MICROBIOLOGY

RESEARCH ARTICLE

Harwood and Risser, *Microbiology* 2021;167:001111

DOI 10.1099/mic.0.001111

The primary transcriptome of hormogonia from a filamentous cyanobacterium defined by cappable-seq

Thomas V. Harwood and Douglas D. Risser*

Abstract

Hormogonia are motile filaments produced by many filamentous cyanobacteria that function in dispersal, phototaxis and the establishment of nitrogen-fixing symbioses. The gene regulatory network promoting hormogonium development is initiated by the hybrid histidine kinase HrmK, which in turn activates a sigma factor cascade consisting of SigJ, SigC and SigF. In this study, cappable-seq was employed to define the primary transcriptome of developing hormogonia in the model filamentous cyanobacterium *Nostoc punctiforme* ATCC 29133 in both the wild-type, and *sigJ*, *sigC* and *sigF* mutant strains 6 h post-hormogonium induction. A total of 1544 transcriptional start sites (TSSs) were identified that are associated with protein-coding genes and are expressed at levels likely to lead to biologically relevant transcripts in developing hormogonia. TSS expression among the sigma-factor deletion strains was highly consistent with previously reported gene expression levels from RNAseq experiments, and support the current working model for the role of these genes in hormogonium development. Analysis of SigJ-dependent TSSs corroborated the presence of the previously identified J-Box in the −10 region of SigJ-dependent promoters. Additionally, the data presented provides new insights on sequence conservation within the −10 regions of both SigC- and SigF-dependent promoters, and demonstrates that SigJ and SigC coordinate complex co-regulation not only of hormogonium-specific genes at different loci, but within an individual operon. As progress continues on defining the hormogonium gene regulatory network, this data set will serve as a valuable resource.

INTRODUCTION

Many species of filamentous cyanobacteria develop hormogonia, specialized filaments that are dedicated to motility [1]. The transition from the actively growing vegetative state to these transiently senescent filaments dedicated to motility includes a round of reductive cell division and substantial changes in cellular architecture [1], as well as the assembly of the type IV pilus motor that powers motility [2] and the production of a hormogonium-specific polysaccharide [3, 4]. Hormogonia facilitate dispersal and are capable of phototaxis [5], allowing them to migrate to more favourable light environments. Additionally, motile hormogonia can assemble into supracellular aggregates [6, 7] and are the infective unit for establishing nitrogen-fixing symbioses with several eukaryotic organisms [8, 9].

Current knowledge on the gene regulatory network (GRN) controlling hormogonium development has been defined primarily based on experiments with the model filamentous cyanobacterium *N. punctiforme* ATCC29133/PCC73122. In this organism, initiation of hormogonium development is controlled by the hybrid histidine kinase HrmK, which senses unknown signal(s) and is thought to activate an unidentified response regulator that in turn directly or indirectly promotes transcription of *sigJ* [10]. Subsequently, SigJ initiates a sigma factor cascade by activating transcription of two additional hormogonium-specific sigma factors, *sigC* and *sigF*, as well as a large number of other hormogonium-specific genes [11]. SigC also promotes the expression of a substantial number of hormogonium-specific genes, including those involved in reductive cell division, and may form an indirect positive feedback loop with *sigJ* by enhancing transcription of *hrmK* [10, 11]. The primary role of SigF appears to be promoting

Received 03 August 2021; Accepted 21 September 2021; Published 15 November 2021
Author affiliations: [1]Department of Biology, University of the Pacific, Stockton, CA 95211, USA.
***Correspondence:** Douglas D. Risser, drisser@pacific.edu
Keywords: cyanobacteria; hormogonia; development; cappable-seq; transcriptional start sites.
Abbreviations: CapSeq, cappable-sequencing; GRN, gene regulatory network; 5'-RACE, 5'-rapid amplification of cDNA ends; RRS, relative read score; TSS, transcriptional start site.
Cappable-seq sequence data were deposited in the NCBI GEO database (GSE180983).
Supplementary data is available with the online version of this article.

transcription of *pilA*, which encodes the major pilin of the type IV pilus system that powers motility [11]. Promoters dependent on SigJ contain a J-Box with the consensus sequence GGGAATACT in the proximity of the putative −10 regions, with the trio of Gs likely to comprise an extended −10 region that typically interacts with sigma factor domain 3, while the AATACT presumably interacts with sigma factor domain 2, and contains the critical residues that are flipped out of the double-stranded DNA to facilitate open complex formation [11].

While a recent study employing RNAseq substantially advanced the understanding of the role the sigma factor cascade plays in hormogonium development [11], the experiments employed are inadequate to confidently resolve transcriptional start sites (TSSs) to single-base resolution. The primary transcriptome has been defined for the closely related strain *Nostoc* sp. strain PCC 7120 [12], however, this strain has lost the ability to form hormogonia, likely due to prolonged laboratory culture [11], and is therefore insufficient to identify TSSs specifically expressed in hormogonia or the role of SigJ, SigC and SigF in their regulation. Here, the recently developed Cappable-seq (CapSeq) method for identifying the primary transcriptome in prokaryotic organisms [13] was employed to globally define transcriptional start sites in wild-type *N. punctiforme* and *sigJ*, *sigC* and *sigF* mutant strains.

METHODS

Strains and culture conditions

N. punctiforme ATCC 29133 and its derivatives, Δ*hrmK* (UOP139) [10], Δ*sigJ* (UOP132), Δ*sigC* (UOP131) and Δ*sigF* (UOP141) [11] were cultured in Allen and Arnon medium diluted fourfold (AA/4) [14], without supplementation of fixed nitrogen for all experimental conditions employed in this study, as previously described [15], with the exception that liquid cultures were supplemented with 4 mM sucralose and solid medium was supplemented with 10 mM sucralose, to inhibit hormogonium formation [16]. For hormogonium induction, the equivalent of 300 µg ml^{-1} chlorophyll *a* (Chl *a*) of cell material from cultures at a Chl *a* concentration of 10–20 µg ml^{-1} was harvested at 2000 ***g*** for 3 min, washed two times with AA/4 and resuspended in 50 ml of fresh AA/4 without sucralose.

Cappable-Seq

Total RNA was extracted from the equivalent of 300 µg ml^{-1} Chl *a* of cell material from each strain at 6 h following hormogonium induction, using previously published methods [15]. Cappable-seq library preparation was performed as previously described [13] by Vertis Biotechnologie AG and the libraries single-read sequenced on an Illumina NextSeq 500 system using 75 bp read length. Total reads for each library were as follows: wild-type=9380885; Δ*sigJ*=9916873, Δ*sigC*=9310957; Δ*sigF*=10544296. CapSeq data were deposited in the NCBI GEO database (GSE180983). Reads were mapped to the *N. punctiforme* genome and TSS were defined as previously described [13] following the protocols for TSS identification without a non-enriched control library (https://github.com/Ettwiller/TSS/). TSS were defined, using the library from the wild-type strain, as those with a relative read score (RRS) of 1 or higher and subsequently clustered using a cutoff of 5.

To discriminate between TSSs that were likely the result of low levels of spurious transcriptional activation and those that were likely to produce biologically relevant transcripts, the TSSs for 22 genes known to be involved in hormogonium development were manually identified. All 22 of these genes were associated with at least one TSS upstream of the translational start site with an RRS of 36.5 or higher, and therefore an RRS of 35 or higher was chosen as the cutoff for biologically relevant TSSs. TSSs were then associated with protein-coding genes based on the criteria that the TSS was within 500 bp upstream of the start codon of a gene and in the same orientation. This yielded a total of 1544 TSSs associated with protein-coding genes.

To determine the relative expression of these TSSs in each of the sigma-factor mutant strain libraries, each of these libraries was mapped to the *N. punctiforme* genome and TSS defined as described above (RRS of 1 or higher). The wild-type/sigma-factor mutant strain RRS ratio was calculated using the RRS for each TSS in each library, and this ratio was subsequently Log2 transformed. It should be noted that these ratios were calculated using the RRS at the nucleotide position for each TSS prior to clustering, as we found that clustering of TSS in the sigma-factor mutant strains often resulted in a positional shift of +/-1 nucleotide relative to the same TSS in the wild-type when the expression of a TSS was substantially reduced in one of the sigma-factor deletion strains. Hierarchical cluster analysis (average group linkage) was performed using Genesis [17]. WebLogo [18] was used to display nucleotide conservation within promoters from each cluster and MEME [19] (default settings with 'search given strand only' selected) was used to identify conserved motifs within the −10 regions (20 bp upstream of TSS). To scan for the F-Box motif in the *pilA* promoter regions from other cyanobacteria, the best hit homologue from BLAST searches using the *N. punctiforme* PilA sequence for query were identified, and the 100 bp immediately upstream of the start codon for each homologue were scanned using MEME (default settings with 'search given strand only' selected and with a maximum motif width of 12). Visualization of mapped reads was performed using Integrated Genomics Viewer [20].

RESULTS

Identification of *N. punctiforme* TSSs by Cappable-seq

To define the primary transcriptome in wild-type *N. punctiforme* and the three hormogonium-specific sigma factor mutants, Cappable-seq [13] was employed on a single biological replicate from each strain 6 h following hormogonium induction. This time point was chosen because both genes that show prolonged induction in developing hormogonia as well as those that are only transiently upregulated typically display robust increases in transcription at 6 h of development

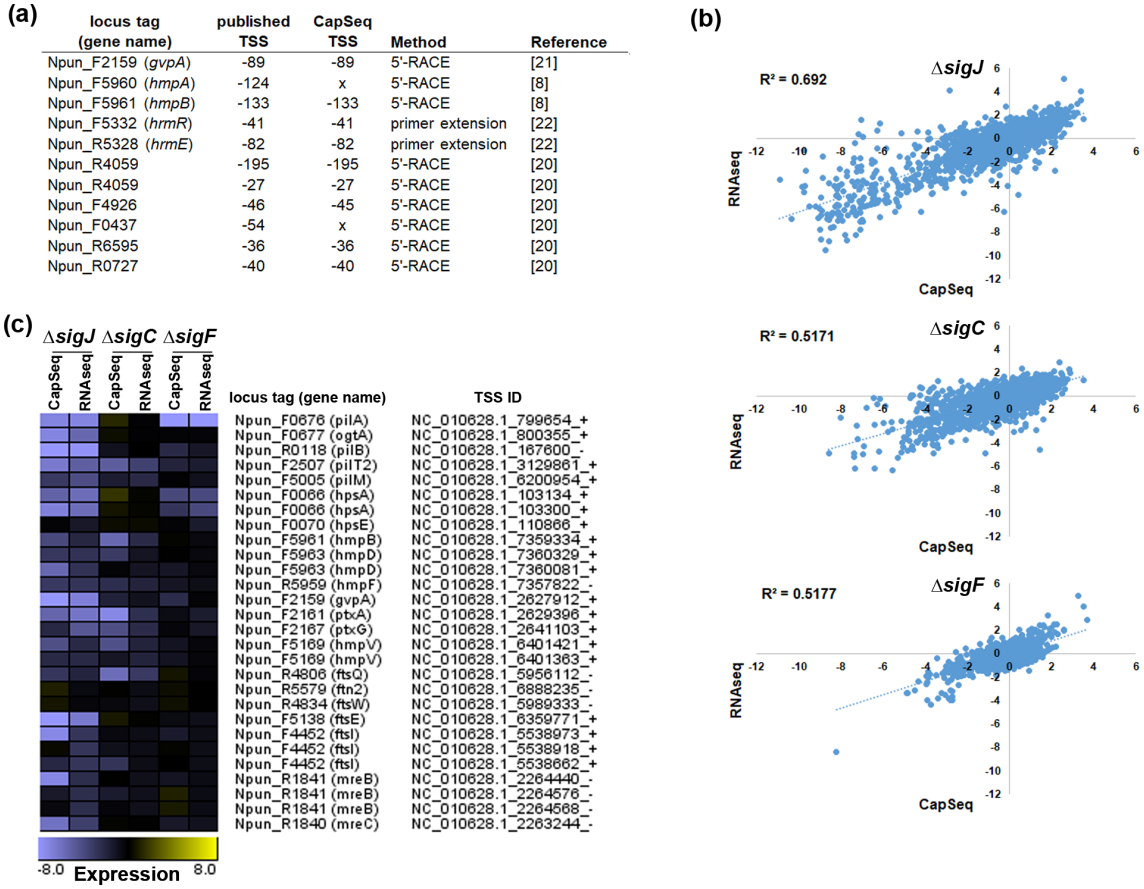

Fig. 1. Comparison of the CapSeq data set to previously published TSSs and RNAseq data. (a) A table comparing TSSs determined by CapSeq to previously published TSSs defined by primer extension or 5′RACE. (b) Comparison of TSS expression defined by CapSeq to total gene expression previously defined by RNAseq for all 1544 TSSs associated with protein-coding genes in each sigma-factor deletion strain. Expression=Log2(sigma-factor mutant strain/wild-type strain). (c) Comparison of TSS expression defined by CapSeq to total gene expression previously defined by RNAseq for selected genes in each sigma-factor deletion strain. Expression=Log2(sigma-factor mutant strain/wild-type strain).

[11]. Initially, TSSs were defined in the wild-type strain using a low-stringency cutoff, yielding a total of 12660 TSSs (Data Set S1, available in the online version of this article). A biology-driven approach was subsequently used to distinguish between TSSs that were likely to produce biologically relevant transcripts versus those that were the result of spurious low-level transcriptional activation. A set of 22 genes known to play key roles in hormogonium development and motility were selected, and their transcriptional start sites manually identified (Data Set S2). Collectively, each gene had at least one TSS with a relative read score (RRS) of 36.5 or higher, so an RRS score of 35 was chosen as the cutoff for TSSs deemed to be biologically relevant. These TSSs were subsequently assigned to specific protein-coding genes based on the criteria that the TSS was within 500 bp upstream of the gene's translational start site and in the same orientation. Collectively, 1544 TSS were identified and assigned to annotated protein-coding genes (Data Set S3). The expression of each of these TSSs was then compared between the wild-type and each sigma-factor deletion strain (Data Set S3).

To assess the quality of this data set, we first compared TSSs determined by CapSeq to 11 previously published TSSs defined by 5′-rapid amplification of cDNA ends (5′-RACE) or primer extension experiments (Fig. 1a) [8, 21–23]. Because several of these genes are not expected to be induced in developing hormogonia, we analysed the entire CapSeq data set of TSSs (Data Set S1), rather than just those with RRS scores above 35. Of the 11 TSSs identified in these publications, eight were an exact match and one was within 1 bp of a TSS at the same position in the CapSeq data set. Notably, the putative TSSs identified in previous publications that were missing from the CapSeq data set were both determined by 5′-RACE using experimental protocols that do not distinguish between bona fide TSSs and processing sites. These results indicate that the CapSeq experiments accurately identified TSSs, including those not expected to be induced at high levels under the experimental conditions employed, indicating that CapSeq is highly sensitive, and this data set can be informative for biological processes aside from hormogonium development, which is the focus of this study.

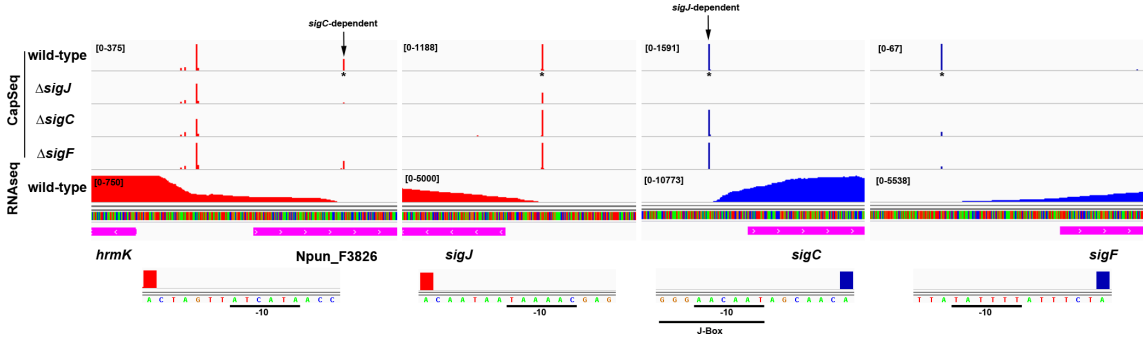

Fig. 2. TSSs of *hrmK* and hormogonium-specific sigma factors. Depicted is read coverage from CapSeq data for the wild-type and sigma-factor deletion strains, and RNAseq for the wild-type strain. For each gene, the sequence in the −10 region for TSS indicated by an asterisk is depicted below. Read coverage scale for CapSeq and RNAseq for each gene is indicated in brackets.

To further assess the quality of the CapSeq data set the expression of each TSS was compared to that of its associated gene as previously determined by RNAseq [11] in each sigma-factor mutant (Fig. 1b). For each strain, TSS expression was highly correlated with total transcript abundance. It should be noted that not all TSS expression levels would be expected to correlate closely with total gene expression because many genes may have more than one TSS, only some of which would be differentially regulated in each sigma-factor mutant. Manual inspection of 28 TSSs associated with 21 genes known to be upregulated in developing hormogonia also indicated a high degree of correlation between TSS levels and total levels of gene expression (Fig. 1c). For each of these genes, expression of at least one associated TSS was differentially regulated in the sigma-factor mutants in a manner similar to that observed for total gene expression. In cases where only a single TSS was associated with a given gene, such as *pilA* and *pilB*, TSS expression was highly similar to total gene expression. In contrast, for genes with multiple TSSs, such as *ftsI* and *mreB*, expression of some TSSs did not correlate with total gene expression, but at least one TSS in each case did, usually displaying a more pronounced change in expression compared to total gene expression. These findings indicate that the CapSeq data set, although derived from a single biological replicate, is highly correlated to the total gene expression levels determined by RNAseq from a biological triplicate data set, and therefore can be useful for identifying differentially expressed TSSs in each of the sigma-factor mutants. However, it should be noted that future CapSeq experiments incorporating additional replicates and experimental conditions, such as un-induced vegetative filaments or additional time points during hormogonium development may be necessary to support the conclusions derived from this data.

Investigation of TSSs in the hrmK/sigma-factor cascade

The GRN promoting hormogonium development is initiated by the hybrid histidine kinase HrmK, which in turn is thought to directly or indirectly lead to the phosphorylation of an unknown response regulator [10]. Once activated, this response regulator initiates transcription of *sigJ*, which then promotes transcription of *sigC* and *sigF* [11]. Each sigma factor subsequently activates a regulon of genes required for hormogonium development, and in the case of *sigC*, may form an indirect positive feedback loop with *sigJ* by enhancing transcription of *hrmK* [10, 11]. Therefore TSSs in the promoter regions for each of these genes were analysed (Fig. 2). The *hrmK* promoter is complex, with four distinct TSSs. Three of these are located in the intergenic region between *hrmK* and Npun_F3826, and do not appear to be affected by any of the three sigma factors. The fourth TSS is more distal, being located within the coding region of the upstream gene, and expression is drastically reduced in the Δ*sigJ* strain, and nearly undetectable in the Δ*sigC* strain, indicating that this TSS is stringently SigC dependent. Read coverage from RNAseq corresponds closely with the location of these TSSs. These data corroborate the finding from RNAseq experiments that SigC influences *hrmK* expression.

The promoter of *sigJ* contains a single TSS, the expression of which is moderately reduced in the *sigJ*-deletion strain, but not in any of the other sigma-factor mutants. The promoter for this TSS does not contain an obvious J-Box, and therefore is unlikely to be a direct target for SigJ. These results indicate that *sigJ* may be positively autoregulated indirectly, although it is also conceivable that deletion of the *sigJ*-coding region has a *cis*-acting effect on the upstream TSS as well, such as decreased mRNA stability due to a reduction in the length of the mRNA. The promoter region of *sigC* contains a single TSS, which is stringently dependent on SigJ and the associated −10 sequence is highly similar to the consensus J-Box (GGGAACAAT vs GGGAATACT for consensus J-Box), indicating that *sigC* is a direct target for regulation by SigJ. The promoter region for *sigF* also contains a single TSS. This TSS is undetectable in the *sigJ* mutant, and shows decreased abundance in the Δ*sigC* and Δ*sigF* strains. The −10 region for this TSS does not contain a consensus J-Box, indicating that it is not likely to be a

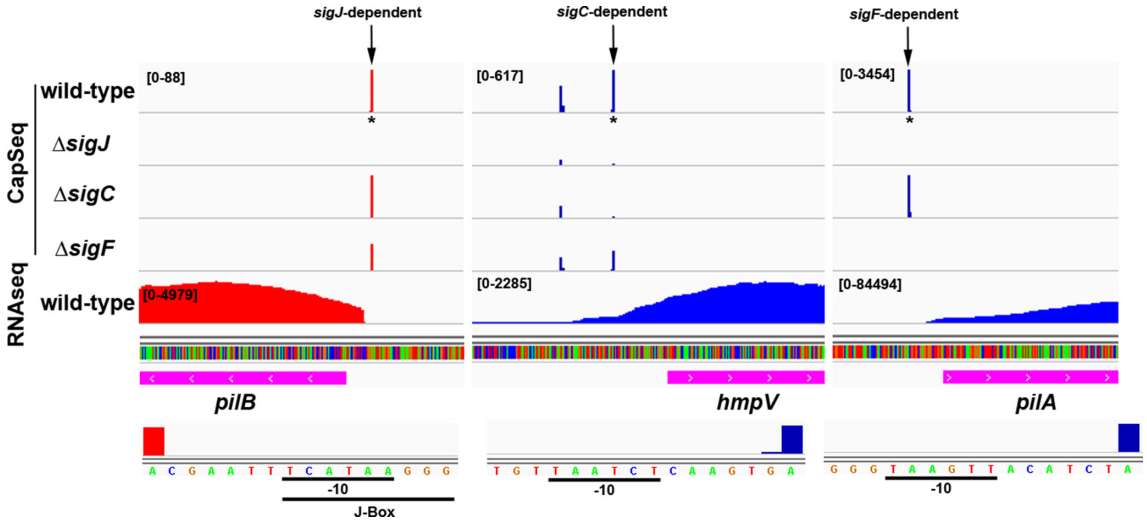

Fig. 3. TSSs of hormogonium-specific sigma-factor target genes. Depicted is read coverage from CapSeq data for the wild-type and sigma-factor deletion strains, and RNAseq for the wild-type strain. For each gene, the sequence in the −10 region for TSS indicated by an asterisk is depicted below. Read coverage scale for CapSeq and RNAseq for each gene is indicated in brackets.

target for direct regulation by SigJ. Collectively, these results support the previously posited model for the involvement of *hrmK* and these sigma factors in the hormogonium GRN.

Investigation of the TSSs for downstream targets of the hormogonium-specific sigma factors

We next investigated the promoter regions for previously identified downstream targets for each sigma factor (Fig. 3). Upregulation of *pilB* in developing hormogonia is stringently dependent on SigJ and the promoter region contains a consensus J-Box [11]. Analysis of the promoter region by CapSeq indicated the presence of a single TSS that was absent in the Δ*sigJ* strain, and the −10 region aligned precisely with the previously identified J-Box. The *hmpV* gene was identified as dependent on SigC for upregulation in hormogonia [11]. The promoter region for *hmpV* contains a pair of TSSs. The more distal TSS is present in all of the sigma-factor deletion strains, although expression is moderately reduced, whereas the more proximal TSS is substantially reduced in both the *sigJ* and *sigC* mutants, consistent with this TSS being most directly dependent on SigC. Notably, this finding is consistent with the results observed from an HmpV-GFP fusion expressed from the native chromosomal locus, which shows HmpV-GFP is present in vegetative filaments but accumulates at much higher levels upon hormogonium induction [24]. The only gene clearly found to be stringently dependent on SigF in developing hormogonia is *pilA* [11]. The promoter region for *pilA* contains a single TSS with drastically reduced expression in both the *sigJ* and *sigF* mutants consistent with direct regulation by SigF. The −10 region for this TSS contains a sequence similar to that of a J-Box, but distinct from any of the −10 regions for other TSSs identified as directly SigJ dependent (discussed further below). This sequence may represent an F-Box that specifically targets SigF to its cognate promoter.

For each of the genes investigated, the TSS analysis is highly consistent with the published RNAseq data, supporting the regulatory model previously proposed.

Investigation of the hmp chemotaxis-like system locus, an example of complex operon regulation

The *hmp* chemotaxis-like system locus consists of *hmpA-E*, and *hmpF*, which is upstream and in the opposite orientation to the other *hmp* genes. A previous study indicated that *hmpA-E* comprise an operon, with transcription initiated at one of two TSSs [8]. One TSS is upstream of *hmpA* and produces a transcript including *hmpA-E*, while the other TSS is upstream of *hmpB*, and therefore produces a transcript encoding *hmpB-E*. However, results from RNAseq indicated that while *hmpA* and *hmpB* are more dependent on SigC for expression, *hmpC-E* are more dependent on SigJ [11]. This observation could be accounted for if there are one or more previously unidentified TSSs within this locus. To determine this the TSSs identified by CapSeq within this locus were investigated (Fig. 4). As previously noted, none of the TSSs identified by CapSeq corresponded to the one previously identified by 5′-RACE upstream of *hmpA*. There were several other potential TSSs in the intergenic region upstream of *hmpA*, but all of them were below the 35 RRS threshold for likely biological relevance in hormogonia, and read coverage for *hmpA* from RNAseq was also extremely low compared to *hmpB-E*. In contrast, two closely spaced TSSs were detected in the intergenic region upstream of *hmpB*, one of which corresponded precisely with that previously identified by 5′-RACE. Expression of both TSSs was substantially reduced upon deletion of either *sigJ* or *sigC*, indicating that these TSSs are sigC-dependent. A third, previously unidentified TSS was also found within the *hmpC*-coding region. This TSS is stringently dependent upon SigJ and the −10 region

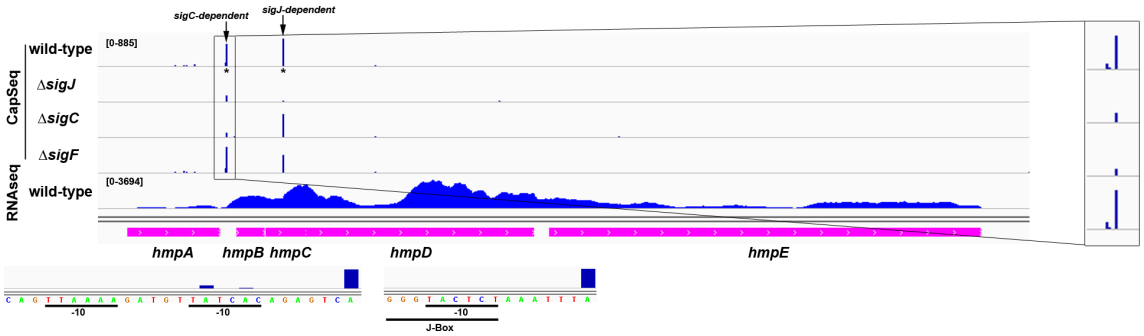

Fig. 4. TSSs of the *hmp* locus. Depicted is read coverage from CapSeq data for the wild-type and sigma-factor deletion strains, and RNAseq for the wild-type strain. For each gene, the sequence in the −10 region for TSS indicated by an asterisk is depicted below. Read coverage scale for CapSeq and RNAseq for each gene is indicated in brackets.

contains a putative J-Box. These findings indicate that there is complex co-regulation of gene expression by SigJ and SigC during hormogonium development, even within operons of genes encoding proteins with related functions.

Global analysis of TSSs

To identify consensus promoter sequences among groups of TSSs a hierarchical cluster analysis was performed based on the expression profiles in each of the sigma-factor deletion strains (Fig. 5a). Clusters with similar expression profiles were subsequently manually assigned cluster numbers. The TSS for *pilA* exhibited a unique expression profile, with drastic reduction in both the Δ*sigJ* and Δ*sigF* strains, and did not cluster with any other TSSs, indicating that SigF has a remarkably limited regulon (cluster 1, Data Set S3). Cluster 7 appears to be dependent on SigJ, given that expression of these TSSs is substantially downregulated in the Δ*sigJ* strain but not the other sigma-factor mutants. Two other small clusters, 4 and 15 are also downregulated primarily in the Δ*sigJ* strain, but to a lesser extent than cluster 7. Both clusters 5 and 14 may be dependent on SigC, as downregulation of these TSSs is observed in both the Δ*sigJ* and Δ*sigC* strains, but is more pronounced in Δ*sigC*. The primary distinction between these clusters is the amplitude of the downregulation, with that of cluster 5 more dramatic than that of cluster 14.

The promoter regions upstream of each TSS were used to generate sequence logos for each cluster (Fig. 5b). Because only a single TSS belonged to cluster 1, no sequence logo could be generated. Promoters from all of the other clusters contained an AT-rich sequence within the −10 region, but there was generally little sequence conservation within the putative −35 region for any of them. As expected, the SigJ-dependent promoters in cluster 7 contained a G-rich sequence immediately upstream of the −10 region. Notably, the −10 sequences from the putative SigC-dependent clusters 5 and 14 appeared more similar to one another than any of the other clusters.

The promoter sequences 100 bp upstream of each TSS for each cluster were subsequently subjected to MEME analysis to search for conserved motifs, however, only the SigJ-dependent cluster 7 contained a consensus motif that was identified in greater than 50% of the sequences, and this motif was identical to the previously identified J-Box. A more targeted search within the −10 regions using the 20 bp most proximal to the TSS identified motifs conserved within greater than 50% of the sequences in 12 of the 16 clusters analysed (Fig. 6a). For the SigJ-dependent cluster 7, the previously identified J-Box was found. Cluster 15, which also appears to be dependent on SigJ, contains a motif similar to the J-Box, including a highly conserved trio of Gs, although this motif was only found in slightly less than half of the promoters. As described above, the −10 region for the *pilA* TSS is similar, but distinct from that found for SigJ-dependent TSSs. Manual inspection of the promoters from the SigJ-dependent cluster seven did not reveal any promoter sequences that were identical to the −10 sequence for that of the sigF-dependent *pilA* promoter (Data Set S3). However, expression of many TSSs in the SigJ-dependent cluster was reduced in the Δ*sigF* strain (Fig. 5a), although to a much lesser extent than in the Δ*sigJ* strain, implying that there may be some overlap in promoter recognition between SigJ and SigF.

In most of the other clusters, a similar motif with the consensus sequence TANNNT was present, with the A and T most proximal to the TSS showing nearly absolute conservation. It is likely that these residues are the ones that are flipped out by the sigma factor to initiate open complex formation. However, in the SigC-dependent clusters 5 and 14, there was a marked reduction in the conservation of the T most proximal to the TSS, especially in cluster 5. Therefore, the tolerance for nucleotides other than T in this position may be a hallmark of SigC-dependent promoters.

In an additional attempt to define a consensus sequence for SigF-dependent promoters a comparative genomics approach was applied based on the rationale that strains encoding both *sigF* and *pilA* might exhibit similar regulation and therefor conserved motifs within the *pilA* promoter region. Of 55 strains investigated, 44 contained a motif nearly identical to the putative F-Box within the −10 region of the *N. punctiforme pilA* promoter (Fig. 6b, Data Set S4). The consensus sequence of this motif was GGGTAAGTT, with

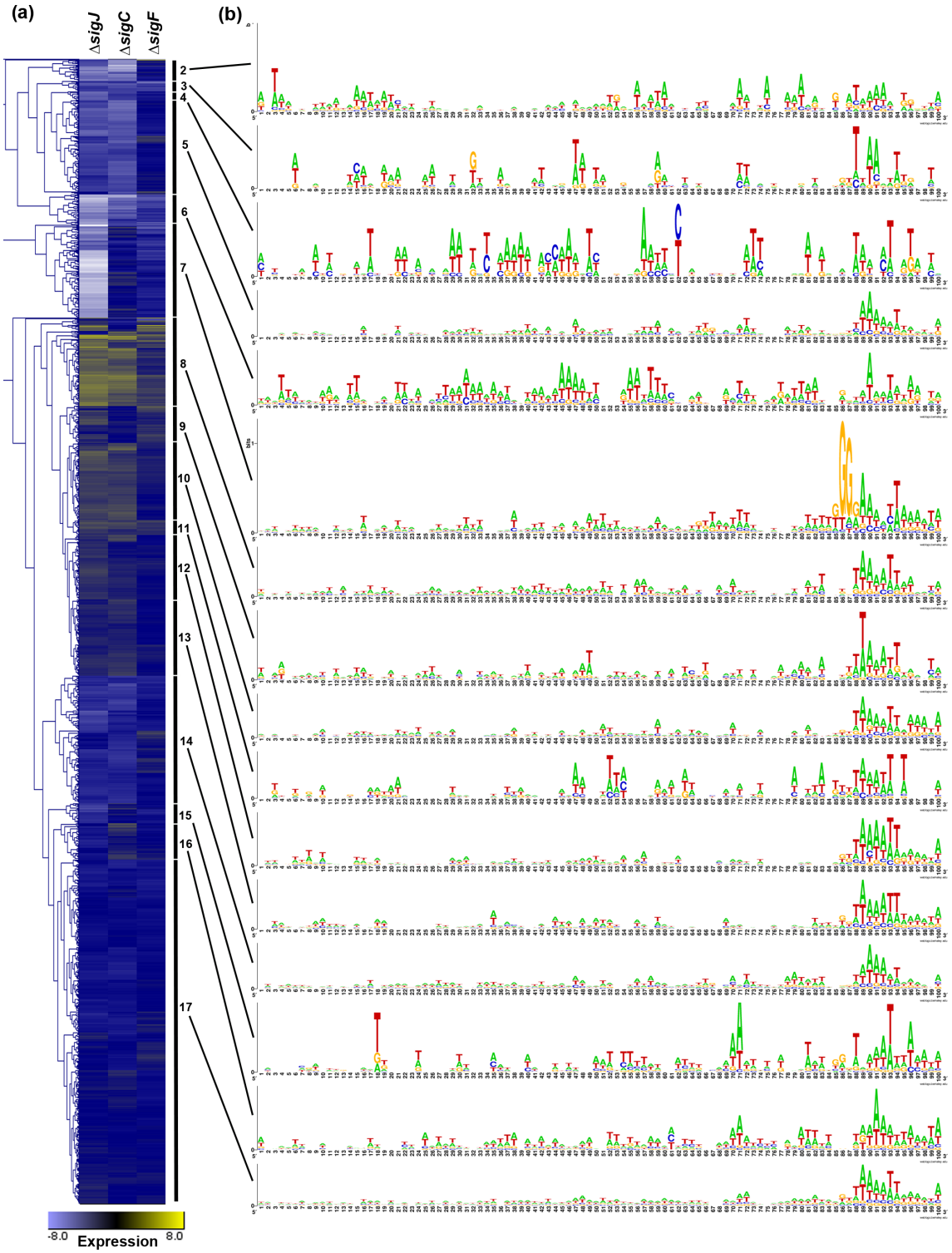

Fig. 5. Global analysis of TSS expression in sigma-factor mutants and conservation of promoter sequences. (a) A heat map depicting the expression of each TSS in each sigma-factor mutant strain. TSSs are organized based on hierarchical cluster analysis. Expression=Log2(sigma-factor mutant strain/wild-type strain). (b) Sequence conservation among promoter regions from different clusters with similar expression profiles.

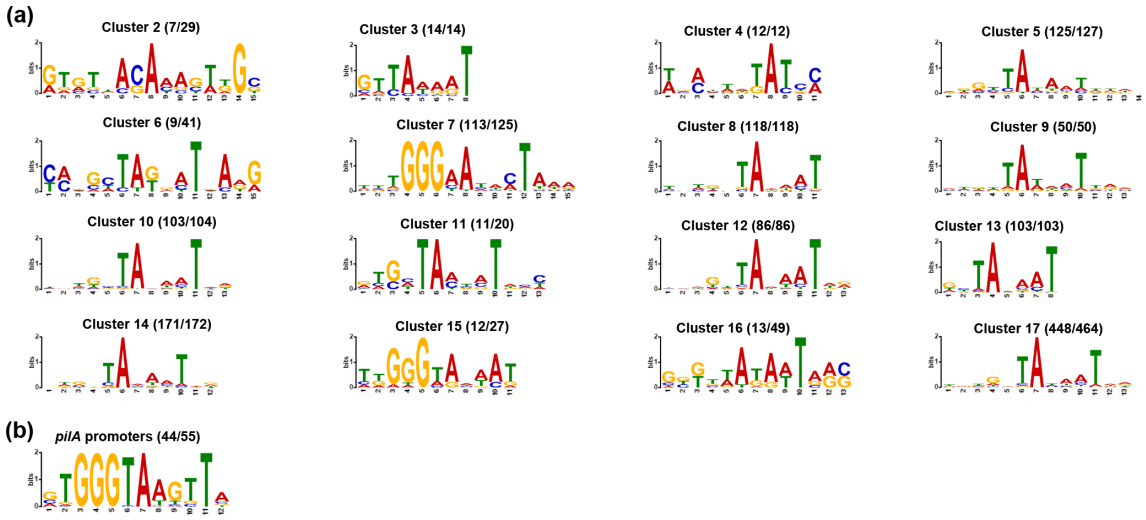

Fig. 6. Identification of conserved motifs among TSS with similar expression profiles. (a) The −10 region (20 bp upstream of TSS) for promoters from each cluster defined in Fig. 5 were subjected to MEME analysis to search for conserved motif. The most frequently conserved motif within each cluster is depicted. The numbers indicated in parentheses=number of promoters containing motif/total number of promoters in cluster. (b) The *pilA* promoter regions (100 bp upstream of the start codon) from 55 cyanobacteria encoding both *sigF* and *pilA* were subjected to MEME analysis to search for conserved motifs. The most frequently conserved motif is depicted. The numbers indicated in parentheses=number of promoters containing motif/total number of *pilA* promoters scanned.

the underlined residues showing nearly absolute conservation. Furthermore, in *Synechocystis* sp. strain PCC 6803, where the *pilA* promoter has been mapped [25], the position of the F-Box aligns precisely with the −10 region. The *pilA* promoter region of seven cyanobacteria known to encode homologues of *N. punctiforme* PilA [2], but lacking SigF, were also scanned for this motif, but none contained a sequence with the absolutely conserved residues of the putative F-Box (Data Set S4). This data lends additional support to the theory that the F-Box is important for promoter targeting by SigF in both *N. punctiforme* as well as other cyanobacteria.

DISCUSSION

The results from this study support the current working model for the role of *hrmK, sigJ, sigC* and *sigF* in the hormogonium gene regulatory network and provide further evidence that the previously identified J-Box motif comprises a −10 and extended −10 region present in SigJ-dependent promoters. Additionally, the data presented provides new insight on target promoters for SigC- and SigF-dependent genes. For SigF, the target regulon is remarkably limited, consisting of only a single gene demonstrating stringent dependence, *pilA*, and the promoter for this gene contains a −10 sequence remarkably similar, yet divergent from known SigJ-target promoters. We propose that this sequence, herein designated an F-Box, is critical for SigF promoter targeting. The presence of this motif in the *pilA* promoter regions from other cyanobacteria containing both *sigF* and *pilA* lend further support to this theory although additional experiments will be required to definitively confirm the essentiality of both the J-Box and F-Box in promotor targeting by their cognate sigma factors.

The −10 region for most other promoters, including those dependent on SigC, contain a consensus motif TANNNT, but promoters that are most stringently dependent on SigC appear to have a greater tolerance for other nucleotides in the position of the T most proximal to the TSS. This tolerance could be critical for the specificity of SigC for its target promoters, but like the putative F-Box, more experiments will be needed to verify this hypothesis.

SigJ and SigF are homologous and thought to have arisen from a gene duplication event [26], and the similarity in promoter recognition sequences is consistent with this theory. With the exception of the marine picocyanobacteria, most cyanobacterial genomes encode both *sigJ* and *sigF*, while a smaller subset encodes *sigF* alone [11]. It is likely that in cyanobacteria that harbour only *sigF*, this sigma factor promotes transcription of a much larger set of genes that includes those identified as *sigJ*-dependent in *N. punctiforme*. In fact, a recent study on a *sigF* mutant in the unicellular cyanobacterium *Synechocystis* PCC 6803 indicates that SigF has a much larger regulon in this cyanobacterium [27]. The rather unusual association of a sigma factor with a single target gene may be due to the requirement for extremely high levels of PilA protein in hormogonia, given that the *pilA* TSS is the second most highly expressed TSS in the entire genome during hormogonium development (Data Set S3).

While the present study focused on defining the TSS for protein-coding genes expressed during hormogonium development, for which a robust complementary total RNAseq data set was already available [11], the CapSeq data set provided here is likely to be of use for additional analyses moving

forward. The fact that CapSeq was able to identify previously described TSS not be expected to be induced in hormogonia indicates that the data can be of use in exploring additional biological phenomenon in *N. punctiforme* aside from hormogonium development. Moreover, the combination of both the CapSeq and previously published RNAseq datasets should allow for the future annotation and quantitation of non-coding RNAs in *N. punctiforme*. Thus, moving forward, this data should be valuable in furthering our understanding of hormogonium development as well as other aspects of the biology of filamentous cyanobacteria.

Funding information
This work was supported by NSF award no. 1753690 to D.D.R.

Author contributions
T.V.H. and D.D.R. designed experiments, performed experiments, analysed data and prepared the manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Meeks JC, Campbell EL, Summers ML, Wong FC. Cellular differentiation in the cyanobacterium Nostoc punctiforme. *Arch Microbiol* 2002;178:395–403.
2. Khayatan B, Meeks JC, Risser DD. Evidence that a modified type IV pilus-like system powers gliding motility and polysaccharide secretion in *Filamentous cyanobacteria*. *Mol Microbiol* 2015;98:1021–1036.
3. Risser DD, Meeks JC. Comparative transcriptomics with a motility-deficient mutant leads to identification of a novel polysaccharide secretion system in *Nostoc punctiforme*. *Mol Microbiol* 2013;87:884–893.
4. Zuniga EG, Boateng KKA, Bui NU, Kurnfuli S, Muthana SM, *et al*. Identification of a hormogonium polysaccharide-specific gene set conserved in *Filamentous cyanobacteria*. *Mol Microbiol* 2020;114:597–608.
5. Castenholz RW. Motility and taxis. In: Carr NG and Whitton BA (eds). *The Biology of Cyanobacteria*. Oxford: Blackwell Science; 1982. pp. 413–419.
6. Shepard RN, Sumner DY. Undirected motility of filamentous cyanobacteria produces reticulate mats. *Geobiology* 2010;8:179–190.
7. Tzubari Y, Magnezi L, Be'er A, Berman-Frank I. Iron and phosphorus deprivation induce sociality in the marine bloom-forming cyanobacterium *Trichodesmium*. *ISME J* 2018;12:1682–1693.
8. Risser DD, Chew WG, Meeks JC. Genetic characterization of the hmp locus, a chemotaxis-like gene cluster that regulates hormogonium development and motility in *Nostoc punctiforme*. *Mol Microbiol* 2014;92:222–233.
9. Pratte BS, Thiel T. Comparative genomic insights into culturable symbiotic cyanobacteria from the water fern Azolla. *Microb Genom* 2021;7.
10. Zuniga EG, Figueroa NM, Gonzalez A, Pantoja AP, Risser DD, *et al*. The hybrid histidine kinase HrmK is an early-acting factor in the hormogonium gene regulatory network. *J Bacteriol* 2020;202:e00675-19.
11. Gonzalez A, Riley KW, Harwood TV, Zuniga EG, Risser DD. A tripartite, hierarchical sigma factor cascade promotes hormogonium development in the *Filamentous cyanobacterium Nostoc punctiforme*. *mSphere* 2019;4:e00231-19.
12. Mitschke J, Vioque A, Haas F, Hess WR, Muro-Pastor AM. Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in Anabaena sp. PCC7120. *Proc Natl Acad Sci U S A* 2011;108:20130–20135.
13. Ettwiller L, Buswell J, Yigit E, Schildkraut I. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics* 2016;17:199:199-z.:.
14. Allen MB, Arnon DI. Studies on nitrogen-fixing blue-green algae. I. growth and nitrogen fixation by anabaena cylindrica lemm. *Plant Physiol* 1955;30:366–372.
15. Campbell EL, Summers ML, Christman H, Martin ME, Meeks JC. Global gene expression patterns of *Nostoc punctiforme* in steady-state dinitrogen-grown heterocyst-containing cultures and at single time points during the differentiation of akinetes and hormogonia. *J Bacteriol* 2007;189:5247–5256.
16. Splitt SD, Risser DD. The non-metabolizable sucrose analog sucralose is a potent inhibitor of hormogonium differentiation in the filamentous cyanobacterium *Nostoc punctiforme*. *Arch Microbiol* 2015;198:137–147.
17. Sturn A, Quackenbush J, Trajanoski Z. Genesis: cluster analysis of microarray data. *Bioinformatics* 2002;18:207–208.
18. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res* 2004;14:1188–1190.
19. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;37:W202-8.
20. Thorvaldsdóttir H, Robinson JT, Mesirov JP. tegrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;14:178–192.
21. Bell N, Lee JJ, Summers ML. Characterization and in vivo regulon determination of an ECF sigma factor and its cognate anti-sigma factor in *Nostoc punctiforme*. *Mol Microbiol* 2017;104:179–194.
22. Campbell EL, Hagen KD, Chen R, Risser DD, Ferreira DP, *et al*. Genetic analysis reveals the identity of the photoreceptor for phototaxis in hormogonium filaments of *Nostoc punctiforme*. *J Bacteriol* 2015;197:782–791.
23. Campbell EL, Wong FCY, Meeks JC. DNA binding properties of the HrmR protein of *Nostoc punctiforme* responsible for transcriptional regulation of genes involved in the differentiation of hormogonia. *Mol Microbiol* 2003;47:573–582.
24. Riley KW, Gonzalez A, Risser DD. A partner-switching regulatory system controls hormogonium development in the filamentous cyanobacterium *Nostoc punctiforme*. *Mol Microbiol* 2018;109:555–569.
25. Asayama M, Imamura S. Stringent promoter recognition and autoregulation by the group 3 sigma-factor SigF in the cyanobacterium *Synechocystis* sp. strain PCC 6803. *Nucleic Acids Res* 2008;36:5297–5305.
26. Imamura S, Asayama M. Sigma factors for cyanobacterial transcription. *Gene Regul Syst Bio* 2009;3:65–87.
27. Flores C, Santos M, Pereira SB, Mota R, Rossi F, *et al*. The alternative sigma factor SigF is a key player in the control of secretion mechanisms in Synechocystis sp. PCC 6803. *Environ Microbiol* 2019;21:343–359.

Edited by: M. Brockhurst and J. P. J. Hall