An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803

Jan Mitschke[a,1], Jens Georg[a,1], Ingeborg Scholz[a], Cynthia M. Sharma[b], Dennis Dienst[c], Jens Bantscheff[a], Björn Voß[a], Claudia Steglich[a], Annegret Wilde[d], Jörg Vogel[b], and Wolfgang R. Hess[a,e,2]

[a]Faculty of Biology and Freiburg Initiative in Systems Biology, University of Freiburg, D-79104 Freiburg, Germany; [b]Institute for Molecular Infection Biology, University of Würzburg, D-97080 Würzburg, Germany; [c]Institute of Biology, Humboldt University Berlin, D-10115 Berlin, Germany; [d]Institute of Microbiology and Molecular Biology, Justus-Liebig University Giessen, D-35392 Giessen, Germany; and [e]Zentrum für Biosystemanalyse, University of Freiburg, D-79104 Freiburg, Germany

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved December 21, 2010 (received for review October 8, 2010)

There has been an increasing interest in cyanobacteria because these photosynthetic organisms convert solar energy into biomass and because of their potential for the production of biofuels. However, the exploitation of cyanobacteria for bioengineering requires knowledge of their transcriptional organization. Using differential RNA sequencing, we have established a genome-wide map of 3,527 transcriptional start sites (TSS) of the model organism *Synechocystis* sp. PCC6803. One-third of all TSS were located upstream of an annotated gene; another third were on the reverse complementary strand of 866 genes, suggesting massive antisense transcription. Orphan TSS located in intergenic regions led us to predict 314 noncoding RNAs (ncRNAs). Complementary microarray-based RNA profiling verified a high number of noncoding transcripts and identified strong ncRNA regulations. Thus, ∼64% of all TSS give rise to antisense or ncRNAs in a genome that is to 87% protein coding. Our data enhance the information on promoters by a factor of 40, suggest the existence of additional small peptide-encoding mRNAs, and provide corrected 5′ annotations for many genes of this cyanobacterium. The global TSS map will facilitate the use of *Synechocystis* sp. PCC6803 as a model organism for further research on photosynthesis and energy research.

gene expression regulation | promoter prediction | RNA polymerase

Cyanobacteria are important primary producers but also are considered promising resources for the production of biofuels such as hydrogen (1), ethanol (2), isobutyraldehyde and isobutanol (3), ethylene (4), volatile isoprene hydrocarbons (5), and alkanes (6). The unicellular model species *Synechocystis* sp. PCC6803 (*Synechocystis* 6803) has an ∼3.6-Mbp genome encoding 3,172 proteins and was the first phototrophic organism to have its genome sequenced (7).

DNA-based annotation has a limited ability to determine the transcriptional organization of a genome and therefore increasingly is being complemented by experimental high-throughput discovery of transcriptional start sites (TSS). A global TSS mapping can help identify transcripts from seemingly empty genomic regions that might be regulatory RNAs or mRNAs of short peptides, both of which are not commonly covered by traditional gene annotation; it also detects potential antisense transcripts (asRNA) originating from the reverse complementary strand of annotated genes. Using a recently developed differential RNA-sequencing approach (dRNA-seq) that is selective for the 5′ ends of primary transcripts (8), we present a genome-wide map of *Synechocystis* 6803 with more than 3,500 experimentally mapped TSS. The annotated primary transcriptome of *Synechocystis* 6803 will facilitate the use of this genetically tractable organism as a model for biofuel-producing microalgae, photosynthesis research, and systems biology.

Results

Large-Scale Mapping of Primary 5′ Ends Using dRNA-Seq. According to the published dRNA-seq protocol (8), we sequenced two cDNA libraries prepared from the same total RNA, one referred to as "(−)"covering both primary and processed transcripts and the other, "(+)," in which primary transcripts were enriched by the use of terminator exonuclease. In total 358,083 sequence reads were obtained by pyrosequencing, and 8.7 million bases of cDNA were mapped to the *Synechocystis* 6803 chromosome and its four megaplasmids.

Fig. 1 gives an overview of our genome-wide TSS mapping and an example of differential cDNA coverage as shown for the protein-coding *trxA* gene. In total, we identified 3,213 chromosomal TSS and 314 TSS on the four megaplasmids pSYSA, pSYSG, pSYSM, and pSYSX (*SI Appendix*, Table S1). All TSS were classified based on their location (Fig. 1*B*) upstream of annotated genes (gTSS) (mostly mRNAs) or in intergenic spacers (nTSS) [noncoding RNA (ncRNA) TSS] or by their inverse orientation to annotated genes (aTSS) (suggesting antisense transcription). TSS in sense orientation located internally within annotated genes were designated "iTSS." For 185 of these TSS, we found associations with more than one category (Fig. 1*C*).

Our data confirmed 44 of 64 TSS that previously had been mapped for 59 genes or operons (*SI Appendix*, Table S2); the other 20 TSS/genes were expressed too weakly for our analysis. Besides the already published ones, we identified additional TSS for the photosystem I gene *psaD*, *Synechocystis* noncoding RNA 2 (SyR2), the type 3 sigma factor gene *sigF*, and the ammonium transporter gene *amt1*. The two TSS of *amt1* are in close proximity to each other. As in some other cyanobacteria (9), the −35 element of one of them (TSS2) overlaps with the binding site of the nitrogen-responsive regulatory protein, NtcA. Thus, our data agree well with previous promoter analyses and often seem to be more sensitive. In ∼89% of all transcripts, transcription started on a purine (2,274 A; 751 G); C or T almost equally marked the first nucleotide in the remaining 502 TSS. Elements with similarity to the enterobacterial −35 box (5′-TTG***-3′) were detected upstream of 19.7% of all TSS (*Dataset S1*, Table S3), a percentage similar to our previous low-scale study of the marine cyanobacterium *Prochlorococcus* MED4 (10). Transcripts originating at 10 nTSS were positive in a BLASTX search against the National Center for Biotechnology Information (NCBI) protein database and seem to encode small proteins of 31–94 amino acids

Author contributions: W.R.H. designed research; J.G., I.S., C.M.S., D.D., J.B., C.S., and A.W. performed research; J.M., J.G., I.S., C.M.S., B.V., C.S., and W.R.H. analyzed data; and J.M., J.G., I.S., C.M.S., A.W., J.V., and W.R.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The microarray data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession nos. GSE16162 and GSE14410). Supplementary data files 1, 2, 3, and 4 are available at http://www.cyanolab.de/Supplementary.html.

[1]J.M. and J.G. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: wolfgang.hess@biologie.uni-freiburg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1015154108/-/DCSupplemental.

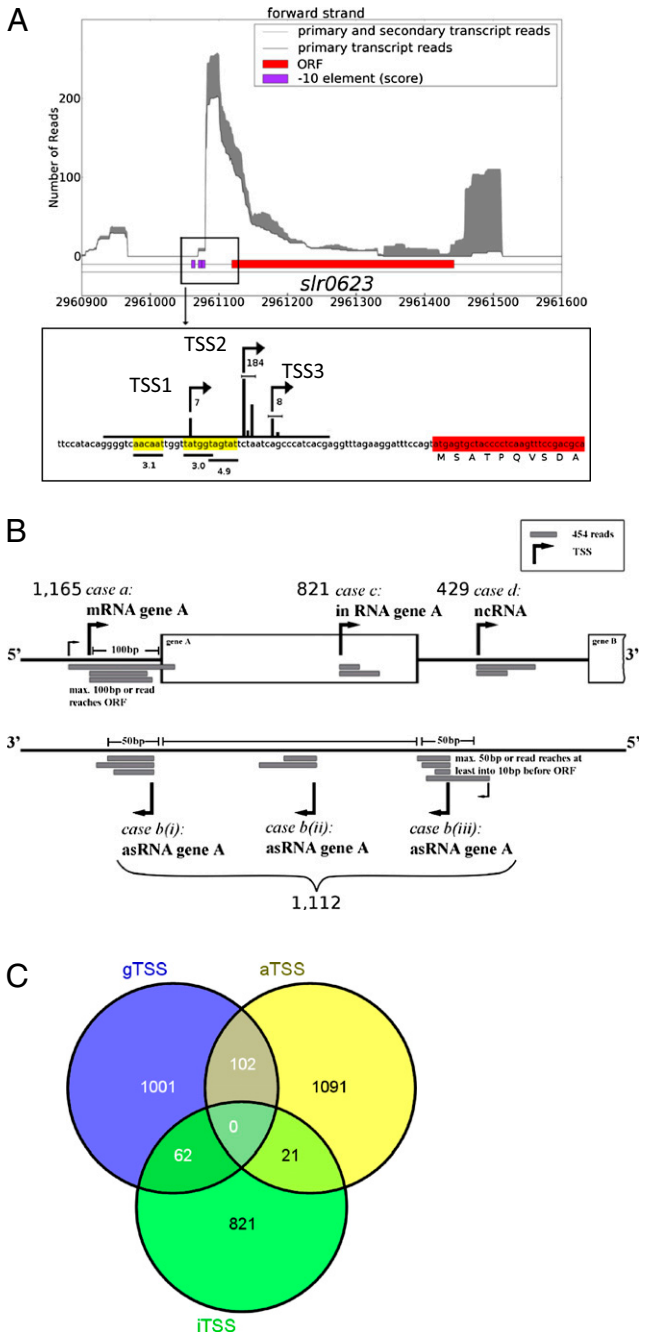

Fig. 1. TSS in the chromosome and plasmids pSYSA, G, M, and X of *Synechocystis* 6803. (*A*) (*Upper*) Three closely spaced TSS are found in the region upstream of *trxA*/slr0623, encoding one of four thioredoxins in *Synechocystis* 6803, illustrated by the 199 primary (+) sequencing reads starting in this region. These reads start at position −49 (TSS1, seven reads), −38/−36 (TSS2, 184 reads), and −32 (TSS3, eight reads). (*Lower*) All three TSS exhibit a reliable score for a −10 element at a distance of six plus or minus one nucleotides from the first transcribed nucleotide. TSS2 and TSS3 correspond to two previously detected major primer extension products (40). The total number of sequencing reads from the (+) and (−) cDNA libraries is shown in light gray along the length of the gene in the upper panel. (*B*) Details of annotation and classification of 3,527 TSS into 1,165 gTSS giving rise to mRNA, 1,112 aTSS producing asRNA, 821 iTSS for internal sense transcripts, and 429 nTSS for candidate ncRNAs. Case a: A TSS was classified as gTSS if the TSS was located 0–100 nt upstream of an ORF or if at least 1 of the 454 reads reached into the coding sequence or the 10 nt in front of it. Case b: The TSS is located antisense to an annotated gene or within ≤50 bp of its 5′ or 3′ UTR. Case c: The TSS is positioned within an annotated sequence. Case d: Putative ncRNA (nTSS). (*C*) Overlaps between different categories of TSS. Many TSS associate with multiple categories according to Fig. 1*B*. Thus, 102 of the 1,165 gTSS actually are located antisense, and 62 are located in sense orientation within another annotated gene. One example is the slr1470 mRNA starting from an iTSS within *rnpA* (slr1469). We also observed 21 aTSS that were located gene-internally because of overlapping transcription. An example is the aTSS for the asRNA to slr0320, which starts internally within the *rpoD* (sll0306) coding sequence. None of the 429 nTSS overlapped one of the other three categories; therefore these TSS are not shown in the diagram.

(*SI Appendix*, Table S4), prompting their reclassification as gTSS. The associated reading frames were named "Norf1–8" (Norf1 and Norf5 have two TSS each). All Norfs except Norf2 have closely related annotated genes in other cyanobacteria, and Norf3 and Norf8 appear to be pseudogenized copies of transposase genes belonging to the ISY120 and IS3/IS911 families of insertion elements.

We focused on the chromosomally located TSS. A global and semiquantitative overview of the occurrence of all TSS along a linear plot of the chromosome is given in Fig. 2. The highest number of reads was associated with the nTSS of the ncRNAs SyR10–SyR12 and members of the Yfr2 ncRNA family, which is ubiquitous within the phylum cyanobacteria (11).

TSS of Protein-Coding and Noncoding Genes. The distances between the identified gTSS and start codons of protein-coding genes ranged from 0–278 nt with a median distance of 42 nt. A high number of reads was found for the gTSS in front of several genes that encode ribosomal, photosynthetic, pilin, cell division, and hypothetical proteins (*Dataset S1*, Table S5). We also noticed several overlapping but divergently transcribed promoters such as for *psbN*/*psbH* in which the two −10 elements almost fully overlap on the plus and minus strand of DNA.

Ten gTSS mapped to the start of reading frames, coinciding with either the A of start codon AUG or the preceding nucleotide. Based on the presence of N-terminally shorter homologs in other bacteria, five of these gTSS indicated misannotated start codons, and another was associated with transposase pseudogene sll1542. However, the four remaining gTSS of *rps12* (sll1096), *apcE* (slr0335), slr1079, and slr0846 (a transcriptional regulator) clearly indicated that these genes are translated from leaderless mRNAs.

For several genes such as *rfbB* (slr0809), a gene-internal iTSS hinted at an incorrectly annotated start codon (Fig. 3*A*). The *rfbB* gene encodes dTDP-glucose 4,6-dehydratase and is annotated in several databases as an ORF of 987 nt; by contrast, our data suggest a shorter reading frame of 930 nt, preceded by a 5′ UTR of 21 nt. In total, start sites were reannotated for 58 reading frames (*SI Appendix*, Table S6). The remaining 732 iTSS might represent TSS of genes located downstream. If so, their positions should be biased toward the 3′ end of the genes in which they are found. However, most of these iTSS are located in the first 3% of their associated genes and otherwise are distributed throughout annotated reading frames (Fig. 3*B*). Thus, these iTSS might yield primarily short sense transcripts and truncated alternative mRNAs. For example, the gTSS of *ntcA* previously was mapped to a constitutive promoter at position −408 (12) (*SI Appendix*, Table S1), indicating cotranscription with the gene ssl2781 located upstream (Fig. 3*C*). In addition, we mapped a second gTSS at −51 and two iTSS, at +618 and +249. The combined results of dRNA-seq, microarray detection, and Northern blot probing demonstrate two abundant short sense transcripts which originate from the iTSS at +249 (Fig. 3 *C* and *D*). Comparison with available tiling microarray data using direct labeling of RNA samples (13) supports our dRNA-seq–based predictions of short sense transcripts from iTSS (Fig. 3*E*) and reveals their frequent differential regulation as compared with the respective full-length mRNA (Supplementary data file 3, available at http://www.cyanolab.de/Supplementary.html).

The set of 370 chromosomally located nTSS led us to predict 314 ncRNA candidates, some of which have more than one TSS. This class of transcripts contains several TSS with highest cDNA coverage (*Dataset S1*, Table S8), among them the nTSS for the

MICROBIOLOGY

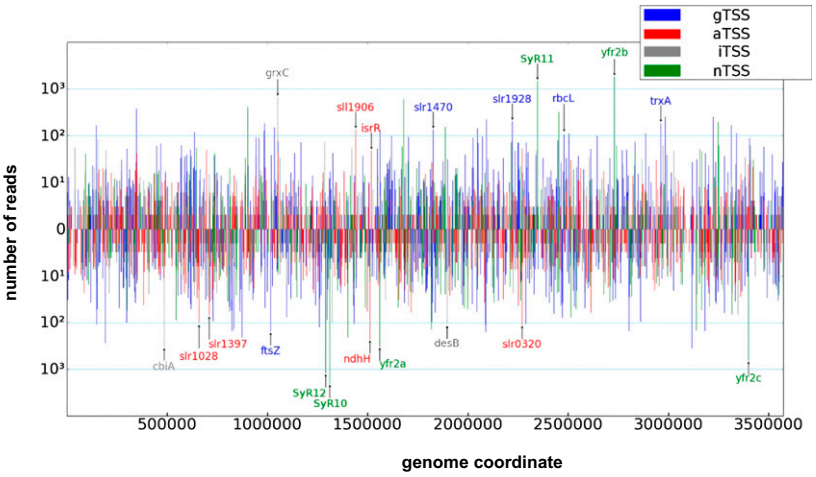

Fig. 2. Occurrence of 3,213 TSS along a linear plot of the *Synechocystis* 6803 chromosome. The genome position is drawn along the *x* axis and is given in nucleotides. Mapped TSS for the forward strand are plotted above the *x* axis, and mapped TSS for the reverse strand are plotted below the *x* axis. The number of sequence reads is given as a proxy for gene expression on the *y* axis (logarithmic scaling). The location of each of the TSS according to Fig. 1*B* served for classification of the respective TSS as gTSS from which an mRNA would originate (blue), nTSS for a putative ncRNA (green), aTSS for antisense transcripts (red), or iTSS (gray).

Yfr2b ncRNA which is known to be transcribed from a strong promoter (14). The nTSS for the RNase P RNA (*rnpB*), 4.5S RNA (*ffs*), 6S RNA, and tmRNA (*ssrA*) matched the published experimental data (15, 16) or genome annotation. We also detected the very short 6S RNA-associated product RNA (pRNA) (17) starting at position c1686546, which corresponds to the bulge-internal adenosine position of *Escherichia coli* 6S RNA (17). Although the cyanobacterial pRNAs are slightly longer (19–30 nt) than their counterparts in *E. coli* (14–20 nt), their detection in a phylogenetically distant cyanobacterium suggests that 6S RNA-mediated regulation of RNA polymerase activity is very widely conserved.

Regulation Under Conditions Relevant for a Photosynthetic Organism. Complementary microarray transcript profiling strongly supported the dRNA-seq results. Cultures were grown under high light, darkness, CO_2 depletion, and standard growth conditions (Table 1), and typical markers such as the high light induction of HliA (encoded by ssl2542) (Supplementary data file 3) (18) were used to verify differential mRNA expression under the chosen conditions. Fig. 4 shows the hydrogenase operon which encodes all proteins required for pentameric bidirectional Ni-Fe hydrogenase (HoxEFUYH). This operon uses a single TSS located 168 nt upstream of *hoxE* (19), and its regulation is well studied (19–21). Intriguingly, we found that the operon is framed by two highly expressed ncRNAs, one of which [SyR1 (14)], was strongly upregulated under high light, whereas the other (ncr0700) was maximally expressed in the dark (Fig. 4*B*). Moreover, overexpression of SyR1 from an inducible promoter caused a severe phenotype accompanied by loss of pigmentation (Fig. 4*C*). Thus, it is tempting to speculate that the two ncRNAs have roles in the dark/light adaptation of the cyanobacterial cell.

A previous study indicated widespread antisense transcription in *Synechocystis* 6803 (13). Our comprehensive mapping of 1,011 aTSS in the chromosome revealed that a quarter of all chromosomal genes are subject to antisense transcription (*Dataset S1*, Table S9). At the extreme end, the slr1028 gene encoding a giant ~418-kDa protein of unknown function is associated with 15 aTSS over its length of ~12 kb (*SI Appendix*, Fig. S1). Several aTSS are associated with genes with key functions in photosynthesis and produce, for example the highly abundant as_ndhH and IsrR species (13, 22). Furthermore, antisense RNA as_ndhB affects a gene (sll0223) that is conserved from bacteria to plants. Interestingly, the plant ortholog, which usually is chloroplast borne, also has an asRNA (23). An asRNA overlapping the very 5′ end was discovered for the *psbA* gene family that encodes photosystem II protein D1 (*Dataset S1*, Table S9). Specifically, we detected aTSS within the 5′ UTRs of *psbA2* and *psbA3*, located 19 nt upstream of the respective start codons. Intriguingly, the aTSS is opposite RNase E cleavage sites in the 5′ UTR of *psbA2* which previously were implicated in *psbA* mRNA destabilization in the dark (24). We also noticed a potential asRNA to the *psaA-psaB* intergenic spacer, originating at position c944069. In *Synechocystis* 6803, photosystem I genes *psaA* and *psaB* are cotranscribed from well-characterized promoters (25), but under some conditions *psaA* accumulates predominantly as a monocistronic mRNA (26). Thus, it is conceivable that differential mRNA accumulation from the *psaAB* dicistron is determined by the intergenic asRNA, similar to the reported activity of GadY sRNA in the *E. coli gad* operon (27). Another intriguing example is an aTSS within *furA* (sll0567) encoding the ferric uptake regulator. Expression of the *furA* ortholog in the cyanobacterium *Anabaena* PCC7120 is fine-tuned by antisense transcription (28), with important consequences for iron homeostasis (29).

A global comparison of our dRNA-seq results with microarray profiling data confirmed the high complexity of the *Synechocystis* 6803 transcriptome and also revealed a high degree of consistency between the two technologies (Supplementary data file 3). In the microarray analyses, 404 asRNAs and 166 ncRNAs were verified as significantly expressed, although asRNAs generally seemed less prone to regulation than ncRNAs and mRNAs (Table 1). However, the calculation of ratios between the absolute microarray transcript abundances of all asRNA:mRNA pairs revealed a number of very characteristic expression changes (Supplementary data file 4, available at http://www.cyanolab.de/Supplementary.html). For example, the as_*ndhB:ndhB* (sll0223) pair showed a ratio >1 under dark incubation and <<1 under high light and CO_2 depletion and thus induction of the CO_2 uptake system (*SI Appendix*, Fig. S2*A*). The mRNAs of several other proteins involved in the uptake of carbon, for example, the *ndhF3*/*ndhD3*/*cupA*/*orf133* genes encoding the NDH-1_3 complex and *ccmK* encoding a protein of the carbon concentrating mechanism (*SI Appendix*, Fig. S2*B*), also were associated with asRNAs and showed similar ratio changes. Together, these asRNAs might help turn off gene expression rapidly when cells reenter noninducing conditions.

Other important functions are associated with asRNAs, such as the NADH-dependent glutamate synthase small subunit (*gltD* gene; *SI Appendix*, Fig. S2*C*) and ATPase (*SI Appendix*, Fig. S2*D*). The very good concordance between dRNA-seq and microarray results is illustrated further for internal sense transcripts (*SI Appendix*, Fig. S2 *E* and *F*) as well as for several ncRNAs and the riboswitch elements RF00442 and RF00379 (*SI Appendix*, Fig. S3).

Discussion

Within the last 2 y, RNA-seq technology has revolutionized the global identification of TSS in prokaryotes and has triggered a wave of studies that are setting new standards in this field (8, 30, 31). Our data provide insight into the complexity of the primary transcriptome of *Synechocystis* 6803, revealing the location of 3,527 TSS on the chromosome and the four megaplasmids of *Synechocystis* 6803. In addition to the TSS map of cyanobacteria, we provide an annotation of 5′ UTRs and a reannotation of

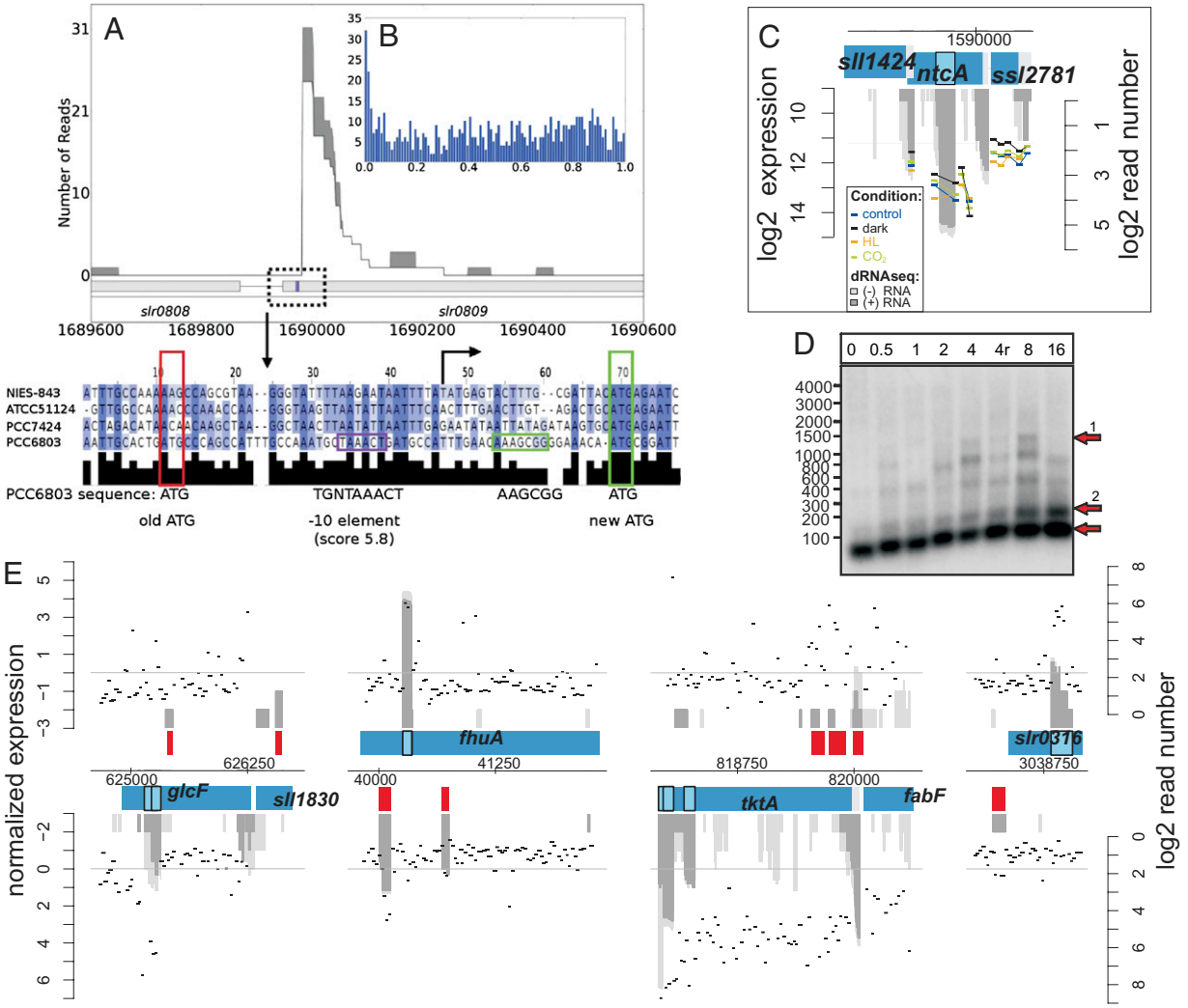

Fig. 3. Gene-internally located TSS (iTSS). (*A*) Example for the reannotation of start codons based on a newly discovered TSS. (*Upper*) Reads of (+) and (–) libraries mapping to the *rfbB* (slr0809) region are shown according to Fig. 1*A*. (*Lower*) An alignment of *rfbB* sequences (5′ region) from the cyanobacteria *Microcystis aeruginosa* NIES 843, *Synechococcus elongatus* PCC7942, *Cyanothece* sp. ATCC 51142, and *Synechocystis* 6803 is shown. The here discovered TSS is displayed by an arrow downstream of the original start codon (boxed in red). The highly conserved reannotated AUG and the possible Shine–Dalgarno element are boxed in green. The −10 box of the new TSS is boxed in purple. (*B*) Distribution of 732 iTSS relative to their positions within the reading frames in which they are located. The number of iTSS belonging to each percentile ia plotted along the *y* axis, and the nucleotide positions are plotted along the *x* axis (in %). (*C*) The *ntcA* gene and its four mapped TSS. The light blue box shows the region from which a very abundant sense transcript originates. Microarray probes are indicated by short vertical tabs linked by short colored lines, and their expression values are plotted on the left *y* axis. (*D*) Northern hybridization. RNA was extracted from cultures grown in a time-course experiment following the removal of nitrate (in hours). The induction of the *ntcA* full-length mRNA (arrow 1) at 4 h can be reversed by the readdition of nitrate (4r). The sense transcripts originating at the iTSS at +249 accumulate in the form of two dominant bands (arrows 2). (*E*) Comparison of dRNA-seq data with previous tiling microarray analysis (13) validates the accumulation of short sense transcripts from iTSS within genes encoding the glycolate oxidase subunit F (*glcF*), the ferrichrome-iron receptor (*fhuA*), or the transketolase (*tktA*). The gene regions from which short sense transcripts originate are drawn in light blue; confirmed asRNAs are shown in red. The normalized expression values (left *y* axis) are plotted for microarray probes (black dots). Read numbers for the (+) and (–) libraries are plotted in dark and light gray, respectively (log_2 scale on right *y* axis).

reading frames for 58 genes. It is likely that additional TSS will be identified in future studies that go beyond the present bacterial culture under standard conditions and include the induction of stress or starvation. Nonetheless, our present study already has identified strongly transcribed promoters without a detectable −10 element that usually are associated with nonstandard growth conditions.

The 1,098 identified chromosomally located gTSS will improve the analysis of mRNA promoter regulation significantly. Moreover, some of the nTSS might give rise to both ncRNAs and long 5′ UTRs of mRNAs, as in the case of Yfr2b and Yfr2c, which are transcribed from an nTSS that also could be a gTSS for the downstream slr0199 and sll1477 genes (14). However, even if some of the 370 chromosomal nTSS are gTSS, the majority seem to drive the transcription of independent RNA species. With the use of microarray analysis to corroborate the dRNA-seq data, 404 of 537 asRNAs and 166 of 194 ncRNAs were verified as significantly expressed. We found that 29.8% and 13.2% of all mRNAs had a significantly reduced or enhanced expression level, respectively, in any of the three conditions (high light, CO_2 depletion, or dark), as compared with the control. Similar percentages were found for the ncRNAs (30.9% and 15.5%, respectively), whereas only 13% and 6.7%, respectively, of all asRNAs showed significantly altered expression (Table 1). Thus, similar fractions of ncRNAs and mRNAs are regulated under these three conditions (which are highly relevant for a photosynthetic organism), strongly suggesting that the

Table 1. Number of differentially regulated transcripts on the *Synechocystis* 6803 transcriptome microarray under three different conditions

	Dark	High light	CO_2	%	Total
mRNA	−560	−211	−170	29.8	3,152
	336	252	129	13.2	
asRNA	−12	−48	−10	13.0	537
	22	8	6	6.7	
ncRNA	−25	−16	−19	30.9	194
	13	7	10	15.5	

Each condition [dark incubation for 1 h, incubation under high light (500 μmol photons $m^{-2}·s^{-1}$) for 30 min, or under CO_2 depletion] is compared against standard growth conditions (50 μmol photons $m^{-2}·s^{-1}$). For each condition, the number of features with significantly changed expression and the total number of genes, asRNAs, and ncRNAs represented on this array are indicated. Negative fold change indicates reduced expression and positive fold change represents enhanced expression against standard conditions. A graphical overview on the combined results of microarray and 454 analyses is presented in Supplementary data file 3.

ncRNAs identified here are functionally relevant. These ncRNAs might act to regulate the turnover or processing of di- or multicistronic mRNAs of key photosynthesis operons; however, given the notoriously high false-positive rates of available algorithms (32), we abstain here from predicting possible ncRNA–mRNA interactions.

The 1,013 aTSS in the *Synechocystis* 6803 chromosome with 3,172 annotated ORFs (NCBI annotation) demonstrate the level of complexity in the transcriptome of this model organism. To exclude the possibility that the aTSS were experimental artifacts of nonspecific priming during cDNA synthesis (33), the dRNA-seq results were compared with available tiling array data (13) and with a transcriptome array in which RNA was labeled directly to avoid such artifacts. This comparison verified the existence of a plethora of antisense transcripts in this organism and recapitulates recent findings of massive antisense transcription in other prokaryotes (34). The compact genome of the human pathogen *Helicobacter pylori* transcribes asRNAs for 46% of all annotated ORFs (8), and in the archaeon *Sulfolobus solfataricus* 6.1% of all genes were associated with asRNAs (30). Thus, many of the antisense transcripts are likely to have important roles in gene regulation in *Synechocystis* 6803 as well.

Gene-internal transcripts initiating at iTSS might give rise to alternative mRNAs, resulting in the synthesis of more than one polypeptide from the same gene. In *Synechocystis* 6803, a second isoform of the ferredoxin:NADP oxidoreductase is generated by an in-frame initiation of translation from the *petH* gene (35). Although such cases generally have been sparse in bacteria (36), we speculate

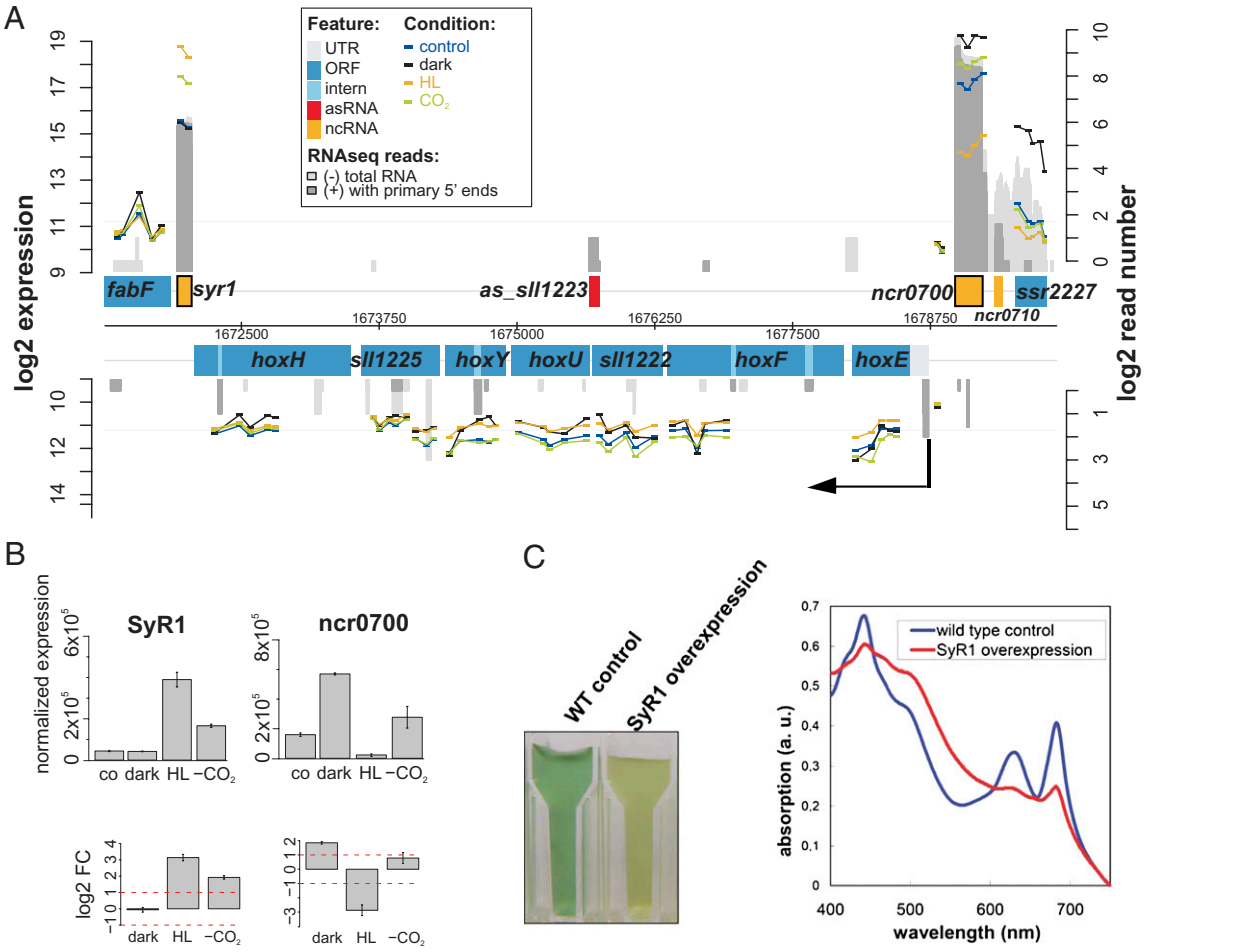

Fig. 4. The hydrogenase (*hox)* operon is framed by two highly expressed ncRNAs, SyR1 (14) and ncr0700. (*A*) Both strands are shown with the location of annotated genes (blue boxes), 5′ UTRs (light gray), internal sense RNAs (light blue), asRNAs (red), and intergenic ncRNA genes (yellow). The TSS of the *hox* operon (19) is indicated by the black arrow. The read numbers for the enriched library (+) are plotted in dark gray, and reads for the untreated library (–) are in light gray and are given in log_2 scale (right *y* axis). The normalized log_2 expression values of four different microarray experiments are plotted in blue (control), black (dark incubation), yellow (incubation at high light), and green (CO_2 depletion). The scale for the microarray data are given at the left *y* axis. All probes of a single RNA feature are connected by lines. (*B*) Mean expression (*Upper*) and mean fold changes (FC) of the two short ncRNAs SyR1 and ncr0700 (*Lower Right*) under control conditions (co), dark incubation, incubation at high light (HL), and CO_2 depletion ($-CO_2$). (*C*) The SyR1 RNA was overexpressed under control of the *petJ* promoter, causing a severe phenotype with a loss of pigmentation (*Left*, cuvettes; *Right*, whole-cell absorption spectra).

that some of the iTSS reported here could produce shorter isoforms of *Synechocystis* 6803 proteins. Alternatively, some internal sense transcripts may have a regulatory scavenger function, acting as target mimicry for ncRNAs, as has been reported for plant miRNAs (37), or may act as independent ncRNAs.

In summary, the annotated primary transcriptome of *Synechocystis* 6803, together with its recently modeled metabolic network (38), will greatly facilitate the use of this organism as a simple photosynthetic model in fundamental and systems biology and for the establishment of biofuel-producing microalgae.

Methods

Full protocols are available in *SI Appendix and SI Methods*.

Growth Conditions and Mutagenesis. *Synechocystis* 6803 was grown at 30 °C in BG11 medium under 50 μmol photons $m^{-2}{\cdot}s^{-1}$ of white light. SyR1 was overexpressed from the conjugative plasmid pVZ-spec under control of the *petJ* promoter. Exconjugants were selected on BG11 agar plates containing 40 $\mu g{\cdot}mL^{-1}$ kanamycin and 20 $\mu g{\cdot}mL^{-1}$ spectinomycin.

Preparation of RNA, Pyrosequencing and Expression Analysis. Total RNA was isolated as previously described (22). Details of the dRNA-seq method are provided in ref. 8. The cDNA libraries were prepared and analyzed on a Roche FLX sequencer as previously described (39). After addition of 5′ linkers with unique tags for each library and poly-A-tailing, the RNA was converted into cDNA. A total of 169,360 and 188,723 sequence reads were obtained for the (–) and (+) populations, respectively. From these populations, 129,346 and 148,767 sequence reads, respectively, were ≥18 nt in length, and 106,018 and 131,943 sequence reads, respectively, matched the sequences of the genome or one of the four megaplasmids of *Synechocystis* 6803. In addition, TSS for 80 different genes were determined manually by 5′ RACE as described (10).

The microarray design, hybridization procedure, and data analysis have been described previously (13). The microarray data are available in the GEO database (accession nos. GSE16162 and GSE14410). Features are stated as significantly expressed if at least one probe at one condition passed the threshold of $2^{11.12}$ after subtraction of the SD. The threshold was defined by the mean of non-*Synechocystis* control probes (after adding the SD of the control probes).

Computational Methods. For the (+) population, 95,413 sequencing reads ≥18 nt (excluding ribosomal sequences) were mapped to the *Synechocystis* 6803 chromosome, and all 5′ ends located within a window of three consecutive nucleotides were joined and considered to be possible TSS. A threshold for the position-specific weight matrix (PSWM) for the −10 element was set at +2.00 based on a computation of the possible gain of true positives against the chance of acquiring more false positives (*SI Appendix*, Fig. S4), the comparison with published data (*SI Appendix*, Table S2), and additional experimental verification (*SI Appendix*, Table S7). We prioritized gTSS over aTSS and iTSS, and all remaining TSS were automatically considered nTSS. All scripts used were written in Python 2.5.2 and Biopython V 1.42 (http://biopython.org/) and are available on request.

ACKNOWLEDGMENTS. This work was supported by the Deutsche Forschungsgemeinschaft Focus Program "Sensory and Regulatory RNAs in Prokaryotes" Grant SPP1258 (to W.R.H., A.W., C.S., and J.V.) and by Federal Ministry of Education and Research Grant 0313921 (to W.R.H.).

1. McKinlay JB, Harwood CS (2010) Photobiological production of hydrogen gas as a biofuel. *Curr Opin Biotechnol* 21:244–251.
2. Deng MD, Coleman JR (1999) Ethanol synthesis by genetic engineering in cyanobacteria. *Appl Environ Microbiol* 65:523–528.
3. Atsumi S, Higashide W, Liao JC (2009) Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. *Nat Biotechnol* 27:1177–1180.
4. Takahama K, Matsuoka M, Nagahama K, Ogawa T (2003) Construction and analysis of a recombinant cyanobacterium expressing a chromosomally inserted gene for an ethylene-forming enzyme at the *psbAI* locus. *J Biosci Bioeng* 95:302–305.
5. Lindberg P, Park S, Melis A (2010) Engineering a platform for photosynthetic isoprene production in cyanobacteria, using *Synechocystis* as the model organism. *Metab Eng* 12:70–79.
6. Schirmer A, Rude MA, Li X, Popova E, del Cardayre SB (2010) Microbial biosynthesis of alkanes. *Science* 329:559–562.
7. Kaneko T, et al. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions (supplement). *DNA Res* 3:185–209.
8. Sharma CM, et al. (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 464:250–255.
9. Herrero A, Muro-Pastor AM, Flores E (2001) Nitrogen control in cyanobacteria. *J Bacteriol* 183:411–425.
10. Vogel J, Axmann IM, Herzel H, Hess WR (2003) Experimental and computational analysis of transcriptional start sites in the cyanobacterium *Prochlorococcus* MED4. *Nucleic Acids Res* 31:2890–2899.
11. Gierga G, Voss B, Hess WR (2009) The Yfr2 ncRNA family, a group of abundant RNA molecules widely conserved in cyanobacteria. *RNA Biol* 6:222–227.
12. Aichi M, Takatani N, Omata T (2001) Role of NtcB in activation of nitrate assimilation genes in the cyanobacterium *Synechocystis* sp. strain PCC 6803. *J Bacteriol* 183:5840–5847.
13. Georg J, et al. (2009) Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. *Mol Syst Biol* 5:305.1–305.17.
14. Voss B, Georg J, Schön V, Ude S, Hess WR (2009) Biocomputational prediction of non-coding RNAs in model cyanobacteria. *BMC Genomics* 10:123.1–123.15.
15. Vioque A (1992) Analysis of the gene encoding the RNA subunit of ribonuclease P from cyanobacteria. *Nucleic Acids Res* 20:6331–6337.
16. Tous C, Vega-Palas MA, Vioque A (2001) Conditional expression of RNase P in the cyanobacterium *Synechocystis* sp. PCC6803 allows detection of precursor RNAs. Insight in the in vivo maturation pathway of transfer and other stable RNAs. *J Biol Chem* 276:29059–29066.
17. Wassarman KM, Saecker RM (2006) Synthesis-mediated release of a small RNA inhibitor of RNA polymerase. *Science* 314:1601–1603.
18. He Q, Dolganov N, Bjorkman O, Grossman AR (2001) The high light-inducible polypeptides in *Synechocystis* PCC6803. Expression and function in high light. *J Biol Chem* 276:306–314.
19. Gutekunst K, et al. (2005) LexA regulates the bidirectional hydrogenase in the cyanobacterium *Synechocystis* sp. PCC 6803 as a transcription activator. *Mol Microbiol* 58:810–823.
20. Oliveira P, Lindblad P (2005) LexA, a transcription regulator binding in the promoter region of the bidirectional hydrogenase in the cyanobacterium *Synechocystis* sp. PCC 6803. *FEMS Microbiol Lett* 251:59–66.
21. Oliveira P, Lindblad P (2008) An AbrB-Like protein regulates the expression of the bidirectional hydrogenase in *Synechocystis* sp. strain PCC 6803. *J Bacteriol* 190:1011–1019.
22. Dühring U, Axmann IM, Hess WR, Wilde A (2006) An internal antisense RNA regulates expression of the photosynthesis gene *isiA*. *Proc Natl Acad Sci USA* 103:7054–7058.
23. Georg J, Honsel A, Voss B, Rennenberg H, Hess WR (2010) A long antisense RNA in plant chloroplasts. *New Phytol* 186:615–622.
24. Horie Y, et al. (2007) Dark-induced mRNA instability involves RNase E/G-type endoribonuclease cleavage at the AU-box and SD sequences in cyanobacteria. *Mol Genet Genomics* 278:331–346.
25. Takahashi T, Nakai N, Muramatsu M, Hihara Y (2010) Role of multiple HLR1 sequences in the regulation of the dual promoters of the *psaAB* genes in *Synechocystis* sp. PCC 6803. *J Bacteriol* 192:4031–4036.
26. Muramatsu M, Sonoike K, Hihara Y (2009) Mechanism of downregulation of photosystem I content under high-light conditions in the cyanobacterium *Synechocystis* sp. PCC 6803. *Microbiology* 155:989–996.
27. Opdyke JA, Kang JG, Storz G (2004) GadY, a small-RNA regulator of acid response genes in *Escherichia coli*. *J Bacteriol* 186:6698–6705.
28. Hernández JA, et al. (2006) Identification of a *furA* cis antisense RNA in the cyanobacterium *Anabaena* sp. PCC 7120. *J Mol Biol* 355:325–334.
29. Hernández JA, et al. (2010) Mutants of *Anabaena* sp. PCC 7120 lacking alr1690 and alpha-furA antisense RNA show a pleiotropic phenotype and altered photosynthetic machinery. *J Plant Physiol* 167:430–437.
30. Wurtzel O, et al. (2010) A single-base resolution map of an archaeal transcriptome. *Genome Res* 20:133–141.
31. Cho BK, et al. (2009) The transcription unit architecture of the *Escherichia coli* genome. *Nat Biotechnol* 27:1043–1049.
32. Backofen R, Hess WR (2010) Computational prediction of sRNAs and their targets in bacteria. *RNA Biol* 7:33–42.
33. Perocchi F, Xu Z, Clauder-Münster S, Steinmetz LM (2007) Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res* 35:e128.1–e128.7.
34. Thomason MK, Storz G (2010) Bacterial antisense RNAs: How many are there, and what are they doing? *Annu Rev Genet* 44:167–188.
35. Thomas JC, Ughy B, Lagoutte B, Ajlani G (2006) A second isoform of the ferredoxin: NADP oxidoreductase generated by an in-frame initiation of translation. *Proc Natl Acad Sci USA* 103:18368–18373.
36. McNamara BP, Wolfe AJ (1997) Coexpression of the long and short forms of CheA, the chemotaxis histidine kinase, by members of the family Enterobacteriaceae. *J Bacteriol* 179:1813–1818.
37. Franco-Zorrilla JM, et al. (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet* 39:1033–1037.
38. Knoop H, Zilliges Y, Lockau W, Steuer R (2010) The metabolic network of *Synechocystis* sp. PCC 6803: Systemic properties of autotrophic growth. *Plant Physiol* 154:410–422.
39. Sittka A, et al. (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet* 4:e1000163.
40. Navarro F, Martín-Figueroa E, Florencio FJ (2000) Electron transport controls transcription of the thioredoxin gene (*trxA*) in the cyanobacterium *Synechocystis* sp. PCC 6803. *Plant Mol Biol* 43:23–32.