Published online 8 February 2018 *Nucleic Acids Research, 2018, Vol. 46, No. 5* **2265–2278**
doi: 10.1093/nar/gky075

Selection, periodicity and potential function for Highly Iterative Palindrome-1 (HIP1) in cyanobacterial genomes

Minli Xu[1], Jeffrey G. Lawrence[2] and Dannie Durand[1,3,*]

[1]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [2]Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA and [3]Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Received May 17, 2017; Revised December 22, 2017; Editorial Decision January 08, 2018; Accepted January 25, 2018

ABSTRACT

Highly Iterated Palindrome 1 (HIP1, GCGATCGC) is hyper-abundant in most cyanobacterial genomes. In some cyanobacteria, average HIP1 abundance exceeds one motif per gene. Such high abundance suggests a significant role in cyanobacterial biology. However, 20 years of study have not revealed whether HIP1 has a function, much less what that function might be. We show that HIP1 is 15- to 300-fold over-represented in genomes analyzed. More importantly, HIP1 sites are conserved both within and between open reading frames, suggesting that their overabundance is maintained by selection rather than by continual replenishment by neutral processes, such as biased DNA repair. This evidence for selection suggests a functional role for HIP1. No evidence was found to support a functional role as a peptide or RNA motif or a role in the regulation of gene expression. Rather, we demonstrate that the distribution of HIP1 along cyanobacterial chromosomes is significantly periodic, with periods ranging from 10 to 90 kb, consistent in scale with periodicities reported for co-regulated, co-expressed and evolutionarily correlated genes. The periodicity we observe is also comparable in scale to chromosomal interaction domains previously described in other bacteria. In this context, our findings imply HIP1 functions associated with chromosome and nucleoid structure.

INTRODUCTION

Repetitive sequences are pervasive features of genomes throughout the bacterial kingdom, and are highly diverse in length, mobility, abundance and spatial organization (1,2). DNA repeats act as substrates for deletion, duplication and rearrangement of genomic regions (3), contributing to genome evolution and plasticity. Some repeats are selfish elements; others have important cellular functions. Repetitive motifs play roles in chromosome compaction and maintenance through their ability to mediate a variety of DNA–protein interactions. For example, Repetitive Extragenic Palindromic sequences (REPs) can act as binding sites for gyrases, helicases, DNA polymerase and Integration Host Factors (IHFs) (reviewed in (1)). Chi sites are strand-biased, non-palindromic octamers that direct DNA double-strand break repair (4). Architecture IMparting Sequences, also strand-biased, non-palindromic octamers, provide chromosome architecture, allowing for orderly replication and segregation (5).

Bacterial repetitive sequences can act as gatekeepers for genetic exchange. DNA Uptake Short Sequences, first found in *Haemophilus influenzae*, are ∼10 bp long motifs that bind preferentially to DNA uptake machinery (6). In contrast, Clustered Regularly Interspaced Palindromic Repeats act as a prokaryotic adaptive immune system that provides resistance to alien genetic material (7). Repetitive sequences also contribute to regulation of genes and gene products. Some repetitive motifs are preferentially located near transcription start sites or near the 3′-ends of genes (8,9), consistent with roles in the regulation of transcription. Similarly, the spatial distribution of Enterobacterial Repetitive Intergenic Consensus sequences suggests a role in accelerating mRNA decay (10). Repeats may exert a regulatory influence by virtue of their ability to form DNA or RNA secondary structures, such as the G-rich quadruplex structures formed by G4 repeats (9), or small stem loops arising from palindromic base pairing in REPs (11,12).

Highly Iterative Palindrome-1 (HIP1) is an octamer palindrome (GCGATCGC) that appears in high frequencies in most cyanobacterial genomes; high HIP1 abundance

*To whom correspondence should be addressed. Tel: +1 412 268 6036; Fax: +1 412 268 7129; Email: durand@cmu.edu
Present address: Minli Xu, Blizzard Entertainment, Irvine, CA 92618, USA.
Disclaimer: Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

has not been observed in genomes outside the cyanobacteria (13). HIP1 frequency can be as high as one per 350 nt; at this frequency, every gene, on average, will be associated with more than one HIP1 motif. Such widespread overabundance implies an important role for HIP1 in the biology of cyanobacterial genomes. However, the functional and molecular roles of HIP1, if any, as well as the forces driving its high abundance, remain unresolved (14–16).

HIP1 has been examined from various perspectives since it was first identified in the early 1990s. HIP1 has long been used for strain identification and polymerase chain reaction-based sequence amplification (14,17,18) and has been considered as a potential tool for site-specific recombination (19). Based on a scan of short mobile elements within cyanobacterial genomes, Elhai *et al*. (20) reported that HIP1 could be an insertion site for Small Dispersed Repeat-5 (SDR5) sequences in the Nostocaceae lineage, but no SDR or SDR-like repeat was found in any HIP1-enriched cyanobacterial genome outside of the Nostocaceae. Krishna *et al.* (21) proposed that HIP1 is a transcription factor binding site in *Synechococystis* sp. PCC6803. However, this hypothesis is not supported by previous or subsequent electrophoretic mobility shift assays (EMSAs) (15,22). Based on an analysis of the co-occurrence of PFAM domains and HIP1 abundance, Delaye and Moya (13) proposed that HIP1 may be functionally linked to *OpcA*, the glucose 6-phosphate dehydrogenase assembly protein. However, the recent discovery of possible HIP1 variants in genomes previously thought to lack HIP1 abundance ((23) and this study) calls for a re-examination of function predictions based on the phylogenetic distribution of HIP1-abundant genomes. Based on the observation that the central nucleotides of the HIP1 motif (GATC and CGATCG) are methylation sites, Elhai (23) proposed that HIP1 sites are introduced into the genome by a unidirectional mutational ratchet that is driven by a methyl-directed mismatch repair system. This model suggests a potential mechanism whereby HIP1 motifs are created, but why HIP1 sequences are maintained, if at all, remains an open question.

Despite 20 years of study, fundamental questions about HIP1 remain unanswered. What processes act to maintain high HIP1 frequency in the genome; is HIP1 under selection or does constant generation of new HIP1 sites by a neutral process offset mutational decay? If HIP1 is under selection, what are its functional roles? In this study, we test the hypothesis that HIP1 motifs are under selection using a comparative genomic approach. We show that HIP1 motifs are more conserved than expected and that this conservation is not a by-product of codon conservation. Our results support the hypothesis that HIP1 motifs are maintained by selection, suggesting that HIP1 motifs likely perform a biological function. We find no evidence for a function related to transcriptional or post-transcriptional regulation. Rather, we observe a statistically significant periodicity in the spatial distribution of HIP1 motifs within many cyanobacterial genomes and verify that this periodicity is not caused by a periodicity in genome background composition. Our results suggest that selection for HIP1 abundance is not acting on individual HIP1 instances, but instead on its distribution along the chromosome, implying a potential HIP1 function associated with chromosomal structure or maintenance. The mechanisms by which chromosome structure is maintained within cyanobacteria are largely unknown. In *Escherichia coli*, between 10 and 20 different low-molecular-mass DNA-binding nucleoid-associated proteins (NAPs) play a central role in shaping chromosome structure (24). However, with the exception of HU and IHF, the phylogenetic range of these NAPs is restricted to proteobacteria (25–27).

MATERIALS AND METHODS

Genomes and dataset

The enrichment, conservation and spatial distribution of HIP1 motifs in the 71 cyanobacterial genomes that were completely sequenced before April 2015 were investigated. Genome sequences (Supplementary Table S1) were obtained from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/). Protein-coding regions were identified using the annotation provided (*.ptt files). Motif coordinates were extracted from these sequences (*.fna files). Plasmids were excluded from this study.

Species relationships

A maximum likelihood phylogeny was constructed from 16S rRNA sequences with 100 bootstrap replicates using PhyML (28) via TOPALi 2.5 (29). Three *Escherichia* genomes (*E. coli* K12 MG1655, *E. coli* O81 ED1a and *Escherichia fergusonii* ATCC 35469) were used to root the tree (not shown). Multiple sequence alignments were constructed using MUSCLE (30), followed by manual refinement. Model selection was performed in TOPALi 2.5 using ModelGenerator (31); the best model was GTR+I+Γ according to the Akaike Information Criteria (AIC1 and AIC2). The resulting phylogeny (Figure 1) is in agreement with recently published cyanobacterial phylogenies (32–35). The phylogenetic diagram, annotated with HIP1 abundance and enrichment statistics, was generated using the iTOL web application (36).

Estimation of the expected number of motifs

To assess whether the observed HIP1 abundance is due to underlying sequence composition, the expected number of HIP1 motifs was calculated using a second order Markov model of sequence composition to account for background tri-nucleotide frequencies (37), as described in Supplementary Methods. Since oligonucleotide frequencies differ between protein-coding and non-coding regions, the expected number of motifs was estimated separately for coding and non-coding regions and summed to obtain the genome-wide total. In coding regions, the expected number of HIP1 motifs is the sum of the expected number in each reading frame. For each annotated open reading frame (ORF), the downstream non-coding region was defined to be the segment between the 10th base downstream of the stop codon and the start of the next annotated feature. To exclude potential unannotated features, non-coding regions longer than 2000 bp were truncated at position 2000.

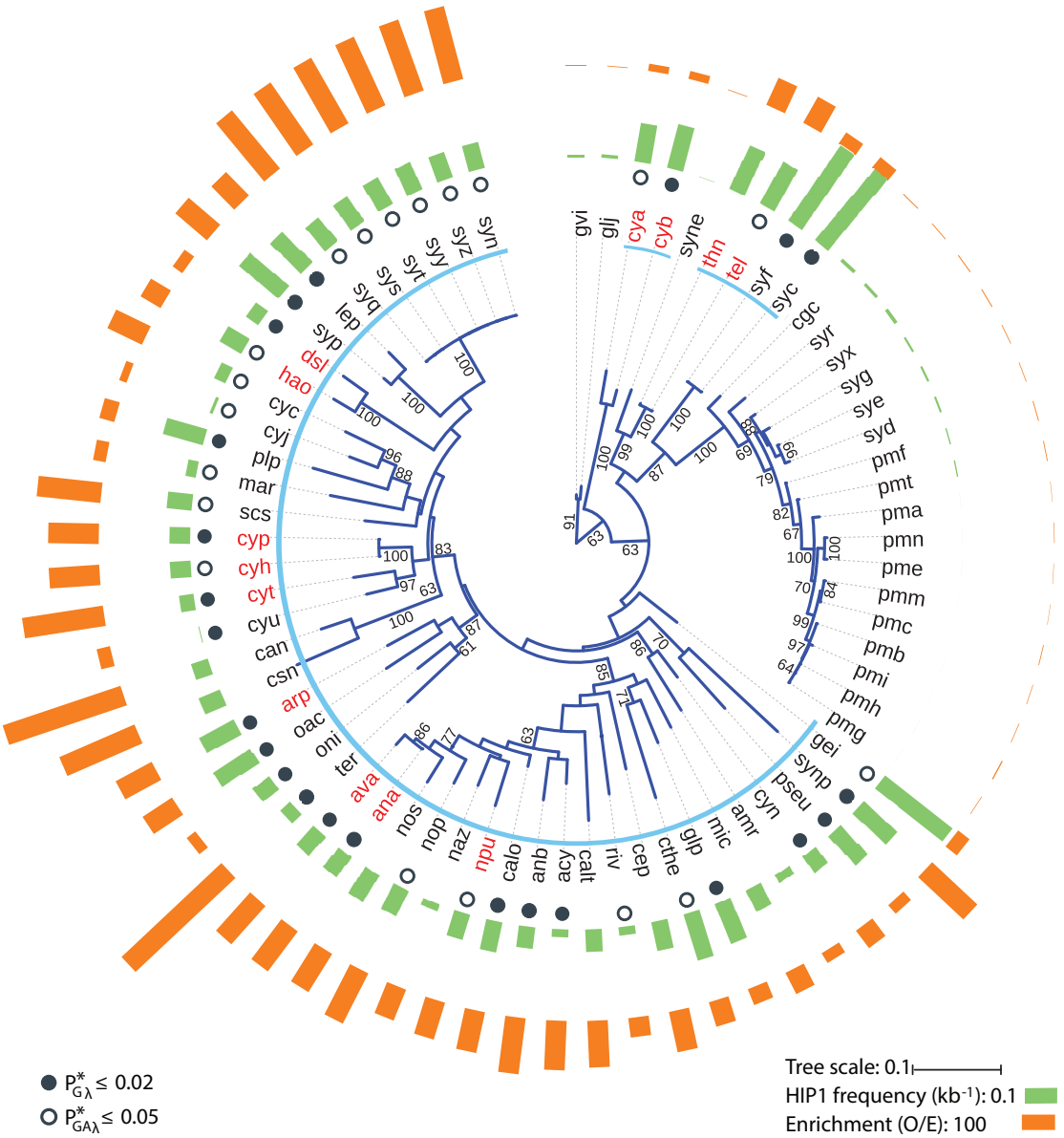

Figure 1. Maximum likelihood phylogeny of the 71 complete genomes used in this study. Tree constructed from 16S rRNA sequences with 100 bootstrap replicates (see 'Materials and Methods' section); bootstrap values below 60 not shown. Genomes used in conservation analyses labeled in red. (Three incomplete genomes that were used in the conservation analyses are not included in this tree.) Charcoal dots indicate genomes with significant HIP1 periodicity (filled dots; $P^*_{G\lambda} \leq 0.02$, open dots; $P^*_{GA\lambda} \leq 0.05$). Light blue ring indicates HIP1-rich genomes (O/E > 16). Full binomial species names given in Supplementary Table S1.

Conservation assessment

To determine whether selection is acting on HIP1 motifs, motif conservation was assessed in genome pairs selected to represent a range of evolutionary distances. Pairwise genome divergence was quantified by the average divergence at synonymous sites (K_S) over all orthologous pairs of protein-coding genes (38), weighted by ORF length, as calculated by DNA Master (http://cobamide2.bio.pitt.edu/computer.htm). Orthologous pairs were predicted using unique reciprocal best BLASTP hits, where the alignment must cover more than 80% of both genes. Eight pairs of genomes were selected with evolutionary divergences ranging from $K_S = 0.03$ to $K_S = 1.2$. These eight pairs include 13 genomes from our primary dataset (highlighted in red in Figure 1) and three partially sequenced genomes (Supplementary Table S4), added to provide genome pairs with a range of evolutionary divergences that are appropriate for assessing conservation.

Motif conservation in these eight pairs was assessed over all aligned blocks inferred from pairwise, whole genome alignments, which were constructed using the progressive alignment function in MAUVE version 2.3.1 (39), using default parameter settings. For a pair of genomes, conservation was quantified using the Jaccard index (i.e. the fraction of possible motif sites at which a motif in one genome is perfectly aligned with a motif in another genome), as de-

scribed in Supplementary Methods. Motif conservation was assessed separately in protein-coding and non-coding regions; within protein-coding regions, reading frame-specific conservation scores were also calculated. Confidence intervals were calculated from standard errors estimated using the jackknife function in the R Bootstrap package (version 2015.2).

To determine whether selection is acting on degenerate HIP1 motifs, we also assessed the conservation of octamers that differ from HIP1 at exactly one position (collectively denoted HIP1* motifs). For assessing motif conservation relative to overall nucleotide conservation in the genome pair, a heterogeneous set of control motifs was used, consisting of all palindromic octamers with the same GC content as HIP1. For both HIP1* and control motifs, a site is conserved if a motif in one genome is perfectly aligned with the *same* motif in the other genome.

Motif conservation, if observed, could be due to selection acting on codon usage. To rule out this possibility, we compared the conservation of each in-frame codon found in HIP1 motifs with the conservation of the same codon outside HIP1 motifs. This analysis was restricted to the regions within the alignable blocks that were annotated as coding regions in both genomes.

Correlation between expression and motif abundance

Potential HIP1 functions related to the regulation of genes and gene products were investigated using mRNA transcript data measured by direct RNA sequencing in *Synechococcus elongatus* PCC 7942 (syf) (16). From this dataset, we obtained the genomic coordinates and the normalized abundance of each transcript in cells grown in constant light conditions. The data were pre-processed to remove some outliers and complete overlaps as described in Supplementary Methods. To determine whether the observed HIP1 conservation could be due to a functional link with an RNA motif, motif enrichments in transcribed and non-transcribed regions were compared. To assess a potential role for HIP1 as a transcription factor binding site, we calculated the distance from each motif to the closest predicted transcription start site (16).

To investigate whether HIP1 could be related to mRNA expression, the Pearson correlation coefficients of transcript expression level and transcript motif frequency were calculated for HIP1 and control motifs. Expression level was quantified in two ways: transcript abundance, in units of mRNA molecules per cell, reflects expression under specific conditions. Codon usage bias reflects selection acting on expression levels under many different conditions. The codon usage bias for each transcript was estimated by the mean Adaptive Codon Enrichment (ACE_u) (40) of the protein coding genes in the transcript, calculated with DNAMaster (http://cobamide2.bio.pitt.edu/computer.htm).

To examine the impact of transcript length on observed correlations between expression and motif frequency, we performed linear regression on transcript motif frequency and transcript length. We then assessed the relationship between motif frequency and expression, independent of transcript length, by calculating the Pearson correlation coefficient of the residual motif frequencies with both measures of expression (transcript abundance and transcript ACE_u).

Observed correlations between HIP1 frequency and expression level could be influenced by biased usage of the codons that make up HIP1. To determine the magnitude of this bias, preferred codons were identified by comparing codon usage in the core genome (f_N, reflecting mutational biases) with codon usage in ribosomal genes (f_O, reflecting mutation and selection) in *S. elongatus* PCC 7942 (syp). Codon preference was measured as enrichment, the ratio of the residue-specific frequencies between the two sets (f_O/f_N); preferred codons have ratios >1. The enrichment values of the six codons found in the canonical HIP1 motif were GCG: 0.721, CGA: 0.305, GAT: 0.674, ATC: 1.239, TCG: 1.088 and CGC: 0.963. When weighted by the relative abundance of the six codons in the *syf* genome, the average enrichment is 0.83.

Assessing periodicity

Determining the periodicity of inter-motif spacings by fitting them to a damped sine wave. To investigate the possibility that HIP1 has a function related to overall genome structure, we asked whether the chromosomal distribution of HIP1 motifs is periodic. A detailed description of the computational analysis used to assess motif periodicity is given in Supplementary Methods. Briefly, we calculated the distance, in base pairs, between every pair of motifs separated by at most 1000 kb and constructed a binned distribution (i.e. a histogram) of these inter-motif spacings, where b is the size of each bin. Periodicity was assessed by fitting a damped sine wave, with period λ and amplitude A, to the Pearson autocorrelation function of this spacing distribution. The parameters of the sine wave were estimated by minimizing a variance-normalized χ^2 function, G, used to assess goodness of fit.

Assessing the significance of the quality of the fit to the sine wave. To determine whether the inter-motif spacing distribution fits a sine wave better than expected by chance, the analysis was repeated with 200 replicates of randomized motif positions. For each replicate, a damped sine wave was fit to each randomized genome, wherein HIP1 sites were randomly reassigned within the genome, with the restriction that sites must be separated by at least 8 bp. The significance of the fit to the genuine data was quantified by $P(G)$, the fraction of randomized genomes with at least as good a fit as that obtained with the genuine data, and $P(A)$, the fraction of randomized genomes with amplitudes at least as great as the amplitude of the sine wave fit to the genuine genome.

Estimating the genome-wide, global period. If the HIP1 distribution is truly periodic, then the estimate of the period should be insensitive to the bin size used to construct the binned distribution of HIP1 spacings. To obtain a global estimate of the genomic period, independent of bin size, the analysis was carried out for bin sizes ranging from 25 to 7000 bp in 25 bp intervals. This resulted in 280 estimates, each with period, λ_b, amplitude, A_b and goodness of fit, G_b. From these, we obtained a bin-size independent estimate of

the period by identifying values of λ_b that are most representative of the entire set of estimates; that is, values of λ_b that maximize the number of estimates of the period that are within 10% of λ_b. We define the arithmetic mean of the most representative periods to be the genome-wide, *global* period, denoted λ^*.

Assessing the stability and significance of the estimate of the global period. Since random data can manifest apparent organization, occasional observation of a periodic signal in a randomized genome is not unexpected. However, consistent good fits with the same period over many bin sizes is unlikely to occur by chance. To determine the stability of the estimated global period, λ^*, we examined every interval of at least 60 contiguous bin sizes for which the mean period, averaged over all bin sizes in the interval, is within 10% of λ^*. For each such interval, the mean goodness of fit, $\bar{G}$, the mean amplitude, $\bar{A}$, and the standard error of the period, $SE(\lambda)$, were calculated from the individual values of G_b, A_b and λ_b in that interval for the genuine and randomized genomes, and compared. The quality of fit and consistency of the inferred periods for a given interval were then assessed by $P(\bar{G}, SE(\lambda))$, the fraction of randomized genomes for which the best fits in that interval yield lower values of both $\bar{G}$ and $SE(\lambda)$ than the best fits to the genuine data over the same range of bin sizes. If none of the 200 randomized genomes meet these criteria, we assign the interval a *P*-value of 0.0025, which is half of the lowest non-zero probability obtainable with a sample of 200 genomes.

The significance of the periodicity of the genome as a whole is defined to be $P^*_{G\lambda}$, the smallest value of $P(\bar{G}, SE(\lambda))$ over all intervals of at least 60 contiguous bin sizes that possess a mean period within 10% of the genomic period. If no such interval exists, then the periodicity of the genome is not significant and $P^*_{G\lambda}$ is arbitrarily assigned a value of 1.

We also assessed significance with a test statistic that includes the additional requirement that the amplitude be as strong in the randomized genomes as in the genuine data. For a given interval of bin sizes, we define $P(\bar{G}, \bar{A}, SE(\lambda))$ to be the fraction of randomized genomes for which $\bar{G}$ and $SE(\lambda)$ are lower and the mean amplitude, $\bar{A}$, is higher than the corresponding values for the genuine data. Then, $P^*_{GA\lambda}$, is defined to be the smallest value of $P(\bar{G}, \bar{A}, SE(\lambda))$ over all such intervals.

To determine whether periodicity is a general feature of HIP1 distributions, we applied the analysis described above to the 50 HIP1-rich genomes (excluding *cyu*) in our dataset (Figure 1 and Supplementary Table S7) and calculated the global period and the overall significance of the HIP1 periodicity in each genome.

Ruling out periodicity due to mutational bias. To exclude the possibility that the observed periodicity is due to an underlying genomic periodicity arising from mutational bias, the same approach for assessing periodicity was applied to non-HIP1 motifs. If significant HIP1 periodicity reflects a general genomic property, then other motifs in the same genome should also be significantly periodic and all motifs should share a common period. If this is the case, the inferred periods of motifs in the same genome will be correlated and this correlation will be strongest in genomes in which HIP1 periodicity is highly significant. Similarly, the *P*-values of these motifs will be most significant in genomes in which HIP1 periodicity is highly significant. To test this, for each genome, the five hexamer palindromes with abundances closest to the abundance of HIP1-like motifs were selected as controls (Supplementary Table S8); hexamers rather than octamers were used to obtain motifs with sufficiently high abundance. The periodicity analysis described above was applied to each of the five hexamer controls to assess the significance of its periodicity and obtain an estimate of its period. (Note that given *any* set of genomic sites as input, this analysis will return the period of the best-fit damped sine wave, regardless of the quality of the fit.) Within the same genome, the consistency of periods across motifs was assessed by the coefficient of variation (CV = σ/μ), where μ and σ are the mean and the standard deviation of the periods of six motifs (HIP1-like motifs plus 5 hexamer palindromes) for each genome. Low values of CV indicate that HIP1-like motifs and hexamer controls have similar periods. Tests for significant differences between CV values and mean *P*-values for control motifs associated with genomes with highly significant periodicity ($P^*_{G\lambda} \leq 0.01$) and without significant periodicity ($P^*_{G\lambda} > 0.1$) were performed using one-way ANOVA.

RESULTS

Widespread over-representation of HIP1-like motifs in cyanobacterial genomes

We investigated the frequency and enrichment of all octamer palindromes in 71 complete cyanobacterial genomes (Figure 1 and Supplementary Table S1) to identify octamer palindromes that are exceptionally prevalent. These genomes represent all five major cyanobacterial ecological subsections (41), with genome sizes varying from 1.44 mb in *Candidatus* Atelocyanobacterium thalassa ALOHA (cyu) to 8.36 mb in *Acaryochloris marina* MBIC11017 (ama), and GC content ranging from 30.8% in *Prochlorococcus marinus* MIT 9515 (pmc) to 68.7% in *Cyanobium gracile* PCC 6307 (cgc).

The canonical HIP1 motif (GCGATCGC) is indeed exceptionally abundant in most genomes (Figure 2 and Supplementary Table S1), with frequencies ranging from 0.2 to 2.73 kb^{-1}. Genomes with low HIP1 frequencies ($f < 0.2$ kb^{-1}) are either reduced genomes of pico-cyanobacteria or found in the earliest branching species (Figure 1). HIP1 abundance is also low in *Ca.* Atelocyanobacterium thalassa (cyu), *Cyanothece sp.* PCC 7822 (cyj) and *Synechococcus sp.* PCC 6312 (syne).

To identify motifs that are more abundant than expected, we calculated the enrichment, defined here to be the ratio of observed (O) to expected (E) motifs, of all octamer palindromes (Supplementary Table S1 and Figure S1). The number of expected motifs was estimated separately for each reading frame in predicted ORFs and for non-coding regions, using a second order Markov model to account for underlying genomic sequence composition. Genomes with high HIP1 abundance also have high enrichment (Figure 3A and Supplementary Table S1) with frequencies ranging from 16- to 317-fold greater than expected. Two genomes with low HIP1 abundance are also enriched for HIP1: *Ca.*

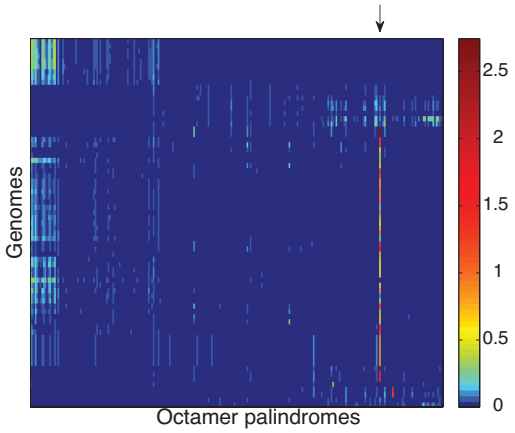

Figure 2. Frequencies (kb^{-1}) of all octamer palindromes in 71 cyanobacterial genomes. Rows represent genomes ordered according to the phylogeny in Figure 1. Columns are in order of increasing octamer GC content. Arrow indicates the column corresponding to the canonical HIP1 motif (GCGATCGC).

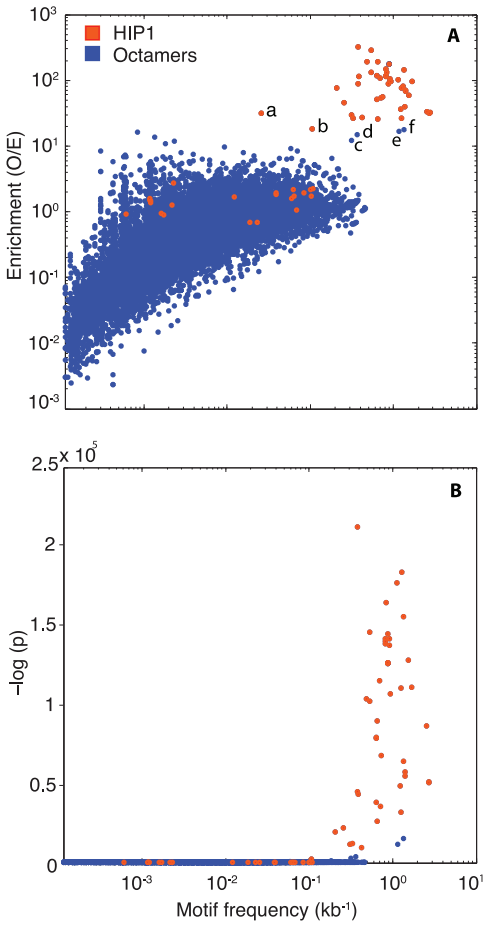

Figure 3. Enrichment versus abundance of octamer palindromes in all 71 genomes. (**A**) Enrichment (O/E) as a function of motif frequency (kb^{-1}). Each dot represents one octamer palindrome from one genome. (**B**) The significance of enrichment (-log(p), χ^2 test) as a function of motif frequency (kb^{-1}). The canonical HIP1 motif (GCGATCGC) is shown in red; other octamer palindromes in blue. (a) GCGATCGC in *Candidatus* Atelocyanobacterium thalassa ALOHA (cyu); (b) GCGATCGC in *Cyanothece sp*. PCC 7822 (cyj); (c) TCGATCGA in *Dactylococcopsis salina* PCC 8305 (dsl); (d) CAGGCCTG in *Synechococcus sp*. PCC 6312 (syne); (e) GGGATCCC in *Synechococcus sp*. JA-3–3Ab (cya); (f) GGGATCCC in *Synechococcus sp*. JA-2–3B'a (cyb).

Atelocyanobacterium thalassa (cyu) and *Cyanothece sp.* PCC 7822 (cyj) (*a* and *b* in Figure 3A).

When motif enrichment is plotted against motif frequency (Figure 3A), non-HIP1 octamer palindromes (blue) form a dense cluster, with motif frequencies below 0.2 kb^{-1} and <10-fold enrichment. HIP1 motifs (red) from genomes with low HIP1 frequency ($f < 0.2$ kb^{-1}) also fall in this region. In contrast, HIP1 motifs from most HIP1-enriched genomes form a distinct cluster to the upper right. The observed enrichment is highly significant for all motifs with at least 16-fold enrichment ($P \leq$ 1E-300, χ^2 test; Figure 3B). The observed enrichment indicates that the high HIP1 abundance cannot be explained by the background sequence composition.

We next considered the extent to which HIP1 abundance is exceptional among octamer palindromes. HIP1-poor genomes (O/E < 16) possess no other highly abundant octamer palindromes, with three exceptions (Supplementary Table S2). *Synechococcus sp*. JA-3–3Ab (cya) and *Synechococcus sp*. JA-2–3B'a (cyb), two closely related thermophilic strains isolated from a hot spring in Yellowstone National Park (42), possess a motif (GGGATCCC) that is very similar to the canonical HIP1 motif (GCGATCGC) in sequence, abundance and enrichment (e and f, Figure 3A). These similarities, combined with the near zero frequency of the canonical HIP1 motif in those genomes, suggest that the motif GGGATCCC is a variant form of HIP1 that plays the same role in the Yellowstone strains that the canonical HIP1 motif plays in HIP1-rich genomes. In a third HIP1-poor genome, *Synechococcus sp.* PCC 6312 (syne, a freshwater isolate (41)), the motif CAGGCCTG is moderately abundant ($f = 0.37$ kb^{-1}) and enriched (O/E = 14.72) (c, Figure 3A). However, while CAGGCCTG has the same GC content as HIP1, this motif does not have the central AT dinucleotide and aligns with the HIP1 motif at only two of the eight positions. HIP1 was the sole motif with high abundance and enrichment among all HIP1-rich (O/E >16) genomes except *Dactylococcopsis salina* PCC 8305 (dsl). In addition to abundant HIP1 ($f = 0.43$, O/E = 27.26), the genome of this isolate from a heliothermal saline pool (43) possesses a second high-abundance motif, TCGATCGA ($f = 0.32$ kb^{-1}, O/E = 12.22; d in Figure 3A).

For the rest of this study, we focus on 51 HIP1-rich genomes (blue ring in Figure 1), which we define to be all genomes with significant HIP1 enrichment (O/E >16, $P \leq 1 \times 10^{-300}$, χ^2 test) and the Yellowstone strains *Synechococcus sp*. JA-3–3Ab (cya) and *Synechococcus* sp. JA-2–3B'a (cyb). In the Yellowstone strains, the putative HIP1 variant GGGATCCC is used in all subsequent analyses. The 20 HIP1-poor genomes (O/E < 16) are not considered further.

HIP1 motifs are conserved between genomes

A motif will be conserved in related genomes if it is maintained by selection against mutational loss, but not if motif abundance is continually replenished by neutral processes. To investigate whether selection maintains HIP1

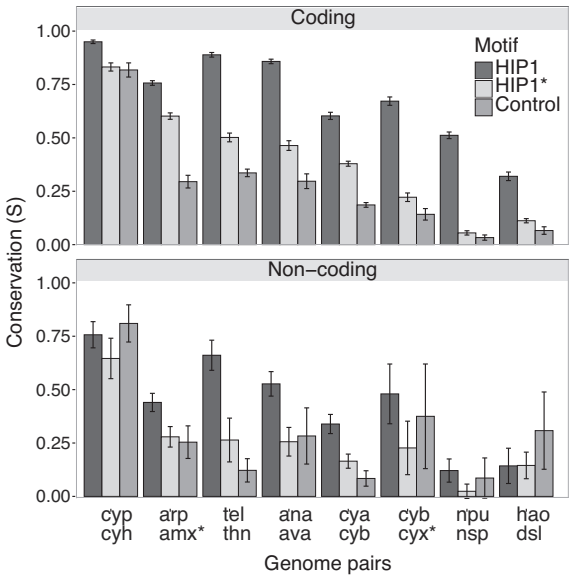

Figure 4. Motif conservation in genome pairs at varying levels of sequence divergence. Conservation of HIP1, HIP1* and control motifs in coding regions and non-coding regions. Error bars indicate 95% confidence intervals based on jackknife replicates. Genome pairs in order of increasing divergence (Supplementary Table S3).

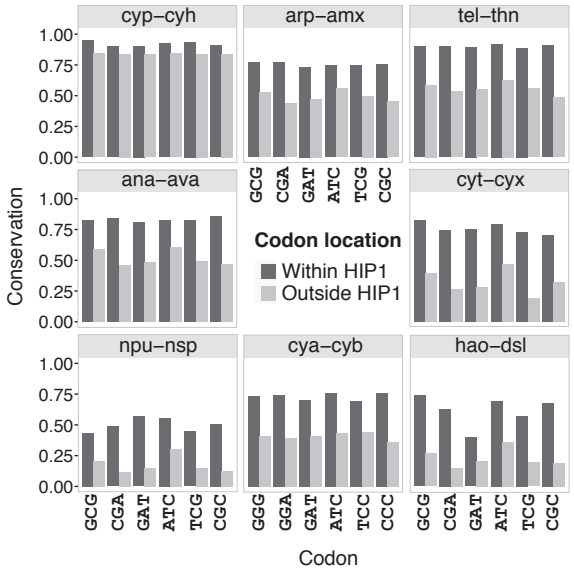

Figure 5. Codon conservation (*S*) within HIP1 motifs and outside HIP1 motifs for selected pairs of genomes. For the *cya–cyb* pair, codon conservation is assessed for the proposed Yellowstone HIP1 variant (GGGATCCC).

abundance, the conservation of HIP1 motifs was assessed in eight pairs of genomes selected at varying evolutionary distances (Supplementary Table S3). Each pair of genomes was aligned and a motif conservation score, *S*, was calculated, where *S* is the fraction of motif sites in aligned blocks that are perfectly aligned.

Because the target of selection is unclear, we also calculated the conservation score of HIP1* motifs (octamers that differ from HIP1 at one position) to determine whether degenerate HIP1 sites are under selection. To assess whether the resulting conservation scores simply reflect the overall level of conservation in the genome pair, we compared them with the conservation scores of control motifs, consisting of all non-HIP1 octamer palindromes with 75% GC content.

HIP1 is significantly more conserved than control palindromes in protein-coding regions in all the eight pairs of genomes ($P = 0.0039$, binomial test), regardless of the divergence between the genomes (Figure 4 and Supplementary Table S5). HIP1* is also more conserved than control motifs in protein-coding regions, although it is less conserved than HIP1 (Figure 4, $P = 0.0039$, binomial test). These trends are observed in all three reading frames, when analyzed separately (Supplementary Figure S4). In non-coding regions, HIP1 is significantly more conserved than controls in the three pairs of genomes in which the HIP1 motif counts in non-coding regions provide sufficient sample sizes (arp-amx, cya-cyb, ana-ava). These results are consistent with the hypothesis that both HIP1 and HIP1* are maintained by selection, and that the selection acting on HIP1* is weaker. The apparent HIP1* conservation could be due to parallel mutation from HIP1 in the ancestral genome, but we deem this unlikely as only ∼4% of double mutations independently affecting orthologous HIP1 motifs in each genome would produce the same HIP1* variant in both genomes. In the Yellowstone strains, the HIP1-like GGGATCCC motif is conserved to a degree comparable to HIP1 in other species. In coding regions, GGGATCCC is more conserved than motifs that differ from GGGATCCC at one position (i.e. the equivalent of HIP1* motifs), which in turn are more conserved than control motifs. Further, these levels of conservation are consistent with conservation of HIP1 and HIP1* in other genomes considering the degree of overall genomic divergence. This suggests both that GGGATCCC is under selection in those genomes and that GGGATCCC is an alternative form of HIP1 in those strains.

HIP1 could appear to be preferentially conserved if it were enriched for preferred codons relative to other palindromes. If codon selection is driving HIP1 conservation, then the codons that appear in HIP1 motifs should be equally conserved within and outside of HIP1 motifs. However, in each of the eight genome pairs tested, all six constituent codons were substantially more conserved within HIP1 than outside of HIP1 ($P = 0.016$, binomial test), ruling out the possibility that HIP1 conservation is an artifact of selection on codon preference (Figure 5). As expected, this effect becomes more pronounced with greater genome divergence.

Lack of evidence for gene-level functions for HIP1

Given HIP1 conservation, we considered which genomic features might be targets of selection. We first considered whether selection could be acting on an amino acid motif. If so, HIP1 would only be conserved in the reading frame corresponding to that motif. However, HIP1 is both enriched and conserved in all three reading frames (Supplementary Figures S3 and 4), suggesting either that the coding potential of HIP1 is not the reason for its conservation, or that amino acid motifs associated with all three

reading frames are under selection, which we deem unlikely. Moreover, HIP1 is both enriched (44) and conserved in non-coding regions (Figure 4), providing further evidence that HIP1 conservation is not maintained by selection acting on an amino acid motif.

If selection acts to conserve HIP1 via an RNA motif, we would expect to see higher HIP1 enrichment in transcribed regions. However, analysis of *S. elongatus* PCC 7942 (syf) transcripts (16) revealed even higher enrichment in non-transcribed than in transcribed regions (O/E = 37 versus O/E = 25), as compared with low enrichment of control motifs in both regions (O/E = 0.67 versus O/E = 0.65) (Supplementary Table S6).

We next considered whether selection is acting on HIP1 in the context of transcription factor binding. Many well-known transcription factor binding motifs are abundant palindromes (e.g. LexA and Crp binding sites in *E. coli* (45)). If HIP1 acts as a binding site for the transcription machinery, the distribution of HIP1 motifs relative to transcription start sites should differ from that of control motifs. To assess this, we estimated these distributions using transcription start sites predicted in the *syf* genome (16). The resulting distributions for HIP1 and control motifs are not significantly different ($P < 0.91$, one-sided Kolmogorov-Smirnov test; Supplementary Figure S5). We also compared the distributions of the distances from each motif to the nearest predicted start codon in 20 HIP1-rich genomes and again observed no significant difference in the HIP1 and control motif distributions (Supplementary Figure S6).

Finally, we studied the relationship between the abundance of an RNA transcript and its HIP1 content, using the *syf* expression dataset (16). A weak, but significant, negative correlation (R = -0.37, $P = 7.51 \times 10^{-27}$, Pearson test) was found between the expression level of a transcript (molecules per cell) and its HIP1 frequency (Supplementary Figure S7); that is, more abundant transcripts harbored significantly fewer HIP1 motifs. Since transcript abundance obtained from direct sequencing only captures expression under the specific experimental conditions tested, we also considered codon usage bias, which provides a measure of selection acting on expression levels under many different conditions. The degree of codon selection of a transcript and its HIP1 frequency were also anti-correlated (ACE_u: $R = -0.18$, $P = 4.15 \times 10^{-7}$, Supplementary Figure S8). Control motifs do not show these correlations.

Both HIP1 and control motifs exhibit a weak, but significant, negative correlation between motif frequency and transcript length (HIP1: $R = -0.22$, $P = 4.75 \times 10^{-10}$; control: $R = -0.27$, $P = 4.13 \times 10^{-15}$), suggesting that transcript length might be responsible for the observed negative correlations. However, a correction for transcript length did not eliminate the negative correlation of motif frequency with respect to transcript abundance (HIP1: $R = -0.34$, $P = 5.09 \times 10^{-23}$; control: $R = 0.03$, $P = 0.34$, Supplementary Figure S7) or transcript ACE_u (HIP1: R = -0.17, $P = 1.19 \times 10^{-6}$; control: $R = 0.07$, $P = 6.52 \times 10^{-2}$, Supplementary Figure S8).

These negative correlations are not unexpected as HIP1 is comprised of codons that are non-preferred in the *syf* genome (the average enrichment of HIP1 codons in highly expressed genes is 0.83). Therefore, HIP1 would be avoided in genes with strong codon usage bias. While HIP1 abundance is more strongly anti-correlated with transcript expression level than with codon usage bias, it is not clear whether this represents under-performance of metrics of codon selection in predicting gene expression level, or selection against HIP1 within highly transcribed regions in excess of selection acting on codon usage.

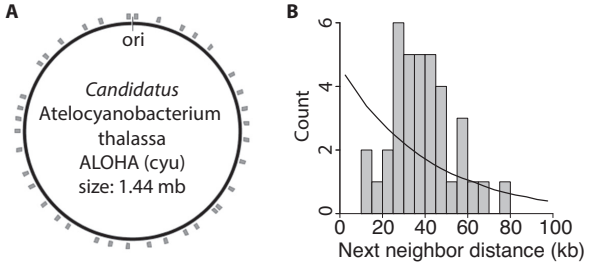

Figure 6. Spatial distribution of HIP1 motifs in *Candidatus* Atelocyanobacterium thalassa ALOHA (cyu). (**A**) Position of HIP1 motifs along the chromosome (gray). (**B**) Observed (gray) and expected (black) HIP1 next neighbor distances. The observed distribution is significantly different than expected ($P = 2 \times 10^{-4}$, two-sided Kolmogorov-Smirnov test). Expected next neighbor distances estimated from 1000 randomizations of motif sites.

Taken together, these analyses find no evidence of selection on particular HIP1 sites encoding amino acid, RNA or DNA motifs associated with promotors within cyanobacterial genomes and do not suggest a function related to the regulation of genes or gene products.

HIP1 is periodically distributed within genomes

Lacking a clear gene-level function for HIP1, we considered potential functions related to chromosomal architecture and maintenance. We posited that selection does not act on individual HIP1 motifs, but rather acts to maintain an idiosyncratic distribution of HIP1 along the chromosome. For example, the 37 HIP1 motifs in the reduced genome of the symbiont *Ca.* Atelocyanobacterium thalassa (cyu) appear to be regularly spaced, with a mean next neighbor distance of ∼39 kb (Figure 6A). This distribution is significantly non-random (Figure 6B; $P = 2 \times 10^{-4}$, two-sided Kolmogorov-Smirnov test). In contrast, the distribution of the second most abundant octamer palindrome with 75% GC content in *cyu* (GCTGCAGC) is neither regularly spaced (Supplementary Figure S9), nor significantly non-random ($P = 0.646$, two-sided Kolmogorov-Smirnov test).

Given the periodic nature of the HIP1 distribution in *cyu*, we hypothesized that periodicity is a general feature of HIP1 distributions. In *cyu*, HIP1 is so sparse that its periodic distribution is readily apparent. To assess periodicity in genomes with greater HIP1 abundance, we examined the distribution of HIP1-like octamers, defined to be the combined set of HIP1 and HIP1* motifs, because both HIP1 and HIP1* motifs are conserved by selection (Figure 4). We exemplify this approach with the *Synechococcus sp*. PCC 7002 (syp) genome.

The spacings between all pairs of HIP1-like octamers, not simply next neighbors as in Figure 6, were assessed throughout the genome. Figure 7A shows the distribution of ob-

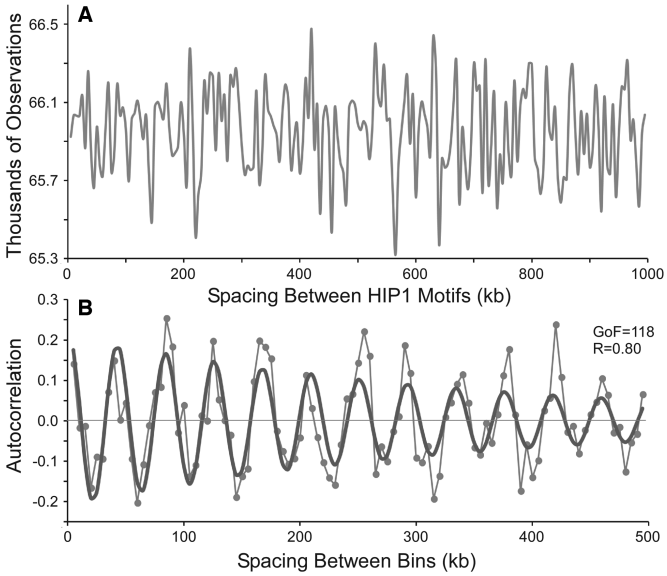

Figure 7. Detection of periodicity in the spatial distribution of HIP1-like motifs in *Synechococcus sp.* PCC 7002 (syp). (**A**) Histogram of all inter-HIP1 spacings up to 1000 kb. (**B**) Autocorrelation (light gray) of the histogram in (A) and the damped sine wave (dark gray) with the optimal fit to the autocorrelation array.

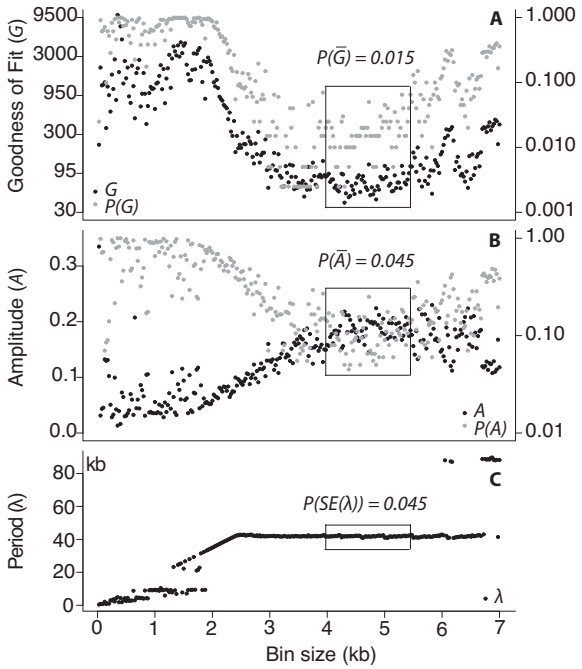

Figure 8. Inferred periodicity of HIP1-like motif spacings over varying bin sizes in *Synechococcus sp.* PCC 7002 (syp). (**A**) Goodness of fit, (**B**) amplitude and (**C**) period (black dots) of best-fit damped sine wave for 280 bin sizes, ranging from 25 to 7000 bp by 25 bp intervals. The chance probability of observing goodness of fit, G and amplitude, A, for each bin size shown in grey. The central box in each subfigure indicates an interval of 60 contiguous bin sizes. The probability of observing the mean values of G and A in the same interval in 200 randomized genomes are $P(\bar{G}) = 0.015$ and $P(\bar{A}) = 0.045$, respectively. The probability of observing the standard error of the period in the same interval in randomized genomes is $P(SE(\lambda)) = 0.045$.

served spacings, up to 1000 kb in length, for the 6299 HIP1-like motifs in the *syp* genome; each data point represents the number of spacings within a 5 kb bin, resulting in 200 bins. While this histogram is strongly suggestive of a periodic distribution of spacings between HIP1-like motifs, background noise precludes accurate assessment of periodicity within this distribution.

To reduce noise, we employed an autocorrelation analysis, wherein the Pearson correlation coefficient was calculated for all pairs of 5 kb bins separated by a fixed distance, for distances ranging from 5 kb (adjacent bins) to 500 kb. When the Pearson correlation coefficient (R) values are plotted against the distance between the bins (Figure 7B), any periodicity in the abundance of HIP1 spacings should manifest itself as periodicity in the autocorrelation values, where higher levels of noise would simply be reflected in lower R values. A periodic signal is clearly apparent in the autocorrelation analysis for *syp* (Figure 7B).

Fourier transform analysis is a typical approach to quantifying this signal. However, the 100 data points in Figure 7B do not provide a large enough sample to obtain a robust signal with a Fourier transform. We avoid this difficulty by introducing an alternate approach in which a damped sine wave is fit to the autocorrelation data. The resulting best-fit damped sine wave provides a basis for assessing the strength of the periodicity of the HIP1 distribution. A low value of the goodness-of-fit, G, (see Supplementary Methods) signals a good fit to a periodic function. Similarly, the amplitude (A) is a measure of the strength of the periodicity, because decreased noise in the spacing histogram increases the magnitude of the autocorrelation. For example, the visibly poor fit of a damped sine wave to the hexamer palindrome closest in abundance to HIP1 results in a value of G that is substantially higher and amplitude that is noticeably lower than those associated with HIP1 (Supplementary Figure S10). The period of the best-fit damped sine wave provides an estimate of the period, λ_b, of the HIP1 distribution for the bin size used to construct the spacing histogram (b = 5 kb). For example, in Figure 7B, λ_b is ∼42 kb, which is similar to the period estimated in *Ca.* Atelocyanobacterium thalassa (Figure 6).

To assess how this estimate is influenced by the bin size used to aggregate data, we repeated the analysis for bin sizes ranging from 25 to 7000 bp in 25 bp intervals. We observe a substantial range of bin sizes for which the estimates are stable (see Supplementary Methods for a discussion of the factors that determine this range). For example, the grey boxes in Figure 8 encompass an interval of 60 consecutive bin sizes, representing bin sizes varying by up to 1.5 kb in length, in which the goodness of fit is consistently good (i.e. the lowest values of G; Figure 8A, black dots) and the amplitude of the sine wave is consistently high (Figure 8B, black dots).

The estimates of the period in this intermediate range are also largely in agreement (Figure 8C). All bin sizes greater than ∼2.5 kb result in an inferred period of ∼42 kb. The standard error of the inferred period, $SE(\lambda)$, is 0.05 kb over the boxed interval in Figure 8C. The stability of the period over many bin sizes provides a basis for obtaining an estimate that is independent of bin size and reflects the in-

trinsic period, if any, of HIP1 in the genome. The period was estimated for each of the 280 bin sizes. Each estimate was compared with all other estimates; two estimates are considered similar if they differ by at most 10%. Estimates of the period with similarity to the largest number of other estimates are considered most representative of the period of the HIP1 spacing distribution. These 'most representative' periods were averaged to obtain a global estimate of the HIP1 period, λ^*. For *syp*, the global HIP1 period is ~39.4 kb, which agrees with 185 of the 280 bin sizes.

While visually compelling, the apparent good fit of a damped sine wave to the autocorrelation data does not allow us to conclude that selection has maintained a periodic distribution of HIP1-like motifs. First, we must demonstrate that the fit of the sine function to genuine data is better than chance alone. Second, we must demonstrate that the observed HIP1 periodicity, if significant, does not reflect an intrinsic periodicity of mutational biases that would be evident in the spacings of all oligomers in the genome, not just HIP1. We address each of these points in turn.

Periodicity of HIP1 is statistically significant. The significance of the goodness-of-fit, $P(G)$, and of the amplitude, $P(A)$, were assessed for each of the 280 bin sizes by applying the autocorrelation analysis described above to 200 randomized genomes and comparing the resulting fits to the fit of the genuine data. For a bin size of 5 kb, the HIP1 spacings in the *syp* genome had a significantly better fit to a damped sine function than the spacings in randomized genomes ($P(G) = 0.045$). In fact, for a broad range of bin sizes, the fit of a sine function to the *syp* data is significantly better than expected by chance alone: 28% of bin sizes showed $P(G) \leq 0.02$, and 27 bin sizes showed $P(G) \leq 0.005$ (Figure 8A, grey dots).

If the HIP1 spacings are truly periodic, then good fits to a damped sine wave should be obtained over many different bin sizes; in addition, these fits should yield consistent estimates of the period. We assessed the probability of observing consistent periodicity over many bin sizes by chance by comparing $\bar{G}$, the mean goodness of fit over a range of bin sizes, with the value of $\bar{G}$ calculated over the same range in randomized genomes. For the boxed interval of 60 bin sizes in Figure 8A, this yields a P-value of $P(\bar{G}) = 0.015$. Analogous P-values for this range were calculated for the mean amplitude ($P(\bar{A})$ =0.045, Figure 8B) and the standard error of the period ($P(SE(\lambda)) = 0.045$, Figure 8C). Longer intervals will yield even more significant P-values. The joint probability of observing values of $\bar{G}$ and $SE(\lambda)$ that are both better in randomized data than in the genuine data for that interval is $P(\bar{G}, SE(\lambda)) < 0.005$, i.e. the minimum probability possible for 200 replicates.

To determine whether periodicity is a general feature of HIP1 distributions, we used this approach to assess HIP1 periodicity in the 50 HIP1-rich genomes (other than *cyu*) in our dataset (Supplementary Table S7). Estimates of the global HIP1 period were calculated for all genomes. The overall significance of the HIP1 periodicity was assessed at two levels of stringency. A total of 22 of the 50 genomes showed significant periodicity of HIP1-like motifs ($P^*_{G\lambda} \leq 0.02$, Figure 1, closed circles); only one genome in 50 is expected by chance at that level of significance. The periods estimated for these genomes ranged from 18 to 94 kb. If we further require that the amplitude also be as strong in the randomized genomes as in the genuine data, then an additional nine genomes are significant at $P^*_{GA\lambda} \leq 0.02$ (Supplementary Table S7) and 40 genomes, in total, are significant at $P^*_{GA\lambda} \leq 0.05$ (Figure 1, open circles). Therefore, we conclude that significant periodicity of HIP1 within cyanobacteria is widespread, and not a feature of only one or a few genomes. These genomes are taxonomically diverse, representing all major cyanobacterial lineages wherein HIP1 is enriched (Figure 1), and include genomes across the ranges of genome size and GC content in our study set.

Periodicity of HIP1 does not reflect periodicity in mutational biases. Before concluding that periodicity is a specific feature of HIP1, we must exclude the possibility that the spatial distribution we observe is simply due to periodicity in the underlying mutational biases of these genomes, as has been observed in some proteobacterial genomes (46,47). If the periodicity we detect reflects global mutational processes, then other oligomers within the same genome will also be periodic. In this case, $P^*_{G\lambda}$ reflects the significance of the overall periodicity of the genome, not just the HIP1 periodicity. Thus, in genomes with highly significant HIP1 periodicity (i.e. low $P^*_{G\lambda}$), other oligomers should also be significantly periodic, with periods similar to the HIP1 period. In genomes lacking significant HIP1 periodicity (i.e. high $P^*_{G\lambda}$), other oligomers within the same genome are not likely to be significantly periodic and the inferred periods of oligomers in that genome, whether significant or not, will not generally agree. Thus, if HIP1 periodicity reflects periodicity in mutational biases, we expect that the consistency of oligomer periods will decrease as HIP1 significance decreases (i.e. $P^*_{G\lambda}$ increases), as illustrated in the abstract model in Figure 9A and B. Similarly, the P-values of oligomers in the same genome will be correlated (Figure 9C). In contrast, if the periodicity we detect is a specific manifestation of HIP1 function, then the agreement, or lack thereof, between the inferred periods of oligomers in a given genome should not be correlated with the significance of HIP1 in that genome. In other words, oligomer periods should be no more correlated in a genome with a low value of $P^*_{G\lambda}$ than in a genome with a high value of $P^*_{G\lambda}$.

To test whether HIP1 periodicity is a manifestation of a global periodicity due to mutational biases, we identified five hexamer palindromes in each genome with abundances similar to the HIP1 abundance and inferred their global periods (Supplementary Table S8), as described for HIP1-like motifs. We used the CV to assess the consistency of periods across motifs (HIP1 and hexamer controls) within a genome. In this case, we expect greater consistency (lower CV) in genomes with more significant periodicity (lower values of $P^*_{G\lambda}$); genomes lacking significant periodicity will have high CV's, because the inferred periods for all oligomers in those genomes simply reflect noise. However, plotting CV as a function of $P^*_{G\lambda}$ shows no relationship between the two quantities (Figure 9D); genomes with highly significant periodicity ($P^*_{G\lambda} \leq 0.01$) and those without ($P^*_{G\lambda} \geq 0.1$) do not have significantly different CV's ($P = 0.98$, ANOVA). Moreover, we observe no correlation between the significance of the periodicity of HIP1 and that

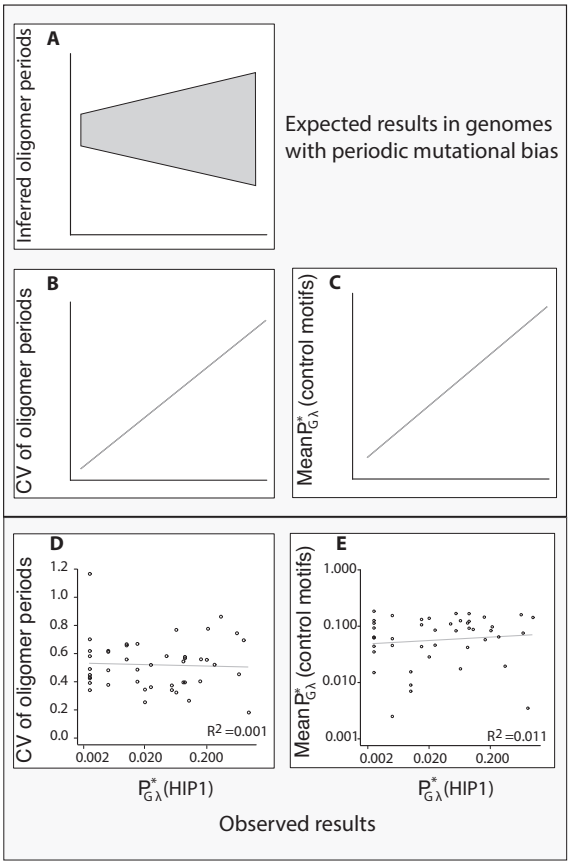

Figure 9. Variation of oligomer periods as a function of the significance of HIP1 periodicity. (**A**–**C**) Expected relationships under the null hypothesis that HIP1 periodicity results from an underlying periodicity in mutational bias. (A) Oligomer periods exhibit greater variation in genomes in which HIP1 periodicity is less significant. (B) The CV of the periods of oligomers in the same genome increases as the significance of the HIP1 periodicity decreases (i.e. the *P*-value increases). (C) The *P*-values of oligomers in the same genome are correlated. (**D** and **E**) Observed relationships: (D) For each genome, the CV among the inferred periods for six motifs—HIP1 and five hexamers—as a function of $P^*_{G\lambda}$, the significance of HIP1 periodicity. (Periods given in Supplementary Table S8). (E) For each genome, the average of $P^*_{G\lambda}$ for five control hexamers as a function of the significance of HIP1 periodicity.

of hexamer controls in the same genome (Figure 9E); the values of $P^*_{G\lambda}$ for hexamer controls are not significantly different in genomes with and without significant HIP1 periodicity ($P = 0.66$, ANOVA). Therefore, we conclude that oligomers in the same genome do not share a common period and the significant periodicity of HIP1 that we observe within cyanobacterial genomes is a specific feature of HIP1, and likely reflects its function.

DISCUSSION

We provide evidence that HIP1 is under selection, and therefore serves a functional role, within cyanobacterial genomes. HIP1 motifs are more conserved between genomes than similar octamer palindromes in both protein-coding and non-coding regions. HIP1 sites with a single mismatch are also conserved, although to a lesser extent. This conservation is not due to use of preferred codons or selection on amino acid motifs.

We found no evidence that HIP1 is associated with the regulation of genes or gene products via proximity to promoter sequences, but within many genomes we observed significant periodicity in their chromosomal distribution that is not a byproduct of periodicity in underlying mutational biases (46,47). Small sample sizes are an inherent difficulty in assessing kilobase scale periodicities in bacterial genomes, simply because of the relative lengths of the period and the genome. The discrete Fourier transform, an effective approach to quantifying periodic signals on much smaller scales, is sensitive to noise and edge effects with such small sample sizes (48). Here, we demonstrate an approach, based on fitting an autocorrelation function to a damped sine wave, that is robust in this regime.

HIP1 showed statistically significant periodicity in a broad range of genomes (Figure 1), including both large and small genomes, GC-rich and GC-poor genomes, and genomes from taxa representing varied cellular morphologies, habitats and metabolic capabilities spanning the phylogenetic diversity of our sample. For this reason, we believe that the periodicity of HIP1 is required for HIP1 function across the range of genomes wherein it is overabundant, enriched and under selection. This periodicity, on the scale of tens of kilobases, suggests that HIP1 may contribute to chromosome architecture, physiology or maintenance.

Functional implications of HIP1 periodicity

This is, to our knowledge, the first report of kilobase-scale periodicity of a repetitive motif, although periodic patterns on similar scales have been described for other genomic features. In *E. coli,* the distances between regulators and their targets were observed to be multiples of 92.8 kb (49). Wright *et al*. (50) identified a set of 22 500 'statistically correlated' gene pairs that tend to be co-located within the genome and to co-occur in the same genomes; the spatial distribution of these pairs in *E. coli* exhibits a periodicity of 117 kb. The same period is observed when full length transcriptional units are considered (51). Genes with extreme codon-bias exhibit a similar periodicity and a smaller one of 33 kb (52). Periodic patterns in expression levels have been reported with periods of roughly 100 kb and of 600–700 kb (53,54); this periodic behavior was disrupted in a gyrase mutant (54), suggesting a link between periodicity, gene expression and chromosomal interaction domains (CIDs). It has been posited that periodicity on this scale would place genes and their regulators in close 3D proximity in the cell (49), allowing for efficient regulation and expression of core genes (54). Indeed, superposition on a 3D interaction map obtained from genome conformation capture shows greater than chance proximity among genes that are co-regulated, associated with the same biological process or encode interacting proteins (55).

These periods are also on a scale commensurate with structural features of the nucleoid, defined in terms of topological domains in *E.coli* (∼10 kb (56)), torsional barriers in *E. coli* (40–90 kb (57)), and plectonemic barriers in *Salmonella typhimurium* (∼80 kb (58)). High-throughput chromosomal conformation capture (Hi-C) studies report

CIDs, consisting of multiple plectonemic loops separated by supercoiling diffusion barriers, with mean lengths of ∼120 kb in *Caulobacter crecentus* (59,60). CIDs on this scale have also been observed in *Bacillus subtilis* (61,62), *Vibrio cholerae* (63) and *E. coli* and *Pseudomonas aeriginosa* (cited in (59)).

The similarity in the scales of the observed HIP1 periods and chromosomal domain sizes is suggestive; HIP1 may create or regulate chromosomal architecture in cyanobacterial genomes, or may serve an alternative, as-yet-undescribed function. One plausible model is that HIP1 acts as a binding site for a protein that modulates chromosomal structure. The earliest HIP1 reports posited that HIP1 could be a protein binding site, but EMSA did not identify a protein that specifically binds HIP1 (15). However, failure to detect a potential HIP1-binding protein could be due to the limited sensitivity of EMSA, or the absence of either a functional HIP1-binding protein or a suitable HIP1-bearing substrate (e.g. large plectonemic complexes) under the experimental conditions used.

The length of the periods we observed ranged from 18 to 94 kb in different genomes. Since the role of the periodicity is not clear, it is difficult to interpret this variation. However, it is known that plectonemic domain size does vary with growth rate (58) and nutrient conditions (57) in heterotrophs. Therefore, this range of periodicities may reflect different growth regimes of these cyanobacterial autotrophs.

Interestingly, γδ resolution assays reveal that highly transcribed regions act as supercoiling diffusion barriers between plectonemic regions in *S. typhimurium* (64,65); this has also been observed with Hi-C in *C. crecentus* (59,60) and to some extent in *B. subtilis* (61). The observed paucity of HIP1 motifs in highly transcribed regions (i.e. in potential diffusion barriers), beyond that predicted by codon usage bias alone, is consistent with a potential role for HIP1 in organizing chromosomal domains.

Very little is known about the mechanisms that organize chromosomal domains in cyanobacteria. In *E. coli* and other proteobacteria, chromosomal structure is shaped by so-called NAPs, abundant, low molecular weight proteins that bridge and bend DNA (24). However, with the exception of HU and IHF, NAP homologs are absent from cyanobacterial genomes (25–27). Thus, chromosomal architecture in cyanobacteria is likely maintained by lineage-specific mechanisms that have yet to be discovered. The relevant cyanobacterial proteins, and their associated binding sites, may be quite different from their proteobacterial counterparts. If HIP1 serves to organize chromosomal domains, our results predict that high-resolution mapping of cyanobacterial chromosome organization with Hi-C will reveal regular physical association of chromosomal loci with upstream and downstream regions separated by a fixed distance; the scale of this distance should be commensurate with the HIP1 period associated with that genome.

Evolution of HIP1-like motifs

While HIP1 motifs are conserved by selection, the precise target of selection is not clear. Not only is the canonical HIP1 motif both enriched and conserved by selection, but variant sequences are also conserved and/or enriched. First, we have demonstrated that HIP1* sequences—those with a single-base mismatch—are also conserved. If this shared conservation reflects shared selection, then the function provided by HIP1 motifs allows some flexibility in their sequence identity. This, in turn, could allow for drift in sequence identity, resulting in enrichment of HIP1 variants among genomes. For example, two octamers, HIP1 and TCGATCGA, are comparably overrepresented in *Dactylococcopsis salina* PCC 8305 (dsl). A similar co-occurrence has been reported in *Geminocystis herdmanii* PCC 6308 (23), a genome not included in our dataset. One step further, the canonical HIP1 is neither abundant nor over-represented in *Synechococcus* sp. strains JA-3–3Ab (cya) and JA-2–3B'a (cyb); instead, a similar motif, GGGATCCC, is both enriched and conserved, suggesting that it has adopted HIP1 function. A different octamer palindrome (CAGGCCTG) is overrepresented in *Synechococcus* sp. PCC 6312 (syne), another genome in which the canonical HIP1 motif is not over-represented. The conservation of this motif was not assessed because a genome suitable for comparison has not been sequenced. However, it is tempting to speculate that in this genome, this HIP1-like palindrome has a functional role analogous to that of HIP1. Taken together, these data are consistent with HIP1 enrichment and conservation representing a balance of mutation and selection.

In this context, it is not surprising that evidence for periodicity of HIP1-like motifs is not statistically significant, using our methods, in a number of genomes wherein HIP1 is enriched. We assessed periodicity using the spacing between motifs drawn from the combined set of HIP1 and HIP1* sequences. This is beneficial in most cases, because HIP1* sites as a whole are under selection (albeit weaker than HIP1) and their inclusion increases the size of the dataset. In many genomes, the resulting increase in statistical power allows the detection of a periodic signal when none is evident using the smaller dataset of HIP1 sites alone. However, selection likely does not act on all HIP1* sites; as a result, the datasets will contain varying numbers of non-selected sites which will confound our ability to extract a periodic signal. Indeed, while the periodicity of pooled HIP1 and HIP1* sites was not significant in *Cyanobacterium aponinum* PCC 10605 ($P^*_{G\lambda} = 0.09$; can), and *Stanieria cyanosphaera* PCC 7437 ($P^*_{G\lambda} = 0.085$; scs), significant periodic signals were detected in these genomes when HIP1 motifs alone were considered. In addition, periodic distributions may be disrupted by recent insertions, deletions or rearrangements of chromosomal regions.

While the conservation of HIP1 may explain their maintenance within cyanobacterial genomes, it does not explain their origin. Elhai (23) has proposed that new HIP1 motifs could be generated by a unidirectional mutational ratchet associated with methyl-directed mismatch repair. It is currently unclear whether the methylation machinery required for this model is present in cyanobacterial genomes, but this type of mechanism could, in principle, generate new HIP1-like motifs that are potential targets of selection. Regardless of their mechanism of origin, it is clear both that HIP1 motifs experience selection for their retention and that they remain periodically distributed within many cyanobacterial genomes. These results provide a framework that ties HIP1

to machinery associated with other periodic signatures in bacterial genomes and suggest that HIP1 offers a promising direction for future investigations of chromosomal architecture in cyanobacteria.

DATA AND SOFTWARE AVAILABILITY

The coordinates of HIP1-like and hexamer control motifs in randomized genomes, as well as the parameters resulting from fitting a damped sine wave to both genuine and randomized data, are available via the Dryad Digital Repository (https://doi.org/10.5061/dryad.b301d). The coordinates, abundance and ACE_u scores for the transcript data used in this study are also available from this repository.

Scripts for estimating the expected number of motif instances from a genome sequence, for calculating the conservation score, *S,* from a Mauve alignment and for assessing periodicity by fitting inter-motif spacings to a damped sine wave are freely available at https://github.com/minli-xu/HIP1.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Annette McLeod and Dr. Maureen Stolzer for extensive help with figure preparation.

FUNDING

National Science Foundation [DBI1262593 to D.D.]; National Institute of General Medical Sciences [GM116884 to J.L.]. Funding for open access charge: Institutional support; Unrestricted funds stemming from partial recovery of indirect costs from the university administration.
Conflict of interest statement. None declared.

REFERENCES

1. Treangen,T.J., Abraham,A.L., Touchon,M. and Rocha,E.P. (2009) Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol. Rev.*, **33**, 539–571.
2. Delihas,N. (2011) Impact of small repeat sequences on bacterial genome evolution. *Genome Biol. Evol.*, **3**, 959–973.
3. Bzymek,M. and Lovett,S.T. (2001) Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 8319–8325.
4. Smith,G.R. (2012) How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. *Microbiol. Mol. Biol. Rev.*, **76**, 217–228.
5. Hendrickson,H. and Lawrence,J.G. (2006) Selection for chromosome architecture in bacteria. *J. Mol. Evol.*, **62**, 615–629.
6. Smith,H.O., Gwinn,M.L. and Salzberg,S.L. (1999) DNA uptake signal sequences in naturally transformable bacteria. *Res. Microbiol.*, **150**, 603–616.
7. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.
8. Mrazek,J., Gaynon,L.H. and Karlin,S. (2002) Frequent oligonucleotide motifs in genomes of three streptococci. *Nucleic Acids Res.*, **30**, 4216–4221.
9. Bochman,M.L., Paeschke,K. and Zakian,V.A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.*, **13**, 770–780.
10. De Gregorio,E., Silvestro,G., Petrillo,M., Carlomagno,M.S. and Di Nocera,P.P. (2005) Enterobacterial repetitive intergenic consensus sequence repeats in yersiniae: genomic organization and functional properties. *J. Bacteriol.*, **187**, 7945–7954.
11. Higgins,C.F., Ames,G.F., Barnes,W.M., Clement,J.M. and Hofnung,M. (1982) A novel intercistronic regulatory element of prokaryotic operons. *Nature*, **298**, 760–762.
12. Stern,M.J., Ames,G.F., Smith,N.H., Robinson,E.C. and Higgins,C.F. (1984) Repetitive extragenic palindromic sequences: a major component of the bacterial genome. *Cell*, **37**, 1015–1026.
13. Delaye,L. and Moya,A. (2011) Abundance and distribution of the highly iterated palindrome 1 (HIP1) among prokaryotes. *Mobile Genet. Elem.*, **1**, 159–168.
14. Robinson,N.J., Robinson,P.J., Gupta,A., Bleasby,A.J., Whitton,B.A. and Morby,A.P. (1995) Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria. *Nucleic Acids Res.*, **23**, 729–735.
15. Robinson,P.J., Cranenburgh,R.M., Head,I.M. and Robinson,N.J. (1997) HIP1 propagates in cyanobacterial DNA via nucleotide substitutions but promotes excision at similar frequencies in *Escherichia coli* and *Synechococcus* PCC 7942. *Mol. Microbiol.*, **24**, 181–189.
16. Vijayan,V., Jain,I.H. and O'Shea,E.K. (2011) A high resolution map of a cyanobacterial transcriptome. *Genome Biol.*, **12**, R47.
17. Neilan,B.A., Saker,M.L., Fastner,J., Torokne,A. and Burns,B.P. (2003) Phylogeography of the invasive cyanobacterium *Cylindrospermopsis raciborskii*. *Mol. Ecol.*, **12**, 133–140.
18. Orcutt,K.M., Rasmussen,U., Webb,E.A., Waterbury,J.B., Gundersen,K. and Bergman,B. (2002) Characterization of *Trichodesmium spp.* by genetic techniques. *Appl. Environ. Microbiol.*, **68**, 2236–2245.
19. Akiyama,H., Kanai,S., Hirano,M. and Miyasaka,H. (1998) A novel plasmid recombination mechanism of the marine cyanobacterium *Synechococcus sp.* PCC7002. *DNA Res.*, **5**, 327–334.
20. Elhai,J., Kato,M., Cousins,S., Lindblad,P. and Costa,J.L. (2008) Very small mobile repeated elements in cyanobacterial genomes. *Genome Res.*, **18**, 1484–1499.
21. Krishna,P.S., Rani,B.R., Mohan,M.K., Suzuki,I., Shivaji,S. and Prakash,J.S. (2013) A novel transcriptional regulator, Sll1130, negatively regulates heat-responsive genes in *Synechocystis sp.* PCC6803. *Biochem. J.*, **449**, 751–760.
22. Cheregi,O. and Funk,C. (2015) Regulation of the *scp* genes in the cyanobacterium *Synechocystis sp.* PCC 6803–What is new? *Molecules*, **20**, 14621–14637.
23. Elhai,J. (2015) Highly Iterated Palindromic sequences (HIPs) and their relationship to DNA methyltransferases. *Life*, **5**, 921–948.
24. Dorman,C.J. (2013) Genome architecture and global gene regulation in bacteria: making progress towards a unified model? *Nat. Rev. Microbiol.*, **11**, 349–355.
25. Badrinarayanan,A., Le,T.B. and Laub,M.T. (2015) Bacterial chromosome organization and segregation. *Annu. Rev. Cell. Dev. Biol.*, **31**, 171–199.
26. Browning,D.F., Grainger,D.C. and Busby,S.J. (2010) Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. *Curr. Opin. Microbiol.*, **13**, 773–780.
27. Landick,R., Wade,J.T. and Grainger,D.C. (2015) H-NS and RNA polymerase: a love-hate relationship? *Curr. Opin. Microbiol.*, **24**, 53–59.
28. Guindon,S., Dufayard,J.F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.
29. Milne,I., Lindner,D., Bayer,M., Husmeier,D., McGuire,G., Marshall,D.F. and Wright,F. (2009) TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics*, **25**, 126–127.
30. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
31. Keane,T.M., Creevey,C.J., Pentony,M.M., Naughton,T.J. and Mclnerney,J.O. (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc

assumptions for choice of matrix are not justified. *BMC Evol. Biol.*, **6**, 29.
32. Schirrmeister,B.E., Antonelli,A. and Bagheri,H.C. (2011) The origin of multicellularity in cyanobacteria. *BMC Evol. Biol.*, **11**, 45.
33. Latysheva,N., Junker,V.L., Palmer,W.J., Codd,G.A. and Barker,D. (2012) The evolution of nitrogen fixation in cyanobacteria. *Bioinformatics*, **28**, 603–606.
34. Larsson,J., Nylander,J.A. and Bergman,B. (2011) Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC Evol. Biol.*, **11**, 187.
35. Shih,P.M., Wu,D., Latifi,A., Axen,S.D., Fewer,D.P., Talla,E., Calteau,A., Cai,F., Tandeau de Marsac,N., Rippka,R. *et al.* (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 1053–1058.
36. Letunic,I. and Bork,P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.*, **44**, W242–W245.
37. Karlin,S. and Brendel,V. (1992) Chance and statistical significance in protein and DNA sequence analysis. *Science*, **257**, 39–49.
38. Li,W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.*, **36**, 96–99.
39. Darling,A.E., Tritt,A., Eisen,J.A. and Facciotti,M.T. (2011) Mauve assembly metrics. *Bioinformatics*, **27**, 2756–2757.
40. Retchless,A.C. and Lawrence,J.G. (2011) Quantification of codon selection for comparative bacterial genomics. *BMC Genomics*, **12**, 374.
41. Rippka,R., Deruelles,J., Waterbury,J.B., Herdman,M. and Stanier,R.Y. (1979) Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J. Gen. Microbiol.*, **111**, 1–61.
42. Steunou,A.S., Bhaya,D., Bateson,M.M., Melendrez,M.C., Ward,D.M., Brecht,E., Peters,J.W., Kuhl,M. and Grossman,A.R. (2006) *Insitu* analysis of nitrogen fixation and metabolic switching in unicellular thermophilic cyanobacteria inhabiting hot spring microbial mats. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 2398–2403.
43. Walsby,A.E., Van Rijn,J. and Cohen,Y. (1983) The biology of a new gas-vacuolate cyanobacterium, *Dactylococcopsis salina* sp.nov., in Solar Lake. *Proc. R. Soc. Lond. B.*, **217**, 417–447.
44. Xu,M. (2015) *Comparative genomics reveals forces driving the evolution of Highly Iterated Palindrome-1 (HIP1) in cyanobacteria*. Ph.D. Thesis, Computational Biology Department, Carnegie Mellon University, Pittsburgh.
45. Gama-Castro,S., Salgado,H., Santos-Zavaleta,A., Ledezma-Tejeida,D., Muniz-Rascado,L., Garcia-Sotelo,J.S., Alquicira-Hernandez,K., Martinez-Flores,I., Pannier,L., Castro-Mondragon,J.A. *et al.* (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.*, **44**, D133–D143.
46. Foster,P.L., Hanson,A.J., Lee,H., Popodi,E.M. and Tang,H. (2013) On the mutational topology of the bacterial genome. *G3*, **3**, 399–407.
47. Dillon,M.M., Sung,W., Sebra,R., Lynch,M. and Cooper,V.S. (2017) Genome-wide biases in the rate and molecular spectrum of spontaneous mutations in *Vibrio cholerae* and *Vibrio fischeri*. *Mol. Biol. Evol.*, **34**, 93–109.
48. Junier,I., Herisson,J. and Kepes,F. (2010) Periodic pattern detection in sparse boolean sequences. *Algorithms Mol. Biol.*, **5**, 31.
49. Kepes,F. (2004) Periodic transcriptional organization of the *E.coli* genome. *J. Mol. Biol.*, **340**, 957–964.
50. Wright,M.A., Kharchenko,P., Church,G.M. and Segre,D. (2007) Chromosomal periodicity of evolutionarily conserved gene pairs. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 10559–10564.
51. Junier,I., Herisson,J. and Kepes,F. (2012) Genomic organization of evolutionarily correlated genes in bacteria: limits and strategies. *J. Mol. Biol.*, **419**, 369–386.
52. Mathelier,A. and Carbone,A. (2010) Chromosomal periodicity and positional networks of genes in *Escherichia coli*. *Mol. Syst. Biol.*, **6**, 366.
53. Allen,T.E., Herrgard,M.J., Liu,M., Qiu,Y., Glasner,J.D., Blattner,F.R. and Palsson,B.O. (2003) Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. *J. Bacteriol.*, **185**, 6392–6399.
54. Jeong,K.S., Ahn,J. and Khodursky,A.B. (2004) Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biol.*, **5**, R86.
55. Xie,T., Fu,L.Y., Yang,Q.Y., Xiong,H., Xu,H., Ma,B.G. and Zhang,H.Y. (2015) Spatial features for *Escherichia coli* genome organization. *BMC Genomics*, **16**, 37.
56. Postow,L., Hardy,C.D., Arsuaga,J. and Cozzarelli,N.R. (2004) Topological domain structure of the *Escherichia coli* chromosome. *Genes Dev.*, **18**, 1766–1779.
57. Sinden,R.R. and Pettijohn,D.E. (1981) Chromosomes in living *Escherichia coli* cells are segregated into domains of supercoiling. *Proc. Natl. Acad. Sci. U.S.A.*, **78**, 224–228.
58. Staczek,P. and Higgins,N.P. (1998) Gyrase and Topo IV modulate chromosome domain size *in vivo*. *Mol. Microbiol.*, **29**, 1435–1448.
59. Le,T.B. and Laub,M.T. (2016) Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries. *EMBO J.*, **35**, 1582–1595.
60. Le,T.B., Imakaev,M.V., Mirny,L.A. and Laub,M.T. (2013) High-resolution mapping of the spatial organization of a bacterial chromosome. *Science*, **342**, 731–734.
61. Marbouty,M., Le Gall,A., Cattoni,D.I., Cournac,A., Koh,A., Fiche,J.B., Mozziconacci,J., Murray,H., Koszul,R. and Nollmann,M. (2015) Condensin- and replication-mediated bacterial chromosome folding and origin condensation revealed by Hi-C and super-resolution imaging. *Mol. Cell*, **59**, 588–602.
62. Wang,X., Le,T.B., Lajoie,B.R., Dekker,J., Laub,M.T. and Rudner,D.Z. (2015) Condensin promotes the juxtaposition of DNA flanking its loading site in *Bacillus subtilis*. *Genes Dev.*, **29**, 1661–1675.
63. Val,M.E., Marbouty,M., de Lemos Martins,F., Kennedy,S.P., Kemble,H., Bland,M.J., Possoz,C., Koszul,R., Skovgaard,O. and Mazel,D. (2016) A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*. *Sci. Adv.*, **2**, e1501914.
64. Booker,B.M., Deng,S. and Higgins,N.P. (2010) DNA topology of highly transcribed operons in *Salmonella enterica* serovar Typhimurium. *Mol. Microbiol.*, **78**, 1348–1364.
65. Deng,S., Stein,R.A. and Higgins,N.P. (2004) Transcription-induced barriers to supercoil diffusion in the *Salmonella typhimurium* chromosome. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 3398–3403.