNAs that could potentially be involved in the differentiation of heterocysts, a genome-wide prediction of Rho-independent terminators within all intergenic regions of *Anabaena* sp. PCC 7120 was carried out. Such a strategy has previously been productive for the discovery of unannotated transcripts in intergenic spacers of the cyanobacterium *Synechocystis* sp. PCC 6803[30] and resulted herein in the prediction of 678 possible Rho-independent terminators (Supplementary Table 1). Several of these are just transcriptional terminators of annotated protein-coding genes, such as element i1848_1 (positions 4519541-4519575 in the genome sequence), which is located 29-64 nt directly downstream of *psbA* (*alr3742*) and therefore identified as its Rho-independent terminator. Other elements are located further away or in inverse orientation with respect to neighboring annotated genes. One such example is element i18_1 (positions 56491-56529), which is in the same orientation as the two adjacent genes, however, 208 nt downstream from the closest gene. Indeed, element i18_1 is the terminator of previously identified ncRNA Yfr1.[31]

Closer inspection of those predicted elements located proximal to genes involved in heterocyst differentiation, identified a whole series of closely spaced possible terminators situated upstream of the heterocyst differentiation gene *hetF*, but in the reverse orientation (Figure 1). The sequence of these predicted terminators (i1605_2 to i1605_12 in Supplementary Table 1) is, in eight cases, identical. In fact, these elements belong to a series of direct sequence repeats, which are conserved and found at a corresponding position in the genomes of all other sequenced Nostocales (Figure 1). The repeat sequences present in the different cyanobacteria are not identical, with the highest divergence found between *Anabaena* 7120 and *Nostoc azollae*, and exhibit a

higher degree of conservation towards the second half of each repeat (an alignment of all 51 repeat sequences from sequenced filamentous strains is shown in Supplementary Figure S1). For each strain, the peripheral repeats are less conserved than the central ones. In phylogenetic analyses, the central repeats 3-9 exhibit a tendency to cluster species-specifically together, whereas the location of repeats, 1, 10, 11 and 12 is consistent with the phylogenetic relationships among the investigated species (Supplementary Figure S2).

In the genome of *Anabaena* 7120, the repeats start on the reverse strand 225 nt upstream of the annotated reading frame for *hetF*. Repeats 3-9 are 133-134 nt long and are almost identical (Figure 2), whereas repeats 1, 2, and 10-12 are less conserved, as observed in the interspecies comparison. Repeats 1, 2, and 12 each include a short sequence insertion that is not present in the other repeats, and thus they are slightly longer than repeats 3-11 (Figure 2). Due to the proximity to *hetF*, an overlap between the repeats and the *hetF* 5' UTR appeared possible. Indeed, 5' RACE analysis identified the *hetF* transcriptional start site (TSS) at position -403 with respect to the *hetF* start codon. Thus, the *hetF* 5' UTR overlaps repeat 1 and part of repeat 2 (Figures 1 and 2).

The direct repeats upstream of *hetF* are transcribed as a small non-coding RNA and its expression is under the control of NtcA and HetR

The association of the repeats with *hetF* points to a possible role related to heterocyst differentiation. Therefore, expression of the region was determined under a nitrogen step-down time course. Indeed, a rapidly induced and abundant transcript of about 60 nt

was observed upon nitrogen deprivation, with a maximum expression observed 10 - 12 h after nitrogen step-down (Figure 3a), several hours before heterocyst differentiation is complete (20-24 h after N step-down under these conditions). As the expression of this RNA was induced under nitrogen depletion, we designated it as Nitrogen Stress Induced RNA 1 (NsiR1). Two less abundant bands of ~200 and ~53 nt were co-induced with NsiR1. In a longer time course, the expression of NsiR1 and of co-induced bands decreased as heterocyst maturation progressed (Figure 3b). The expression of NsiR1 was further analyzed in mutants of the two major regulators governing responses to nitrogen deficiency in *Anabaena* 7120, the global nitrogen regulator NtcA (strain CSE2[11]) and the cell differentiation regulator HetR (strain 216[32]). Compared to the wild-type, the *ntcA* mutant displayed very weak NsiR1 expression (only visible in overexposed images; Figure 3c), whereas the *hetR* mutant showed no expression at all (Figures 3a and 3c). As similar repeats are detectable in all 5 Nostocales for which a genome sequence is available (Figure 1), we evaluated the expression of NsiR1-like RNAs in additional heterocyst-forming cyanobacteria belonging to the cyanobacterial sections IV and V senso Rippka et al. (1979).[33] In all cases a small RNA in the range of ~53 to ~70 nt was detected (Supplementary Figure S3), suggesting the presence of an NsiR1-like transcript in, altogether, 19 different species.

In the primer extension assays carried out with primer T4 (location of primers in Figure 2) and using total RNA from *Anabaena* 7120, two extension products of 94 nt (less abundant) and 155 nt (more abundant) were obtained (Figure 3d) that fit two previously established criteria for NsiR1 transcripts, namely, they were transiently induced upon nitrogen step-down and were absent in a *hetR* mutant but barely detectable in an *ntcA*

mutant (cf. Figures 3a and 3c). The corresponding 5' ends mapped 72 and 133 nt

upstream from the 3' end of primer T4. This finding indicates that the major 5' end

corresponds to a position located about one full repeat upstream from the 3' end of the

primer. The identification of these two extension products was confirmed using a

different primer (see position of primer 9 in Figure 2). In this case, the sizes of the

products were 100 nt (less abundant) and 161 nt (more abundant), leading to the same

previously identified 5' positions after subtracting the length of the primer (Figure 3d). It

should be taken into account that, because both primer T4 and 9 are complementary to

sequences in all repeats, the primer extension products could correspond, in principle,

to transcripts originating at any of the 12 repeats found in the chromosome of *Anabaena*

7120. Precise determination of TSS for the promoter of one particular repeat requires

the analysis of *gfp* transcripts originating from promoter-probe vectors, as shown below

in Figure 5.

Expression of NsiR1 is specific for (pro)heterocysts

To further characterize expression of NsiR1, a fragment containing all 12 repeats

(except for the last terminator, see white triangle in Figure 2) was cloned under the

copper-inducible promoter of the *petE* gene from *Anabaena* 7120. As a reporter of

transcription, a *gfp* gene, coding for the green fluorescent protein (GFP), was inserted

downstream of the repeats, yielding pCSAM207 (Figure 4a). For comparison, control

cell lines were constructed carrying promoter-probe plasmid pCSAM201 (containing a

promoterless *gfp*) or plasmid pCSAM202 (containing a P$_{petE}$::*gfp* fusion) (Figure 4a).

After conjugation into *Anabaena* 7120, the control strain bearing promoter-probe plasmid pCSAM201 showed very low, background fluorescence, whereas the strain bearing plasmid pCSAM202 showed strong expression of GFP in all cells of the filaments in the presence of copper, which induces transcription from the *petE* promoter, but not when the cells were grown in the absence of copper. This GFP fluorescence was observed both in the presence and in the absence of combined nitrogen (not shown).

Surprisingly, the strain bearing plasmid pCSAM207 (transcriptional fusion of all 12 repeats to *gfp* under *petE* promoter control) exhibited, despite the presence of copper, very weak expression of *gfp* when combined nitrogen was present (not shown), but showed strong cell-specific fluorescence upon nitrogen deprivation (Figure 4b). Moreover, the observation of filaments at different time points after nitrogen step-down indicated that GFP fluorescence appeared exclusively in specific cells that were becoming heterocysts (proheterocysts), as well as in mature heterocysts. Based on these observations, we concluded that transcripts initiated at the *petE* promoter in pCSAM207 did not include the *gfp* gene and that multiple arrays of the predicted terminators within the repeats were indeed terminating those transcripts. Interestingly, these data suggest the presence of nitrogen-regulated, cell-specific promoter(s) within the repeats, whose transcriptional activity reaches the *gfp* gene. In order to further test this possibility, a fragment containing repeats 10, 11, and 12 (without the last terminator, see Figure 2) was inserted into the promoter-probe vector pCSAM201 yielding pCSAM208 (Figure 4a). pCSAM208 was transferred into the wild-type strain and mutant derivatives CSE2 (*ntcA*) and 216 (*hetR*). As shown in Figure 4c, cell-specific GFP fluorescence was observed in the wild-type strain bearing plasmid pCSAM208, but not

in any of the two mutant strains, confirming that the expression from the NsiR1 promoter(s) was cell-specific and controlled through NtcA and HetR.

Predicted NsiR1 secondary structure and model for NsiR1 transcription

Because, as mentioned above, primers complementary to the repeats can anneal to any of them, and in order to precisely define the TSS for one single particular repeat, 5' ends of *gfp* transcripts were analyzed in strains bearing pCSAM207. Consistent with the data obtained by fluorescence microscopy, *gfp* transcripts were not detected in samples from ammonium-grown cells (time 0) but 5' ends that correspond to almost identical positions (see TSS (11) and TSS (12) in Figure 2) in repeats 11 and 12 (located closer to the *gfp* gene), could be readily identified in samples obtained after 6-8 hours upon nitrogen depletion (Figure 5a). In fact, the position of both TSSs is consistent with the TSS previously estimated for transcripts originating at the repeats present in the chromosome (Figure 3d). Additionally, extension products corresponding to positions located around the predicted terminator for repeat 11 were observed. These signals could have resulted from premature stops of the primer extension at the secondary structure of the terminator stem-loop. However, processing of long transcripts covering more than one repeat could also produce 5' ends corresponding to this position (see below). No 5' ends that could correspond to transcripts starting upstream from repeat 11 were observed.

3' RACE analysis was carried out in order to identify the precise 3' end of NsiR1 transcripts. All sequences obtained resulted from transcripts corresponding to the conserved repeats (repeats 3-9) and ended with 5, 6, or 7 uridine residues after the

predicted terminator loop. The combination of 5' and 3' mapping in *Anabaena* yielded an NsiR1 sequence of 5'-(U)GGUAGAUGCACCCUGAUAACUAACUCCCCUAGCUGGCUAACACCGACUGG GGGCUUUUU(UU)- 3', corresponding to most of the identical portion of repeats 3-9. This sequence fits the observed length of the most abundant NsiR1 band in northern blots. The predicted secondary structure of NsiR1 is shown in Figure 6a. It consists of a short stem-loop at the 5' end, the predicted Rho-independent terminator at the 3' end, and a likely unpaired region in between. A prediction for a consensus secondary structure based on the 59-62 nt region that is conserved in all available Nostocales (see alignment in Supplementary Figure S1) is presented in Supplementary Figure S4.

The observation that both northern hybridization with an oligonucleotide probe (Figure 3) and primer extension analysis (Figures 3d and 5a) identify precursors that are longer than one NsiR1 unit, suggested that some read-through occurs in the repeat-internal terminators. Based on all available data, we suggest a simple model for the transcription of NsiR1 (Figure 6b). Note that, although only three repeats are included in Figure 6b, this model can be extended to include all 12 repeats. The model depicts a situation in which, for example, three arbitrarily chosen consecutive repeats are mostly transcribed as single NsiR1 molecules. In addition, a minor proportion of transcripts that are not terminated at the corresponding terminator might actually end at the terminator for the next repeat (leading to transcripts of about 200 nt, as those observed in northern blots). Such molecules would contain two NsiR1 units and might eventually be processed by RNases recognizing the terminator loop as substrate. In fact, signals that can be interpreted as premature stops in primer extension assays but might also result from

such processing were observed in primer extension experiments carried out both with oligonucleotides for the repeats (cf. bands of 94 and 100 nt in Figure 3d) and for the *gfp* gene in pCSAM207 (band marked with an asterisk in Figure 5a).

Discussion

Filamentous cyanobacteria exhibit complex developmental alternatives. The differentiation of akinetes, hormogonia and heterocysts requires precise regulatory circuits leading to alternative transcriptional patterns under certain growth conditions and/or in specific cells of filaments. The molecular mechanisms leading to cell-specific expression of genes whose products are involved in early steps of differentiation are, to a large extent, unknown. The initiation of heterocyst differentiation is controlled by a regulatory loop established by NtcA, the global nitrogen regulator, and HetR, a master regulator of cellular differentiation. In response to nitrogen deficiency, induction of the expression of *ntcA* and *hetR* is mutually dependent on each other[34] and takes place mostly in cells that are in the process of differentiation.[13,35] Because non-coding RNAs are frequently involved in regulation of processes leading to cell differentiation of eukaryotic cells, we have carried out a genome-wide search in order to identify non-coding RNAs that could be involved in the process of heterocyst differentiation. The search strategy was focused on the prediction of Rho-independent terminators within all intergenic regions of *Anabaena* 7120. We have previously used such a simple strategy in the unicellular cyanobacterium *Synechocystis* sp. PCC 6803, and in that case, further rigorous testing of the prediction results by using tiling microarrays indicated that this

12

approach is relatively successful for the discovery of unannotated transcripts originating from intergenic spacers, whereas it is only partially productive for the identification of *cis*-antisense RNAs.[30] Accordingly, we expect the number of 678 predicted elements in *Anabaena* 7120 to be an underestimation of the real number of Rho-independent terminators in this organism, whereas the number of false positives is probably low. In the following we focused on a series of repeated elements located in the intergenic region *all3545-hetF* that are transcribed as a short non-coding RNA that we have named NsiR1.

In a recent survey of the occurrence of certain repeats in the Nostocales, eight families of repeats were described.[19] Those repeats are small, in the range of 21 to 27 nt, dispersed, and NsiR1-encoding repeats do not belong to this class. Another class of repeats that has recently gained substantial attention are CRISPRs.[23] Several CRISPRs were identified in *Anabaena* 7120,[36] but the NsiR1 region does not belong to any of them. The observation that the structure of the NsiR1 region is strikingly similar within the available genome sequences of heterocystous cyanobacteria and that NsiR1-like transcripts are detected in 19 different strains suggests that NsiR1 plays some functional role in these cyanobacteria. When comparing the repeat sequences of *Anabaena* 7120 with those available from other Nostocales, the alignment can be split into two segments. The second half (65-70 nt) of each repeat is highly conserved (Figure 2 and Supplementary Figure S1) while the first half shows various levels of conservation. According to northern blots the major product of transcription of the repeated elements corresponds to the highly conserved second half of the repeats, consistent with the results of primer extension assays, which suggest, at least for repeats 11 and 12, that

there is one, centrally located, TSS per repeat. 3' RACE experiments indicate that NsiR1 transcripts are in fact terminated at the predicted Rho-independent terminators located at the end of each repeated segment. Therefore, the recognition sequences for the transcriptional machinery would likely be located within the first half of the repeated elements. Because these sequences exhibit different degrees of sequence divergence, it is possible that some repeats are not transcribed or become activated under conditions that were not tested herein. In fact, transcription from repeat 10 was not observed in strains bearing pCSAM207. Expression of NsiR1 takes place specifically in cells that are in the process of differentiation, and the identified promoters are regulated by NtcA and HetR. However, similar to other heterocyst-specific promoters that depend on both NtcA and HetR,[34,37] no sequences matching the consensus for NtcA-activated promoters (for a review see Luque and Forchhammer (2008))[38] could be identified upstream of the TSS(s), suggesting that the observed regulation is likely mediated by some NtcA-regulated factor rather than by NtcA itself.

The abundance of NsiR1 is very high during the first steps of heterocyst differentiation. While it is not yet clear why such a high concentration of this RNA is required, one possibility is that the effect of gene dosage has driven the amplification of the NsiR1 region in order to generate a high amount of NsiR1 during a short time window early in the differentiation process. It is worth noting that the attempts to overexpress NsiR1 from the copper-inducible *petE* promoter did not significantly increase NsiR1 levels (not shown). We observed that, even at time-points in which NsiR1 expression is highest, *gfp* transcript levels in filaments bearing pCSAM202 (in which *gfp* is transcribed from the *petE* promoter in all cells of the filaments) are similar to those observed in filaments

bearing pCSAM207 (in which *gfp* is transcribed from TSS (11) and TSS (12) specifically in differentiating cells, that account for less than 10 per cent of all cells). Thus, the cumulative transient per-cell transcription from the up to 12 NsiR1 promoters per chromosome is likely to be much higher than that from the *petE* promoter in our constructs.

The observed regulation and cell-specific expression, together with the observation that the position of NsiR1-encoding repeats is conserved, strongly points to a function of NsiR1 in the context of heterocyst development. However, the targets of NsiR1 are completely unknown at present. Unfortunately, because, at least in *Anabaena* 7120, repeat 1 overlaps with the 5' UTR of *hetF*, it is not possible to delete all NsiR1-encoding repeats without simultaneously altering expression of *hetF*, a gene whose encoded protein is involved in the regulation of cellular levels of HetR.[18] Regulation of gene expression through antisense RNA has previously been described in *Anabaena* 7120, contributing to the control of *furA* expression, which codes for the ferric uptake regulator FurA.[39] In the unicellular cyanobacterium *Synechocystis* sp. PCC 6803, the antisense RNA IsrR controls the expression of the *isiA* gene under a variety of stress conditions.[40] More recently, the number of experimentally confirmed antisense RNAs in *Synechocystis* rose to 73.[30] Therefore, it is one possibility that the sequence complementarity between NsiR1 and the *hetF* 5' UTR is functionally relevant.

Another possibility is that NsiR1 acts on one or several mRNAs in trans. The correct computational prediction of the targets of bacterial sRNAs still constitutes a major challenge since these targets are frequently encoded far away, at different genomic loci, and some sRNAs have single targets whereas others control several different mRNAs.

The interacting sequence elements are frequently only short sequence stretches of imperfect similarity, which can reside in any part of the sRNA and can even be formed through the joining of sequence elements from two separate domains (for a recent review see Backofen and Hess (2010).[41] Nevertheless, several algorithms have been designed for target prediction, from these sRNATarget[42] and IntaRNA[43] have been employed here. There are 167 mRNAs identified as potential targets by both programs with stringent parameters (Supplementary Table S3). From these 61 have an annotation and 106 code for "hypothetical proteins". Among the 61 annotated proteins are several with a relation to nitrogen metabolism and six, HetP, HepA, HepB, PatB, NifN and NtcA, with a demonstrated implication in heterocysts differentiation. However, such predictions need to be taken with caution and are likely to contain many false-positives[41]. Nevertheless they provide a starting point for rigorous experimental analysis in the future.

NsiR1 is the first known non-coding RNA present in bacteria that is induced in a cell-type-specific manner under nitrogen deprivation. Its dependence on NtcA and HetR establishes that NsiR1 belongs to the regulatory network that leads to heterocyst differentiation. The precise position of NsiR1 within this network constitutes a challenging subject for further research.

Materials and Methods

Strains and growth conditions

Wild-type *Anabaena* 7120 and mutant derivatives CSE2[11] and 216[32] were grown photoautotrophically at $30^{\circ}C$ in BG11$_0$C (BG11[33] lacking NaNO$_3$ and supplemented with 10 mM NaHCO$_3$) containing 6 mM NH$_4$Cl plus 12 mM *N*-tris (hydroxymethyl) methyl-2-aminoethanesulfonic acid (TES)-NaOH buffer (pH 7.5), bubbled with a mixture of CO$_2$ and air (1 % v/v), and supplemented with 2 μg·ml^{-1} of streptomycin and 2 μg·ml^{-1} of spectinomycin in the case of strain CSE2. Exconjugants containing replicative plasmids were selected and maintained on BG11 solidified with 1% Difco agar in the presence of 5 μg·ml^{-1} of streptomycin and 5 μg·ml^{-1} of spectinomycin (SmSp^R plasmids) or 50 μg·ml^{-1} of neomycin (Nm^R plasmids). Other filamentous cyanobacteria were cultivated in BG11[33]. All media used contained 0.3 μM Cu^{2+}, except for the experiment shown in Fig. 4b, in which regular BG11$_0$ was further supplemented with 0.3 μM CuSO$_4$ (to achieve a final concentration of 0.6 μM Cu^{2+}).

RNA analysis

Cells growing exponentially (~3-5 μg of chlorophyll *a*·ml^{-1}) in BG11$_0$C plus NH$_4$Cl were harvested at room temperature and either processed directly (time 0) or washed and resuspended in BG11$_0$C and further incubated as indicated in each experiment. Samples were collected by filtration or centrifugation and stored frozen. RNA was isolated and analyzed by northern hybridization as previously described[34; 44] with the following modifications: membranes were pre-hybridized for 45 min at 65 °C in 50% deionized formamide, 7% SDS, 250 mM NaCl, 120 mM phosphate buffer pH 7.2. Following the addition of probe, the membranes were incubated overnight at the same temperature (transcript probe) or 48 $^{\circ}$C (oligonucleotide probe), after which they were

washed with the following three buffers for 10 min each (buffer 1: 2X SSC, 1% SDS; buffer 2: 1X SSC, 0.5% SDS; buffer 3: 0.1X SSC, 0.1% SDS). Images of radioactive filters were obtained and analyzed using a Cyclone storage phosphor system and OptiQuant image analysis software (Packard). Primer extension was carried out as described[45] using 15 µg of RNA per sample. The oligonucleotides used as primers were T4, primer 9 and GFP4 (sequences in Supplementary Table S2, positions in Figures 2 and 5).

Rapid amplification of cDNA ends (RACE)

5' and 3' RACE experiments were conducted as described by Steglich *et al*. (2008).[44] Sequences of all primers and linkers used are included in Supplementary Table S2. For 5' RACE, primer 5'RACE_RT_hetF was used for cDNA synthesis, and a linker-specific primer together with 5'RACE_PCR_hetF was employed for subsequent PCR. For 3' RACE, linker-specific primer 3' RACE_DNA-1 was used for cDNA synthesis, and linker-specific 3'RACE_DNA-2 together with NsiR1-specific primer 3 were used for PCR. Relevant bands were extracted from gels, cloned into *E. coli,* and sequenced.

Prediction of ncRNA candidates

The prediction was based on the identification of Rho-independent terminators as described in Georg *et al*. (2009)[30] using the program Rnall.[46] Such a terminator consists of the characteristic GC-rich hairpin followed by a U-rich region, the so called U-tail. The terminator can be subdivided into a proximal (first five bases) and a distal part (four bases after the proximal part). In order to increase stringency of the prediction we applied several filters according to the following rules: 1. At least four G-C or G-U pairs;

2. At most two nt spacer between stem and U-tail; 3. No 'A' at first position of the terminator helix. 4. At least one 'U' at position 2 or 3 in the proximal part; 5. No four purines or four cytosines in the distal region; at least four 'U's in the proximal and the distal region; 6. No multiloops and at most 1 bulge loop with at most 3 unpaired nucleotides. 7. Free energy of the stem-loop at most -7.4 kcal/mol. Rules 1-5 were taken from Lesnik *et al.* (2001)[47] and rules 6 and 7 were defined by us. Free energies were calculated by RNAshapes[48] as Rnall provides a heuristic structure prediction, leading to artifacts in the computed free energy.

Biocomputational prediction of possible targets for NsiR1

For every protein-coding gene, the sequence of the coding region and 100 upstream nucleotides was analyzed for its possible recognition by NsiR1 using the algorithm IntaRNA with a window length of 140 and, otherwise, default parameters.[43] All hits were ordered according to the calculated free energy values gained by the formation of mRNA:NsiR1 duplexes in kcal/mol. Only those hits with values lower than -10 kcal/mol were considered further. In parallel, the program sRNATarget[42] provided at http://ccb.bmi.ac.cn/srnatarget/ was employed, using default parameters. For sRNAtarget, a minimum score of 0.5 was required (maximum is +1.0).

Plasmids

All replicative plasmids constructed are derivatives of vector pDUCA7.[32] Plasmid pCSAM200 is a $Nm^S/Sm^R/Sp^R$ derivative in which the *Pst*I fragment containing part of the *npt* gene has been replaced with a *Pst*I fragment containing the C.S3 cassette (*Pst*I-excised from a plasmid containing C.S3 cloned in the unique *Hind*III site of polylinker

L.EHE2; nomenclature of Elhai and Wolk (1988)[49]), cloned in the same orientation as *npt*. A *Bam*HI fragment containing a *gfp* gene plus corresponding Shine-Dalgarno sequence (thus suitable for transcriptional fusions) was excised from plasmid pAM1956[50] and cloned into the *Bam*HI site of pIC20R[51] rendering pCSAM170. The orientation of the *gfp* gene in this construct is the same as that of the β-galactosidase gene. An *Eco*RI fragment containing the *gfp* gene and the complete polylinker of pCSAM170 was cloned into the unique *Eco*RI site of plasmids pCSAM200 or pDUCA7, yielding Sm^R/Sp^R promoter-probe vector pCSAM201 or Nm^R promoter-probe vector pCSAM211, respectively. The promoter region of the *petE* gene was amplified using oligonucleotides petE1 (containing a *Cla*I site) and petE2 (containing a *Hind*III site) (Supplementary Table S2) and chromosomal DNA from *Anabaena* 7120 as template. The resulting *Cla*I-*Hind*III fragment was cloned between the *Cla*I and *Hind*III sites of the polylinker of pCSAM170, resulting in pCSAM173, in which the *gfp* gene is located downstream of the P$_{petE}$. An *Eco*RI fragment containing the P$_{petE}$::*gfp* construct and the complete polylinker of pCSAM173 was cloned into the unique *Eco*RI site of plasmid pCSAM200 rendering pCSAM202. A fragment containing all 12 repeats was amplified by PCR using oligonucleotides All_repeats_FWR and primer 9 (Supplementary Table S2) and cloned into the *Hind*III site of pCSAM173, between P$_{petE}$ and the *gfp* gene. The whole construct was excised with *Eco*RI and cloned into the unique *Eco*RI site of pCSAM200, rendering pCSAM207. To construct pCSAM208, a fragment containing repeats 10, 11 and 12 (incomplete) was amplified with oligonucleotides TS1 (containing a *Cla*I site) and TS2 (containing an *Xho*I site) (Supplementary Table S2) and chromosomal DNA from *Anabaena* 7120 as template. The resulting *Cla*I-*Xho*I site was

cloned between the *Cla*I and *Xho*I sites of promoter-probe pCSAM211, rendering pCSAM208. The main features of plasmids pCSAM201, 202, 207 and 208 are depicted in Figure 4a.

Fluorescence microscopy

The accumulation of GFP was analyzed by laser confocal microscopy as described.[52] GFP was excited at 488 nm by an argon ion laser, and the fluorescent emission was monitored by collection across windows of 500-540 nm (GFP imaging) and 630-700 nm (cyanobacterial autofluorescence).

Acknowledgments

Figure legends

Figure 1. Organization and conservation of DNA repeats upstream of *hetF* in 5 different cyanobacteria. A scheme of the region upstream of *hetF* in the genomes of *Anabaena* 7120, *Anabaena variabilis* ATCC 29413*, Nodularia spumigena* CCY9414, *Nostoc punctiforme* PCC 73102 and *Nostoc azollae* strain D708 is shown. Homologous ORFs are presented in identical colors. Repeats in *Anabaena* 7120 are numbered 1-12 with respect to the direction of transcription. In the unfinished genome of *N. azollae*, this region is split between two different contigs, and therefore the distance between the first and the second repeat is unknown. The color gradients reflect the divergence between a certain repeat and repeats 3-9 of *Anabaena* 7120. Boxed repeats are almost identical to each other within one genome. The *hetF* transcriptional start is indicated by TSS (*hetF*).

Figure 2. Sequence alignment of repeats 1-12 in *Anabaena* 7120. Insertions of 28, 22 and 54 nt in repeats 1, 2 and 12 compared to repeats 3-11 are indicated at the respective sites. Mapped transcriptional start sites for *hetF* and repeats 11 and 12 are labeled by bent arrows. A position that could correspond to a processing site for transcripts originating at repeat 11 (as determined by primer extension, see Figure 5) is indicated by an asterisk. The positions of several oligonucleotides are indicated: for primer extension below the alignment, for northern analysis shaded in grey. A black triangle indicates the 5' end of the fragment cloned into pCSAM208, a white triangle indicates the 3' end of the fragment cloned into pCSAM207. The 3' end of the fragment cloned in pCSAM208 is located 24 nt downstream from the white triangle. Genome coordinates are provided for the first and the final nucleotide position of the region.

Figure 3. Identification of NsiR1 in cyanobacteria. (a) Expression of NsiR1 in wild-type *Anabaena* 7120 and in the mutant strains CSE2 (*ntcA*) and 216 (*hetR*) after nitrogen depletion. 15 µg of RNA (first nine lanes) or 11 µg RNA (other lanes) was loaded. (b) Expression of NsiR1 over 36 h after nitrogen depletion (15 µg RNA per lane). A 5' end labeled oligonucleotide was used as a probe (cf. Figure 2). The sizes of the major bands are indicated. The lower panels show 5S rRNA as a loading control. RNA was isolated from ammonium-grown cells (lane 0) or from ammonium-grown cells incubated in combined nitrogen-free medium for the indicated number of hours. (c) Overexposure of a section from (a). (d) Primer extension analysis of the expression of the NsiR1 region in *Anabaena* 7120 and in the mutant strains CSE2 (*ntcA*) and 216 (*hetR*). RNA was isolated from ammonium-grown cells (lane 0) or from ammonium-grown cells incubated in combined nitrogen-free medium for the indicated number of hours. Expression of NsiR1 precursors upon nitrogen deprivation in the wild-type strain (left panel) and comparison of primer extension products for RNA samples from the wild-type and mutants CSE2 and 216 strains (middle panel) are shown. Samples from an independent experiment were analyzed with a different primer (right panel). The sizes of the bands referred to in the text are indicated. WT, wild-type strain *Anabaena* 7120. Names of the used oligonucleotides are indicated (see Figure 2 for position and sequence).

Figure 4. Cell-specific transcription from the NsiR1 region. (a) Scheme of the different *gfp* fusions constructed in the replicative plasmids pCSAM201 (promoter-probe), and pCSAM202 (P$_{petE}$::*gfp*) and the repeat-containing plasmids pCSAM207 and pCSAM208. (b) Expression of GFP in wild-type cells carrying plasmid pCSAM207. Micrographs of filaments subjected to nitrogen step-down were taken after 5.5 h in nitrogen-free

medium. Images corresponding to red autofluorescence (left panel), GFP fluorescence (center panel), and overlay of both images (right panel) are shown. (c) Overlay of autofluorescence and GFP fluorescence in filaments of the wild-type strain and in the mutants CSE2 and 216 bearing plasmid pCSAM208 after 8 h in nitrogen-free medium.

Figure 5. Analysis of *gfp* transcripts originating at repeats 11 and 12. (a) Primer extension assays carried out with RNA samples from the wild-type strain bearing plasmid pCSAM207 and oligonucleotide GFP4 (see position in (b)). Approximate positions of the different parts of transcripts are indicated on the left. The 5' ends identified for repeats 11 and 12 are indicated by bent arrows, as in Figure 2. (b) Schematic of the region analyzed. The position of oligonucleotide GFP4 is indicated by an arrow under the *gfp* gene. A primer extension signal possibly indicating a processing site is labeled by an asterisk. MCS, multiple cloning site.

Fig. 6. Predicted secondary structure of NsiR1 and model for transcription and processing. (a) Predicted secondary structure of NsiR1. White triangles point to the 5' ends of NsiR1 as determined for repeats 11 and 12 (cf. Figures 2 and 5a). The different lengths of the polyU tail observed in 3' RACE are indicated by (UU). (b) Model for the transcription and processing of any three contiguous expressed repeats. Asterisks indicate positions at which, according to primer extension products, long transcripts bearing more than one repeat might be processed.

References

1.	Berman-Frank, I., Lundgren, P. & Falkowski, P. (2003). Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria. *Res. Microbiol.* **154**, 157-164.

2.	Gallon, J. (2001). N_2 fixation in phototrophs: adaptation to a specialized way of life. *Plant Soil* **230**, 39-48.

3.	Gallon, J. (1992). Reconciling the incompatible: N_2 fixation and O_2. *New Phytol.* **122**, 571-609.

4.	Berman-Frank, I., Quigg, A., Finkel, Z. V., Irwin, A. J. & Haramaty, L. (2007). Nitrogen-fixation strategies and Fe requirements in cyanobacteria. *Limnol. Oceanogr.* **52**, 2260-2269.

5.	Wolk, C. P., Ernst, A. & Elhai, J. (1994). Heterocyst metabolism and development. In *The Molecular Biology of Cyanobacteria* (Bryant, D. A., ed.), pp. 769-823. Kluwer Academic Publishers, Dordrecht, The Netherlands.

6.	Aldea, M., Kumar, K. & Golden, J. (2008). Heterocyst development and pattern formation. In *Chemical communication among Bacteria* (Winans, S. & Bassler, B., eds.), pp. 75-90. ASM Press, Washington, D.C.

7.	Xu, X., Elhai, J. & Wolk, C. P. (2008). Transcriptional and developmental responses by *Anabaena* to deprivation of fixed nitrogen. In *The Cyanobacteria. Molecular Biology, Genomics and Evolution* (Herrero, A. & Flores, E., eds.), pp. 383-422. Caister Academic Press, Norfolk, UK.

8.   Zhang, C. C., Laurent, S., Sakr, S., Peng, L. & Bédu, S. (2006). Heterocyst differentiation and pattern formation in cyanobacteria: a chorus of signals. *Mol. Microbiol.* **59**, 367-375.

9.   Zhao, J. & Wolk, C. P. (2008). Developmental biology of heterocysts. In *Myxobacteria: Multicellularity and Differentiation* (Whitworth, D. E., ed.), pp. 397-418. ASM Press, Washington, D.C.

10.   Flores, E. & Herrero, A. (2010). Compartmentalized function through cell differentiation in filamentous cyanobacteria. *Nat. Rev. Microbiol.* **8**, 39-50.

11.   Frías, J. E., Flores, E. & Herrero, A. (1994). Requirement of the regulatory protein NtcA for the expression of nitrogen assimilation and heterocyst development genes in the cyanobacterium *Anabaena* sp. PCC 7120. *Mol. Microbiol.* **14**, 823-832.

12.   Wei, T. F., Ramasubramanian, T. S. & Golden, J. W. (1994). *Anabaena* sp. strain PCC 7120 *ntcA* gene required for growth on nitrate and heterocyst development. *J. Bacteriol.* **176**, 4473-82.

13.   Black, T. A., Cai, Y. & Wolk, C. P. (1993). Spatial expression and autoregulation of *hetR*, a gene involved in the control of heterocyst development in *Anabaena*. *Mol. Microbiol.* **9**, 77-84.

14.   Buikema, W. J. & Haselkorn, R. (1991). Characterization of a gene controlling heterocyst differentiation in the cyanobacterium *Anabaena* 7120. *Genes Dev.* **5**, 321-330.

15.   Shi, Y., Zhao, W., Zhang, W., Ye, Z. & Zhao, J. (2006). Regulation of intracellular free calcium concentration during heterocyst differentiation by HetR and NtcA in *Anabaena* sp. PCC 7120. *Proc. Natl. Acad. Sci. USA* **103**, 11334-11339.

16. Zhou, R., Wei, X., Jiang, N., Li, H., Dong, Y., Hsi, K. L. & Zhao, J. (1998). Evidence that HetR protein is an unusual serine-type protease. *Proc. Natl. Acad. Sci. USA* **95**, 4959-4963.

17. Huang, X., Dong, Y. & Zhao, J. (2004). HetR homodimer is a DNA-binding protein required for heterocyst differentiation, and the DNA-binding activity is inhibited by PatS. *Proc. Natl. Acad. Sci. USA* **101**, 4848-4853.

18. Risser, D. D. & Callahan, S. M. (2008). HetF and PatA control levels of HetR in *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* **190**, 7645-7654.

19. Elhai, J., Kato, M., Cousins, S., Lindblad, P. & Costa, J. L. (2008). Very small mobile repeated elements in cyanobacterial genomes. *Genome Res.* **18**, 1484-1499.

20. Meeks, J. C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., Predki, P. & Atlas, R. (2001). An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. *Photosynth. Res.* **70**, 85-106.

21. Tóth, G., Gáspári, Z. & Jurka, J. (2000). Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.* **10**, 967–981.

22. van Belkum, A., Scherer, S., van Alphen, L. & Verbrugh, H. (1998). Short-sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev.* **62**, 275–293.

23. Sorek, R., Kunin, V. & Hugenholtz, P. (2008). CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.* **6**, 181-186.

24. Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., Dickman, M. J., Makarova, K. S., Koonin, E. V. & van der Oost, J. (2008).
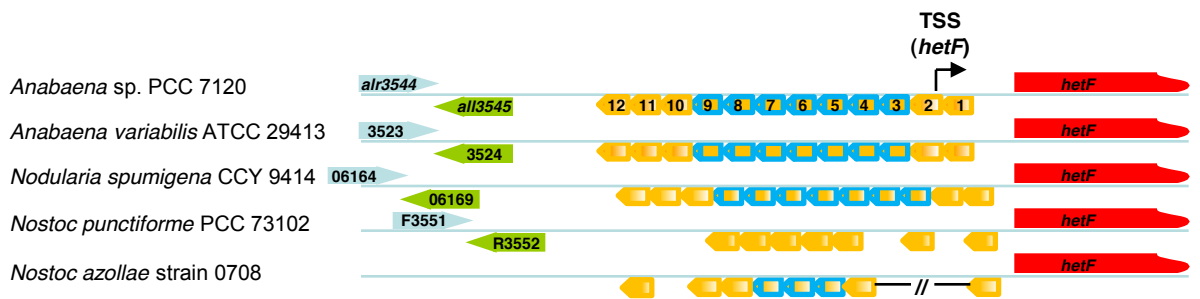
Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960-964.

25. Chen, K. & Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.* **8**, 93-103.

26. Gottesman, S. (2004). The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu. Rev. Microbiol.* **58**, 303-328.

27. Wolk, C. P., Fan, Q., Zhou, R., Huang, G., Lechno-Yossef, S., Kuritz, T. & Wojciuch, E. (2007). Paired cloning vectors for complementation of mutations in the cyanobacterium *Anabaena* sp. strain PCC 7120. *Arch. Microbiol.* **188**, 551-563.

28. Wong, F. C. & Meeks, J. C. (2001). The *hetF* gene product is essential to heterocyst differentiation and affects HetR function in the cyanobacterium *Nostoc punctiforme*. *J. Bacteriol.* **183**, 2654-2661.

29. Wong, F. C. & Meeks, J. C. (2002). Establishment of a functional symbiosis between the cyanobacterium *Nostoc punctiforme* and the bryophyte *Anthoceros punctatus* requires genes involved in nitrogen control and initiation of heterocyst differentiation. *Microbiology* **148**, 315-323.

30. Georg, J., Voss, B., Scholz, I., Mitschke, J., Wilde, A. & Hess, W. R. (2009). Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. *Mol. Syst. Biol.* **5**, 305.

31. Voss, B., Gierga, G., Axmann, I. M. & Hess, W. R. (2007). A motif-based search identifies the ortholog of the small RNA Yfr1 in all major lineages of cyanobacteria. *BMC Genomics* **8**, 375.

32. Buikema, W. J. & Haselkorn, R. (1991). Isolation and complementation of nitrogen fixation mutants of the cyanobacterium *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* **173**, 1879-1885.

33. Rippka, R., Deruelles, J., Waterbury, J. B., Herdmann, M. & Stanier, R. Y. (1979). Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J. Gen. Microbiol.* **111**, 1-61.

34. Muro-Pastor, A. M., Valladares, A., Flores, E. & Herrero, A. (2002). Mutual dependence of the expression of the cell differentiation regulatory protein HetR and the global nitrogen regulator NtcA during heterocyst development. *Mol. Microbiol.* **44**, 1377-85.

35. Olmedo-Verd, E., Muro-Pastor, A. M., Flores, E. & Herrero, A. (2006). Localized induction of the *ntcA* regulatory gene in developing heterocysts of *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* **188**, 6694-9.

36. Godde, J. S. & Bickerton, A. (2006). The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J. Mol. Evol.* **62**, 718-729.

37. Muro-Pastor, A. M., Flores, E. & Herrero, A. (2009). NtcA-regulated heterocyst differentiation genes *hetC* and *devB* from *Anabaena* sp. strain PCC 7120 exhibit a similar tandem promoter arrangement. *J. Bacteriol.* **191**, 5765-74.

38. Luque, I. & Forchhammer, K. (2008). Nitrogen assimilation and C/N balance sensing. In *The Cyanobacteria. Molecular Biology, Genomics and Evolution* (Herrero, A. & Flores, E., eds.), pp. 335-382. Caister Academic Press, Norfolk, UK.
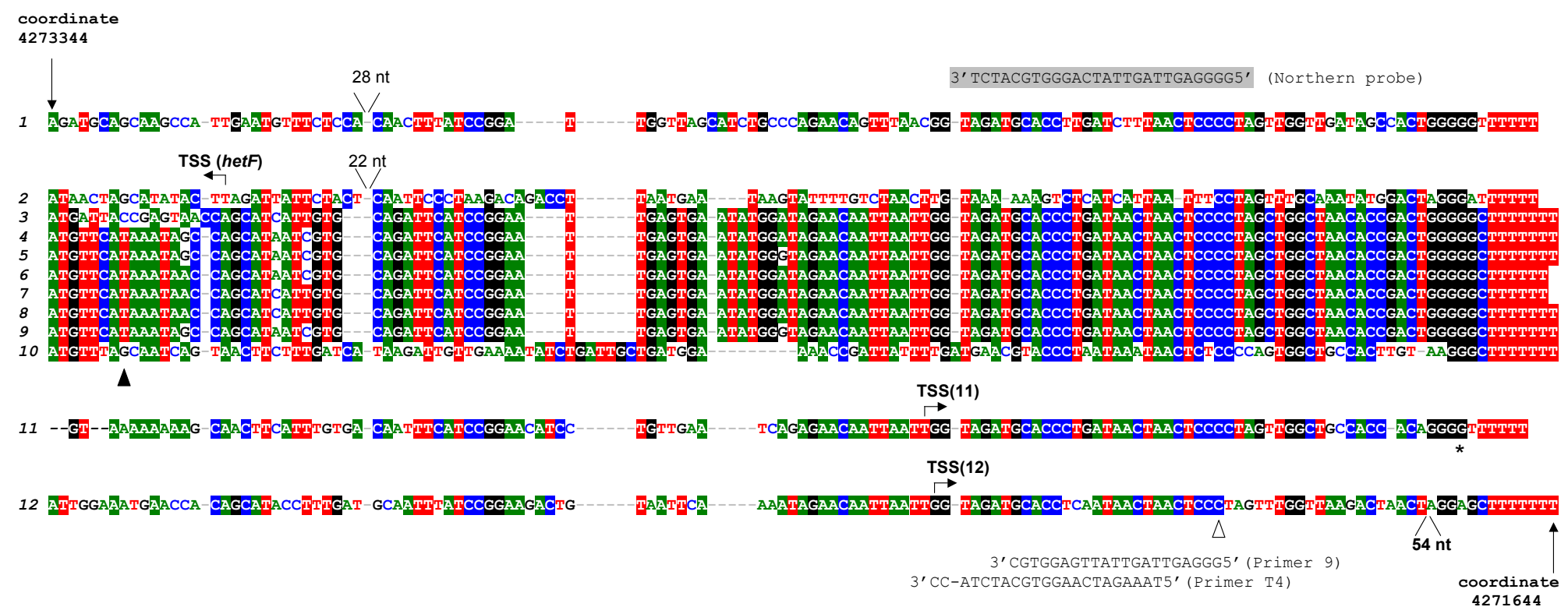
39.   Hernández, J. A., Muro-Pastor, A. M., Flores, E., Bes, M. T., Peleato, M. L. & Fillat, M. F. (2006). Identification of a *furA cis* antisense RNA in the cyanobacterium *Anabaena* sp. PCC 7120. *J. Mol. Biol.* **355**, 325-334.

40.   Dühring, U., Axmann, I. M., Hess, W. R. & Wilde, A. (2006). An internal antisense RNA regulates expression of the photosynthesis gene *isiA*. *Proc. Natl. Acad. Sci. USA* **103**, 7054-7058.

41.   Backofen, R. & Hess, W. R. (2010). Computational prediction of sRNAs and their targets in bacteria. *RNA Biology* **7**, 1-10.

42.   Cao, Y., Zhao, Y., Cha, L., Ying, X., Wang, L., Shao, N. & Li, W. (2009). sRNATarget: a web server for prediction of bacterial sRNA targets. *Bioinformation* **3**, 364-6.

43.   Busch, A., Richter, A. S. & Backofen, R. (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* **24**, 2849-56.

44.   Steglich, C., Futschik, M. E., Lindell, D., Voss, B., Chisholm, S. W. & Hess, W. R. (2008). The challenge of regulation in a minimal photoautotroph: non-coding RNAs in *Prochlorococcus*. *PLoS Genet.* **4**, e1000173.

45.   Muro-Pastor, A. M., Valladares, A., Flores, E. & Herrero, A. (1999). The *hetC* gene is a direct target of the NtcA transcriptional regulator in cyanobacterial heterocyst development. *J. Bacteriol.* **181**, 6664-6669.

46.   Wan, X.-F. & Xu, D. (2005). Intrinsic terminator prediction and its application in *Synechococcus* sp. WH8102. *Internat. J. Comp. Sci. Technol.* **20**, 465-482.

47. Lesnik, E. A., Sampath, R., Levene, H. B., Henderson, T. J., McNeil, J. A. & Ecker, D. J. (2001). Prediction of *rho*-independent transcriptional terminators in *Escherichia coli. Nucleic Acids Res.* **29**, 3583-94.

48. Giegerich, R., Voss, B. & Rehmsmeier, M. (2004). Abstract shapes of RNA. *Nucleic Acids Res.* **32**, 4843-4851.

49. Elhai, J. & Wolk, C. P. (1988). A versatile class of positive-selection vectors based on the nonviability of palindrome-containing plasmids that allows cloning into long polylinkers. *Gene* **68**, 119-38.

50. Yoon, H. S. & Golden, J. W. (1998). Heterocyst pattern formation controlled by a diffusible peptide. *Science* **282**, 935-8.

51. Marsh, J. L., Erfle, M. & Wykes, E. J. (1984). The pIC plasmid and phage vectors with versatile cloning sites for recombinant selection by insertional inactivation. *Gene* **32**, 481-5.

52. Muro-Pastor, A. M., Olmedo-Verd, E. & Flores, E. (2006). All4312, an NtcA-regulated two-component response regulator in *Anabaena* sp. strain PCC 7120. *FEMS Microbiol. Lett.* **256**, 171-177.

Figure

Ionescu et al., FIG. 1

TSS
(*hetF*)

Anabaena sp. PCC 7120

Anabaena variabilis ATCC 29413

Nodularia spumigena CCY 9414

Nostoc punctiforme PCC 73102

Nostoc azollae strain 0708

Ionescu et al., FIG 2

Ionescu et al., FIG. 3

Ionescu et al., FIG 4

(a)

pCSAM201

pCSAM202   P*petE*   *gfp*

pCSAM207   P*petE*   1 2 3 4 5 6 7 8 9 10 11 12   *gfp*
coordinates 4273391 to 4271730

pCSAM208   10 11 12   *gfp*

(b)

(c)

wild-type strain
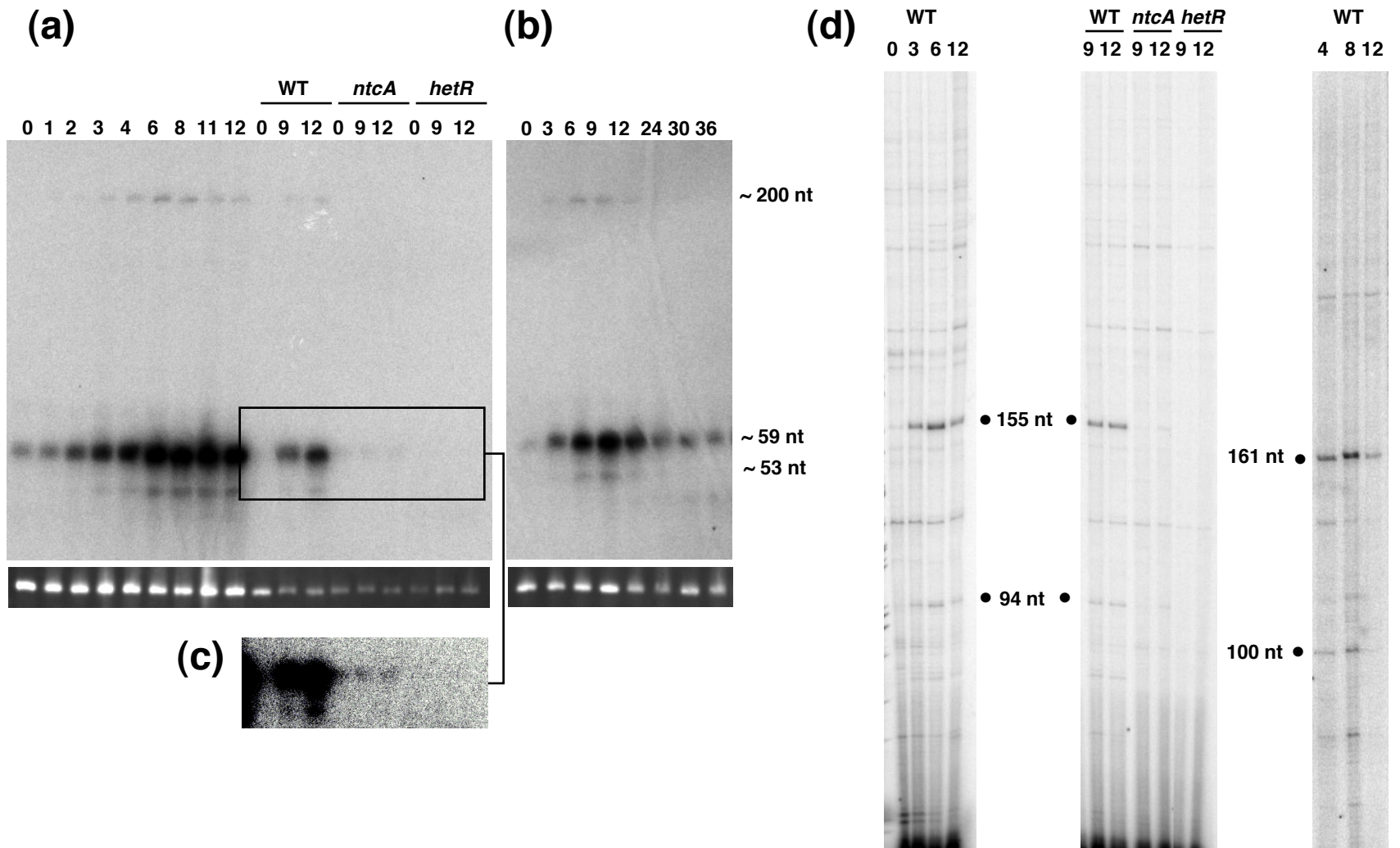+ pCSAM208

strain CSE2 (*ntcA*)
+ pCSAM208

strain 216 (*hetR*)
+ pCSAM208
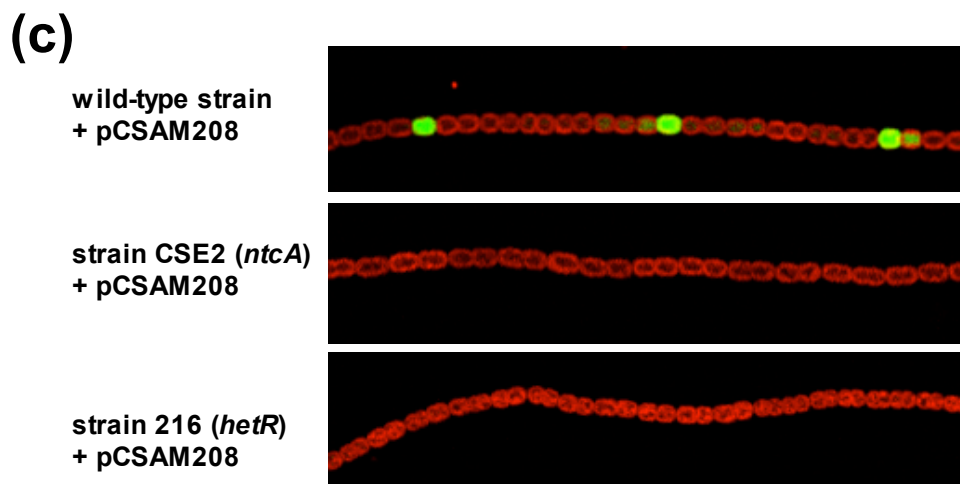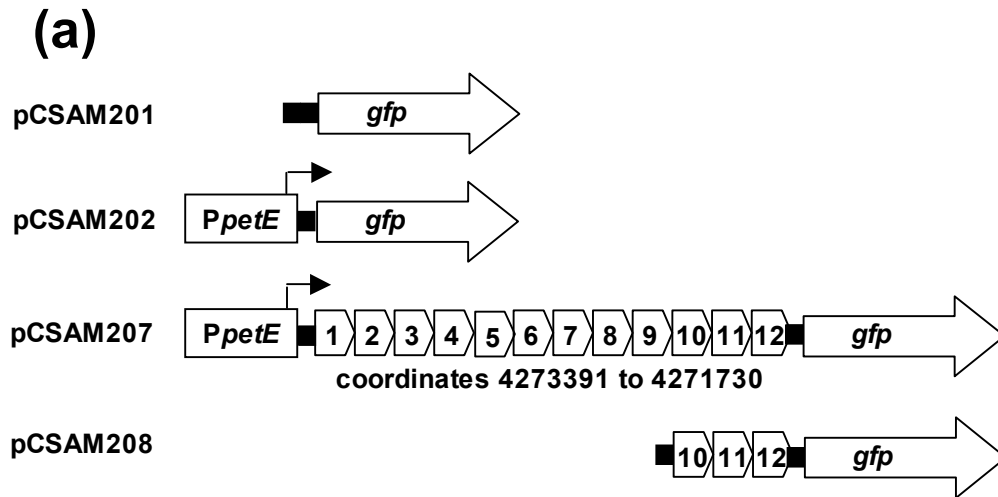
Ionescu et al., FIG 5

Ionescu et al., FIG 6

(a)

```
                                        40
                                        |
                                      A A
                                    U     C
                                      C   A
                                      G-C
                                      G-C
                                      U-G
                                      C A
                                      G-C
              A U                     A-U
            G     G          30 — U-G
            A   C                    C-G — 50
             U-A — 10                C-G
             G-C                     C-G
    5' (U)G-CCUGAUAACUAACUC-GCUUUUU(UU) 3'
        Δ  Δ                      |
                                 20
```
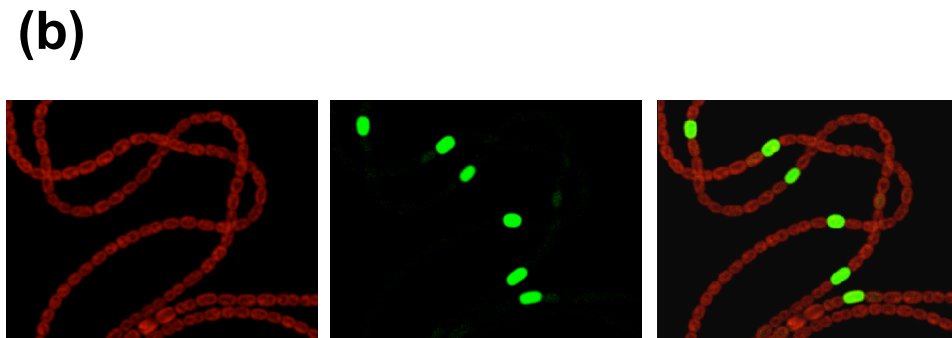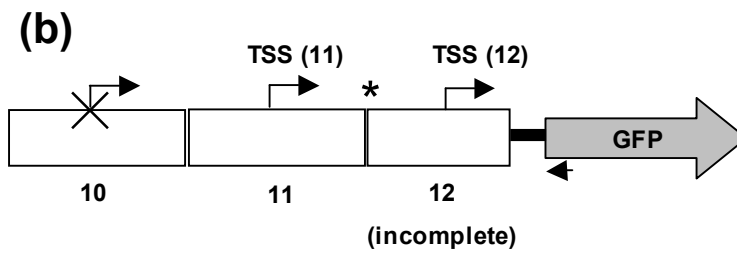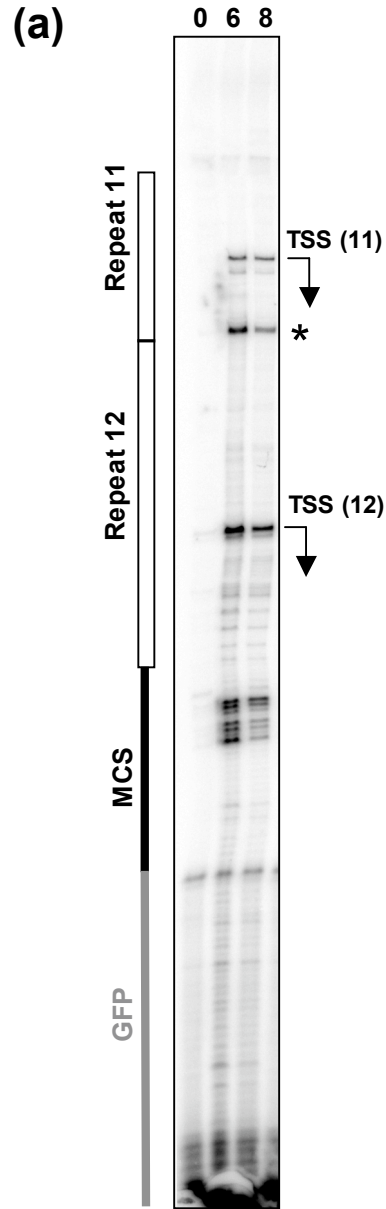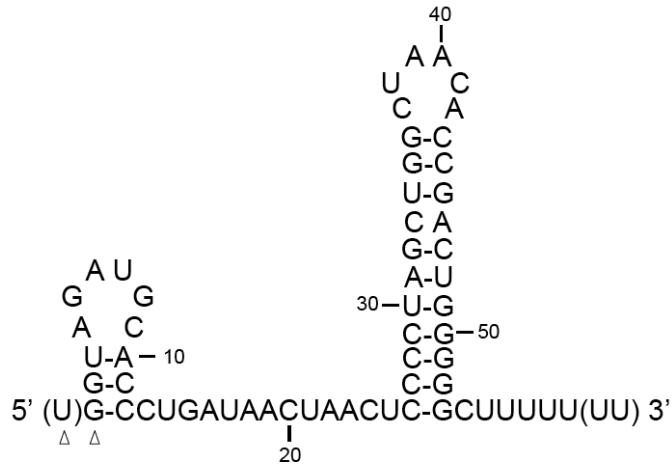
(b)

NsiR1 transcripts (59-62 nt)

Transcripts containing two NsiR1 units (~200 nt)

Ionescu et al.: Supplementary Information, table and figure legends and Table S2

Figure S1. Alignment of repeats from all cyanobacteria. Sequences are given as the corresponding RNA transcripts, their IDs consist of the strain name, prefixed by the number of the repeat. The alignments were generated using RNAforester,[1] which lines up the structured region at the 3' end, followed by manual optimization. Alignment editing and the coloring were done with RALEE.[2] The color scheme is as follows: yellow indicates >40% conservation, blue indicates >60% conservation and red indicates >80% conservat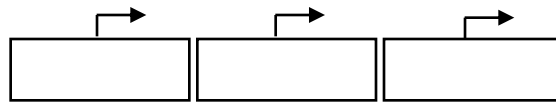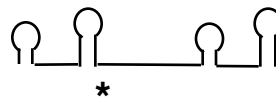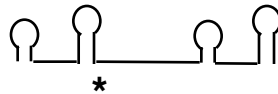ion. The conserved secondary structure at the 3' end (predicted terminators) is indicated in dot-bracket notation in the last line. The position of TSS(s) determined for *Anabaena* 7120 is indicated.

Figure S2. Minimum evolution tree of all repeats from Figure S1. The distance matrix was calculated using the Jukes Cantor algorithm, as implemented in MEGA 4.0.[3]

Figure S3. Northern blot with an RNA probe potentially covering the whole repeats region. (a) A time-course is shown for *Anabaena* 7120 subjected to nitrogen step-down for up to 24 h. (b) Identification of NsiR1-like transcripts in 18 other cyanobacterial strains. All cyanobacterial strains were grown in BG11 medium[4] without agitation. The experiment was done over the course of 12 h of nitrogen depletion and sampled every 3 h. Only selected time points are shown. Labeled RNA probes were prepared using 100 ng of purified PCR products corresponding to the whole repeats region obtained with primers Northern-F and Northern T7-R (Supplementary Table S2) and the T7 Maxi

Script kit (Ambion) according to the manufacturer instructions. For each reaction 50 µCi of α^{32}P-UTP was used.

Figure S4. Secondary structure of NsiR1. (a) Base conservation and (b) basepairing probabilities in the secondary structure of NsiR1 as calculated using RNAalifold.[5] The structure was built based on an alignment of NsiR1 sequences from all the available Nostocales (Supplementary Figure S1). Highly divergent sequences were not used. The color bar represents probabilities for basepairing as shown in panel (b).

Table S1. Genome-wide prediction of *Rho*-independent terminators in intergenic regions of *Anabaena* 7120.

Table S2. Sequences of all oligonucleotide primers. Introduced restriction sites are underlined. Sequence corresponding to the T7 promoter is in lower case.

Primer name	Sequence 5' – 3'	Experiment
5' RACE_RT_hetF	CAATGGGTCATTCATCAAATGCC	5' RACE (RT)
5' RACE_PCR_hetF	CATAGAAAGAAACGCATCTCCC	5' RACE (PCR)
5' linker-specific primer	ATATGCGCGAATTCCTGTAGAACGA	5' RACE (PCR)
5' RACE RNA linker	AUAUGCGCGAAUUCCUGUAGAACGAACACUAGAAGAAA	5' RACE
3' RACE RNA linker	phosphate-UUCACUGUUCUUAGCGGCCGCAUGCUC-idT	3' RACE
3' RACE_DNA-1	GAGCATGCGGCCGCTAAG	3' RACE (RT)
3' RACE_DNA-2	GGCCGCTAAGAACAGTG	3' RACE (PCR)
Primer 3	GCACCTCAATAACTAACTCCC	3' RACE (PCR)
Primer 9	GGGAGTTAGTTATTGAGGTGC	Primer extension, cloning
Primer T4	TAAAGATCAAGGTGCATCTACC	Primer extension
GFP4	TGTATATCTCCTTCTTAAA	Primer extension
oligonucleotide probe	GGGGAGTTAGTTATCAGGGTGCATCT	Northern blot
Northern-F	CATTGTGCAGATTCATCCGGAA	Northern blot
Northern T7-R	taatacgactcactatagggaCAGGGTGCATCTACCAATTAATTG	Northern blot
petE1 (*Cla*I)	<u>ATCGAT</u>GGACTCAGAACACAGTACTC	Cloning
petE2 (*Hind*III)	A<u>AAGCTT</u>CCATGGCGTTCTCCT	Cloning
All_repeats_FWD	CTTGAAACGCCAAAGTTTCAATTC	Cloning
TS1 (*Cla*I)	<u>ATCGAT</u>GCAATCAGTAACTTCTTTG	Cloning
TS2 (*Xho*I)	<u>CTCGAG</u>CACCAGTTAGTCTTAACC	Cloning

Supplementary Table S3. List of mRNAs predicted as possible targets for NsiR1 in trans. The sequence of NsiR1 was used to search for possible targets among all genes of *Anabaena* 7120 using the algorithms IntaRNA[6] and sRNATarget[7]. All hits were ordered according to the calculated free energy values gained by the formation of mRNA:NsiR1 duplexes in kcal/mol (mRNA Energy (IntaRNA)). Moreover, the sRNATarget scores (maximum +1.0) are given, together with the annotation for each predicted target mRNA, its location, strand (+, forward strand; -, reverse strand), the length of the respective protein, its database ID (PID) and associated number from the clusters of orthologous genes (COG). Heterocyst-related proteins are highlighted in boldface letters. Note that *hetF* is not in this list since the overlap with NsiR1 is in the elongated 5'UTR only, from position -116 to -174. The predicted energy of this interaction is -37.4086 kcal/mol (IntaRNA).

Supplementary Information References

1. Höchsmann, M., Töller, T., Giegerich, R. & Kurtz, S. (2003). *IEEE Bioinformatics Conference*.

2. Griffiths-Jones, S. (2005). RALEE--RNA ALignment editor in Emacs. *Bioinformatics* **21**, 257-9.

3. Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596-1599.

4. Rippka, R., Deruelles, J., Waterbury, J. B., Herdmann, M. & Stanier, R. Y. (1979). Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J. Gen. Microbiol.* **111**, 1-61.

5. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R. & Hofacker, I. L. (2008). The Vienna RNA websuite. *Nucleic Acids Res.* **36**, W70-4.

6. Busch, A., Richter, A. S. & Backofen, R. (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* **24**, 2849-56.

7. Cao, Y., Zhao, Y., Cha, L., Ying, X., Wang, L., Shao, N. & Li, W. (2009). sRNATarget: a web server for prediction of bacterial sRNA targets. *Bioinformation* **3**, 364-6.

Nodularia spumigena CCY 9414 Repeat 2
Nodularia spumigena CCY 9414 Repeat 5
Nodularia spumigena CCY 9414 Repeat 6
Nodularia spumigena CCY 9414 Repeat 3
Nodularia spumigena CCY 9414 Repeat 8
Nodularia spumigena CCY 9414 Repeat 7
Nodularia spumigena CCY 9414 Repeat 4
Nodularia spumigena CCY 9414 Repeat 1
Nodularia spumigena CCY 9414 Repeat 9
Nostoc azollae strain 0708 repeat 7
Nostoc azollae strain 0708 repeat 6
Nostoc azollae strain 0708 repeat 4
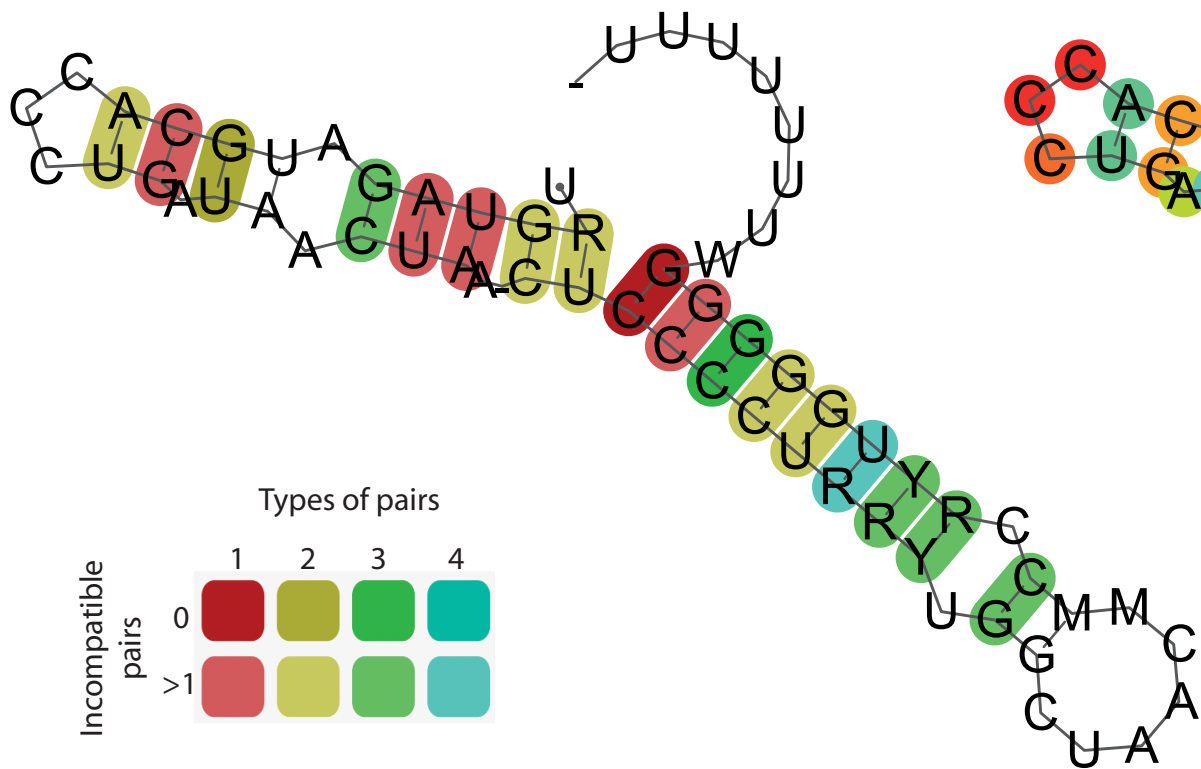*Nostoc azollae* strain 0708 repeat 5
Nostoc punctiforme PCC 73102 repeat 2
Nostoc punctiforme PCC 73102 repeat 3
Nostoc azollae strain 0708 repeat 8
Anabaena sp PCC 7120 Repeat 11
Anabaena variabilis ATCC 29413 repeat 11
Nodularia spumigena CCY 9414 Repeat 11
Nodularia spumigena CCY 9414 Repeat 10
Nostoc azollae strain 0708 repeat 2
Anabaena sp PCC 7120 Repeat 10
Anabaena variabilis ATCC 29413 repeat10
Anabaena sp PCC 7120 Repeat 3
Anabaena variabilis ATCC 29413 repeat 3
Anabaena sp PCC 7120 Repeat 5
Anabaena sp PCC 7120 Repeat 9
Anabaena sp PCC 7120 Repeat 4
Anabaena sp PCC 7120 Repeat 6
Anabaena sp PCC 7120 Repeat 7
Anabaena sp PCC 7120 Repeat 8
Anabaena variabilis ATCC 29413 repeat 6
Anabaena variabilis ATCC 29413 repeat 9
Anabaena variabilis ATCC 29413 repeat 4
Anabaena variabilis ATCC 29413 repeat 5
Anabaena variabilis ATCC 29413 repeat 7
Anabaena variabilis ATCC 29413 repeat 8
Nostoc azollae strain 0708 repeat 3
Nodularia spumigena CCY 9414 Repeat 12
Anabaena sp PCC 7120 Repeat 12
Anabaena variabilis ATCC 29413 repeat 12
Nostoc punctiforme PCC 73102 repeat 7
Nostoc punctiforme PCC 73102 repeat 5
Nostoc punctiforme PCC 73102 repeat 6
Anabaena sp PCC 7120 Repeat 1
Anabaena variabilis ATCC 29413 repeat 1
Nostoc punctiforme PCC 73102 repeat 1
Nostoc azollae strain 0708 repeat 1
Nostoc punctiforme PCC 73102 repeat 4
Anabaena sp PCC 7120 Repeat 2
Anabaena variabilis ATCC 29413 repeat 2

Repeat 11
Repeat 10
Repeat 12
Repeat 1

0.1

Ionescu et al., supplementary Fig. S3

(a)

(b)