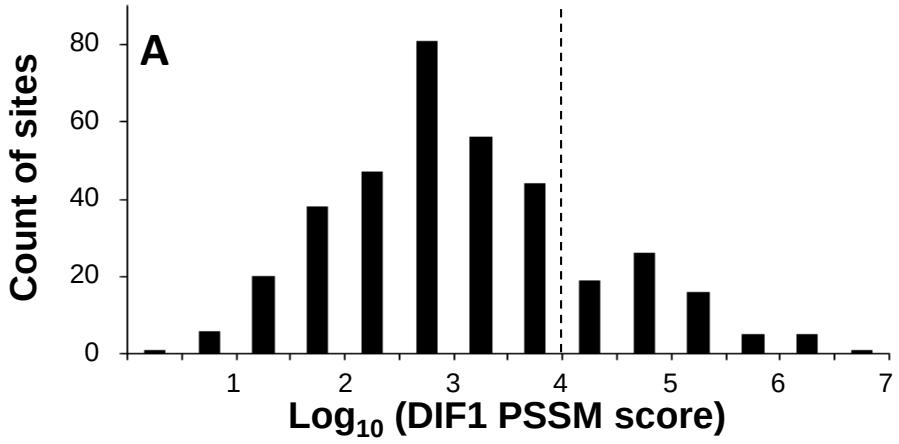
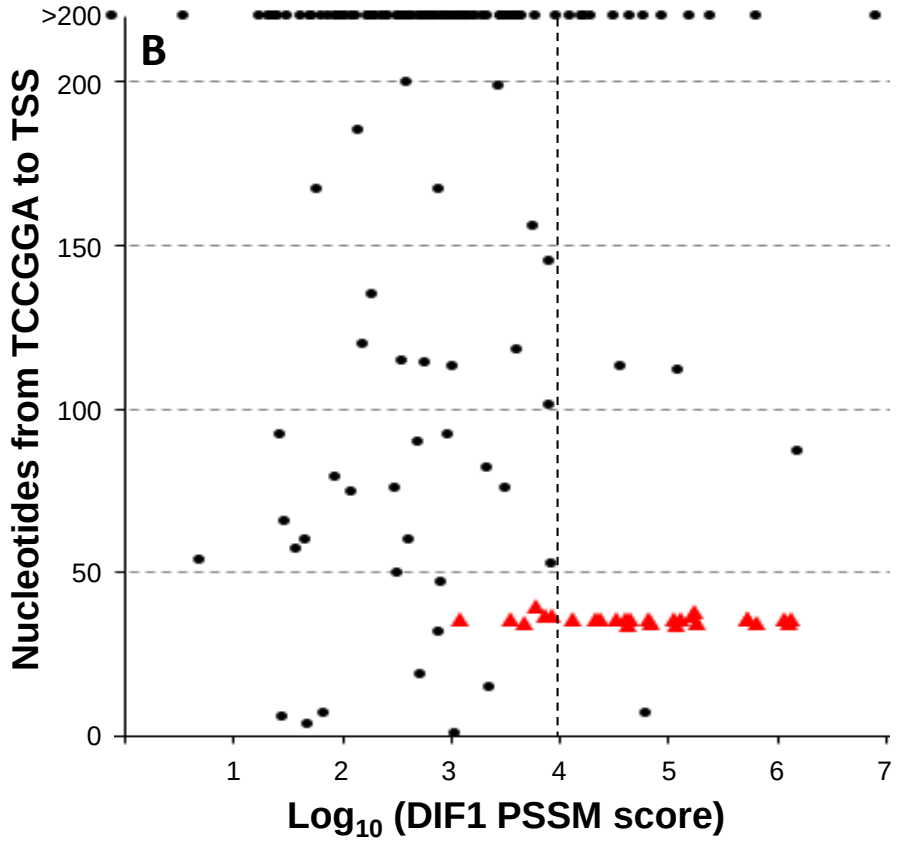
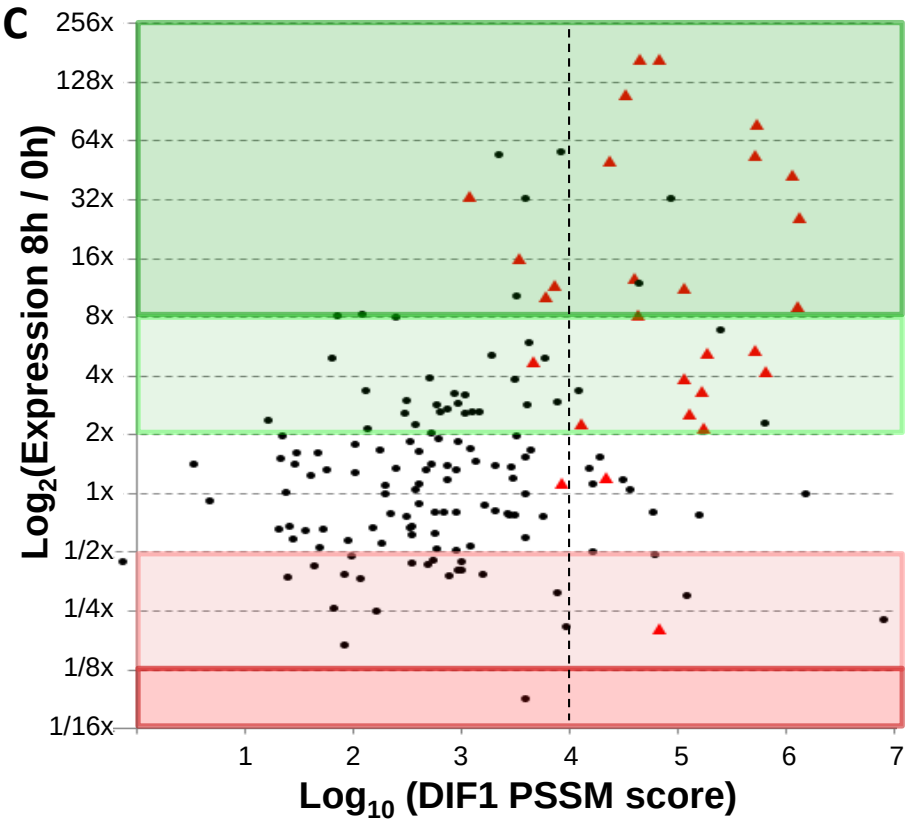
Supporting Information Figure S9: Characteristics of TCCGGA sites







D Characteristics of different classes of sites

	All sites	PSSM>4	-35 sites	-35 & PSSM>4
Total N	161	39	28	22
> 8x induced	25 (16%)	14 (36%)	16 (57%)	12 (55%)
> 2x induced	61 (38%)	25 (64%)	25 (89%)	20 (91%)
DIF-3 class	18 (11%)	13 (33%)	15 (54%)	12 (55%)
DIF-1 class	31 (19%)	17 (44%)	20 (71%)	16 (73%)
Avg PSSM score	3.23	5.02	4.82	5.14
Avg induction	0.86	1.83	3.33	3.41

Figure S9. Characteristics of TCCGGA sites in genome of *Anabaena* PCC 7120.

Candidate *patS* and *patX* genes are generally preceded by DIF1 motifs, TCCGGA or TCCGGA variants. Most TCCGGA sites in a genome are presumably situated randomly and have no regulatory significance. The conserved position of these sites relative to *patX* across filamentous cyanobacteria is itself a potent argument for function, but we seek here additional evidence, drawn from the characteristics of sequences surrounding the TCCGGA sites. If these characteristics are similar to those of TCCGGA sites shown by experiment to be associated with regulated transcription, that would be an additional argument for function. The goal of this analysis is therefore to determine whether the regulation of downstream transcription can be predicted from the sequences surrounding TCCGGA sites, using the well studied transcriptional start sites of *Anbaena* PCC 7120 (Mitschke *et* al., 2011). It is necessary to test this in order to give proper weight to the similarity scores used as a tool to identify relevant DIF1 motifs upstream from candidate *patS* and *patX* genes (Supporting Figures S7 and S8).

The context of each of the 365 TCCGGA sites found in the genome of *Anabaena* PCC 7120 were evaluated by means of a PSSM built from the contexts surrounding the 58 DIF1 motifs described by Mitschke *et al.* (2011) (see Experimental Procedures for explanation of PSSMs). The context was defined as the TCCGGA site plus 12 nucleotides upstream and 36 or 37 nucleotides downstream. In the 15 cases where a TCCGGA site was part of the 58 motifs of Mitschke *et al*., the sequence was evaluated with a PSSM in which the sequence had been removed. The maximal DIF1 PSSM score was determined (as decribed in the Experimental Procedures) for both strands. A match therefore has direction, with the score on one strand generally much greater than the score on the other. Note that the scores are displayed above on a Log_{10} scale. The larger the score, the closer the sequence matches the positional nucleotide frequencies of the DIF1-set of Mitschke *et al.* (2011), as compared to the nucleotide frequencies of a thousand nucleotides containing the TCCGGA site. All underlying data is available in Supporting Table S9.

(**A**) Distribution of PSSM scores. The DIF1 PSSM scores are shown in bins of 0.5 log units. Note the bimodal distribution.

(**B**) Distances of TCCGA from transcriptional start sites (TSSs) as a function of DIF1 PSSM scores. 162 TCCGGA sites were considered, all sites that are associated with TSSs reported by Mitschke, *et al.* (2011). The sequencing protocol of Mitschke *et al.* was not designed to include the small (~60 nt) transcripts of the noncoding RNA NsiR1, and in fact they reported transcripts from only one (TSS1) of the twelve copies, even though two others (TSS11 and TSS12) have been experimentally confirmed (Ionescu *et al.*, 2010). For this reason, only the TCCGGA site from *nsiR1* (TSS1) was included in the set. A TCCGGA site was considered to have an associated TSS if it points in the same direction as transcription from the TSS immediately downstream, regardless of distance. Distances greater than 200 nt were compressed to a single line. Special attention was given to those TCCGGA sites between 33 and 44 nt upstream from a TSS (red triangles) because this is the range of the DIF1 motifs reported by Mitschke *et al.* (2011) that are associated with differentiation-regulated promoters.

(**C**) Inducibility of TSSs associated with TCCGGA sites as a function of DIF1 PSSM scores. 161 TCCGGA sites were considered (one less than in (**B**) because one of the TSSs has unmeasurable expression after induction). Inducibility of a TSS is defined as the level of expression 8 h after N-deprivation relative to expression before deprivation, using data from Mitschke et al (2011). Red triangles indicate inducibility of the special class of sites described in (**B**).

(**D**) Characteristics of different TCCGGA sites. Induction is as defined in (**C**). This and other characteristics are presented for four categories of TCCGGA sites: all sites associated with TSSs, as defined in (**B**); sites with DIF1 PSSM scores greater than 4; -35-related sites, whose associated TSSs lie within the same range of nucleotides as DIF1 motif sites described by Mitschke *et al.* (2011); and the intersection of the last two categories. The DIF-3 class is defined by complex rules put forth by Mitschke *et al.* (2011) in their supplemental methods, most notably that those TSSs that are induced by N-deprivation at least 3 Log_2 units (8-fold) and there is lesser or no induction when HetR is absent. The DIF-1 class follows the same definition, except that only 2-fold induction is required.

It is evident from the data shown above that knowing that a TCCGGA site is about -35 from a TSS is a good predictor (~90%) that the TSS is induced at least 2-fold, regardless of whether this information is used alone or in combination with the DIF1 PSSM score. Predictions from the PSSM score alone is more modestly successful, with a success rate of 64%. However, 80% of the sites predicted from that score are -35 from the TSS. Perhaps the sequence characteristics of -35-associated sites don't carry over to TCCGGA sites that are more distant from associated TSSs, or perhaps there are few if any such sites in *Anabaena* PCC 7120.

The PSSM score therefore does not significantly increase the already high confidence that a TCCGGA site situated -35 from a TSS is inducible (or inducible and dependent on HetR). The position of the TSS may be suspected positional cues, as seen with *patX* upstream sequences. In cases where the TSS is unknown (e.g. upstream from *patS* genes outside of *Anabaena* PCC 7120), a PSSM score greater than 4 may be informative. While the alignment of upstream sequences of *patS* (Supporting Fig. S7) and *patX* (Supporting Fig. S8A) in heterocyst-forming cyanobacteria makes it likely that the PSSM will be informative outside of *Anabaena* PCC 7120, this remains to be demonstrated.