environmental microbiology

Environmental Microbiology (2017) **19**(11), 4700–4713 doi:10.1111/1462-2920.13934

Biogeographic conservation of the cytosine epigenome in the globally important marine, nitrogen-fixing cyanobacterium *Trichodesmium*

Nathan G. Walworth, David A. Hutchins, Egor Dolzhenko, Michael D. Lee, Feixue Fu, Andrew D. Smith and Eric A. Webb*
Department of Biological Sciences, University of Southern California, Los Angeles, CA, 90089, USA.

Summary

Cytosine methylation has been shown to regulate essential cellular processes and impact biological adaptation. Despite its evolutionary importance, only a handful of bacterial, genome-wide cytosine studies have been conducted, with none for marine bacteria. Here, we examine the genome-wide, C^5-Methyl-cytosine (m5C) methylome and its correlation to global transcription in the marine nitrogen-fixing cyanobacterium *Trichodesmium*. We characterize genome-wide methylation and highlight conserved motifs across three *Trichodesmium* isolates and two *Trichodesmium* metagenomes, thereby identifying highly conserved, novel genomic signatures of potential gene regulation in *Trichodesmium*. Certain gene bodies with the highest methylation levels correlate with lower expression levels. Several methylated motifs were highly conserved across spatiotemporally separated *Trichodesmium* isolates, thereby elucidating biogeographically conserved methylation potential. These motifs were also highly conserved in *Trichodesmium* metagenomic samples from natural populations suggesting them to be potential *in situ* markers of m5C methylation. Using these data, we highlight predicted roles of cytosine methylation in global cellular metabolism providing evidence for a 'core' m5C methylome spanning different ocean regions. These results provide important insights into the m5C methylation landscape and its biogeochemical implications in an important marine N_2-fixer, as well as advancing evolutionary theory examining methylation influences on adaptation.

Received 4 April, 2017; revised 7 August, 2017; accepted 30 August, 2017. *For correspondence. E-mail eawebb@usc.edu; Tel. (213) 740-7954; Fax (213) 740-8123.

Introduction

DNA methylation is a type of epigenetic modification that has been shown to regulate key physiological processes in the cell including gene expression, imprinting, cell differentiation and gene silencing (Krueger *et al.*, 2012). Theoretical and empirical studies have also demonstrated it to be important in environmental adaptation via transgenerational epigenetic inheritance (Jablonka and Raz, 2009), thereby serving as a mechanism to generate phenotypic diversity in the absence of genetic mutation (Schmitz *et al.*, 2011; Geoghegan and Spencer, 2012; Kronholm and Collins, 2016). Thus, adaptive phenotypes have the potential to arise prior to genetic changes, which may then be fixed upon adaptive mutation through a process called genetic assimilation (Klironomos *et al.*, 2013; Ehrenreich and Pfennig, 2015; Kronholm and Collins, 2016). Therefore, epigenetic variation has the potential to affect rates of adaptive fitness increases.

DNA methylation of adenine or cytosine in bacteria has been primarily investigated as part of restriction-modification (R-M) systems that protect against phages and other foreign DNA (Loenen *et al.*, 2014), although recent observations suggest alternative roles for R-M systems in regulating global gene expression (Vasu and Nagaraja, 2013; Doberenz *et al.*, 2017). Accordingly, *Trichodesmium* contains syntenic homologs for 3 genes that are required for Type I restriction systems namely *hsdR* for restriction (Tery_2422), *hsdM* for methylation (Tery_2418) and *hsdS* for sequence specificity (Tery_2421) as described in *Escherichia coli* (Roer *et al.*, 2015). The best-studied methyltransferases in bacteria are DNA adenine methyltransferases (Dam and CcrM homologs in Gammaproteobacteria and Alpha-proteobacteria respectively) that either target the GA^mTC motif in Gammaproteobacteria or the GA^mNTC in Alphaproteobacteria (Kahramanoglou *et al.*, 2012; Sánchez-Romero *et al.*, 2015) and thereby influence transcriptional regulation (Waldron *et al.*, 2002), replication (Campbell and Kleckner, 1990), cell cycle (Reisenauer *et al.*, 1999), virulence(Heithoff *et al.*, 1999) and DNA mismatch repair (Glickman and Radman, 1980). *Trichodesmium* indeed harbours a single-copy homolog to the *E. coli dam* gene (Tery_3905) although DNA adenine methylation has not been studied in *Trichodesmium*.

Cytosine methylation by a second type of DNA methyltransferase (Dcm) involves the addition of a methyl group to either the C4 or C5 carbon residue (i.e., m4C and m5C respectively) (Blow *et al*., 2016) and has been shown to regulate membrane-, stress- and stationary phase-related gene expression in *Escherichia coli* (Kahramanoglou *et al*., 2012; Sánchez-Romero *et al*., 2015) and motility, adhesion and virulence in *Helicobacter pylori* (Kumar *et al*., 2012). Additionally, the *E. coli* m5C Dcm is homologous to the methyltransferase of the plasmid-encoded EcoRII R-M system and both have been shown to exclusively target the CCWGG motif (Boyer *et al*., 1973; Kahramanoglou *et al*., 2012). Thus, a functional Dcm confers resistance against parasitism because EcoRII-target sites will be kept methylated even when the R-M system is perturbed. DNA methyltransferases also exist as solitary (i.e., orphan) enzymes that methylate DNA sequences at specific sites but are not associated with restriction enzymes (Palmer and Marinus, 1994). However, the mechanisms by which prokaryotic cytosine methylation affect gene expression are still unknown, as the activities of transcriptional regulatory machinery may depend on the methylation state of the targeted DNA substrate and/or a change in the mechanical properties of methylated DNA deriving from methylated cytosines (Severin *et al*., 2011). Furthermore, the handful of studies investigating genome-wide, prokaryotic m5C cytosine methylation unassociated with R-M systems has been either confined to a few free-living and pathogenic species or limited in scope due to the type sequencing. For example, one recent study using single molecule real-time sequencing (SMRT) examined both adenine (m6A) and cytosine (m4C and m5C) methylation in over 200 prokaryotes (Blow *et al*., 2016). However, due to only modest sensitivity in detecting m5C methylation using SMRT sequencing, comprehensive analysis was not possible. Furthermore, virtually no studies have investigated m5C methylation in biogeochemically important marine microbes and their potential impacts to keystone metabolisms such as nitrogen (N_2) fixation.

Hence, characterization of epigenetic regulation in globally distributed marine phytoplankton including N_2-fixing cyanobacteria (diazotrophs) provides a tractable route to study methylation patterns under the context of global biogeochemical cycles, thereby defining a role for methylation in ecological fitness and adaptation. Marine microbial systems offer a variety of useful properties for these studies due to their worldwide distributions, large diversity, relatively fast generation times, commercial uses and significance in global carbon and nutrient cycles. For example, dinitrogen (N_2) fixation by the cyanobacterial genus *Trichodesmium* makes a globally important contribution of bioavailable nitrogen to ocean food webs with some estimates suggesting it to carry out as much as half of the total N_2 fixation in the vast, subtropical gyre biomes (Sohm *et al*., 2011; Hutchins *et al*., 2015). These inputs of new N to surface waters exert control over N-limited primary production, which in turn influences carbon export to the deep ocean where CO_2 can be sequestered from the atmosphere from hundreds to thousands of years (i.e., the biological pump) (Sohm *et al*., 2011). Hence, the ecological and biogeochemical roles of phytoplankton such as *Trichodesmium* necessitate a deeper understanding of the mechanisms they use to respond and/or adapt to environmental change (Collins *et al*., 2013).

Here, we characterize the genome-wide, C^5-Methylcytosine methylome of *Trichodesmium erythraeum* IMS101 (hereafter IMS101) and its biogeographic conservation across spatiotemporally separated isolates. One other study previously identified a relatively high degree of adenine methylation in *Trichodesmium* (Zehr *et al*., 1991). However, adenine methylation cannot be evaluated using the bisulfite sequencing method used in this study. To the best of our knowledge, this is the first m5C characterization for any marine diazotroph; A couple other eukaryotic m5C papers came out during this review process. We characterize both global methylation and specific motifs, thereby identifying potential methylation-specific genomic signatures of gene regulation in IMS101. Transcriptional analysis of IMS101 demonstrates that the genes with the highest methylation levels exhibit lower expression. Additionally, we identify highly conserved motifs in samples from natural populations and other isolates. Thus, we highlight both potential roles of cytosine methylation associated with biogeochemically critical metabolic pathways in *Trichodesmium* and show biogeographic conservation of this methylation in *Trichodesmium* obtained from different biogeochemical regimes.

Results and discussion

Methyltransferases in Trichodesmium

Both genomic analysis and IMG annotation (https://img.jgi.doe.gov) indicate that IMS101 harbours five m5C cytosine methyltransferase genes (*dcm*) that each reside within separate phylogenetic clusters upon maximum phylogenetic likelihood sequence analysis with other microbial *dcm* homologs (Methods; Fig. 1). Upon searching each IMS101 Dcm protein sequence against NCBI Refseq (Tatusova *et al*., 2015) using BLASTP (Altschul *et al*., 1990), all top high-scoring pairs to IMS101 Dcm homologs are all annotated m5C Dcm bacterial protein sequences (Supporting Information File S1). Furthermore, each Dcm harbours a m5C-specific Dcm pfam00145 domain (e-value $\leq 10^{-60}$) suggesting them to be m5C methyltransferases. Although *E. coli* genomes typically only harbour one m5C *dcm* (e.g., WP_001157239.1 in *E. coli* K12 MG1655; Kahramanoglou *et al*., 2012), bacteria across a broad phylogenetic range harbour multiple phylogenetically distinct m5C *dcm* copies

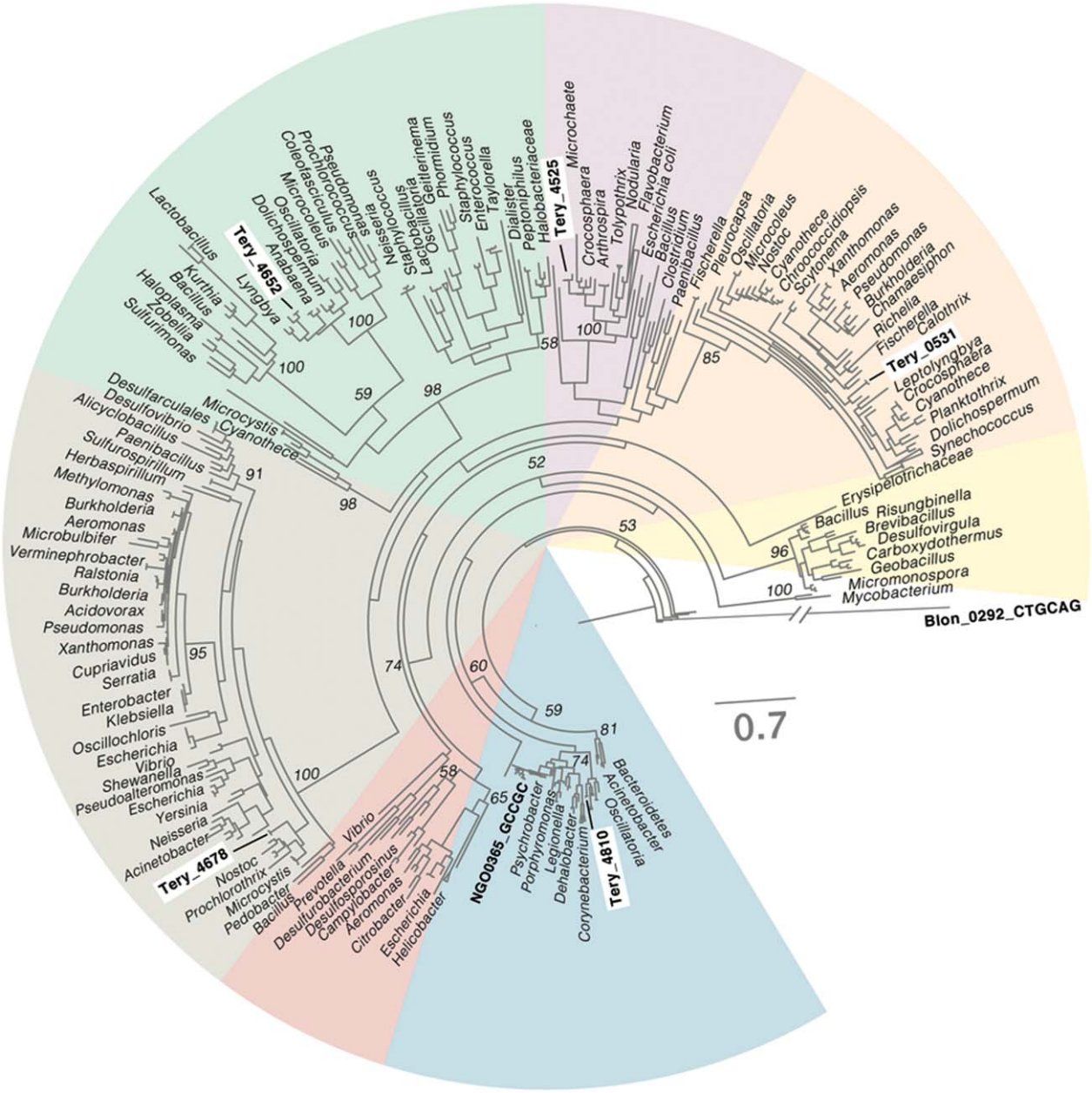

Fig. 1. Maximum likelihood phylogeny of DNA methyltransferases.
Shown is a maximum likelihood analysis of the phylogenetic distribution of IMS101 and 2175 Dcm protein sequences among other bacterial Dcm homologs. Representative bacterial genera residing within phylogenetic clusters are labelled. *Trichodesmium* proteins are noted within white boxes and represent both IMS101 and 2175 homologs as they always tightly clustered together. Bootstrap values >50 at major branch points are denoted. Different colours denote different clusters.

within a single genome, which can associate with different methylated motifs (Blow *et al.*, 2016). Blow and colleagues (2016) used single molecular real time (SMRT) sequencing, which measures the rate of base incorporation on a single DNA molecule by DNA polymerase via fluorescence pulses. The interpulse duration (IDP) (i.e., speed of incorporation) yields certain kinetic signatures that can distinguish between m6A, m4C and m5C modifications, but due to the small size of the m5C methyl group and its positioning in the major groove, only subtle changes in polymerase kinetics are produced thereby yielding inconsistent m5C detection (Clark *et al.*, 2013; Huo *et al.*, 2015). Alternatively, we used bisulfite sequencing which uses sodium bisulfite to chemically deaminate unmethylated m5C cytosine residues to uracil while leaving mostly all methylated cytosines as cytosines. This method thus

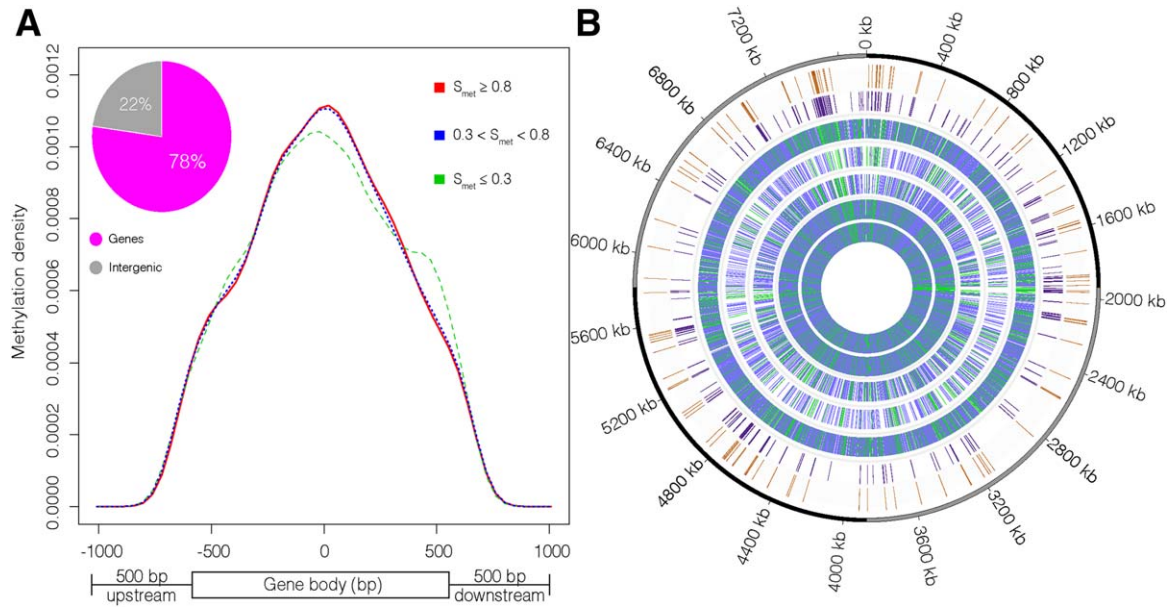

Fig. 2. Gene body methylation profiles and distribution of methylated cytosines between genic/intergenic regions.
A. Inset is a pie chart showing the distribution of methylated cytosines in genic and intergenic regions. Larger image is a histogram of m5C methylation density along gene bodies and 500 bp intergenic flanking regions. Green, blue and red colours represent lowly, partially and highly methylated residues. S_{met} values (# of reads with a methylated cytosine/total # of reads mapping to that cytosine) denote methylation levels.
B. Genome-wide map showing all methylated (m5C) cytosines (MCs) in each of the conserved sequence contexts (see below). In tracks with both green and blue colours: green = intergenic regions and blue = genes. Concentric rings from in to out: (1) All MCs on the minus strand (2) All MCs on the positive strand (3) All MCs in the CCG (GC^mCGC) context (4) All MCs in the CWG ($CTGC^mAG$) context (5) All MCs in the CG (CpG) context (6) All transposase sequences elements in genome [repetitive elements with transposase homology from Walworth and colleagues (2015)] (7) All methylated transposase sequences.

enables reliable detection of m5C methylation when compared with a reference genome (Clark *et al*., 2013; Huo *et al*., 2015). Hence, although Blow and colleagues (2016) were unable to comprehensively address m5C methylation, two m5C *dcm* genes were predicted to associate with two consistently methylated motifs also identified in this study (see below).

Genome-wide cytosine methylation of IMS101

Characterizing the methylation landscape of IMS101 was facilitated using Illumina bisulfite sequencing (Walworth *et al*., 2015) in which general bisulfite-conversion and sequence statistics can be found in Supporting Information Table S1. Cultures of IMS101 grown in a modified Aquil medium (Hutchins *et al*., 2015) were filtered and flash frozen in liquid nitrogen in biological triplicate during the middle of the 12 h photoperiod. A total of 15 652 methylated residues were detected (~1% of total cytosines) with ~78% and ~22% residing in total gene and intergenic regions respectively (Supporting Information File S2). Notably, ~60% of the *Trichodesmium* genome codes for protein, which is rare for the broad majority of free-living prokaryotes that typically average ~80% (Walworth *et al*., 2015). Upon plotting the locations of methylated cytosines (MCs) on both forward and reverse strands, MCs within genic and intergenic exhibit broad distributions throughout the genome (Fig. 2B). Interestingly, a handful of MCs in intergenic regions form clusters in and around repetitive elements with sequence homology to transposases. Many of these repetitive elements with MCs have been annotated as intergenic regions yet many retain transposase sequence homology suggesting them to be ancient pseudogenized transposases (Walworth *et al*., 2015). Hence, MCs overlapping with these sequences may potentially regulate their propagation as seen in other systems (Miura *et al*., 2001; Kato *et al*., 2003; Lister *et al*., 2008; see below for further discussion).

Similar to what has been observed in *E. coli* K12 (Kahramanoglou *et al*., 2012), the genome-wide, site-specific methylation level (# of reads with a methylated cytosine/total # of reads mapping to that cytosine) average (S_{met}) was ~0.74 (median = 0.84). Upon binning methylated residues into lowly ($S_{met} \leq 0.3$), partially ($0.3 < S_{met} < 0.8$) and highly ($S_{met} \geq 0.8$) methylated sites, we observed a distribution of ~8%, ~37% and ~55% of total methylated residues respectively. Methylation was detected in ~54% (2751/5076) of annotated genes (Supporting Information File S3), but of the promoter regions ($n = 1543$) with experimentally confirmed transcriptional start sites in IMS101 (Pfreundt *et al*., 2014), only ~5% were observed to be

methylated. The detected upstream promoter regions had an average length of 75 bp, which is similar to the 100 bp length typically used to estimate promoter region overlap (Blow *et al.*, 2016). Furthermore, methylated residues were not found to be enriched among hundreds of previously analysed prokaryotic epigenomes, which include m6A, m4C and m5C methylation (Kahramanoglou *et al.*, 2012; Blow *et al.*, 2016). The majority of methylated residues were observed to reside within gene bodies (defined as the genic region between the start and stop codon) and tended to be located towards the centre of genes (Fig. 2A). These data demonstrate gene body methylation (Veluchamy *et al.*, 2013) to be the primary mode of m5C methylation in *Trichodesmium*. Of the 2751 genes containing $\geq$1 methylated cytosine, $\sim$92% harbour $<$10 methylated residues yielding an average methylation level (# of methylated cytosines/# bp per gene) of 0.03 (median = 0.02) per methylated gene (Supporting Information Fig. S1 A,B). Hence, extensive methylation across the length of genes was not observed and methylation levels (R_{met}) (i.e., methylation density) showed no significant increase with increasing cytosine content or gene length (Supporting Information Fig. S1 C,D).

m5C sequence contexts and functionality

Methylated cytosines were identified in the sequence contexts of CCG, CHH, CWG and CG (where 'H' denotes adenines, thymines or cytosines and 'W' denotes adenines or thymines). Most of the observed methylation was detected in the CG (i.e., CpG) context (75%) followed by CWG (12.7%), CCG (12%) and CHH (0.3%) (Supporting Information Fig. S2). Downstream, genome-wide analysis of the CHH motif was omitted due to its underrepresentation in the IMS101 genome. Upon aligning $\pm$5 base pairs (bp) flanking the methylated cytosine in each of the remaining three sequence contexts, highly conserved five (GC^mCGC) and six ($CTGC^mAG$) bp motifs were observed for CCG and CWG respectively, while only the dinucleotide CG was primarily conserved in CpG methylation. However, from the CG motif plot, there was evidence of other less conserved sub-motifs (e.g., NC^mGN), and we have outlined the frequencies with which these sub-motifs are methylated within the CpG context (Supporting Information File S4). The CG dinucleotide was 1.6X more numerous in coding regions versus intergenic regions, and of the 261 208 detected, $\sim$3% were methylated. Both GC^mCGC and $CTGC^mAG$ motifs were 3X more numerous in coding regions than intergenic regions, and of the 1936 GC^mCGC and 2084 $CTGC^mAG$ motifs detected, 95% and 98% were methylated respectively (Supporting Information File S4). Interestingly, neither sequence occurred more frequently than expected by chance in either genes or intergenic regions, yet $>$95% were methylated. Thus, these data suggest that the latter motifs could be cytosine methylation target sites in IMS101. Of the $>$200 bacterial taxa analysed in a recent prokaryotic epigenomic study (Blow *et al.*, 2016), only one instance of the GC^mCGC motif was predicted, which associated with a m5C *dcm* in *Neisseria gonorrhoeae* FA 1090 (NGO0365_GCCGC, bold in Fig. 1) while another instance of $CTGC^mAG$ was predicted to be associated with one methyltransferase gene of unknown specificity in of *Bifidobacterium longum subsp. infantis* ATCC (Blon_0292_CTGCAG, bold in Fig. 1). Based on our phylogenetic analyses of the Dcm homologs in *Trichodesmium*, the Tery_4810 methyltransferase may have the same recognition site as the *dcm* of *N. gonorrhoeae* (i.e., GC^mCGC) as they both reside within the same phylogenetic cluster.

Hypergeometric tests with Benjamini and Hochberg's FDR correction (Benjamini and Hochberg, 1995) (FDR $\leq$ 0.1; Methods) were conducted to identify enriched Gene Ontology (GO) categories within each of the three sequence contexts (Supporting Information File S4). Significantly enriched pathways shared among the three motifs involved energy, signalling, transport and core metabolic processes, consistent with enriched GO pathways in the m5C cytosine methylome of the diatom *P. tricornutum* (Veluchamy *et al.*, 2013). For example, GO categories were associated with oxidation/reduction activity, protein folding, protein binding, transporter activity, respiration, signal transduction, kinase activity, and so forth. (Fig. 3). Hence, although only $\sim$1% of cytosines are methylated genome-wide, approximately half of IMS101 genes (n = 2751; Supporting Information File S3) harbour methylation in significantly enriched GO categories (see hypergeometric test above) associated with broad energetic, carbon and signalling processes, suggesting that m5C methylation is associated with central but also specific pathways in core cellular metabolism.

Interestingly, a collection of genes involved in nitrogen assimilation and transformation tied to energy, pyrimidine and arginine metabolisms were not highly methylated but did retain multiple methylated sequence contexts within a single gene body. For example, glutamine synthetase (Tery_3834; GS) and ferredoxin-dependent glutamate synthase (Tery_0466; Fd-GOGAT) involved in ammonia assimilation (Kameya *et al.*, 2007) contained methylated residues within each of the three motifs. Two ammonium transporter genes (Tery_1788 and Tery_4477) harboured methylated cytosines in both the CCG and CWG sequence contexts (Fig. 3) while glutamate racemase (Tery_0161; *murI*) essential for both cell wall biosynthesis and gyrase inhibition (involved in DNA supercoiling) (Sengupta *et al.*, 2008) only retained methylated cytosines in the CWG context. Similar to GS and GOGAT, carbamoyl phosphate synthetase (Tery_2877; *carB*) catalysing the ATP-dependent synthesis of carbamoyl phosphate representing

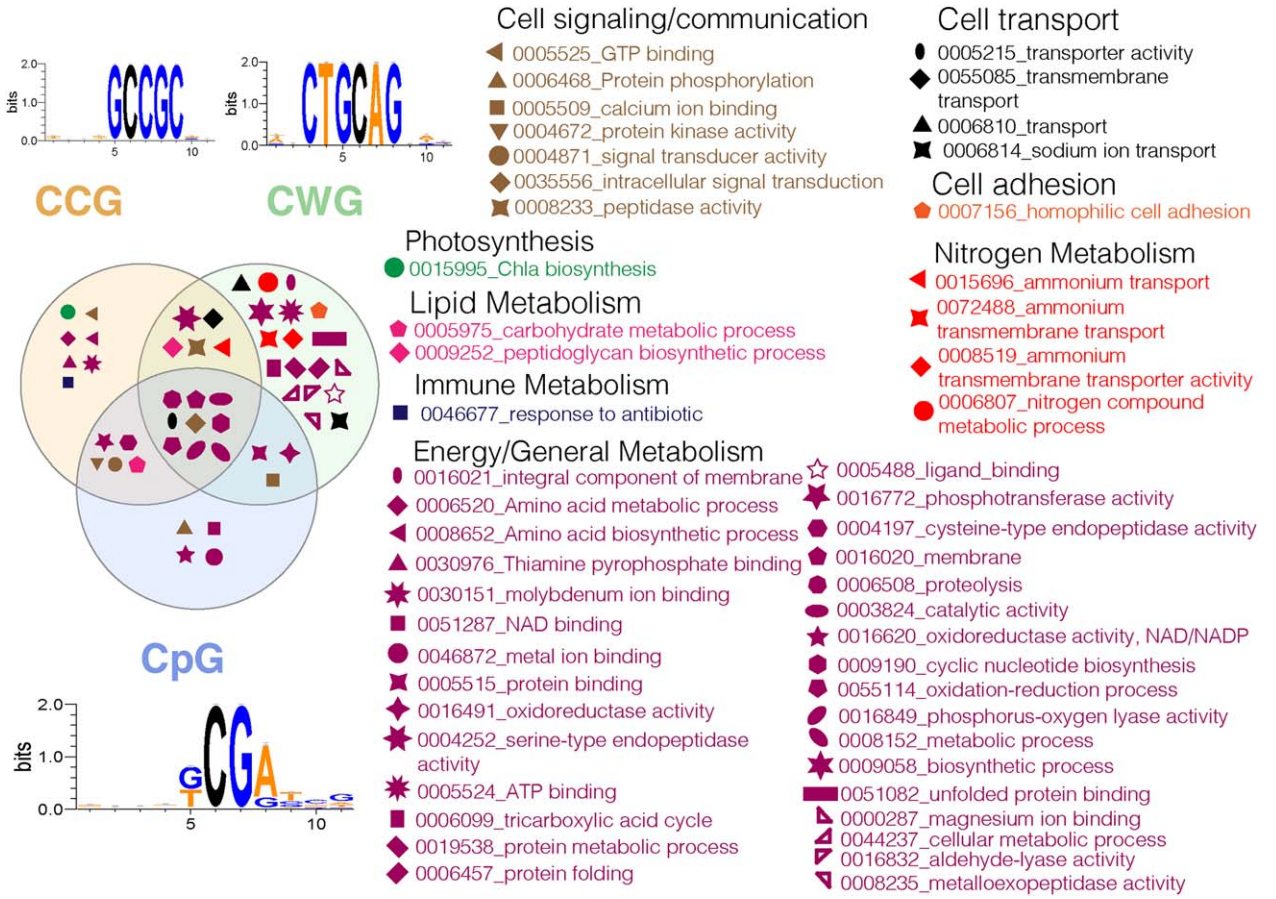

Fig. 3. m5C conserved motifs and Gene Ontology (GO) enriched pathways The Venn diagram displays the three sequence contexts (CCG, CWG and CpG) comprising > 99% of genome-wide methylated cytosines and the conservation of motif patterns for each of them. The black 'C' in the motif graphs represents the methylated cytosine. The symbols denote significantly GO-enriched pathways per sequence context, which represent pathways that are most highly methylated in the genome relative to their abundance (see Gene Ontology (GO) enrichment analysis).

the first committed step in pyrimidine and arginine biosynthesis (Holden *et al.*, 1999) harboured methylated cytosines in all three sequence motifs. Accordingly, arginine side groups are critical components of the nitrogen storage polymer, cyanophycin, in cyanobacteria (H. Li *et al.*, 2001). Other genes associated with nitrogen and energy metabolism retained methylated cytosines within all 3 motifs including urease (Tery_0752; *ureA*) involved in urea hydrolysis and NAD synthetase (Tery_1984; *nadE*) involved in electron transport and redox reactions. While it is tempting to speculate on the potential roles of methylated cytosines within each of the sequence contexts of different genes, more empirical studies are necessary. Nonetheless, taken together, the enrichment of methylated genes associated with these pathways suggests a possible role for m5C cytosine methylation in nitrogen assimilation and transformation linked to amino acid biosynthesis, energy metabolism and nitrogen storage. Conversely, the lack of enrichment of genes for nitrogen fixation (e.g., *nif*) suggests that cytosine methylation is not involved in their transcriptional control. However, since external inputs of nitrogen (e.g., environmentally supplied ammonia, urea or nitrate) can negatively impact *nif* expression, nitrogenase enzyme synthesis and nitrogenase activity (Bergman *et al.*, 2013), cytosine methylation may indirectly influence N_2 fixation via its potential regulatory roles in environmental nitrogen assimilation (e.g., GS/GOGAT pathway and urea hydrolysis). Future studies can examine methylation levels in nitrogen assimilation genes corresponding to fluctuations of environmentally supplied nitrogen.

Cytosine methylation correlation to genes and expression levels

We observed methylated residues to significantly overlap with genes more than expected if assuming independence (standard two-sided binomial test, $p < 0.01$; Methods) suggesting cytosine methylation to be significantly associated

with gene bodies in IMS101. Conversely, methylated loci (query) overlap significantly less than expected (negative correlation) with intergenic intervals ($p < 0.01$). To examine the relationship between methylation and length-normalized gene expression levels under normal culture conditions (replete nutrients in Aquil medium; Methods), Illumina RNA-Seq was performed and expression was compared across a gradient of R_{met} values. Upon binning the range of R_{met} values (0.0002–0.03) into deciles and plotting normalized expression values per bin, genes residing in the 10th decile with the highest R_{met} values (0.008–0.03) exhibited much lower expression levels than the rest of the bins (Fig. 4A). Each decile retained similar distribution shapes and were heavily skewed right (i.e., non-normal) due to a group of highly expressed genes shifting the overall distribution as commonly seen in expression data (Love *et al.*, 2014). Due to these skews, methods using quantiles (e.g., median, upper-quartile, etc.) are typically used to more adequately analyse significant changes in expression data (Robinson and Oshlack, 2010; Risso *et al.*, 2011; Love *et al.*, 2014). Hence, the nonparametric Kruskal–Wallis test was conducted to test whether length-corrected, median expression levels of bins with different R_{met} ranges were significantly different, thereby suggesting a shift in overall expression between bins. This test yielded a very low probability (i.e., highly statistically significant) that all bins contained the same median expression levels ($p < 10^{-13}$, Fig. 4). Next, the post hoc Dunn test with Benjamini–Hochberg correction (Benjamini and Hochberg, 1995) yielded that the median of bin 10 (Supporting Information File S5) with the highest R_{met} but lowest expression values was significantly different than all other bins (Fig. 4). This significant shift in length-corrected, median expression may suggest that these genes could be in part transcriptionally regulated by methylation rather than merely protected by it from R-M. An analogous result ($p < 10^{-16}$) is observed when expression values are plotted against deciles calculated from cytosine-specific methylation levels (R_c; # of methylated cytosines/# of per-gene cytosines; Supporting Information Fig. S3) as global R_{met} and R_c values generally show strong correlation ($R^2 = 0.89$; Supporting Information Fig. S1D, inset). Interestingly, transcriptional levels seem to begin to reduce in genes with R_{met} values approaching 0.01. Hence, to test if the median of expression levels in genes harbouring R_{met} values ≥ 0.01 (Supporting Information File S5) was significantly different than those with R_{met} values ≤ 0.01, we conducted a two-sample permutation ($n = 2000$) test with Monte Carlo simulation on the medians of each R_{met} group. This test was highly significant ($p < 10^{-4}$) and suggested significantly reduced expression levels were associated with R_{met} values ≥ 0.01. Further study is needed to determine if the R_{met} value of 0.01 may be an indicator of transcriptional regulation by cytosine methylation in *Trichodesmium* rather than, for example, merely a protective measure. Nonetheless, these data suggest that IMS101 genes may not need to be extensively methylated across the gene body (i.e., ~1% of total gene length) to be transcriptionally impacted by cytosine methylation. Conversely, only genes with dense methylation across the entire gene body correlated with reduced transcription in the diatom, *Phaeodactylum tricornutum* (Veluchamy *et al.*, 2013). Hence, future studies examining methylation changes across gene bodies relative to changes in transcription under differing conditions may

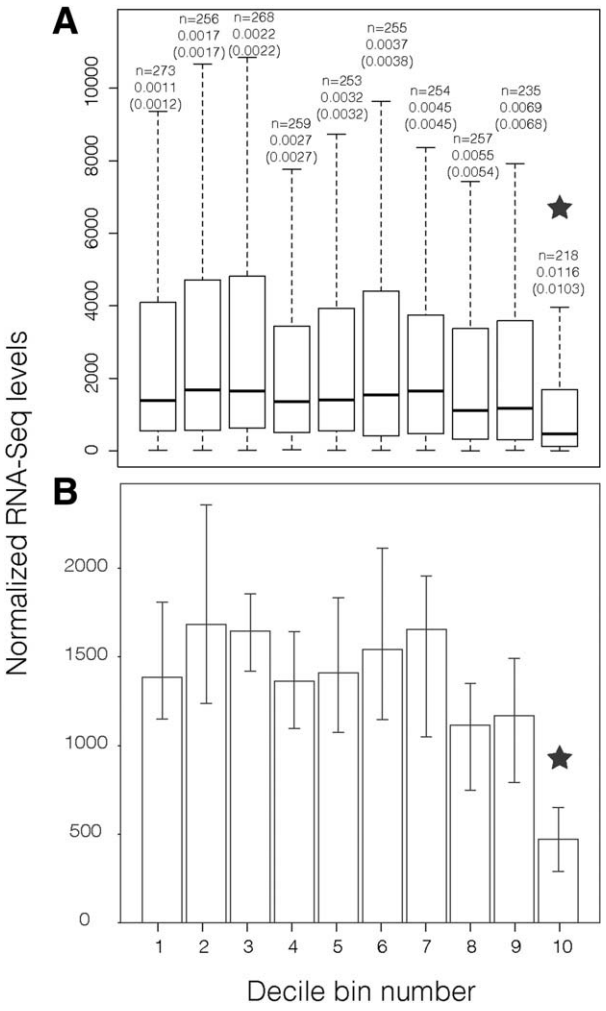

Fig. 4. Expression profiles of methylated genes (A) R_{met} values of methylated genes are distributed into deciles and boxplots of normalized expression values are plotted for each decile. Above each boxplot are 3 metrics and from top to bottom are: # of genes in that decile, mean R_{met} value and median R_{met} value in parentheses. The star indicates significantly different medians in that expression bin.
B. The same bins with their medians plotted. The star indicates statistical significance as above. Error bars display 95% confidence intervals.

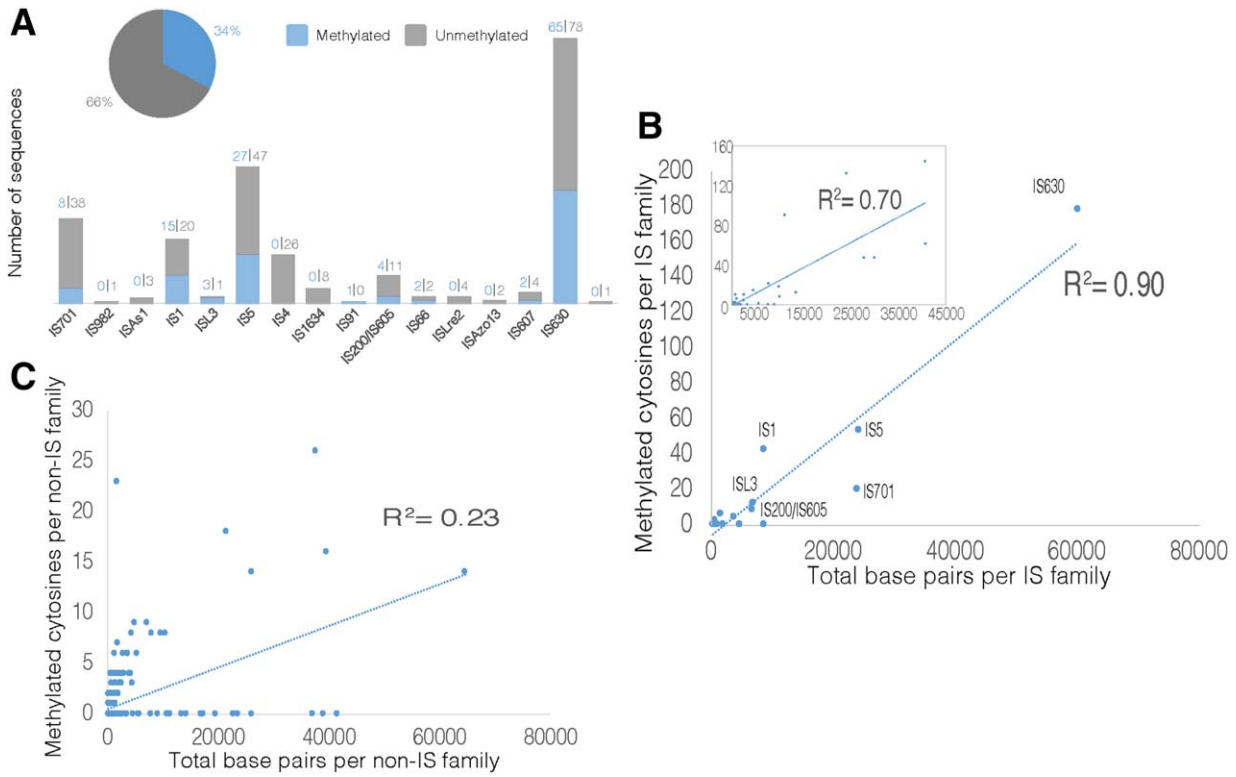

Fig. 5. Methylation and insertion sequences (IS) (A) The pie chart shows the distribution of methylated to total sequences for all IS families. The stacked bar plot shows the number of sequences per IS family with the numbers above each bar denoting: the # of methylated sequences (left/blue) and the # of unmethylated sequences (right/grey) per IS family.
B. Scatterplot shows the # of methylated cytosines as a function of total base pairs (bp) per IS family. Similarly, the inset shows the # of methylated cytosines as a function of total bp per IS class (i.e., family) identified from an in-house pipeline (Methods) (c) Scatterplot shows the # of methylated cytosines as a function of total bp per non-IS family (i.e., IGR family).

elucidate the degree to which gene body methylation influences expression.

m5C in insertion sequences and other repeat loci

The IMS101 genome harbours a large number of selfish DNA elements, or insertion sequences (IS) that may aid in its adaptation (Lin *et al.*, 2011) and also be in part responsible for the genome's low coding percentage (~60%) (Pfreundt *et al.*, 2014; Walworth *et al.*, 2015). Cytosine methylation has been shown to control both the transcription and transposition of transposable elements (TE) in plant genomes (Miura *et al.*, 2001; Kato *et al.*, 2003; Lister *et al.*, 2008), and methylation across transposable elements has been observed in other algae (Feng *et al.*, 2010; Veluchamy *et al.*, 2013). In IMS101, we observed significant overlap of intragenic methylated loci with IS loci (standard two-sided binomial test; $p < 0.001$) indicating IS gene features to be enriched within the methylated gene pool (Fig. 2B).

To further examine if cytosine methylation is associated with IS loci relative to other repetitive DNA elements devoid of IS sequence homology (i.e., intergenic repeating sequences, hereafter IGR), we analysed correlations between methylation and either IS or IGR (i.e., non-IS) sequences respectively. IGR sequences serve as a control group representing a random assortment of other repetitive elements in the genome. We observed ~34% of annotated IS (i.e., transposase) genes to be methylated (Fig. 5A) with methylated sites per IS family ($n = 16$) to be positively correlated with increasing number of sequences per family ($R^2 = 0.93$) and total base pairs (bp) per family ($R^2 = 0.90$; Fig. 5B). We also identified IS classes (i.e., families) using an in-house pipeline that clusters genome-wide transposase sequences via sequencing identity ($n = 69$) and observed a positive ($R^2 = 0.70$) correlation to total bp per class (Fig. 5B, inset; Methods). To examine whether methylation was substantially more associated with IS families than with IGR and whether the increase in methylated sites per IS family (or class) was merely due to the proliferation of repetitive DNA (i.e., increasing copy number or total

base pairs per IS family), we examined methylated sites per IGR family devoid of IS sequence homology. From these analyses, methylated residues per IGR family showed little correlation with either number of sequences ($R^2 = 0.22$) or total base pairs per IGR family ($R^2 = 0.23$; Fig. 5C). The substantially weaker correlation between methylation density and total base pairs (or sequence copy number) of IGR families relative to the stronger correlation between methylation density and total base pairs (or sequence copy number) of IS families (or classes) suggests that cytosine methylation is more associated with IS families. Furthermore, these data suggest that this higher correlation may not solely be due to repetitive DNA proliferation and may have a regulatory role in IS family (im)mobilization as seen in eukaryotes (Miura *et al*., 2001; Kato *et al*., 2003; Lister *et al*., 2008). From these data, it seems methylation at certain motifs are either associated to specific IS families (or classes) or proliferative in families with high copy numbers. Hence, more targeted studies assessing transposition and m5C methylation are needed to confirm these trends.

Genus-level conservation of m5C methylation

To help verify methylated residues in our IMS101 culture, we bisulfite-sequenced another IMS101 isolate (IMSB) that has been separately maintained in culture since first being isolated from coastal Atlantic waters in 1991 (Prufert-Bebout *et al*., 1993). We mapped bisulfite-treated DNA reads from IMSB onto the IMS101 reference genome (Methods) to obtain highly conserved methylated cytosines shared by both isolates ($n = 14\ 934$; 95% of the methylated calls in IMS101; Supporting Information File S6).

We also bisulfite-sequenced another strain of the *T. erythraeum* species (strain 2175) isolated from the Tropical Atlantic Ocean in 2006 and mapped reads onto the draft 2175 genome (Walworth *et al*., 2015). Similar numbers of methylated residues were detected between IMS101 and 2175 for the sequence contexts CG (11 753 and 11 469 respectively), CCG (1886 and 1636 respectively) and CWG (1972 and 2175 respectively) with similar proportions distributed between genic and intergenic regions (Supporting Information Fig. S2; Supporting Information Fig. S5; Supporting Information File S4; Supporting Information online text). Analogous motif patterns to those in IMS101 were also detected in 2175 for CCG (GC^mCGC), CWG ($CTGC^mAG$) and CG along with similar frequencies of methylation (3%, 98% and 100% of the CG, $CTGC^mAG$ and GC^mCGC respectively) in each of the sequence contexts (Supporting Information Fig. S4 and Supporting Information File S4). Furthermore, 2175 also harbours several gene *dcm* copies that tightly cluster phylogenetically with those of IMS101 relative to other bacterial taxa (Fig. 1, see legend). We also detected a similar number of methylated genes in IMS101 and 2175 (2752 and 2918 respectively) in which the majority were shared homologs ($n = 2244$; 82% relative to IMS101), suggesting cytosine methylation to affect similar metabolic pathways in both isolates (Supporting Information File S7). Of the 674 genes that were methylated in 2175 but not in IMS101, 502 had homologs in IMS101 in which 1/3 were either hypothetical or had unknown function (Supporting Information File S8). Most other homologs included genes involved in signal transduction, translation, transcription and ATP binding. Taken together, even though 2175 was isolated thousands of miles from the coastal Atlantic and several years later, these data suggest widespread mechanistic conservation of cytosine methylation in 2175 relative to IMS101. This global conservation could represent either constitutively methylated sites that protect certain regions from restriction enzymes and/or conserved regulatory roles methylated cytosines may have on gene regulation.

Finally, we bisulfite-sequenced a *Trichodesmium thiebautii* strain VI-1 (hereafter VI-1) isolated from the Sargasso Sea in 1998 (Hynes *et al*., 2012) as well as two natural populations of *Trichodesmium* sampled *in situ* on a cruise transect in the western South Atlantic (Stations 6 and 8 respectively) that were collected on a 130 μm net, handpicked and frozen (Webb *et al*., 2007) (also morphologically *T. thiebautii*). We then searched for the highly-conserved methylation sequence contexts (GC^mCGC and ($CTGC^mAG$) in these data by stringently mapping reads onto homologous regions of the IMS101 genome and identifying positive methylation calls (Methods). These positive methylation calls only occurred at the conserved methylation sequence contexts identified in IMS101 and 2175 while unmethylated cytosines occurred within no specifically conserved motifs (see above; Supporting Information Fig. S4). In aggregate, these conserved motifs detected *in situ* may also be potentially conserved markers of m5C methylation in natural populations as is the case with IMS101, IMSB and 2175 ($n = 3730$; Supporting Information File S9). Hence, this motif conservation detected across nearly a decade of time and space suggests the notion of a 'core' epigenome in *Trichodesmium*.

Conclusion

To the best of our knowledge, we present the only known genome-wide m5C methylation map for any marine diazotroph and its environmental conservation. Overall, the genome exhibits both low global methylation (~1% of total cytosines) and per-gene R_{met} levels (~0.03). Much of this methylation overlaps with essential core metabolic pathways, suggesting m5C methylation to associate with globally important processes in the cell. Conserved motifs were detected for CCG and CWG and were methylated >95% of the time, while the dinucleotide CG was

methylated ~3% of the time. Methylation was found to significantly overlap with genes relative to intergenic regions as seen in other algae and plants (Feng *et al*., 2010). Additionally, a positive relationship was observed between number of methylated residues and total base pairs of IS families, while little correlation was observed to that of IGR families. This suggests that cytosine methylation may play a role in transposition (im)mobilization, as seen in other systems (Miura *et al*., 2001; Kato *et al*., 2003; Lister *et al*., 2008). Finally, we identified highly conserved methylated sites overlapping with core metabolic pathways between spatiotemporally separated *Trichodesmium* isolates. This degree of conservation in global cellular metabolism provides evidence for the notion of a 'core' epigenome conserved within natural populations of *Trichodesmium*.

The methodology and scope described here elucidates the metabolic potential in specific biochemical pathways that overlaps with methylated residues. To more explicitly examine the regulation of, for example, nitrogen metabolism by cytosine methylation, more targeted assays of cytosine methylation under different nitrogen regimes is necessary and/or methylation studies using genetically tractable (e.g., *Anabaena*) N_2-fixers containing the capacity for m5C cytosine methylation. *Dcm* genes involved in cytosine methylation are also conserved in other globally distributed bacteria (e.g., *Synechococcus*) (Fig. 1). Hence, despite m5C methylation being broadly conserved across an extensive range of photoautotrophs, only a handful of studies have examined its role in metabolism within an ecological context. Accordingly, these data expand the limited information available on m5C methylation for both microbes and phytoplankton, and are thus important for not only the bio-ecology and biogeochemical implications of a globally distributed marine N_2-fixer but also for general evolutionary theory examining epigenetic impacts to adaptation in biological systems.

Experimental procedures

Culturing methods

Trichodesmium erythraeum strain IMS101 (IMS101) was maintained in a modified Aquil medium devoid of combined nitrogen containing standard vitamins and trace metals with 500 nM iron and 20 μM phosphate (Hutchins *et al*., 2015). Cultures were grown under a light intensity of 120 μmol photons per meter squared per second with a light-dark cycle of 12:12 light:dark in 26°C incubators. Cultures were continuously bubbled with 0.2 μm-filtered prepared air/CO_2 mixtures (Praxair) to maintain stable CO_2 concentrations of 380 μatm. Semi-continuous culturing methods were used on six replicate cell lines per treatment and each replicate was diluted individually based on the growth rate calculated for the respective replicate (Hutchins *et al*., 2007; 2013). Cultures were kept optically thin to avoid self-shading, nutrient limitation and perturbations to targeted CO_2 levels, and total population size in each biological replicate was approximately $7 \times 10^5 - 1.1 \times 10^6$ cells, depending on growth stage, based on microscopic cell counts. IMSB, 2175 and VI-1 were maintained in batch YBC-II medium as previously described (Chappell and Webb, 2010).

DNA/RNA sampling and isolation for Illumina sequencing

For DNA sampling, three randomly chosen biological replicates were gently filtered onto 5 μm polycarbonate filters (Whatman) during the middle of the photoperiod, immediately flash frozen and stored in liquid nitrogen until extraction DNA extraction. Samples for RNA analysis were simultaneously subjected to the same sampling procedure in biological duplicate. Sampling details for other *Trichodesmium* samples can be found in the following studies: IMSB (Prufert-Bebout *et al*., 1993), 2175 (Walworth *et al*., 2015), *T. theibautii* VI-1(Hynes *et al*., 2012) and *Trichodesmium* natural populations from Stations (St.) 6 and 8 on an Atlantic cruise transect (Webb *et al*., 2007). DNA was extracted from frozen filters with the FastDNA Spin Kit for Soil (MP Biomedicals, Santa Ana, CA, USA) following the manufacturer's protocol. Extracted DNA was then sent to the USC Epigenome Center for library construction and sequencing. Briefly, ~100 ng of DNA was bisulfite treated with the Zymo Gold kit (Zymo Research) and libraries were constructed using the Ovation Ultra-Low Methyl-Seq library kit (NuGEN) followed by sequencing on the NextSeq (Illumina). Genome coverage and read mapping statistics can be found in Supporting Information Table S1.

RNA was extracted using the Ambion MirVana miRNA Isolation Kit (Thermo Fisher Scientific) in an RNAse free environment according to the manufacturer's instructions followed by two incubations with Ambion's Turbo DNA-free kit to degrade trace amounts of DNA. Extracted RNA was sent to the UC San Diego IGM Genomics Center for library construction and Illumina sequencing. Briefly, rRNA was removed from total RNA using the Ribo-Zero rRNA Removal Kit (Illumina), and libraries were constructed with the TruSeq Stranded mRNA Library Prep Kit (Illumina) followed by 50 base pair, single end sequencing with the Illumina HiSeq.

Methylation bioinformatics

The methods for this section pertain to *Trichodesmium* isolates with available reference genomes including IMS101, IMSB and 2175. Raw reads were quality processed and mapped using both the MethPipe pipeline (Song *et al*., 2013) and the BSMAP package (Xi and W. Li, 2009) with default settings. A cytosine was deemed methylated if the residue had a combined coverage of $\geq$5 reads and if methylation was detected in all biological replicates with at least 20% of total reads being methylated (Veluchamy *et al*., 2013). Methylated cytosines identified by both mapping packages were kept and cytosine methylation levels were estimated with MethPipe. A combination of coverage, biological replicates and bisulfite conversion rates (the rate at which unmethylated cytosines appear as thymines in sequenced reads) are used to determine the confidence in positively identifying methylated cytosines in which a combined coverage of 15–30$\times$ in biological duplicate and a bisulfite conversion rate of $>$0.99 is advised (Song *et al*., 2013; Ziller *et al*., 2014). Hence, to confidently assign

methylation in IMS101, we deeply sequenced 3 biological replicates (380 μatm CO_2 treatment) yielding a combined genome-wide coverage of $>700\times$ and a per-sample bisulfite conversion rate of >0.99 (Supporting Information File S10). In order to determine if methylated cytosines overlapped with promoter regions containing empirically determined transcriptional start sites (TSS), TSS coordinates were downloaded from Pfreundt and colleagues (2014) and compared with methylated residue genome coordinates. Motif bar graphs were generated by aligning ± 5 base pairs surrounding the methylated cytosine, and motif conservation was determined and visualized using Weblogo (v 2.8.2) (Crooks *et al.*, 2004).

Gene ontology (GO) enrichment analysis

Gene Ontology (GO) annotations for *Trichodesmium* were downloaded from the Genome2D web server (http://pepper.molgenrug.nl/index.php/bacterial-genomes). Next, hypergeometric tests were conducted using the 'phyper' function in 'R' (R Core Team, 2014) in order to test for significant enrichment of GO categories among methylated cytosines. *P*-values were corrected with the Benjamin and Hochberg method (FDR ≤ 0.1) (Love *et al.*, 2014). Finally, enriched GO categories were manually checked.

Maximum likelihood phylogeny

IMG-annotated (https://img.jgi.doe.gov/) cytosine DNA methyltransferases (*dcm*) were extracted from the IMS101 genome and searched against the 2175 genome using BLASTn (Altschul *et al.*, 1990) to obtain 2175 *dcm* homologs. *Dcm* sequences were then searched against the NCBI nonredundant protein database (June 2016) using BLASTx and high-scoring pairs covering $\geq 70\%$ of the original protein length with an e-value $< 1 \times 10^{-5}$ were kept for phylogenetic analysis. Protein multiple sequence alignments were generated using MUSCLE (v3.8.31) (Edgar, 2004) and trimmed using trimal (1.2rev59) (Capella-Gutiérrez *et al.*, 2009), and RAxML (v8.0.0) (Stamatakis, 2014) was used to construct maximum likelihood phylogeny with 100 bootstrap replicates and the following parameters: -f a -p 12 345 -m PROTCATLG -N 100 -x 12 345 -T 4.

Correlation of methylation to genetic elements

The GenometriCorr package (Favorov *et al.*, 2012) was used to test for correlations between methylated loci and genetic features. In order to test for significant overlap of methylated loci with genetic elements (genes or intergenic regions), a standard two binomial test was used via the 'Projection test'.

RNA-Seq normalization

Raw fastq files were quality trimmed and filtered as previously described (Walworth *et al.*, 2015) and mapped onto IMS101, IMG-called genes (https://img.jgi.doe.gov/) using Bowtie2 v2.2.6(Ben Langmead and Salzberg, 2012) with default settings. Genes containing an average of less than 10 counts across libraries were removed, and the remaining were first normalized by gene length via a 'Loess regression' in the EDASeq package (Risso *et al.*, 2011) followed by normalization using calculated size factors per library in the DESeq2 package(Love *et al.*, 2014).

Methylation in insertion sequences and other repeat loci

In order to test for correlations between m5C methylation and insertion sequences/transposases, CDS elements identified as transposases by ISfinder (Siguier *et al.*, 2006) were downloaded from Pfreundt and colleagues2014 (2014) to obtain genome-wide IS loci coordinates per IS family (Methods), which were then tested for significant overlap with intragenic methylated loci via the exact binomial test (Favorov *et al.*, 2012). Methylated residues overlapping with IS CDS's were then summed across each IS family and plotted as a function of total base pairs per family (Fig. 5B).

We also used an in-house method to quantify methylated cytosines per IS class (i.e., family). IS sequences identified by Walworth and colleagues2015 (2015) were clustered at 70% identity using USEARCH (Edgar, 2010), which yielded representative centroid sequences for each of the resulting 69 classes. Next, BLASTn (Altschul *et al.*, 1990) was used to search for all paralogous sequences for each centroid (class) within the IMS101 genome (e-value $\leq 1\ e-5$ and a minimum length threshold of $>=$ 70% of the original centroid sequence length). Methylated cytosines overlapping with IS paralogs were then summed across each IS class and plotted as a function of total base pairs per class (Fig. 5B, inset).

In order to construct intergenic repeat (e.g., IGR) families devoid of IS sequence homology, centroid (i.e., representative) sequences representing repetitive genetic elements in IMS101 were downloaded from Walworth and colleagues2015 (2015) and those exhibiting homology to IS elements were removed via USEARCH (Edgar, 2010). Paralogs for each centroid were then searched for against the IMS101 scaffold using BLASTn. High-scoring pairs spanning $\geq 50\%$ of the original centroid sequence length with an e-value $< 1 \times 10^{-5}$ were kept and methylated residues overlapping with paralogs were summed across each IGR family and plotted as a function of total bp per IGR family (Fig. 5C).

Data access

All physiological data are in Supporting Information File S11. Raw read files for both bisulfite-sequenced (BS) libraries and RNA-Seq libraries have been deposited in the NCBI's

Gene Expression Omnibus (Edgar *et al.*, 2002) and are accessible through GEO Series accession number

GSE86992 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE86992).

Acknowledgments

The authors thank S. Collins for discussions on epigenetics and adaptation. Funding was provided by the U.S. National Science Foundation OCE 1260490 and OCE 1143760 to D.A. Hutchins, E.A. Webb and F.-X. Fu. The authors declare no competing interests.

Author contributions

Experiments were conceived and carried out by N.G. Walworth, D.A. Hutchins, E.A. Webb and F.-X. Fu with experimental and analytical contributions from E. Dolzhenko, M.D. Lee, A.D. Smith. N.G. Walworth wrote the manuscript with assistance from M.D. Lee, D.A. Hutchins and E.A. Webb.

References

Altschul, S., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* **57:** 289–300.

Bergman, B., Sandh, G., Lin, S., Larsson, J., and Carpenter, E.J. (2013) *Trichodesmium*– a widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiol Rev* **37:** 286–302.

Blow, M.J., Clark, T.A., Daum, C.G., Deutschbauer, A.M., Fomenkov, A., Fries, R., *et al*. (2016) The Epigenomic Landscape of Prokaryotes. *PLoS Genet* **12:** e1005854.

Boyer, H.W., Chow, L.T., Dugaiczyk, A., Hedgpeth, J., and Goodman, H.M. (1973) DNA substrate site for the Eco RII. Restriction endonuclease and modification methylase. *Nature* **244:** 40–43.

Campbell, J.L., and Kleckner, N. (1990) *E. coli* oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. *Cell* **62:** 967–979.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25:** 1972–1973.

Chappell, P.D., and Webb, E.A. (2010) A molecular assessment of the iron stress response in the two phylogenetic clades of *Trichodesmium*. *Environ Microbiol* **12:** 13–27.

Clark, T.A., Lu, X., Luong, K., Dai, Q., Boitano, M., Turner, S.W., *et al*. (2013) Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol* **11:** 4.

Collins, S., Rost, B., and Rynearson, T.A. (2013) Evolutionary potential of marine phytoplankton under ocean acidification. *Evol Appl* **7:** 140–155.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res* **14:** 1188–1190.

Doberenz, S., Eckweiler, D., Reichert, O., Jensen, V., Bunk, B., Spröer, C., *et al*. (2017) Identification of a *Pseudomonas aeruginosa* PAO1 DNA methyltransferase, its targets, and physiological roles. *mBio* **8:** e02312–16.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792–1797.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460–2461.

Edgar, R., Domrachev, M., and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30:** 207–210. http://doi.org/10.1093/nar/30.1.207.

Ehrenreich, I.M., and Pfennig, D.W. (2015) Genetic assimilation: a review of its potential proximate causes and evolutionary consequences. *Ann Bot* **117:** 769–779.

Favorov, A., Mularoni, L., Cope, L.M., Medvedeva, Y., Mironov, A.A., Makeev, V.J., and Wheelan, S.J. (2012) Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput Biol* **8:** e1002529.

Feng, S., Cokus, S.J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M.G., *et al*. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci USA* **107:** 8689–8694.

Geoghegan, J.L., and Spencer, H.G. (2012) Population-epigenetic models of selection. *Theor Popul Biol* **81:** 232–242.

Glickman, B.W., and Radman, M. (1980) *Escherichia coli* mutator mutants deficient in methylation-instructed DNA mismatch correction. *Proc Natl Acad Sci USA* **77:** 1063–1067.

Heithoff, D.M., Sinsheimer, R.L., Low, D.A., and Mahan, M.J. (1999) An essential role for DNA adenine methylation in bacterial virulence. *Science* **284:** 967–970.

Holden, H.M., Thoden, J.B., and Raushel, F.M. (1999) Carbamoyl phosphate synthetase: an amazing biochemical odyssey from substrate to product. *Cell Mol Life Sci* **56:** 507–522.

Huo, W., Adams, H.M., Zhang, M.Q., and Palmer, K.L. (2015) Genome modification in enterococcus faecalis OG1RF assessed by bisulfite sequencing and single-molecule real-time sequencing. *J Bacteriol* **197:** 1939–1951.

Hutchins, D.A., Fu, F., Zhang, Y., Warner, M.E., Feng, Y., Portune, K., *et al*. (2007) CO_2 control of *Trichodesmium* N_2 fixation, photosynthesis, growth rates, and elemental ratios. *Limnol Oceanogr* **52:** 1293–1304.

Hutchins, D.A., Fu, F.-X., Webb, E.A., Walworth, N., and Tagliabue, A. (2013) Taxon-specific response of marine nitrogen fixers to elevated carbon dioxide concentrations. *Nat Geosci* **6: 1–6:**

Hutchins, D.A., Walworth, N.G., Webb, E.A., Saito, M.A., Moran, D., McIlvin, M.R., *et al*. (2015) Irreversibly increased nitrogen fixation in *Trichodesmium* experimentally adapted to elevated carbon dioxide. *Nat Commun* **6:** 8155.

Hynes, A.M., Webb, E.A., Doney, S.C., and Waterbury, J.B. (2012) Comparison of cultured *Trichodesmium* (Cyanophyceae) with species characterized from the field[1]. *J Phycol* **48:** 196–210.

Jablonka, E., and Raz, G. (2009) Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Q Rev Biol* **84:** 131–176.

Kameya, M., Ikeda, T., Nakamura, M., Arai, H., Ishii, M., and Igarashi, Y. (2007) A novel ferredoxin-dependent glutamate synthase from the hydrogen-oxidizing chemoautotrophic bacterium Hydrogenobacter thermophilus TK-6. *Journal of Bacteriology* **189:** 2805–2812. http://doi.org/10.1128/JB.01360-06.

Kahramanoglou, C., Prieto, A.I., Khedkar, S., Haase, B., Gupta, A., Benes, V., *et al*. (2012) Genomics of DNA cytosine methylation in *Escherichia coli* reveals its role in stationary phase transcription. *Nat Commun* **3:** 886–889.

Kato, M., Miura, A., Bender, J., Jacobsen, S.E., and Kakutani, T. (2003) Role of CG and non-CG methylation in immobilization of transposons in *Arabidopsis*. *Curr Biol* **13:** 421–426.

Klironomos, F.D., Berg, J., and Collins, S. (2013) How epigenetic mutations can affect genetic evolution: model and mechanism. *BioEssays* **35:** 571–578.

Kronholm, I., and Collins, S. (2016) Epigenetic mutations can both help and hinder adaptive evolution. *Mol Ecol* **25:** 1856–1868.

Krueger, F., Kreck, B., Franke, A., and Andrews, S.R. (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat Methods* **9:** 145–151.

Kumar, R., Mukhopadhyay, A.K., Ghosh, P., and Rao, D.N. (2012) Comparative transcriptomics of *H. pylori* strains AM5, SS1 and their hpyAVIBM deletion mutants: possible roles of cytosine methylation. *PLoS One* **7:** e42303.

Langmead, B., and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Li, H., Sherman, D.M., Bao, S., and Sherman, L.A. (2001) Pattern of cyanophycin accumulation in nitrogen-fixing and non-nitrogen-fixing cyanobacteria. *Arch Microbiol* **176:** 9–18.

Lin, S., Haas, S., Zemojtel, T., Xiao, P., Vingron, M., and Li, R. (2011) Genome-wide comparison of cyanobacterial transposable elements, potential genetic diversity indicators. *Gene* **473:** 139–149. http://doi.org/10.1016/j.gene.2010.11.011.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133:** 523–536.

Loenen, W.A.M., Dryden, D.T.F., Raleigh, E.A., Wilson, G.G., and Murray, N.E. (2014) Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res* **42:** 3–19.

Love, M.I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550.

Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H., and Kakutani, T. (2001) Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature* **411:** 212–214.

Palmer, B.R., and Marinus, M.G. (1994) The dam and dcm strains of *Escherichia coli* — a review. *Gene* **143:** 1–12.

Pfreundt, U., Kopf, M., Belkin, N., Berman-Frank, I., and Hess, W.R. (2014) The primary transcriptome of the marine diazotroph *Trichodesmium erythraeum* IMS101. *Sci Rep* **4:** 6187–6111.

Prufert-Bebout, L., Paerl, H.W., and Lassen, C. (1993) Growth, nitrogen fixation, and spectral attenuation in cultivated *Trichodesmium* species. *Appl Environ Microbiol* **59:** 1367–1375.

R Core Team. (2014) *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. https://www.R-project.org/.

Reisenauer, A., Kahng, L.S., McCollum, S., and Shapiro, L. (1999) Bacterial DNA methylation: a cell cycle regulator? *J Bacteriol* **181:** 5135–5139.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011) GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12:** 480.

Robinson, M.D., and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11:** R25.

Roer, L., Aarestrup, F.M., and Hasman, H. (2015) The EcoKI Type I restriction-modification system in *Escherichia coli* affects but is not an absolute barrier for conjugation. *J Bacteriol* **197:** 337–342.

Sánchez-Romero, M.A., Cota, I., and Casadesús, J. (2015) DNA methylation in bacteria: from the methyl group to the methylome. *Curr Opin Microbiol* **25:** 9–16.

Schmitz, R.J., Schultz, M.D., Lewsey, M.G., OMalley, R.C., Urich, M.A., Libiger, O., *et al*. (2011) Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334:** 369–373.

Sengupta, S., Ghosh, S., and Nagaraja, V. (2008) Moonlighting function of glutamate racemase from *Mycobacterium tuberculosis*: racemization and DNA gyrase inhibition are two independent activities of the enzyme. *Microbiology* **154:** 2796–2803.

Severin, P.M.D., Zou, X., Gaub, H.E., and Schulten, K. (2011) Cytosine methylation alters DNA mechanical properties. *Nucleic Acids Res* **39:** 8740–8751.

Siguier, P., Pérochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34:** D32–D36.

Sohm, J.A., Webb, E.A., and Capone, D.G. (2011) Emerging patterns of marine nitrogen fixation. *Nat Rev Micro* **9:** 499–508.

Song, Q., Decato, B., Hong, E.E., Zhou, M., Fang, F., Qu, J., *et al*. (2013) A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* **8:** e81148.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313.

Tatusova, T., Ciufo, S., Federhen, S., Fedorov, B., McVeigh, R., O'Neill, K., *et al*. (2015) Update on RefSeq microbial genomes resources. *Nucleic Acids Res* **43:** D599–D605.

Vasu, K., and Nagaraja, V. (2013) Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol Mol Biol Rev* **77:** 53–72.

Veluchamy, A., Lin, X., Maumus, F., Rivarola, M., Bhavsar, J., Creasy, T., *et al*. (2013) Insights into the role of DNA methylation in diatoms by genome-wide profiling in *Phaeodactylum tricornutum*. *Nat Commun* **4:** 1–10.

Waldron, D.E., Owen, P., and Dorman, C.J. (2002) Competitive interaction of the OxyR DNA-binding protein and the Dam methylase at the antigen 43 gene regulatory region in *Escherichia coli*. *Mol Microbiol* **44:** 509–520.

Walworth, N., Pfreundt, U., Nelson, W.C., Mincer, T., Heidelberg, J.F., Fu, F., *et al*. (2015) *Trichodesmium* genome maintains abundant, widespread noncoding DNA in situ, despite oligotrophic lifestyle. *Proc Natl Acad Sci USA* **112:** 4251–4256.

Webb, E.A., Jakuba, R.W., and Moffett, J.W. (2007) Molecular assessment of phosphorus and iron physiology in *Trichodesmium* populations from the western Central and western South Atlantic. *Limnol Oceanogr* **52:** 2221–2232.

Xi, Y., and Li, W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10:** 232.

Zehr, J.P., Ohki, K., Fujita, Y., and Landry, D. (1991) Unique modification of adenine in genomic DNA of the marine cyanobacterium Trichodesmium sp. strain NIBB 1067. *J Bacteriol* **173:** 7059–7062.

Ziller, M.J., Hansen, K.D., Meissner, A., and Aryee, M.J. (2014) Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nat Methods* **12:** 230–232.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Gene-specific methylation statistics. Shown are the distribution of methylated cytosines per gene and corresponding per-gene methylation levels (A, B) as well as methylated cytosines to total gene length and methylation level to total cytosines per gene (C, D). The histogram in (A) shows the frequency of methylated cytosines per gene while the histogram in (B) shows the frequency of methylation levels (R_{met}; # of methylated cytosines/ total base pairs per gene) per gene. The scatterplot in (C) shows methylated cytosines per total gene length while the scatterplot in (D) shows methylation levels per total cytosines per gene. The inset in (D) shows the positive correlation between R_{met} levels (*y*-axis) and cytosine-specific methylation levels (R_c; # of methylated cytosines/ # of per-gene cytosines).

Fig. S2. Distribution of methylated cytosines. Shown are the intra- and intergenic distributions of methylated cytosines within the analysed sequence contexts for both the IMS101 (blue) and 2175 (orange) genomes respectively.

Fig. S3. Boxplots of normalized expression values per R_c decile. Shown are boxplots of the normalized expression values of genes split into deciles based on their per-gene cytosine-specific methylation levels (R_c; # of methylated cytosines/ # of per-gene cytosines). Genes in the 10^{th} decile with the highest R_c levels retain the lowest gene expression. The star indicates significantly different medians in that expression bin. See main text for discussion.

Fig. S4. Conserved motifs associated with methylated cytosines per sequence context and no motifs associated with unmethylated cytosines per sequence context. Shown are the detected conserved motifs for each of the sequence contexts surrounding methylated cytosines for IMS101/2175 (left panel) and natural populations (middle panel). The right panel shows the lack of conserved motifs surrounding unmethylated cytosines (right panel) in each of the sequence contexts. Methylated cytosines are coloured black while unmethylated are red.

Fig. S5. Surrounding bases around methylated cytosines in the CHH context in both IMS101 and 2175. Shown are surrounding nucleotides for methylated cytosines detected in the CHH context for IMS101 and 2175 respectively. Here, the methylated cytosine is the large blue 'C'.

Table S1. General bisulfite-converted DNA sequencing statistics. Shown are general bisulfite-converted DNA and sequencing statistics for each sample containing a reference genome. From left to right is the sample name, estimated bisulfite conversion rate (as estimated by the MethPipe pipeline), the fraction of covered cytosines and the average genome coverage.