EPIGENETIC REGULATORY POTENTIAL

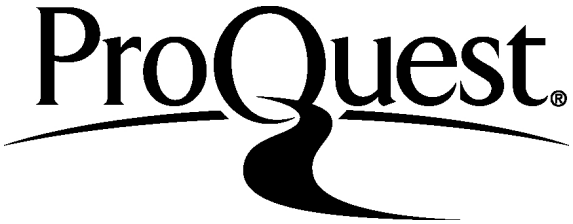
AT THE METAGENOMIC AND GENOMIC SCALE

by

Ian M. Rambo

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree Master of Science in Marine Studies

Spring 2016



EPIGENETIC REGULATORY POTENTIAL

AT THE METAGENOMIC AND GENOMIC SCALE

by

Ian M. Rambo

Approved: _____

Jennifer F. Biddle, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____

Mark A. Moline, Ph.D.
Director of the School of Marine Science and Policy

Approved: _____

Mohsen Badiey, Ph.D.
Dean of the College of Earth, Ocean, and Environment

Approved: _____

Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

I would like to thank Dr. Adam Marsh (College of Earth Ocean and Environment, University of Delaware) for providing access to his methylation profiling platform, Dr. Thomas Hanson (College of Earth Ocean and Environment, University of Delaware) for ion chromatography technical support, Dr. Rovshan Mahmudov (College of Engineering, University of Delaware) for technical support in gas chromatography analysis, Dr. Christopher Sommerfield and Kaitlin Tucker (College of Earth Ocean and Environment, University of Delaware) for radionuclide dating technical support, Kenneth Hoadley for motif analysis scripts, Annamarie Pasqualone for 16S rRNA gene amplicon and Illumina library preparation, Dr. Patrick Gaffney (College of Earth Ocean and Environment, University of Delaware) for help with statistical analysis, Dr. Rosa Leon-Zayas and Glenn Christman for bioinformatics guidance, and the Biddle lab as a whole for your support. Christoph Gohlke wrote the DNA curvature script. Support from the University of Delaware Center for Bioinformatics and Computational Biology Core Facility [and/or use of the BioHen compute cluster] was made possible through funding from Delaware INBRE (NIGMS GM103446), Delaware EPSCoR (NSF EPS-0814251, NSF IIA-1330446), the State of Delaware, and the Delaware Biotechnology Institute. I would also like to thank the Center for Dark Energy Biosphere Investigations (C-DEBI) for funding this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Marine sediments harbor a vast amount of Earth's microbial biomass, yet little is understood regarding how cells subsist in these low-energy environments. Since growth in these environments is expected to be slow because the overall energy pool is low, cells may require additional methods for conserving energy. Gene regulation is a potential necessity due to the high energy requirements of transcription, and this process could be influenced by epigenetic modification via DNA methylation. In this study, changes in the methylation states of CpG sites were profiled within metagenomes from an estuarine sediment core using a next-generation sequencing strategy. Additionally, the presence of epigenetic patterns and conserved sequence structures within gene promoter and coding regions was determined for microbial genomes representative of metagenome target taxa to determine how widespread these signatures may be. The results of this study suggest the presence of dynamic shifts in CpG methylation within these sediment microbial communities, as well as conserved trends in DNA methylation target motif frequency and intrinsic DNA curvature within known bacteria. The analyses of these phenomena further highlight the dynamic roles of epigenetic modifications within microbial genomes.

Chapter 1

COMPOSITION, FUNCTION, AND DNA METHYLATION OF BROADKILL RIVER SEDIMENT COMMUNITIES

1.1 Introduction

Marine sediments are some of the largest reservoirs of microbial biomass on Earth (Whitman et al. 1998; Kallmeyer et al. 2012), and describing the relationships between community structure, activity, and ecosystem function in these habitats remains a challenge (Fuhrman 2009; Orsi et al. 2013). The majority of sedimentary bacteria and archaea are unable to be successfully cultured in a laboratory setting, and if they are able to be cultivated, they likely do not exist in physiological states representative of those found within their natural habitats (Hoehler and Jørgensen 2013). Next-generation sequencing technologies enable researchers to overcome the constraints of cultivation by directly analyzing environmental DNA and RNA. These technologies are employed in subsurface microbiology to provide information regarding community dynamics and ecological roles (Hua et al. 2014), classify rare or uncultured species (Albertsen et al. 2013; Seitz et al. 2016), and describe potential microbial activity (Orsi et al. 2013).

Determining the drivers that govern activity in the subsurface is key to understanding the relationships between these microbes and their environments.

Models suggest that many marine subsurface cells should be sporulated due to the low availability of energy (Lomstein et al. 2012), yet metatranscriptome analysis of anaerobic sediments from the Peru Margin suggests that microbes may not express high levels of genes for spore formation (Orsi et al. 2013). Subseafloor metagenome analysis supports these findings, as endospore-specific genes were not frequent in deep-sea sediment communities off Japan and Peru (Kawai et al. 2015). This suggests that instead of forming spores, cells in the subsurface possibly suspend certain life processes though other means in order to survive, subsisting at low levels of activity, and growing on geological timescales.

A positive relationship between microbial activity and gene or transcript abundance is often seen by previous studies (Muttray et al. 2001; Schippers et al. 2005; DeAngelis et al. 2010; Gaidos et al. 2011; Hunt et al. 2013). The analysis of rRNA is a popular technique for determining growth or activity via gene expression, but one caveat of this method is the presence of DNA and RNA from inactive cells (Blazewicz et al. 2013). Of 415 studies in which abundances of genes or transcripts for carbon and nitrogen cycling enzymes were quantified, only 59 (14%) of these publications provided both abundance and corresponding process rates (Rocca et al. 2014). The equation of abundance with activity is a possible false positive in many studies due to genes or transcripts being present, but not necessarily expressed, within a genome. In the deep biosphere, a large disconnect was seen between genes discovered via metagenomics, and genes detected via metatranscriptomics (Orsi et al.

2013). As such, we are interested in developing methods to examine potential genetic signals that affect expression in the deep biosphere and can possibly allow low cellular activity over geological timescales.

In less extreme sediment environments such as estuarine sediments, it is possible that microbes can regulate life processes to acclimate themselves to less-than-favorable environmental conditions as the sediment ages. Epigenetic regulation is a potential microbial survival strategy within low-energy sediment, allowing for cell maintenance and rapid acclimation to environmental stressors (Bird 2002; Casadesús and Low 2006; Low and Casadesús 2008). While there are many mechanisms for genetic regulation, transcriptional silencing or activation by DNA methylation has been shown to act as a mode of epigenetic regulation in microbes (Kumar and Rao 2012; Wion and Casadesus 2006; Low et al. 2001).

DNA methylation is present in all domains of life (Singal and Ginder 1999; Clouet-d'Orval, Gaspin, and Mougin 2005; Brunet et al. 2011; Casadesús and Low 2006; Heinrichs 2014), and involves the addition of a methyl group via a methyltransferase (MTase) to either the carbon 5 position of a cytosine ring (resulting in 5-methylcytosine (m5C)), the nitrogen 4 position of cytosine (resulting in N4-methylcytosine (m4C)), or the nitrogen 6 position of adenine (resulting in N6-methyladenine (m6A)) (Ratel et al. 2006). While m5C and m6A are found in many bacterial, protist, and fungal genomes, m4C is only known to occur in bacteria (Cheng 1995). The methylated bases of a genome, known as a methylome (Murray et al.

2012), can differ greatly depending on life experiences and life stage (Szyf 2009; Suderman et al. 2012; Gonzalez et al. 2014). Methylation fulfills multiple functional roles within bacterial systems, the most widely known being an association with restriction-modification (RM) systems. Methylation-dependent RM systems are a critical component of bacterial defense mechanisms, as they allow for the identification of self vs. foreign (e.g. viral) DNA based on the presence or lack thereof of methylated bases at target sites (Fang et al. 2012). Most bacterial MTases are associated with RM systems, with MTases being partnered with a respective restriction endonuclease (RE) that cleaves at a non-methylated target sequence (Roberts et al. 2010). However, several MTases such as Dam and CCrM are categorized as 'orphan' MTases due to their lack of an associated RE (Løbner-olesen et al. 2005; Marinus and Casadesus 2009). Like their RE counterparts, methyltransferases also have respective DNA target sequences. Microbes methylate cytosines at 5'-CG-3' (CpG) sequence contexts as is common in eukaryotes (Wojciechowski et al. 2012), but also perform methylation at highly diverse target sequences (e.g. Dam and CCrM methylate the N6 position of the internal adenine of 5'-GATC-3' and 5'-GANTC-3', respectively; Dcm methylates the C5 position of the internal cytosine of 5'-CCWGG-3') (Low et al. 2001).

Aside from providing a means of defense against foreign DNA in microbes, methylation catalyzed by both RM-associated and orphan MTases is known to affect cell processes at the transcriptional level. Methylome analysis of 213 bacterial and 13

archaeal species from 19 different phyla and 37 different classes by Single Molecule, Real-Time (SMRT) sequencing has shown DNA methylation to be highly prevalent and evolutionarily conserved in microbes at diverse target sites, with widespread methylation by orphan MTases, suggesting a role for methylation in gene regulation across many microbial lineages (Blow et al. 2016).

Methylated bases can influence the interactions between regulatory proteins and DNA through direct steric effects (i.e. steric hindrance) (Low et al. 2001) or indirect effects via the alteration of DNA curvature and thermodynamic stability (Diekmann 1987). Both types of effects act as signals for gene-protein interactions and transcriptional regulation (Low and Casadesús 2008). For instance, RNA polymerase and transcription factors are able to differentiate fully-methylated and hemimethylated DNA at promoter regions, and this discrimination of differentially methylated DNA acts as a method for determining which genes are transcribed at specific stages in the cell cycle (Low and Casadesús 2008; Gonzalez et al. 2014; Collier 2009). In *Escherichia coli*, two GATC sequences are present in the upstream regulatory region of the *pap* operon, and the differential methylation of these target sequences (i.e. one methylated and one non-methylated) acts as an "on/off" switch for phase variation between pilus expression and non-expression states (Braaten et al. 1994). The methylation states of target sites in populations of cells do not necessarily fall into binary "methylated" or "non-methylated" classifications, but undergo dynamic shifts (Marsh and Pasqualone 2014) that allow for a stable, efficient, and specific mode of

gene regulation through targeted silencing (Jeltsch et al. 2007). This system of genetic "switches" regulated by DNA methylation could potentially be a viable mechanism for both long and short-term transcriptional silencing for microbes inhabiting the sediments of dynamic environments such as estuaries.

Estuaries are highly productive ecosystems whose microbial communities are responsible for significant geochemical turnover vital to global nutrient cycling (Bauer et al. 2013). Due to their roles as both sinks and sources of atmospheric carbon (Chmura 2013) and nitrogen (Moseman-Valtierra et al. 2011), estuarine environments are highly important for carbon and nitrogen cycle regulation. Dynamic relationships with tidal and riverine flow effects govern biotic and abiotic factors such as sediment geochemistry (Hardison et al. 2011) and deposition (Kemp et al. 2012), and these factors can act as stressors within the microbial communities that populate estuarine sediments.

To better understand if sediment microbes can potentially acclimate to environmental stressors though DNA methylation-induced regulation, I utilized an Illumina sequencing-based assay to identify dynamic shifts in CpG methylation within sediment metagenomes from the Broadkill River estuary system. This assay has been previously used for analyzing CpG methylation at HpaII target sites (5'-CCGG-3') within the Antarctic polychaete *Spiophanes tcherniai* (Marsh and Pasqualone 2014). While this assay has not been used prior for microbial analysis, we opted to utilize it due to the presence of cytosine methylation at CpG sites in microbes and the

anticipation of a heterogeneous, dynamic community. Since adenine methylation is considered to be more widespread than cytosine methylation in microbes (Ratel et al. 2006), this choice of motif potentially serves to reduce signal saturation. To our knowledge, this is the first report on DNA methylation within metagenomic sequence data, and is the first to utilize this method of CpG methylation analysis in an environmental application.

1.2 Materials and Methods

1.2.1 Core collection

Sediment cores were sampled from the Oyster Rocks site of the Broadkill River, Milton, DE, USA (38.802161, -75.20299) at low tide in July 2012 and 2014. The 2012 core was sectioned into 3 cm sections and immediately frozen at -80°C. Three cores were sampled from the same area, ~5 m from the riverbank in 2014: a 32 cm radionuclide dating core (R), and 25 cm (S) and 30 cm (L) cores for DNA extraction and pore water ion chromatography, methane flame ionization gas chromatography, and porosity measurements. Cores L and S were sliced into 3 cm depth samples and immediately frozen at -80 °C, while Core R was immediately processed.

1.2.2 Radionuclide dating

Core R was sectioned into 1 cm thick intervals from 0-10 cm, and 2 cm thick intervals from 10-32 cm. Samples were dried at 60 °C for 48 hours, and transferred to a 25 °C desiccation chamber for storage until further processing. Wet and dry sample weights were recorded and used in subsequent porosity calculations. Dried samples were crushed with a mortar and pestle, and ground into a fine powder with an IKA Werke M20 mill (IKA Werke, Staufen, Germany). Powdered samples were transferred to 60 ml plastic jars and compressed at $3.4x10^3$ kPa with a manual hydraulic press. Radionuclide counting of compressed samples was performed for 24 hours on a Canberra Instruments Low Energy Germanium Detector (Canberra Industries, Meriden, CT, USA). Levels of ^{7}Be ($t_{1/2}$ = 53.22 days), ^{210}Pb ($t_{1/2}$ = 22.20 years), and ^{137}Cs ($t_{1/2}$ = 30.17 years) activity were measured by gamma spectroscopy of the 478, 46.5, and 662 kEV photopeaks, respectively (Igarashi et al. 1998; Cutshall et al. 1983; Wallbrink et al. 2002).

1.2.3 Porosity

Porosity was calculated as:

$$\varphi = \frac{M_w/\rho_w}{\frac{M_w}{\rho_w} + \left(\frac{Md - S * \frac{M_w}{1000}}{\rho_{ds}}\right)}$$

where M_w is the mass of the water lost on drying, M_d is the mass of the dried sediment, ρ_w is the density of pure water (defined as 1 g/cm^3), Pds is the density of dry sediment calculated as $M_d/\pi r^2 h$, where r is the radius and h is the height of the core slice, and S is salinity in grams per kilogram (Lloyd et al. 2011). A constant salinity value of 26 g/kg was used.

1.2.4 Porewater ion chromatography

Porewater was extracted from 50 mL sediment samples by centrifugation at 13,000 G for 30 minutes. Porewater ions were measured with a Metrohm 850 Professional ion chromatograph (Metrohm, Herisau, Switzerland). Dilutions were measured to determine a standard curve. Samples were diluted to ensure signal within the standard curve.

1.2.5 Methane

Methane concentrations were determined for cores L and S as previously described (Biddle et al. 2012). Core subsamples (volume = 305 cm^3) were extracted from each core slice using a 5 ml syringe whose top had been removed with a razor blade. Core subsamples were transferred into 20 mL amber glass vials. 1 mL 1 M NaOH was added to each vial to halt microbial activity. Vials were crimped, shaken, and stored for 10 days at 25°C.

A standard curve was calculated from 500, 1000, and 5000 ppm standards. Methane headspace concentrations were measured via flame ionization gas chromatography using a 5890 Series II gas chromatograph equipped with a flame ionization detector (Hewlett-Packard, Palo Alto, California, USA). 100 µL gas extractions were run in triplicate to determine mean retention times.

Methane concentrations were calculated as:

$$[CH_4] = \frac{P_{CH_4} * V_{headspace}}{RTV_{sed}1000}$$

Where P_{CH_4} is the partial pressure of methane in ppm, $V_{headspace}$ is the volume of the headspace, R is the universal gas constant, T is the temperature in Kelvin, and V_{sed} is the volume of the sediment added to the vial.

1.2.6 Illumina library preparation

Illumina libraries were prepared from the 2012 core sections. Sediment genomic DNA was extracted from 0.5 g of sediment with a MoBio PowerSoil (MoBio, Valencia, CA) kit according to the manufacturer's instructions. Purified gDNA was digested with the methylation-sensitive RE HpaII, which cleaves at the unmodified internal cytosine of a 5'-CCGG-3' motif. Digested DNA was cleaned with a QIAquick PCR purification kit (Qiagen, Hilden, Germany), sheared to a median size of 300 bp using a Covaris focused-ultrasonicator (Covaris, Woburn, MA, USA), and

cleaned again with QIAquick. Digested extracts were immediately transferred to -20°C until Illumina library preparation. Illumina libraries were prepared using the NEBNext Ultra Library Prep Kit for Illumina (New England BioLabs, Ipswich, MA, USA).

Illumina libraries were sequenced with an Illumina Hi-Seq 2500 (Illumina, San Diego, California, USA) at the Delaware Genomics and Biotechnology Institute (Newark, DE, USA). Single-read sequencing was performed for all samples, with 150-cycle sequencing for the 3-6 cm and 12-15 cm samples, and 50-cycle sequencing for the 24-27 cm sample.

1.2.7 16S rRNA gene amplicon sequencing

DNA was extracted, purified, and digested using the previously described method for Illumina libraries. Purified DNA was quantified and tested for successful PCR reactions for the bacterial 16S rRNA gene. 16S rRNA gene amplicon library preparation and Ion Torrent PGM sequencing were performed by Molecular Research, LP (Clearwater, Texas, USA). The V4 variable region 515F/806R (Caporaso et al. 2011) PCR primers were used to perform single-step, 30-cycle PCR using the HotStarTaq Plus Master Mix Kit (Qiagen, USA) under the following conditions: 94°C for 3 minutes, followed by 28 cycles (5 cycle used on PCR products) of 94°C for 30 seconds, 53°C for 40 seconds and 72°C for 1 minute, after which a final elongation step at 72°C for 5 minutes was performed. Libraries were sequenced on an Ion Torrent

PGM (Thermo-Fisher Scientific, USA).16S rRNA gene analysis was performed with QIIME 1.8.0 (Caporaso et al. 2010). Dereplication, abundance sorting, and discarding reads < 2 bp was performed with the USEARCH7 algorithm (Edgar 2013). Chimeras were filtered with UCHIME (Edgar et al. 2011) using the RDP Gold Classifier training database v9 (Cole et al. 2014). Operational taxonomic unit (OTU) picking was performed at 97% similarity with the UCLUST algorithm (Edgar 2010). Non-chimeric sequences were chosen as the representative set of sequences for taxonomic assignment and alignment. Taxonomic assignments were performed with the UCLUST algorithm (Edgar 2010) using the Greengenes V13.8 database for 97% OTUs (DeSantis et al. 2006). OTU tables were rarefied by QIIME from 2000 to 9500 sequences per sample by steps of 100, with 10 iterations performed at each step.

1.2.8 Metagenome assembly and annotation

All metagenome sequence reads were trimmed to 51 bp and quality controlled with a custom Python script by Dr. Adam G. Marsh. Reads with Phred score nucleotide confidence < 95% were removed. Quality-controlled reads were assembled in IDBA (Peng et al. 2010) with kmer sizes ranging from 18 to 36 and increasing by 2 with each iteration, with 97% similarity for alignment (Supplementary Table 2).

Phylogenetic annotation was performed with PhymmBL (Brady and Salzberg 2009; Brady and Salzberg 2011) and Kraken (Wood and Salzberg 2014). PhymmBL combines the Phymm (Brady and Salzberg 2009) with BLAST (Altschul et al. 1990)

algorithms for increased accuracy. Aside from aligning and annotating contigs, PhymmBL provides identity confidence scores. A 65% identity confidence score threshold was imposed for Order-level assignments. Comparative taxonomic classifications were performed with Kraken (Wood and Salzberg 2014) using the standard database comprised of complete RefSeq bacterial, archaeal, viral, and fungal genomes. Contigs assigned to viral or fungal genomes were considered contaminants and were removed for downstream analyses. Marker gene annotation of filtered contigs was performed with Phylosift (Darling et al. 2014).

Open reading frame (ORF) prediction was performed in six reading frames with MetaGene (Noguchi et al. 2006). ORFs were annotated for KEGG Orthology (KO) families (Kanehisa et al. 2016) in HMMER 3.0 (Eddy 2011) using the Functional Ontology Assignments for Metagenomes (FOAM) hidden Markov model database (Prestat et al. 2014). An e-value acceptance threshold of 1e-4 was enforced. In the case of multiple KO assignments per contig, the result with the best e-value and bitscore was chosen to represent that contig. Contigs that did not receive a protein annotation from these software were manually annotated with BLASTX (Altschul et al. 1997). Alignments were performed against the non-redundant protein database and scored with the BLOSUM62 matrix (Henikoff and Henikoff 1992), with a maximum expectation value of 1e-4 and a word size of 3.

1.2.9 Methylation scoring

CpG methylation was calculated with a proprietary Python pipeline designed by Dr. Adam G. Marsh. A combined assembly of IDBA-assembled contigs was performed to identify CpG sites that could be directly compared between all depths. Methylated (*met*) and unmethylated (*umt*) continuous score metrics were derived for each CpG site based on the proportional representation of methylated and non-methylated copies within an aligned assembly. These two metrics allow for the quantitative scoring of methylation states relative to (x,y) positions on a 2D coordinate plane**.** The magnitude of a site's shift in methylation state (M_{shift}) between samples was measured as the Euclidean distance, as follows:

$$M_{shift} = \sqrt{(umt_2 - umt_1)^2 + (met_2 - met_1)^2}$$

1.2.10 Statistical analyses

Statistical analyses were conducted with R. The presence of outlier contigs from IDBA assemblies was determined with Bonferroni outlier tests. Further tests were conducted for *met* and *umt* CpG methylation scores of abundant phyla (Actinobacteria, Bacteroidetes, Chloroflexi, Firmicutes, and Proteobacteria). Modalities of *met* and *umt* scores were tested with Hartigans' dip test for unimodality (Hartigan and Hartigan 1985) using the diptest package. Bootstrap standard errors (SE) and coefficients of variation (CV) were estimated (n = 10,000) for *met* and *umt*

score metrics. Random sampling with replacement was performed with the dplyr package, with the randomization seed set at 2016. Variances of *met* and *umt* were tested with a Brown-Forsythe Levene-type test (Brown and Forsythe 1974) using the lawstat package. Two-tailed Jonckheere-Terpstra trend tests (Jonckheere 1954; Terpstra 1952) with 1000-permutation reference distributions were performed with the clinfun package to determine down-core trends in *met* and *umt* scores.

1.3 Results

1.3.1 Sediment properties

Sediment dating constraints of core R via ^{7}Be, ^{210}Pb, and ^{137}Cs show that the Oyster Rocks site is comprised of a top layer of recently deposited tidally mixed or bioturbated sediment (~ 4 cm, sediment age < 106 days) situated above older sediment (50-100+ years old) (Supplementary Table 1). Sulfate concentrations were more varied between 0-3 cm and 3-6 cm (Figure 1.1 A) for Core L, but concentrations were higher in deeper samples from 6-9 cm to 27-30 cm. Methane concentrations of Core L were shown to increase with depth, with higher variance between 0-12 cm and lower variance from 15-30 cm (Figure 1.1 B). Porosity for Core R was shown to be far lower within older sediments (Figure 1.1 C). The higher variability of SO_4^{-2} and CH_4 in more recently deposited sediments could be due to bioturbation and tidal forcing.

1.3.2 16S rRNA gene analysis

16S rRNA gene Chao1 diversity index was generally higher at 3-6 cm than the 12-15 cm and 24-27 cm samples (Figure 1.2). The 12-15 cm and 24-27 cm samples had similar profiles for rarefied Chao1 diversity and observed OTU counts. The 12-15 cm and 24-27 cm samples had a higher presence of Dehalococcoidetes and sulfate-reducing Deltaproteobacteria, as well as Marine Crenarchaoetal Group and Marine Hydrothermal Vent Group archaea (Figure 1.3). OTUs are clearly shared between the three depths, with corresponding abundance changes suggesting taxa which prefer anaerobic conditions inhabit the deeper depths (Figure 1.3).

1.3.3 Metagenome taxonomic composition and function

Sulfate-reducing Deltaproteobacteria were in all three metagenome samples, although relative abundances increased within the 12-15 cm and 24-27 cm samples (Figure 1.4). Common soil-inhabiting bacteria such as the Actinomycetales were present in high abundance at all depths. The most abundant classes seen in all three samples were the Actinobacteria, Alphaproteobacteria, Bacilli, Betaproteobacteria, Clostridia, Deinococci, Deltaproteobacteria, and Gammaproteobacteria (Figure 1.4, 1.5). Contigs with both higher-scoring PhymmBL annotations and Kraken annotations further indicate a prevalence of anaerobic taxa within the 12-15 and 24-27 cm samples. These annotations show an increased presence of obligate anaerobes such as Clostridiales and Synergistales, as well as anaerobic halorespirers such as the

Dehalococcoidales (Figure 1.5**).** Both Kraken and PhymmBL annotations show a presence of methanogenic archaea at the 12-15 cm and 24-27 cm depths. Kraken will not classify a sequence if sufficient evidence does not exist, and this allows for higher precision at default settings compared with PhymmBL (Wood and Salzberg 2014). This is intended as a control for false positives assigned by PhymmBL.

Phylosift marker gene annotations support the presence and higher relative abundance of Dehalococcoidia within sediments at 12-15 cm and 24-27 cm (Figure 1.6). These annotations also support the presence of methanogenic archaea with high similarity to the Methanomicrobiales and Methanosarcinales at these depths. Additionally, KEGG Orthology annotations obtained from the FOAM database suggest that the deeper communities have the genetic potential to perform anaerobic metabolic processes (Figure 1.7). Genes involved in sulfate reduction (formate dehydrogenase, adenylyl sulfate kinase, NADH dehydrogenase, and heterodisulfide reductase) and methanogenesis (trimethylamine corrinoid protein co-methyltransferase) were present.

1.3.4 Metagenome CpG methylation

From these metagenome data, a total of 6254 CpG sites that could be directly compared between all three samples were mapped to 3743 contigs (4.33% of all three unprocessed IDBA assemblies). Differential methylation states were observed in 1173 sites, while the remaining 5081 had equivalent methylation states. Of these CpG sites,

4235 (67.7%) were identified within contigs with PhymmBL annotation Order confidence scores greater than or equal to 0.650. CpG methylation states at the phylum level show that the majority of sites are either highly methylated (MET) or unmethylated (UMT). An overall trend of CpG site shifts to lower methylation states with depth can be seen in the Proteobacteria (Figure 1.10), Firmicutes (Figure 1.11), and Actinobacteria (Figure 1.12). Most differential shifts are from MET to UMT states and vice versa (Table 1.1). The presence or shifts to unknown (UNK) states is mostly due to coverage being too low (< 5 fold).

The methylation shift behaviors of individual CpG sites are varied and highly dependent upon their original states. Comparable CpG sites exhibit both differential and non-differential methylation shifts across the three samples (Figure 1.13, 1.14, 1.15)**.** The CpG sites represented in these figures comprise the six most numerous shifts from 3-6 cm to 12-15 cm, excluding those to or from UNK states. Shift behaviors between 12-15 cm to 24-27 cm of the same sites that are not to UNK states are displayed. Sites that evidence non-differential shifts within highly methylated states (MET:MET) tend to cluster at specific positions on the coordinate plane, with the majority of shifts occurring between similar methylation proportions and being of low magnitude (Figure 1.13, 1.14, 1.15: A, G). The low spread and magnitude of these shifts indicates a preference for methylated CpG sites to remain at highly similar methylation states. CpG sites that shift from MET to UMT states are also shown to cluster together in the MET state, but these methylated groupings shift into a variety of

UMT states (Figure 1.13, 1.14, 1.15: C, I). Inversely, multiple UMT CpG sites of varying methylation scores shift to highly similar MET states (Figure 1.13: E, K; Figure 1.14, 1.15: D, L**)**. A change from 100% of a CpG site's copies being methylated to 100% of them being unmethylated (and vice versa) is indicative of greater cellular responses. This relationship also indicates that similarly methylated CpG sites can experience varying degrees of methylation loss, but cluster at similar methylation proportions when shifting from UMT to MET states. CpG sites can also have mixed states of methylation prominence (MIX) resulting from partial methylation gains or losses. Proteobacteria and Actinobacteria were observed to have higher instances of shifts from MET to MIX states, indicating that these CpG sites undergo partial methylation losses (Figure 1.13, 1.15: B, H). Partial methylation gains of unmethylated CpG sites (UMT:MIX) were also observed in the Proteobacteria from 12-15 to 24-27 cm (Figure 1.13 M**)**.

Hartigans' dip test for unimodality was used to test the null hypothesis that the distribution of a variable is unimodal, with an alternative hypothesis that the distribution is at least bimodal. Results for *met* and *umt* scores support a non-unimodal distribution for most target Phyla (Supplementary Table 4). P-values for Bacteroidetes *met* scores are highly non-significant, indicating a unimodal score distribution for this phylum. *Met* scores for Chloroflexi CpG sites are also unimodal at 24-27 cm ($p = 0.5416$). P-values for *met* scores are consistently higher with depth across all five phyla in accordance with the overall trend towards unimodality (Figure 1.9 A). In contrast,

p-values for *umt* scores decrease with depth, indicating less unimodal distributions that are supported by overall trends towards distinct peaks (Figure 1.9 B).

Bootstrap SE and CV of CpG site *met* and *umt* were lowest overall for the Actinobacteria. An overall trend of increasing SE for *met* scores with depth was seen in all analyzed Phyla. SE for *umt* scores did not follow this trend, as the SE for 12-15 cm scores was highest overall amongst all Phyla. *Met* SE was highest for 24-27 cm scores, while *umt* SE was highest for 12-15 cm scores. CV was shown to be generally higher for 3-6 cm and 12-15 cm scores, with a trend of decreasing CV with depth (Supplementary Table 3).

The Brown-Forsythe test compares group variances using the median as opposed to the mean used in the classic Levene's test. *Met* and *umt* variances were tested against the null hypothesis H_0: variances of two or more groups are equal. Results reject the null hypothesis and support the alternative hypothesis that variances are unequal across depths ($p < 0.01165$), supporting the presence of dynamic shifts in methylation states (Supplementary Table 5).

Jonckheere-Terpstra trend tests were used to compare score metric medians across all three samples against the null hypothesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$, where θ_i is the population median for the *i*th population. The alternative hypothesis is that the population medians have an *a priori* ordering with at least one strict inequality, e.g. H_a: $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_k$. The *a priori* ordering of these sample depths gives the Jonckheere-Terpstra test more statistical power than the Kruskall-Wallis H test.

Results show that *umt* scores for Bacteroidetes did not change consistently with depth ($p = 0.47$), while *met* scores did ($p = 0.002$). Results for Chloroflexi show that both *umt* and *met* scores decreased with depth ($p = 0.008$, $p = 0.002$, respectively). Both *met* and *umt* scores for the other four phyla decreased with depth ($p = 0.002$) (Supplementary Table 5).

Only 35 CpG sites were traced back to contigs receiving a KEGG Orthology annotation. Chitinase gene annotations were recovered for 14 comparable CpG sites that could be traced back to six contigs with PhymmBL annotation Order confidence scores greater than or equal to 0.650 (Figure 1.16, 1.17**)**. While the number of recovered CpG sites for chitinase-annotated contigs is low, it should be noted that 12 quantifiable CpG sites were identified at several positions within contigs assigned to Actinomycetales and Thermoanaerobacterales. These sites exhibit both methylation gains and losses with depth. Quantifiable states of 73 CpG sites were recovered for transposase genes identified by BLASTX alignments (Figure 1.18, 1.19). Transposase CpG sites tended to mostly remain in unmethylated states between both depth shifts. However, several CpG sites had highly weighted shifts from UMT to MET states, or vice-versa.

Table 1.1: Recovered CpG site methylation state shifts

Shift	Overall	Proteobacteria	Actinobacteria	Firmicutes	Bacteroidetes	Chloroflexi
	1491	690	277	150	20	53
MET:MET	657	311	129	65	9	21
	26	9	9	2	2	-
MET:MIX	36	21	4	2	1	4
	472	204	90	59	5	15
MET:UMT	913	399	169	105	14	40
	159	66	25	24	6	6
MET:UNK	291	147	49	26	3	12
	10	7	1	1	-	1
MIX:MET	15	6	3	3	-	-
	5	3	1	1	-	-
MIX:MIX	2	1	9	-	-	-
	3	1	1	-	-	-
MIX:UMT	26	11	2	2	-	-
	-	-	-	-	-	-
MIX:UNK	3	-	-	-	-	-
	282	136	46	34	6	16
UMT:MET	294	128	58	36	6	8
	12	5	3	2	-	-
UMT:MIX	23	9	1	6	-	2
	1336	622	232	127	28	45
UMT:UMT	1376	640	241	137	21	47
	44	21	9	5	-	1
UMT:UNK	156	65	31	15	6	5
	114	45	27	13	1	7
UNK:MET	96	43	23	14	-	2
	3	1	1	-	-	-
UNK:MIX	7	4	1	1	-	-
	38	15	8	8	-	2
UNK:UMT	121	49	19	19	5	2
	240	101	51	29	6	3
UNK:UNK	219	92	42	24	7	6
SUM	4235	1926	781	455	141	149

Shaded rows indicate methylation state shift counts between the 3-6 cm to the 12-15 cm sample. Non-shaded rows indicate counts for the 12-15 cm to 24-27 cm shift.

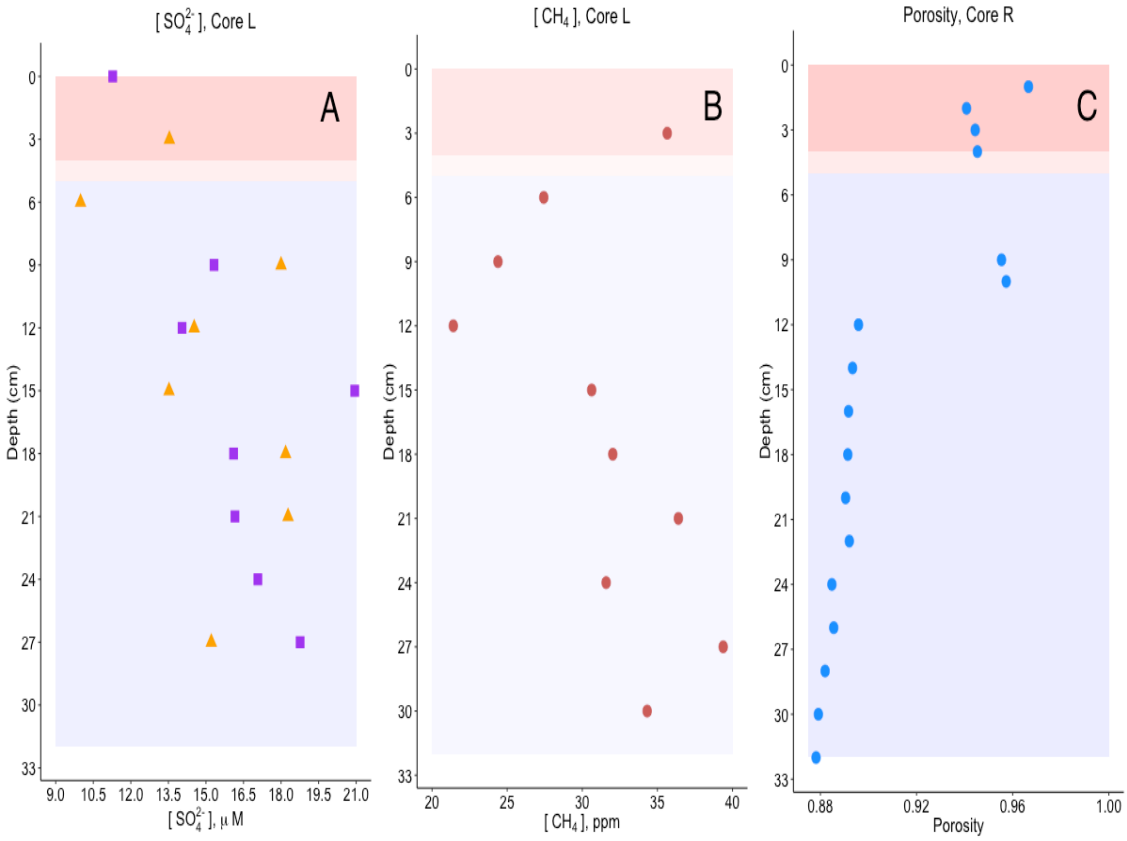


Figure 1.1: Sediment core sulfate, methane, and porosity. Sulfate (Panel A, purple squares = 1:6 dilution, orange triangles = 1:11 dilution) and methane (Panel B) concentrations are displayed for Core L. Porosity (Panel C) is displayed for Core R. Shaded bars are representative of radionuclide age constraints along the depth of Core R (peach = youngest sediment, age < 106 days; pink = extent of ^{210}Pb from 0 cm to transition zone; blue = sediment 50-100+ years old).

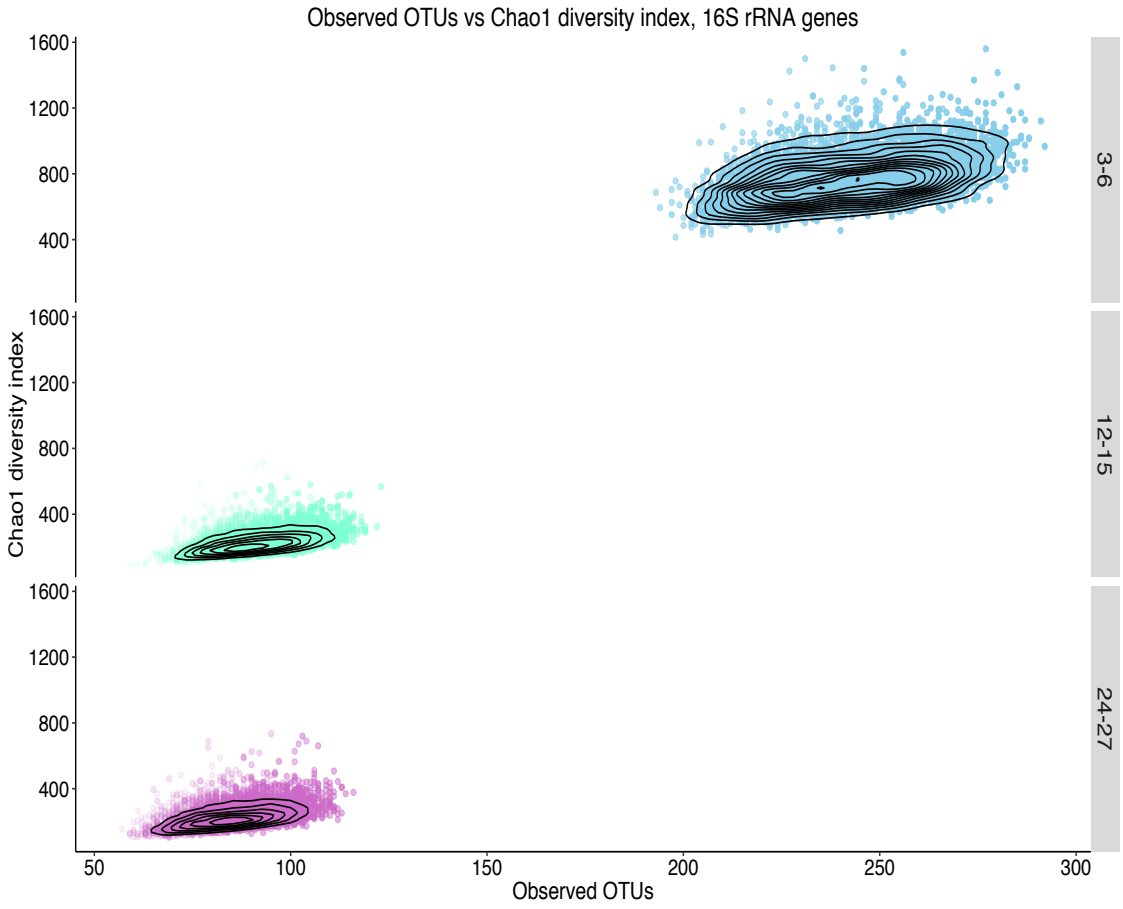


Figure 1.2: Chao1 alpha diversity for rarefied 16S rRNA gene OTU tables. Alpha diversity was calculated for rarefied OTU tables using the Chao1 index (Chao 1984). Community diversity is higher within recently deposited surface sediments (3-6 cm) and lower within older, established sediments (12-15 cm, 24-27 cm).

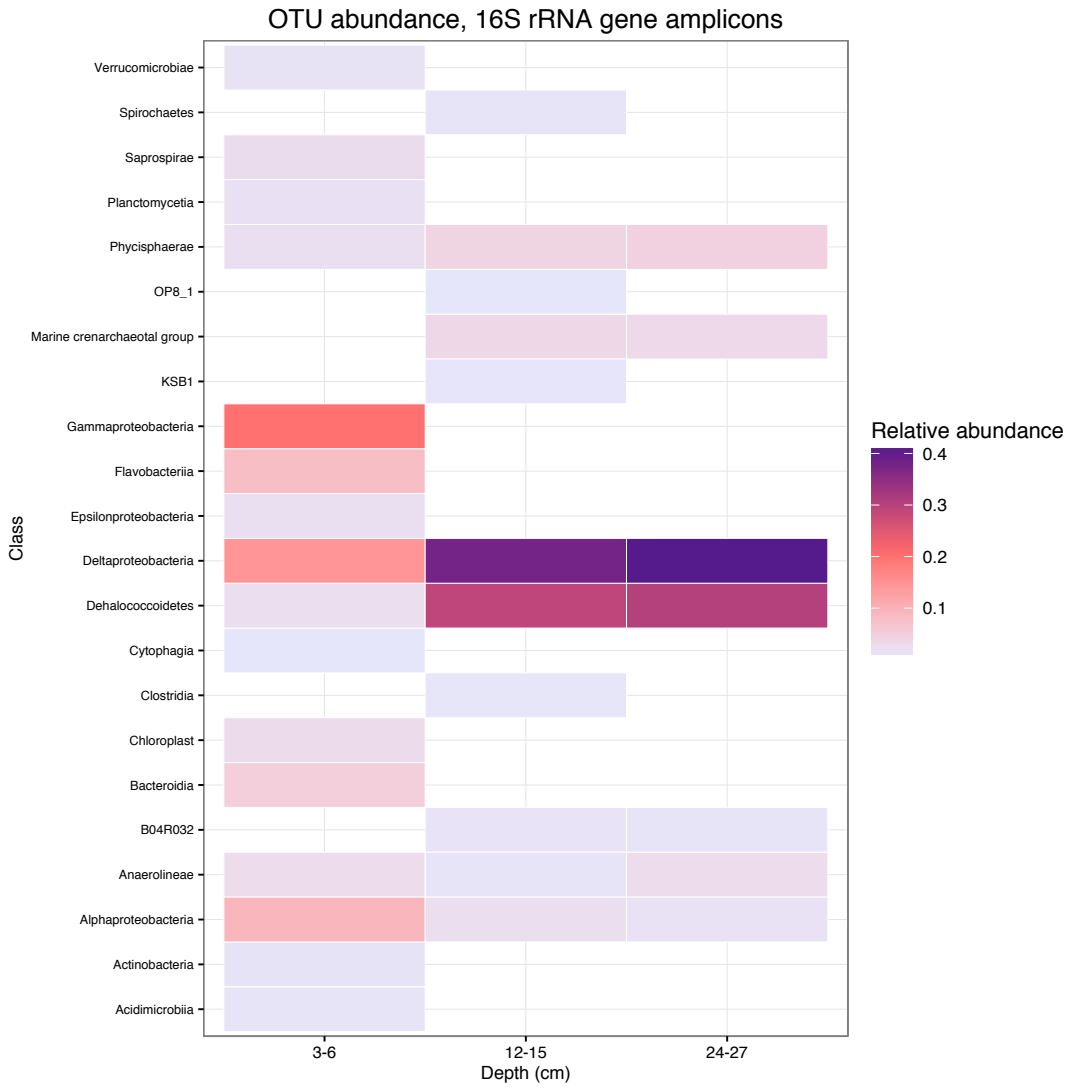


Figure 1.3: OTU class relative abundance of 16S rRNA gene amplicons. Analysis of 16S rRNA gene amplicons shows a drastic shift in community diversity between the shallow (3-6 cm) sample and deeper (12-15, 24-27 cm) samples. This drastic shift can be associated with the differences in sediment age between the shallow and deep samples. The shallow sample is host to a wide variety of microbes, both aerobic and anaerobic, while anaerobic taxa such as Deltaproteobacteria and Dehalococcoidetes have greater abundance in the deeper samples.

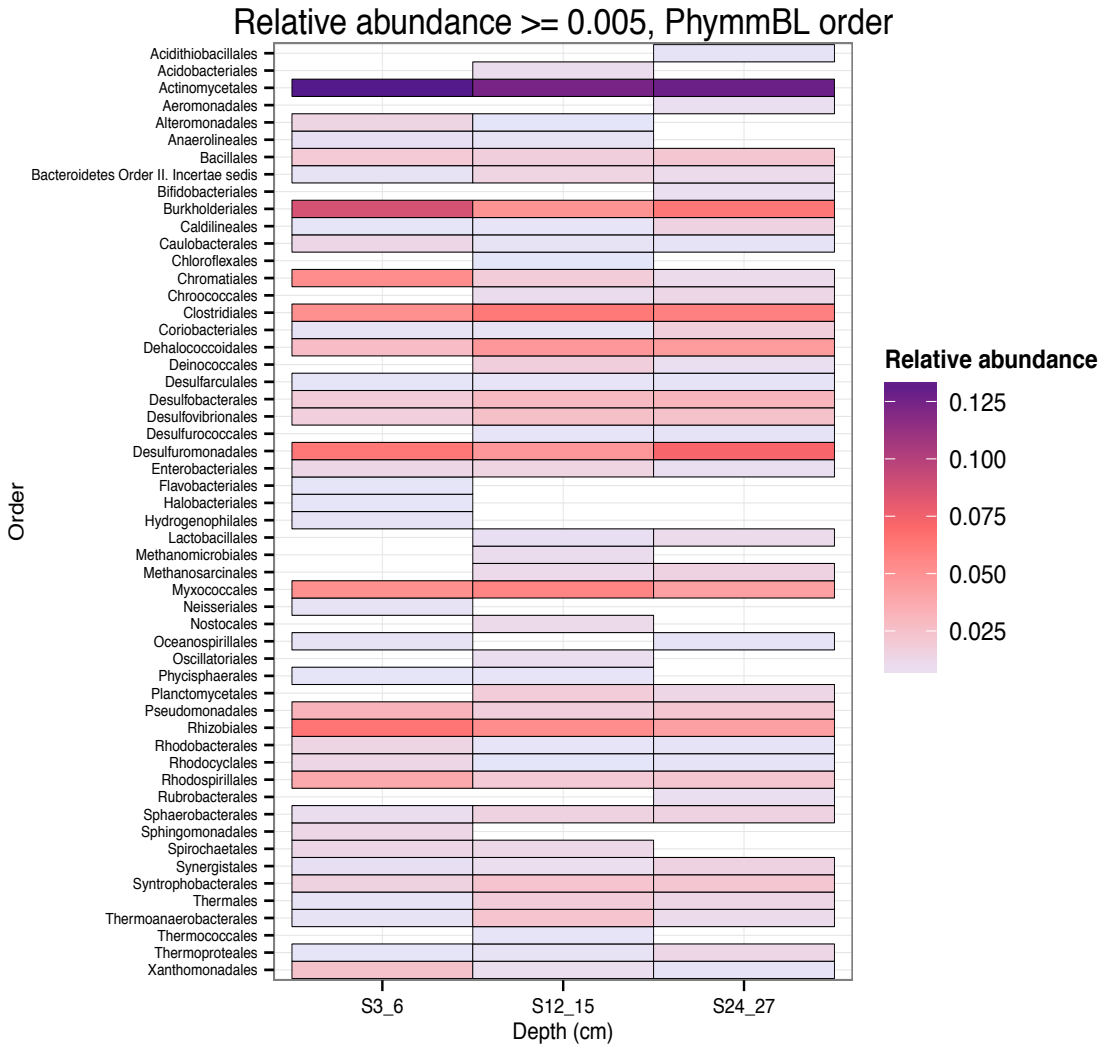


Figure 1.4: Taxonomic abundances of PhymmBL-annotated metagenome contigs. Annotated contigs have an Order confidence score greater than or equal to 0.650. Members of Actinomycetales are common in soils and sediments, and were present at all depths.

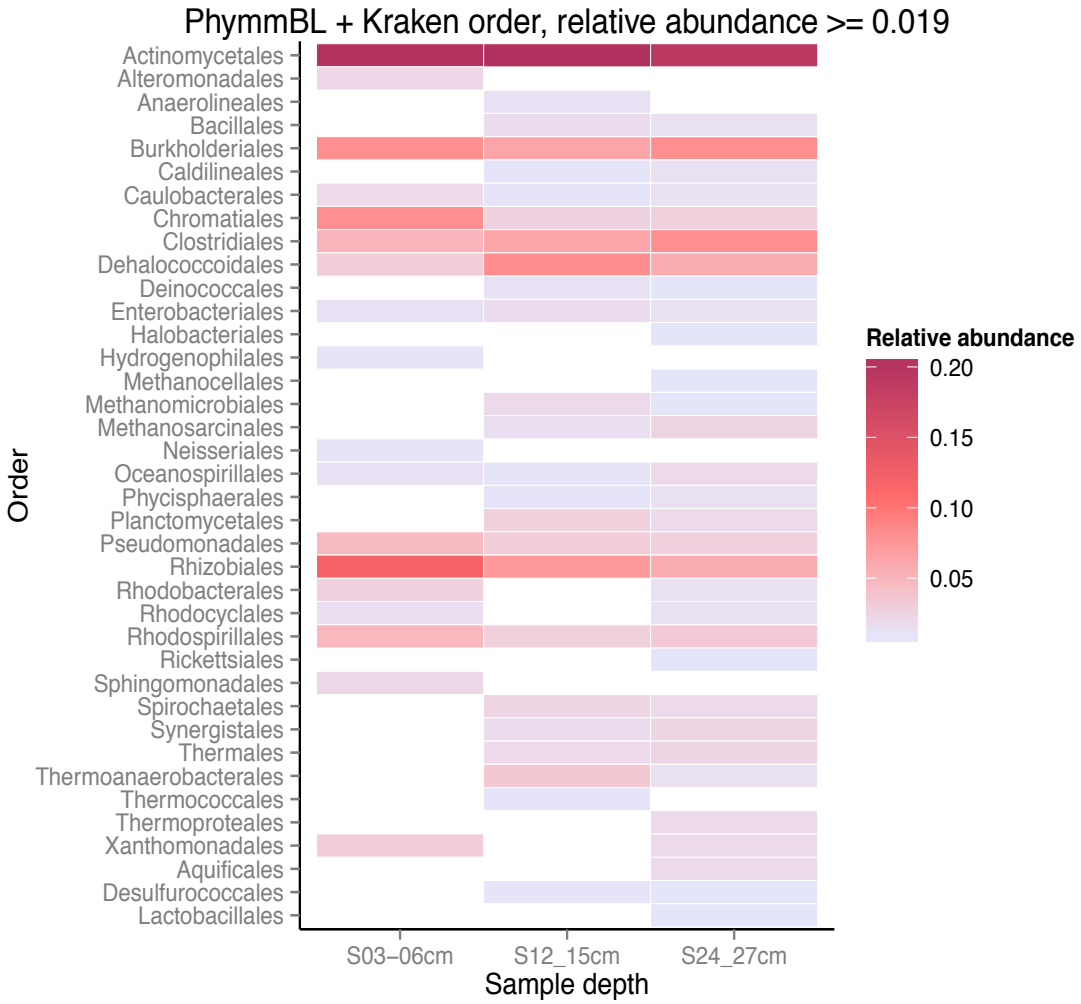


Figure 1.5: Taxonomic abundances of metagenome contigs with both Kraken and higher-confidence PhymmBL assignments. PhymmBL annotations were paired with Kraken annotations to select against potential false positives annotated by PhymmBL. These annotations suggest a presence of Actinomycetales at all depths, and an increased abundance of anaerobic Orders (Clostridiales, Dehalococcoidales). Methanogenic archaea (Methanosarcinales) are present in the 12-15 cm and 24-27 cm samples.

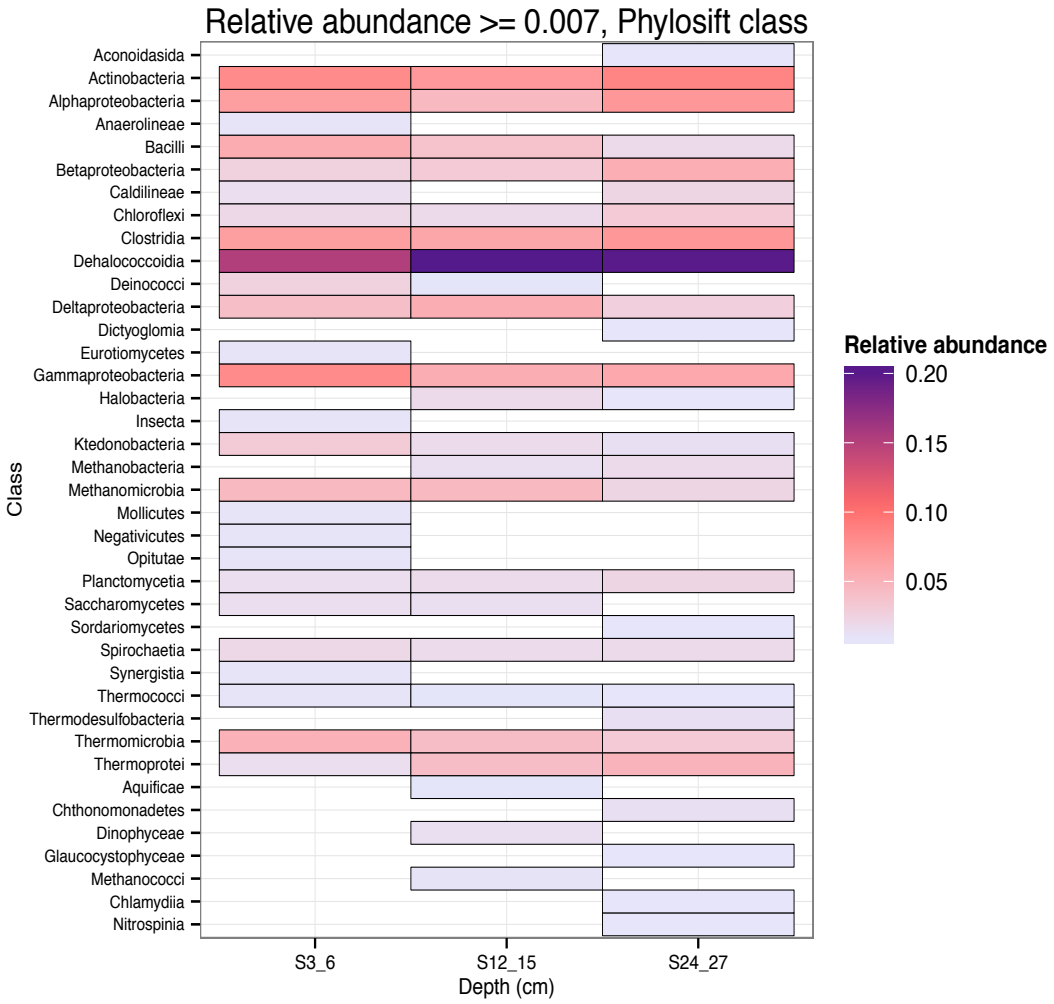


Figure 1.6: Class-level abundance of marker genes annotated with Phylosift. Marker gene annotation suggests an increased abundance of several anaerobic classes in deeper samples, which could be indicative of a more anaerobic sediment environment at 12-15 cm and 24-27 cm.

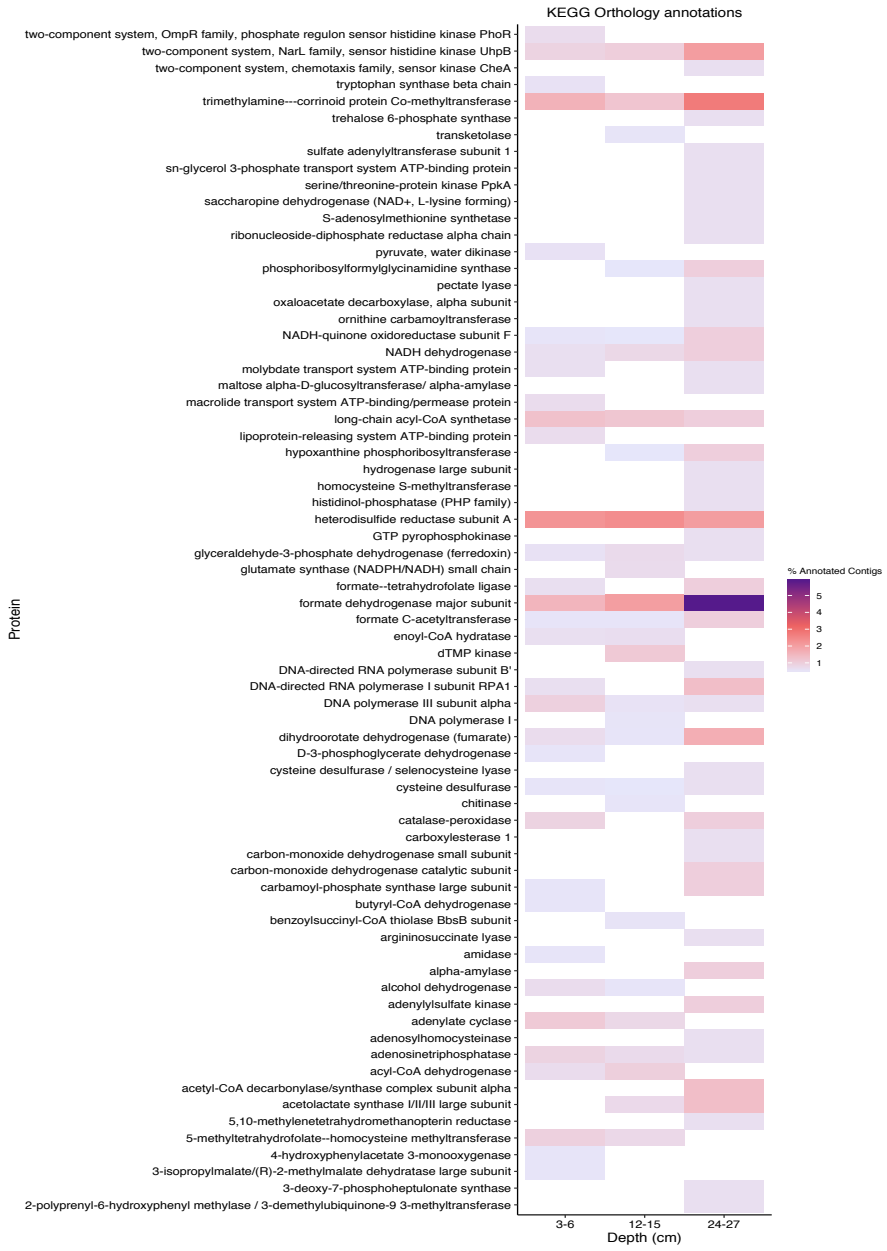


Figure 1.7: Abundance of KEGG Orthology functional assignments across sample depths. KEGG Orthology annotation results show that Oyster Rocks sediment communities at 12-15 cm and 24-27 cm have higher enzymatic potential for anaerobic metabolism.

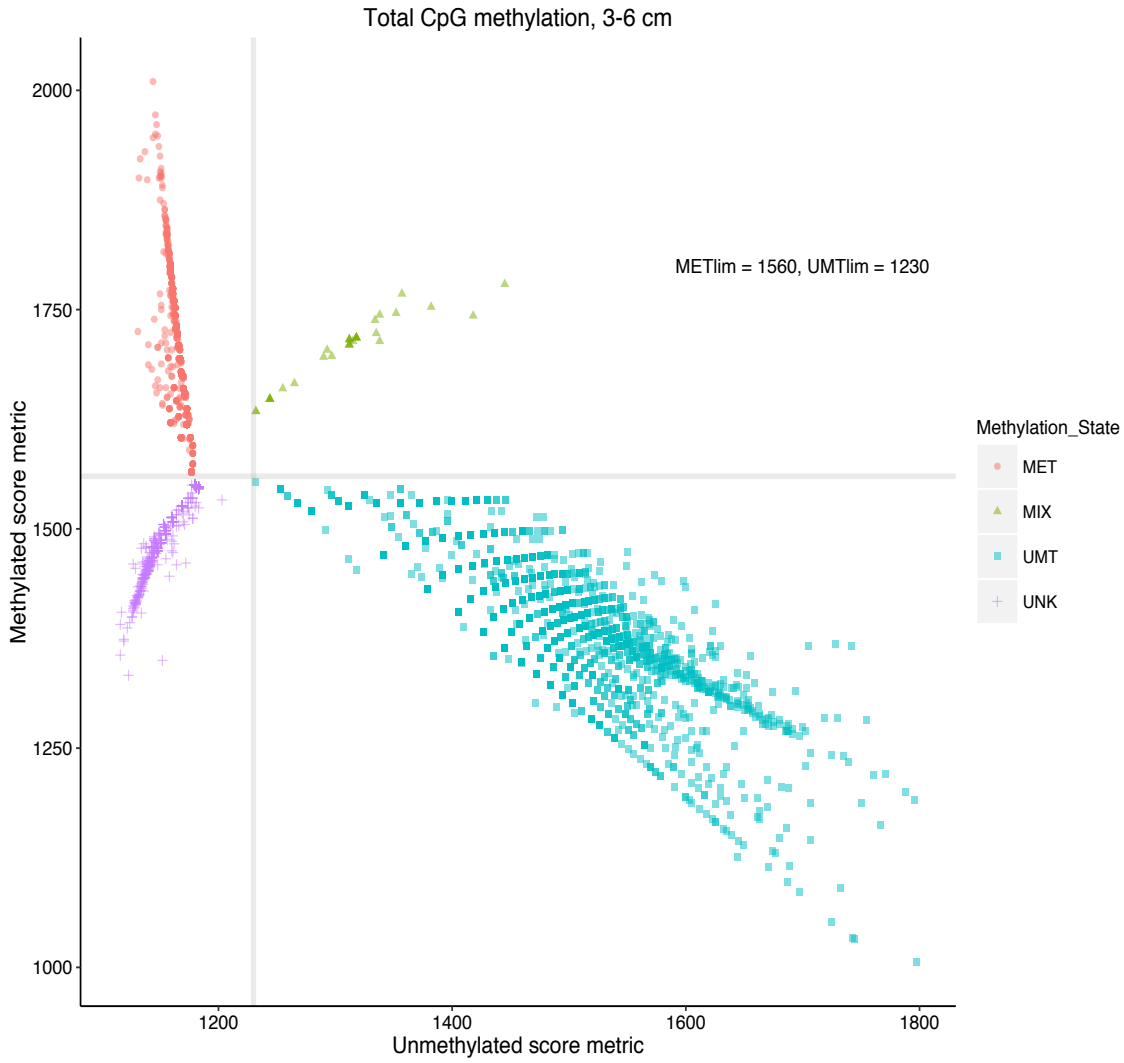


Figure 1.8: Overall CpG methylation score profile for 3-6 cm sample. Each point is representative of a single CpG site. Methylated (y-axis) and unmethylated (x-axis) score metrics are determined for each site by the proportional representation of methylated and non-methylated copies. An algorithm within the methylation scoring software divides the data into response groups, thus imparting multi-modality. The gray lines determine the boundaries set for grouping points into methylated (MET), mixed methylation (MIX), unmethylated (UMT), or unknown or resolved (UNK) categories based on their methylation scores. Orange circles = methylated sites, green triangles = mixed methylation sites, blue squares = unmethylated sites, purple crosses = unknown or unresolved sites.

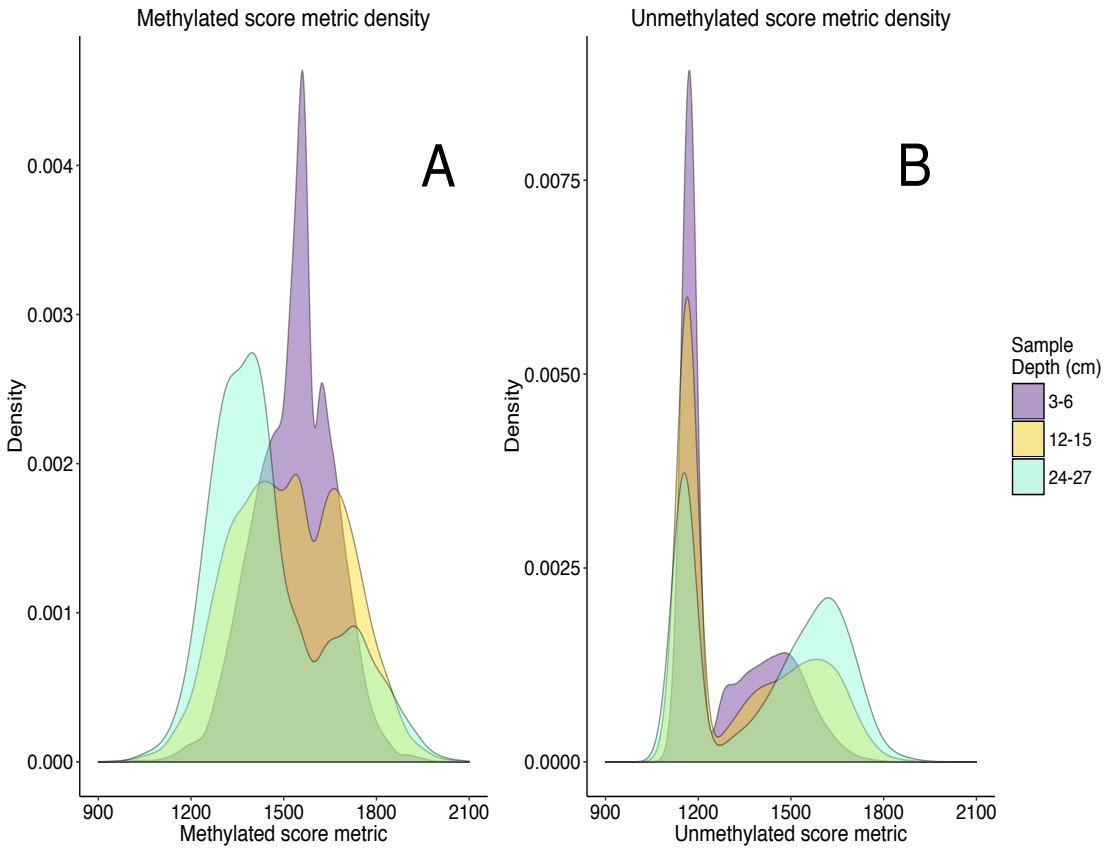


Figure 1.9: Density distributions of pan-metagenomic methylation scores. Score distributions for comparable CpG sites are multimodal and shift between samples, although methylated score metrics become increasingly unimodal with depth. Panel A shows density distributions of the methylated score metric (*met*), while Panel B shows the unmethylated score metric (*umt*).

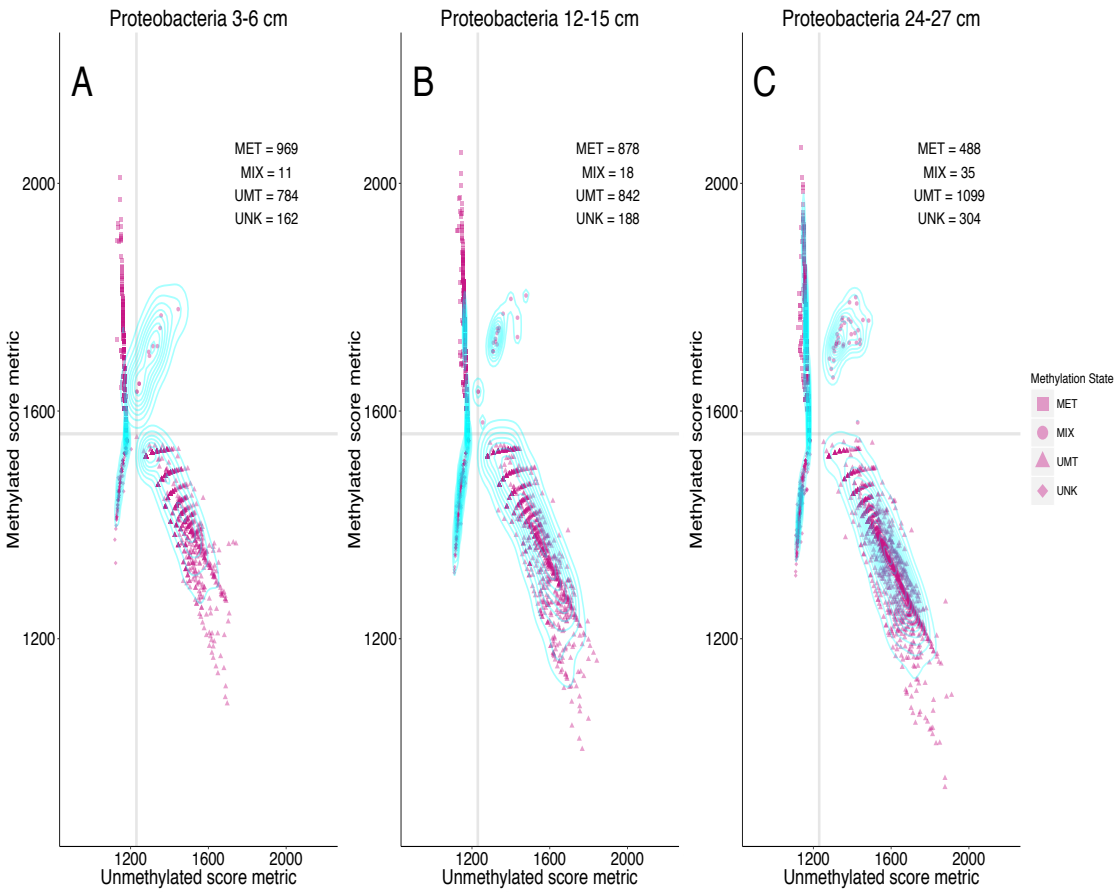


Figure 1.10: Proteobacteria CpG methylation across depths. A total of 1,926 CpG sites were recovered that could be compared between the three depths. Gammaproteobacteria represented 46.08% of these sites, with Alphaproteobacteria, Deltaproteobacteria, and Betaproteobacteria comprising 21.24%, 16.82%, and 15.32% of these sites, respectively. The vast majority of CpG sites shift between methylated and unmethylated states from the 3-6 cm sample (Panel A) to the 12-15 cm sample (Panel B), suggesting binary "on/off" shifts. There appears to be an increase in mixed methylation states with depth, indicating that some CpG sites shift towards fractional methylation. There is noticeable loss of methylated sites in the transition from 12-15 cm to 24-27 cm sample (Panel C), and an increase in unmethylated sites. Squares = methylated sites, triangles = unmethylated sites, circles = mixed methylation sites, diamonds = unknown or unresolved sites.

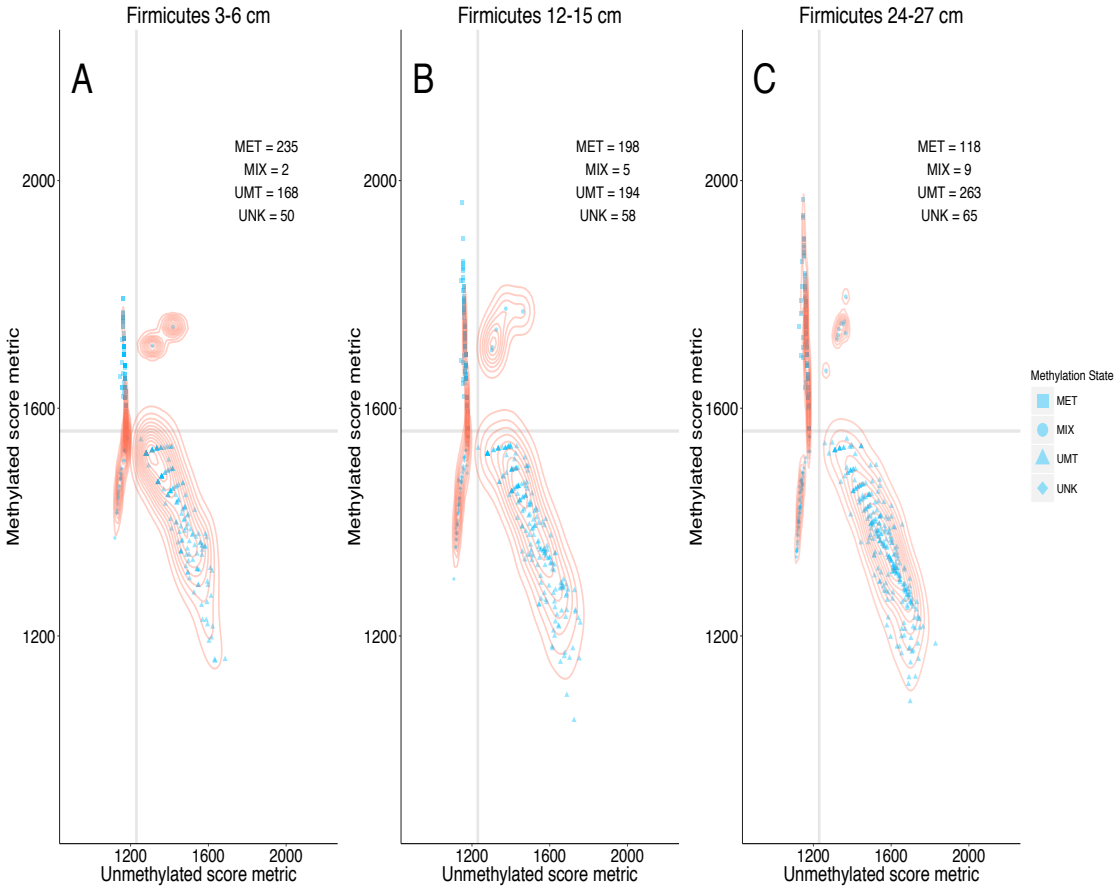


Figure 1.11: Firmicutes CpG methylation across depths. A total of 455 CpG sites were recovered from Firmicutes. The majority of these CpG shifts occur from highly methylated to unmethylated states, which is suggestive of binary "on/off" switches. The most prominent orders comprising these data include members of Clostridia (Clostridiales, Thermoanaerobacterales), Bacilli (Bacillales, Lactobacillales) and Negativicutes (Selenomonadales). Squares = methylated sites, triangles = unmethylated sites, circles = mixed methylation sites, diamonds = unknown or unresolved sites.

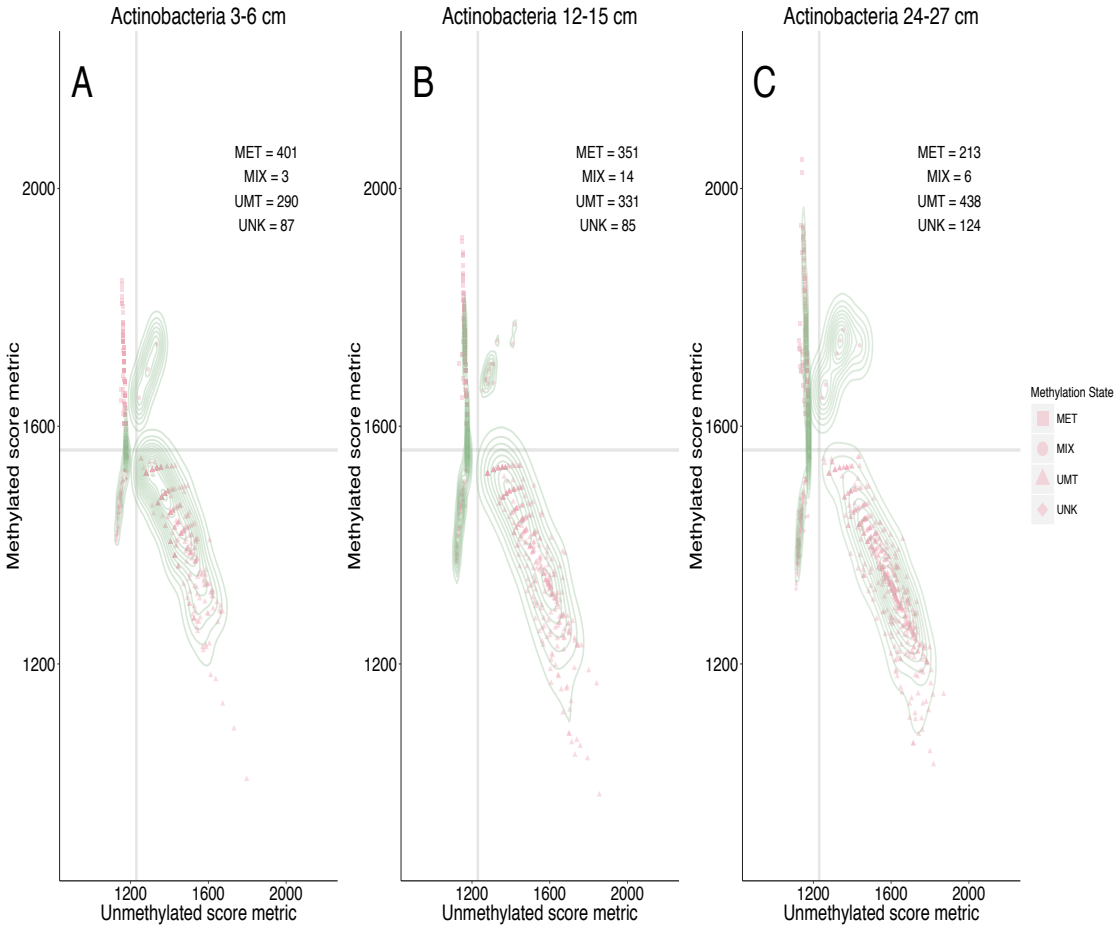


Figure 1.12: Actinobacteria CpG methylation across depths. Of the total 781 CpG sites recovered for Actinobacteria, 696 (89.11%) were mapped to contigs annotated as Actinomycetales. Remaining CpG sites were mapped to Coriobacteriales, Rubrobacterales, and Bifidobacteriales. While most sites shift towards unmethylated states downcore, it should be noted that many UMT sites congregate into a high-density region (Panel C). Squares = methylated sites, triangles = unmethylated sites, circles = mixed methylation sites, diamonds = unknown or unresolved sites.

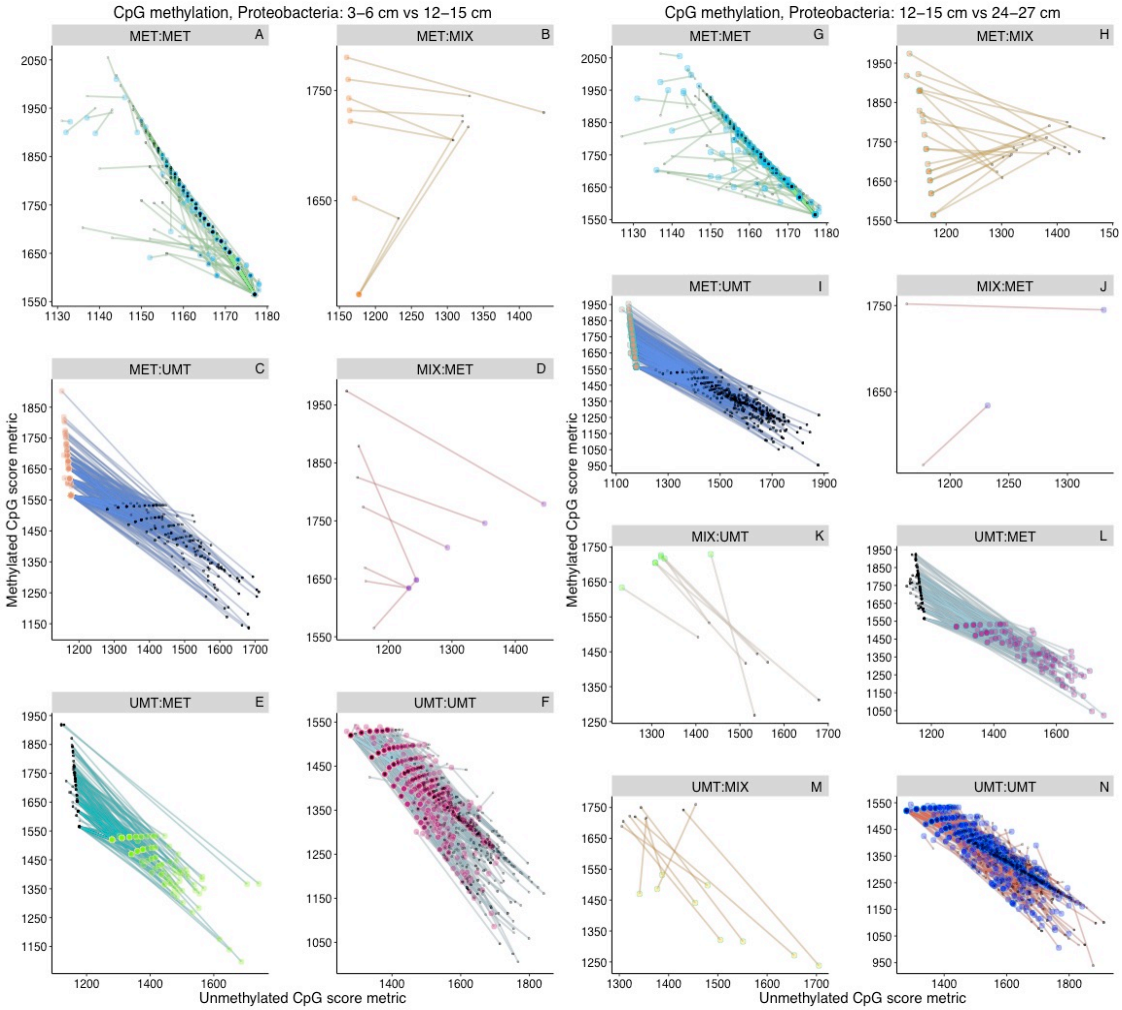


Figure 1.13: Proteobacteria CpG site methylation shift dynamics. Scores for individual CpG sites are linked by a line and separated into differential and non-differential methylation state shifts based on coordinate positions. Each point is a CpG site that has evidenced a change in methylation state between samples while maintaining coverage > 5x. The larger, colored point represents the quantitative state within the shallower sample, with higher opacity indicating multiple CpG sites with similar methylation proportions. The line extends to the state of the same CpG site within the deeper sample, represented by a smaller black point. The shading of the smaller black point is dependent upon the magnitude of the shift (M_{shift}) between depths, with greater shifts being darker. More intense line coloration indicates multiple similar shifts, and the length of the line is determinant upon the magnitude of the shift.

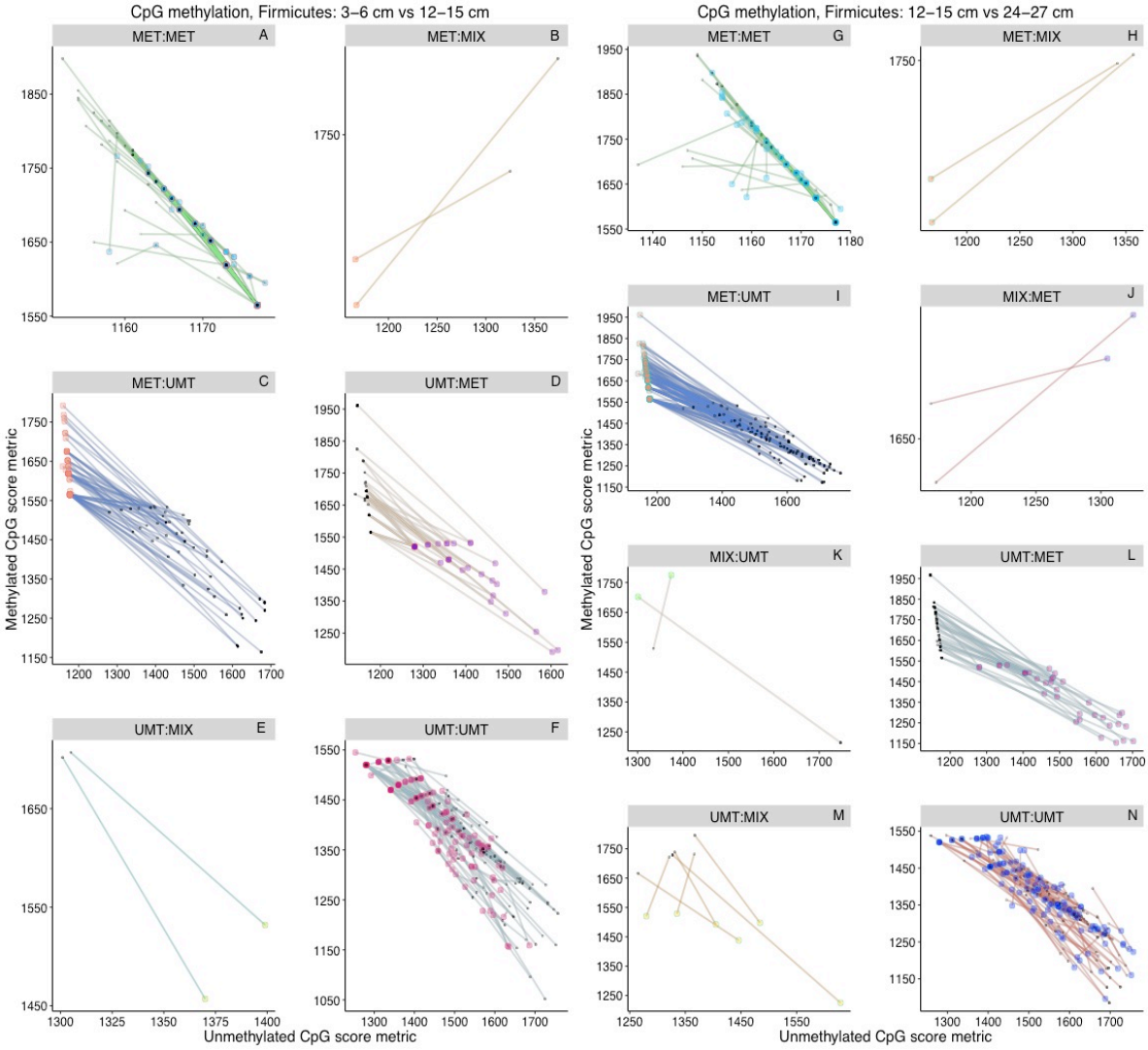


Figure 1.14: Firmicutes CpG site methylation shift dynamics. See description of Figure 1.13 for interpretation guidelines. A greater number of CpG sites shift from methylated to unmethylated states or remain in unmethylated states between these depths.

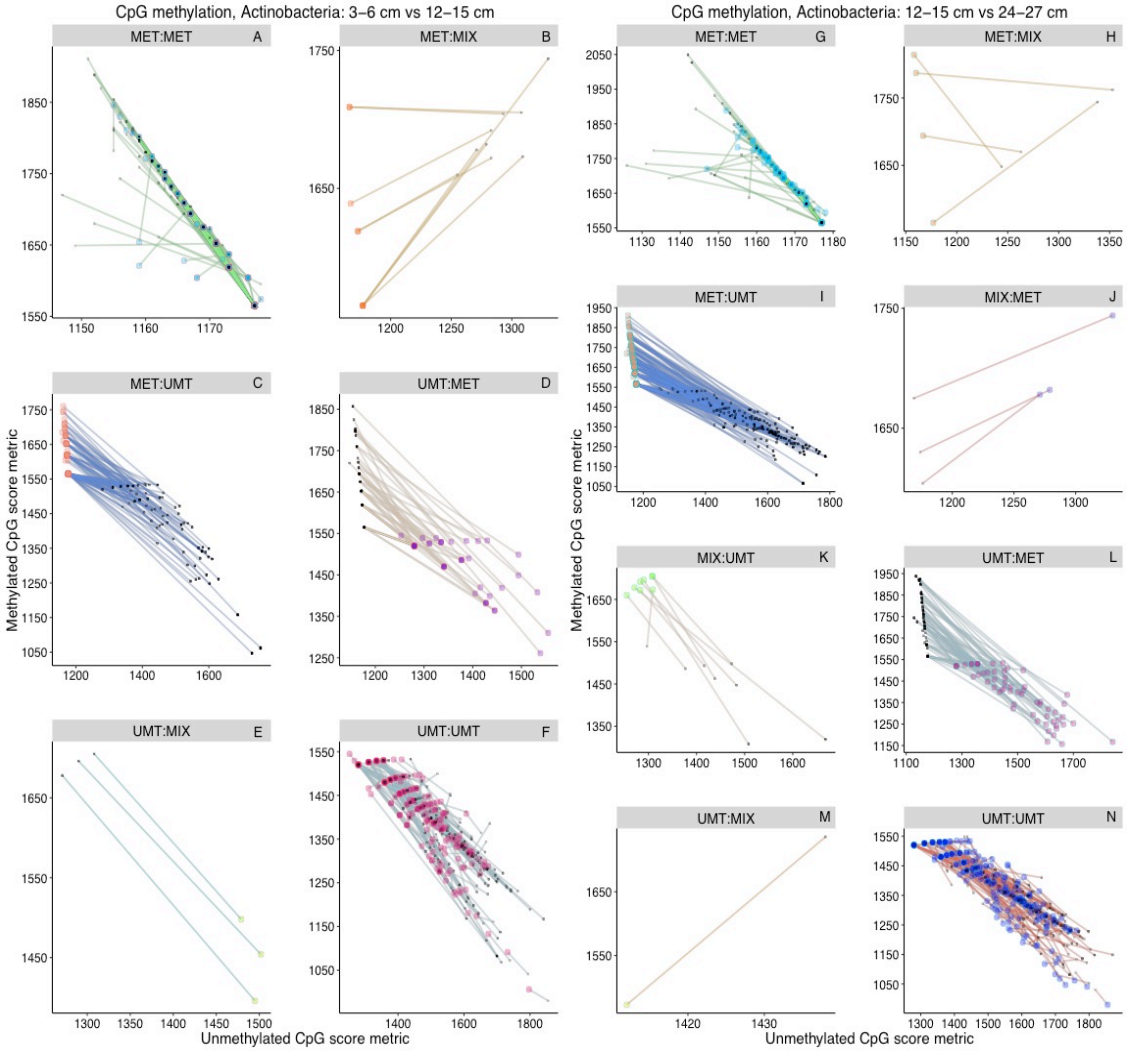


Figure 1.15: Actinobacteria CpG site methylation shift dynamics. See description of Figure 1.13 for interpretation guidelines. Actinobacteria had the lowest CV for *met* and *umt* scores, which is attributed to the tendencies of CpG sites to cluster at similar states.

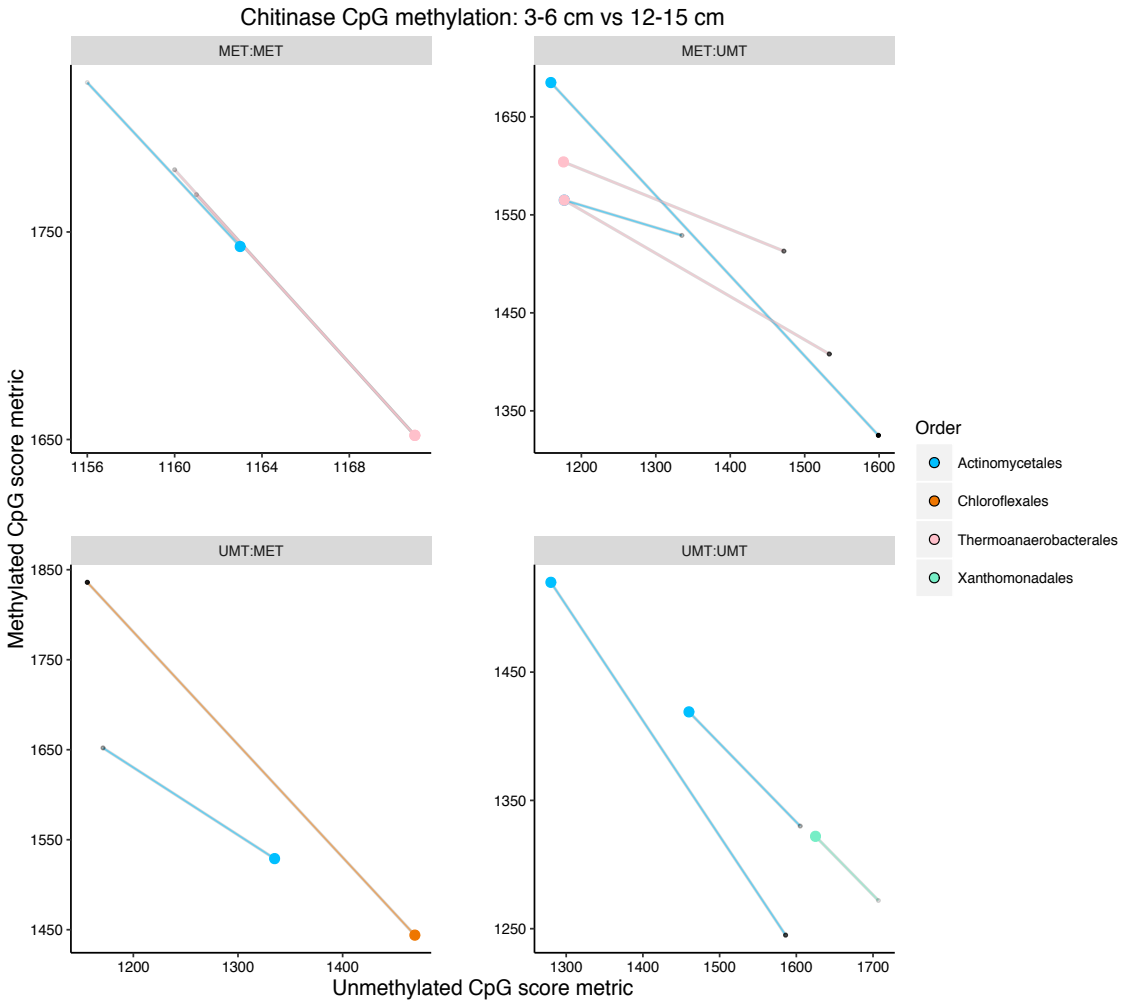


Figure 1.16: CpG methylation shifts for chitinases, 3-6 vs 12-15 cm. Genes were assigned to KO within the FOAM database. Actinobacteria experienced more drastic methylation losses as opposed to methylation gains. The points in this graph represent the same CpG sites represented in Figure 1.15 (12-15 vs. 24-27 cm), with the smaller point representing the same coordinate position as the larger point in the later figure.

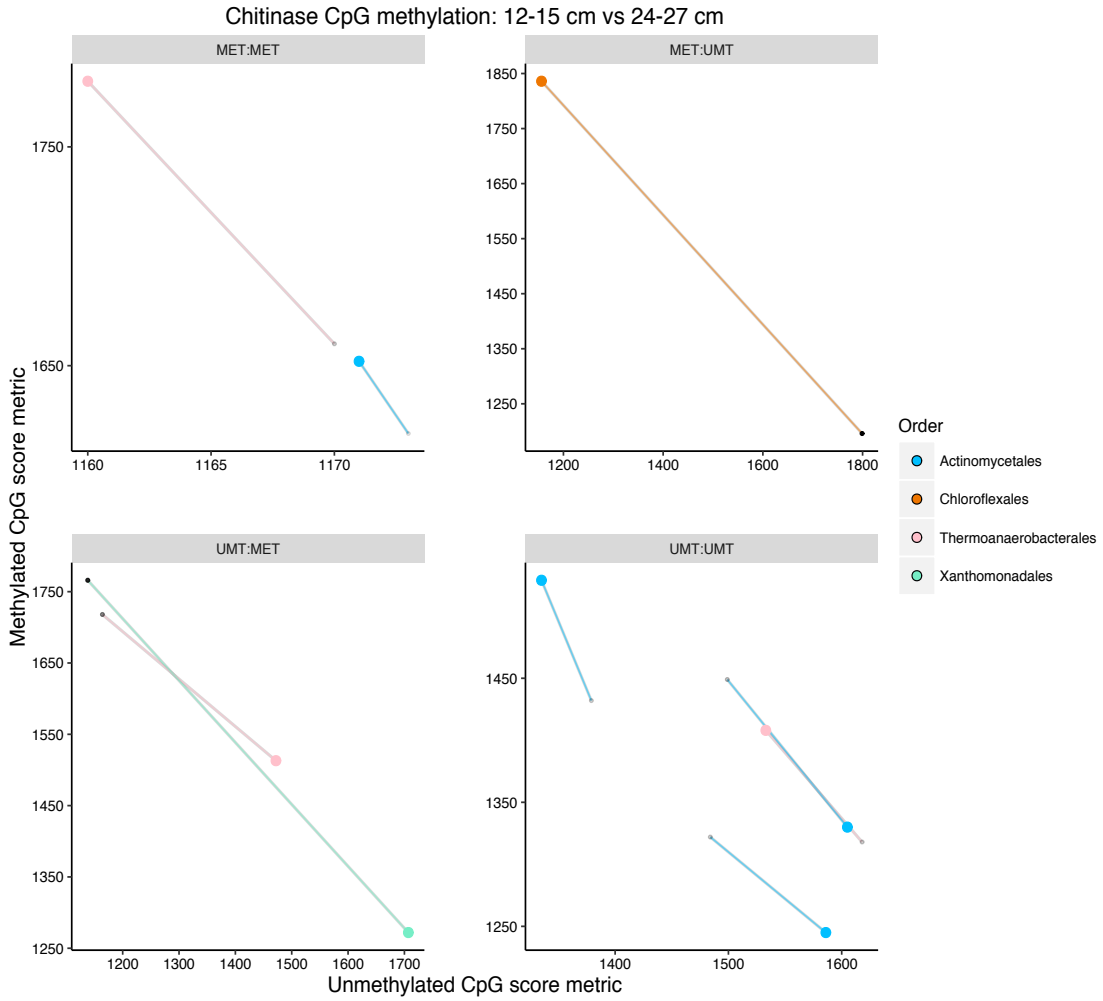


Figure 1.17: CpG methylation shifts for chitinases, 12-15 vs 24-27 cm. CpG sites are the same as those seen in Figure 1.14 (3-6 cm vs 12-15 cm), except displaying their shifts in methylation state from 12-15 cm to 24-27 cm depths, with the larger point representing the same coordinate position as the smaller point in the previous figure. CpG sites seen within the Actinomycetales are generally less methylated. The Chloroflexales CpG site has higher-magnitude binary shifts between high and low methylation states.

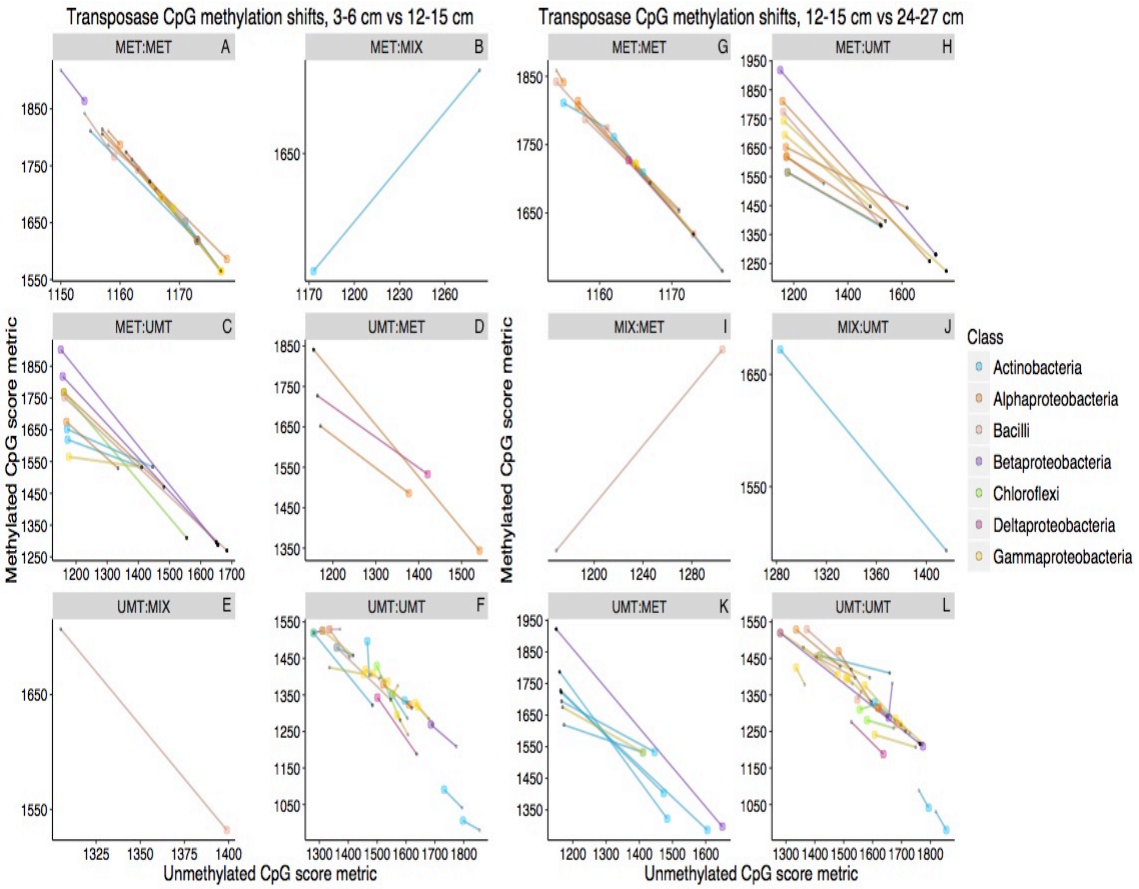


Figure 1.18: Transposase CpG methylation state shifts by class. Figure interpretation guidelines are the same as for Figures 1.13-1.17. CpG sites that shift to and from UNK states are excluded in this figure. Higher-magnitude shifts from methylated to unmethylated states are seen in the Chloroflexi and Betaproteobacteria. Gammaproteobacteria CpG sites remain in high or low-methylation states. Shifts to other states are more binary, with higher-magnitude shifts from fully methylated to unmethylated states, and vice-versa.

1.4 Discussion

DNA methylation is widespread in bacteria (Blow et al. 2016), and plays important roles in the protection of DNA in RM systems (Casadesús and Low 2006). Growing evidence indicates that many bacteria use methylation as a means of regulating gene expression (Marinus and Casadesus 2009; Løbner-olesen, Skovgaard, and Marinus 2005; Wion and Casadesus 2006; Reisenauer and Shapiro 2002; Srikhanta et al. 2005; Srikhanta et al. 2009; Low et al. 2001; Collier 2009; Brunet et al. 2011;Casadesus and Low 2008). The ability of DNA methylation to rapidly regulate gene expression makes it a potential acclimation mechanism for subsurface microbes faced with environmental stressors. Methylation would not require any evolution of gene regulators and could allow organisms that grow under surface conditions to adapt to the purported slow growth in deeper sediments. This study provides the first glimpse into dynamic CpG methylation of sediment microbial communities using an Illumina-based method that allows for total metagenome descriptions and changes in methylation.

Sediment populations observed through 16S rRNA gene and metagenomic gDNA sequencing do not appear to be governed highly by factors such as porosity and sediment porewater geochemistry, and this notion is supported by previous research of the Broadkill River (Cheng 2013) and other estuarine microbe communities (Koretsky et al. 2005). However, this study suggests there are shifts in community composition

more closely related to the drastic change in sediment age suggested by radionuclide constraints. 16S rRNA gene amplicons suggest that the 3-6 cm sample contains greater diversity than the deeper samples. It is likely that surface sediments are more aerobic than deeper sediments due to regular cycling and deposition, and these higher oxygen levels could be a factor in this higher diversity. Obligate anaerobes and facultative aerobes were observed in the shallow sample as well, suggesting communities poised to increase once environmental variables stabilize. The 3-6 cm sample also encompasses the transition zone from young, fresh sediment to older, established sediment at 4-5 and 5-6 cm, so overlap in communities was expected. Our data clearly shows increases in age from 3-6 to 12-15 cm, and the deeper depth of 24-27 cm is certainly older although our tests could not measure an exact age between 15-24 cm. A consideration to be taken into account is that sediment cores for geochemistry and radionuclide activity were not sampled concurrently with the core used for DNA extraction and sequencing. However, results from these cores support the presence of a drastically shifting downcore age gradient with higher anaerobic community potential at depth. Methane and porewater ion profiles are more varied within surface sediments, suggesting a bioturbated or tidally mixed region of fresh sediment in line with ^{7}Be and ^{210}Pb activity constraints.

A discrepancy in the relative abundance and dispersion of Actinobacteria sequences exists between 16S rRNA gene amplicon and metagenome results. Actinobacteria are shown to have high relative abundance through all three depths in

the metagenome results, yet 16S rRNA gene results show Actinobacteria present only in the 3-6 cm sample and existing at lower abundances. This could be due to HpaII-digested DNA being used in 16S gene amplicon libraries. Actinomycetes have been shown to possess multiple HpaII target sites within their 16S genes (Tolba et al. 2013), and this digestion could have affected PCR gene amplificiation resulting in a lack of reliable signatures. However, it should be noted that this is likely true for several taxa. 16S rRNA gene results indicate that anaerobic taxa increase with depth, but do not reliably indicate exact abundances or lack of signal. Due to this, Phylosift annotations are a more reliable indicator of marker gene-based taxonomic abundances.

Based on functional annotations in the assembled metagenome and community composition, the sediment community clearly transitions to a mostly anaerobic environment with increasing depth. The presence of methanogenic archaea and sulfate-reducing Deltaproteobacteria within the 12-15 and 24-27 cm samples suggest these communities support active anaerobic lifestyles (Oremland and Polcin 1982). The majority of taxa shifted by the 12-15 cm depth, but a few taxa were shared across all depths tested.

Methylation scoring results show that 82% of CpG sites had equivalent methylated or unmethylated states between depths. The majority of CpG sites remained in MET or UMT states for the Proteobacteria, Firmicutes, and Actinobacteria. Two possible explanations are presented here. First, DNA sequences with methylated CpG sites are propagated in the methylated state due to maintenance

methylation, leading to a higher number of conserved methylated sites. Second, non-methylated CpG sites are not substrates for maintenance DNA methylases, and as such tend to remain non-methylated (Casadesús and Low 2006). These results suggest a prevalence of conserved CpG methylation states across depths. Of the remaining 18% of CpG sites that were different, the majority occurred between high-methylation and low-methylation states. This is representative of the standard binary response associated with the concept of an epigenetic on/off switch (Marsh and Pasqualone 2014). Shifts between highly methylated (MET) and fractionally methylated (MIX) states suggest the presence of dynamic CpG sites that result in a mixed population of methylation states (Marsh and Pasqualone 2014). However, we cannot rule out the potential effect of whole genome duplication on methylation scoring, as newer genomes would contain fewer methylated bases. This ability to score fractional methylation of CpG sites within metagenomic DNA opens up future possibilities for better understanding how microbes in the environment can respond to various conditions and stressors, and for also potentially assessing the genomic age of intact cells from sediments. Assay results also provide evidence of dynamic CpG methylation shifts at the phylum level.

The methylation dynamics of individual CpG sites were analyzed for the Proteobacteria, Firmicutes, Bacteroidetes, Actinobacteria, and Chloroflexi. Actinobacteria CpG sites were shown to cluster together at higher densities in the UMT state at 24-27 cm, which explains in part why this phylum has the lowest

bootstrapped CV and SE. It should be noted, however, that 696 of 781 (89.11%) recovered CpG sites representative of this phylum were mapped to Actinomycetales, which is present in considerable abundance within all samples. The high representation of this order also accounts for a lower CV within the Actinobacteria population for each depth. However, when variances of the three depths are compared, Brown-Forsythe test results show the Actinobacteria exhibiting the highest variation between the three samples. Representative CpG sites for Proteobacteria and Firmicutes were from more diverse taxa, and potentially account for higher CV and SE within these phyla. There is a general trend of decreasing CV for both *met* and *umt* scores with depth, and this is influenced by the overall trends towards bimodal score distributions. The vast majority of methylation shifts between MET and UMT states for Actinobacteria, Proteobacteria, and Firmicutes were small. Multiple CpG sites experience these low-magnitude shifts to shared states of methylation loss and gain. These shared behaviors in methylation shifts are potentially due to the increasing age of genomes with sediment depth. More normal distributions of *met* scores indicate hemimethylation within the 3-6 cm sample, which is indicative of cell growth and newer genomes. Shifts towards binary methylation distributions are likely due to decreased hemimethylation, resulting in most methylation being attributed to maintenance factors.

CpG methylation states were also shown to vary for specific genes, including chitinase and transposase. Chitin, the long-chain polymer of the glucose derivative N-

acetylglucosamine, is a highly abundant renewable polymer in marine ecosystems present in bivalves, cephalopods, arthropods, fungi (Bhattacharya et al. 2007) and diatoms (Durkin et al. 2009). There are no reports of quantitatively significant long-term accumulation of chitin in nature, implying efficient degradation and turnover (Tracey 1957; Gooday 1990). Most chitin is produced at or near the surface of aquatic environments (Gooday 1990), and previous studies provide evidence of bacterioplanktonic chitinase presence and activity within the coastal estuarine waters of Delaware Bay (Cottrell et al. 1999; Cottrell et al. 2000). Chitin hydrolysis rates are higher within the water column than in the sediment, and sediments only contain traces of chitin (Souza et al. 2011). However, chitinolytic bacteria such as the Actinomycetes are widely distributed in sediment environments (Bhattacharya et al. 2007; Souza et al. 2011) and are responsible for converting this insoluble source of carbon and nitrogen into a form that can be utilized by the entire marine food web (Gooday 1990; Gooday et al. 1991). It has been previously noted that within the first 10 cm of sediment, chitin is rapidly removed from an estuary (Gooday 1990). CpG site methylation was characterized in chitinase genes. From the shallow depth to 12 cm, the most abundant shifts are from methylated to unmethylated states. Notably, these changes are occuring in lineages of Actinobacteria and Thermoanaerobacter, both known as anaerobic lineages that utilize chitin. While some chitinases also annotated for these groups remain methylated, other CpG sites certainly shift. I interpret the variations in these signals to suggest regulation, and not just a signature

of cellular replication, considering that methylation losses are greater as these genes enter the anaerobic, 12-15 cm environment. These same CpG sites are again methylated within the 24-27 cm depth as they leave the assumed zone of available chitin. While I did not concurrently measure chitin, the often noted correlation between cultivable chitinolytic bacteria and chitin abundances suggests that this process is one that would not be maintained if chitin were not present. The evidence for anaerobic organisms only removing methylation from chitinase CpG sites within the anaerobic sedimentary horizons suggests that this is a valid glimpse into methylation-based regulation of an energetically costly process. This study has provided an initial glimpse into how marine sediment microbes potentially utilize DNA methylation to regulate biogeochemical processes that are vital for nutrient cycling.

While DNA methylation can silence gene transcription, its can also regulate transposons and the expression of associated transport genes. Transposons are common within microbes, and usually consist of sequences < 2000 bp known as insertion sequences (IS) (Voet et al. 2008). An IS usually contains a transposase (Nagy and Chandler 2004), which binds to the end of the transposon and catalyzes its movement within a genome (Voet et al. 2008). Transposition is regulated due to its potential to activate genes upon upstream insertion, as well as damage the integrity of a genome (Nagy and Chandler 2004). DNA methylation has been shown to regulate transposition in bacteria, and differential N6-adenine methylation at GATC sites is

known to regulate IS expression and DNA binding in *E. coli* (Casadesús and Low 2006). While much of the literature concerning bacterial transposase regulation by DNA methylation is focused on Dam MTase methylation at GATC sites in *E. coli* (Dodson and Berg 1989; Reznikoff 1993; Roberts et al. 1985; Yin et al. 1988; Spilemann-Ryser et al. 1991), the regulatory mechanisms of one model organism do not necessarily apply to the entire bacterial domain. CpG methylation has been shown to silence plant transposons (Miura et al. 2001), and could potentially do the same in microbes due to the conservation of this methylation pattern. There is evidence for differential methylation of multiple CpG sites within sediment bacteria transposases. A lack of methylation within CpG sites or shifts into hemimethylated states could also hint at the potential activity of transposases. The regulation of intracellular transposases and transposon mobility could play a role in rapid acclimation responses by influencing transcriptional activity and acting as mobile genes that can be inserted into a genome. Horizontal gene transfer is speculated to occur in estuarine sediment microbes (Angermeyer et al. 2016), and extracellular transposases could play a potential role in this process due to substantial numbers of horizontally-transferred genes in several bacteria species being attributed to foreign DNA such as transposons (Ochman et al. 2000). Due to the known influence of DNA methylation within transposons and the results of this study, we speculate that DNA methylation could potentially act as a regulator of transposition within the subsurface.

The Illumina-based methylation-scoring assay utilized in this study was originally designed for determining DNA methylation in eukaryotes, and has been used to identify dynamic CpG shifts in the Antarctic polychaete *Spiophanes tcherniai* (Marsh and Pasqualone 2014). While this approach does not give a complete view of bacterial and archaeal DNA methylation due to the high diversity of MTase target sequences and methylated nucleotides in bacteria and archaea, it does provide community-level insight into the dynamic behavior of a well-known and conserved methylation site. Since adenine methylation is considered to be a main driver in bacterial gene regulation (Ratel et al. 2006), the environmental microbiology application of this assay could be augmented by methods to quantify adenine methylation. Proportions of N6-adenine methylation at GATC sites could be attained by digesting extracted DNA with DpnI restriction endonuclease, which cuts at the recognition site Gm6A^TC (Lacks 1980). DpnI differs from most restriction endonucleases in that it cleaves at a methylated instead of a non-methylated recognition site. Modifying the assay for use with paired-end sequences would aid in assembly and alignment, and would improve the overall quality of the data set. Modern sequencing and computing methods allow for the reconstruction of genomes obtained from the environment (Baker et al. 2016; Baker et al. 2015; Seitz et al. 2016; Eisen et al. 2002; Singer et al. 2013; Tully et al. 2014), allowing for detailed insights into the mechanisms that drive microbial communities and individuals alike. A benefit of this Illumina assay is that it requires less DNA than single-molecule approaches. Future modifications tailored for metagenomic DNA could pave the way for the reconstruction of dynamic methylation profiles within genomes obtained from the environment.

Chapter 2

TRENDS IN EPIGENETIC REGULATORY SIGNALS FOR KNOWN BACTERIAL GENOMES

2.1 Introduction

Transcription initiation and gene expression are heavily influenced by the interactions between DNA regulatory sequences and transcription factors. These interactions are not governed solely by the recognition of particular nucleotide motifs, but are also affected by multiple processes that affect the conformation and intermolecular forces of a DNA sequence. DNA methylation and DNA curvature are two conserved epigenetic processes that play important roles in microbial gene structure and transcriptional regulation.

DNA methylation is conserved within bacteria and archaea, and involves the modification of adenine or cytosine by a DNA methyltransferase (MTase) into N6-methyladenine (n6A), 5 methylcytosine (m5C), and/or N4-methylcytosine (n4C) within a specific target sequence (Kumar and Rao 2012). Although DNA methylation is conserved, the target sequences of n6A, m5C, and n4C-specific MTases vary greatly among different taxa (Wion and Casadesus 2006; Casadesús and Low 2006; Butkus et al. 1987; Klimasauskas et al. 1990; Klimasauskas et al. 1989; Cheng 1995; Blow et al.

2016, Murray et al. 2012). These target sequences are generally scarce in microbial genomes due to their length and complexity (Wojciechowski et al. 2012). While many MTases are associated with restriction-modification (RM) systems, methylation activity by non-RM-associated MTases (orphan MTases) is shown to have a significant presence in microbes (Blow et al. 2016). Since methylation is known to occur in or around regulatory regions (Casadesús and Low 2006), characterizing spatial motif frequency highs and lows could aid in identifying potential regulatory sites. Methylome analysis of 213 bacterial genomes via single molecule, real-time sequencing shows a significant enrichment of unmethylated MTase target sites within regulatory regions, supporting the involvement of these sites in regulatory processes (Blow et al. 2016). While methylation is associated with an "on/off" transcriptional switch (Low et al. 2001, Casadesús and Low 2006), the secondary effects imparted upon DNA by methylation highly influence transcription. Methylation can induce or increase the curvature of DNA (Kravatskaya et al. 2011; Ratel et al. 2006; Wion and Casadesus 2006; Asayama and Ohyama 2005; Diekmann 1987; Engel and von Hippel 1978), and this can affect the binding of regulatory proteins due to steric hindrance or alteration of the DNA structure (Casadesús and Low 2006).

DNA curvature is caused by the intrinsic intermolecular forces of the nucleotide sequence itself, or by external forces such as protein binding (Asayama and Ohyama 2005). Several methods exist for analyzing the curvature of a DNA sequence, including X-ray diffraction (Dlakic et al. 1996), FRET (Parkhurst et al. 1996), LRET

(Heydun et al. 1997), and TEB (Vacano and Haerman 1997) spectroscopies, and electrophoretic ligation ladder assays (Ross et al. 1999). Intrinsic curvature is determined by the nucleotide sequence, and certain patterns can have greater effects on curvature. Fourier analysis has shown a correlation between intrinsic curvature and pattern periodicity (Gabrielian and Pongor 1996). Statistical analysis of laboratory results has led to the development of several DNA curvature models based on nucleotide periodicities and the presence of features such as dinucleotide stacks (Bolshoy et al. 1991; Goodsell and Dickerson 1994). The periodicity of stretches of adenines or thymines (termed A-tracts or T-tracts) is assumed to influence how planar or curved a DNA strand is. A-tract and T-tract periodicities < 10.5 bp induce left-handed superhelical conformations, while periodicities > 10.5 bp result in right-handed conformations. Periodicities equal to 10.5 bp result in a planar curvature conformation (Asayama and Ohyama 2005). The axial path of a DNA strand is further determined by the presence of dinucleotide stacks that influence three Eulerian angles: the deflection angle (wedge angle), the helical twist, and the direction of the deflection. The dinucleotide stacks AA/TT, AG/CT, CG/CG, GA/TC, and GC/GC are estimated to have large wedge angles, and are considered to have greater effects on the deflection of the DNA helical axis and curvature of the sequence (Bolshoy et al. 1991).

Intrinsic curvature imparted by the presence or absence of A-tracts, T-tracts, and certain dinucleotide stacks can enhance or inhibit transcription by altering

interactions with transcription factors (Asayama and Ohyama 2005). The presence and conservation of curved DNA structures in or around promoters and origins of replication suggest that curvature plays an important role in the regulation of basic genetic processes within microbes (Jauregui et al. 2003; Asayama and Ohyama 2005; Kanhere and Bansal 2005; Meysman et al. 2014). Heightened DNA curvature within promoter regions has shown to raise the efficiency and DNA binding of an RNA polymerase (RNAP) (Matsushita et al. 1996) by increasing the number of contacts between RNAP and promoter DNA (Pérez-Martín and Espinosa 1994), an important step in transcription initiation (Asayama and Ohyama 2005). The structure of a DNA sequence is not necessarily optimal for the surface of an RNA polymerase, but RNAP counters this by altering promoter DNA curvature into a left-handed superhelical conformation as it wraps around RNAP upon binding (Coulombe and Burton 1999). Aside from the core promoter region, upstream and downstream structural features also influence interactions with transcription factors. While the curvature of the core promoter region is linked to its transcriptional activity and efficiency, previous studies have shown that intrinsic DNA curvature upstream of promoters is related to the activity of the promoter region (Perez-Martin et al. 1994; Katayama et al. 1999; Bracco et al. 1989; McAllister and Achberger 1989; Travers 1989; Liu-Johnson et al. 1986; Matsushita et al. 1996). Intrinsic curvature is lower downstream of the core promoter, and this is linked to curvature being more dependent upon external forces such as protein binding within these regions (Asayama and Ohyama 2005).

Due to the known relationships between DNA methylation and curvature, a question arises concerning the widespread prevalence of these phenomena in relation to potential regulatory regions. Since heightened MTase motif frequency and curvature are known to occur near regulatory regions, identifying spatial highs and lows on a large scale can provide insight into conserved epigenetic trends. In this study we identify global trends in CpG and GATC site frequency, as well as intrinsic curvature, for intergenic upstream and intragenic downstream regions of known genes within sequenced bacterial reference genomes.

2.2 Materials and Methods

2.2.1 TBLASTN alignments of DNA methyltransferase orthologs

A total of 1010 complete RefSeq bacterial genomes representative of classes abundant within the metagenome described in Chapter 1 were obtained from NCBI for analysis of structural patterns within promoter and coding regions, as well as the identification of potential methyltransferase orthologs. *Escherischia* genomes were excluded since methylation and methyltransferases are well documented in this genus. All genomes were derived from chromosomal replicons. Genomes were formatted into a nucleotide BLAST database for local alignment with BLAST+ (Camacho et al. 2009). Amino acid sequences for 151 bacterial C5-cytosine, N4-cytosine, and N6-adenine-specific DNA methyltransferases were obtained from RefSeq, SwissProt, European Molecular Biology Laboratory, GenBank, and Protein Data Bank databases. Recognition sites for these proteins were parsed from the REBASE database (Roberts

et al. 2010). Taxonomic lineages for methyltransferase origin organisms and complete genomes were parsed from the NCBI Taxonomy database (Sayers et al. 2009; Benson et al. 2009). Local alignments of subject genomes and methyltransferase queries were performed with the BLAST+ suite v.2.2.30 (Camacho et al. 2009) using TBLASTN (protein compared against nucleotide subject sequence translated in six reading frames) with the BLOSUM62 substitution matrix (Henikoff and Henikoff 1992). An e-value threshold of 1e-4 was imposed during alignment. Results were filtered to exclude hits of genus-level similarity between the genome subject and the methyltransferase organism of origin.

Table 2.1: Representative genomes and extracted gene fragments

Class	**Number of genomes**	**Gene fragments**
Actinobacteria	119	342,868
Alphaproteobacteria	155	290,146
Bacilli	192	336,988
Betaproteobacteria	133	278,505
Clostridia	82	163,380
Deinococci	14	22,641
Deltaproteobacteria	37	103,613
Gammaproteobacteria	278, 212*	630,659, 536,392*
Total	1010	2,168,800
	944*	2,074,533*

* Indicates number of genomes and gene fragments for motif and curvature analysis.

2.2.2 Methylation motif frequency and intrinsic curvature of gene upstream and coding regions

I utilized a series of custom Perl, Python, and R scripts to perform a class-level analysis of global trends in DNA methylation-associated motif frequency and intrinsic curvature for gene upstream promoter regions and coding sequences (CDS). The density of methyltransferase target sites was used to identify spatial increases and decreases in target site composition along upstream promoter and downstream coding regions. Gene fragments include a -200 bp upstream promoter and a coding sequence (CDS) +350 bp downstream from the transcriptional start site (TSS). These data only include fragments that do not overlap with other CDS sequences, meet a required length (550 bp), and whose first three nucleotides of the CDS are not a stop codon (TAA, TAG, TGA).

GATC, CpG and mirror GpC motif frequencies were calculated for 500 bp sequences extending -175 bp upstream and +325 bp downstream from the TSS (position 0) for each gene fragment. Calculations utilized a standard weighted moving average with a 51 bp window. Within each window, the score attributed to a CpG, GpC, or GATC motif was weighted based on the motif's position relative to the central nucleotide in the window.

CpG or GpC motif frequencies weighted to their position within a window of size ϖ were calculated as:

$$CpG|GpC_{freq} = \frac{1}{1 + |\left(\frac{\varpi}{2}\right) + [CG|GC]_{pos}|}$$

$[CG|GC]_{pos}$ is the position of the CpG or GpC motif within the window nucleotide string.

GC content was calculated as:

$$GC = \frac{G_{freq} + C_{freq}}{\varpi}$$

where G_{freq} and C_{freq} are the frequencies of guanine and cytosine, respectively.

CpG frequency normalized to GC content was calculated as:

$$CpG_{norm} = (\frac{CpG_{freq}}{\varpi})/(GC + 1e^{-5})$$

Expected GATC motif frequency was calculated as:

$$GATC_{exp} = \left(\frac{GA_{freq} * AT_{freq} * TC_{freq}}{G_{freq} * A_{freq} * T_{freq} * C_{freq}}\right) * (\varpi - \mathrm{GATC}_{len} + 1)$$

where ϖ is the window size, $GATC_{len} == 4$ (aka the length of sequence GATC), and $[GATC]_{freq}$ is the nucleotide or dinucleotide frequency.

DNA curvature was calculated for gene fragments using the nearest-neighbor dinucleotide wedge model (Bolshoy et al. 1991).

Results were structured into Hierarchical Data Format, version 5 (HDF5) using a custom Python script. In the case of this project, being able to select certain data sets allowed for this large quantity of data to be worked with in R, especially since R stores files in memory. Data sets were organized by taxonomic rank (phylum to NCBI accession number), with a data set for each CDS.

2.3 Results

2.3.1 TBLASTN alignments of DNA methyltransferase orthologs

A total of 4065 TBLASTN alignments passing selection criteria were obtained for both orphan and RE-paired DNA methyltransferases. Potential orthologs of the m5C-specific orphan methylase Dcm were observed within all classes except the Deinococci. Overall results for m5C-specific MTase alignments show that query coverage was higher within Actinobacteria, Gammaproteobacteria, Betaproteobacteria, and Bacilli, although percent similarity ranged from 21.07% to 100% (Figure 2.1). Mean alignment coverage of Dcm hits was high, with a range of 93% to 98%. Amino acid similarity for these classes was lower, with a range of 53.28% to 68.62%. These alignments suggest that m5C methylation activity at 5'-CCWGG-3' sites is conserved among these classes. Potential Dam orphan methylase orthologs were observed in all classes other than the Deinococci. While alignment coverage was high (> 93%), amino acid similarity was generally lower (between 47-64%). These scores were higher within the Gammaproteobacteria, and could be due to the prominence of Dam GATC methylation within this class. The RE-paired type II MTases M.MboIA and M.MboIB, which also methylate GATC motifs, were seen to have significant alignments within the Clostridia and Gammaproteobacteria (Supplementary Table 5). Overall, there appears to be potential for N6-adenine methylation at GATC sites within these seven classes based on these significant

alignments. N4-cytosine specific MTase alignments were seen in Actinobacteria, Alphaproteobacteria, Bacilli, Betaproteobacteria, Clostridia, and Gammaproteobacteria. Most of these alignments had low similarities. M.ScaI aligned to members of Actinobacteria, Bacilli, and Clostridia, although these alignments had low similarity. M.PvuII was shown to have higher similarity and coverage in the Betaproteobacteria and Gammaproteobacteria (Supplementary Table 8).

2.3.2 Methylation motif frequency and intrinsic curvature of gene upstream and coding regions

A total of 2,168,800 gene fragments were extracted from 3,248,452 genes specified within general feature format (GFF) files. Of the gene fragments excluded from analysis, 580,720 (53.79%) were ignored due to overlaps with a previous CDS, while 458286 (42%) were shorter than 550 bp and 40,646 (4%) contained invalid start codons. It should be noted that the number of extracted Gammaproteobacteria gene fragments was reduced to only include those from orders present within the Oyster Rocks metagenome (Acidithiobacillales, Aeromonadales, Alteromonadales, Chromatiales, Enterobacteriales, Methylococcales, Oceanospirillales, Pseudomonadales, Thiotrichales, and Xanthomonadales) (Table 2.1). Motif frequency and intrinsic curvature suggest conserved structural features within these eight classes. However, these features are highly dependent upon nucleotide position. Certain trends were seen upstream of the core promoter, within the core promoter region, and in the downstream CDS.

2.3.2.1 Upstream of the core promoter

Regions upstream of the core promoter region (-175 to -35 bp) are characterized by gradually increasing curvature (Figure 2.9). GATC tetranucleotide frequency appears to remain steady or decrease (Figure 2.4). $GATC_{expected}$ (determined by GA/AT/TC dinucleotide frequencies) appears stationary at the upstream region while sharply increasing close to the core promoter (Figure 2.2). GC content decreases in this region until it reaches a minimum within the core promoter region. However, GC content increased near the core promoters of Bacilli and Clostridia (Figure 2.8). This increased GC content, but lack of GA/AT/TC motifs, could signify a higher prevalence of A-tracts or T-tracts, as well as other influential dinucleotide stacks within this upstream region that promote increased curvature. The directional angles of GA and AT dinucleotides are 0 ° and 120 °, which could result in a more bent conformation. CpG frequency normalized to GC content is higher further upstream, but decreases closer to the core promoter (Figure 2.6). The coefficient of variation (CV) for GATC and CpG motif frequencies is generally higher or gradually increases (Figure 2.5, 2.7), suggesting that there is higher variability in the potential DNA methylation motifs present within this region.

2.3.2.2 Core promoter region

An increased frequency of GA/AT/TC dinucleotides likely accounts for the rise in curvature within the core promoter region (-35 to 0 bp (transcription start site)),

as the GA/TC dinucleotide stack has a large wedge angle. There is a dip in GC content for all classes except the Clostridia and Bacilli, and this increase in AT content could increase the probability of A-tracts or T-tracts occurring along with AA/TT dinucleotide stacks. The increase of these features could account for the peak in curvature seen within this region. The higher curvature of this region can also be attributed to the preferential binding of RNAP to curved DNA.

2.3.2.3 Downstream coding region

The downstream CDS region spanning from 0 to +150 bp appears to have conserved spatial patterns. Frequency of GA/AT/TC dinucleotides is highest around +35. CpG frequency and GC content are higher within the +35 to +150 bp substring, and GA/AT/TC dinucleotide frequency decreases. Intrinsic curvature also dips drastically in this region, resulting in a more planar region. The increased GC content at this site likely contributes to a decrease in the probability of A-tracts and/or T-tracts, and an increased probability of CG/CG and GC/GC dinucleotide stacks. These dinucleotide stacks are considered to have greater effects on the deflection of the DNA sequence due to their higher wedge angles (Bolshoy et al. 1991). The directional angles of the CG and GC dinucleotides are respectively 0° and 180°, and these angles result in a more planar DNA conformation. The CpG normalized CV is lower overall within the CpG valley, suggesting that individuals within these classes have more conserved CpG frequencies within this region. Higher GC content could result in

decreased A and T-tract periodicity due to a more balanced nucleotide distribution within the downstream region.

There is a reduced probability of matching a CpG site in the +35 to +150 region due to lower GC content, and this would decrease the prevalence of CpGs located closer to the central nucleotide of the sliding window. Inversely, GC content increases within the valley and the CV for CpG normalized to GC content is lower (Figure 2.8 and 2.7). This would impart a greater chance of matching a CpG site with consistently higher weight, as an increase in CpG frequency would allow for more sites closer to the center of the sliding window.

A structural feature is seen around +150 bp characterized by a sharp increase in curvature, a slight increase in GA/AT/TC nucleotide frequency, a sharp decrease in CpG frequency, a sharp peak in weighted GATC tetranucleotide motif frequency, and a decrease in GC content. These features potentially indicate the presence of a conserved downstream regulator containing a GATC site. Downstream GATC sites are methylated in *E. coli* and as a result increase the efficiency of transcription factor DNA binding (Casadesús and Low 2006). A similar pattern at this conserved site could act to regulate RNAP. Decreasing curvature and increasing GC content characterize gene sequences downstream of the +150 peak to the +325 position. CpG frequency also gradually increases within this region.

2.3.2.4 Actinobacteria

Actinobacteria has the highest GC content among the eight classes analyzed (Figure 2.8). They had one of the highest normalized CpG frequencies, as well as having consistently low CV for this metric, indicating a higher similarity of CpG frequency among members of this class (Figure 2.6). Actinobacteria displayed the lowest mean GA/AT/TC dinucleotide frequency of classes containing a large number of genomes. There was a higher $GATC_{expected}$ CV, due to potentially higher probability of CpG or GpC dinucleotides imparted by higher overall GC content. While this class followed conserved trends in curvature structure, it had the lowest overall intrinsic curvature (Figure 2.9).

2.3.2.5 Alphaproteobacteria

The Alphaproteobacteria had higher $GATC_{expected}$, and was only surpassed by the AT-rich Clostridia and Bacilli. The $GATC_{expected}$ mean based off dinucleotide frequency is high, and the $GATC_{expected}$ CV at the core promoter region is lowest of all classes. This class also had the highest GATC motif frequency and the lowest GATC motif CV, especially within the core promoter region. Overall GC content is higher, ~60%. This could account for lower overall intrinsic curvature, potentially due to decreased A and T-tract periodicities that result from more even nucleotide frequencies.

2.3.2.6 Bacilli

Bacilli were observed to have AT-rich genomes with high intrinsic curvature. This class had the highest expected GATC frequency based off GA/AT/TC nucleotides, and followed conserved trends by having a peak in this metric within the core promoter region. Frequency of the GATC tetranucleotide was lower overall, however. TBLASTN alignments suggest that the Bacilli have potential orthologs for Dam, M.FokI, and M.KpnI MTases, with respective recognition sites of GATC, GGATC, and GGTACC. The presence of the latter two MTases suggests that the Bacilli might utilize these target motifs in transcriptional regulation, and their presence within the core promoter region could account for the increased $GATC_{expected}$ frequency. It should be noted that CV increases for $GATC_{expected}$ in the core promoter region, signifying that dinucleotide frequencies are more varied overall in the -35 to 0 region.

2.3.2.7 Betaproteobacteria

Representative genomes of the Betaproteobacteria had lower intrinsic curvature and higher GC content. Members of this class displayed higher GATC tetranucleotide frequencies compared to more AT-rich classes. The potential conservation of Dam orthologs within this class based on TBLASTN alignments could explain the higher GATC tetranucleotide frequency. Average $GATC_{expected}$ frequencies

have a much higher peak at the core promoter region, indicating that GATC could be a common regulatory motif within this site.

2.3.2.8 Clostridia

The Clostridia were observed to have a higher $GATC_{expected}$ frequency within the core promoter region, but the CV for this class was shown to increase at this site. GATC tetranucleotide frequencies were lowest for this class, and the CV for this metric was highest overall at the core promoter region and potential +150 bp regulatory site. This could indicate that the Clostridia methylate a variety of target motifs at the core promoter region, and TBLASTN analysis suggests that several N6-adenine-specific MTase orthologs are present within this class (Supplementary Table 5). CpG methylation was shown to be lowest overall within this class. This class has the highest intrinsic curvature, and is likely related to AT-rich genomes.

2.3.2.9 Deinococci

GA/AT/TC dinucleotide frequency was lowest within the Deinococci, and is possibly linked to the lack of alignments to N6 adenine-specific MTases targeting GATC motifs (Supplementary Table 5). Based off these results, it is possible that the Deinococci tend to utilize other methylation motifs within their genomes aside from GATC. High-scoring significant alignments were seen for M.AbrI, M.BstVI, and M.XhoI, all of which have a 5'-CTCGAG-3' target sequence. Increases in this target

sequence at the core promoter region could partially account for the increase in GA/AT/TC dinucleotide frequency, as it contains the dinucleotides GA and TC. This class has one of the highest mean GC contents (Figure 2.8) and lowest intrinsic curvatures (Figure 2.9), which could be due to a lack of A-tracts, T-tracts, and/or AA/TT dinucleotide stacks. It should be noted that the $GATC_{expected}$ CV is highest for the Deinococci (Figure 2.3) and is likely caused by a lower number of genomes (Table 2.1). However, the dip in $GATC_{expected}$ CV mirrors the increased $GATC_{expected}$ frequency at the core promoter region **(**Figure 2.2), suggesting that members of this class have more similar GA/AT/TC dinucleotide frequencies at this site.

2.3.2.10 Deltaproteobacteria

Deltaproteobacteria were shown to have lower CpG and GATC motif frequency than the Alphaproteobacteria, although their GC contents are very similar. N6-adenine-specific orthologs were present for Dam methylase, but the majority of high-scoring TBLASTN hits for this MTase family were for RE-paired MTases having 5'-CTCGAG-3' target motifs (Supplementary Table 5). It is possible that GATC is a utilized methylation motif in this class, although other adenine methylation motifs might be prevalent.

2.3.2.11 Gammaproteobacteria

The Gammaproteobacteria had the highest intrinsic curvature of all the Proteobacteria classes examined in this study. GATC methylation is well documented in this class; however, GATC tetranucleotides were less frequent compared to the Betaproteobacteria and Alphaproteobacteria. TBLASTN analysis shows that potential orthologs for several N6-adenine, m5-cytosine, and N4-cytosine specific MTases are seen within this class. Due to this, the methylation specificities of this class are likely to be far more varied than Dam GATC methylation.

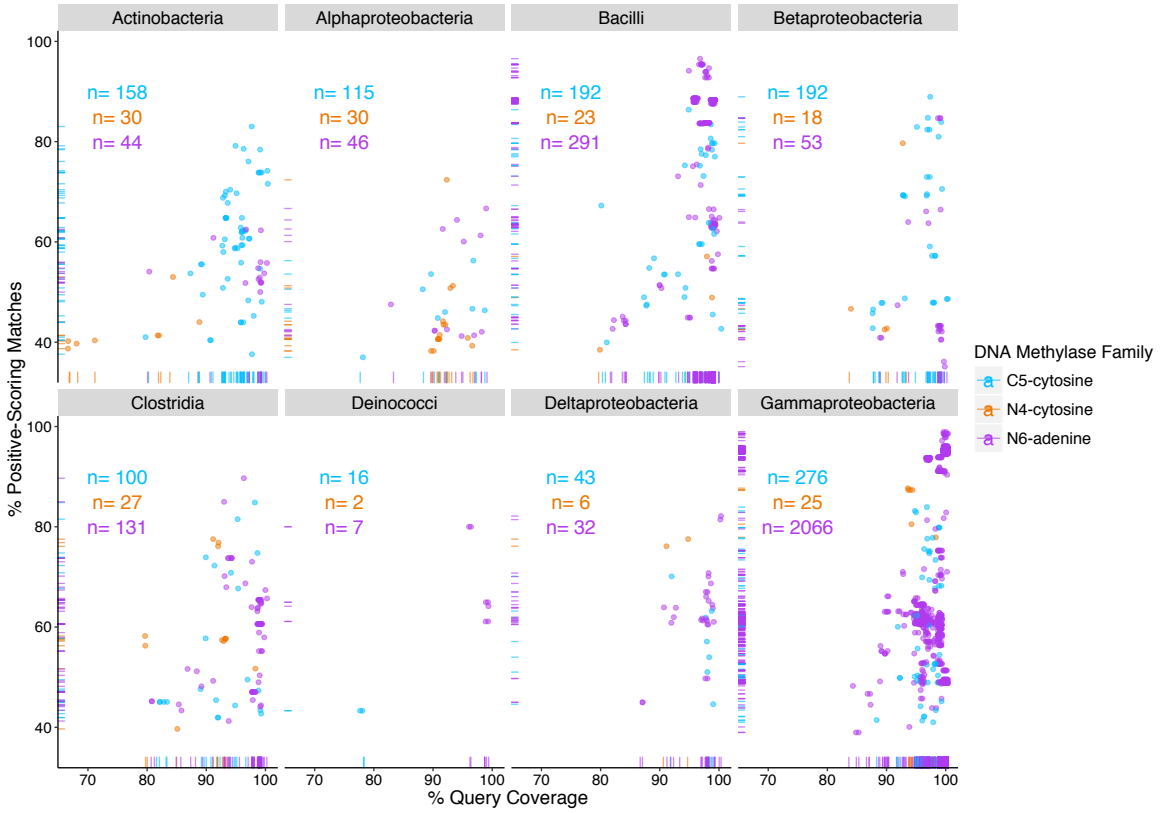


Figure 2.1: TBLASTN DNA methyltransferase alignment results. Hits were selected for those with alignment lengths greater than or equal to 90% of the query length. The majority of n6A MTases are highly conserved in the Gammaproteobacteria, potentially due to many known n6A MTases originating from class members such as *E. coli*. M5C MTases had higher coverage and similarities for Actinobacteria, Bacilli, and Gammaproteobacteria. Alignments to n4C MTases had generally lower similarity and query coverage.

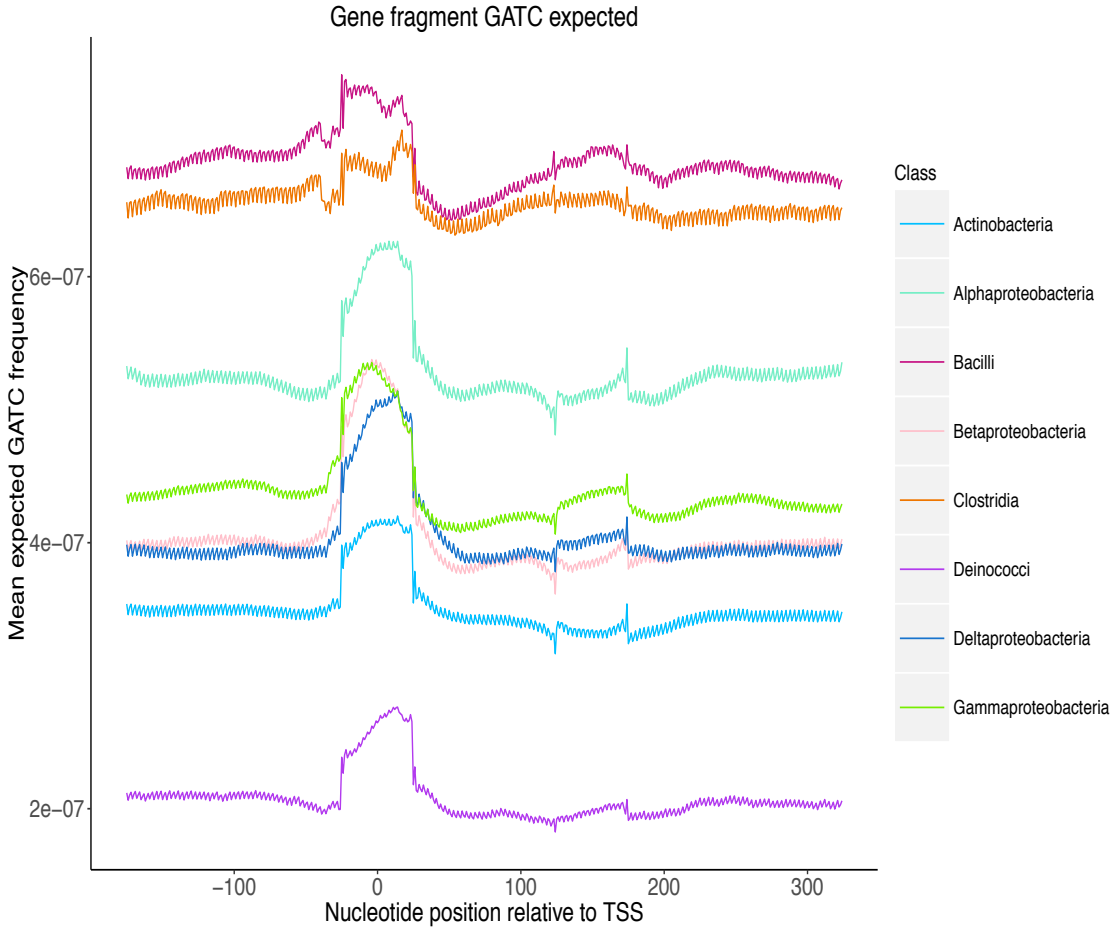


Figure 2.2: Overall mean gene fragment expected GATC. The transcription start site (TSS) is at point 0. Expected GATC motif frequency was calculated as a measure of the frequency of GA/AT/TC dinucleotides within a 51 bp sliding window. There is a conserved increase in the frequency of these dinucleotides from -35 to +35 bp.

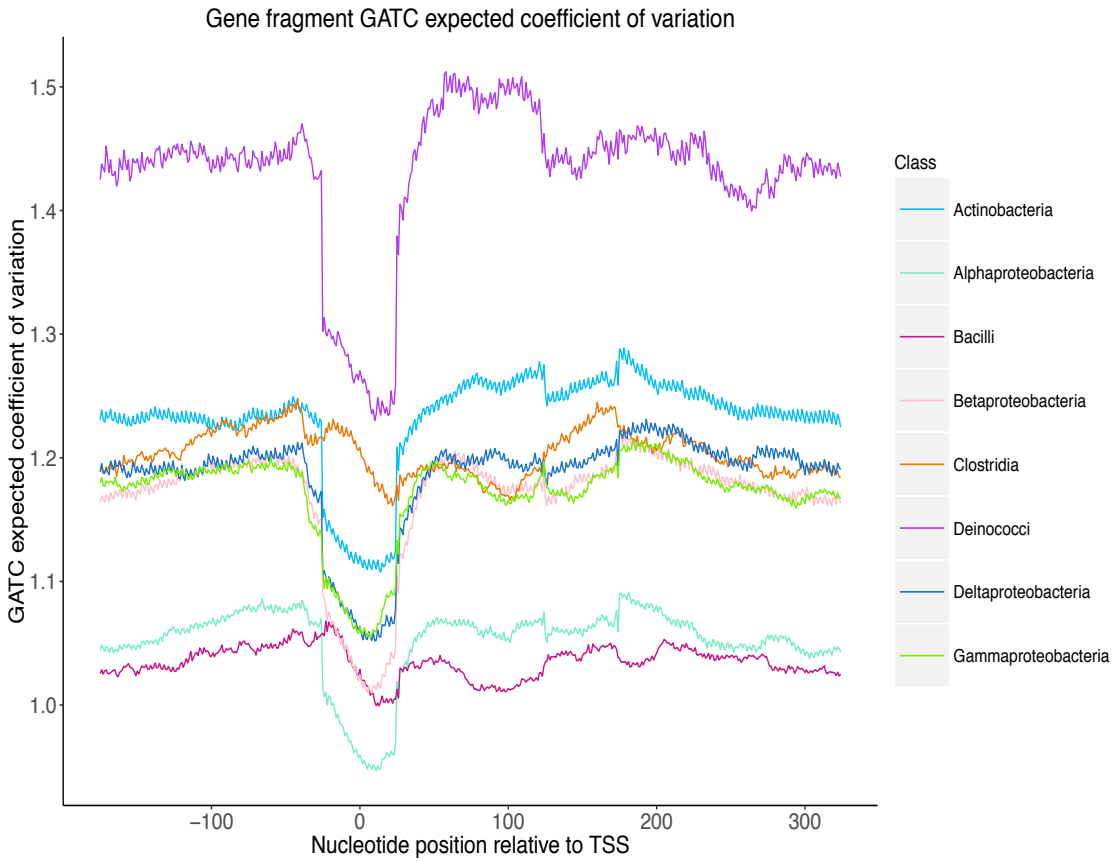


Figure 2.3: Coefficient of variation for expected GATC. The transcription start site (TSS) is at point 0. A higher coefficient of variation (CV) indicates that there is more variability in expected GATC frequency at a site, while a lower CV means that expected GATC frequency is more consistently the same amongst members of a class. There appears to be a lower CV at the core promoter region, indicating that the increase in expected GATC at this site is more consistently the same among members of a class.

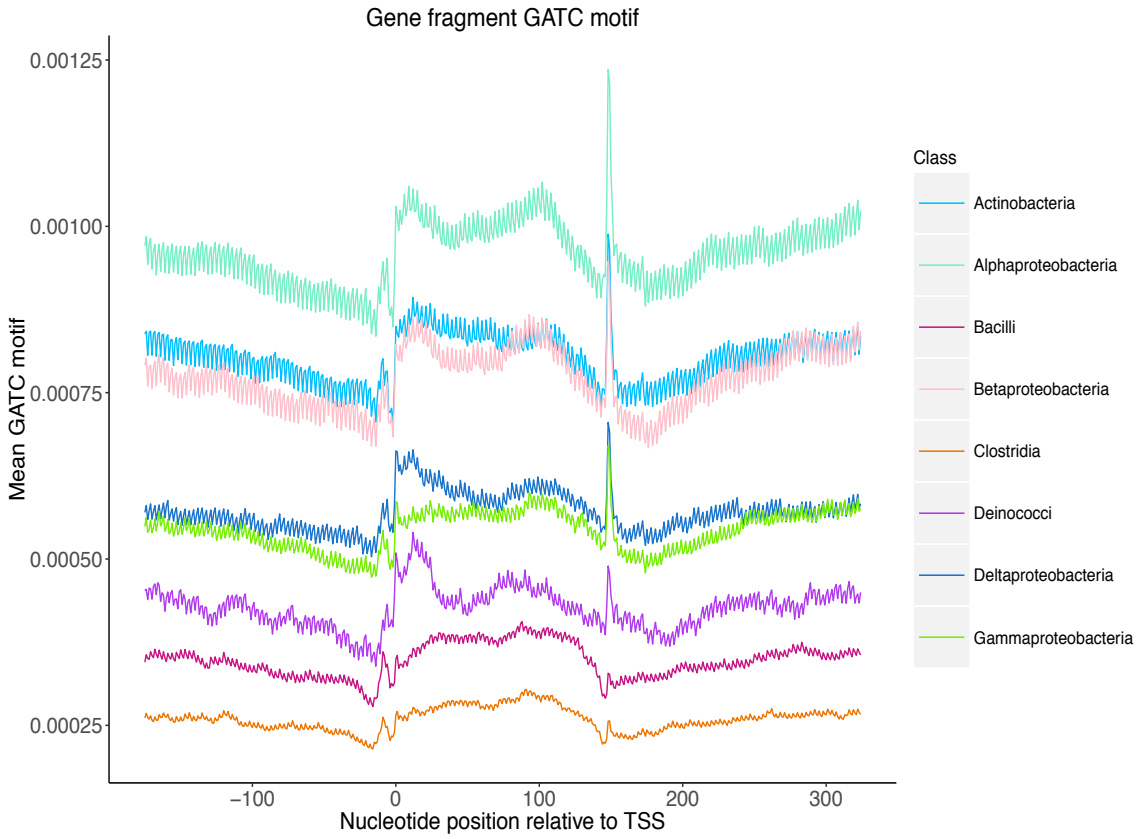


Figure 2.4: Overall mean gene fragment weighted GATC frequency. The transcription start site (TSS) is at point 0. The score attributed to a GATC site is weighted based on its position relative to the central nucleotide of a 51 bp sliding window. These are exact pattern matches, as opposed to a score based on dinucleotide frequencies.

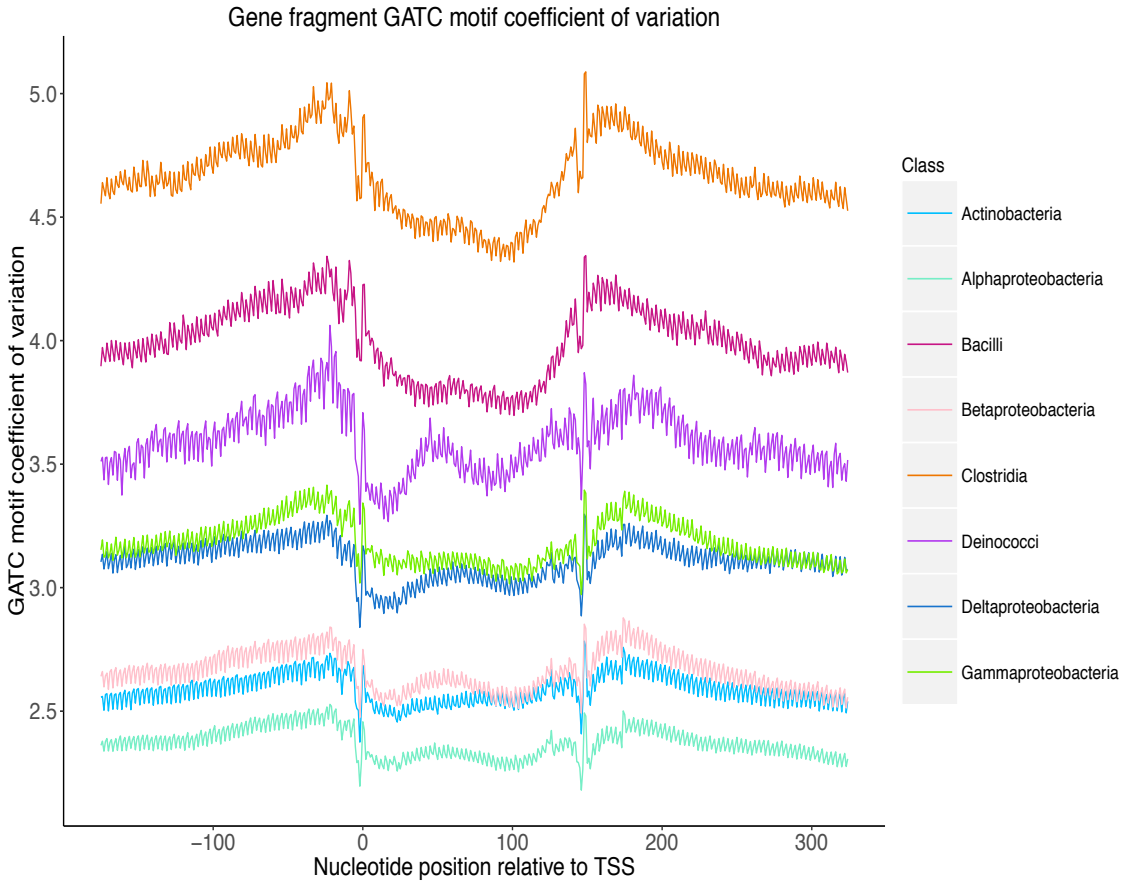


Figure 2.5: Coefficient of variation for gene fragment GATC frequency. The transcription start site (TSS) is at point 0. There is a lower CV within the downstream region associated with an increase in GATC tetranucleotide motif frequency (Figure 2.4). GATC motifs are far less prevalent than CpG motifs (Figure 2.6), which also increase in this region.

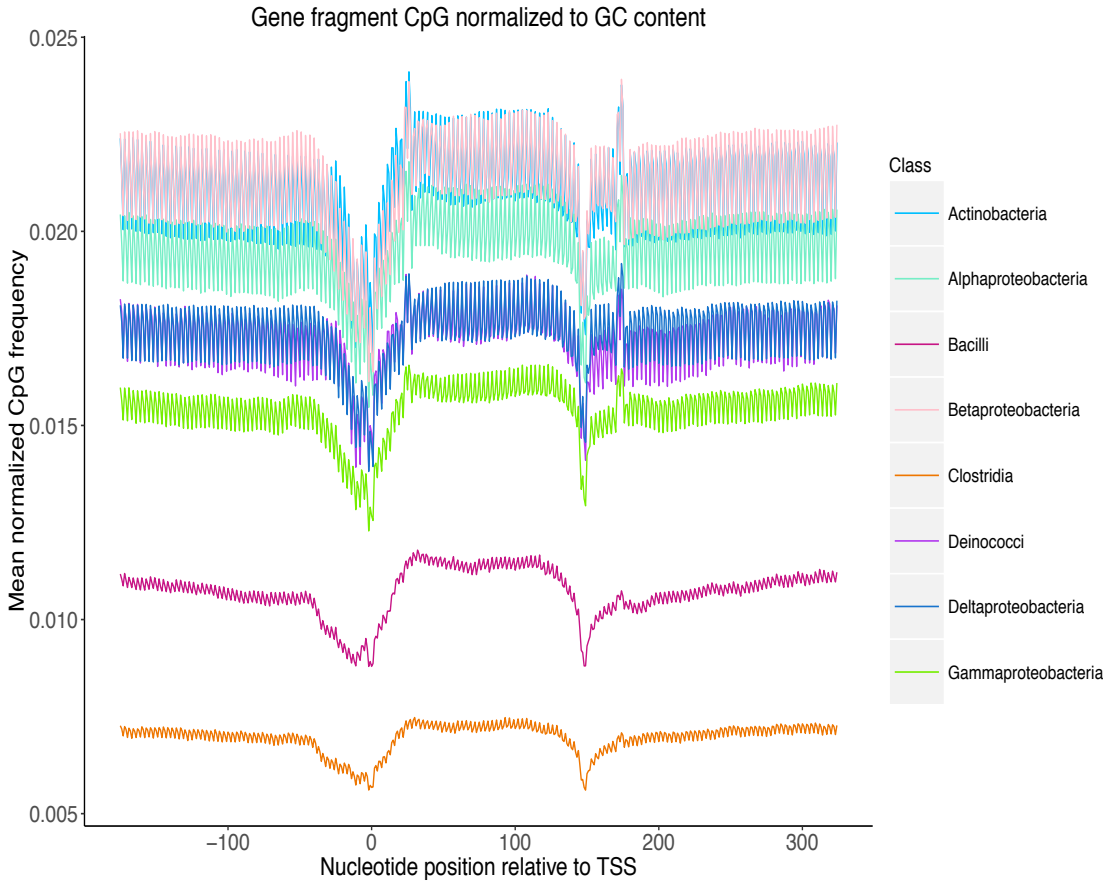


Figure 2.6: Overall mean gene fragment CpG normalized to GC content. CpG motif frequency is lower within the core promoter region (around -35 bp until the TSS at position 0). However, there is an increase in frequency downstream of the TSS. This increase and plateau in CpG motif frequency coincides with increases in GC content (Figure 2.8) and decreases in curvature (Figure 2.9).

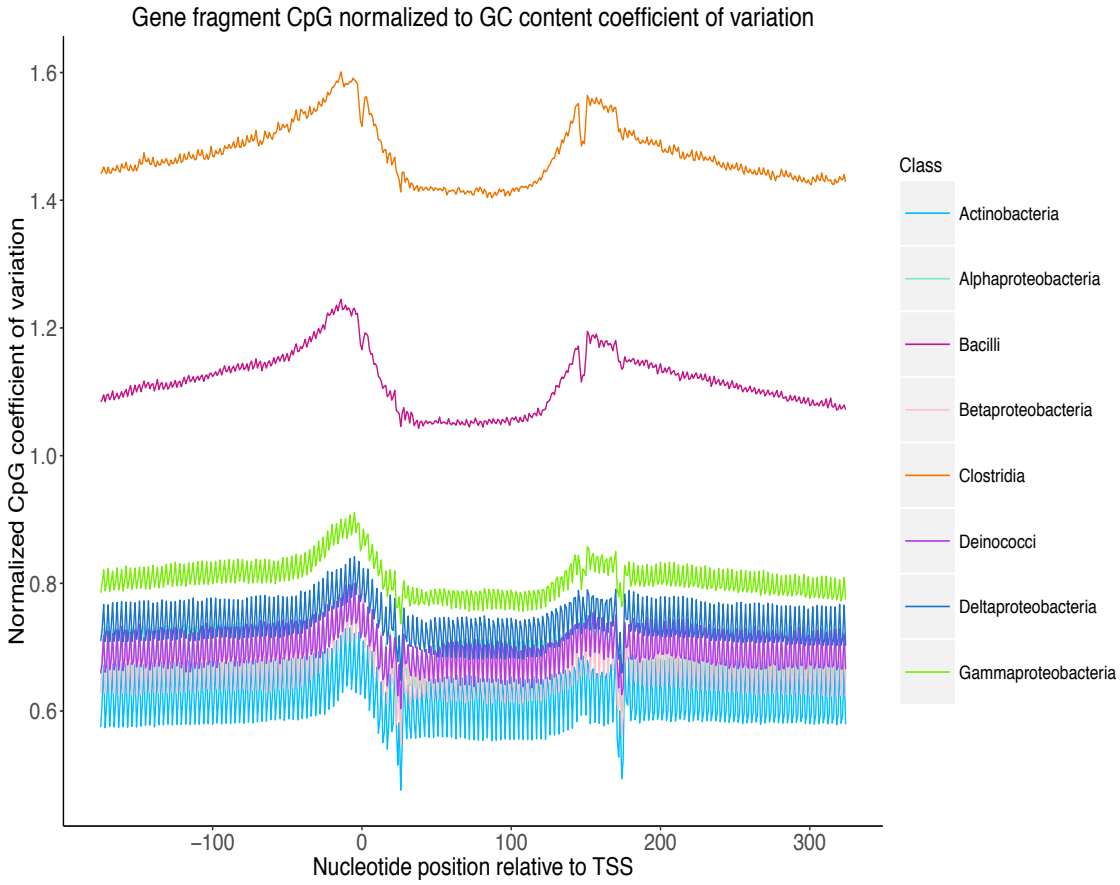


Figure 2.7: Coefficient of variation for CpG normalized to GC content. The transcription start site (TSS) is at point 0. Nucleotide positions within the plateau of higher normalized CpG motif frequency (Figure 2.6) and valley of decreased curvature (Figure 2.9) spanning from ~ +35 bp to +150 bp have a lower CV for CpG motif frequency, indicating that higher CpG frequency is a consistent and conserved pattern at this site, and likely influences curvature.

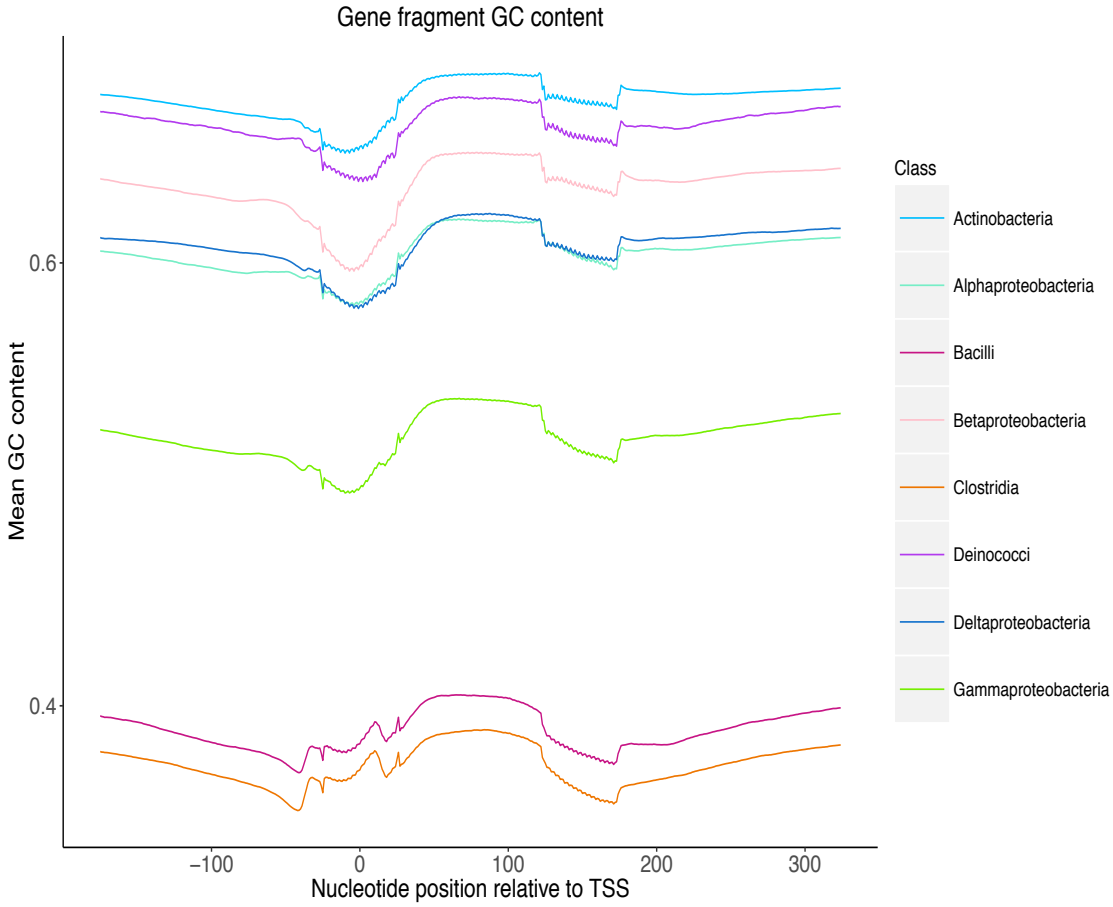


Figure 2.8: Overall gene fragment mean GC content. The transcription start site (TSS) is at point 0. The downstream region ~ +35 to +150 has higher GC content, and this accounts for lower curvature in this region (Figure 2.9) due to a decrease in the probability of A-tracts and T-tracts.

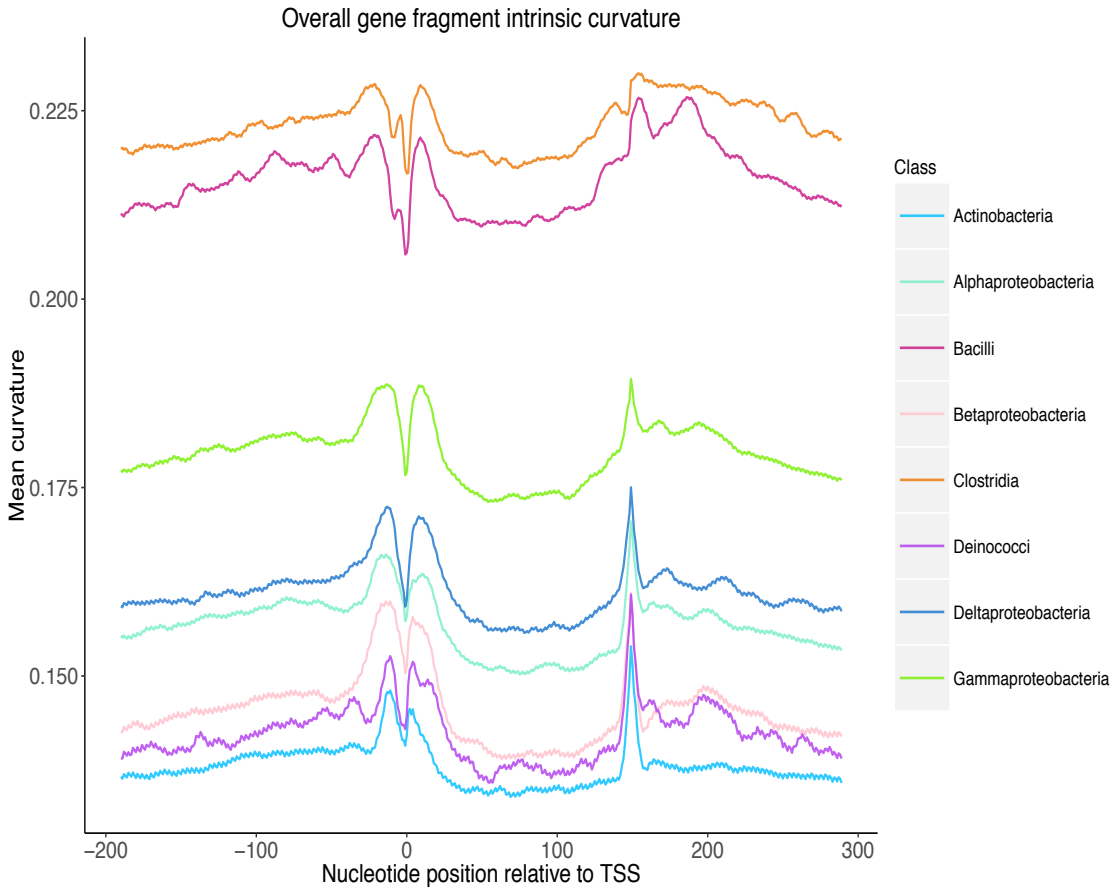


Figure 2.9: Overall mean gene fragment intrinsic DNA curvature. The transcription start site (TSS) is at point 0. Curvature was calculated for 2,168,800 gene fragments using the nearest-neighbor dinucleotide wedge model (Bolshoy et al. 1991). Higher curvature is conserved within the core promoter region starting ~ -35 bp upstream. Curvature decreases as the TSS is approached, but increases again immediately downstream of the TSS. There is a noticeable decrease in curvature downstream of the core promoter from ~ +35 to +150 bp across these classes.

2.4 Discussion

Microbes are able to regulate transcriptional activity through the direct interaction of transcription factors with specific nucleotide sequences, yet a host of epigenetic phenomena can further influence enzymatic DNA binding and efficiency within a genome. Epigenetic regulation in microbes is not governed by a single biological phenomenon, but is a dynamic system involving the interplay of multiple processes that act upon and influence each other. DNA methylation and DNA curvature act as important transcriptional regulators within genes, and this study has identified conserved sites associated with these phenomena on a large scale.

N6-methyladenine, 5-methylcytosine, and N4-methylcytosine-specific DNA methyltransferase orthologs were present within all target classes. A higher proportion of N6-adenine specific MTase orthologs were aligned within the Gammaproteobacteria, which could be attributed to the prevalence of known methyltransferases isolated from members of this class, particularly *E. coli*. Overall alignment results suggest that the Actinobacteria contain orthologs for several m5-cytosine specific MTases (Figure 2.1, Supplementary Table 4). A preference for this type of methylation could be due to the lower GC content of this Class.

There appear to be conserved global trends in methylation motif spatial highs and lows across the eight bacterial classes analyzed in this study. Low frequencies of GATC motifs were observed, but this scarcity is common in microbial genomes (Wojciechowski et al. 2012). A higher frequency of expected GATC sites is conserved

within and immediately downstream of the core promoter region across these classes (Figure 2.2), and could signify that a higher number of genes utilize GATC methylation or other methylation motifs with GA/AT/TC dinucleotides at promoter regions to regulate transcription. It should be noted, however, that these sites were generally scarce. Along with conserved trends of increased GATC motif and $GATC_{expected}$ content within core promoter regions, there is also an increase in curvature within the core promoter region (Figure 2.9) that is consistent with previous studies (Asayama and Ohyama 2005).

CpG site frequency normalized to GC content appears to increase and plateau from +35 to +150 bp downstream from the TSS, with low frequencies within the core promoter region (Figure 2.6). This contrasts with heightened GATC frequency within core promoter regions, and suggests that methylation at CpG sites or CpG-containing motifs could play a greater role within downstream gene coding sequences. A sharp decrease in CpG site frequency also coincides with an increase in curvature (Figure 2.9) and a relative increase in GATC frequency (Figure 2.4) around 150+ bp downstream, suggesting a potential regulatory site.

The CpG_{norm} metric provides insight into CpG representation amongst these Classes due to its dependency on both CpG density and GC content. Clostridia are shown to exhibit CpG underrepresentation on the basis of exhibiting the lowest GC contents (Figure 2.8) and CpG site frequencies (Figure 2.6) of these eight Classes. These lower mean CpG_{norm} scores indicate reduced CpG density if GC contents are

low as well. Clostridia were also shown to have the highest CV for CpG_{norm}, as the presence of a CpG site would be less probable and result in a higher CpG_{norm} score within a low GC genome. Genome-wide CpG underrepresentation is exhibited in *Clostridium perfringens* (Wojciechowski et al. 2012), and the results of this study support CpG depletion within the Clostridia. Another explanation for CpG underrepresentation in Clostridia is due to the possibility that cytosine methyltransferases with 5'-CG-3' or 5'-CCGG-3' target sequences can shape dinucleotide frequencies. Significant BLAST alignments to the MTases M.SssI and M.HhaI were observed within the Clostridia, which respectively target 5'-CG-3' and 5'-CCGG-3' (Supplementary Table 6). The lack of CpG sites is also related to the spontaneous deamination of m5C to thymine. Since m5C deamination forms thymine and not uracil (an illegitimate DNA base), it is not corrected during DNA repair (Wojciechowski et al. 2012) and can lead to potentially deleterious transition mutations (Chahwan et al. 2010). Thus, reduced frequencies of CpG sites could lessen the chance of mutations caused by m5C deamination.

GATC motif frequency increased around 150+ bp downstream from the TSS in the Actinobacteria, Alphaproteobacteria, Betaproteobacteria, and Deltaproteobacteria. These increases coincide with increased curvature at this site (Figure 2.9). The higher probability of GA/AT/TC dinucleotides could be due to A-tracts, T-tracts, AA/TT stacks, or GA/TC stacks, all of which highly influence intrinsic curvature (Asayama and Ohyama 2005).

Curvature increases are not nearly as profound for the Bacilli and Clostridia. This could be due to the generally higher intrinsic curvature of genes within these classes, as well as less drastic increases in curvature at 150+ as opposed to the other classes (Figure 2.9). The Bacilli and Clostridia appear to have a higher frequency of GA/AT/TC dinucleotides within the core promoter region (Figure 2.2), but a lower frequency of GATC tetranucleotides when compared with the other classes (Figure 2.4). Based on TBLASTN alignments, other N6-adenine-specific motif sites aside from GATC could be present within the promoter regions of these two classes (Supplementary Table 5). High-quality alignments for M.EcoRV and M.MunI were present in the Clostridia (target sites GATATC and CAATTG, respectively), as well as alignments for M.FokI (GGATG). The presence of these target motifs would increase $GATC_{expected}$ at the site due to GA/AT/TC dincueltodies contained within these motifs, and high wedge angle dinucleotides (GA, AA, GG) and AA/TT dinucleotide stacks could be responsible for higher curvature.

A curvature peak at the core promoter region could be due to the presence and periodicity of A-tracts, T-tracts, or AA/TT dinucleotide stacks that cause the formation of a more highly curved region with a right-handed superhelical conformation (Asayama and Ohyama 2005). Formation of a right-handed curvature conformation within the promoter could aid in RNAP binding; for instance, *E. coli* RNAP favors DNA with a right-handed superhelical conformation when binding (Hirota and Ohyama 1995). Based on these results, there is likely a conserved increase in

methylation-associated motif frequency and intrinsic curvature at the core promoter region that would aid in RNAP binding.

Several additions can be made to expand the scope of this study. The variety and complexity of bacterial MTase target sequences expands far beyond CpG and GATC, and furthering this global motif identification approach for shared, high-scoring MTase ortholog target sites such as 5'-CCWGG-3' (Dcm and M.MvaI) would give a better picture of conserved methylation motif trends. Stressors such as temperature can affect DNA curvature (Kozobay-avraham et al. 2008), and tying global motif frequency and curvature trends to these metadata could allow for a better understanding of microbial epigenetic regulation based on stressors.

DNA methylation has been considered to play a role in microbial gene regulation for over 40 years (Holliday and Pugh 1975), yet our understanding of this phenomenon and its effects on other epigenetic processes such as curvature is undergoing a paradigm shift due to the advent of next-generation sequencing technologies and big data analysis techniques. While this study provides a class-level overview of intrinsic curvature and CpG and GATC motif frequencies for a portion of known bacterial genomes, it does further the connections between epigenetic phenomena supporting the notion of dynamic genomes that are regulated by multiple factors outside of the nucleotide sequence itself. The further development of holistic – omics methodologies that account for these phenomena will vastly improve our understanding of these complex epigenetic systems.

Chapter 3

CONCLUSIONS

While not generally regarded as a widespread mechanism for gene regulation in bacteria, this thesis shows the widespread potential for epigenetic regulation in microbes. Epigenetic regulation incorporates a variety of biological processes that alter both the chemical composition and structure of DNA. This study has provided evidence of microbial regulatory potential by dynamic CpG methylation at a community level, as well as identified conserved curvature and DNA methylation motif trends within known bacterial genomes. While epigenetic regulators such as DNA methylation and curvature have been studied in bacteria since the late 1970s, laboratory techniques for analyzing these phenomena are being rapidly supplemented by sequencing technologies capable of detecting base alterations (Blow et al. 2016; Burgess 2013). These phenomena may help explain microbial activities in complex, slow-growing environments such as estuary sediments.

This study utilized a methylation-sensitive assay based on Illumina Hi-Seq sequencing to identify dynamic shifts in m5C at 5'-CCGG-3' sites. While this method identifies base modifications at a single target site and was developed for single genome analysis, we show it can be adapted to the uneven read depths and complex community compositions associated with metagenomic sequence data. The ability to

utilize this assay with comparatively low concentrations of DNA makes it well suited for environmental applications. These features have enabled us to provide a first glimpse into dynamic microbial methylation at a community scale. While this approach could be modified for different target sites and paired-end sequencing, the proprietary nature of the DNA methylation analysis pipeline may possibly inhibit the future development of this assay for metagenomics. Since current single-molecule sequencing methods provide researchers with means of accurately determining methylomes for pure-culture organisms, and these methods can also be adapted for metagenomics.

Pacific Biosciences single-molecule, real-time (SMRT) sequencing has been used to map the methylomes of several bacterial species (Murray et al. 2012; Fang et al. 2012), yet these studies were performed on pure culture organisms raised in a laboratory setting. Utilizing SMRT sequencing for methylome analysis requires a high concentration of clean, high molecular weight DNA, which can be difficult to obtain in sediment and soil environments. Another caveat of this method lies in the determination of base modifications by measuring nucleotide addition rates by DNA polymerase along a single strand of DNA. DNA polymerase is inhibited by several contaminants common to soils and sediments such as humic acid, and the effects of these inhibitors could be detrimental to the accuracy of nucleotide incorporation kinetic data. Aside from SMRT sequencing, Oxford Nanopore MinION is a recently introduced single-molecule DNA sequencing technology potentially offering read

lengths of tens of kilobases (Laver et al. 2015). Proof-of-principle experiments have demonstrated the ability of Nanopore sequencing to detect 5-methylcytosine within DNA strands (Schreiber et al. 2013; Laszlo et al. 2013), yet its performance in bacterial methylome and metagenomic applications has not been thoroughly tested.

The current ability to reconstruct complete or nearly complete genomes from environmental sequence data is a testament to the enormous strides made in both sequencing and computational technologies over this past decade, and now we are able to develop methods that look deeper into the regulatory mechanisms governing microbial activity in the environment. Developing a better understanding of community epigenetic dynamics influenced by factors such as seasonality can further explain biogeochemical fluxes vital to nutrient cycling and atmospheric composition. However, the limited and generally unfavorable options for analyzing these phenomena in environmental sequence data present a current hurdle for researchers. Future developments of Illumina or Nanopore-based methods for environmental DNA methylation analysis could allow for researchers to not only reconstruct the genomes of uncultured microbes, but to construct their methylomes as well, and ideally tie these with metatranscriptomes and metaproteomes to show true regulation of function. The identification of conserved sequence structural features could also be used to better determine epigenetic regulation shared across bacterial, archaeal and eukaryotic domains.

This study has also taken a large-scale approach to the determination of conserved structural patterns related to DNA curvature and methylation motif frequency and their relationship to anticipated transcriptional start sites. Similar approaches have been used for gene calling software such as MetaGene (Noguchi et al. 2006), which utilizes di-codon frequencies based on GC content. The results of this study show that trends are seen in curvature, motif frequency, and GC content across bacterial classes, and the identification of these trends could be utilized in more robust software for gene calling in environmental sequence data, potentially including an epigenetic basis.

Sequencing technologies revolutionized life science as we know it, and their development has rapidly increased our knowledge of what phenomena lay beyond the genome itself. Emerging technologies that characterize epigenetic shifts in microbes are shedding new light on the genetic mechanisms that influence microbial systems such as DNA methylation. Due to the widespread presence of DNA methylation in microbes and its likely role in gene regulation, gaining a better understanding of this phenomenon and its dynamic interplay with curvature and transcriptional regulation will benefit numerous fields such as healthcare and environmental remediation.

REFERENCES

Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L Nielsen, Gene W Tyson, and Per H Nielsen. 2013. "Genome Sequences of Rare, Uncultured Bacteria Obtained by Differential Coverage Binning of Multiple Metagenomes." *Nature Biotechnology* 31 (6): 533–38. doi:10.1038/nbt.2579.

Altschul, Stephen F, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. 1997. "Gapped BLAST and PSI-BLAST : A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215: 403–10.

Angermeyer, Angus, Sarah C Crosby, and Julie A Huber. 2016. "Decoupled Distance – Decay Patterns between dsrA and 16S rRNA Genes among Salt Marsh Sulfate-Reducing Bacteria" 18: 75–86. doi:10.1111/1462-2920.12821.

Asayama, Munehiko, and Takashi Ohyama. 2005. "Curved DNA and Prokaryotic Promoters: A Mechanism for Activation of Transcription." In *DNA Conformation and Transcription*, 37–51.

Baker, Brett J, Cassandre Sara Lazar, Andreas P Teske, and Gregory J Dick. 2015. "Genomic Resolution of Linkages in Carbon , Nitrogen , and Sulfur Cycling among Widespread Estuary Sediment Bacteria." *Microbiome* 3 (14). ??? 1–12. doi:10.1186/s40168-015-0077-6.

Baker, Brett J, Jimmy H Saw, Anders E Lind, Cassandre Sara Lazar, Kai-uwe Hinrichs, Andreas P Teske, and Thijs J G Ettema. 2016. "Genomic Inference of the Metabolism of Cosmopolitan Subsurface Archaea , Hadesarchaea." *Nature Microbiology*, no. February. Nature Publishing Group: 1–7. doi:10.1038/nmicrobiol.2016.2.

Bauer, James E., Wei-Jun Cai, Peter A. Raymond, Thomas S. Bianchi, Charles S. Hopkinson, and Pierre A.G. Regnier. 2013. "The Changing Carbon Cycle of the Coastal Ocean." *Nature* 504 (61): 61–70.

Benson, Dennis A, Ilene Karsch-mizrachi, David J Lipman, James Ostell, and Eric W Sayers. 2009. "GenBank." *Nucleic Acids Research* 37 (October 2008): 26–31. doi:10.1093/nar/gkn723.

Bhattacharya, Debaditya, Anand Nagpure, Rajinder K Gupta, and Debaditya Bhattacharya. 2007. "Bacterial Chitinases : Properties and Potential." *Critical Reviews in Biotechnology* 27: 21–28. doi:10.1080/07388550601168223.

Biddle, Jennifer F, Zena Cardman, Howard Mendlovitz, Daniel B Albert, Karen G Lloyd, Antje Boetius, and Andreas Teske. 2012. "Anaerobic Oxidation of Methane at Different Temperature Regimes in Guaymas Basin Hydrothermal Sediments." *The ISME Journal* 6 (5). Nature Publishing Group: 1018–31. doi:10.1038/ismej.2011.164.

Bird, Adrian. 2002. "DNA Methylation Patterns and Epigenetic Memory." *Genes & Development* 16 (1): 6–21. doi:10.1101/gad.947102.

Blazewicz, Steven J, Romain L Barnard, Rebecca a Daly, and Mary K Firestone. 2013. "Evaluating rRNA as an Indicator of Microbial Activity in Environmental Communities: Limitations and Uses." *The ISME Journal* 7 (11). Nature Publishing Group: 2061–68. doi:10.1038/ismej.2013.102.

Blow, Matthew J, Tyson A Clark, Chris G Daum, Adam M Deutschbauer, Alexey Fomenkov, Roxanne Fries, Jeff Froula, and Dongwan D Kang. 2016. "The Epigenomic Landscape of Prokaryotes." *PLOS Genetics*, 1–28. doi:10.1371/journal.pgen.1005854.

Bolshoy, A, P Mcnamara, R E Harrington, and E N Trifonov. 1991. "Curved DNA without A-A : Experimental Estimation of All 16 DNA Wedge Angles." *Proceedings of the National Academy of Sciences* 88 (6): 2312–16.

Braaten, Bruce A, Xiangwu Nou, Linda S Kaltenbach, and David A Low. 1994. "Methylation Patterns in Pap Regulatory DNA Control Pyelonephritis-Associated Pili Phase Variation in E . Coli." *Cell* 76: 577–88.

Bracco, Laurent, Denise Kotlarz, Annie Kolb, Stephan Diekmann, and Henri Buc. 1989. "Synthetic Curved DNA Sequences Can Act as Transcriptional Activators in Escherichia Coli." *The EMBO Journal* 8 (13): 4289–96.

Brady, Arthur, and Steven Salzberg. 2011. "PhymmBL Expanded: Confidence Scores, Custom Databases, Parallelization and More." *Nature Methods* 8 (5): 2011–13. doi:10.1038/nmeth0511-367.PhymmBL.

Brady, Arthur, and Steven L Salzberg. 2009. "Phymm and PhymmBL: Metagenomic Phylogenetic Classification with Interpolated Markov Models." *Nature Methods* 6 (9). Nature Publishing Group: 673–76. doi:10.1038/nmeth.1358.

Brunet, Yannick R, Christophe S Bernard, Marthe Gavioli, Roland Lloubès, and Eric Cascales. 2011. "An Epigenetic Switch Involving Overlapping Fur and DNA Methylation Optimizes Expression of a Type VI Secretion Gene Cluster." *PLoS Genetics* 7 (7): e1002205. doi:10.1371/journal.pgen.1002205.

Burgess, Darren J. 2013. "Epigenetics: Bacterial DNA Methylation Gets SMRT." *Nature Reviews. Genetics* 14 (1). Nature Publishing Group: 4. doi:10.1038/nrg3389.

Butkus V, L Petrauskiene, Z Maneliene, S Klimasauskas, V Laucys, A Janulaitis. 1987. "Cleavage of Methylated CCCGGG Sequences Containing Either N4-Methylcytosine or 5-Methylcytosine with MspI, HpaII, SmaI, XmaI and Cfr9I Restriction Endonucleases." *Nucleic Acids Research* 15 (17): 7091–7102.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason

Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. “BLAST + : Architecture and Applications” 9: 1–9. doi:10.1186/1471-2105-10-421.
Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, et al. 2010. “Correspondence QIIME Allows Analysis of High- Throughput Community Sequencing Data Intensity Normalization Improves Color Calling in SOLiD Sequencing.” *Nature Publishing Group* 7 (5). Nature Publishing Group: 335–36. doi:10.1038/nmeth0510-335.
Casadesús, Josep, and David Low. 2006. “Epigenetic Gene Regulation in the Bacterial World.” *Microbiology and Molecular Biology Reviews : MMBR* 70 (3): 830–56. doi:10.1128/MMBR.00016-06.
Chahwan, Richard, Sandeep N Wontakal, and Sergio Roa. 2010. “Crosstalk between Genetic and Epigenetic Information through Cytosine Deamination.” *Trends in Genetics* 26 (10). Elsevier Ltd: 443–48. doi:10.1016/j.tig.2010.07.005.
Chao, Anne. 1984. “Nonparametric Estimation of the Number of Classes in a Population.” *Scandinavian Journal of Statistics* 11 (4): 265–70.
Cheng, Bingran. 2013. “Variation of Archaeal Communities in Sediments of Coastal Delaware.”
Cheng, Xiaodong. 1995. “Structure and Function of Dna Methyltransferases.” *Annual Review of Biophysics and Biomolecular Structure* 24: 293–318.
Chmura, Gail L. 2013. “What Do We Need to Assess the Sustainability of the Tidal Salt Marsh Carbon Sink?” *Ocean & Coastal Management* 83 (October). Elsevier Ltd: 25–31. doi:10.1016/j.ocecoaman.2011.09.006.
Clouet-d’Orval, Béatrice, Christine Gaspin, and Annie Mougin. 2005. “Two Different Mechanisms for tRNA Ribose Methylation in Archaea: A Short Survey.” *Biochimie* 87 (9-10): 889–95. doi:10.1016/j.biochi.2005.02.004.
Cole, James R., Qiong Wang, Jordan a. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. 2014. “Ribosomal Database Project: Data and Tools for High Throughput rRNA Analysis.” *Nucleic Acids Research* 42 (D1): 633–42. doi:10.1093/nar/gkt1244.
Collier, Justine. 2009. “Epigenetic Regulation of the Bacterial Cell Cycle.” *Current Opinion in Microbiology* 12 (6): 722–29. doi:10.1016/j.mib.2009.08.005.
Cottrell, Matthew T, Jessica A Moore, and David L Kirchman. 1999. “Chitinases from Uncultured Marine Microorganisms.” *Applied and Environmental Microbiology* 65 (6): 2553–57.
Cottrell, Matthew T, Daniel N Wood, Liying Yu, and David L Kirchman. 2000. “Selected Chitinase Genes in Cultured and Uncultured Marine Bacteria in the - and -Subclasses of the Proteobacteria.” *Applied and Environmental Microbiology* 66 (3): 1195–1201.
Coulombe, Benoit, and Zachary F Burton. 1999. “DNA Bending and Wrapping around RNA Polymerase : A ‘ Revolutionary ’ Model Describing Transcriptional

Mechanisms." *Microbiology and Molecular Biology Reviews : MMBR* 63 (2): 457–78.
Darling, Aaron E, Guillaume Jospin, Eric Lowe, Frederick a Matsen, Holly M Bik, and Jonathan a Eisen. 2014. "PhyloSift: Phylogenetic Analysis of Genomes and Metagenomes." *PeerJ* 2: e243. doi:10.7717/peerj.243.
DeAngelis, Kristen M., Whendee L. Silver, Andrew W. Thompson, and Mary K. Firestone. 2010. "Microbial Communities Acclimate to Recurring Changes in Soil Redox Potential Status." *Environmental Microbiology* 12: 3137–49. doi:10.1111/j.1462-2920.2010.02286.x.
Diekmann, Stephan. 1987. "DNA Methylation Can Enhance or Induce DNA Curvature." *The EMBO Journal* 6 (13): 4213–17.
Durkin, Colleen A, Thomas Mock, and E Virginia Armbrust. 2009. "Chitin in Diatoms and Its Association with the Cell Wall" 8 (7): 1038–50. doi:10.1128/EC.00079-09.
Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLoS Computational Biology* 7 (10). doi:10.1371/journal.pcbi.1002195.
Edgar, Robert C. 2013. "UPARSE : Highly Accurate OTU Sequences from Microbial Amplicon Reads" 10 (10). doi:10.1038/nmeth.2604.
Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster than BLAST." *Bioinformatics* 26 (19): 2460–61. doi:10.1093/bioinformatics/btq461.
Edgar, Robert C., Brian J. Haas, Jose C. Clemente, Christopher Quince, and Rob Knight. 2011. "UCHIME Improves Sensitivity and Speed of Chimera Detection." *Bioinformatics* 27 (16): 2194–2200. doi:10.1093/bioinformatics/btr381.
Eisen, Jonathan a, Karen E Nelson, Ian T Paulsen, John F Heidelberg, Martin Wu, Robert J Dodson, Robert Deboy, et al. 2002. "The Complete Genome Sequence of Chlorobium Tepidum TLS, a Photosynthetic, Anaerobic, Green-Sulfur Bacterium." *Proceedings of the National Academy of Sciences of the United States of America* 99 (14): 9509–14. doi:10.1073/pnas.132181499.
Engel, James Douglas, and Peter H. von Hippel. 1978. "Effects of Methylation on the Stability of Nucleic Acid Conformations." *The Journal of Biological Chemistry* 253 (3): 927–34.
Fang, Gang, Diana Munera, David I Friedman, Anjali Mandlik, Michael C Chao, Onureena Banerjee, Zhixing Feng, et al. 2012. "Genome-Wide Mapping of Methylated Adenine Residues in Pathogenic Escherichia Coli Using Single-Molecule Real-Time Sequencing." *Nature Biotechnology* 30 (12): 1232–39. doi:10.1038/nbt.2432.
Fuhrman, Jed a. 2009. "Microbial Community Structure and Its Functional Implications." *Nature* 459 (7244): 193–99. doi:10.1038/nature08058.
Gabrielian, Andrei, and Sfindor Pongor. 1996. "Correlation of Intrinsic DNA Curvature with DNA Property Periodicity." *Federation of European Biochemical Societies* 393: 65–68.
Gaidos, Eric, Antje Rusch, and Melissa Ilardo. 2011. "Ribosomal Tag Pyrosequencing

of DNA and RNA from Benthic Coral Reef Microbiota: Community Spatial Structure, Rare Members and Nitrogen-Cycling Guilds." *Environmental Microbiology* 13: 1138–52. doi:10.1111/j.1462-2920.2010.02392.x.

Gonzalez, Diego, Jennifer B Kozdon, Harley H McAdams, Lucy Shapiro, and Justine Collier. 2014. "The Functions of DNA Methylation by CcrM in Caulobacter Crescentus: A Global Approach." *Nucleic Acids Research* 42 (6): 3720–35. doi:10.1093/nar/gkt1352.

Gooday, G W. 1990. "The Ecology of Chitin Degradation." *Advances in Microbial Ecology* 11: 387–430.

Gooday, Graham W., James I. Prosser, Kevin Hillman, and Martin G Cross. 1991. "Mineralization of Chitin in an Estuarine Sediment : The Importance of the Chitosan Pathway." *Biochemical Systematics and Ecology* 19 (5): 395–400.

Goodsell, David S, and Richard E Dickerson. 1994. "Bending and Curvature Calculations in B-DNA." *Nucleic Acids Research* 22 (24): 5497–5503.

Hardison, A K, I C Anderson, E A Canuel, C R Tobias, and B Veuger. 2011. "Carbon and Nitrogen Dynamics in Shallow Photic Systems : Interactions between Macroalgae , Microalgae , and Bacteria" 56 (4): 1489–1503. doi:10.4319/lo.2011.56.4.1489.

Hartigan, J.A., and P.M. Hartigan. 1985. "The Dip Test of Unimodality." *Annals of Statistics* 13 (1): 70–84.

Heinrichs, Arianne. 2014. "m6A RNA Methylation in Mammals." *Nature Structural & Molecular Biology* 21 (2). Nature Publishing Group: 117–117. doi:10.1038/nsmb.2773.

Henikoff, Steven, and Jorja G Henikoff. 1992. "Amino Acid Substitution Matrices from Protein Blocks." *Proceedings of the National Academy of Sciences of the United States of America* 89 (November): 10915–19.

Hoehler, Tori M, and Bo Barker Jørgensen. 2013. "Microbial Life under Extreme Energy Limitation." *Nature Reviews. Microbiology* 11 (2). Nature Publishing Group: 83–94. doi:10.1038/nrmicro2939.

Holliday, R, and J E Pugh. 1975. "DNA Modification Mechanisms and Gene Activity during Development." *Science* 187: 226–32.

Hua, Zheng-Shuang, Yu-Jiao Han, Lin-Xing Chen, Jun Liu, Min Hu, Sheng-Jin Li, Jia-Liang Kuang, Patrick Sg Chain, Li-Nan Huang, and Wen-Sheng Shu. 2014. "Ecological Roles of Dominant and Rare Prokaryotes in Acid Mine Drainage Revealed by Metagenomics and Metatranscriptomics." *The ISME Journal*, November. Nature Publishing Group, 1–15. doi:10.1038/ismej.2014.212.

Hunt, Dana E., Yajuan Lin, Matthew J. Church, David M. Karl, Susannah G. Tringe, Lisa K. Izzo, and Zackary I. Johnson. 2013. "Relationship between Abundance and Specific Activity of Bacterioplankton in Open Ocean Surface Waters." *Applied and Environmental Microbiology* 79 (1): 177–84. doi:10.1128/AEM.02155-12.

Jauregui, Ruy, Cei Abreu-Goodger, Gabriel Moreno-Hagelsieb, Julio Collado-Vides,

and Enrique Merino. 2003. “Conservation of DNA Curvature Signals in Regulatory Regions of Prokaryotic Genes.” *Nucleic Acids Research* 31 (23): 6770–77. doi:10.1093/nar/gkg882.

Jeltsch, Albert, Renata Z Jurkowska, Tomasz P Jurkowski, Kirsten Liebert, Philipp Rathert, and Martina Schlickenrieder. 2007. “Application of DNA Methyltransferases in Targeted DNA Methylation.” *Applied Microbiology and Biotechnology* 75 (6): 1233–40. doi:10.1007/s00253-007-0966-0.

Jonckheere, Aimable Robert. 1954. “A Distribution-Free K-Sample Test against Ordered Alternatives.” *Biometrika* 41: 133–45.

Kallmeyer, Jens, Robert Pockalny, Rishi Ram Adhikari, David C Smith, and Steven D’Hondt. 2012. “Global Distribution of Microbial Abundance and Biomass in Subseafloor Sediment.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (40): 16213–16. doi:10.1073/pnas.1203849109.

Kanehisa, Minoru, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. “KEGG as a Reference Resource for Gene and Protein Annotation.” *Nucleic Acids Research* 44 (October 2015): 457–62. doi:10.1093/nar/gkv1070.

Kanhere, Aditi, and Manju Bansal. 2005. “Structural Properties of Promoters : Similarities and Differences between Prokaryotes and Eukaryotes.” *Nucleic Acids Research* 33 (10): 3165–75. doi:10.1093/nar/gki627.

Katayama, Seiichi, Osamu Matsushita, Chang-min Jung, Junzaburo Minami, and Akinobu Okabe. 1999. “Promoter Upstream Bent DNA Activates the Transcription of the Clostridium Perfringens Phospholipase C Gene in a Low Temperature-Dependent Manner.” *The EMBO Journal* 18 (12): 3442–50.

Kawai, Mikihiko, Ikuo Uchiyama, Hideto Takami, and Fumio Inagaki. 2015. “Low Frequency of Endospore-Specific Genes in Subseafloor Sedimentary Metagenomes.” *Environmental Microbiology Reports* 7 (2): 341–50. doi:10.1111/1758-2229.12254.

Kemp, Andrew C., Christopher K. Sommerfield, Christopher H. Vane, Benjamin P. Horton, Simon Chenery, Shimon Anisfeld, and Daria Nikitina. 2012. “Use of Lead Isotopes for Developing Chronologies in Recent Salt-Marsh Sediments.” *Quaternary Geochronology* 12: 40–49. doi:10.1016/j.quageo.2012.05.004.

Klimasauskas, Saulius, Dana Steponavicene, Zita Maneliene, Maryte Petrusyte, and Viktoras Butkus. 1990. “M.SmaI Is an N4-Methylcytosine Specific DNA-Methylase.” *Nucleic Acids Research* 18 (22): 6607–9.

Klimasauskas, Saulius, Albertas Timinskas, Saulius Menkevicius, Danguole Butkiene, Viktoras Butkus, and Arvydas Janulaitis. 1989. “Sequence Motifs Characteristic of DNA[cytosine-N4]methyltransferases: Similarity to Adenine and Cytosine-C5 DNA-Methylases.” *Nucleic Acids Research* 17 (23): 9823–32.

Koretsky, Carla M., Philippe Van Cappellen, Thomas J. DiChristina, Joel E. Kostka, Kristi L. Lowe, Charles M. Moore, Alakendra N. Roychoudhury, and Eric Viollier. 2005. “Salt Marsh Pore Water Geochemistry Does Not Correlate with

Microbial Community Structure." *Estuarine, Coastal and Shelf Science* 62 (1-2): 233–51. doi:10.1016/j.ecss.2004.09.001.
Kozobay-avraham, L, S Hosid, Z Volkovich, and A Bolshoy. 2008. "Prokaryote Clustering Based on DNA Curvature Distributions." *Discrete Applied Mathematics* 157 (10). Elsevier B.V.: 2378–87. doi:10.1016/j.dam.2008.06.049.
Kravatskaya, G I, Y V Kravatsky, V R Chechetkin, and V G Tumanyan. 2011. "Coexistence of Different Base Periodicities in Prokaryotic Genomes as Related to DNA Curvature , Supercoiling , and Transcription." *Genomics* 98 (3). Elsevier Inc.: 223–31. doi:10.1016/j.ygeno.2011.06.006.
Kumar, Ritesh, and Desirazu N Rao. 2012. "Role of DNA Methyltransferases in Epigenetic Regulation in Bacteria." In *Epigenetics: Development and Disease*, 81–102. doi:10.1007/978-94-007-4525-4.
Lacks, Sanford A. 1980. "Purification and Properties of the Complementary Endonucleases DpnI and DpnII." *Methods in Enzymology* 65 (1977): 138–46.
Liu-Johnson, Huei-Nin, Marc R. Gartenberg, and Donald M. Crothers. 1986. "The DNA Binding Domain and Bending Angle of E. Coli CAP Protein." *Cell* 47: 995–1005.
Løbner-olesen, Anders, Ole Skovgaard, and Martin G Marinus. 2005. "Dam Methylation : Coordinating Cellular Processes." *Current Opinion in Microbiology* 8: 154–60. doi:10.1016/j.mib.2005.02.009.
Lomstein, Bente Aa, Alice T Langerhuus, Steven D'Hondt, Bo B Jørgensen, and Arthur J Spivack. 2012. "Endospore Abundance, Microbial Growth and Necromass Turnover in Deep Sub-Seafloor Sediment." *Nature* 484 (7392). Nature Publishing Group: 101–4. doi:10.1038/nature10905.
Low, David A, and Josep Casadesús. 2008. "Clocks and Switches: Bacterial Gene Regulation by DNA Adenine Methylation." *Current Opinion in Microbiology* 11 (2): 106–12. doi:10.1016/j.mib.2008.02.012.
Low, David A, Nathan J Weyand, and Michael J Mahan. 2001. "Roles of DNA Adenine Methylation in Regulating Bacterial Gene Expression and Virulence." *Infection and Immunity* 69 (12): 7197–7204. doi:10.1128/IAI.69.12.7197.
Marinus, Martin G, and Josep Casadesus. 2009. "Roles of DNA Adenine Methylation in Host-Pathogen Interactions : Mismatch Repair, Transcriptional Regulation, and More." *FEMS Microbiology Reviews* 33: 488–503. doi:10.1111/j.1574-6976.2008.00159.x.
Marsh, Adam G, and Annamarie A Pasqualone. 2014. "DNA Methylation and Temperature Stress in an Antarctic Polychaete , Spiophanes Tcherniai." *Frontiers in Physiology* 5 (May): 1–9. doi:10.3389/fphys.2014.00173.
Matsushita, Chieko, Osamu Matsushita, Seiichi Katayama, Junzaburo Minami, Kenichi Takai, and Akinobu Okabe. 1996. "An Upstream Activating Sequence Containing Curved DNA Involved in Activation of the Clostridium Perfringens Plc Promoter." *Microbiology* 142: 2561–66.
McAllister, Carl F, and Eric C Achberger. 1989. "Rotational Orientation of Upstream

Curved DNA Affects Promoter Function in Bacillus Subtilis." *The Journal of Biological Chemistry* 264 (18): 10451–56.

Meysman, Pieter, Julio Collado-vides, Enrique Morett, Roberto Viola, Kristof Engelen, and Kris Laukens. 2014. "Structural Properties of Prokaryotic Promoter Regions Correlate with Functional Features." *PloS One* 9 (2). doi:10.1371/journal.pone.0088717.

Moseman-Valtierra, Serena, Rosalinda Gonzalez, Kevin D. Kroeger, Jianwu Tang, Wei Chun Chao, John Crusius, John Bratton, Adrian Green, and James Shelton. 2011. "Short-Term Nitrogen Additions Can Shift a Coastal Wetland from a Sink to a Source of N2O." *Atmospheric Environment* 45 (26). Elsevier Ltd: 4390–97. doi:10.1016/j.atmosenv.2011.05.046.

Murray, Iain a, Tyson a Clark, Richard D Morgan, Matthew Boitano, Brian P Anton, Khai Luong, Alexey Fomenkov, Stephen W Turner, Jonas Korlach, and Richard J Roberts. 2012. "The Methylomes of Six Bacteria." *Nucleic Acids Research* 40 (22): 11450–62. doi:10.1093/nar/gks891.

Muttray, Annette F., Zhontang Yu, and William W. Mohn. 2001. "Population Dynamics and Metabolic Activity of Pseudomonas Abietaniphila BKME-9 within Pulp Mill Wastewater Microbial Communities Assayed by Competitive PCR and RT-PCR." *FEMS Microbiology Ecology* 38: 21–31. doi:10.1016/S0168-6496(01)00169-6.

Nagy, Zita, and Michael Chandler. 2004. "Regulation of Transposition in Bacteria." *Research in Microbiology* 155: 387–98. doi:10.1016/j.resmic.2004.01.008.

Noguchi, Hideki, Jungho Park, and Toshihisa Takagi. 2006. "MetaGene: Prokaryotic Gene Finding from Environmental Genome Shotgun Sequences." *Nucleic Acids Research* 34 (19): 5623–30. doi:10.1093/nar/gkl723.

Ochman, Howard, Jeffrey G Lawrence, and Eduardo A Groisman. 2000. "Lateral Gene Transfer and the Nature of Bacterial Innovation." *Nature* 405: 299–304.

Oremland, Ronald S., and Sandra Polcin. 1982. "Methanogenesis and Sulfate Reduction : Competitive and Noncompetitive Substrates in Estuarine Sediments." *Applied and Environmental Microbiology* 44 (6): 1270–76.

Orsi, William D, Virginia P Edgcomb, Glenn D Christman, and Jennifer F Biddle. 2013. "Gene Expression in the Deep Biosphere." *Nature* 499 (7457). Nature Publishing Group: 205–8. doi:10.1038/nature12230.

PacBio. 2013. *PACBIO ® GUIDELINES FOR SUCCESSFUL SMRTbell ™ LIBRARIES. Pacific Biosciences*.

Peng, Yu, Henry C M Leung, S. M. Yiu, and Francis Y L Chin. 2010. "IDBA - A Practical Iterative De Bruijn Graph De Novo Assembler." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6044 LNBI: 426–40. doi:10.1007/978-3-642-12683-3_28.

Perez-Martin, Jose, N Timmisg, and Victor De Lorenzo. 1994. "Co-Regulation by Bent DNA." *Journal of Biological Chemistry* 269 (36): 22657–62.

Prestat, Emmanuel, Maude M David, Jenni Hultman, Neslihan Taş, Regina Lamendella, Jill Dvornik, Rachel Mackelprang, et al. 2014. “FOAM (Functional Ontology Assignments for Metagenomes): A Hidden Markov Model (HMM) Database with Environmental Focus.” *Nucleic Acids Research* 42 (19): 1–7. doi:10.1093/nar/gku702.

Ratel, David, Jean-Luc Ravanat, François Berger, and Didier Wion. 2006. “N6-Methyladenine: The Other Methylated Base of DNA.” *BioEssays* 28 (3): 309–15. doi:10.1002/bies.20342.

Reisenauer, Ann, and Lucy Shapiro. 2002. “DNA Methylation Affects the Cell Cycle Transcription of the CtrA Global Regulator in Caulobacter.” *The EMBO Journal* 21 (18): 4969–77.

Roberts, Richard J, Tamas Vincze, Janos Posfai, and Dana Macelis. 2010. “REBASE--a Database for DNA Restriction and Modification: Enzymes, Genes and Genomes.” *Nucleic Acids Research* 38 (Database issue): D234–36. doi:10.1093/nar/gkp874.

Rocca, Jennifer D, Edward K Hall, Jay T Lennon, Sarah E Evans, Mark P Waldrop, James B Cotner, Diana R Nemergut, Emily B Graham, and Matthew D Wallenstein. 2014. “Relationships between Protein-Encoding Gene Abundance and Corresponding Process Are Commonly Assumed yet Rarely Observed.” *ISME J*, December. International Society for Microbial Ecology. http://dx.doi.org/10.1038/ismej.2014.252.

Ross, Eric D, Robert B Den, Philip R Hardwidge, and L James Maher Iii. 1999. “Improved Quantitation of DNA Curvature Using Ligation Ladders.” *Nucleic Acids Research* 27 (21): 4135–42.

Sayers, Eric W, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, et al. 2009. “Database Resources of the National Center for Biotechnology Information.” *Nucleic Acids Research* 37 (October 2008): 5–15. doi:10.1093/nar/gkn741.

Schippers, Axel, Lev N Neretin, Jens Kallmeyer, Timothy G Ferdelman, Barry a Cragg, R John Parkes, and Bo B Jørgensen. 2005. “Prokaryotic Cells of the Deep Sub-Seafloor Biosphere Identified as Living Bacteria.” *Nature* 433: 861–64. doi:10.1038/nature03302.

Seitz, Kiley W, Cassandre S Lazar, Kai-uwe Hinrichs, Andreas P Teske, and Brett J Baker. 2016. “Genomic Reconstruction of a Novel , Deeply Branched Sediment Archaeal Phylum with Pathways for Acetogenesis and Sulfur Reduction.” *The ISME Journal* 1 (10). Nature Publishing Group: 1–10. doi:10.1038/ismej.2015.233.

Singal, Rakesh, and Gordon D. Ginder. 1999. “DNA Methylation.” *Blood* 93 (12): 4059–71.

Singer, Esther, John F Heidelberg, Ashita Dhillon, and Katrina J Edwards. 2013. “Metagenomic Insights into the Dominant Fe (II) Oxidizing Zetaproteobacteria from an Iron Mat at Lo ’ Ihi , Hawai ’ I.” *Frontiers in Microbiology* 4 (March):

1–9. doi:10.3389/fmicb.2013.00052.
Souza, Claudiana P, Bianca C Almeida, Rita R Colwell, and Irma N G Rivera. 2011. "The Importance of Chitin in the Marine Environment." *Marine Biotechnology* 13: 823–30. doi:10.1007/s10126-011-9388-1.
Srikhanta, YN, SJ Dowideit, JL Edwards, ML Falsetta, H-J Wu, OB Harrison, KL Fox, et al. 2009. "Phasevarions Mediate Random Switching of Gene Expression in Pathogenic Neisseria." *PLOS Pathology* 5.
Srikhanta, Yogitha N, Tina L Maguire, Katryn J Stacey, and Sean M Grimmond. 2005. "The Phasevarion : A Genetic System Controlling Coordinated , Random Switching of Expression of Multiple Genes." *Proceedings of the National Academy of Sciences* 102 (15): 5547–51.
Suderman, M., P. O. McGowan, a. Sasaki, T. C. T. Huang, M. T. Hallett, M. J. Meaney, G. Turecki, and M. Szyf. 2012. "Conserved Epigenetic Sensitivity to Early Life Experience in the Rat and Human Hippocampus." *Proceedings of the National Academy of Sciences* 109: 17266–72. doi:10.1073/pnas.1121260109.
Szyf, Moshe. 2009. "The Early Life Environment and the Epigenome." *Biochimica et Biophysica Acta - General Subjects* 1790 (9). Elsevier B.V.: 878–85. doi:10.1016/j.bbagen.2009.01.009.
Terpstra, T.J. 1952. "The Asymptotic Normality and Consistency of Kendall's Test Against Trend, When Ties Are Present in One Ranking." *Indagationes Mathematicae* 14: 327–33.
Tolba, Sahar T.M., A. Abd Allah Nagwa, and Dina Hatem. 2013. "Molecular Characterization of Rare Actinomycetes Using 16S rRNA-RFLP." *African Journal of Biological Sciences* 9 (March): 185–97.
Tracey, M.A. 1957. "Chitin." *Review of Pure and Applied Chemistry* 7: 1–13.
Travers, Andrew. 1989. "DNA Structure. Curves with a Function." *Nature* 341: 184–85.
Tully, Benjamin J, Rohan Sachdeva, Karla B Heidelberg, and John F Heidelberg. 2014. "Comparative Genomics of Planktonic Flavobacteriaceae from the Gulf of Maine Using Metagenomic Data." *Microbiome* 2 (1): 1–14. doi:10.1186/2049-2618-2-34.
Voet, Donald, Judith G. Voet, and Charlotte W. Pratt. 2008. *Fundamentals of Biochemistry*. 3rd ed.
Whitman, William B, David C Coleman, William J Wiebe, and David C Colemant. 1998. "Prokaryotes: The Unseen Majority." *Proceedings of the National Academy of Sciences of the United States of America* 95 (12): 6578–83.
Wion, Didier, and Josep Casadesus. 2006. "N 6 -Methyl-Adenine : An Epigenetic Signal for DNA – Protein Interactions." *Nature Reviews Microbiology* 4 (March): 183–93. doi:10.1038/nrmicro1350.
Wojciechowski, Marek, Honorata Czapinska, and Matthias Bochtler. 2012. "CpG Underrepresentation and the Bacterial CpG-Specific DNA Methyltransferase M.MpeI." *Proceedings of the National Academy of Sciences* 110 (1): 105–10.

doi:10.1073/pnas.1207986110.
Wood, Derrick E, and Steven L Salzberg. 2014. “Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments.” *Genome Biology* 15 (3): R46. doi:10.1186/gb-2014-15-3-r46.

Appendix A

CHAPTER 1 SUPPLEMENTARY MATERIAL

Table A1: Radionuclide activity in Core R

Sample Depth (cm)	$xs^{210}Pb$ (mBq/g)	^{7}Be (mBq/g)	^{137}Cs (mBq/g)
0-1	18.89	8.59	0.00
1-2	40.60	7.34	0.00
2-3	60.86	5.73	0.00
3-4	37.91	3.38	0.00
4-5	8.21	0.00	0.00
5-6	0.00	0.00	0.00
6-7	0.00	0.00	0.00
7-8	0.00	0.00	0.00
8-9	0.00	0.00	0.00
9-10	0.00	0.00	0.00
10-12	0.00	0.00	0.00
12-14	0.00	0.00	0.00
14-16	0.00	0.00	0.00
16-18	0.00	0.00	0.00
18-20	0.00	0.00	0.00
20-22	0.00	0.00	0.00
22-24	0.00	0.00	0.00
24-26	0.00	0.00	0.00
26-28	0.00	0.00	0.00
28-30	0.00	0.00	0.00
30-32	0.00	0.00	0.00

The 0-4 cm section is comprised of recently deposited tidally mixed or bioturbated sediment. 4-5 cm presents a discontinuity in beryllium-7, but excess lead-210 indicates sediment deposited 44.40 years ago at most. The lack Cs-137 and all other radionuclide signatures suggest that 6-32 cm is comprised of sediment deposited 50-100+ years ago.

Table A2: IDBA assembly statistics

	Total	High Abundance	Low Abundance	Singletons	Viral contigs
3-6 cm					
Number of contigs	18800	4617	191	13647	345
Number of bases	2,493,769	1,214,155	283,733	917,933	77,948
Minimum length	34	200	200	34	34
Maximum length	5808	3919	342	199	5808
Median length	73	279	248	54	162
Mean length	132	314	252	67	226
N50 contig length	251	299	249	75	336
N80 contig length	79	251	238	45	250
12-15 cm					
Number of contigs	49214	13815	3386	31538	475
Number of bases	8,218,771	4,841,641	900,041	2,319,312	157,777
Minimum length	34	200	200	34	35
Maximum length	5419	2782	484	281	5419
Median length	95	309	258	65	296
Mean length	167	350	266	73	332
N50 contig length	276	346	261	83	334
N80 contig length	112	267	241	50	264

Assemblies were performed with settings --mink 18 --maxk 36 --step 2 --similar 0.97. Viral and fungal contigs were identified with Kraken, while potential outlier contigs with high coverage were identified with Bonferroni outlier tests. Table continues on next page.

Table A2, continued

24-27 cm					
Number of contigs	18361	4010	1114	13105	132
Number of bases	2,474,539	1,238,974	295,037	893,040	47,488
Minimum length	34	200	201	34	35
Maximum length	5419	995	429	259	5419
Median length	76	282	256	53	287
Mean length	134	309	265	68	360
N50 contig length	250	301	260	76	328
N80 contig length	83	252	241	46	269

Table A3: CpG *met* and *umt* bootstrap standard errors and coefficients of variation

Phylum	Depth (cm)	*met* SE	*umt* SE	*met* CV	*umt* CV
	3-6	2.0161	2.0798	0.0178	0.0126
	12-15	2.2221	2.1513	0.014	0.0094
Actinobacteria	24-27	2.6162	1.5361	0.0137	0.0068
	3-6	6.321	5.568	0.0565	0.0297
	12-15	6.8338	8.6293	0.0487	0.0355
Bacteroidetes	24-27	7.312	4.5488	0.042	0.0182
	3-6	6.3055	5.5369	0.0566	0.0313
	12-15	8.1568	8.1293	0.0584	0.0333
Chloroflexi	24-27	9.8981	5.0229	0.0569	0.0182
	3-6	5.6114	5.6426	0.0503	0.0315
	12-15	7.3147	7.6035	0.0526	0.0316
Firmicutes	24-27	8.7326	6.6896	0.0505	0.0264
	3-6	6.4639	5.4965	0.0549	0.0310
	12-15	7.6221	6.9012	0.0509	0.0286
Proteobacteria	24-27	8.3157	6.1184	0.0478	0.0240

Values were calculated from 10,000 bootstrap replicate estimates.

Table A4: Hartigans' dip test for unimodality results

Phylum	Depth (cm)	*met*		*umt*	
		D statistic	**p-value**	**D statistic**	**p-value**
Actinobacteria	3-6	0.049296	< 2.2e-16	0.051216	< 2.2e-16
	12-15	0.038012	< 2.2e-16	0.042741	< 2.2e-16
	24-27	0.028809	0.0001099	0.079759	< 2.2e-16
Bacteroidetes	3-6	0.036251	0.7084	0.058167	0.05443
	12-15	0.042867	0.4047	0.080687	0.000598
	24-27	0.031532	0.8958	0.094608	1.201e-05
Chloroflexi	3-6	0.04698	0.01705	0.052203	0.004281
	12-15	0.057596	0.0008166	0.050336	0.007174
	24-27	0.028701	0.5416	0.085675	< 2.2e-16
Firmicutes	3-6	0.051648	< 2.2e-16	0.061538	< 2.2e-16
	12-15	0.040966	1.539e-05	0.051472	< 2.2e-16
	24-27	0.019763	0.2671	0.084154	< 2.2e-16
Proteobacteria	3-6	0.041409	< 2.2e-16	0.044161	< 2.2e-16
	12-15	0.038322	< 2.2e-16	0.046205	< 2.2e-16
	24-27	0.018309	0.0001406	0.086463	< 2.2e-16

H_0: distribution is unimodal. H_a: distribution is at least bimodal.

Table A5: Two-tailed Jonckheere-Terpstra trend test and Brown-Forsythe variance test results

Phylum	Metric	Jonckheere-Terpstra		Brown-Forsythe	
		Test statistic	**p-value**	**Test statistic**	**p-value**
Actinobacteria	***met***	723680	0.002	68.286	< 2.2e-16
	umt	1014800	0.002	116.34	< 2.2e-16
Bacteroidetes	***met***	6662	0.002	4.5444	0.01165
	umt	8599.5	0.47	5.6229	0.004156
Chloroflexi	***met***	26873	0.002	16.35	1.408e-07
	umt	37382	0.008	17.643	4.235e-08
Firmicutes	***met***	253740	0.002	40.351	< 2.2e-16
	umt	345770	0.002	57.716	< 2.2e-16
Proteobacteria	***met***	4449700	0.002	140.98	< 2.2e-16
	umt	6127000	0.002	195.9	< 2.2e-16

Explanations of null and alternative hypotheses for these tests can be found in the Chapter 1 Results section *Metagenome CpG methylation*.

Appendix B

CHAPTER 2 SUPPLEMENTARY MATERIAL

Table B1: TBLASTN results for DNA methyltransferase alignments, 5-methylcytosine specific

Class	Target sequence	MTase	Mean % positive matches	Mean query coverage
Actinobacteria	ACCGGT	M.AgeI	48.06	99.00
Actinobacteria	AGCT	M.AluI	55.85	96.00
Actinobacteria	CCNGG	M.SsoII	67.48	92.33
Actinobacteria	CCWGG	Dcm	60.21	95.16
Actinobacteria	CTNAG	M.DdeI	46.43	94.00
Actinobacteria	GAGCTC	M.SacI	77.31	96.50
Actinobacteria	GCCGGC	M.NaeI	73.53	99.33
Actinobacteria	GGCC	M.FnuDI	79.17	95.00
Actinobacteria	GRCGYC	M.HgiDI	70.05	94.50
Actinobacteria	GTGCAC	M.ApaLI	47.96	90.00
Alphaproteobacteria	ACCGGT	M.AgeI	46.51	98.00
Alphaproteobacteria	AGCT	M.AluI	45.41	91.50
Alphaproteobacteria	CCWGG	Dcm	54.93	93.50
Alphaproteobacteria	GRCGYC	M.HgiDI	36.98	78.00

Results are for alignments whose alignment lengths are greater than or equal to 90% of the MTase query length. MTases with an "M" prefix are associated with a respective restriction endonuclease, while those without a prefix are orphan MTases.

Table B1, continued

Class	Target sequence	MTase	Mean % positive matches	Mean query coverage
Bacilli	ACCGGT	M.AgeI	45.52	99.00
Bacilli	CCNGG	M.SsoII	80.26	98.67
Bacilli	CCWGG	Dcm	55.93	98.00
Bacilli	CTNAG	M.DdeI	52.61	91.00
Bacilli	GCAGC	M.BbvI	69.19	94.00
Bacilli	GCNGC	M.Bsp6I	73.94	92.00
Bacilli	GGCC	M.BspRI	39.95	81.00
Bacilli	GGCC	M.BsuRI	78.39	99.00
Bacilli	GGNCC	M.Sau96I	73.15	97.00
Bacilli	RGCGCY	M.NgoBI	59.56	97.00
Betaproteobacteria	CCGG	M.HpaII	74.02	97.00
Betaproteobacteria	CCWGG	Dcm	68.62	93.91
Betaproteobacteria	CGCG	M.BepI	70.56	99.00
Betaproteobacteria	CTNAG	M.DdeI	45.03	88.00
Betaproteobacteria	GCAGC	M.BbvI	57.22	98.00
Betaproteobacteria	GGCC	M.FnuDI	81.96	97.00
Betaproteobacteria	GRCGYC	M.HgiDI	69.37	97.00
Betaproteobacteria	GRCGYC	M.HgiGI	72.95	97.00
Clostridia	ACCGGT	M.AgeI	43.06	99.00
Clostridia	CCGG	M.HhaI	84.83	98.00
Clostridia	CCWGG	Dcm	53.28	93.33
Clostridia	CG	M.SssI	49.51	97.00

Table B1, continued

Class	Target sequence	MTase	Mean % positive matches	Mean query coverage
Clostridia	CTNAG	M.DdeI	52.69	89.50
Clostridia	GGCC	M.FnuDI	81.52	95.00
Clostridia	GGNCC	M.Sau96I	74.77	99.00
Clostridia	GRCGYC	M.HgiDI	67.68	95.00
Clostridia	GRCGYC	M.HgiGI	73.91	90.00
Clostridia	GTGCAC	M.ApaLI	45.06	82.50
Deinococci	GGCC	M.FnuDI	43.31	78.00
Deltaproteobacteria	ACCGGT	M.AgeI	44.60	99.00
Deltaproteobacteria	CCWGG	Dcm	65.06	96.30
Deltaproteobacteria	GAGCTC	M.SacI	51.02	98.00
Gammaproteobacteria	ACCGGT	M.AgeI	47.76	99.00
Gammaproteobacteria	AGCT	M.AluI	60.28	97.00
Gammaproteobacteria	CCGG	M.HpaII	49.54	93.70
Gammaproteobacteria	CCGG	M.MspI	62.68	96.00
Gammaproteobacteria	CCWGG	Dcm	58.84	96.95
Gammaproteobacteria	CGCG	M.BepI	68.22	99.00
Gammaproteobacteria	GAGCTC	M.SacI	42.21	98.00
Gammaproteobacteria	GGCC	M.BsuRI	41.44	88.00
Gammaproteobacteria	GGCC	M.FnuDI	83.02	95.00
Gammaproteobacteria	GGNCC	M.PspPI	75.12	96.50
Gammaproteobacteria	GGWCC	M.SinI	77.71	97.50
Gammaproteobacteria	GRCGYC	M.HgiDI	69.90	96.00
Gammaproteobacteria	GRCGYC	M.HgiGI	73.39	98.00

Table B2: TBLASTN results for DNA methyltransferase alignments, N6-adenine specific

Class	Target sequence	MTase	Mean % positive matches	Mean query coverage
Actinobacteria	CTGCAG	M.BsuBI	60.81	91.00
Actinobacteria	CTGCAG	M.PstI	62.5	97.00
Actinobacteria	GATC	Dam	62.32	99.00
Alphaproteobacteria	CAATTG	M.MunI	42.53	92.00
Alphaproteobacteria	CTCGAG	M.XhoI	66.67	99.00
Alphaproteobacteria	GANTC	M.HhaII	41.55	96.67
Alphaproteobacteria	GATC	Dam	55.49	93.17
Alphaproteobacteria	GTTAAC	M.HpaI	47.54	83.00
Bacilli	GATC	Dam	59.91	96.03
Bacilli	GGATG	M.FokI	65.50	97.67
Bacilli	GGTACC	M.KpnI	49.65	88.50
Betaproteobacteria	GATC	Dam	47.33	97.39
Clostridia	CAATTG	M.MunI	70.43	92.00
Clostridia	CTCGAG	M.BstVI	44.25	99.00
Clostridia	GATATC	M.EcoRV	47.04	98.00
Clostridia	GATC	Dam	64.64	97.71
Clostridia	GATC	M.MboIA	62.21	99.00
Clostridia	GATC	M.MboIB	69.05	93.00
Clostridia	GGTACC	M.KpnI	47.17	85.00
Clostridia	GGTNACC	M.EcaI	43.37	86.00

Results are for alignments whose alignment lengths are greater than or equal to 90% of the MTase query length. MTases with an “M” prefix are associated with a respective restriction endonuclease, while those without a prefix are orphan MTases.

Table B2, continued

Class	Target sequence	MTase	Mean % positive matches	Mean query coverage
Deinococci	CTCGAG	M.AbrI	80.00	96.00
Deinococci	CTCGAG	M.BstVI	61.11	99.00
Deinococci	CTCGAG	M.XhoI	64.67	99.00
Deltaproteobacteria	CTCGAG	M.AbrI	81.78	100
Deltaproteobacteria	CTCGAG	M.BstVI	61.47	95.50
Deltaproteobacteria	CTCGAG	M.XhoI	68.69	99.00
Deltaproteobacteria	CTGCAG	M.BsuBI	68.71	98.00
Deltaproteobacteria	CTGCAG	M.PstI	60.82	98.25
Deltaproteobacteria	GATC	Dam	58.25	94.67
Gammaproteobacteria	CAATTG	M.MunI	48.25	84.00
Gammaproteobacteria	CTCGAG	M.AbrI	82.86	100
Gammaproteobacteria	CTCGAG	M.BstVI	61.62	98.50
Gammaproteobacteria	CTGCAG	M.BsuBI	64.44	97.33
Gammaproteobacteria	CTGCAG	M.PstI	61.50	96.00
Gammaproteobacteria	GANTC	M.HhaII	70.79	93.00
Gammaproteobacteria	GATATC	M.EcoRV	75.25	99.00
Gammaproteobacteria	GATC	Dam	86.15	99.11
Gammaproteobacteria	GATC	M.MboIA	49.44	99.87
Gammaproteobacteria	GATC	M.MboIB	46.29	96.00
Gammaproteobacteria	GGTACC	M.KpnI	64.35	93.50

Table B3: TBLASTN results for DNA methyltransferase alignments, N4-cytosine specific

Class	Target sequence	MTase	Mean % positive matches	Mean query coverage
Actinobacteria	AGTACT	M.ScaI	41.34	82.00
Actinobacteria	GGCCNNNNNGGCC	M.SfiI	44.02	89.00
Alphaproteobacteria	CCWGG	M.MvaI	51.03	93.00
Bacilli	AGTACT	M.ScaI	38.49	80.00
Bacilli	CCWGG	M.MvaI	48.91	99.00
Betaproteobacteria	CAGCTG	M.PvuII	79.69	93.00
Betaproteobacteria	GGCCNNNNNGGCC	M.SfiI	43.97	88.00
Clostridia	AGTACT	M.ScaI	39.67	85.00
Clostridia	CCWGG	M.MvaI	56.55	93.83
Gammaproteobacteria	CAGCTG	M.PvuII	87.43	94.00

Results are for alignments whose alignment lengths are greater than or equal to 90% of the MTase query length.

Appendix C

PROPRIETARY RELEASE STATEMENT

C.1 Proprietary release statement for use of methylation profiling platform

I (Adam Marsh) hereby grant permission to you (Ian Rambo) to utilize the data and analytics derived from my proprietary methylation profiling platform currently under license by University of Delaware to Genome Profiling LLC. You may utilize the data and analytic results in any fashion or form that you deem necessary for your thesis and for the publication of your work. All components of the methylation work that you have compiled into your thesis to date are expressly provided by myself for your open academic use.

This release only encompasses academic uses of the data and analytics consistent with the open dissemination and publication of your current thesis work. Nothing contained in this release statement shall be deemed to grant any right in or to or license under any ideas, know-how, technology, inventions or intellectual property for DNA methylation profiling owned by the University of Delaware and currently under license to Genome Profiling LLC.