Integrative Analysis of DNA Methylation and Gene Expression of Emiliania Huxleyi Reveals Gene Families Related to calcification

# Abstract

# Introduction

Coccolithophores play an important role in the oceanic carbon cycle through the calcification of coccoliths and photosynthesis. Their intricate coccolith structures have attracted the attention of scientists from various fields of material science, geology, biogeography, ecophysiology, and particularly those from the biomedical community for potential biomedical applications related to bone formation and related diseases[1,2]. Biomineralized tissues such as bones and teeth are of fundamental importance in medicine and health care, but little is yet known about the mechanisms that govern the formation of these tissues. Understanding the design principles of coccolith structures may also have potential applications in material science for creating novel materials for periodontal structures, bone scaffolding, biomedical implants, and membranes for high-temperature separations[3]. While many coccolithophores have restricted distributions, one species, *Emiliania huxleyi* (*E. huxleyi*), is exceptional in the breadth of its distribution, making it the dominant bloom-forming coccolithophorid and one of the most ubiquitous and abundant species of oceanic phytoplankton. It has emerged as a model system for studying biosphere-geosphere interactions, and the subject of extensive physiological, biochemical, and ecological research [[3](#ENREF_3), [4](#ENREF_4)]. The sequencing and annotation of its genome[4] made it an ideal model organism for studying biomineralization in coccolithophores.

Coccoliths are synthesized in a matrix-mediated manner in the calcifying *E. huxleyi*. The nature, orientation, size, and shape of the crystalline elements are controlled within a specialized compartment that resides adjacent to the nucleus, known as the coccolith vesicle. A Golgi-derived reticular body attaches to the coccolith vesicle to deliver calcium and key matrix-macromolecules, likely to include proteins, polysaccharides, proteolipids, and proteoglycans. Once the coccolith is completely formed, the reticular body disintegrates, and the vesicle migrates to the cell surface, where the coccolith is extruded in a massive exocytotic event and becomes part of an interlocking sphere of platelets that encapsulate the cell. Cellular proteins presumably coordinate biomineralization and coccolith-genesis and are involved in processes associated with ion transport; crystal nucleation, growth, and patterning; vesicle trafficking, and exocytosis.

DNA methylation is a heritable epigenetic mark in many eukaryotic organisms and plays an important role in many biological functions by regulating genome structure and transcription[5,6]. DNA methylation is an essential epigenetic modification to an organism’s genome that regulates gene expressions in normal cellular differentiation and development, disease pathogenesis, and aging [7,8]. Hypermethylation of CpG sites/islands surrounding the transcription start site (TSS) is often hypothesized to repress the transcription of that corresponding gene [9,10]. However, more recent studies showed that the relations between DNA methylation and gene expression are much more complicated. Simple binary classification of hyper or hypo-methylation in promoter regions only found small negative correlations between methylation and gene expression levels [11–13].

Two isogeneic *E. huxleyi* strains, calcifying M217 and non-calcifying CCMP1516, provide an ideal model system for studying the effects of epigenetic modifications on the gene expression and regulation related to biomineralization. Isolated from the South Pacific (02.6667S 82.767W) in 1991. *E. huxleyi* strain CCMP1516 (EH1516) was maintained in the Bigelow National Center for Marine Algae and Microbiota, where, over time, it lost its ability to calcify. Before becoming a non-calcifier, a subclone of CCMP1516 was sent to the Plymouth Algal Collection and designated M217 (EH217). Unlike CCMP1516, M217 has retained its ability to calcify in culture. CCMP1516 and M217 are isogeneic lines, evidenced by the 100% identity of the nucleotide sequences of the cox3, tufA, and the mitochondrial and plastid 16S rRNA. Whole genome DNA methylation has not been systematically studied for E. huxleyi or other coccolithophores. Being isogeneic strains, methylation is hypothesized to be a key factor for phenotypic differences in calcification between EH1516 and EH217.

To study the correlations between methylation patterns and corresponding gene expression profiles in EH217 and EH1516, we generated libraries of RNA-Seq and whole-genome bisulfite sequencing (WGBS) data for six samples, three biologically replicates for each strain.

The most common approach to associate DNA methylation with gene expression changes is first to identify differentially methylated cytosine sites (DMCs) and differentially methylated regions (DMRs) on the genome and associate them with nearby genes [14]. However, studies often only found weak correlations between hyper or hypo-regulated DMRs and differential gene expression [11–13,15]. A Bayesian linear regression model was used to measure the log2 fold changes (LFC) of gene expression levels according to mean methylation values in DMRs and specific genomic regions, for example, promoter, 5’UTR, and gene body, across 19 human cell types [16]. The model had a near-zero median value and relatively large deviation for the regression coefficients. Therefore, the simple binary or linear assumptions in these DMR based models could be too simplistic for DNA methylation patterns that demand more sophisticated models. Another difficulty of applying DMR methods is that they depend on some a priori thresholds for the size and number of CpGs to determine DMRs. The results produced from different DMR calling programs are often very different, and adjusting these parameters can significantly change the numbers of DMRs identified.

This study analyzed the differential methylation patterns between calcifying and non-calcifying strains of *E. huxleyi* strains and their relations to differential gene expressions.

# RESults

## Genome-wide patterns of DNA methylations in *E. huxleyi*

The average methylation level revealed the general characteristics of the methylome and was computed as the ratio of the number of reads with methylated cytosines to the total number of mapped reads covering the site. The cytosine methylation levels varied depending on the local sequence context (C, CpG, CHG, and CHH) and the *E. huxleyi* strains. The EH1516 strain had an average methylation level of ~7.1% for all C’s, 16.1% for CpG, 5.3% for CHG, and 0.5% for CHH contexts. In comparison, the EH217 strain had an average methylation level of ~7.2%, 15.8%, 5.9%, and 0.6% for all C’s, the CpG, CHG, and CHH contexts (Table 1). It showed that the two E. huxleyi strains had similar average methylation levels, while the EH217 had a slightly higher methylation rate in the CHG context.

The cytosine methylation rates varied according to the local sequence context (CG. CHG or CHH) and the *E. huxleyi* strains. As shown in Figure 1, the two *E. huxleyi* strains had similar overall methylation levels in the CpG context at about 15.9%; EH217 had a slightly higher average methylation level of 5.9% in the CHG context than 5.3% in EH1516 (p-value ???). Because of the very low methylation rate in the CHH context, we did not consider the CHH context further in this study. Figure 2 showed the average methylation levels for the 100 longest scaffolds in the E. huxleyi genome. The mean CpG methylation level per scaffold varied from 12.3% to 18.6%, with the mean of 15.88±0.23 % for EH1516 and 15.68±0.22% for EH217 at 0.95 confidence interval. The differences between the EH217 and EH1516 were insignificant (< 1%) for most scaffolds. The only exception was scaffold 14, on which the EH217 CpG methylation level was 3.8% higher than EH1516. In the CHG context, EH217 had a consistently higher mean methylation level (6.18±0.1%) than EH1516 (5.49±0.1%), although the differences were usually less than 1%. Despite their comparable average methylation levels, the two strains had differentiable methylation profiles across the whole genome. Figure 3. (a) Correlations and (b) PCA plot of *E. huxleyi* methylation profiles in the CpG context; (c) and (d) in the CHG context. Figure 3 showed the correlations and the PCA plots of sample methylation profiles in the CpG and CHG contexts. The samples from the same *E. huxleyi* strain had a much higher average correlation than samples from different strains in both the CpG and CHG methylation profiles. The CpG context had an average correlation of 0.986 between samples of the same strain compared to 0.932 across strains, and CHG context showed more considerable variations of having a mean correlation of 0.967 and 0.857 for samples within and across the strains. The PCA plots of the CpG and CHG methylation profiles in Figure 3 (b) and (d) further revealed that the two *E. huxleyi* strains had genuine differences in genome-wide methylations that might contribute to their differences in gene expression profiles and physiology.

As shown in Figure 4, an overview of the genomic methylation level in 10Kbp windows on scaffolds 1 and 14 revealed that the methylation rates varied widely across the *E. huxleyi* genome. Scaffold 1 was the longest, and both strains shared similar overall methylation profiles along the scaffold. Scaffold 14 had the most significant change of average CpG methylation levels between the strains, showing a marked difference around the 61Kbp position. Supplemental Figure 1 showed that most cytosines in the E. huxleyi genomes were not methylated, with more than 80% of CpG sites and more than 90% of CHG sites having methylation levels lower than 10%. On the other hand, among Cs’ with methylation level > 10%, the methylation levels between 90%-100% were most prominent, including more than 12% of the CpG sites and more than 3% of the CHG sites.

## DNA methylation levels in different regions of the E. huxleyi genome

To compare DNA methylation patterns in the different regions of the E. huxleyi genome, we computed the average methylation levels in various regions around the annotated genes (Figure 5). The promoters were selected to be 1000 bp upstream and 1000 bp downstream from the transcription start sites (TSS), and up2000 and down2000 were 2000bp regions upstream and downstream from TSS.

In the CpG context, exons had the highest average methylation levels in exons and lowest in up2000. The CpG methylation levels were significantly different between the EH1516 and EH217 strains in the up2000, down2000 and promoters (p < 0.01), as well as in the exons (p < 0.05). EH11516 had higher average CpG methylation levels in all those regions. On the other hand, CHG methylation levels were significantly higher (p < 0.01) in EH217 than EH1516 among all regions, while introns had the highest and up2000 had the lowest average values.

The analyses illustrated significant differences in methylation levels between the EH1516 and EH217 strains, especially in the up2000, down2000, and promoter regions. Such significant variations in methylation within those functional elements might explain some of the gene expression and phenotypical differences between the strains.

## Differentially methylated cytosines (DMCs) and regions (DMRs)

Differentially methylated cytosines (DMC) had statistically different methylation levels in the two different strains. Analysis using MOABS[17] (version 1.3.4) detected 303,106 DMCs in the CpG context, about 2.2% of all CpG methylated cytosines; and 200,872 DMCs in the CHG context, about 1.69% of all CHG methylated cytosines. Among the 303,106 CpG DMCs, 185,182 DMCs had lower levels (hypo-methylated) while 117,924 DMCS had higher levels (hyper-methylated) in EH217 compared to EH1516. More DMCs with higher methylation levels in EH1516 were hypothesized to be responsible for reduced expression of associated genes, including genes related to the calcification. In the CHG context, The DMC set included 135,246 hyper-methylated and 65,626 hypo-methylated DMCs in EH217 over EH1516, due to the fact that EH217 had significantly higher mean CHG methylation level than EH1516.

Figure 6 showed the distributions of DMCs varied significantly from scaffolds to scaffolds. The average number of CpG DMCs per 1000bp in the 100 longest *E. huxleyi* scaffolds ranged from 0.98 to 8.45, with a mean of 2.2 and standard deviation of 1.3; that of CHG DMCs per 1000 bp ranged from 0.8 to 3.68, with a mean of 1.42 and small standard deviation of 0.41. The frequencies of CpG and CHG DMCs appeared to be highly correlated with correlation coefficient of 0.85 (p < 2.2e-16).

The distributions of DMCs relative to various E. huxleyi genomic features are shown in Figure 7. Approximately 46% of the CpG DMCs fell in the gene regions, with 33% in exons and 12% in introns.About 44% of the CHG DMCs were located in the gene regions, with 30% in exons and 15% in introns. The relative DMC rate in Figure 7(b) showed the relative frequency of DMCs appearing in feature regions, normalized by the feature lengths. Exons had the highest relative DMC rate of about 1.3 in the CpG and 1.15 in the CHG contexts. The DMCs between EH217 and EH1516 were slightly more concentrated in the gene and exon regions.

The differentially methylated genes (DMGs) overlapped DMRs in their gene regions. 3780 *E. huxleyi* DMGs were identified (Table S1), intersecting about 49% of the 17,185 DMRs in both CpG and CHG contexts. Differential methylations in the promoter regions upstream from TSS affect transcriptional expression. We analyzed the DMRs in the up2000 segments from gene TSS and identified 2,341 differentially methylated promoters (DMPs) associated with *E. huxleyi* genes (Table S?).

Enriched KEGG pathway analysis using KOBAS [18] showed that DMGs were involved in beta-Alanine metabolism and Fatty acid metabolism, while DMPs in Amino sugar metabolism and Vitamin B6 metabolism. Enriched GO analysis showed that DMGs were enriched in GO biological processes such as negative regulation of oxidoreductase activity, peptidyl-amino acid modification, lipid homeostasis, chromatin assembly, cilium organization and DNA conformation change. Furthermore, DMGs were mainly enriched in GO cellular components Golgi-associated retrograde protein (GARP) complex, cilium, DNA packaging complex, and Golgi apparatus. Since calcification in coccolithophores takes place in a Golgi-derived intracellular vesicle [19], the enrichment results suggested that methylation would have impacted the calcification process of *E. huxleyi*.

## Differentially expressed genes (DEGs) between *E. huxleyi* strains

We generated RNA-seq data from three biologically replicated samples of each of the two *E. huxleyi* strains and mapped to the *E. huxleyi* reference genome. Differential gene expression analysis found 14,525 genes that were differentially expressed (DE) between strain EH217 and EH1516 under an FDR threshold of 0.05, out of which 7,160 genes were over-expressed in strain EH217 versus EH1516 and 7,365 were under-expressed. Those DE genes provided a good-sized dataset for studying the associations of DNA methylation and gene expression changes in the *E. huxleyi* strains.

Although strains EH217 and EH1516 were isogeneic, many genes exhibited differential expression levels in the samples, resulting in their phenotypic changes. Figure 13 shows the correlations of gene expression profiles of the two strains between the six samples. The within-the-strain average correlation between EH217 samples was 0.977 and between EH1516 samples was 0.944, significantly higher than the across-the-strains mean correlation of 0.806. PCA plot in Figure 13 (b) showed that samples of the two strains were well separated along the first principal axis of the normalized gene expression profiles.

found 12,711differentially expressed genes (DEGs) between strain EH217 and EH1516 under an FDR threshold of 0.05, out of which 6,234 genes were up-regulated and 6477 were down-regulated in EH217 versus EH1516.

KEGG analysis identified one enriched pathway “Glycosaminoglycan degradation” for DEGs up-regulated in EH217, with p-value < 0.05, probably due to incomplete annotation of KEGG pathways for *E. huxleyi*. DEGs down-regulated in EH217 had 11 enriched KEGG pathways with p-value < 0.05: The two most significantly enriched pathways, Steroid biosynthesis (ehx00100) and AGE-RAGE signaling pathway in diabetic complications (ehx04933), had rich factor over 50% and were probably responding to growing stresses in the EH1516 strain.

GO enrichment analysis showed that the up-regulated and down-regulated genes had intriguingly different sets of enriched GO terms. GO:0030286 (CC: dynein complex) is the most significantly over-represented GO category for the genes up-regulated in EH217, with 40 identified out of 68 genes annotated with this GO category and a fairly low p-value ~ 1.6×10^-15. Dynein is a family of cytoskeletal motor proteins that move along microtubes in cells. Independent knockdown of subunits showed that dynein regulatory complex was required and cilia-driven flow is a key epigenetic factor in controlling otolith biomineralization in vertebrates [20]. The over-represented DEGs associated with dynein complex and cilium gave a plausible explanation of the biological mechanism for the phenotypic differences in coccolith-calcification between the EH217 and EH1516 strains. The EH1516 strain was non-calcifying because of epigenetic changes resulted in the underexpression of genes in dynein complex and cilium and microtubule-based processes.

The most significantly over-represented GO terms for down-regulated DEGs were related to actin filament (CC), actin binding (MF), and actin filament-based process (BP). Those actin filament processes become more active in EH1516, probably responding to the stress of missing coccolith shells.Co-expression analysis of DEGs between different strains and growth conditions

## Correlations between differential methylations and expressions

In this section, we examined how the differential methylations cytosines (DMC) and regions (DMR) related to the differentially expressed genes (DEG). Particularly we compared the frequencies of DMCs in the promoter regions of DEGs vs non-DEGs. Also, we can compare the number of DMRs overlapped with DEG promoters vs non-DEG promoters.

We compared the number of DMCs in the promoter regions of DEGs vs. non-DEGs. In the CpG context, DEGs had an average of 1.859 DMCs per 1000 bp, significantly higher than that of 1.707 for non-DEGs (p = 2.744e-06). Comparable results existed for DMCs in the gene regions, in which DEGs averaged 2.161 DMCs per 1000 bp compared to 1.943 in non-DEGs (p = 1.57e-06). When DMRs were concerned, the DEGs had a higher number of DMRs overlapped with their promoters and gene regions than on-DEGs, averaging 0.152 against 0.128 in promoters (p = 0.0037) and 0.197 against 0.151 in gene regions (p = 0.00024). This analysis demonstrated the DEGs had significantly more methylation changes than non-DEGs, indicating a strong correlation between the methylation and the expression of genes.

However, in the CHG context, DEGs and non-DEGs had close average number of 1.08 vs. 1.09 DMCs per 1000 bp (p = 0.531) in their promoter regions, and 1.22 vs. 1.21 (p = 0.518) in gene regions, respectively. This surprising result indicated that DMCs in the CHG context might not be closely correlated with the gene expression, even though EH217 had significantly higher CHG methylation levels than EH1516.

Furthermore, we compared the number of DMCs in the promoter regions of up-regulated and down-regulated DEGs in EH217 over EH1516. We found that up-regulated DEGs had an average of 2.005 CpG DMCs per 1000 bp, much higher than 1.718 of down-regulated DEGs (p = 1.235e-07). As shown in Figure 12, up-regulated DEGs had a much higher number of DMCs than down-regulated DEGs and non-DEGs, especially in the gene regions. It supported the hypothesis that DNA methylation plays a vital role in regulating the genes over-expressed in EH217, including those responsible for calcification. We then looked more closely at the hyper and hypo-methylated DMCs in the promoter regions. For DMCs hypo-methylated in EH217, up-regulated DEGs had an average of 1.211 DMCs per 1000 bp, much higher than 1.041 in down-regulated DEGs (p = 2.899e-07). Up-regulated DEGs also had more hyper-methylated DMCs of 0.794 than down-regulated DEGs’ 0.677 per 1000 bp in average (p = 2.723e-05). It indicated there was a complex relationship between methylation changes and expression levels, and both hypo and hyper-methylated DMCs might be important for the higher expression of biomineralization-related genes in EH217.

## Differentially methylated regions (DMRs) related to calcification genes

In this section, we will study in more detail about the DMRs for the candidate calcification genes and showcase ones that stand out.

Differentially methylated and expressed genes (DMEGs) are the overlap of DEGs and DMGs, i.e., the genes differentially expressed between EH217 and EH1516 and overlapped with at least one DMR. Total 1389 DMEGs were identified, among which 712 were up-regulated in EH217 and 677 were down-regulated. Enrichment analysis of the 712 up-regulated DMEGs found that “cilium” (GO:0005929: CC) was still the most over-represented (p = 7.3e-6) GO term and “microtubule-based movement” (GO:0007018: BP) is the most over-represented (p = 4.2e-5) GO term for biological process.

## Co-expression Analysis

The set of DEGs between strain EH217 and EH1516 certainly contained many genes belonging to other biological processes than calcification. To investigate the regulation mechanism of the biomineralization process in *E. huxleyi*, we compared the gene expression changes between the calcifying strain EH217 and non-calcifying strain EH1516 versus the changes of the calcifying strain EH217 under the normal growth conditions 9 mM Ca2+ (normal condition) and 9 mM Ca2+ spiked on day 6 with 20 mM NaHCO3- (spiked condition). While calcification was suppressed in the strain EH1516 vs EH217, E. huxleyi grown under spiked condition exhibited much greater calcifications than normal.

RNA-seq data were generated from replicated samples of strain EH217 under the growth conditions 9 and 9S, and 8,528 DEGs were identified between them, using an FDR threshold of 0.05. The intersection of the 12,711 DEGs between strains EH217 and EH1516 and the 8,528 DEGs between growth conditions 9S and 9 resulted in 3,727 overlapped genes differentially expressed in both cases. The Pearson correlation was approximately 0.35 between LFC of EH217 over EH1516, and condition 9S over 9 for those 3,727 intersected DE genes. Most genes in the overlapped set had the same direction of expression change, either up-regulated or down-regulated in both EH217 and 9S, as shown in Table 2.

The 1,532 genes up-regulated in both EH217 and 9S might be of particular relevance to the biomineralization process because of their elevated expressions in the samples associated with stronger calcifications. However, we shouldn’t disregard other DEGs for studying the biomineralization because the mechanism suppressing the calcification in EH1516 might be different from that promoting it in EH217 under the 9S condition. In fact, the highly enriched DEGs over-expressed in EH217 and associated with dynein complex and cilium didn’t appear in the DEGs between the spiked and normal conditions. This suggested that at least some of the biological processes silencing biomineralization in EH1516 differed from those increasing calcifications under the “spiked” condition. On the other hand, they may share DEGs and biological processes related to calcification in *E. huxleyi*. Figure 14 showed the top enriched pathways among these genes, based on the enrich analysis by KOBAS [18].

GO term enrichment analysis identified the following top enriched GO terms in the annotations of this set of genes, as shown in Figure 15.

# Discussion

# Conclusion

# Materials and Methods

## Sample Preparations and Data Generation

*E. huxleyi* strain CCMP1516 was obtained from the Bigelow National Center for Marine Algae and Microbiota and M217 from the Plymouth Algal Collection (Plymouth, UK). Both strains are subclones of the same sample isolated from the South Pacific (02.6667S 82.767W) and considered to be isogenenic. The originally calcifying CCMP1516 lost its ability to calcify over time, but M217 stayed a strong calcifier. Both strains were grown under ???. RNA and DNA samples were extracted at ??? growth stage, with three biological replicates for each strain. The samples were shipped to and sequenced by BGI.

After sequencing, adaptors, contaminations and low-quality reads were filter from raw reads to produce the RNA-seq and WGBS read libraries used in this study. Supplemental Table 1 and 2 show the statistics of the RNA-seq and WGBS libraries, respectively. Each of the six RNA-Seq libraries contained 49 to 54 million clean reads with a length of 100 bp, representing over 50-fold coverage of the *E. huxleyi* transcriptome. Each WGBS library had 38.1 - 41.6 million clean reads with a size of 150bp, offering approximately 34 - 37 fold coverage of the reference *E. huxleyi* genome [4]. The data have been deposited in the NCBI Sequence Read Archive (SRA) (BioProject: PRJNA894788).

## Differential Expression Analysis

The RNA-Seq datasets were mapped to the *E. huxleyi* reference genome using the STAR aligner [21], and a gene expression count matrix was computed according to the numbers of reads mapped to the annotated gene regions using the GenomicAlignments [22] R package. The mapping of the RNA-seq libraries achieved very high efficiencies between 82% to 94%, as shown in Supplemental Table 3. We used DESeq2 [23] and edgeR [24] R packages to identify the differentially expressed genes (DEGs), based on the expression count matrix. The false discovery rate (FDR) < 0.05 was used as the threshold for determining DEGs. Under this threshold, we found 12,711 DEGs between EH217 and EH1516, accounting for about 33.4% of all predicted *E. huxleyi* genes, out of which 6,234 genes were up-regulated and 6477 were down-regulated in EH217.

Enrichment analysis of KEGG pathway for the DEGs were performed using KOBAS [18]. GO enrichment analysis was based on the GO term annotations of *E. huxleyi* genes using Trinotate [25], with the entire set of *E. huxleyi* genes as the background.

## Differential Methylation Analysis

The *E. huxleyi* reference genome has a high CG content of 65.7%, and contains nearly 18 million CpG sites. Bismark (v0.18.0) [26] was used to map the WGBS libraries to the *E. huxleyi* genome, and extract the methylation results for individual cytosines. Table 1 shows the mapping statistics of the bisulfite sequences. The sample “EH1516A” was discarded from further analyses because of its extremely low mapping efficiency compared to other samples, probably due to contaminations. The remaining five samples had mapping efficiency ranged from 45.3% to 52.3%, with a mean rate of 49.74% and high coverages of the genome over 73%, indicating the high reliability and accuracy of the data.

The mapping results showed that 11,966,862 CpG sites had a coverage of at least 3, and 8,219,026 of them have a minimum coverage of 10 in all samples. Only those sites with min coverage of three in all samples were considered in the differential methylation analysis. We investigated different methods to determine DMCs and DMRs on the genome between EH217 and EH1516. As shown in Supplemental Figure 1, the distributions of the *E. huxleyi* methylation rates had two peaks at the extremes near 0% and 100%. An overly stringent DMC criterion was to choose cytosine sites whose methylation levels were <10% (unmethylated) in one strain but > 90% (fully methylated) in the other. This analysis identified 109,847 CpG sites, with 38,871 fully methylated in EH217 but unmethylated in EH1516 and 70,976 fully methylated in EH1516 but unmethylated in EH217. EH1516 had significantly more fully methylated CpG sites than EH217. Another analysis utilized methylKit[27] and found 263,293 DMCs in the CpG context, with a threshold of at least a 25% difference in methylation rate and a q-value of 0.01. MOABS [17] applied a Beta-Binomial hierarchical model to increase the statistical power of detecting differential methylations at individual cytosines. Using the Bismark mapping outputs and default MOABS parameters, we identified 303,106 DMCs in the CpG context, with 185,182 hypo-methylated (lower methylation levels) and 117,924 hyper-methylated (higher methylation levels) in EH217 compared to EH1516.

Differentially methylated regions (DMRs) are genomic segments with significantly different methylation patterns between the strains. In practice, DMRs were often defined as regions that contained a sufficient number DMCs within some max distances, e.g., at least three DMCs with distances between consecutive ones below 300 bp. We found that different software packages for methylation analysis, e.g., DSS [28] and MOABS, often applied various measures and produced vastly different DMRs. Manual inspections found the MOABS results to be the most consistent and sensitive. Therefore, we used the DMCs and DMRs results from MOABS for further analysis.

The analyses using MOABS [17] (version 1.3.4) identified 17,185 total DMRs, including 12,123 DMRs in the CpG context with sizes ranging from 6 to 2243bp and a mean length of 56bp, and 5,062 DMRs in the CHG context with a mean length of 43bp. In the CpG context, 5,252 DMRs were hypermethylated and 6,871 DMRs were hypomethylated in EH217 compared to EH1516; in the CHG context, 3,637 DMRs were hypermethylated and 1,425 were hypomethylated in EH217. Supplemental Figure 2 shows an example DMR and its different methylation patterns among the samples.

ferentially methylated sites (DMCs) were found, abut 2.7% of all highly covered sites. In the CHG context, 4,427,318 sites have at least a coverage of 10 in all samples. The CHG methylation profiles showed even more significant variations between EH1516 and EH217 strains, with cross-strain correlations as low as 0.85. 103,690 DMCs were found in the CHG context, approximately 2.34% of high-covered CHG sites.

We found that often produced vastly varied DMRs, and the results using MOABS were most consistent according to manual inspections.

MOABS was used to identify differentially methylated regions (DMRs).

## Co-expression Analysis

Additional RNA-Seq data were generated for the calcifying EH217 strain under different growth conditions with 9 mM Ca2+ (condition 9) versus 9 mM Ca2+ spiked on day 6 with 20 mM NaHCO3- (condition 9S), with two biological replicates for each condition. The clean reads of each replicate had >30 fold coverage of the *E. huxleyi* transcriptome. The data have been deposited in the NCBI Sequence Read Archive (SRA) with an accession number of ???. The addition of NaHCO3 spike under the 9S condition significantly increased the calcification rate of *E. huxleyi*. The proposition was that calcification related genes would be up-regulated in condition 9S vs. condition 9 as well as in EH217 strain vs. EH1516 strain.

# References

1. Abedin M, Tintut Y, Demer LL. Vascular calcification: mechanisms and clinical ramifications. Arterioscler Thromb Vasc Biol. 2004;24: 1161–1170.

2. Giachelli CM. Inducers and inhibitors of biomineralization: lessons from pathological calcification. Orthod Craniofac Res. 2005;8: 229–231.

3. Gordon R, Parkinson J. Potential roles for diatomists in nanotechnology. J Nanosci Nanotechnol. 2005;5: 35–40.

4. Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, et al. Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature. 2013;499: 209–213.

5. Phillips T. The Role of Methylation in Gene Expression. Nature Education. 2008;1: 116.

6. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet. 2008;9: 465–476. doi:10.1038/nrg2341

7. Smith ZD, Meissner A. DNA methylation: Roles in mammalian development. Nature Reviews Genetics. Nature Publishing Group; 2013. pp. 204–220. doi:10.1038/nrg3354

8. Jin Z, Liu Y. DNA methylation in human diseases. Genes and Diseases. Chongqing Medical University.; 2018. pp. 1–8. doi:10.1016/j.gendis.2018.01.002

9. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, et al. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A. 2010;107: 8689–8694. doi:10.1073/pnas.1002720107

10. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science (1979). 2010;328: 916–919. doi:10.1126/science.1186366

11. VanderKraats ND, Hiken JF, Decker KF, Edwards JR. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. Nucleic Acids Res. 2013;41: 6816–6827. doi:10.1093/nar/gkt482

12. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523: 212–216. doi:10.1038/nature14465

13. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. 2014;15: R37. doi:10.1186/gb-2014-15-2-r37

14. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012;13: 705–719. doi:10.1038/nrg3273

15. Schlosberg CE, VanderKraats ND, Edwards JR. Modeling complex patterns of differential DNA methylation that associate with gene expression changes. Nucleic Acids Res. 2017;45: 5100–5111. doi:10.1093/nar/gkx078

16. Tang B, Zhou Y, Wang C-M, Huang TH-M, Jin VX. Integration of DNA methylation and gene transcription across nineteen cell types reveals cell type-specific and genomic region-dependent regulatory patterns. Sci Rep. 2017;7: 3626. doi:10.1038/s41598-017-03837-z

17. Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, Meong M, et al. MOABS: model based analysis of bisulfite sequencing data. Genome Biol. 2014;15: R38. doi:10.1186/gb-2014-15-2-r38

18. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011;39: W316–W322. doi:10.1093/nar/gkr483

19. Sciences N. The Mechanisms of Calcification in Coccolithophores. 2012.

20. Colantonio JR, Vermot J, Wu D, Langenbacher AD, Fraser S, Chen JN, et al. The dynein regulatory complex is required for ciliary motility and otolith biogenesis in the inner ear. Nature. 2009;457: 205–209. doi:10.1038/NATURE07520

21. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635

22. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. Prlic A, editor. PLoS Comput Biol. 2013;9: e1003118. doi:10.1371/journal.pcbi.1003118

23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-014-0550-8

24. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26: 139–40. doi:10.1093/bioinformatics/btp616

25. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. Gibas C, editor. PLoS One. 2011;6: e21800. doi:10.1371/journal.pone.0021800

26. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27: 1571–2. doi:10.1093/bioinformatics/btr167

27. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012;13: 87. doi:10.1186/gb-2012-13-10-R87

28. Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics. 2016;32: 1446–1453. doi:10.1093/bioinformatics/btw026

Table . Mapping statistics of bisulfite sequences to the E. huxleyi genome.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sample Name | Mapping efficiency | Genome coverage | Bisulfite Conversion Ratio | CpG methylation | CHG methylation | CHH methylation |
| EH1516A | 19.1% | 41.6% | 97.9% | 9.0% | 4.8% | 0.5% |
| EH1516B | 48.2% | 73.9% | 97.9% | 16.3% | 5.4% | 0.5% |
| EH1516C | 45.3% | 73.3% | 98.1% | 15.8% | 5.1% | 0.4% |
| EH217A | 51.7% | 74.1% | 97.6% | 15.8% | 6.0% | 0.6% |
| EH217B | 52.3% | 74.9% | 97.7% | 15.9% | 5.9% | 0.6% |
| EH217C | 51.2% | 73.7% | 97.7% | 15.7% | 5.8% | 0.5% |

Table . The DEGs intersected divided into four groups according to their relative expression changes in compared samples. The table shows the numbers of genes in each group.

|  |  |  |
| --- | --- | --- |
|  | Up-regulated in EH217 | Down-regulated in EH217 |
| Up-regulated in 9S | 1,532 | 657 |
| Down-regulated in 9S | 213 | 1,325 |

Table . List of genes in DEMGs with “cilium” (GO:0005929: CC) annotation

|  |  |  |
| --- | --- | --- |
|  | log2FC | Description |
| 106349 | 1.909857 | membrane attack complex perforin domain-containing |
| 117244 | 1.950292 | cytoplasmic dynein 2 heavy chain 1 isoform X1 |
| 122188 | 5.729132 | cilia- and flagella-associated 61 |
| 194332 | 7.210228 | cilia- and flagella-associated 70 |
| 198139 | 6.098505 | Dynein heavy chain partial |
| 198190 | 5.422113 | central pair associated wd-repeat |
| 207330 | 7.682979 | dyslexia susceptibility 1 candidate gene 1 homolog |
| 211505 | 6.698345 | flagellar outer dynein arm heavy chain beta |
| 212369 | 2.16351 | hypothetical protein EMIHUDRAFT\_212369 |
| 216525 | 7.940246 | monoglycylase TTLL8 |
| 219951 | 7.99493 | myosin heavy |
| 223628 | 7.676854 | fibrocystin-l |
| 229544 | 5.850834 | dynein intermediate chain axonemal |
| 230030 | 6.845327 | Dynein heavy chain axonemal |
| 243453 | 7.289079 | FAM179B isoform X1 |
| 252292 | 2.002519 | TRAF3-interacting 1 |
| 438199 | 1.924581 | Tubulin-tyrosine ligase family |
| 63754 | 6.800667 | Dynein regulatory complex 1 |
| 66449 | 7.778062 | Cilia- and flagella-associated 58 |
| 77679 | 7.837322 | intraflagellar transport 57 homolog |

Figure . Comparison of overall cytosine methylation levels across the *E. huxleyi* samples.

|  |  |
| --- | --- |
| Chart, line chart, scatter chart  Description automatically generated | Chart, histogram  Description automatically generated |

Figure . The average methylation levels in the CpG (left) and CHG (right) contexts of the 100 longest scaffolds of the E. huxleyi reference genome. The blue dots showed the differences in methylation levels between the EH217 strain vs. the EH1516 strain.

|  |  |
| --- | --- |
| Chart  Description automatically generated | Chart, scatter chart  Description automatically generated |
| (a) | (b) |
| Chart, box and whisker chart  Description automatically generated | Chart, scatter chart  Description automatically generated |
| (c) | (d) |

Figure 3. (a) Correlations and (b) PCA plot of *E. huxleyi* methylation profiles in the CpG context; (c) and (d) in the CHG context.

|  |  |
| --- | --- |
| Chart  Description automatically generated | Chart  Description automatically generated |
| Chart  Description automatically generated | Chart  Description automatically generated |

Figure . Methylation levels of strain EH1516 and EH217 in CpG and CHG contexts within 10-kb windows across the scaffold 1 and scaffold 14 of the genome.

|  |  |
| --- | --- |
| Chart  Description automatically generated with medium confidence | Chart, scatter chart  Description automatically generated |

Figure . The CpG and CHG methylation level distributions by region. The error bars showed the 95% confidence interval around the mean methylation level.

Graphical user interface, chart, line chart

Description automatically generated

Figure . Average rates of DMCs per 1000 bp in the first 100 longest scaffolds.

|  |
| --- |
| Pie chart  Description automatically generated |
| (a) |
| Chart, line chart  Description automatically generated |
| (b) |

Figure . (a) The percentages of DMCs in the CpG context overlapped with various feature regions. (b) The relative rate of DMCs in feature regions, normalized by feature sizes.

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated | Chart, scatter chart  Description automatically generated |
| 1. DMGs | 1. DMPs |

Figure . The KEGG pathways enriched in genes up-regulated in (A) DMGs and (B)DMPs. The rich factor reflects the proportion of up-regulated genes in a given pathway. The number of up-regulated genes in the pathway is indicated by the circle area, and the circle color represents the ranges of the corrected P-value.

|  |  |
| --- | --- |
| Chart, bar chart  Description automatically generated | Chart, bar chart  Description automatically generated |
| (A) DMGs | (B) DMPs |
| Diagram, bubble chart  Description automatically generated | Diagram  Description automatically generated |
| 1. DMGs enriched BP GO terms | 1. DMPs enriched BP GO terms |
| Chart, scatter chart, bubble chart  Description automatically generated | Chart, scatter chart, bubble chart  Description automatically generated |
| 1. DMGs enriched CC GO terms | 1. DMPs enriched CC GO terms |

Figure . GO enrichment analysis of DMGs (A) and DMPs (B) showed the top enriched GO subcategories (p-value < 0.01), where the gene numbers were shown on the bar top.

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated | Chart, scatter chart  Description automatically generated |
| 1. Genes up-regulated in EH217 | 1. Genes down-regulated in EH217 |

Figure . The KEGG pathways enriched in DEGs up-regulated (A) or down-regulated (B) in strain EH217 vs. EH1516. The rich factor reflects the proportion of DEGs in a given pathway. The number of DEGs in the pathway is indicated by the circle area, and the circle color represents the ranges of the corrected p-value.

|  |  |
| --- | --- |
| Chart, bar chart  Description automatically generated | Chart  Description automatically generated |
| (A) Enriched GO terms in genes up-regulated in EH217 | (B) Enriched GO terms in genes down-regulated in EH217 |
| Chart, scatter chart, bubble chart  Description automatically generated | Chart, bubble chart  Description automatically generated |
| Enriched CC GO categories in genes up-regulated in EH217 | Enriched CC GO categories in genes down-regulated in EH217 |
| Chart, bubble chart  Description automatically generated | Diagram  Description automatically generated |
| Enriched BP GO categories in genes up-regulated in EH217 | Enriched BP GO categories in genes down-regulated in EH217 |

Figure . Enriched GO term subcategories (over-represented p-value < 0.001) of genes up-regulated (A) and down-regulated (B) in strain EH217 vs. EH1516.

Chart, box and whisker chart

Description automatically generated

Figure . Comparison of CpG DMC count per 1000 bp between up-regulated, down-regulated DEGs, and non-DEGs in different feature regions. The error bars showed the 99% confidence interval around the mean.

|  |  |
| --- | --- |
| Chart  Description automatically generated | Chart, histogram, scatter chart  Description automatically generated |
| (a) | (b) |

Figure . (a) Correlations and (b) PCA plot of normalized gene expression profiles of the six E. huxleyi samples of strain E1516 and E217.

Chart, scatter chart

Description automatically generated

Figure . The KEGG pathways enriched in genes up-regulated in both strain EH217 vs. EH1516 and condition 9S vs. 9. The rich factor reflects the proportion of up-regulated genes in a given pathway. The number of up-regulated genes in the pathway is indicated by the circle area, and the circle color represents the ranges of the corrected P-value.

Chart, bar chart

Description automatically generatedChart, bubble chart

Description automatically generated

Figure . The left picture shows the top GO terms enriched (p-value < 0.001) in genes up-regulated in both strain EH217 vs. EH1516 and condition 9S vs. 9. Circles in closer proximity in the right picture represented biological process (BP) GO terms that are more closely related, analyzed by REVIGO[25]. The size of the circle indicates the frequency of the GO term in the underlying GOA database (bubbles of more general terms are larger). The color of the circle represents the statistical significance of the enriched GO terms based on the over-represented p-value;

Supplemental Table . Statistics of th six WGBS libraris, with three biologically replicated samples of strain EH1516 and EH217.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample Name | Clean Reads | Clean bases | Read length(bp) | Q20 (%) | GC (%) |
| EH1516A | 40,161,824 | 6,024,273,600 | 150 | 97.37% | 31.33% |
| EH1516B | 41,579,220 | 6,236,883,000 | 150 | 96.92% | 34.48% |
| EH1516C | 38,142,830 | 5,721,424,500 | 150 | 98.81% | 33.74% |
| EH217A | 40,114,082 | 6,017,112,300 | 150 | 97.21% | 34.53% |
| EH217B | 41,574,136 | 6,236,120,400 | 150 | 96.82% | 34.80% |
| EH217C | 41,648,326 | 6,247,248,900 | 150 | 96.86% | 34.25% |

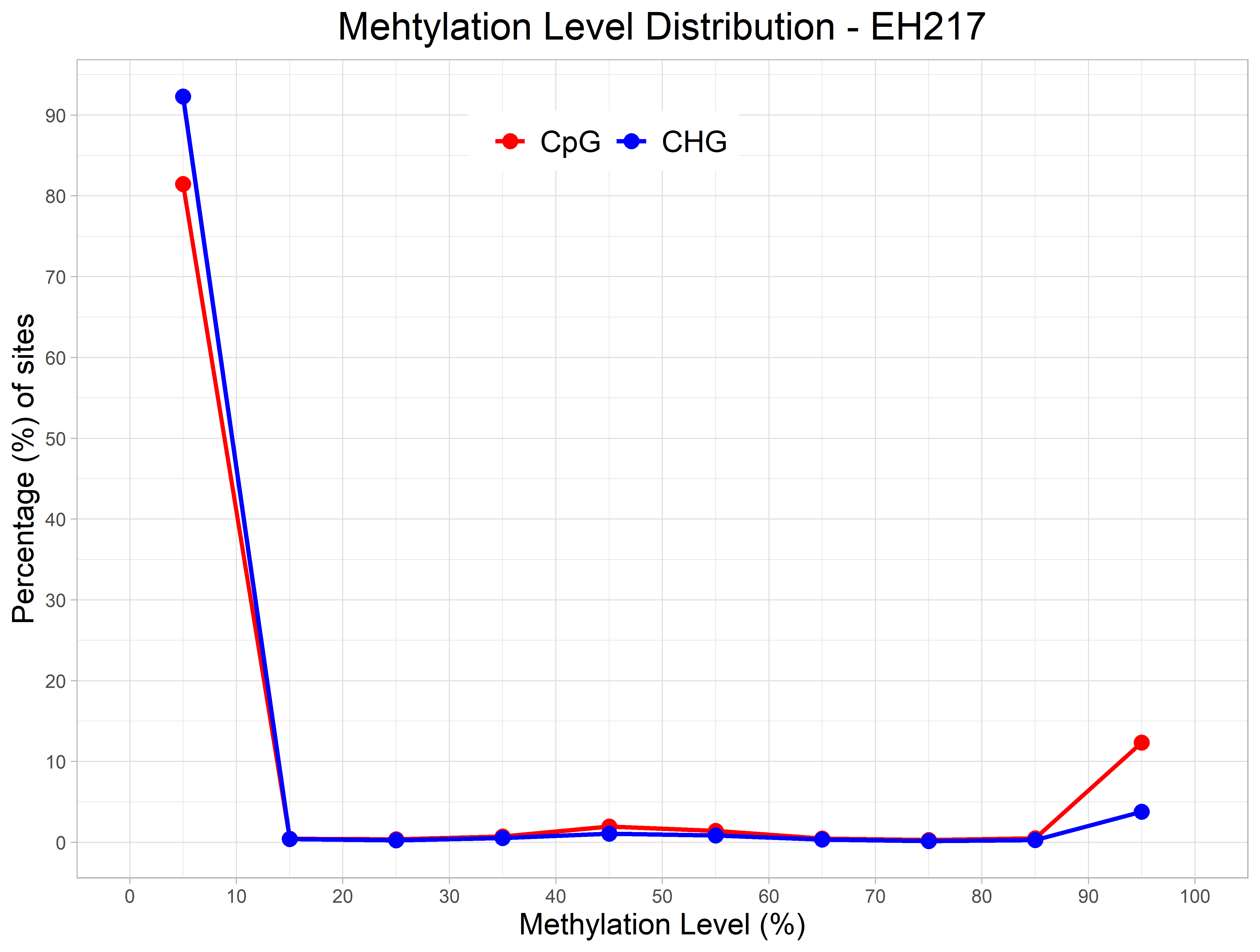
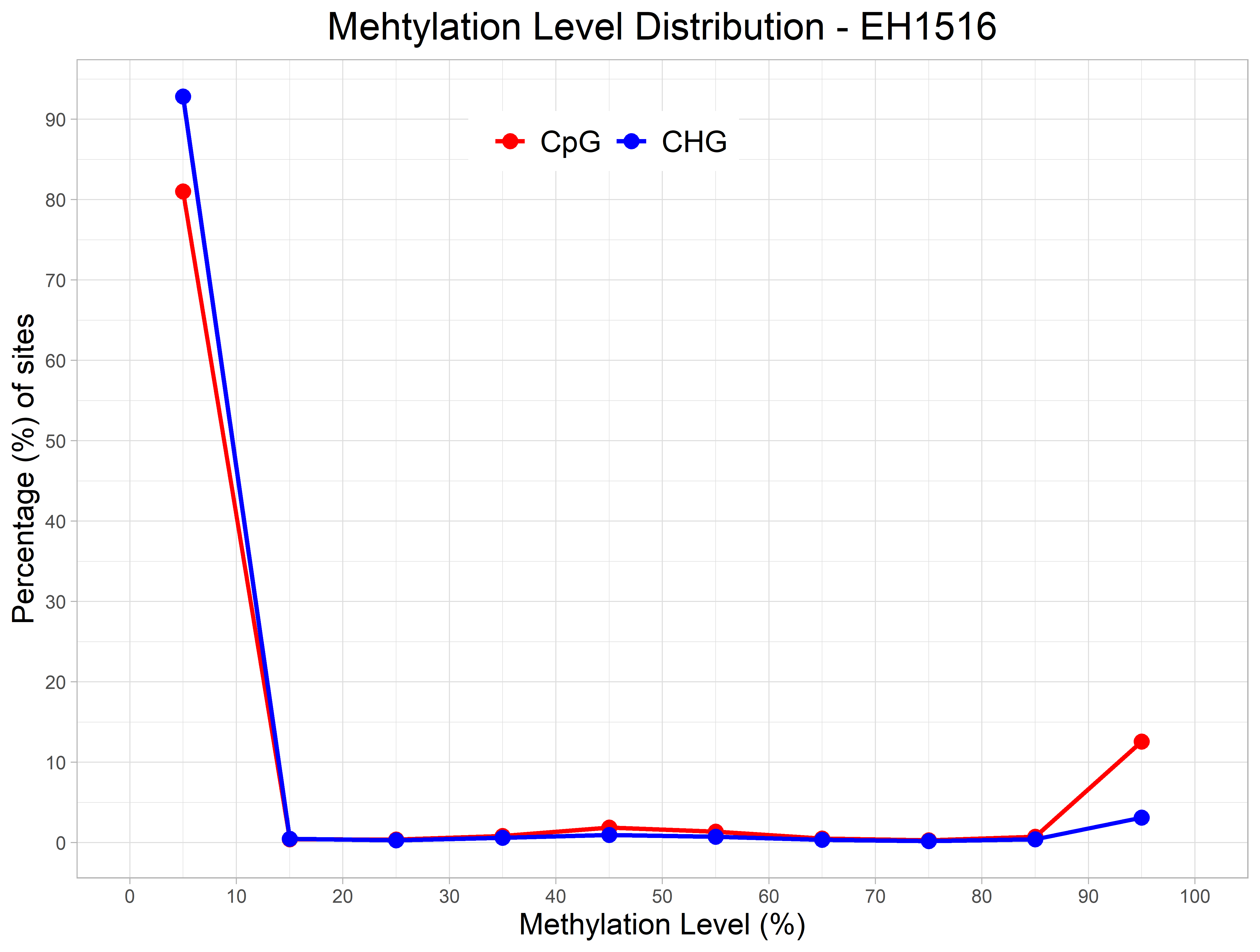
Supplemental Table . Statistics of the six RNA-Seq libraries, with three biologically replicated samples of strain EH1516 and EH217.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample Name | Clean Reads | Clean bases | Read length (bp) | Q20(%) | GC (%) |
| EH1516A | 49,237,758 | 4,923,775,800 | 100 | 95.24% | 66.11% |
| EH1516B | 54,073,336 | 5,407,333,600 | 100 | 94.92% | 67.43% |
| EH1516C | 53,909,048 | 5,390,904,800 | 100 | 94.80% | 67.72% |
| EH217A | 49,297,440 | 4,929,744,000 | 100 | 95.04% | 67.44% |
| EH217B | 51,206,106 | 5,120,610,600 | 100 | 94.86% | 67.51% |
| EH217C | 51,710,230 | 5,171,023,000 | 100 | 94.87% | 67.49% |

Supplemental Table . Mapping statistics of RNA-Seq datasets to the E. huxleyi reference genome.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample Name | % Uniquely mapped | % mapped to multiple loci | % unmapped |
| EH1516A | 56.25% | 25.24% | 17.71% |
| EH1516B | 62.00% | 30.53% | 6.14% |
| EH1516C | 61.70% | 30.54% | 5.96% |
| EH217A | 58.56% | 28.50% | 11.90% |
| EH217B | 57.38% | 28.77% | 13.13% |
| EH217C | 57.42% | 28.75% | 12.39% |

Supplemental Figure . Distribution of Cytosine sites according to their methylation levels under different contexts. Only cytosine sites with coverage of at least 10 were used.



Timeline

Description automatically generated

Supplemental Figure . An example DMR showing the differential methylation patterns between the EH1516 and EH217 strains.

Chart, pie chart

Description automatically generated with medium confidence

Supplemental Figure . The percentages of DMCs in the CHG context overlapped with various feature regions.

|  |  |
| --- | --- |
|  | Chart, box and whisker chart  Description automatically generated |
| (a) | (b) |

Supplemental Figure . Comparison of CpG (a) and CHG (b) DMC count per 1000 bp in different feature regions. The error bars showed the 99% confidence interval around the mean.