Association of multicellularity in cyanobacteria with the transcription factor HetR and its putative regulator PatX, an RGSGR pentapeptide-containing protein

Jeff Elhai[1] and Ivan Khudyakov[2]

1. *Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA*
2. *All-Russia Research Institute for Agricultural Microbiology, Saint-Petersburg 196608, Russia*

Introduction

In 1952, Alan Turing presented a theoretical mechanism by which homogenous cells could be transformed into patterns of differentiated cells owing to the action of two interacting regulators with different diffusion rates.[1] The idea was refined by Gierer and Meinhardt[2,3] into the model shown conceptually in **Figure 1A**, relying on the actions of a slowly-diffusing, autocatalytic regulator (R) and a rapidly diffusing suppressor molecule (S). While the simplicity of the model is appealing, until recently, there has been little evidence from multicellular eukaryotes to support the theory.[4]

The most compelling biological case for the relevance of the model comes arguably from the multicellular prokaryote, ~~*Anabeana*~~ *Anabaena* PCC 7120, which differentiates specialized cells at semi-regular intervals along its filaments (**Figure 2**). These cells, called heterocysts, provide the conditions required for nitrogen-fixation in the presence of molecular oxygen.[5,6] ~~*Anabeana*~~ *Anabaena* PCC 7120 has been shown to synthesize two proteins, HetR and PatS, that seem to possess the characteristics called for by Gierer and Meinhardt's R and S morphogens. In addition, a third component, NtcA, ties the morphogenetic machinery to the nitrogen status of cells, as is physiologically appropriate, and a fourth, HetN, serves as an S morphogen tailored to pattern maintenance rather than creation. Recent models of heterocyst differentiation have expanded on the ideas of Turing and Gierer and Meinhardt, acknowledging the non-diffusibility of HetR, interactions amongst multiple actors, and their discrete concentrations in a series of cells.[7,8,9] A prevailing model of the regulation of heterocyst differentiation is summarized in **Figure 1B** and fleshed out below.

Comment [j1]: IK: Anabaena!!
JE: Global replace to fix all. Just so you don't get too worried, I made the mistake only once, in a global replace of Ana7120 with Anabeana.

The gene encoding HetR was identified as part of a hunt for mutants of ~~*Anabeana*~~ *Anabaena* PCC 7120 unable to sustain heterocyst differentiation,[10] and the protein was soon found to have the characteristics expected from an R morphogen. HetR appears to be a master regulator of heterocyst differentiation. First, differentiation was abolished by a point mutation in HetR (S179N) and a deletion or disruption of the gene.[11,12] Second, ~~3~~three-fold more heterocysts were formed (including multiple contiguous heterocysts seldom seen in wild-type *Anabaena*) when HetR was expressed from a multicopy plasmid,[11] from a regulatable promoter,[13] or from a mutant allele~~variant~~ of HetR, R223W.[14] HetR ~~is required for~~affects the expression of hundreds of genes, with heightened expression after nitrogen deprivation and repression in nitrogen-replete medium.[15,16] Expression of HetR is autocatalyic -- the protein is required for the increase in its own synthesis after nitrogen deprivation[12,13,17] HetR acts as a DNA-binding protein,[18,19] in a

tetrameric state that is regulated by phosphorylation,[20] and is in a positive feedback loop with NtcA.[21]

The characteristics of PatS protein is suggestive of its functioning as an S morphogen. PatS was discovered from the ability of a small DNA fragment on a multicopy plasmid to inhibit heterocyst differentiation in ~~*Anabeana*~~*Anabaena* PCC 7120.[22] Inhibition required the expression of an 11- to 17-amino acid open reading frame (ORF), called *patS*. Strains lacking this ORF and mutants with ORFs altered in one of the last five codons produced aberrant, though non-random, spacing of heterocysts, including multiple contiguous heterocysts.[22~~22~~~~21~~,23] Overexpression of *patS* reduces transcription from an inducible *hetR* promoter.[24] Exogenous application of a pentapeptide consisting of the last five amino acids of PatS (RGSGR) blocks heterocyst differentiation[22~~22~~~~21~~] and reduces the level of HetR protein.[25] This peptide also binds to HetR protein, preventing its binding to DNA.[18~~18~~~~17~~,26] Maximal binding affinity is achieved by a six-amino acid peptide ending in RGSGR, where the identity of the first amino acid is not critical.[26~~26~~~~25~~] With the finding that PatS expression is localized to developing cells[23~~2~~~~22~~] and is dependent upon HetR,[18~~18~~~~17~~] the matching of PatS characteristics to those of an S morphogen is complete, with one exception: diffusion. Clever experiments have suggested the diffusion of PatS-derived signals to adjacent cells,[25~~25~~~~24~~,27,28] but direct evidence for diffusion of any morphogen in ~~*Anabeana*~~*Anabaena* PCC 7120 remains elusive.

A second putative S morphogen was recognized within the protein HetN, initially identified in a similar fashion as PatS: (1) its presence on a multicopy plasmid suppressed heterocyst differentiation in ~~*Anabeana*~~*Anabaena* PCC 7120, and (2) its interruption led to multiple contiguous heterocysts.[29,30] Although *hetN* encodes a protein similar to short chain dehydrogenases (Pfam PF00106) and polyketide synthases (Pfam PF008659), most of the protein could be deleted without affecting its ability to suppress heterocyst differentiation.[31] However, an RGSGR sequence found within HetN[32] proved to be essential for the protein's effectiveness as a suppressor.[31~~31~~~~30~~,33] HetN differs from PatS in two important respects. First, its expression starts at a late stage of differentiation and persists in mature heterocysts (as opposed to induction at an early stage and downregulation in mature heterocysts).[34,35] Second, HetN is important in the maintenance of the pattern of heterocysts but not its initial formation.[34~~34~~~~33~~]

The functions of HetR, PatS, and HetN in *Anabaena* PCC 7120, combined with the Turing/Meinhardt model for pattern formation (**Figure 1B**), provide an appealing explanation for the appearance of spaced heterocysts in response to nitrogen deprivation. However, there are several observations that do not obviously square with this view. HetN-like proteins bearing the RGSGR motif are found in only a small fraction of heterocyst-forming cyanobacteria.[33~~33~~~~32~~] ORFs capable of producing a PatS-like protein were also reported to be absent in heterocyst-forming *Cylindrospermopsis raciborskii* CS-505[36] and *Anabaena* 90.[37] Finally, proteins antigenically similar to HetR are found even in filamentous cyanobacteria that don't make heterocysts,[38] consistent with the presence in non-heterocyst-forming cyanobacteria of DNA that hybridized to a hetR probe.[11] Cyanobacteria without the postulated machinery evidently produce spaced heterocysts, and cyanobacteria with at least part of the machinery do not. These observations prompted us to look systematically in cyanobacterial genomes for genes that may encode the proteins that make up the machinery of the Turing/Meinhardt model.

Methods

Cyanobacterial genomes and genome analysis

127 cyanobacterial genomes, including plasmids, from all major groupings, were accessed through BioBIKE.[39] 11 additional cyanobacterial genomes were also considered specifically with regards to HetR and/or PatX. The origins and other characteristics of these genomes are shown in **Suppl. Table S1**.

Phylogenetic trees

Organismal trees were built by analyzing concatenated alignments of 29 proteins found in all 127 cyanobacteria considered in this study. The names and coordinates of the orthologous proteins for each organism are given in **Suppl. Table S2**. Most of the proteins were readily obtained by the BioBIKE's ORTHOLOG-OF function, but in 16 cases (0.4% of the total number), a protein was not found in an organism, either because it had not been annotated or the ORF was broken by an apparent frame shift, either in Nature or in the sequencing and assembly of the genome. In such cases (noted in **Suppl. Table S2**), the ORF was detected using TBlastN[40] via the SEQUENCE-SIMILAR-TO function (protein vs translated DNA) and repaired digitally. 21 additional proteins (0.6% of the total number) had start codons apparently miscalled, truncating the annotated protein relative to other orthologs. In these cases the gene sequence was extended upstream to the presumably correct start codon matching those used by orthologs.

Alignments of each set of proteins (sets provided in **Suppl. Table S3**) were made through Clustal W[41] accessed within BioBIKE and concatenated using an ad hoc BioBIKE script. The most informative columns were extracted using Gblocks,[42] and the final tree was made by PhyML 3.0,[43] with LG as the substitution model, NNI as the type of tree improvement, and 100 bootstraps. The tree was visualized and manipulated using FigTree 1.4.2.[44]

Individual protein trees were made in an analogous fashion.

Identification of members of protein families

Orthologous proteins were obtained through BioBIKE's ORTHOLOG-OF function, which defines an ortholog by bidirectional best hit with a threshold of 10^{-10}. In other words, a protein A in organism X is defined as orthologous to protein B in organism Y if B is the best Blast hit of A against Y and A is the best Blast hit of B against X (E values < 10^{-10}). Amino acid sequences of HetR-like, HetN-like, NtcA-like sequences, and proteins encoded by genes typically flanking *patS* and *patX* genes were found within BioBIKE by ORTHOLOG-OF and SEQUENCE-SIMILAR (both protein vs protein and protein vs translated DNA). Identification was confirmed by examination of protein alignments. RGSGR-containing proteins were identified by examining the output of a BioBIKE expression that found all ORFs (whether annotated or not) containing RG[SGT]GR. Candidates were excluded if they were encoded by ORFs that were out of frame within conserved genes. Coordinates of genes encoding orthologous proteins of interest and the exemplar proteins used to find the orthologs are shown in **Suppl. Tables S4, S5, and S6**.

Analysis of protein characteristics

Transmembrane domains were predicted with TMHMM 2.0 (TransMembrane-Hidden Markov Model),[45,46] a program that compares the sequence characteristics of a given protein to those of a training set of 160 well studied membrane spanning regions of proteins from eukaryotes and prokaryotes. The program employs hidden Markov models, which use the propensities of regions of the training set to predict the likelihood that an amino acid follows a given amino acid string, much like using the high incidence "t" after "ich" to predict that one is probably looking at

German text rather than English. The program reports membrane-spanning regions, but it is easily fooled by signal sequences.

Putative signal sequences were identified using SignalP~~,~~[47,48]~~–~~.[49,50] The program uses neural networks built from one of three data sets to distinguish signal sequences from non-signal sequences. Neural networks build decision-making processes from known exemplars. SignalP employs data sets using known signal sequences from human proteins (representing eukaryotes), from Bacillus subtilis (representing Gram-positive eubacteria), and from Escherichia coli (representing Gram-negative eubacteria). The networks may be trained positively with proven signal peptides and negatively with transmembrane sequences or trained only positively, relying on the user to assert that the input sequences have no confounding transmembrane sequences. The program also offers a strict threshold (default) and a permissive threshold (sensitive), or a ~~knowledgable~~knowledgeable user can specify any threshold. SignalP therefore offers 12 choices: 3 data sets x 2 training regimes x 2 levels of set sensitivity.

Cyanobacteria are cytologically Gram-negative, but they are phylogenetically no closer to E. coli than to B. subtilis, so it is not obvious which neural network to choose. To address this question, we took the sequences from 162 proteins from Synechocystis PCC 6803 whose true N-termini had been determined,[51] 22 with apparent signal peptides and 140 whose N-terminus is the methionine at the predicted translational start site or the amino acid next to it. Running these proteins through the three neural networks of SignalP allowed us to assess the false positive and false negative rates and in this way judge the best of the 12 choices for determining cyanobacterial signal peptides. The Gram-positive data set (and, surprisingly, the eukaryotic data set) outperformed the Gram-negative data set. The best condition was to use the Gram-~~posistive~~positive data set with no training on transmembrane regions and high sensitivity (5% false negative and 1% false positive). We used this condition in predicting signal sequences but also show results from the same data set trained on transmembrane regions.

Comment [j2]: IK: Where are they shown?
JE: Fig 7B, but if I say so now, I'd have to put that figure in before Figures 3-6, which would be very confusing. I don't think I can refer to a figure in the text as something that will appear later.

Analysis of upstream DNA sequence motifs

Sequences upstream from candidate *patX* genes were collected with BioBIKE's SEQUENCES-UPSTREAM-OF function. To search for DIF^{+} motifs ~~regions~~ upstream from candidate *patS* genes, 1000 nucleotides upstream from the start site were scanned for sequences with no more than one mismatch in TCCGGA using BioBIKE's MATCHES-OF-PATTERN function.

To distinguish biologically functional matches from spurious matches, the characteristics of the DNA sequences surrounding the TCCGGA sites were compared to characteristics of a training set, 54-nucleotide sequences containing TCCGGA sites preceding ~~*Anabeana*~~*Anabaena* PCC 7120 genes known to be dependent on HetR (DIF+ genes).[15] The training set was used to construct a position-specific scoring matrix (PSSM), a lookup table giving the adjusted probability of a given nucleotide at a given position in a set of aligned sequences. The probabilities were adjusted by adding a constant number of counts (called pseudocounts) to all nucleotides, to minimize the effect of low counts for a nucleotide owing to a small sample size. BioBIKE's APPLY-PSSM-TO function was used to consider each 54-nucleotide segment within an upstream region, calculating a joint probability for that segment based on the product of unit probabilities from the PSSM and comparing that to a joint probability calculated from the product of individual nucleotide frequencies within the training set. The ratio is expressed as a logarithm, where a number close to 0 is expected if the sequence of a segment arose by chance. Otherwise, a positive number is expected that increases in magnitude as the fragment increases

in sequence similarity to the sequences of the training set. In calculating the probabilities, only those positions were considered where the information in the training set exceeds 0.2. Other PSSMs were constructed in an analogous fashion. Some PSSMs were constructed based on training sets of variable length, owing to a gap of 17 or 18 nucleotides separating conserved regions. In such cases, sequences were forced to the same length by deleting when appropriate a nucleotide at the center of the gap region (after determining that this position has negligible information content).

The information (a measure of order) at a certain position in an alignment is related to entropy (a measure of disorder) and is defined as $\mathbf{E_{max}} - \mathbf{E}$, where **E** (entropy) is $-\Sigma\ \mathbf{p_i} \log_2 \mathbf{p_i}$ summed over all four nucleotides, $\mathbf{p_i}$ is the frequency of a given nucleotide at the position, $\log_2 \mathbf{p_i}$ is taken to be 0 when $\mathbf{p_i}$ is 0, and $\mathbf{E_{max}}$ is the maximum possible value of **E**. The maximum value occurs when there is an even distribution of nucleotides, i.e. $-\Sigma\ (1/4) \log_2 (1/4) = 2$. The maximum information therefore is 2, occurring when $\mathbf{p_i}$ for one nucleotide is 1 and for the other three, 0. Information for alignments was calculated using the INFORMATION-OF function of BioBIKE.

The information of positions within an alignment was visualized using WebLogo.[52] Since the number of amino acids differs from the number of nucleotides, the maximum information value for amino acid sequences differs as well. That value is $-\Sigma\ (1/20) \log_2 (1/20) = 4.3$. For both nucleotide and amino acid logos, perfectly aligned positions may not have the maximum information value, because the program applies a correction for small sample size when the number of nucleotide sequences is less than 20 or amino acid sequences is less than 40.

Results

Phylogeny of cyanobacteria used in this study

In order to place the presence or absence of NtcA, HetR, PatS, and HetN in a logical context, we developed a phylogenetic tree of the 127 cyanobacterial genomes used in this study (see **Suppl. Table S1** for description of the genomes). The tree is based on alignments of 29 proteins, a subset of the 32 proteins used by Howard-Azzeh et al (2014).[53] Trees based on 16S rRNA sequences provide significantly less information on which to base a tree, and consequently there are far fewer nodes with good bootstrap information than with trees based on many conserved proteins.[54] Not surprisingly, considering its basis, the tree is completely concordant with that of Howard-Azzeh et al, with respect to the 100 genomes used by both groups and nodes enjoying strong bootstrap support. It also concordant with three other trees based on different sets of concatenated protein alignments.[5454,55,56] All of these trees differ in important respects from that of Uyeda et al (2016),[57] who attempted to avoid artifacts resulting from long branch attraction. Their tree differs in the placement of the picocyanobacteria, *Prochlorothrix hollandica* PCC 9006, branching heterocyst-forming cyanobacteria, and the *Calothrix* PCC 6303 and PCC 7103 pair. However, none of these differences in predicted phylogeny affect the arguments we will present. The intermixing of branched and non-branched (Sections IV and V) heterocyst-forming cyanobacteria is discussed later.

Fig. 3 shows one of many possible interpretations of the cyanobacterial phylogenetic tree. The tree is rooted by *Gloeobacter violaceus* PCC 7421, owing to the early divergence of Gloeobacter from the rest of the cyanobacterial lineage.[58,] If that rooting is accurate, then it is evident that unicellularity is the original morphotype of cyanobacteria. Filamentous strains appear to have arisen early and include in their number the coherent clade of heterocyst-forming cyanobacteria.

Whether filamentarity appeared once as shown in **Fig. 3** or multiple times is an open question, one that is discussed later.

Fig. 4 shows details of the phylogenetic tree, split between heterocyst-forming cyanobacteria (**Fig. 4A**) and the rest (**Fig. 4B**), and also lists the genome abbreviations used in this work.

Appearance of HetR in cyanobacteria

Orthologs of HetR were found in almost all filamentous strains and in almost no unicellular strains (**Fig. 4**). The filamentous exceptions are the Pseudanabaenas (Clade 8 in **Figs. 3** and **4B**) and *Geitlerinema* PCC 7105. The absence of HetR in *Geitlerinema* PCC 7105 may be the result of an incomplete sequence or mis-assembly of the genome. The gene order surrounding *hetR* (**Suppl. Fig. 1**) is conserved in the two closest available genomes, those of *Phormidium* OSCR and *Phormidium* BDU 130791. Several genes near *hetR* in the two Phormidia are completely missing from *Geitlerinema* PCC 7105, including five ribosomal proteins presumably required for life that are found in all other cyanobacterial genomes (except in one case where the gene cluster is split between two contigs). We conclude that a segment containing these genes as well as *hetR* is almost certainly present in *Geitlerinema* PCC 7105 but missing from its available genome assembly.

The only phenotypically unicellular strains possessing a HetR ortholog are *Synechococcus* PCC 7002 and *Synechococcus* PCC 7335. *Synechococcus* PCC 7002 is closely related to *Leptolyngbya* PCC 7376, a filamentous strain. It was formerly called *Agmenellum quadruplicatum* PR6[59] because of its propensity to grow as four-cell filaments, and a variant has been found that forms long filaments at 24^{o}C (Don Bryant, personal communication). *Synechococcus* PCC 7335 lies phylogenetically within a clade otherwise consisting of filamentous cyanobacteria (**Fig. 4B**) and is most closely related to *Leptolyngbya Heron Island J*. HetR may therefore be a holdover from a time in the recent evolutionary past when the ancestors of these two strains were filamentous.

The tight association of HetR with filamentous strains raises the possibility that the protein may be important in the filamentous life style (at least in the clade that excludes the Pseudanabaena). If so, then one might expect the phylogeny of the HetR protein to match the organismal phylogeny, if filamentarity arose only once, but not if multicellularity arose several times. In fact, though the HetR phylogenetic tree (**Suppl. Fig. 2**) lacks sufficient bootstrap support to be definitive, it matches the phylogenetic tree as well as can be expected. In particular, HetR from Clade 1A (Figs. 3 and 4A), containing all the heterocyst-forming cyanobacteria, appear to have a common ancestor. HetR from *Synechococcus* PCC 7002, *Leptolyngbya* PCC 7376, and *Spirulina subsalsa* PCC 9445, all in Clade 2 (Figs. 3 and 4B), form a coherent group distinct from other HetR proteins., all with bootstrap support and consistent with the phylogenetic tree. It is also worth noting that there are seven organisms represented in **Suppl. Fig. 2** with more than one apparent copy of HetR. In each case, there is one copy of HetR (termed "primary") that has a typical amino acid sequence (see below), while the other copies (termed "secondary") have less conserved sequences and cluster together (**Suppl. Fig 2**).

While the evolutionary connection of all the HetR sequences is beyond dispute, it is an open question as to whether the HetR proteins in the phenotypically unicellular strains and the secondary HetR proteins have the same function as primary HetR proteins in the filamentous strains or indeed any function at all. To address this question, all available HetR sequences were aligned (**Fig. 5** and **Suppl Fig. 3**). There is overwhelming amino acid sequence conservation in

the 75 primary HetR proteins from filamentous cyanobacteria (including heterocyst-forming). Of the 299 amino acid positions (allowing for frayed N- and C-termini), 172 (the green columns) are highly conserved as defined in **Fig. 5**. Of this latter group, 23 residues have been implicated in DNA- (19) or PatS-binding (4), from analyses of crystal structures[60,61,62] and in vitro assays[26~~25~~,60~~057~~,61~~6158~~,62~~6259~~,63] and in vivo phenotypes of site-specific mutants.[11,14,18~~1817~~,61~~6158~~,62~~6259~~,63~~6360~~,64,65] Only four primary HetR sequences in filamentous cyanobacteria have mutations in any residue implicated in DNA- or PatS-binding, and three of the mutations are conservative.

The two phenotypically unicellular strains present a different picture. To avoid observation bias (there are far more available sequences of Nostocs and Anabaenas than sequences phylogenetically close to the unicellular strains), we compared each of the two strains to its closest relative: *Synechococcus* PCC 7335 to *Leptolyngbya* Heron Island J and *Synechococcus* PCC 7002 to *Leptolyngbya* PCC 7376 (**Table 1**). *Synechococcus* PCC 7335 has more than four-times the number of mutations in conserved positions as does *Leptolyngbya* PCC Heron Island J, and 24% are non-conservative substitutions, compared to 0% for *Leptolyngbya* PCC Heron Island J. With the more distant *Synechococcus* PCC 7002 / *Leptolyngbya* PCC 7376 pair, the former has 2.6-times more mutations than the latter and 62% non-conservative substitutions, compared to 42% for *Leptolyngbya* PCC 7376. HetR's from the unicellular strains are evidently under lesser or different selection than those from the related filamentous strains. Similarly, secondary HetR's have a much higher number of deviations and non-conservative deviations than their primary counterparts (**Table 1** and data not shown).

If the high number of deviations in unicellular and secondary HetR's were due to drift in the absence of selection, then one would expect to find deviations spread randomly across the functional categories, but this is not observed (**Table 1**). In both cases, the amino acids implicated in DNA-binding are significantly less likely to deviate from the standard residue. Two secondary HetR's, $HetR_{Lep6406\text{-}b}$ and $HetR_{Lyn141951\text{-}b}$, have the number of deviations expected by chance, but the other secondary HetR's have far fewer (data not shown). In contrast, the HetR's from unicellular cyanobacteria are significantly more likely to experience deviations in residues associated with PatS-binding. In addition, HetR from *Synechococcus* PCC 7002 also carries an R223A mutation, in a residue implicated in the phenotypic sensitivity of HetR to PatS and HetN.[14] Evidently, mutation is not random in secondary HetR's and those in unicellular cyanobacteria, indicating maintained selective pressure during at least part of the period since separating from their primary filamentous homologues, presumably owing to retained DNA-binding function.

Appearance of HetN and PatS in cyanobacteria

If HetN is defined as a protein (a) similar in sequence to HetN of *~~Anabeana~~Anabaena* PCC 7120 and (b) possessing RGSGR, then its incidence is limited to *~~Anabeana~~Anabaena* PCC 7120 and its three closest relatives (**Fig. 4A**), plus two distantly related unicellular cyanobacteria (**Fig. 4B**). In addition, two strains of Chlorogloeopsis carry a HetN-like protein with the sequence ERGSGH, one off from the conventional motif (**Fig.4A** and **Suppl. Fig. 4**). There is good reason to doubt the significance of these proteins in heterocyst regulation, as a mutation of the *~~Anabeana~~Anabaena* PCC 7120 HetN motif from RGSGR to RGSGK results in loss of function in *~~Anabeana~~Anabaena* PCC 7120.[31~~30~~] A phylogenetic analysis of HetN-like proteins (**Suppl. Fig. 4**) shows a well-supported cluster of *~~Anabeana~~Anabaena* PCC 7120 HetN and its three relatives, lying distinct from a second well supported cluster that includes the two

Formatted: Font: Italic

Comment [j3]: IK: I'd rather use allele designations $HetR_{Lep6406\text{-}b}$ and $HetR_{Lyn141951\text{-}b}$
JE: Done for these two. I'll try to change all other similar instances.

Formatted: Subscript

Formatted: Subscript

Chlorogloeopsis HetN candidates. The two unicellular HetN candidates lie in a distant cluster (not shown).

It is much more difficult to identify putative PatS proteins. Only three genomes amongst the 127 genomes considered in this work have had *patS* genes annotated within them: ~~*Anabeana*~~*Anabaena* PCC 7120,[222221] *Nostoc punctiforme* ATCC 29133,[66] and *Nodularia spumigena* CCY9414.[67] Two other genomes, *Leptolyngbya* NIES 3755 and *Arthrospira* PCC 8005, have genes misannotated as *patS*. The lack of annotated *patS* genes is to be expected, since most automated gene-calling processes exclude ORFs of the size of *patS*. Scanning genomes for ORFs containing RGSGR is also unsatisfactory, as the rate of false positives is far too high. The genomes considered in this study average 11.1 RGSGR-containing ORFs, of which 82% are in called genes (80% of these are in the wrong reading frame). The high-GC genome of *Cyanobium gracile* PCC 6307 provides an extreme example: it has 68 RGSGR-containing ORFs. It is highly unlikely that any of them have regulatory function in this unicellular organism. To identify true orthologs of PatS, we therefore also considered genetic context.

The *patS* gene from ~~*Anabeana*~~*Anabaena* PCC 7120 (*asl2301*) is preceded by two genes (*all2302* and *all2303*) encoding proteins annotated as patatin and dihydroorotase, respectively. On the downstream side is a gene (*alr2300*) annotated as *hetY*, encoding a protein described as necessary for timely heterocyst differentiation.[68] We found 26 genomes, all from heterocyst-forming cyanobacteria, with a short RGSGR-containing ORF situated near at least one of the typical upstream or downstream genes (**Fig. 6**). All but one of the encoded RGSGR motifs are preceded by a glutamate (E) residue. From the sequence characteristics of these 26 putative PatS proteins, four other candidate PatS proteins were identified, all in genomes of heterocyst-forming cyanobacteria. One, from *Hapalosiphon* MRB220, is so similar in sequence to the putative PatS, from *Fischerella* PCC 9339, that it was added to the list. The remaining three were tagged as possible PatS instances but suspicious (**Fig. 6**).

PatS from ~~*Anabeana*~~*Anabaena* PCC 7120 terminates in RGSGR, but this is true for only 18 of the 30 identified PatS proteins. In the case of three related cyanobacteria, the PatS sequence extends 69 amino acids beyond RGSGR. In the remaining nine cases, the RGSGR appears shortly before the end of the protein. In all but one case, PatS from *Rivularia* PCC 7116, the RGSGR motif lies within nine amino acids from the putative N-terminus.

All intergenic regions of all cyanobacterial genomes were searched for ORFs with ERGSGR motifs. Only 11 new instances were found, 9 of which were in genomes of heterocyst-forming cyanobacteria. They fall into four families (**Suppl. Fig. 5**). One family, consisting of three Fischerella and a related strain, all have similar ORFs with ERGSGR preceded by two amino acids. A family of two related Scytonemas and Tolypothrix strains plus a distantly related Cylindrospermum strain have more conventional looking PatS candidates, as does an ORF from Rivularia. The only two alternative candidates from non-heterocyst-forming strains have sequences quite dissimilar to the rest. ORFs from both *Leptolyngbya* PCC 6406 and *Oscillatoria* PCC 10802 have ERGSGR motifs embedded within ORFs with large N-terminal and C-terminal extensions. These ORFs were not considered any further.

All 39 of the cyanobacteria in Clade 1A (all the heterocyst-forming cyanobacteria) have a PatS-candidate protein, with the following exceptions. The eight cyanobacteria most closely related to *Anabaena cylindrica* PCC 7122 lack a candidate, as does *Scytonema tolypothrichoides* VB61278, unless one counts the candidate shown in **Suppl. Fig. 5**. The latter genome is unusual

in that it has 788,708 nt that are given as ambiguous (8% of the total), most of them in contiguous chunks. This is more than any of the other 126 genomes we considered. It is possible that a PatS-like peptide is encoded in a region missing from the assembly. The *patS* genes of the two closest relatives of *Scytonema tolypothrichoides* VB61278, *Mastigocladopsus repens* PCC 10914 and *Tolypothrix campylonemoides* VB511288, are both flanked by genes encoding dihydroorotase and HetY. However, in the available genome assembly of *Scytonema tolypothrichoides* VB61278, these two genes are distant from one another.

Appearance of PatX in cyanobacteria

The absence of both HetN and PatS in the *Anabaena cylindrica* PCC 7122 clade despite normal heterocyst spacing could be explained if these genomes possess a third RGSGR-containing protein. We could identify only one protein containing RGSGR that can be found in the genomes of multiple cyanobacteria, including members of the *Anabaena cylindrica* PCC 7122 clade. That protein, termed PatX, is poorly conserved in overall sequence, and so we turned again to genetic context to guide discovery of other members of the family.

Genes encoding RG[ST]GR-bearing proteins were found in the genomes of 37 heterocyst-forming cyanobacteria (**Fig. 7A**), near at least one of six linked genes that include three known to be related to heterocyst differentiation or function: *hetR*, *sepJ* (also known as *fraG*, encoding a protein required for filament integrity under N-fixing conditions[69]), and *glnA* (encoding glutamine synthetase,[70] which catalyzes the first step in the assimilation of fixed nitrogen[71]). In ~~*Anabeana*~~*Anabaena* PCC 7120, these genes are alr2339, all2338, and alr2328, respectively. Immediately upstream of the gene is a gene encoding a protein that is highly conserved in Groups 1-6 and that may be an FAD-dependent oxidoreductase (All2333 in ~~*Anabeana*~~*Anabaena* PCC 7120). Amongst the 37 proteins are the two postulated by Stucken et al (2010) to substitute for PatS[36~~36~~35] and another misidentified as PatS.[38~~38~~37] Antonaru and Nürnberg recently recognized proteins they called alternative PatS that share the characteristics of PatX.[72] All heterocyst-forming strains have *patX* genes in the proper genetic context (**Fig. 7A**), except for *Richelia intracellularis* HH01, which has a plausible *patX* gene but disconnected from the usual flanking genes, and the two Chlorogloeopsis strains which lack any sign of *patX*. The sequence of PatX in ~~*Anabeana*~~*Anabaena* PCC 7120 and its two closest relatives carry an RGTGR motif. All the others carry RGSGR. It is not clear whether an RGTGR motif would be effective in regulating HetR. Mutating the central S to A reduced the activity of HetN[**Error! Bookmark not defined.**~~31~~30] and abolished the activity of PatS,[27~~27~~26] but perhaps a mutation to T would have less effect.

From this collection of proteins, certain structural generalities stand out (**Fig. 7A** and **Fig. 8B**) and may be contrasted with those of PatS (**Fig. 6** and **Fig. 8A**). First, of course, the proteins possess the RGSGR motif, except those carrying RGTGR instead. The motif lies close to the C-terminus of the protein and is usually preceded by a H or Y (heterocyst-forming strains) or H, E, or D (non-heterocyst-forming strains) and followed by R. The motif is often preceded by a proline-rich region. The N-termini are rich in hydrophobic residues, and all filamentous organisms except *Microcoleus* PCC 7113 have at least one candidate PatX protein with a signal peptide identified by SignalP. The N-terminus exhibits a striking pattern, *PxxxPPxxxPPxxx*, where P is a polar residue, S, T, or G, and x is any hydrophobic residue. This motif is found in proteins with one transmembrane domain that form homodimeric complexes.[73] Only *Richelia intracellularis* HH01 amongst the heterocyst formers lacks a candidate PatX protein with a motif of this sort. However, there is no good evidence that the region forms a transmembrane domain.

TMHMM predicts such regions in 18% of the PatX sequences, but the program is often confused by signal sequences.[45~~45~~~~44~~]

PatX is equally well represented in the genomes of filamentous strains that don't make heterocysts, identifying genes by the presence of RG[SGT]GR and their proximity to *hetR* and/or an all2333 ortholog **(Fig. 7B).** By this definition, all filamentous strains have PatX, with the exception of *Oscillatoria* PCC 6304 and *Spirulina subsalsa* PCC 9445. These two strains have ORFs that have many characteristics of PatX but are not near either of the signature genes. The PatX candidates from non-heterocyst-forming strains also have an N-termini identified by SignalP as signal sequences, but the *PxxxPPxxxPPxxx* motif is absent, and prolines are less prominent in the region preceding the RGSGR motif. The amino acids immediately preceding and following the RGSGR motif in general follow the tendencies of those in PatX sequences from heterocyst-forming strains (**Fig. 8B** and **8C**).

The PatX sequences of the four Planktothrix strains carry RGGGR motifs. However, the same strains have PatX-like genes that in all respects except genetic context better fit the characteristics of PatX genes found in heterocyst-forming strains: a similar pattern of polar residues, MR…*PPxxxPPxxxPxx*, at the N terminus, and an H residue preceding an RGSGR motif. Five of the thirty PatX sequences from non-heterocyst forming strains have one to three extra RGSGR motifs at spaced intervals, far more than one would expect by chance.

Only one unicellular strain, *Synechococcus* PCC 7335, has an identifiable PatX ORF. RGSGR-containing ORFs were not found outside of conserved genes in *Synechococcus* PCC 7002 nor in four close relatives with completely sequenced genomes **(Suppl. Table S1)**.

Comment [Э4]: If mentioned, they should figure somewhere in supplementary materials ***[They are described in Suppl Table 1 (but I needto complete this and add the additional organisms now represented in the HetR tree)]***

Sequences upstream from PatS and PatX genes

We report elsewhere (Khudyakov and Gladkov, submitted) that the expression of *patX*$_{Ana7120}$, like that of *patS*$_{Ana7120}$, is induced by nitrogen deprivation and its expression is confined to differentiating cells. This finding prompted an examination of the sequences upstream from these genes, to consider whether shared elements might serve as the basis for their common regulation. Mitschke et al (2011) reported two transcriptional start sites upstream from *patS*$_{Ana7120}$, one at ~~-~~-580 that is induced by nitrogen deprivation and another at -692 that is not[15] (these differ from the two 5' ends reported earlier[23~~23~~~~22~~]). The inducible start site is preceded by a consensus DIF^{+} motif (TCCGGA),[15] beginning at -35 (**Suppl. Fig. 6**).

Comment [j5]: IK: (TCCGGA)[15]
JE: Done. But I'm going to remove it (in a new version) when I format for Mol Microbiol, because three instances of "Mitchke et al, 2012" in close proximity is too great a price to pay.

In order to learn what characteristics to look for in a functional ~~DIF motif~~DIF^{+} motif, we turned to a collection of such motifs (with no more than one mismatch) that has been shown to precede 58 inducible genes in ~~*Anabeana*~~*Anabaena* PCC 7120, beginning 33 to 38 nucleotides before the transcriptional start site or (three outliers) beginning at -43 or -44.[15] Considering just the 55 regions of the ~~second~~ first group, the ~~DIF motif~~DIF^{+} motif is followed 17 or 18 nucleotides downstream by a less conserved motif (G[TA]ANA) around 10 nucleotides before the transcriptional start site (**Suppl. Fig. 6** and **Fig. 9D**). One might surmise from previous studies[15,74] that ~~DIF motif~~DIF^{+} motifs are followed by classical -10 regions recognized by SigA (consensus TATAAT), however the similarity of the -10 region to TATAAT is low, with a median score (L_2(-10) in **Suppl. Fig. 6**) of 0.55 compared to a median score of 4.30 over all Anabaena transcriptional sites.[15] Since the scale is based on $\log_2$, the two median scores differ by a factor of 13. There is no obvious correlation between on one hand either the -10 scores, the quality of the ~~DIF sequence~~DIF^{+} motif, or the similarity of the -10 region to G[TA]ANA and on

Comment [j6]: IK: DIF^{+} motif
JE: Global replace, done. (distinguishing between the motif and the region containing the motif.

the other hand the number of transcripts at 8h after N-deprivation (8hRds) or degree of induction (L_2(8h/0h)).

The 55 promoter regions with DIF$^+$ motifs ~~regions~~ were used as a training set to construct a position specific scoring matrix (see **Methods**) to identify possible DIF$^\pm$-motif-containing regions (DIF$^+$ regions) in sequences upstream from *patS* and *patX*, those better predicted by the sequence characteristics of the training set than by the overall nucleotide frequencies of the ~~DIF regions~~training set. This strategy led to the identification of candidate regions (with a $\log_{10}$ odds score better than 4) in 15 of the 30 putative *patS* genes (**Suppl. Fig. 7**). Using these 15 regions as the training set led to the discovery of an additional putative DIF$^\pm$ region. Regions identified in this manner clustered consistent with the phylogenetic relationships shown in **Fig. 4**. In all cases, the ~~DIF motif~~DIF$^+$ motif lies in an intergenic region contiguous with the beginning of the putative *patS* gene. 88% of the regions carry exact matches to the TCCGGA motif. In contrast, only 26% of Mitschke et al's set have exact matches. The consensus -10 region of the putative DIF$^\pm$ region upstream of *patS* genes is similar to that of the training set of DIF$^\pm$ regions (compare **Fig. 9A** with **9D**): GTAGAGA vs G[TA]ANA.

It must be stressed that the mere presence of a DIF$^\pm$ ~~region~~ motif defined as no more than one nucleotide off from TCCGGA is not significant without additional sequence or positional information. Such sequences are found by actual count on average once every 106 to 218 nt over the range of heterocyst-forming cyanobacteria.

The same training set of 55 DIF$^\pm$ regions was used to identify putative DIF$^\pm$ regions upstream from *patX* genes. Candidate regions with good scores were found upstream from *patX* genes in all 37 of the heterocyst-forming cyanobacteria that have *patX* (**Suppl. Fig. 8A**). All ~~DIF sequence~~DIF$^+$ motifs were positioned 57nt upstream from the translational start site (except for one case where it is 58nt), and all were perfect TCCGGA sequences. The regions had a large number of conserved sequences (**Fig. 9B**), including a conserved -10 region (always 18 nt from DIF), GTAnnAG, preceded by a conserved A.

Similarly, each of the 28 non-heterocyst-forming cyanobacteria showed plausible ~~DIF sequence~~DIF$^+$ regions close to the translational start site of *patX* (**Suppl. Fig. 8B**). Although there are far fewer well conserved positions in the upstream sequences (to be expected, given the greater phylogenetic range of this cyanobacterial grouping – see **Fig. 4**), there is a conserved cluster of nucleotides near the -10 position, very similar to those of heterocyst-forming cyanobacteria (**Fig. 9C**). In the seven cases where cyanobacteria have two copies of *patX*, six bear DIF$^\pm$ regions with the same characteristics as those of other non-heterocyst-forming cyanobacteria. The seventh case, ProH9006-a, may have a degenerate form. The *patX* genes of two strains of Lyngbya have two DIF$^\pm$ ~~regions~~ motifs one after the other preceding the translational start site. Considering all 37 ~~DIF sequence~~DIF$^+$ motifs, only 11% have perfect TCCGGA sequences, while 65% have TCCTGA, spread over the full range of non-heterocyst-forming strains. The patX gene of *Synechococcus* PCC 7335, the only unicellular strain~~s~~ possessing one, is preceded by a respectable DIF$^\pm$ region.

A striking feature of sequences upstream from *patX* genes is the presence of NtcA-binding sites (GTAN$_8$TAC)[75] in 33 of 37 heterocyst-forming cyanobacteria, always 13-16nt from the ~~DIF sequence~~DIF$^+$ motif (**Suppl. Fig. 8A**). The site preceding *patX* from ~~*Anabeana*~~*Anabaena* PCC 7120 has been shown experimentally to bind NtcA.[75~~75~~~~69~~] No such sites are found upstream from the ~~DIF site~~DIF$^+$ motifs of non-heterocyst-forming cyanobacteria nor from those upstream

from *patS* genes in heterocyst-forming strains (**Suppl. Figs. 7** and **8B**). Only two genes of the 57 in Mitschke et al's set of DIF^{+}-bearing genes have ~~DIF site~~DIF^{+} motifs preceded by NtcA-binding sites (15nt before the ~~DIF site~~DIF^{+} motif of *all0935* and 22nt before the ~~DIF site~~DIF^{+} motif of *asr1775*).

Discussion

The Turing/Meinhardt model, expanded by later models,[9] calls for an R morphogen that causes developmental action and for a diffusible S morphogen that inhibits it. HetR and PatS, modulated by the actions of NtcA and HetN, may be a realization of this model, leading to the appearance of spaced heterocysts in ~~*Anabeana*~~*Anabaena* PCC 7120. In many other cyanobacteria, however, these actors are insufficient to explain patterned differentiation. PatS is not present in a clade containing many heterocyst-forming cyanobacteria, including ~~the model organism~~ *Anabaena cylindrica* PCC 7122. HetN is absent from most heterocyst formers, confined to only close relatives of ~~*Anabeana*~~*Anabaena* PCC 7120. Their regulatory burden may be taken up by a third protein, PatX, one that like PatS and HetN carries the RGSGR motif but is almost universally distributed amongst cyanobacteria possessing HetR. PatX is therefore likely to be the original counterweight to HetR in a Turing/Meinhardt-like regulatory system, partially supplanted in relatively recent times by PatS in most heterocyst forming strains and by HetN in a small subset. A biological role for PatX has been shown experimentally in ~~*Anabeana*~~*Anabaena* PCC 7120 (Khudyakov and Gladkov, submitted).

A significant problem in assessing the prevalence of PatS and PatX is that they cannot be recognized reliably by the usual automated methods applied to genomes. PatS poses a formidable challenge for automated methods because of its small size (median size 13 amino acids, **Fig. 6**). As a result, while *patS* genes used to be annotated in the two genomes found at NCBI with physiological evidence for the function of PatS (~~*Anabeana*~~*Anabaena* PCC 7120 and *Nostoc punctiforme* ATCC 29133) plus one more (*Nodularia spumigena* CCY9414) without such evidence, reannotation efforts by NCBI[76] that paid no heed to published results have discarded these annotations. The genes can be recognized, however, by the presence of an encoded RGSGR motif in a standard genetic context (**Fig. 6**) as discussed above.

Genes encoding PatX generally appear in recently annotated genomes, but the low level of sequence conservation does not permit associating them with orthologous genes. They are readily determined by genetic context and sequence characteristics: a generally C-terminal RGSGR motif, an N-terminal region, either membrane-spanning or a signal sequence, and both separated by a spacer region generally rich in prolines (**Fig. 7**). We imagine that the protein might be transported out of the heterocyst and acted on by a peptidase or tethered to the membrane and digested near the site of transport out of the cell. The proline-rich region might maintain PatX in a disordered structure to ensure that the RGSGR region is available to a peptidase.

Both *patS* and *patX* genes are preceded by conserved upstream regions with the following characteristics (**Fig. 9** and **Suppl. Figs. 7-8**). Both have ~~DIF motif~~DIF^{+} motifs (exact TCCGGA in the case of heterocyst-forming organisms, TCC[GT]GA in the case of non-heterocyst-forming organisms) at a position that is probably -35 to the transcriptional start site. At the -10 position

Comment [j7]: IK: Used to be a model organism
JE: Needlessly cruel. Ruanbao Zhou certainly considers it still a model organism, and Peter Wolk might as well. Too much controversy. I'll remove the phrase.

Comment [j8]: IK: Really? Can you give a reference for physiological evidence for the function of PatS in *Nostoc punctiforme*?
JE: ***Risser et al (2012). Effectiveness of en exogenous RGSGR and Ana7120 PatS in Nostoc is evidence.***

there is another conserved motif: GTAGAGA (*patS*) or GTAnnAG (*patX*). In the case of *patX* in heterocyst-forming cyanobacteria, the ~~DIF motif~~DIF^{+} motif is preceded by an NtcA-binding site. Transcription from the *patX* promoter in ~~*Anabeana*~~*Anabaena* PCC 7120 ~~follows the same contours~~responds to nitrogen deprivation in the same way as transcription from the *patS* promoter, but the latter is an order of magnitude lower.[15] However, the level of transcription of the two genes is comparable.[19~~19~~18~~]

Having defined PatS and PatX in this way, we can educe the following generalities. Most importantly, PatX and HetR are present together in almost all filamentous cyanobacteria except for the distantly related Pseudoanabaena (where both are lacking) and absent in almost all unicellular cyanobacteria (**Fig. 4**). HetR from the two exceptional unicellular cyanobacteria, *Synechococcus* PCC 7002 and *Synechococcus* PCC 7335, both are atypical, with many differences relative to closely related filamentous strains and defects in conserved residues associated with the binding of RGSGR (**Fig. 5**).

These atypical HetR proteins may be nonfunctional (in a state of decay), may retain HetR-like function, or may have transitioned to a function different from canonical HetR. The latter is most likely, as an analysis of mutations indicates continued selection for binding to DNA but not to PatS (**Fig. 5** and **Table 1**). For similar reasons, secondary HetR, though clearly deviant from canonical HetR protein, probably retains function as transcriptional regulatory proteins, perhaps supplementing HetR or perhaps serving some other purpose. It should be noted that all of the organisms bearing secondary HetR proteins, except *Prochlorothrix hollandica* PCC 9006, are closely related to one another. That and the clustering of secondary HetR proteins (**Suppl. Fig. 2**) is consistent with one or perhaps two acquisitions of the alternative form.

The exceptional filamentous strains (**Fig. 7**) include two strains of Chlorogloepsis that lack PatX and three strains of Anabaena that have versions of PatX carrying RGTGR instead of RGSGR. These five strains are amongst the six cyanobacteria that carry HetN (if the HetN-like proteins of the Chlorogloeopsis strains are functional). Perhaps HetN partially substitutes for the function of PatX in these strains. It is also possible that Chlorogloeopsis strains need no HetN nor PatX at full strength to control HetR, since they are rarely in a filamentous state.[59~~59~~~~56~~,77,78] There is one more filamentous strain, *Crinalium epipsammum* PCC 9333, whose PatX protein has RGTGR in place of RGSGR and four closely related strains of Planktothrix with a motif of RGGGR. In the latter cases, the strains also possess a second gene similar in sequence characteristics and upstream sequence to conventional PatX (**Suppl. Fig. 8B**) but in non-standard genetic contexts. In short, except in the case of *Crinalium epipsammum* PCC 9333, all filamentous cyanobacteria that lack PatX or have nonstandard PatX possess a protein that may conceivably compensate for the defect.

From these considerations, it is possible to propose a plausible sequence of evolutionary events that led to the cyanobacteria present today (**Fig. 4**). Filamentarity arose from a primordial unicellular state, and not long after the divergence of the Pseudoanabaena, HetR and PatX entered the lineage, conferring some advantage to filamentous cyanobacteria (the juxtaposition of the two genes in most filamentous strains may reflect a primordial genetic linkage). Perhaps they direct ordered rupturing of filaments (Khudyakov and Gladkov, submitted) to ensure that

Comment [j9]: IK: I cannot apprehend how the **contour** of transcription from the patS promoter can be an order of magnitude lower, while the **level** of transcription of the two genes is comparable
JE: Contour: Same general form, though not necessarily same magnitude. I'll try a different way of saying it.
Promoter activity not reflected in gene RNA levels: (1) Attenuation (2) mRNA decay that affects the 3' end but not the 5' end, or (3) multiple promoters..

Comment [Э10]: ???

Comment [j11]: Look at Fig. S6 of Koch et al (as well as the text)

propagules that break off from the mass maintain the advantages of filamentarity. The innovation of expensive but oxygen-resistant nitrogen fixation in heterocysts was enabled by tying the expression of HetR/PatX to nitrogen availability (through NtcA) on one hand and on the other hand to the regulation of heterocyst-related genes. The appearance of PatS made possible a greater degree of control, but the protein was lost in the common ancestor of the clade that includes *Anabaena cylindrica* PCC 7122. In the common ancestor of *Nostoc* PCC 7524, ~~*Anabeana*~~*Anabaena* PCC 7120, and its two closest relatives, an allele of a short-chain dehydrogenase/reductase appeared that had gained RGSGR within its sequence, resulting in a protein designated HetN. Subsequently, the RGSGR motif in PatX carried by the common ancestor of three of these strains mutated to RGTGR, leading possibly to diminished function of the protein. If one considers only the required elements of HetN[33~~32~~] – its membrane associated N-terminus and its RGSGR motif – it looks very much like PatX and might well substitute for it (but see an alternate view of Higa et al (2012)).[31]

In passing, we might note that *Mastigocladops~~i~~is repens* PCC 10914, placed within Section V (true-branching heterocyst-forming) in classical cyanobacterial taxonomy,[595~~956~~] clusters with Section IV cyanobacteria (linear heterocyst-forming) in the tree shown in Fig. 4, i.e. with Tolypothrix and Scytonema strains rather than with Fischerella and Hapalosiphon strains. This tree is concordant with earlier trees based on fewer genomes, used (wrongly, we believe) to support the monophyletic origin of Section V.[545~~451~~,565~~653~~] The polyphyletic nature is seen also in the 16S rDNA-based tree from an extended set of strains.[79]

Just as true-branching is polyphyletic, so is filamentarity, having been lost several times, every time with the concomitant loss of HetR and PatX or, in the cases of *Synechococcus* PCC 7002 and *Synechococcus* PCC 7335, the degradation (or repurposing) of HetR. (Needless to say, this does not imply that HetR and PatX are required for filamentarity -- clearly not the case in *Anabaena* PCC 7120 (Ref 11 and Khudyakov and Gladkov, submitted)). **Fig. 4** interprets events as having a single acquisition of filamentarity, HetR, and PatX and multiple losses. It is more parsimonious to envision multiple acquisitions of filamentarity (and HetR and PatX), as others have suggested,[565~~653~~] but if acquisition of filamentarity is more difficult than its loss, then simple parsimony may be a poor guide. The test is whether proteins such as HetR that are associated with filamentarity appear to have arisen in *Synechococcus* PCC 7002 and *Synechococcus* PCC 7335 by descent from a filamentous ancestor or by horizontal gene transfer. In the case of *Synechococcus* PCC 7335, the answer is clear from the phylogenetic tree of HetR (**Suppl. Fig. 2**) in favor of descent, but the evidence is only suggestive, not definitive, in the case of *Synechococcus* PCC 7002. An examination of other proteins associated with filamentarity might tip the scales. We found one protein, Alr4863, with orthologs in all filamentous strains except RicHH1 and orthologs in no unicellular strains, except *Synechococcus* PCC 7002, *Synechococcus* PCC 7335, and *Chamaesiphon minutus* PCC 6605 (~~Cha6605;~~ another unicellular strain with a close filamentous relative). The phylogenetic tree of this protein is nearly superimposable upon the organismal tree (**Fig. 4**), after erasing the unicellular strains except *Synechococcus* PCC 7002, *Synechococcus* PCC 7335, and *Chamaesiphon minutus* PCC 6605 ~~Cha6605~~ (result not shown), compatible with the lineal descent of filamentarity.

Comment [Э12]: It's a questionable matter, according to Higa et al. (2012) its hydrophobic N-terminus is not required for pattern maintenance, and the course of its expression differs, and so it looks very much **unlike** PatX, although can partially substitute for it.
02.01.18. Jeff, you did not react on this remark. I prefer to remove this sentence.
Sorry. I left this for last and forgot to go back to it. This matter is discussed in Corrales-Guerrero et al (2014). They (we) noted the difference between our results and those of Higa et al. They're not as incompatible as you might think. I've added a caveat at the end.

Comment [j13]: IK: Such reasoning can be misinterpreted as HetR or HetR-PatX being essential for filamentarity. We need to state clearly here or somewhere previously that the loss/inactivation of hetR does not lead to the loss of filamentarity or visible phenotypic defects in nitrogen-replete medium.
JE: OK. I added a sentence.

Formatted: Not Highlight

Field Code Changed

Formatted: Font: Italic

Summary

Our view of the regulation of heterocyst differentiation has been colored by the peculiarities of a single, atypical strain, Anabaena PCC 7120. While ~~*Anabeana*~~*Anabaena* PCC 7120 possesses regulatory elements (HetR, PatS, and HetN) that fit well with a Turing/Meinhardt model of patterned differentiation, other heterocyst-forming cyanobacteria possess different elements (HetR, PatX, and maybe PatS), though the patterns of heterocysts in many of these are indistinguishible from those of ~~*Anabeana*~~*Anabaena* PCC 7120. Two of the elements, HetR and PatX, appear to be primordial, as they exist together in practically almost all filamentous strains.

Comment [Э14]: practically? *[done]*

Acknowledgements: We thank Laura Antonaru, Jim Golden, Dennis Nürnberg, Doug Risser, and Karina Stucken for useful discussions. We also thank Peter Wolk for recognizing our common research directions and putting us together.

Author Contributions: JE and IK both conceived of the project and acquired, analyzed, and interpreted sequence data, first independently and then together. JE wrote the article with major contributions from IK.

… a few additional comments:

References #28 and #30 site the same article, but under different years

There are now many other duplications. I'll fix them once we freeze the text. Perhaps soon?

[Fixed]

REFERENCES

Warning! In this rough draft I've made no attempt to put references in a common format or to rid the list of duplicates. That will come once the draft is finalized

80 81 82

[1] Turing A M (1952). The chemical basis of morphogenesis. Phil Trans Royal Soc B 237:37-72.
[2] Gierer A, Meinhardt H (1972). A theory of biological pattern formation. Kybernetik 12:30-39.
[3] Meinhardt H (2008). Models of biological pattern formation: From elementary steps to the organization of embryonic axes. Curr Top Dev Biol 81:1-63.
[4] Marcon L, Sharpe J (2012). Turing patterns in development: What about the horse part? Curr Opin Genet Dev 22:578-584.
[5] Kumar K, Mella-Herrera RA, Golden JW (2010). Cyanobacterial heterocysts. Cold Spring Harb Perspect Biol 2:a000315. http://cshperspectives.cshlp.org/content/2/4/a000315.abstract.html
[6] Maldener I, Sumers ML, Sukenik A (2014). Cellular differentiation in flamentous cyanobacteria. In: *The Cell Biology of Cyanobacteria*. Eds: Flores E, Herrero A. Caister Academic Press. pp.263-291.
[7] Gerdtzen ZP, Salgado JC, Osses A, Asenjo JA, Rapaport I, Andrews BA (2009). Modeling heterocyst pattern formation in cyanobacteria. BMC Bioinform 10(Suppl 6):516.
[8] Muñoz-Garcia J, Ares S (2016). Formation and maintenance of nitrogen-fixing cell patterns in filamentous cyanobacteria. Proc Natl Acad Sci USA 113:6218-6223.
[9] Herrero A, Stavans J, Flores E (2016). The multicellular nature of filamentous heterocyst-forming cyanobacteria. FEMS Microbiol Rev 40:831-854.
[10] Buikema WJ, Haselkorn R (1991). Isolation and complementation of nitrogen fixation mutants of the cyanobacterium Anabaena sp. strain PCC 7120. J Bacteriol 173:1879-1885.
[11] Buikema WJ, Haselkorn R (1991). Characterization of a gene controlling heterocyst differentiation in the cyanobacterium Anabaena 7120. Genes Devel 5:321-330.
[12] Black TA, Cai Y, Wolk CP (1993). Spatial expression and autoregulation of *hetR*, a gene involved in the control of heterocyst development in *Anabaena*. Mol Microbiol 9:77-84.
[13] Buikema WJ, Haselkorn R (2001). Expression of the *Anabaena hetR* gene from a copper-regulated promoter leads to heterocyst differentiation under repressing conditions. Proc Natl Acad Sci USA 98:2729-2734. (www.pnas.orgycgiydoiy10.1073ypnas.051624898)
[14] Khudyakov IY, Golden JW (2004). Different functions of HetR, a master regulator of heterocyst differentiation in *Anabaena* sp.PCC 7120, can be separated by mutation. Proc Natl Acad Sci USA 101**:**16040–16045.
[15] Mitschke J, Vioque A, Haas F, Hess WR, Muro-Pastor AM (2011). Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena* sp. PCC7120. Proc Natl Acad Sci USA 108:20130-20135. www.pnas.org/cgi/doi/10.1073/pnas.1112724108
[16] Videau P, Ni S, Rivers OS, Ushijima B, Feldmann EA, Cozy LM, Kennedy MA, Callahan SM (2014). Expanding the direct HetR regulon in Anabaena sp. strain PCC 7120. J Bacteriol 196:1113-1121.
[17] Cai Y, Wolk CP (1997). *Anabaena* sp. strain PCC 7120 responds to nitrogen deprivation with a cascade-like sequence of transcriptional activations. J Bacteriol 179:267-271.

[18] Huang X, Dong Y, Zhao J: HetR homodimer is a DNA-binding protein required for heterocyst differentiation, and the DNA-binding activity is inhibited by PatS. Proc Natl Acad Sci U S A 2004, 101:4848–4853.
[19] Flaherty BL, Johnson DBF, Golden JW (2014). Deep sequencing of HetR-bound DNA reveals novel HetR targets in *Anabaena* sp. strain PCC7120. BMC Microbiol 13:255. http://www.biomedcentral.com/1471-2180/14/255
[20] Valladares A, Flores E, Herrero A (2016). The heterocyst differentiation transcriptional regulator HetR of the filamentous cyanobacterium Anabaena forms tetramers and can be regulated by phosphporylation. Mol Microbiol 99:808-819. doi:10.1111/mmi.13268
[21] Muro-Pastor AM, 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 44:1377-1385.
[22] Yoon H-S, Golden JW (1998). Heterocyst pettern formation controlled by a diffusible peptide. Science 282:935-938.
[23] Yoon Y-S, Golden JW (2001). PatS and products of nitrogen fixation control heterocyst pattern. J Bacteriol 183:2605-2613.
[24] Rajagopalan R, Callahan SM (2010). Temporal and spatial regulation of the four transcription start sites of hetR from Anabaena sp. strain PCC 7120. J Bacteriol 192:1088-1096. Doi:10.1128/JB.01297-09
[25] Risser DD, Callahan SM (2009). Genetic and cytological evidence that heterocyst patterning is regulated by inhibitor gradients that promote activator decay. Proc Natl Acad Sci USA 106:19884-19888.
[26] Feldmann EA, Ni S, Sahu ID, Mishler CH, Levengood JD, Kushnir Y, McCarrick RM, Lorigan GA, Tolbert BS, Callahan SM, Kennedy MA (2012). Differential binding between PatS C-terminal peptide fragments and HetR from *Anabaena* sp. PCC 7120. Biochemistry 51:2436–2442.
[27] Corrales-Guerrero, L, Mariscal V, Flores E, Herrero A (2013). Functional dissection and evidence for intercellular transfer of the heterocyst-differentiation PatS morphogen. Mol Microbiol 88:1093-1105. doi:10.1111/mmi.12244
[28] Rivers OS, Videau P, Callahan SM (2014). Mutation of sepJ reduces the intercellular signal range of a hetN-dependent paracrine signal, but not of a patS-dependent signal, in the filamentous cyanobacterium Anabaena sp. strain PCC 7120. Mol Microbiol 94:1260-1271.
[29] Black TA, Wolk CP (1994). Analysis of a Het - Mutation in Anabaena sp. Strain PCC 7120 Implicates a Secondary Metabolite in the regulation of heterocyst spacing. J Bacteriol 176:2282-2292.
[30] Bauer CC, Ramaswamy KS, Endley S, Scappino LA, Golden JW, Haselkorn R (1997). Suppression of heterocyst differentiation in *Anabaena* PCC 7120 by a cosmid carrying wild-type genes encoding enzymes for fatty acid synthesis. FEMS Microbiol Lett 151:23-39.
[31] Higa KC, Rajagopalen R, Risser DD, Rivers OS, Tom SK, Videau P, Callahan SM (2012). The RGSGR amino acid motif of the intercellular signalling protein, HetN, is required for patterning of heterocysts in Anabaena sp. strain PCC 7120. Mol Microbiol 83:682-693.
[32] Li B, Huang X, and Zhao J (2002). Expression of hetN during heterocyst differentiation and its inhibition of hetRup-regulation in the cyanobacterium Anabaena sp. PCC 7120. FEBS Lett 517: 87–91.

[33] Corrales-Guerrero L, Mariscal V, Nürnberg DJ, Elhai J, Mullineaux CW, Flores E, Herrero A (2014). Subcellular Localization and Clues for the Function of the HetN Factor Influencing Heterocyst Distribution in Anabaena sp. Strain PCC 7120. J Bacteriol 196:3452-3460. http://jb.asm.org/content/196/19/3452

[34] Callahan SM, Buikema WJ (2001). The role of HetN in maintenance of the heterocyst pattern in *Anabaena* sp. PCC 7120. Mol Microbiol 40:941-950.

[35] Videau P, Oshiro RT, Cozy LM, Callahan SM (2014). Transcriptional dynamics of developmental genes assessed with an FMN-dependent fluorophore in mature heterocysts of Anabaena sp. strain PCC 7120. Microbiol 160:1874-1881. DOI 10.1099/mic.0.078352-0

[36] Stucken K, John U, Cembella A, Murillo AA, Soto-Liebe K, Fuentes-Valdés JJ, Friedel M, Plominsky AM, Vásquez M, Glöckner G (2010). The Smallest Known Genomes of Multicellular and Toxic Cyanobacteria: Comparison, Minimal Gene Sets for Linked Traits and the Evolutionary Implications. PLoS One 5:e9235. doi:10.1371/journal.pone.0009235

[37] Wang H, Sivonen K, Rouhiainen L, Fewer DP, Lyra C, Rantala-Ylinen A, Vestola J, Jokela J, Rantasärkkä K, Li Z, Liu B (2012). Genome-derived insights into the biology of the hepatotoxic bloom-forming cyanobacterium Anabaena sp. strain 90. BMC Genomics 13:613. http://www.biomedcentral.com/1471-2164/13/613

[38] Zhang J-Y, Chen W-L, Zhang C-C (2009). hetR and patS, two genes necessary for heterocyst pattern formation, are widespread in filamentous nonheterocyst-forming cyanobacteria. Microbiol 155:1418-1426. DOI 10.1099/mic.0.027540-0

[39] Elhai J, Taton A, Massar JP, Myers JK, Travers M, Casey J, Slupesky M, Shrager J (2009). BioBIKE: A web-based, programmable, integrated biological knowledge base. Nucleic Acids Res 37:W28–W32. doi:10.1093/nar/gkp354.

[40] Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.

[41] Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673–4680, doi:10.1093/nar/22.22.4673.

[42] Talavera, G.; Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **2007**, *56*, 564–577, doi:10.1080/10635150701472164.

[43] Guindon, S.; Dufayard, J.F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321, doi:10.1093/sysbio/syq010.

[44] Rambaud, A. Fig Tree. Available online: http://tree.bio.ed.ac.uk/software/figtree/ (accessed on 10 March 2015).

[45] Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (1998). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes.

[46] Krogh A, Sonnhammer E (2000). TMHMM. http://www.cbs.dtu.dk/services/TMHMM/

[47] ~~Petersen TN, Brunak S, von Heijne G, Nielsen H (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 8:785-786. 10.1038/nmeth.1701~~

[48] ~~SignalP 4.1. http://www.cbs.dtu.dk/services/SignalP/~~

[49] Petersen TN, Brunak S, von Heijne G, Nielsen H (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 8:785-786. 10.1038/nmeth.1701

[50] SignalP 4.1. http://www.cbs.dtu.dk/services/SignalP/

[51] Sazuka T, Yamaguchi M, Ohara O (1999). Cyano2Dbase updated: Linkage of 234 protein spots to corresponding genes through N-terminal microsequencing. Electrophoresis 20:2160-2171.

[52] Crooks GE, Hon G, Chandonia JM, Brenner SE (2004). WebLogo: A sequence logo generator, *Genome Research*, 14:1188-1190. http://weblogo.berkeley.edu/logo.cgi

[53] Howard-Azzeh M, Shamseer L, Schellhorn HE, Gupta RS (2014). Phylogenetic analysis and molecular signatures defining a monophyletic clade of heterocystous cyanobacteria and identifying its closest relatives. Photosynth Res 122:171-185.

[54] Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, Calteau A, Cai F, Tandeau de Marsac N, Rippka R, Herdman M, Sivonen K, Coursin T, Laurent T, Goodwin L, Nolan M, Davenport KW, Han CS, Rubin EM, Eisen JA, Woyke T, Gugger M, Kerfeld CA (2014). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proc Natl Acad Sci USA 110:1053-1058. doi:10.1073/pnas.1217107110.

[55] Sánchez-Baracaldo P, Ridgwell A, Raven JA (2014). A neoproterozoic transition in the marine nitrogen cycle. Curr Biol 24:6520657. http://dx.doi.org/10.1016/j.cub.2014.01.041

[56] Schirrmeister BE, Gugger M, Donoghue PCJ (2015). Cyanobacteria and the great oxidation event: Evidence from genes and fossils. Paleontol 58:769-785.

[57] Uyeda JC, Harmon LJ, Blank CE (2016). A Comprehensive Study of Cyanobacterial Morphological and Ecological Evolutionary Dynamics through Deep Geologic Time. PloS ONE 11(9): e0162539. https://doi.org/10.1371/journal.pone.0162539

[58] Saw JHW, Schatz M, Brown MV, Kunkel DD, Foster JS, Shick H, Christensen S, Hou S, Wan X, Donachie SP (2013). Cultivation and Complete Genome Sequencing of Gloeobacter kilaueensis sp. nov., from a Lava Cave in Kılauea Caldera, Hawai'i. PloS ONE 8:e76376.

[59] Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY (1979). Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria. J Gen Microbiol 111:1-61.

[60] Kim Y, Joachimiak G, Ye Z, Binkowski TA, Zhang R, Gornicki P, Callahan SM, Hess WR, Haselkorn R, Joachimiak A (2011). Structure of transcription factor HetR required for heterocyst differentiation in cyanobacteria. Proc Natl Acad Sci USA **108,** 10109–10114.

[61] Kim Y, Ye Z, Joachimiak G, Videau P, Young J, Hurd K, Callahan SM, Gornicki P, Zhao J, Haselkorn R, Joachimiak A (2013). Structures of complexes comprised of *Fischerella* transcription factor HetR with *Anabaena* DNA targets. Proc Natl Acad Sci USA 110:E1716–E1723.

[62] Hu H-X, Jiang Y-L, Zhao M-X, Cai K, Liu S, Wen B, Lv P, Zhang Y, Peng J, Zhong H, Yu H-M, Ren Y-M, Zhang Z, Tian C, Wu Q, Oliveberg M, Zhang C-C, Chen Y, Zhou C-Z (2015). Structural insights into HetR-PatS interaction involved in cyanobacterial pattern formation. Sci Reports 5:16470. 10.1038/srep16470

[63] Risser DD, Callahan SM (2007). Mutagenesis of hetR reveals amino acids necessary for HetR function in the heterocystous cyanobacterium Anabaena sp. strain PCC 7120. J Bacteriol 189:2460–2467.

[64] Dong Y, Huang X, Wu X-Y, Zhao J. (2000). Identification of the active site of HetR protease and its requirement for heterocyst differentiation in the cyanobacterium *Anabaena* sp. strain PCC 7120. J. Bacteriol. 182:1575–1579.

[65] Feldmann EA, Ni S, Sahu ID, Mishler CH, Risser DD, Murakami JL, Tom SK, McCarrick RM, Lorigan GA, Tolbert BS, Callahan SM, Kennedy MA (2011). Evidence for direct binding between HetR from *Anabaena* sp. PCC 7120 and PatS-5. Biochemistry 50:9212–9224.
[66] Meeks JC, Elhai J, Thiel T, Potts M, Larimer F, Lamerdin J, Predki P, Atlas R (2002). An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. Photosyn Res 70:85-106.
[67] Voß B, Bolhuis H, Fewer DP, Kopf M, Möke F, Haas F, El-Shehawy R, Hayes P, Bergman B, Sivonen K, Dittmann E, Scanlan DJ, Hagemann M, Stal LJ, Hess WR (2013). Insights into the Physiology and Ecology of the Brackish-Water-Adapted Cyanobacterium *Nodularia spumigena* CCY9414 Based on a Genome-Transcriptome Analysis. PLoS ONE 8: e60224. https://doi.org/10.1371/journal.pone.0060224
[68] Yoon H-S, Lee MH, Xiong J, Golden JW (2003). *Anabaena* sp. strain PCC 7120 *hetY* gene influences heterocyst development. J Bacteriol 185:6995-7000. DOI: 10.1128/JB.185.23.6995–7000.2003
[69] Nayar AS, Yamaura H, Rajagopalan R, Risser DD, Callahan SM (2007). FraG is necessary for filament integrity and heterocyst maturation in the cyanobacterium Anabaena sp. strain PCC 7120. Microbiol 153:601-607. DOI 10.1099/mic.0.2006/002535-0
[70] Tumer NE, Robinson SJ, Haselkorn R (1983). Different promoters for the *Anabaena* glutamine synthetase gene during growth using molecular or fixed nitrogen. Nature 306:337-342. https://www.nature.com/articles/306337a0
[71] Flores, E., and A. Herrero. 1994. Assimilatory nitrogen metabolism and its regulation, p.487–517. In D. A. Bryant (ed.). The molecular biology of cyanobacteria. Kluwer Academic Publishers, Boston, Mass.
[72] Antonaru LA, Nürnberg DJ (2017). Role of PatS and cell type on the heterocyst spacing pattern in a filamentous branching cyanobacterium. FEMS Microbiol Lett 364, fnx154 DOI: 10.1093/femsle/fnx154
[73] Dawson JP, Weinger JS, Engleman DM (2002). Motifs of serine and threonine can drive association of transmembrane helices. J Mol Biol 316:799-805. doi:10.1006/jmbi.2001.5353
[74] Li X, Sandh G, Nenninger A, Muro-Pastor AM, Stensjö K (2015). Differential transcriptional regulation of orthologous *dps* genes from two closely related heterocyst-forming cyanobacteria. FEMS Microbiol Lett 362 (doi: 10.1093/femsle/fnv017).
[75] Picossi S, Flores E, Herrero A (2014). ChIP analysis unravels an exceptionally wide distribution of DNA binding sites for the NtcA transcription factor in a heterocyst-forming cyanobacterium. BMC Genomics 15:22. http://www.biomedcentral.com/1471-2164/15/22
[76] National Center for Biotechnology Information (2015). Prokaryotic RefSeq genome re-annotation project. https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/reannotation/ (retrieved 14 July 2017).
[77] Waterbury JB (2006). The cyanobacteria – isolation, purification, and identification. In: *The Prokaryotes: Bacteria: Firmicutes, Cyanobacteria*, 3rd edition, Vol. 4. Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E, eds. Springer. pp.1053-1073. DOI:10.1007/0-387-30744-3_38
[78] Koch R, Kupczok A, Stucken K, Ilhan J, Hammerschmidt K, Dagan T (2017). Plasticity first: molecular signatures of a complex morphological trait in filamentous cyanobacteria. BMC Evol Biol 17:209. DOI 10.1186/s12862-017-1053-5

[79] Gugger MF, Hoffmann L (2004). Polyphyly of true branching cyanobacteria (Stigonematales). Intl J Syst Evol Microbiol 54:349-357. DOI 10.1099/ijs.0.02744-0

[80] Henikoff S, Henikoff JG (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915-10919.

[81] Chao K-M, Zhang, L (2009). Sequence Comparison: Theory and Methods. In: Scoring Matrices, Chapter 8. Springer-Verlag, London. pp.149-172. https://link.springer.com/content/pdf/10.1007/978-1-84800-320-0_8.pdf

[82] Monera OD, Sereda TJ, Zhou NE, Kay CM, Hodgesis TS (1995). Relationship of Sidechain Hydrophobicity and α-Helical Propensity on the Stability of the Single-stranded Amphipathic α-Helix. J Peptide Sci 1:319-329.