

Fall 2013

Patterns of cytosine methylation in the genome of *Caenorhabditis elegans*

Kazufusa Okamoto

Follow this and additional works at: <https://scholars.unh.edu/dissertation>

Recommended Citation

Okamoto, Kazufusa, "Patterns of cytosine methylation in the genome of *Caenorhabditis elegans*" (2013). *Doctoral Dissertations*. 751.
<https://scholars.unh.edu/dissertation/751>

This Dissertation is brought to you for free and open access by the Student Scholarship at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact nicole.hentz@unh.edu.

PATTERNS OF CYTOSINE METHYLATION IN THE GENOME OF *CAENORHABDITIS ELEGANS*

BY

KAZUFUSA OKAMOTO

Baccalaureate Degree (B.S.) Colorado State University, 2004

DISSERTATION

Submitted to the University of New Hampshire

in Partial Fulfillment of

the Requirements for the Degree of

Doctor of Philosophy

In

Biochemistry

September, 2013

UMI Number: 3575992

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3575992

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.

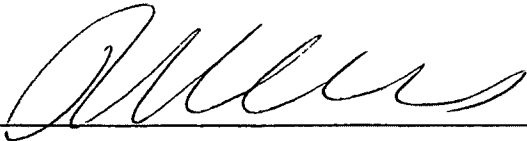
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code.




ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

This thesis has been examined and approved.



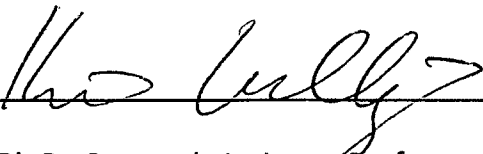
Thesis Director, W. Kelley Thomas, Ph.D., Professor (Biochemistry)



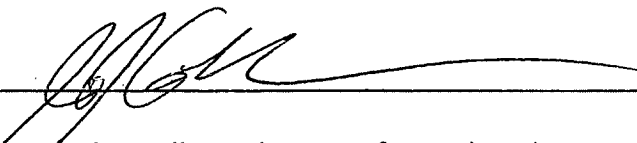
Rick Cote, Ph.D., Professor (Biochemistry)



Feixia Chu, Ph.D., Assistant Professor (Biochemistry)



Kevin Culligan, Ph.D., Research Assistant Professor (Genetics)



John Collins, Ph.D., Professor (Biochemistry)

8/20/2013

Date

ACKNOWLEDGMENTS

I dedicate this dissertation to the many people that supported and believed in me. For without you, this would not be possible.

My deepest appreciation and gratitude goes to my adviser, Dr. W.Kelley Thomas who shaped the scientist in me with patient guidance to find truth through process and fearlessness.

I would like to thank my mother who always pushed me to do better. Finally, I would like to thank the love of my life, Danielle, for the constant support and encouragement.

LIST OF TABLES

TABLE	PAGE
Table. 1 Candidate DNA Methyltransferases.....	18
Table. 2 Assembly Statistics of Genome Wide Bisulfite Sequencing.....	52
Table. 3 Assembly Statistics of Genome Wide Bisulfite Sequencing.....	55
Table. 4 Expected Shared Sites of MeC Between Datasets	66
Table. 5 Expected verses Observed Frequency of Shared Sites Enriched and GWBS Datasets.....	76

LIST OF FIGURES

FIGURE	PAGE
Figure. 1 Normalized frequencies of DNA methylation context and evolution.....	9
Figure. 2 Overview of the MEME Suite.....	16
Figure. 3 Worm Base ID Y75B8A.6 Organization of Motifs.....	19
Figure. 4 Worm Base ID T09A5.8 Organization of Motifs.....	20
Figure. 5 Asymmetric vs. Symmetric Methylation.....	31
Figure. 6 Normalized Levels of DNA Methylation per Chromosome.....	33
Figure. 7 Distribution of Methylated Cytosines per Chromosome.....	34
Figure. 8 Genic vs. Intergenic Methylation per Chromosome	36
Figure. 9 Categorical Distribution of Methylation	37
Figure. 10 Constitutive vs. Facultative Methylation per Chromosome per Category	42
Figure. 11 Constitutive vs. Facultative Methylation per Chromosome per Category	43
Figure. 12 Ratios of Symmetric vs. Asymmetric Constitutive Methylation.....	53
Figure. 13 Comparison of Enriched vs. GWBS Constitutive vs. Facultative Methylation per Chromosome per Category.....	56

Figure. 14 Comparison of the Distribution of Constitutive vs. Facultative Methylation per Chromosome.....	59
Figure. 15 Distribution of MeC per Chromosome I-III N2 Enriched vs. GWBS.....	61
Figure. 16 Distribution of MeC per Chromosome IV-X N2 Enriched vs. GWBS.....	62
Figure. 17 Comparison of Constitutively Methylated Sites per Category N2 GWBS vs. N2 Enriched.....	63
Figure. 18 Comparison of All MeC Sites vs. Constitutive Sites per Category N2 GWBS vs. PB306 vs. VC2864.....	65
Figure. 19 Clustal W Alignment MeC Conformation PCR Product Against the Reference.....	68
Figure. 20 Normalized Frequencies of DNA Methylation Context and Evolution.....	71
Figure. 21 <i>C.elegans</i> Intergenic vs. Genic Mutation Rates.....	78

TABLE OF CONTENTS

ACKNOWLEDGMENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

CHAPTER	PAGE
INTRODUCTION.....	1
I. IDENTIFICATION OF DNA METHYLTRANSFERASES IN <i>C. ELEGANS</i>	11
Background.....	11
Methods.....	16
Results and Discussion.....	21
Conclusion.....	22
II. PATTERN OF C5 DNA METHYLATION IN <i>C. ELEGANS</i>	23
Background.....	23
Methods.....	27
Results and Discussion.....	29

Conclusion.....	44
III. SHIFTING PATTERNS OF DNA METHYLATION.....	46
Background.....	46
Methods.....	48
Results and Discussion.....	50
Conclusion.....	69
IV. THE MUTAGENIC CONSEQUENCES OF DNA METHYLATION.....	72
Background.....	73
Methods.....	74
Results and Discussion.....	75
Conclusion.....	79
REFERENCES.....	80

ABSTRACT

PATTERNS OF CYTOSINE METHYLATION IN THE GENOME OF *CAENORHABDITIS ELEGANS*

BY

KAZUFUSA OKAMOTO

University of New Hampshire, September 2013

Recent large-scale comparative analysis of cytosine DNA methylation across diverse eukaryotes suggest that early features of DNA methylation present in the last common ancestor of all eukaryotes some 1.6 to 1.8 billion years ago included the methylation of gene bodies and transposable elements (Zemach, McDaniel et al. 2010; Parfrey, Lahr et al. 2011). These potentially ancient patterns may reflect a primitive role of methylation in transcriptional fidelity and as a mechanism to protect the germ line from transposon, or repeat, mediated mutation. Because spurious transcription and mutation are hypothesized to be among the critical limiting factors to genome size, an ancient role for methylation in support of fidelity of transcription and genome stability suggests a possible link with the origin of eukaryotes. As a consequence,

understanding the roles of methylation across diverse eukaryotes will be critical to understanding the evolution of methylation and its role in the evolution of genome complexity.

In light of these observations it is perplexing that one of our key model eukaryotes, the nematode (*Caenorhabditis elegans*) is assumed to lack active DNA methylation. In fact, *C. elegans* is often invoked to suggest the dispensability of methylation in multicellular animals (Feng, Cokus et al. 2010; Zemach, McDaniel et al. 2010). Historically, this view has been based on crude assays using methylation sensitive restriction enzymes (Simpson, Johnson et al. 1986) that lack the sensitivity to identify low levels of methylation.

While it is clear that the genome of *C. elegans* is not highly methylated, in this thesis we used comparative genomics and genome wide bisulfite sequencing to show that: 1) The genome of *C. elegans* appears to encode at least three DNA methyltransferases and a DNA methyltransferase associated protein; 2) the genome of *C. elegans* is methylated in a pattern consistent with the proposed basal eukaryotic pattern and 3) that cytosine methylation is not a major contributor to the basal rate and pattern of mutation in the genome of *C. elegans*. Based on these observations we contend that *C. elegans* represents an ideal model for the study of the basal roles of DNA methylation shared by all eukaryotes.

INTRODUCTION

When comparing complete genomic sequences across diverse phylogenetic lineages, a general pattern emerges where there is an increase in genome size from prokaryotes to multicellular eukaryotes. The changes include increase in gene number, resulting from the retention of duplicate genes, and an increases in the abundance of spliceosomal introns and mobile genetic elements (Lynch and Conery 2003). This trend of evolving increased genome size and ultimately genome complexity may arise from a change in the drift selection balance. In that hypothesis, the balance can be shifted towards drift and the power of selection can be dampened when population sizes decrease, a common feature associated with increased complexity. Here we propose that epigenetic factors can also contribute to the evolution of complexity by reducing the deleterious effects of increased genome complexity by suppressing spurious transcription and the spread of transposable elements.

DNA methylation is perhaps the best characterized epigenetic mechanism. DNA methylation is found in the genomes of diverse organisms including both prokaryotes and eukaryotes. In prokaryotes, DNA methylation occurs on both cytosine and adenine bases and encompasses part of the host restriction system (Wilson and Murray 1991). In eukaryotes, methylation seems to be confined primarily to cytosine bases and is associated with a repressed chromatin state and inhibition of gene expression (Bird and

Wolffe 1999). DNA methylation has been proven to be involved in a number of biological processes such as regulation of imprinted genes, X chromosome inactivation, and tumor suppressor gene silencing in cancerous cells. It also acts as a protection mechanism against pathogen DNA and transposable elements (Chandler and Walbot 1986; Yoder, Walsh et al. 1997; Matzke, Mette et al. 2000) and DNA methylation is essential for viability in mice, since targeted disruption of the DNA methyltransferase enzymes results in lethality (Li, Bestor et al. 1992; Okano, Bell et al. 1999).

DNA Methyltransferases and Associated Proteins in *C. elegans*

The most extensively studied DNA methyltransferase enzymes are that of mammals. Mammalian cytosine DNA methyltransferases fit into two general classes based on the DNA substrate they prefer (Klose and Bird 2006). The *de novo* methyltransferases DNMT3a and DNMT3b are mostly responsible for cytosine methylation at previously unmethylated sites, whereas the maintenance methyltransferase DNMT1 copies pre-existing methylation patterns onto the new DNA strand during DNA replication (Okano, Xie et al. 1998). A fourth DNA methyltransferase, DNMT2, shows weak DNA methyltransferase activity *in vitro* (Hermann, Schmitt et al. 2003) and targeted deletion of the DNMT2 gene in mouse embryonic stem cells causes no detectable effect on DNA methylation. This suggests that this enzyme has little involvement in setting DNA methylation patterns (Okano, Xie et al. 1998). In mouse DNMT3L is a DNMT-related protein that does not contain DNA methyltransferase

activity, but physically associates with DNMT3a and DNMT3b and modulates their catalytic activity (Suetake, Shinozaki et al. 2004). In combination, these *de novo* and maintenance methyltransferases constitute the core enzymatic components of the DNA methylation system in mammals (Klose and Bird 2006).

Prior to this study it was not clear if *Caenorhabditis elegans* is capable of DNA methylation. However, some recent studies (Lyko 2001; Zhang, Yazaki et al. 2006; Schaefer and Lyko 2007; Pomraning, Smith et al. 2009) have spurred interest in re-evaluating organisms that have been long believed to live in the absence of DNA methylation. For example, *Drosophila melanogaster* was also once considered a classic example of an organism that functions without DNA methylation (Bird and Tweedie 1995) yet it was subsequently reported that *Drosophila* possesses a functioning DNA methylation system and low levels of genomic methylation were discovered (Lyko 2001). This is in spite of the fact that *Drosophila* does not encode homologs of any of the known DNMT genes.

The most recent DNA methylation study in *C. elegans* by Gutiérrez and Sommer (2004) proposed a recent loss of the DNA methylation system in *C. elegans*. This study was based on a BLAST search for orthologous sequences to the *Drosophila* dnmt-2 gene in the EST and genomic DNA sequences of three nematode species; *C. elegans*, *C. briggsae*, and *P. pacificus*. Although orthologous sequences were found in all nematodes surveyed, expression of the gene was only confirmed in *P. pacificus* leading

to the suggestion that functional methylation may have been lost in the lineage leading to *C. elegans*. More recently DNA methylation in a parasitic nematode (*Trichinella spiralis*) was shown to be stage specific (Gao, Liu et al. 2012). This study also concluded that the *C. elegans* genome contains a Dnmt1, but out of the 11 species of nematodes tested *T. spiralis* was the only one encoding a DNMT3 homologue. The lack of a clear Dnmt3 homologue and thus questionable capacity for de novo methylation is nevertheless a common feature across eukaryotes shown to actively methylate their genomes (Jeltsch 2010).

The Roles of DNA Methylation

There are two general mechanisms by which DNA methylation inhibits gene expression. First, modification of cytosine bases can inhibit the association of some DNA-binding factors with their corresponding DNA recognition sequences (Watt and Molloy 1988). Second, proteins that recognize methyl-CpG can elicit the repressive potential of methylated DNA (Boyes and Bird 1991). In mammals methyl-CpG-binding proteins (MBPs) use transcriptional co-repressor molecules to silence transcription and to modify surrounding chromatin, providing a link between DNA methylation and chromatin remodeling and modification (Hendrich and Bird 1998; Jones, Thomas et al. 1998).

Context of Heritable DNA Methylation

Cytosine residues at CpG dinucleotides are the preferred targets for DNA methylation in mammals, while methylation at both CpG and CpNpG (where N is any base) sequence contexts is also common in plants (Gruenbaum, Naveh-Manly et al. 1981) (Fig. 1). The symmetry of the CpG and CpNpG sites was proposed to be important for stable maintenance of methylation patterns throughout DNA replication cycles. After replication, a maintenance methyltransferase could readily methylate C residues in the newly synthesized strand, if the parental strand contained an MeC in the complementary sequence (Gruenbaum, Cedar et al. 1982). This semi-conservative model predicts that the methylation pattern at non-symmetrical sequence contexts would not be efficiently maintained and should be lost after several cell divisions. However, cytosine methylation of non-symmetrical sequence contexts were reported in mammals (Ramsahoye, Binizskiewicz et al. 2000; Lister, Pelizzola et al. 2009), in fungi (Selker, Fritz et al. 1993; Goyon, Nogueira et al. 1994) and in plants (Cao, Aufsatz et al. 2003) and could contribute to the regulation of gene expression (Cao, Aufsatz et al. 2003). Therefore, non-symmetrical methylation patterns have to be maintained by a mechanism different to that proposed in the semi-conservative model or they have to be established *de novo* after each DNA replication cycle (Pélissier, Tutois et al. 1996).

Since, little is known about the molecular mechanisms that target DNA sequences for *de novo* methylation, it is not clear if the processes involved in *de novo* methylation of symmetrical sequences are different from those taking place in the *de*

de novo methylation of non-symmetrical sequence contexts (Pélissier, Tutois et al. 1996). Because of the difficulty in analyzing cells where *de novo* methylation is initiated, the frequent appearance of symmetrical methylation patterns may simply reflect that only these patterns are efficiently maintained (Pélissier, Tutois et al. 1996).

DNA Methylation and Mutation

Methylation of cytosine residues was first demonstrated to be mutagenic in *E. coli* (Coulondre, Miller et al. 1978). These initial studies identified methylated cytosines as hotspots for spontaneous base substitutions. Mutations which occur at CpG dinucleotides are easily recognized because of the nature of base substitutions. Deamination of MeC at CpG dinucleotides results in the formation of TpG. Alternatively, if deamination occurs on the complementary DNA strand CpA is generated. The conversion of MeC to T is believed to be more likely the result of endogenous mutagenic processes rather than mutagenesis caused by exogenous factors (Rideout, Coetzee et al. 1990). Methylation of cytosine at a CpG dinucleotide increases the probability of a C→T or corresponding G→A transition mutation between 12- and 42-fold (Cooper and Youssoufian 1988).

5-Methyl cytosine (MeC) in DNA is genetically unstable. Methylated CpG (mCpG) sequences frequently undergo mutation resulting in a general depletion of this dinucleotide sequence in mammalian genomes. In human genetic disease and cancer relevant genes, mCpG sequences are mutational hotspots. It is an almost universally

accepted that these mutations are caused by random deamination of MeC (Gonzalzo and Jones 1997). However, it is plausible that mCpG transitions are not only caused by spontaneous deamination of MeC in double-stranded DNA but by other processes including, for example, mCpG-specific base modification by endogenous or exogenous mutagens or carcinogens (Pfeifer 2006). When adjacent to another pyrimidine, MeC preferentially undergoes photo-induced pyrimidine dimer formation (Pfeifer 2006). Furthermore, certain polycyclic aromatic hydrocarbons form guanine adducts and induce G to T transversion mutations with high selectivity at mCpG sequences (Gonzalzo and Jones 1997).

The increased deamination rate of MeC relative to C, however, still does not account for the high frequency of mutagenesis observed at CpG sites. One explanation may be that G-T mispairs resulting from deamination of MeC are more difficult for the cell to repair than G-U mispairs which can result from the deamination of cytosine, since thymine (unlike uracil) is a normal component of DNA. A higher efficiency of repair of G-U but not G-T mismatches by the well characterized uracil-DNA glycosylase (UDG) enzyme may also contribute to the increased frequency of mutagenesis caused by MeC deamination (Gonzalzo and Jones 1997). Excision of U has been found to be as much as 6000-fold more efficient than excision of T at identical template sites using extracts from human colonic mucosa (Schmutte, Yang et al. 1995).

Phylogenetic Distribution of DNA Methylation

In animals, the level and pattern of methylation varies dramatically among major lineages. It was believed that the nematode *Caenorhabditis elegans* has little to no methylated DNA, since the genome lacks detectable methylated cytosine (MeC) and does not encode a conventional DNA methyltransferase (Regev, Lamb et al. 1998; Lyko 2001; Kunert, Marhold et al. 2003; Gutierrez and Sommer 2004; Vandegehuchte, Lemière et al. 2009). Another invertebrate, *Drosophila melanogaster*, long thought to be devoid of methylation, has since been shown to have a DNA methyltransferase-like gene (Hung, Karthikeyan et al. 1999) and is reported to contain very low MeC levels (Lyko, Ramsahoye et al. 2000), although mostly in the CpT dinucleotide rather than in CpG.

With the exception of *Drosophila melanogaster* and other insects, most other eukaryotic genomes have moderately high levels of methyl-CpG concentrated in large domains of methylated DNA separated by equivalent domains of unmethylated DNA (Colot and Rossignol 1999; Klose and Bird 2006). This mosaic methylation pattern has been confirmed at higher resolution in the sea squirt, *Ciona intestinalis* (Simmen, Leitgeb et al. 1999). In vertebrate genomes, which have the highest levels of MeC found in the animal kingdom, methylation is dispersed over much of the genome, a pattern referred to as global methylation (Klose and Bird 2006).

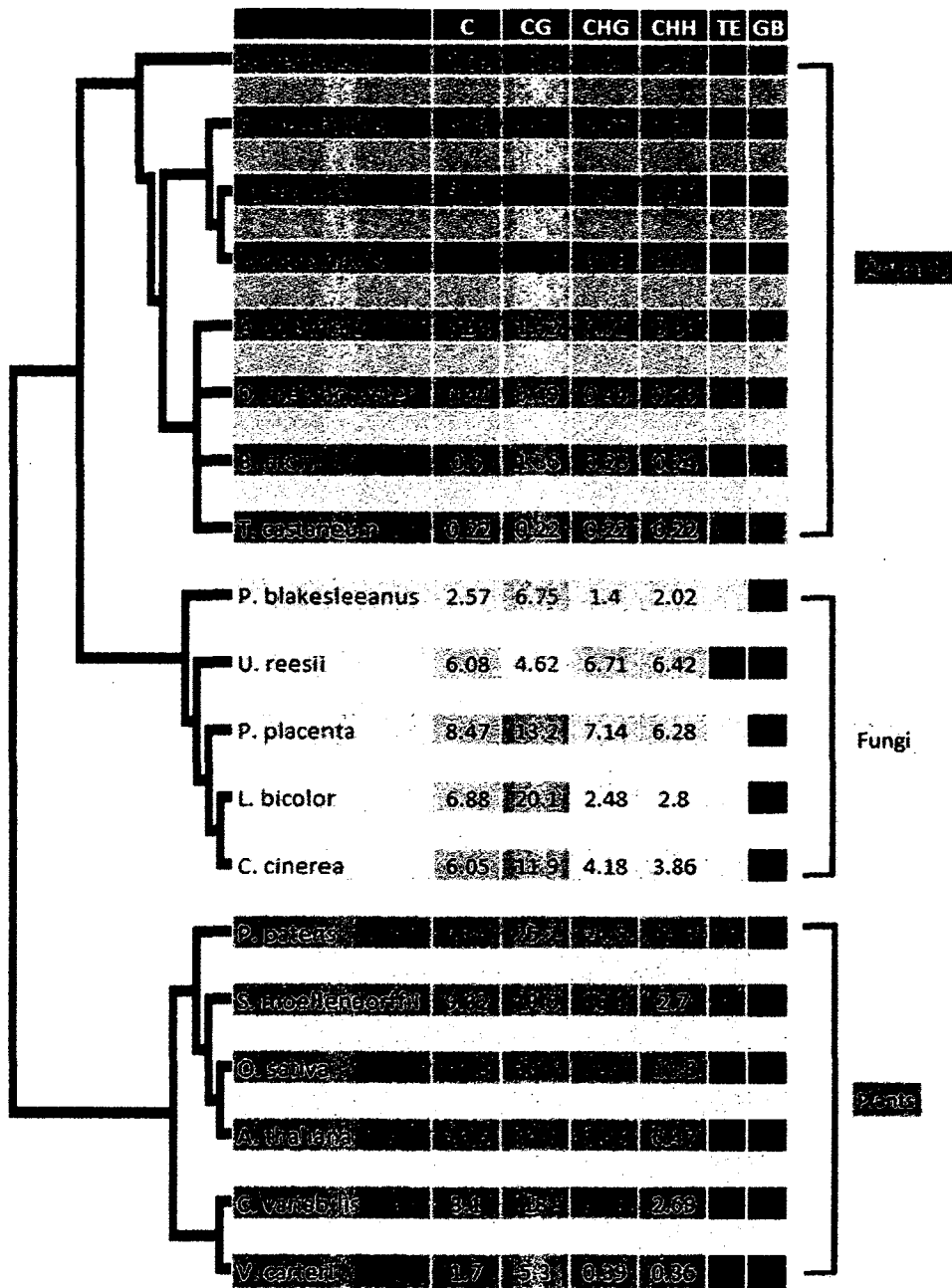


Fig. 1 Normalized frequencies of DNA methylation context and evolution

The phylogenetic tree was based on the NCBI Taxonomy Browser. The values represent the normalized fraction in percent of methylated Cs per motif. The filled boxes on the right indicate high methylation of gene bodies (GB) and transposable elements (TE). Data obtained from; Lister et al., 2009, Zemack et al., 2010, Su et al., 2011 and this study.

Even though methylation levels and contexts (CpG or non CpG) differ from organism to organism, methylation of transposons and gene bodies are common across eukaryotes (Zemach, McDaniel et al. 2010). The conservation of the specific methylation of transposons(TE) and gene bodies (GB) suggest that this is a basal pattern and likely crucial in the evolutionary process of eukaryotic genomes and puts forth the hypothesis that a complex transcriptome requires DNA methylation to suppress transcription errors and stabilize the genome. Therefore, it is critical to conduct a rigorous analysis of methylation in the *C. elegans* genome since it is an excellent model to study the basal mode of DNA methylation.

CHAPTER I

IDENTIFICATION OF DNA METHYLTRANSFERASES IN *C. ELEGANS*

Background

DNA methyltransferases are distinguished as either maintenance (DNMT1 family) or *de novo* methyltransferases (DNMT3 family), depending on their preference for hemimethylated or unmethylated DNA, respectively (Bestor 2000). DNA methyltransferase 1 (DNMT1), for example, is generally considered to maintain DNA methylation patterns associated with DNA replication (Leonhardt, Page et al. 1992), and it has a stronger preference for hemimethylated DNA; however, this may not always be the case since DNMT1 also acts on unmethylated targets (Okano, Xie et al. 1998). Additional evidence for *de novo* activity of DNMT1 in chromatin at sites of homologous recombination has recently been proposed (Cuozzo, Porcellini et al. 2007). The DNA methyltransferase associated protein (DMAP) is a co-repressor that forms a complex with DNMT1 and targets replication foci in the S phase of Vero (primate) cells (Rountree, Bachman et al. 2000). In Human cells, DMAP1 participates in the epigenetic reprogramming that was previously shown to be involved in homology-directed DNA repair (Cuozzo, Porcellini et al. 2007). This means that DMAP1 acts as a co-repressor in global maintenance (Rountree, Bachman et al. 2000), as well as cooperating with DNMT1 in epigenetic alterations associated with repair of DS DNA breaks. It has been

shown that DMAP1 has a strong binding preference for hemimethylated DNA and stimulates DNA methylation mediated by DNMT1 in maintenance methylation as well as *de novo* methylation activity *in vitro* (Lee, Fitzpatrick et al. 2001).

DNMT3L is another gene that shares homology with DNMT3 family methyltransferase genes. DNMT3L is required for the establishment of methylation imprints in mammalian oocytes (Hata, Okano et al. 2002). DNMT3L, which by itself has no detectable DNA methyltransferase activity, appears to regulate methylation of imprinted genes through its interaction with DNA methyltransferases, DNMT3a and DNMT3b (Hata, Okano et al. 2002). DNMT3L binds and colocalizes with DNMT3a and DNMT3b in the nuclei of mammalian cells. Accordingly, DNMT3L^{-/-} mutants, (DNMT3a^{-/-}, DNMT3b^{+/-}) female mice also fail to establish maternal methylation imprints (Hata, Okano et al. 2002). These results provide genetic evidence that DNMT3 family methyltransferases and a potential cofactor DNMT3L are required for *de novo* methylation of imprinted genes in the female mammalian gamete. Thus, the establishment of a DMAP and DNMT3L homologues in *C. elegans* would provide evidence for the existence of essential parts to the DNA methylation machinery. However, we must also keep in mind that most eukaryotes have only an identifiable DNMT1 homolog and some have no homologues to known DNMTs yet actively methylate their genomes.

Furthermore, it is not enough to include only well-established

mammalian DNA methylation machinery in this search. It is of equal importance to include all classes of DNA methyltransferases, since one cannot predict the mechanism in which *C. elegans* methylates DNA. In plants, *Arabidopsis thaliana* is best studied model for DNA methylation and has at least three classes of DNA methyltransferase genes: the *MET* class, the *CMT* class, and the *DRM* class (Finnegan and Kovac 2000). *MET1*, like its mammalian homolog *Dnmt1* (Bestor, Laudano et al. 1988), encodes the major *Arabidopsis* CpG maintenance methyltransferase (Finnegan, Peacock et al. 1996; Ronemus, Galbiati et al. 1996; Kishimoto, Sakai et al. 2001). When *Met1* was tested in a RNA directed DNA methylation (RdDM) system where a 35S:GFP transgene was methylated and silenced by homologous RNA virus sequences, CpG methylation of the 35S promoter sequence was heritable in the absence of an RNA trigger and was dependent on the activity of *MET1* (Jones, Ratcliff et al. 2001). However, suppression of *MET1* activity did not block the establishment of RNA-directed CpG methylation in this system. These results suggest that *MET1* is important in the maintenance of gene silencing that is caused by RdDM, but probably not in the initiation of RdDM. *CMT*-like genes are specific to the plant kingdom and encode methyltransferase proteins containing a chromodomain (Henikoff and Comai 1998). *Arabidopsis CMT3* loss-of-function mutants show a large decrease in CpNpG methylation and more subtle and locus-specific effects on asymmetric methylation (Lindroth, Cao et al. 2001; Cao and Jacobsen 2002). The *DRM* genes share homology with mammalian *DNMT3* genes that encode de novo methyltransferases (Cao, Springer et al. 2000). Previous work showed

that a double mutant of *drm1* and *drm2* showed a lack of de novo DNA methylation normally associated with transgene silencing of the *FWA* and *SUPERMAN* genes (Cao and Jacobsen 2002). It was also observed that *drm1 drm2* double-mutant plants show major losses of asymmetric methylation and more subtle and locus-specific effects on CpNpG methylation at endogenous *Arabidopsis* loci (Cao and Jacobsen 2002).

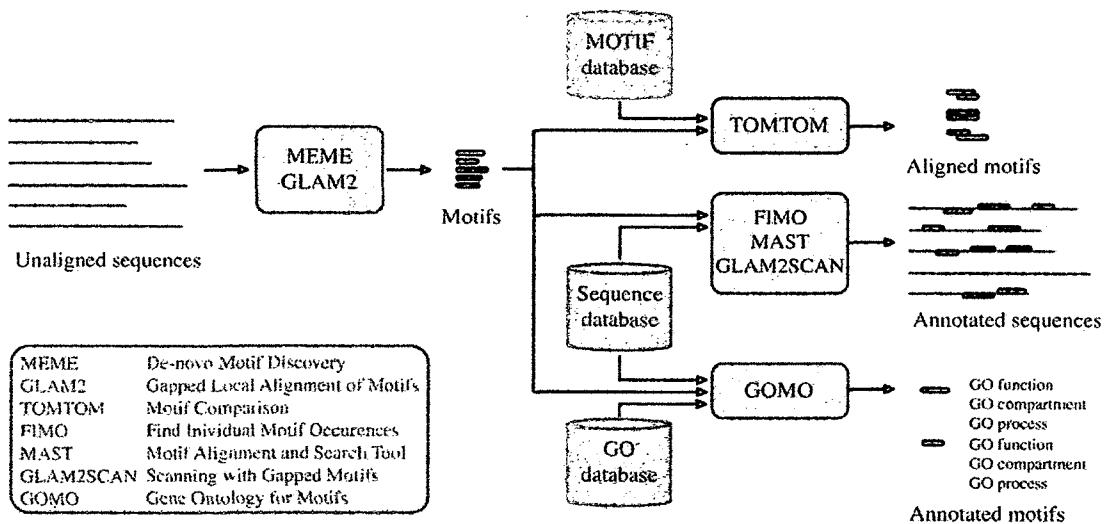
Neither *drm* nor *cmt3* mutants affect the maintenance of pre-established RNA-directed CpG methylation. However, when *drm* is mutated there is a nearly complete loss of asymmetric methylation and a partial loss of CpNpG methylation (Cao, Aufsatz et al. 2003). The remaining asymmetric and CpNpG methylation was dependent on the activity of *CMT3*, showing that *DRM* and *CMT3* act redundantly to maintain non-CpG methylation (Cao, Aufsatz et al. 2003). It was shown that these DNA methyltransferases appear to act downstream of siRNAs, since *drm1 drm2 cmt3* triple mutants show a lack of non-CpG methylation but elevated levels of siRNAs demonstrating that *DRM* activity is required for the initial establishment of RdDM in all sequence contexts including CpG, CpNpG, and asymmetric sites (Cao, Aufsatz et al. 2003).

Previous evaluations of the *C. elegans* genome have either not recognized any DNA methyltransferase homologs (Gutierrez and Sommer 2004; Zemach, McDaniel et al. 2010) or only a single gene homologous to the Dnmt1 family (Gao et al., 2012). To characterize the capacity for the *C. elegans* genome to code for methyltransferase we

searched the genome with homologs of all major DNMT, and DNMT related protein sequences.

Methods

Known DNMT sequences from *Arabidopsis*, *Neurospora*, mouse, and human were queried against the *C. elegans* data base (www.wormbase.org) using BLAST/BLAT on default settings and an e-value cut off of 0.01. Orthology was inferred by reciprocal best BLAST (RBB) (Li, Stoeckert et al. 2003). The motif search and discovery results were gathered using MEME and MAST, a part of a software toolkit that allows for motif discovery and motif database searching (Bailey, Boden et al. 2009).



(Bailey, Boden et al. 2009)

Fig. 2 Overview of the MEME Suite

MEME and GLAM2 are tools for motif discovery, Tomtom searches for similar motifs in databases of known motifs, FIMO, GLAM2SCAN and MAST search for occurrences of motifs in sequence databases, and GOMO provides associations between motifs and GO terms. The components of the MEME Suite are implemented in ANSI C as command line tools. These are published as SOAP (Simple Object Access Protocol) web services using Opal and the Tomcat Java servlet container. Opal provides job management services allowing the MEME Suite to queue multiple simultaneous requests (Bailey, Boden et al. 2009).

Results and Discussion

Initial searches for putative *C. elegans* DNA methyltransferases (DNMTs) yielded a number of prospects. Table 1 represents a list of putative DNMT homologues that was paired down from a larger list by functional inferences based on automatic annotations by InterPro (<http://www.ebi.ac.uk/interpro/>). The potential *C. elegans* DNMTs were then BLASTed against the non-redundant protein database to find annotated sequences that matched the proposed function (e. g. methyltransferase) of the *C. elegans* sequences. Once the annotated sequence was matched with an e-value cut off of at least 0.01, the organism/locus that corresponded to the annotated sequence was subsequently used for RBB analysis.

Although the prospective genes could all potentially be involved in DNA methylation only three (Uniprot # Q81AA7 ; P45968 ; Q9U1S4) were chosen based on their predicted catalytic domains and functions in addition to evidence based on RBB analysis. All putative *C. elegans* DNA methyltransferase and DNA methyltransferase associated genes have transcript evidence confirmed via microarray expression data and matching cDNAs. Furthermore, a recent publication (Gao, Liu et al. 2012) also identified one of the putative DNMTs (Uniprot # P45968 Wormbase ID Y75B8A.6) as a DNMT1 homologue.

WB ID	Uniprot ID	RBB	E-Value	Phenotype	Annotation	RBB Organism
C33C12.9	O16582	RBB Confirmed	6.0E-30	NA	Adenine tranferase	Loa loa
C38D4.9	Q18511	RBB Confirmed	3.0E-15	NA	DNMT-like	Crassostrea
Y43H11AL.1	Q8IAA7	No RBB	NA	NA	C5 methyltransferase	Human
Y62E10A.5	Q2HQK2	No RBB	NA	NA	alkyltransferase	Human
Y71F9AL.1	Q9N4H1	No RBB	NA	NA	RNA methyltransferase	Human
Y105E8A.17	Q8WQA7	RBB Confirmed	5.0E-75	embrionic lethal	DMAP	Human
Y75B8A.6	Q9U1S4	RBB Confirmed	4.8E-08	NA	DNMT1	Mouse
T09A5.8	P45968	RBB Confirmed	6.5E-06	extended life emb lethal	C5 methyltransferase	Arabadosis

Table 1. Candidate DNA Methyltransferases

Columns from left to right: The Wormbase ID of potential DNMTs, the corresponding Uniprot ID, Wether reciprocal best blast was confirmed or not, phenotype associated with the gene knocked out, automated annotation for the non-RBB confirmed or annotation based on orthology for the RBB confirmed, and finally the organism with which RBB was performed.

Since DMAP1 has a strong binding preference for hemimethylated DNA and stimulates DNA methylation mediated by DNMT1, a search for orthologous DMAP sequences in *C. elegans* was also performed. This search yielded two genes (Uniprot # Q8WQ87; A8QE0) with high similarity to mammalian DMAP1 (Table1).

A closer look into the motifs of two of the most conserved putative *C. elegans* DNMTs (Uniprot # Q81AA7 ; P45968) revealed that the motifs required for an active methyltransferase are present. Aside from the domain (seen in red Fig.3 and Fig.4) that was used initially to implicate this protein as a putative DNA methyltransferase, additional motifs have been found by comparing the individual motifs from other known DNA methyltransferases and related proteins using the MEME toolkit (Bailey, Boden et al. 2009). Of the motifs and domains found in the *C. elegans* DNMT homologues we find

sequence conservation as well as conservation in organization. However, not all motifs and domains found in some known DNMTs are found in the *C. elegans* homologue. In fact, none of the ten motifs said to be required for a functional mammalian DNMT (Goll and Bestor 2005) are present in their entirety in *C. elegans*. However, both mouse and *Arabidopsis* DNMTs do not have all ten motifs in any of the DNMT families as well. For instance, the mouse DNMT3 has only two of the ten and six of the ten in DNMT1.

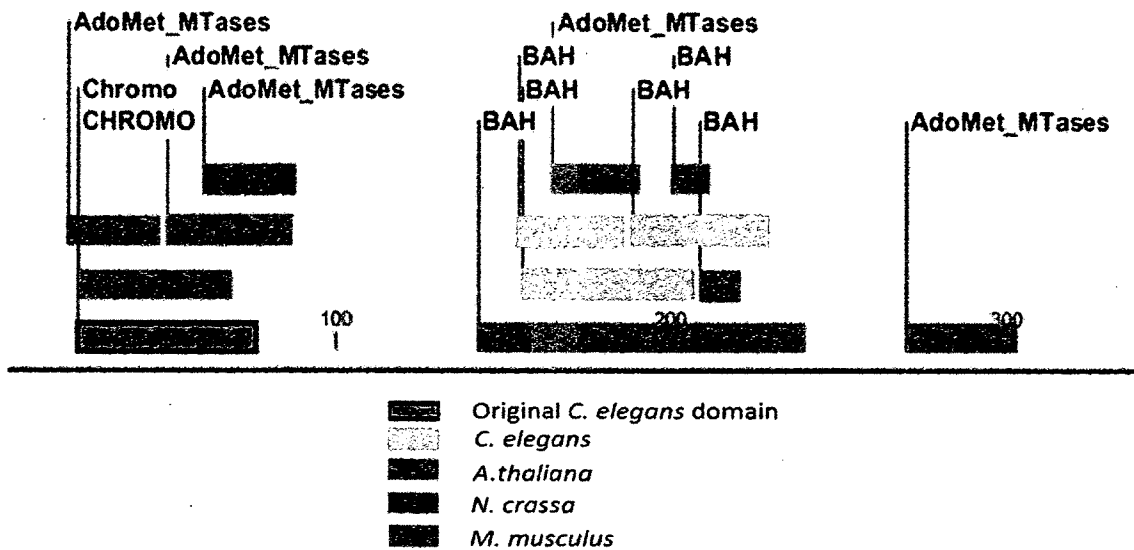


Fig. 3 Worm Base ID Y75B8A.6 Organization of Motifs

The black line above the legend represents the putative *C. elegans* DNMT sequence. The colored boxes above the sequence represent homologous domains and motifs from other organisms labeled below the sequence. The *C. elegans* domain seen in red that was used initially to implicate this protein as a putative DNA methyltransferase. The additional motifs have been found in other known DNA methyltransferases and related proteins. CHROMO is a chromatin organization modifier domain. BAH is the bromo-adjacent homology domain. The AdoMet_MTases are catalytic domains which allow for the S-adenosylmethionine-dependent methyltransferases to interact with DNA.

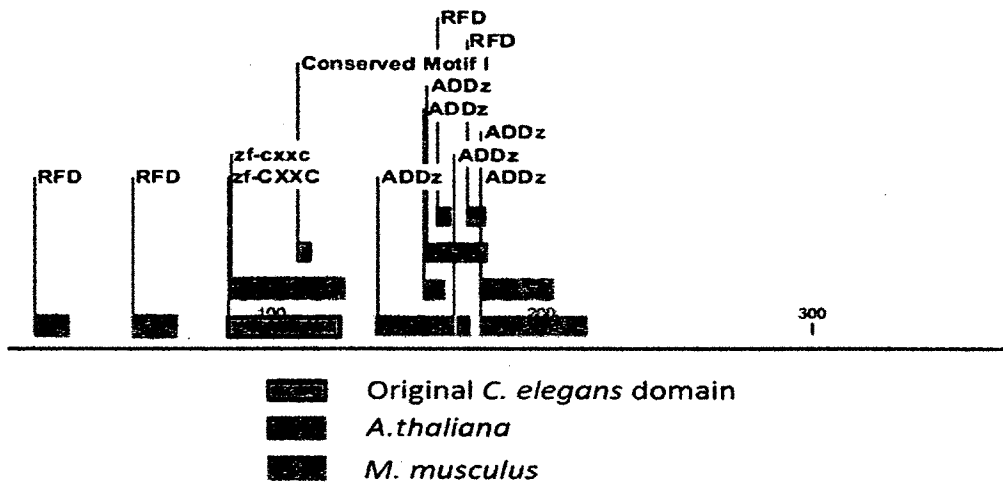


Fig. 4 Worm Base ID T09A5.8 Organization of Motifs

The black line above the legend represents the putative *C. elegans* DNMT sequence. The colored boxes above the sequence represents homologous domains and motifs from other organisms labeled below the sequence. The *C. elegans* domain (seen in red) was used initially to implicate this protein as a putative DNA methyltransferase. The additional motifs have been found in other known DNA methyltransferases and related proteins. zf-CXXC is a zinc-finger motif. RFD is DNA replication foci-targeting sequence. ADDz is involved in protein/chromatin interactions. Conserved Motif I is involved in transfer of methyl from S-adenosylmethionine to cystine.

Upon further investigation conserved DNMT motifs of known functioning DNMTs it was interesting to note that not all of the conserved motifs said to be required for functionality (Goll and Bestor 2005) are actually present in the published DNMT sequences. With this in mind we sought to find a “universal” motif to compare the putative *C. elegans* DNMTs. By using BLAST to find known DNMTs that share high similarity and extracting the sequences we were able to use the MEME toolkit to search

for conserved motifs. The results were then used in the MAST algorithm to identify the motifs in the putative *C. elegans* DNMTs. Once the motifs were found in the *C. elegans* DNMTs the motifs were BLASTed against the Uniprot database to find functional assignments. This resulted in *C. elegans* DNMTs sharing highly similar motifs involved in DNA binding and catalytic activity with other cytosine-specific methyltransferases, specifically DNMT1 (Table.1).

Conclusion

Based on our analysis it appears that *C. elegans* encodes at least three putative DNMTs and one DMAP. It is important to note that these putative methyltransferases were already electronically annotated as DNA methyltransferases in the *C. elegans* database. While our evidence of homology to known DNMTs is strong it remains to be shown that *C. elegans* actively methylates its genome (Chapter 2). Furthermore, the function of these putative DNMT and other loci involved remains to be elucidated by classical functional assays.

CHAPTER – 2

PATTERN OF C5 DNA METHYLATION IN *C. ELEGANS*:

Levels, Motifs and Patterns In an Enriched Genome Dataset

Background

C. elegans is a premier model organism in biology and the first metazoan to have its genome completely sequenced. However, post-synthesis modification of *C. elegans* DNA remains virtually unstudied. Although, *C. elegans* is a relatively simple organism, it shares many essential biological processes and pathways with other multicellular organisms with high genome content and complexity. Therefore *C. elegans* could be an important an important model for the study of DNA methylation and its role in genome evolution.

Recent large-scale comparative analysis of cytosine DNA methylation across diverse eukaryotes suggest that early features of DNA methylation included the methylation of gene bodies and transposable elements (Zemach, McDaniel et al. 2010). These potentially ancient patterns may reflect a primitive role of methylation in transcriptional fidelity and as a mechanism to protect the germ line from transposon (or repeat) mediated mutation. Because spurious transcription and mutation are hypothesized to be among the critical limiting factors to genome size, an ancient role for methylation in support of fidelity of transcription and genome stability suggests a

possible link with the origin of eukaryotes (Bird and Tweedie 1995; Maunakea, Nagarajan et al. 2010).

In light of these observations it is perplexing that one of our key model eukaryotes, the nematode (*Caenorhabditis elegans*) is assumed to lack active DNA methylation. In fact, *C. elegans* is often invoked to suggest the dispensability of methylation in multicellular animals (Feng, Cokus et al. 2010; Zemach, McDaniel et al. 2010). Historically, this view has been based on crude, yet standard, genomic assays using methylation sensitive restriction enzymes (Simpson, Johnson et al. 1986).

However, while it is clear that the genome of *C. elegans* is not highly methylated, methylation maybe limited to a small subset of nuclei (e.g. the germline) as might be expected based on some proposed ancestral functions. It is also possible that some patterns of methylation will be developmentally regulated and perhaps limited to a subset of cell types and specific developmental stages as has been shown in the parasitic nematode *Trichinella* (Gao, Liu et al. 2012).

It is also enigmatic that the genome of *C. elegans* appears to encode multiple DNA methyltransferases (DNMT; uniprot: Q8IAA7, P45968 and Q9U1S4), and a DNA methyltransferase associated protein (uniprot:Q8WQA7) as discussed in Chapter 1. One of these DNMT genes (P45698) was independently identified in a screen for co-suppressors of germline transgenes in *C. elegans*, suggesting a role in repeat inactivation (Robert, Sijen et al. 2005). More recently, this same gene has been implicated as being a DNMT1 homologue of a parasitic nematode in which DNA methylation has been

confirmed (Gao, Liu et al. 2012). It is also noteworthy that *C. elegans* contains transposable elements that actively transpose in the soma yet are suppressed in the germline (Emmons and Yesner 1984). Based on these observations, a more sensitive and detailed examination of the *C. elegans* DNA methylome is warranted.

In this chapter we explore the existence of cytosine DNA methylation in the genome of *C. elegans*. Methods for analysis of DNA methylation can be divided roughly into two types: global and gene-specific. For global methylation analysis, there are methods which measure the overall level of MeCs in genomes such as chromatographic methods and methyl accepting capacity assay (Selker, Tountas et al. 2003). For gene-specific methylation analysis, a number of techniques have been developed. Earlier studies used methylation sensitive restriction enzymes to digest DNA. The digest is followed by Southern hybridization based detection or PCR amplification (Rollins, Haghghi et al. 2006). Recently, bisulfite reaction based methods, such as methylation specific PCR (MSP) or bisulfite genomic sequencing PCR have become popular (Rakyan, Hildmann et al. 2004). For this study, methylation in genome wide or global terms will be the focus. Furthermore, because of the known paucity of methylation in *C. elegans*, for this initial analysis we have employed an enrichment step to focus our sequencing efforts on the DNA sequences containing 5-methyl Cytosines (MeC).

The core method used to detect MeC in our assay is bisulfite treatment. Treatment of DNA with bisulfite converts cytosine residues to uracil, which are read as

thymine residues in the sequencing process. MeC residues, however, are unaffected by bisulfite. Current bisulfite treatment protocols have become incredibly robust with conversion rates greater than 99.9% and inappropriate conversion (conversion of MeC to U) rates less than 0.78% (Genereux, Johnson et al. 2008). Therefore, bisulfite treatment introduces specific changes in the DNA sequence that depend on the methylation status of each cytosine residue, at high accuracy and low error rates yielding single-nucleotide resolution information about the methylation status of a segment of DNA (Rakyan, Hildmann et al. 2004).

Methods

C. elegans (N2) were grown under normal conditions (Brenner 1974) and DNA was extracted from mixed stage worms using the Qiagen genomic tip protocol. Given the fact that we anticipate if *C. elegans* actively methylates its genome that the levels will be low, we chose to conduct this first analysis using an enrichment step where DNA fragments containing MeC are enriched in the sample using a MeC binding protein attached to a substrate. To enrich the sample for methylated strands we fragmented the DNA using the Gene Machine Hydro shear to ~ 500bps and used the Invitrogen MethylMiner DNA enrichment Kit, a methylation binding enzyme attached to magnetic beads to pull down fragments containing methylated C(s). This DNA was subjected to bisulfite treatment using the Invitrogen MethylCode Bisulfite Conversion Kit, which converts non-methylated Cs to T. This MeC enriched and bisulfite treated DNA sample was then sequenced using Illumina Sequencing technology. This method relies on the attachment of randomly fragmented, adapter ligated, genomic DNA to a planar, optically transparent surface. Attached DNA fragments are extended and bridge amplified to create an ultra-high density sequencing flow cell with hundreds of millions of clusters, each containing ~1,000 copies of the same template (Quail, Kozarewa et al. 2008). These templates are sequenced using a four-color DNA sequencing-by-synthesis technology that employs reversible terminators with removable fluorescent dyes. Together with the Illumina data analysis pipeline, this sequencing technology achieves an error rate of less than 0.9% (Quail, Kozarewa et al. 2008).

The resulting sequence data is then analyzed by aligning the bisulfite treated sample sequence to the current published reference genome. From this reference genome, two "*in silico* bisulfite treated" references must be prepared. First is the reference genome with all cytosines changed to thymines and second is a reference with all the guanines changed to adenines to account for the complimentary strand. The C-T, T-C, G-A, and A-G transitions can then be examined to elucidate potential methylated sites (Pomraning, Smith et al. 2009). Mapping high-throughput bisulfite reads to the reference genome is a challenge due to reduced complexity of bisulfite sequence, and asymmetric cytosine to thymine alignments (Xi and Li 2009). BSMAP is based on the open source software SOAP (Short Oligonucleotide Alignment Program) (Li, Li et al. 2008). This analysis results in a report for every C on either strand. This report gives the chromosome, position, the number of times a read mapped to that position and the number of times that read had a C that was not converted to a T. BSMAP parameters were set to a fragment size of 100-280 bps, 8 processors were used, seed size was set to 14, 4 mismatches were allowed in the alignment, and the max number of equal best hits to count was set to 10.

Results and Discussion

Cytosine methylation levels in *C. elegans* - Previous studies have failed to produce any evidence of methylated DNA in *C. elegans*. This may stem from a level of methylation that is too low to detect using HPLC and methylation sensitive restriction enzyme analysis (Simpson, Johnson et al. 1986). To improve our detection ability we set out to enrich the DNA for methylated DNA. We started with fragmented genomic DNA at a total weight of 97.6 ug 98.8 ug and 98.2 ug. The resulting yield after three rounds of methylated DNA enrichment was 1.46 ug, 1.52 ug, and 1.56 ug respectively. The average yield of methylated DNA was a 1.5% most of which is likely unmethylated. This low level of MeC containing DNA is consistent with the lack of detectability in previous attempts to elucidate DNA methylation in *C. elegans* as > 1% DNA methylation would likely be undetectable by differential restriction enzyme analysis. This DNA was then pooled and treated with bisulfite. Illumina libraries (Paired-End 76 base pair) were prepared and sequenced at Vanderbilt University and at Expression Analysis.

Out of the original bisulfite treated reads (74,285,412), a total of 70,004,089 reads were not mapped due to either low quality, being unpaired, or having no match due to the decreased complexity of bisulfite treated DNA. The resulting 4,281,323 mapped reads were included in the analysis with an average read coverage across the genome of 3.81 and the fraction of the reference covered was 0.81. As shown below the successfully mapped reads are not randomly distributed across the genome.

Based on the predicted rate of bisulfite non-conversion (>1%) that would result in a C remaining in a bisulfite treated read and the sequencing error rate of T to C (>0.1%) that would change a converted MeC back to a C we filtered all data to focus only on position in the genome that were covered by at least 3 bisulfite reads that had a putative MeC at a specific position. When filtered for 3 or more methylated C confirming reads, 160,988 putative methylated sites remained. This is about 0.5% of the Cs in the genome. At these 160,998 sites, the average coverage was 31.27.

Methylation patterns and motifs in *C. elegans* - It has been shown that C residues at CpG dinucleotides are the preferred targets for DNA methylation in mammals, while methylation at CG and CHG and asymmetric sites CHH are common in plants fungi and insects (Gruenbaum, Naveh-Many et al. 1981) (Cao, Aufsatz et al. 2003). To evaluate the distribution of putative MeCs among these motifs in *C. elegans* we counted the contexts for each of the 16 triplets beginning with C. In *C. elegans* we find a bias (63%) toward methylation of non-symmetric (CHH) sites where H is any base. However, as can be seen in Figure 5 when normalized for the abundance of each triplet in the genome, context methylation seems to be randomly distributed.

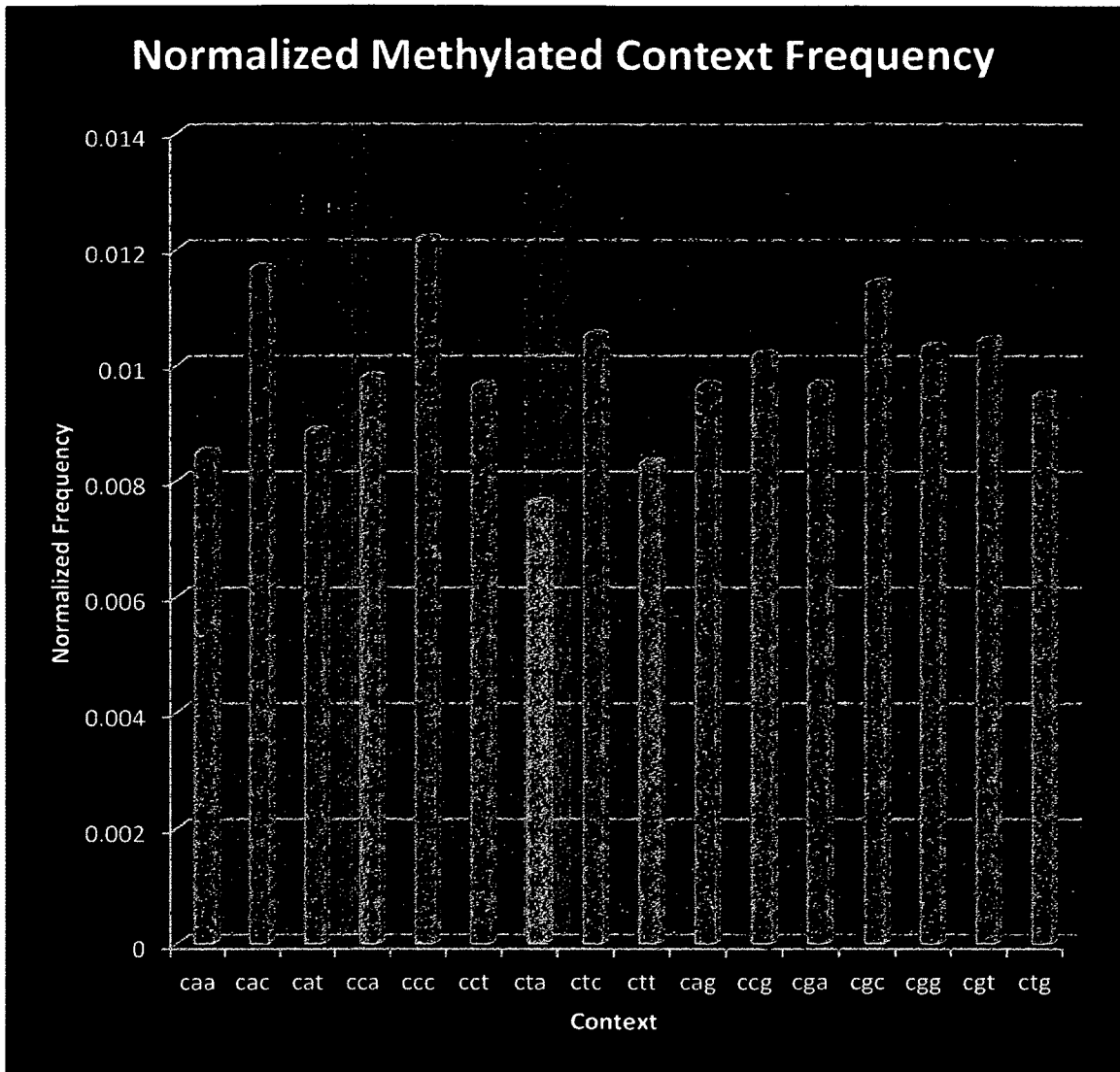


Fig. 5 Asymmetric vs. Symmetric Methylation

Each putative MeC containing site was assigned to one of 16 possible triplets beginning with MeC. In this figure the frequency of each sequence context is shown where the first C meets the criteria as a MeC site. Each triplet motif of MeC was normalized by the number of each triplet occurring in the reference genome.

The distribution of methylation in *C. elegans* – To examine the genome wide distribution of methylated Cs we first determined the number of MeCs for each chromosome. The distribution of DNA methylation per chromosome is significantly different at the 95% confidence levels when normalized against the total number of Gs and Cs per chromosome (Fig.6). The highest level of DNA methylation was found in Chromosome I and the lowest in the X chromosome.

To further investigate the intra-chromosomal spatial pattern of methylation, the putative MeC containing positions were divided into 1 Mb bins along each chromosome by position and the frequency was plotted on the same scale to show the relative levels of methylation across each chromosome (Fig 7). Multiple regions were found to have extremely high frequencies of MeC. As discuss further below, the high density at the end of chromosome I is an artifact and the result of the highly methylated rRNA genes. The high density region in chromosome V appeared to be a consequence of very high coding density of known protein coding genes.. One overall pattern is that the core regions of the autosomes appear to have higher levels of MeC than the arms a pattern not found in the X chromosome. This pattern is correlated with several biological patterns including lower rate or recombination and higher gene densities in the cores regions of autosomes than in the arms (Cutter, Dey et al. 2009; Rockman and Kruglyak 2009).

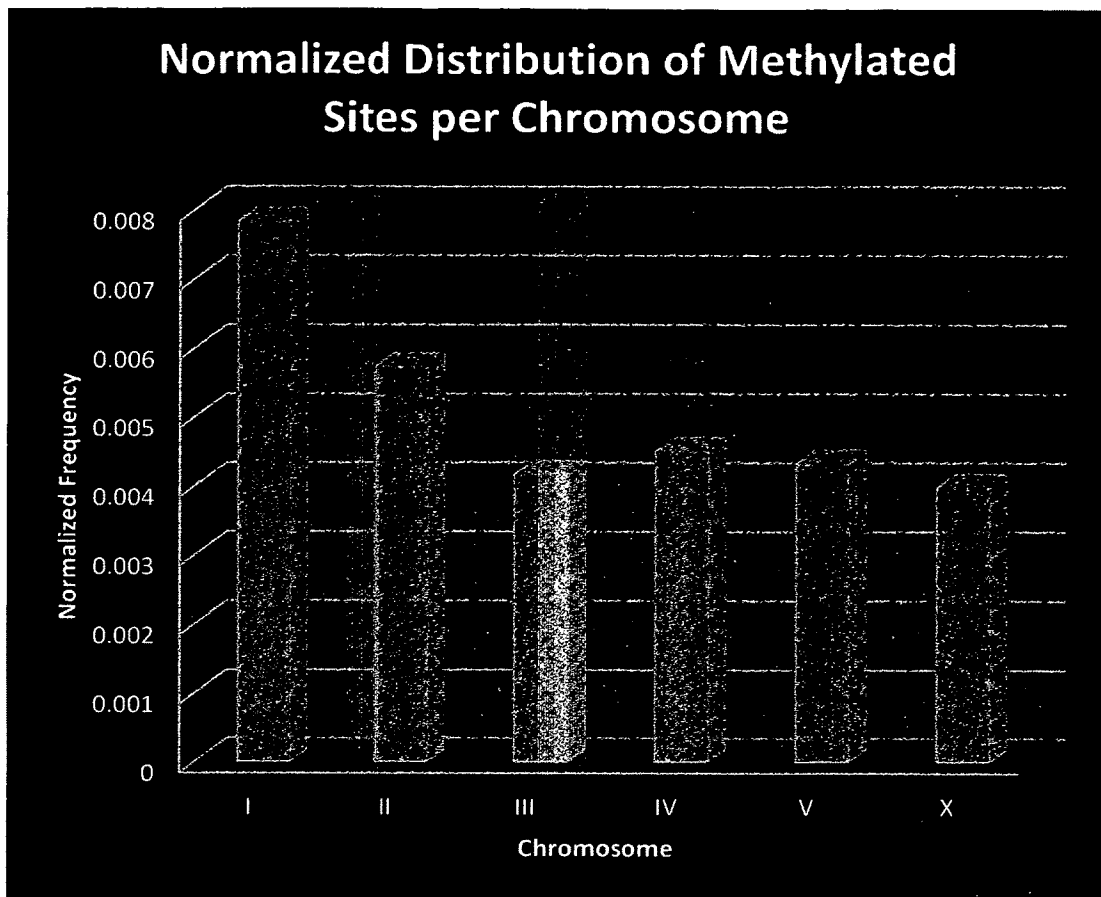


Fig. 6 Normalized Levels of DNA Methylation per Chromosome

Frequencies of MeCs at each position were normalized by correcting for every C/G in each chromosome by dividing the total number of methylated C per chromosome by the total number of C/G occurrences in the reference per chromosome. A 6-sample test for equality of proportions between chromosomes shows that the normalized frequencies are significantly different (p -value $< 2.2e-16$).

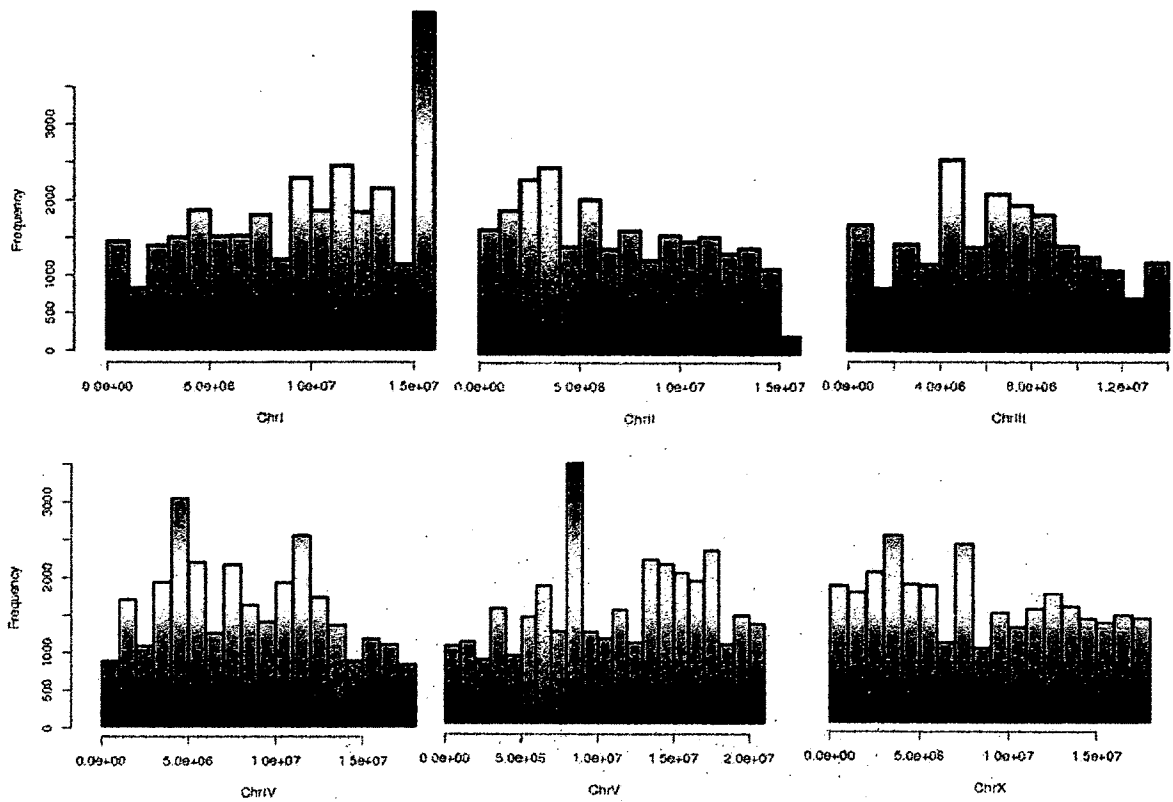


Fig. 7 Distribution of Methylated Cytosines per Chromosome

Methylcytosine counts (MeC sites) were divided into 1Mb bins along each chromosome by position and the frequency was plotted on the same scale to show the relative levels of methylation across each chromosome. The spike at the end of Chromosome I is due to an artifact. The ~55 copies of tandemly repeated ribosomal DNA is not included in the reference, only one repeat is annotated at the end of Chromosome I.

To investigate the distribution of MeC with respect to coding vs non-coding we first assigned each putative methylated position a category based on functional annotation of the genome. Figure 8 clearly shows that when we divide the positions into two categories, genic and intergenic, the proportion of genic are vastly overrepresented (Fig. 8). Based on an analysis of the distribution of MeCs across diverse coding sequence functions we conclude that several functional categories appear to be actively methylated based on their overrepresentation. Most notably the two categories with the highest density of methylation are the transposable elements (TE) and small nuclear RNAs (snRNAs) followed by pseudogenes. In addition, the transcribed regions (gene bodies) of protein coding genes and the transcribed regions of the ribosomal RNA encoding repeat also show significantly greater density of methylation than intergenic regions. Together these observations suggest that while the number of methylated DNA molecules in *C. elegans* may represent only a fraction of the nematode genomes the pattern of methylation is strikingly similar to that expected for all eukaryotes including the details of methylation within protein coding genes which follows the exact same pattern observed in other eukaryotes with the highest density of methylated sites in the exons followed by introns and much reduced methylation in the at the gene termini (Fig. 8).

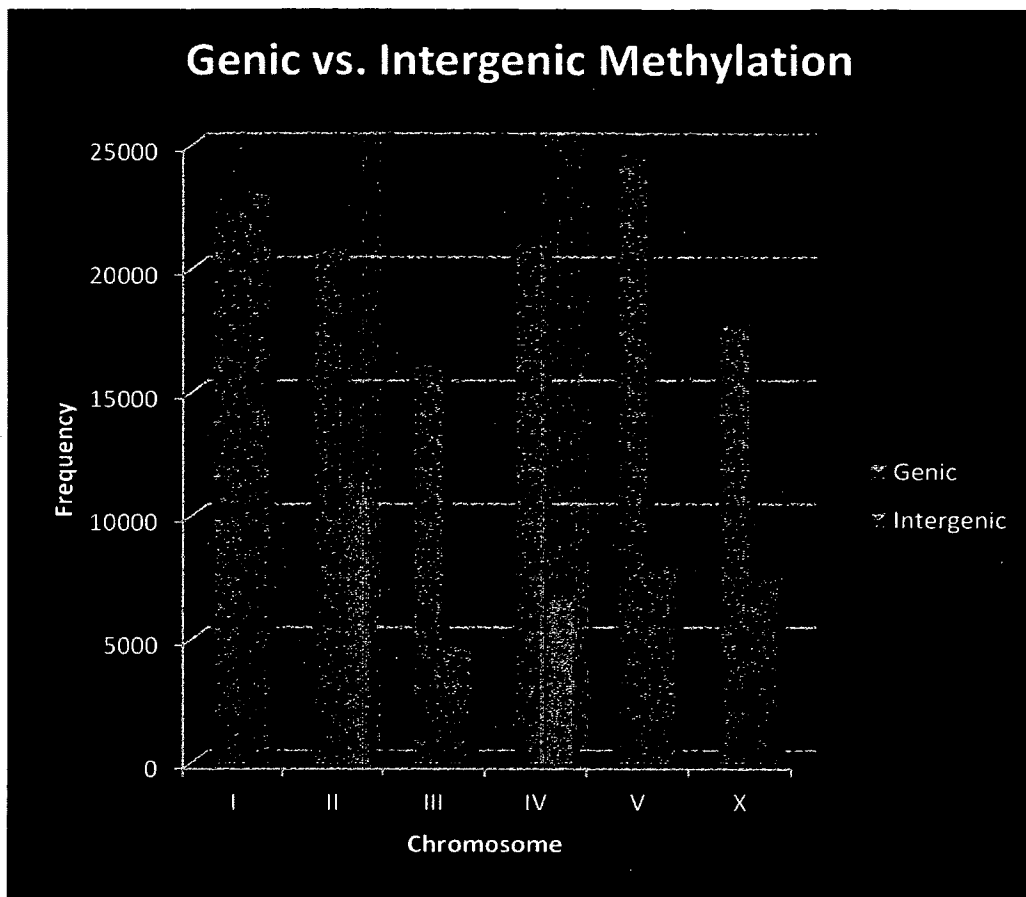


Fig. 8 Genic vs. Intergenic Methylation per Chromosome

To assign each position to category of genic or intergenic we used *the C. elegans* reference WS187 and the corresponding WS187 GFF files for alignment and categorization. All annotated coding genes positions regardless of function were considered genic, regions not assigned a functional coding annotation were inferred to be intergenic.

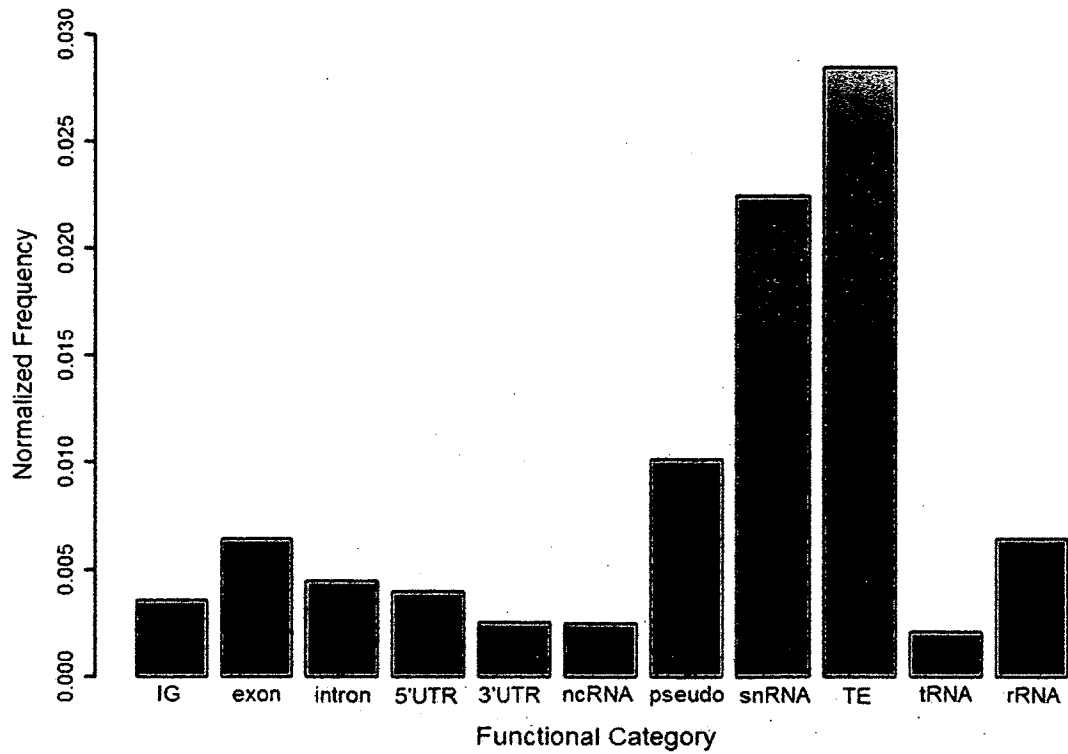


Fig. 9 Categorical Distribution of Methylation

C. elegans reference sequences were used for alignment and categorization of all putative methylated sites. Data were normalized based on all Cs in the *C. elegans* genome. Categories included; Intergenic (IG), exons(coding), 5'UTR s, introns, non-coding RNAs (ncRNA), pseudogenes (pseudo), small nuclear RNAs (snRNA), 3'UTRs, transposable elements (TE), transfer RNAs (tRNA) and ribosomal RNAs (rRNA). A two-sample test for equality of proportions revealed that the fraction of 5-methylcytosines in each category were significantly different from intergenic levels at $\alpha=0.045$ after Bonferroni correction. P-values: exon $<2.2e-16$, 5'utr=0.0001171, intron $<2.2e-16$, ncRNA=9.585e-05, pseudo $<2.2e-16$, snRNA $<2.2e-16$, 3'UTR $<2.2e-16$, TE $<2.2e-16$, tRNA=6.538e-05, rRNA $<2.2e-16$.

One of the criteria used to define putative MeC containing positions is that this call must be made by at least 3 independent (C) reads. However, there can be hundreds of reads covering the base call per locus. In fact, as stated above the average coverage of the putative methylated sites is 31.27. When we plotted the number of reads at a site versus the number that contain C (i.e. putative MeC) we observe 2 distinct groups of MeC containing positions (Figure 10). The first group is comprised of cases where the vast majority of reads contain MeC and the second group where the majority of the reads at a site are not MeC. It is not clear from this data whether the second group represents MeC containing positions in a subset of individual worms and/or cell nuclei within individuals. From here forward we will describe the second group as facultative and the first group as constitutive. When we look at this pattern across the chromosomes it is very clear that chromosomes I and III are extremely different in having a large number of positions that appear to show deep coverage at constitutively methylated sites. While the enrichment step precludes us from making definitive estimates of frequency, the comparisons of chromosomes within this datasets appears to be remarkable. It is potentially noteworthy that these two chromosomes (I and III) were previously shown to be uniquely enriched for histone modifications of transcriptionally active chromatin and to have the most highly expressed genes and the fewest genes with low levels of expression(Liu, Lin et al. 2011). If we do a similar plot of MeC reads vs. coverage by functional category (Figure 10) we see that the sites with very high number of reads (>100) that are nearly all MeC are limited to exons.

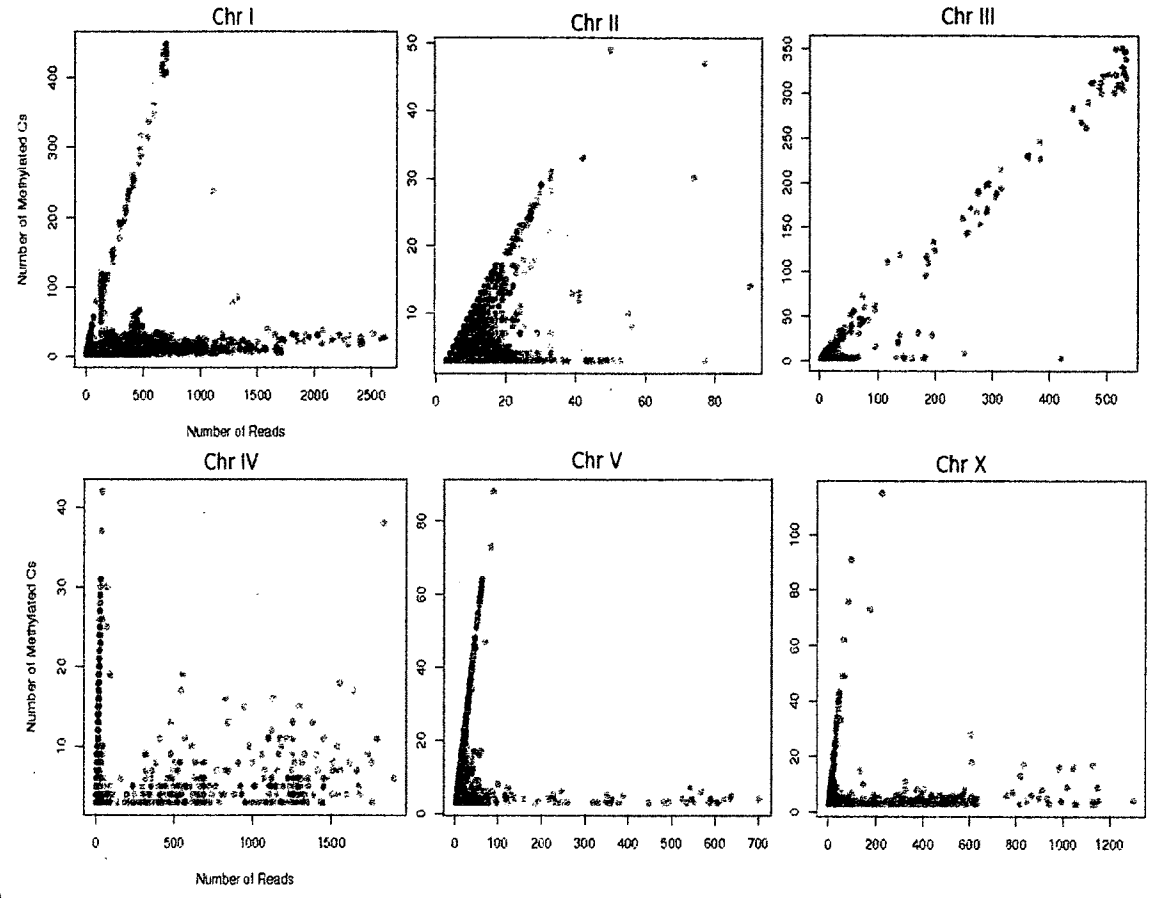
If we quantify the two categories (constitutive and facultative), and greater or less than the 50% of the reads showing MeC respectively we can make broad comparisons across chromosomes and functional categories. A comparison between chromosomes reveals that the DNA methylation observed is dominated (by these definitions) by constitutively methylated sites (Fig.9).

A comparison of facultative and constitutive MeC across functional categories reveals that like the chromosomal comparison the number methylated sites are dominated by constitutive methylation (Fig. 11). However, the rRNA is almost completely facultatively methylated (99.9%) and snRNAs (77%) and tRNAs (51%) also show reduced proportions of constitutive MeC sites. Although these categories only comprise a very small fraction of the total methylated sites, a bias towards facultative methylation in these categories could have strong implications. The observation of a reduced proportion of constitutive sites in chromosome I (Figure 11) is also explained by the fact that Chromosome I encodes ~55 copies of ribosomal DNA which are dominated by facultative MeC patterns.

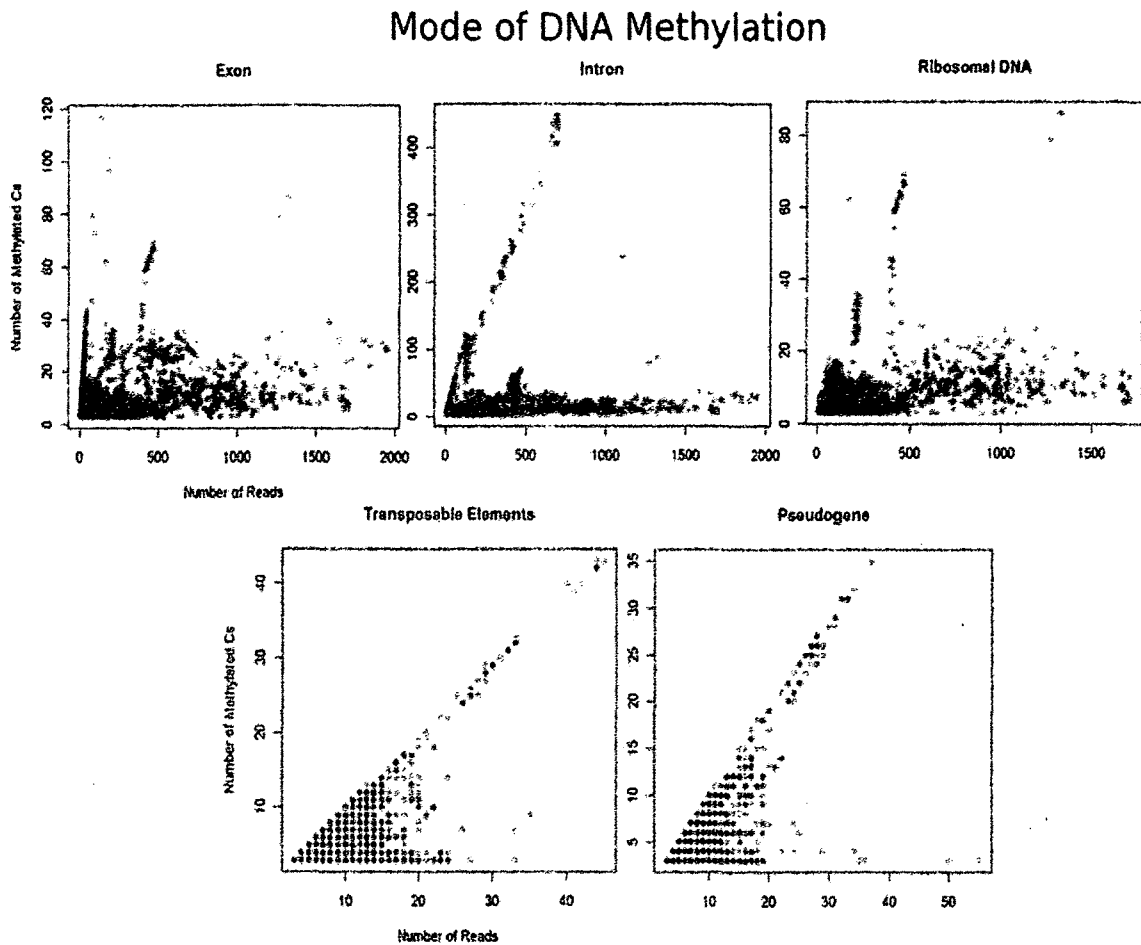
Since rRNA gene transcription accounts for most of the nuclear transcription in an actively growing cell (controlling the pace of ribosome production and subsequent establishment of protein synthesis rates) the role of regulation of this pathway is of great import. It has been shown in *A. thaliana* that rRNA dosage compensation is controlled by DNA methylation (Lawrence, Earley et al. 2004). Related to this rRNA

pathway are the upstream epigenetic switch that is controlled by expression of snRNA, which are required for rRNA maturation (Tycowski, Shu et al. 1994) and the downstream switch, which involves tRNAs. Together, the bias towards facultative methylation of these particular categories of genes is expected and further illustrates the similarity of DNA methylation in *C. elegans* to other organisms.

Mode Of DNA Methylation



A



B

Fig. 10 Constitutive vs. Facultative Methylation per Chromosome per Category

Scatter plot of the density of two distinct groups of methylated Cs. Darker shading represents more cases in that position of the scatterplot. The Y axis scale varies from chromosome to chromosome (A) and among functional categories (B), and is the total number of reads from the same locus that provide evidence for methylation. The X axis is the number of reads covering a specific position.

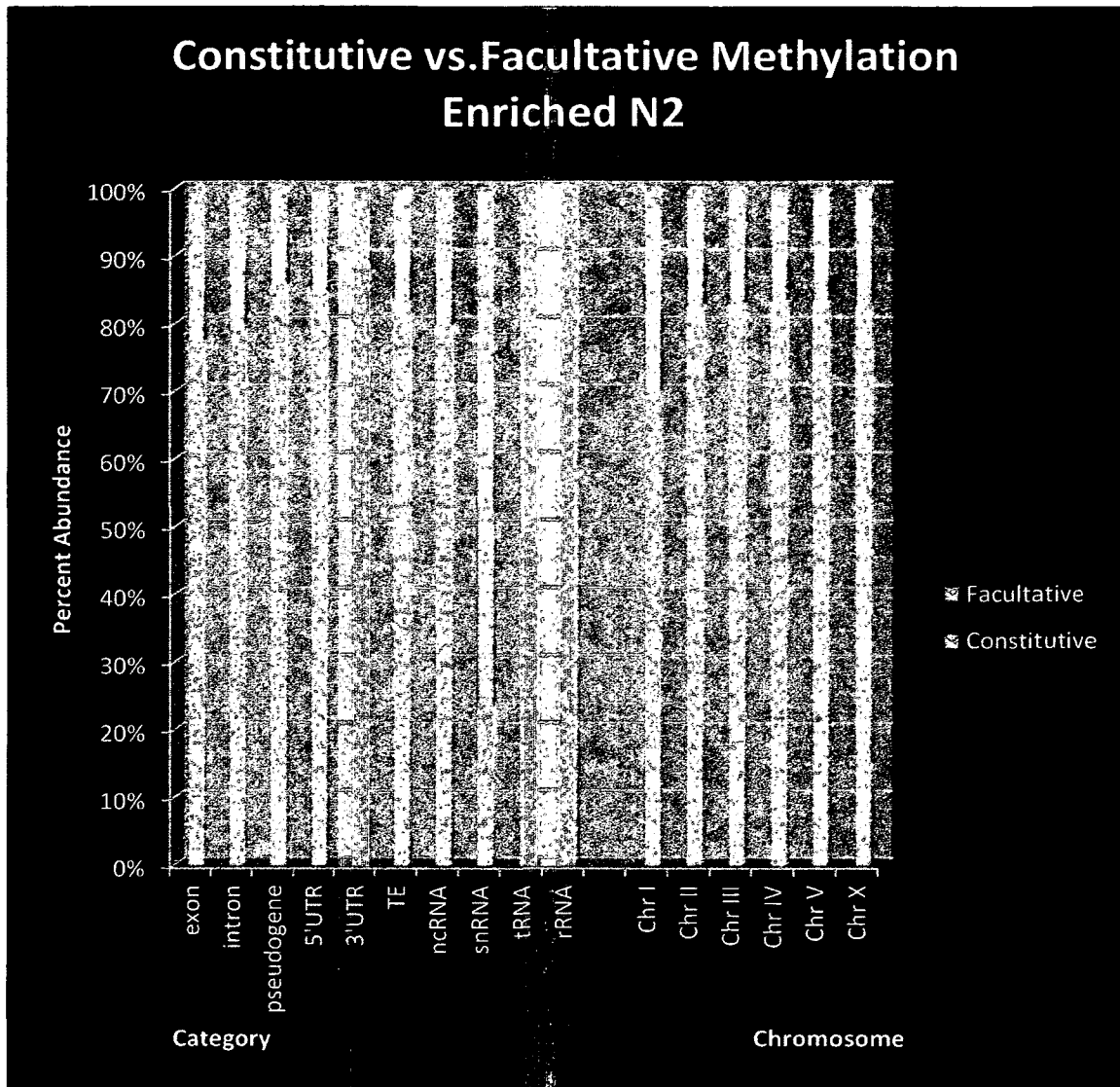


Fig. 11 Constitutive vs. Facultative Methylation per Chromosome per Category

Constitutively methylated Cs were called at a cut off of %50 or more of the total reads for the loci that show evidence for methylation (3 or more putative MeC). Facultative or differentially methylated loci were called at a cut off of %50 at those sites.

Conclusion

While the direct comparison of methylation levels across phylogenetically diverse taxa is complicated by different methods of measurement, *C. elegans* is clearly much less extensively methylated and less biased toward symmetric motifs than canonical methylation systems (humans and *Arabidopsis*). In our analysis of *C. elegans*, only about 0.4% of the Cs are methylated in the enriched population of 5-methylcytosine containing molecules and there is virtually no bias toward symmetrical motifs. This relatively unbiased and low level of methylation is shared by most animals, some plants and some fungi and may in fact be an additional primitive characteristic of eukaryotic methylation (See Figure 1 in Introduction).

Based on their overrepresentation, we conclude that several functional categories appear to be actively methylated. Most notably, the two categories with the highest density of methylation are the transposable elements (TE) and small nuclear RNAs (snRNAs), followed by pseudogenes. In addition, the transcribed regions (gene bodies) of protein-coding genes and the transcribed regions of the ribosomal-RNA encoding repeat also show significantly greater density of methylation than intergenic regions. Together, these observations suggest that while the number of methylated DNA molecules in *C. elegans* may represent only a fraction of the nematode cells, the pattern of methylation is strikingly similar to the basal eukaryotic pattern. Most notably the details of methylation within protein-coding genes follows the same pattern

observed in other eukaryotes with the highest density of methylated sites in the exons followed by introns and much reduced methylation at gene termini.

As a core experimental model, the further characterization of DNA methylation in *C. elegans* represents an important opportunity to test the specific role of this process in transcriptional fidelity and genome stability and the potential role of methylation in the transition to eukaryotic genome complexity.

CHAPTER - 3

SHIFTING PATTERNS OF DNA METHYLATION

Genome Wide Bisulfite Sequencing of Three Strains of *C. elegans*

Background - Preliminary findings from an analysis of methylation of the *C. elegans* genome clearly show low levels of active methylation. The pattern of methylation was relatively unbiased with respect to symmetric and asymmetric motifs however, significant biases were found with respect to the sequence function. In that case the frequency of MeC containing sites was much higher in gene bodies, including transposable elements snRNAs and rRNAs (CH-2, Fig. 9). In addition, in the MeC enriched DNA analysis we found two distinct categories of MeC containing sites, those where the vast majority of reads covering the site contain MeC and those where only a minority of the reads appear to be methylated. While these patterns are clear, there relative proportions and therefor the representation of these patterns in the native genome may be biased by the enrichment step prior to bisulfite treatment.

To test the possibility that the MeC enriched analysis was biased we repeated the process of bisulfite sequenced the entire genome of *C. elegans* using a new culture of the same laboratory strain (N2). This analysis will allow us to not only examine the

biases associated with enrichment but allow us to explore the reproducibility of our analysis albeit on enriched and non-enrich fractions.

Furthermore, because the laboratory strain N2 has been propagated for many generations outside normal selective conditions and with potentially reduced population sizes since it was isolated from compost by Sidney Brenner in 1974 (Brenner 1974), the patterns of methylation found in the laboratory strains could be different from patterns found in natural populations. To explore differences among strains we chose a recently isolated strain that is also one of the most divergent con-specific isolates of *C. elegans* (PB306). This strain displays variation in fecundity (Harvey and Viney 2007) and differs in patterns of natural base-substitution polymorphism (Denver, Wilhelm et al. 2012).

Finally, as discussed in Chapter I we have identified multiple putative DNA methyltransferase genes in *C. elegans*. One such gene has been independently described (Gao, Liu et al. 2012) and appears to be homologous to DNMT1. To explore the potential contribution of this DNMT to the pattern of MeC in the genome we conducted a parallel genome wide bisulfite sequencing analysis of a homozygous deletion strain (VC2864) provided by the *C. elegans* knockout consortium (<http://celeganskoconsortium.omrf.org/>).

Methods

All strains were cultured and propagated using standard methods (Brenner 1974) and DNA was extracted from mixed stages using the Qiagen genomic tip protocol. The Illumina sequencing libraries were constructed using the Nextflex Bisulfite Sequencing kit (Bioo Scientific) and were then sequenced (100bp paired-end Illumina Hiseq 2000). The resulting reads were then mapped to the reference genome WS187 using the BSMAP program detailed in Chapter 2. This report gives the chromosome, position, the number of times a read mapped to that position and the number of times that read had a C that was not converted to a T. BSMAP parameters were set to a fragment size of 100-500bps, 8 processors were used, seed size was set to 14, 5 mismatches were allowed in the alignment, and max number equal best hits to count was set to 10. All methylated Cs included in the analysis must have at least 3 confirming reads that mapped uniquely and also had its paired end map uniquely at an appropriate distance. Maximum coverage was set to 2000 to account for the errors associated with extremely deep coverage of bisulfite sequencing such as read duplicates and sequencing error that may skew the MeC to non-MeC ratio and contribute to false positives and false negatives. Furthermore, the ratio per methylated C site to be included in the analysis was at least 2% of the total coverage must be methylated C to account for any false positives due to incomplete conversion of non-methylated Cs (> 0.01%) and sequencing error (> 0.9%) (Genereux, Johnson et al. 2008; Quail, Kozarewa et al. 2008).

All strains were verified by de novo and reference assembly. The contigs that resulted from the de novo assembly were queried for the deleted gene confirming that the putative DNMT gene Y75B8A.6 in VC2864 was deleted and was not deleted in N2. Similarly, reference assembly showed reads mapping to the putative DNMT gene Y75B8A.6 in N2 and no reads mapping in VC2864. Verification of the PB306 strain was inferred from the identification of known PB306 polymorphisms (Denver, Wilhelm et al. 2012).

Results and Discussion

The initial bisulfite sequencing analysis in Chapter 2 involving the enrichment for DNA fragments containing MeC using an MeC binding protein based approach resulted in bisulfite treated reads that numbered 74,285,412. A total of 70,004,089 reads were not mapped due to low quality, being unpaired, or having no match. 4,281,323 reads were included in the analysis with an average read coverage of 3.81 and the fraction of the reference covered was 0.81. This means 6% of the reads were included in the analysis and resulted in a fraction of the genome with high read coverage as is expected from enriching for areas of methylated DNA in an organism with low levels of methylation. In the enrichment analysis there were 160,988 sites with 3 or more MeC containing reads and an average coverage at those sites of 31.27.

By contrast, in our genome wide bisulfite sequencing (GWBS) the percentage of reads aligned improved dramatically (Table 1). In addition to the number of sites that mapped and the number of positions that met the criteria for inclusion (>3 and > 2% of reads showing evidence of MeC) was almost an order of magnitude greater in N2 and VC2864 samples. This improvement could be a result of a number of variables such as the longer read length with additional complexity per read, going from 76 bp to 100bp, or a more robust library preparation protocol. Also the greater percentage of reads mapped could reflect the unbiased nature of GWBS and the inclusion of all methylated sites not enriched for constitutively or facultatively

methyated sites resulting in a more uniformly distributed mapping of reads to the genome from the greater diversity of sequence. If enrichment by MeC binding protein strongly favors fragments with multiple MeC containing positions this could simultaneously bias the reads to a smaller, more densely methylated fraction of the genome and reduce the mapping efficiency of the enriched reads when mapping to the reference.

The specific filters applied to these analyses that define a putative MeC containing site are different across samples. For the enriched samples, we limit the analysis to sites with 3 or more reads containing a C, while for the N2, VC2864 and PB306 data we required 3 or more reads containing C and greater than 2% of the total reads. This was done because with such deep coverage the errors will contribute erroneously to the generation of false positives due to either non conversion or sequencing error. We also increased the threshold for constitutive verses facultative sites from 50% in the enriched analysis to 80% in the GWBS analysis to again account for the scale of read coverage and the contribution of error that could result from a tenfold increase in reads analyzed.

In this chapter we compare all datasets using an 80% MeC containing reads definition of constitutive. While these filters are stringent based on the estimated rates of errors false positive that do arise will be strongly biased toward positions showing facultative patterns of MeC. Therefore in this analysis we focus primarily on the

differences between facultative and constitutive patterns and compare each across the four datasets.

Strain	Total Read Pairs	Aligned Read Pairs	Percent Aligned	Valid Mappings	Average C Coverage	Analyzed Sites After Filtering	Coverage After Filter
N2 Enriched	74,285,412	4,283,932	6%	4,019,408	3.81	160,988	31.27
N2	37,767,748	25,177,722	67%	44,405,906	16.92	1,010,585	4.70
VC	76,801,941	62,616,272	82%	110,314,818	43.30	1,585,465	11.03
PB	957,524	359,736	38%	656,102	3.12	8,243	6.90

Table. 2 Assembly Statistics of Genome Wide Bisulfite Sequencing

Assembly statistics for GWBS for three strains N2 laboratory strain (N2), the DNMT knock out strain (VC2846) and the natural isolate PB306 (PB). BSMAP parameters were set to a fragment size of 100-500bps, 8 processors were used, seed size was set to 14, 5 mismatches were allowed in the alignment, and max number equal best hits to count was set to 10.

Symmetric vs. Asymmetric Methylation

In Figure 12 the proportion of symmetric vs. asymmetric patterns represented by the MeC positions differ consistently for when all sites are considered, when only facultative sites are considered and when only the constitutive sites are considered between the enriched sample and the three GWBS datasets.

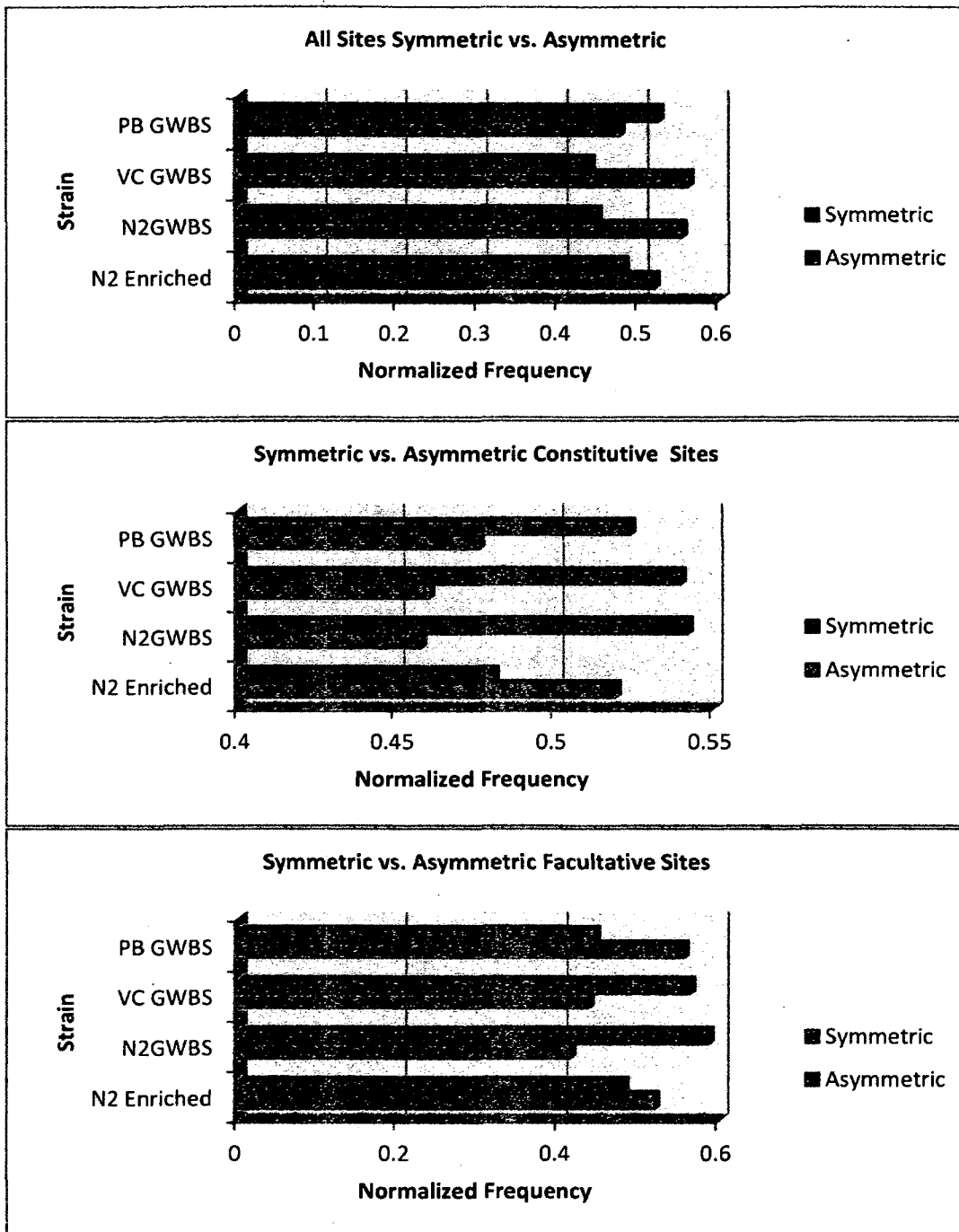


Fig.12 Ratios of Symmetric vs. Asymmetric Constitutive Methylation

Ratios of constitutive methylation are calculated by normalizing counts for asymmetric or symmetric methylation by all asymmetric or symmetric sites in the reference then divided by the total constitutively, facultatively and all methylated sites. Constitutive

sites were defined as >80% of reads containing Cs in all cases. A 3-sample test for equality of proportions reveals that the ratios of asymmetric and symmetric methylation are significantly different when comparing all strains; p-value < 2.2e-16 (asymmetric) and p-value < 2.2e-16 (symmetric).

In a comparison between the initial analysis in Chapter 2 where the sample was enriched for methylated DNA, constitutively methylated sites dominated, however in the GWBS analysis of N2 and VC2684 it appears that facultative methylation is dominant with over 90% of the methylated sites being facultatively methylated in all categories and chromosomes (Figure 15 and 16). One explanation for this is that the GWBS data for N2 and VC2864 includes a more inclusive representation of MeC containing sites not enriched in the DNA methylation binding protein derived dataset. An alternative explanation is that the N2 and VC datasets contain a significant fraction of false positives. In support of this second hypothesis the GWBS analysis of PB306 resulted in a very high proportion of constitutive sites which could be explained by the much lower number of reads and thus much lower level of false positives using the same filtering parameters. However, this observation could also reflect a difference in pattern in the PB306 genome. To test this, we analyzed a random subset of the data for VC2864 and N2 normalized to the coverage for PB306. We found the ratios of constitutive to facultative methylation remained consistent (Table 2). Furthermore, when analyzing the entire dataset for each strain and filtering out the facultative sites by using a cutoff of 80% methylation ratio we find that there are more constitutive sites in PB306 (5,812)

than VC2864 (684) and N2 (745) combined. Taken together these data suggest that while the GWBS datasets for N2 and VC may include some false positives that bias the facultative ratio upward there appears to be a significant increase in both the ratio and number of constitutively methylated sites in the natural isolate PB306.

Strain	Total Read Pairs	Aligned Read Pairs	Percent Aligned	Valid Mappings	Average C Coverage	Analyzed Sites After Filtering	Constitutive sites	Facultative Sites
N2	942,131	534,261	57%	3,064,814	4.15	7,374	606	6,768
VC	1,153,865	655,245	57%	2,953,871	5.1	14,417	863	13,554
PB	957,524	359,736	38%	656,102	3.12	8,243	5,812	2,431

Table. 3 Assembly Statistics of Genome Wide Bisulfite Sequencing

Assembly statistics for GWBS for three strains N2 laboratory strain (N2), the DNMT knock out strain (VC2846) and the natural isolate PB306 (PB). The number of reads per strain were normalized based on a random subset of data from VC2864 and N2. The last two columns of the table show that the constitutive versus facultative site ratios remain consistent with all data included when the analysis is started with a normalized number of reads. BSMAP parameters were set to a fragment size of 100-500bps, 8 processors were used, seed size was set to 14, 5 mismatches were allowed in the alignment, and max number equal best hits to count was set to 10.

As observed in the enriched N2 analysis (Chapter 2), scatter plots (Figure 15) reveal that the GWBS datasets all still show two distinct groups of methylated MeC containing sites. However, in stark contrast to the enriched analysis the large number of positions with deep coverage (up to 300-400 fold) with nearly all MeC base calls observed exclusively in chromosomes I and III are no longer observed in N2, VC or the PB306 datasets. This observation may suggest that the original observation in the enriched data is due to those sites on Chromosome I and III have a much higher affinity for the enrichment step.

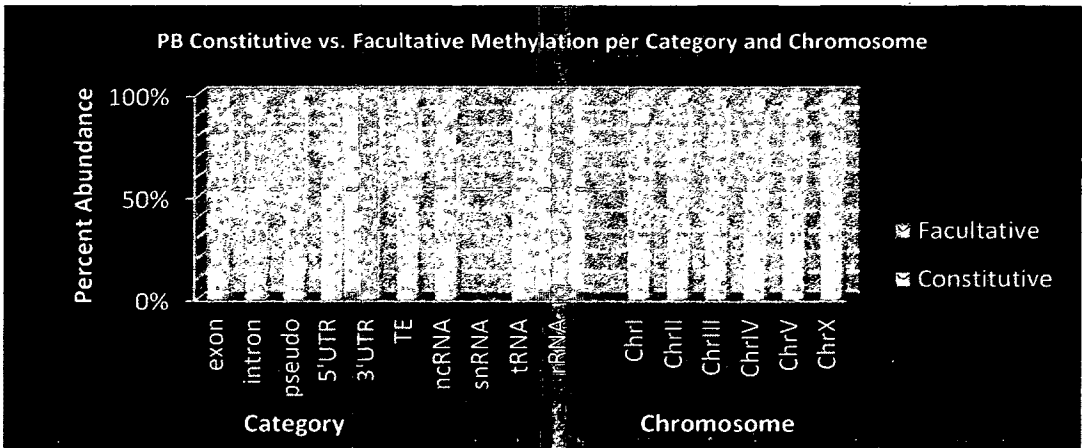
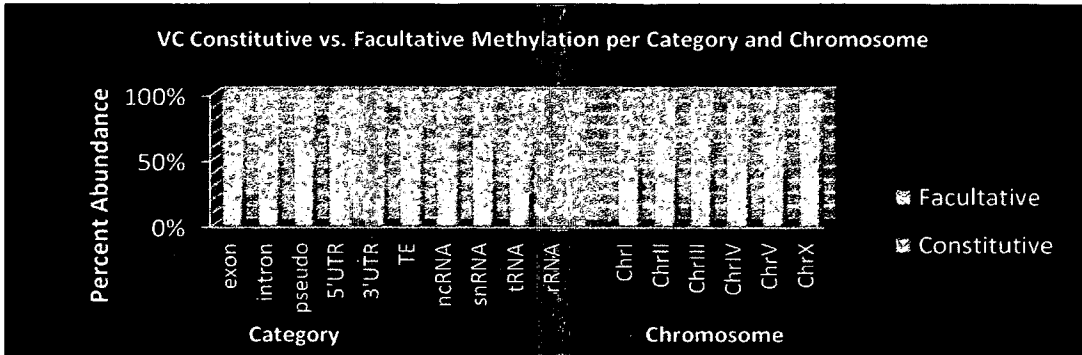
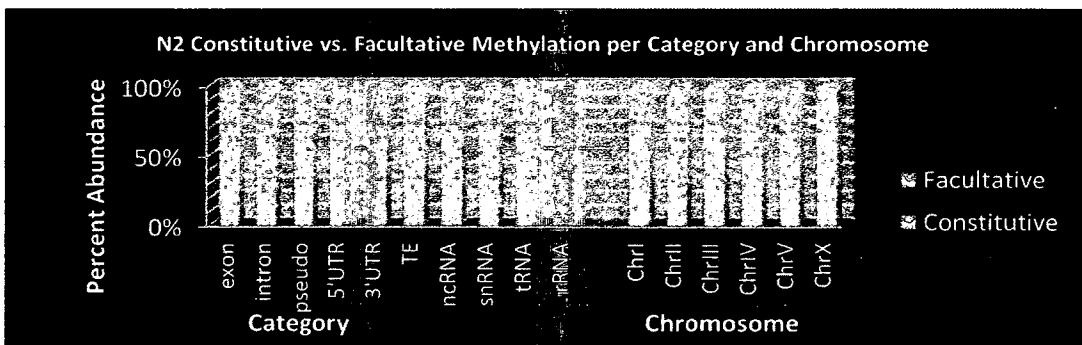
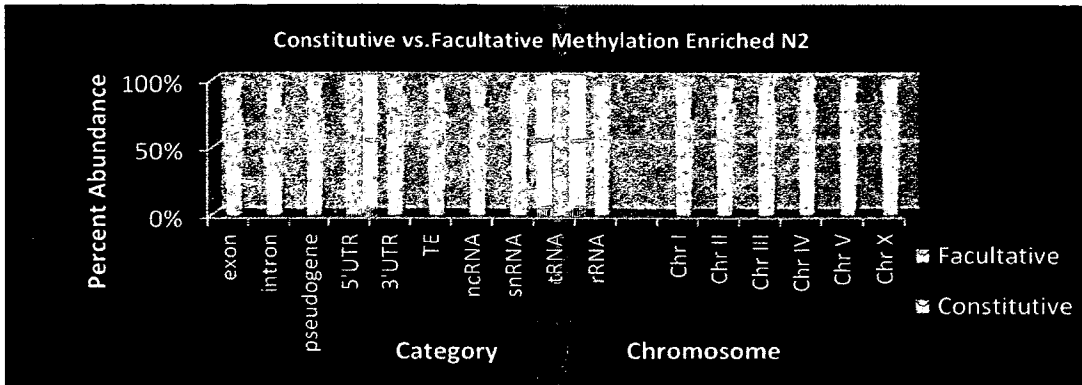
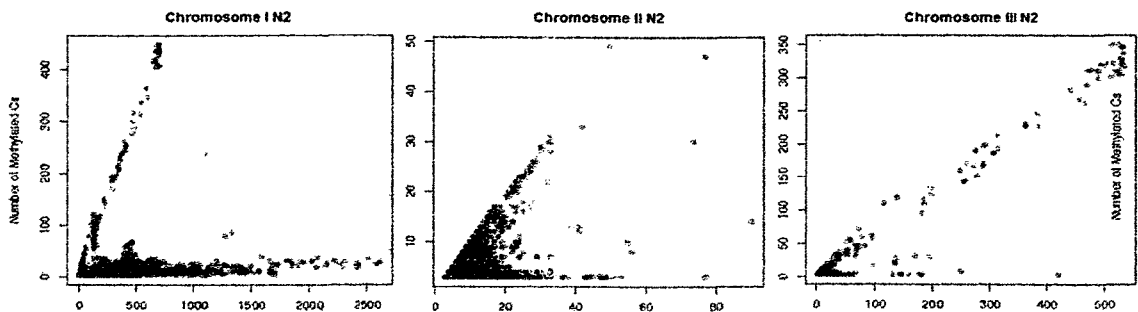


Fig. 13 Comparison of Enriched vs. GWBS Constitutive vs. Facultative Methylation per Chromosome per Category

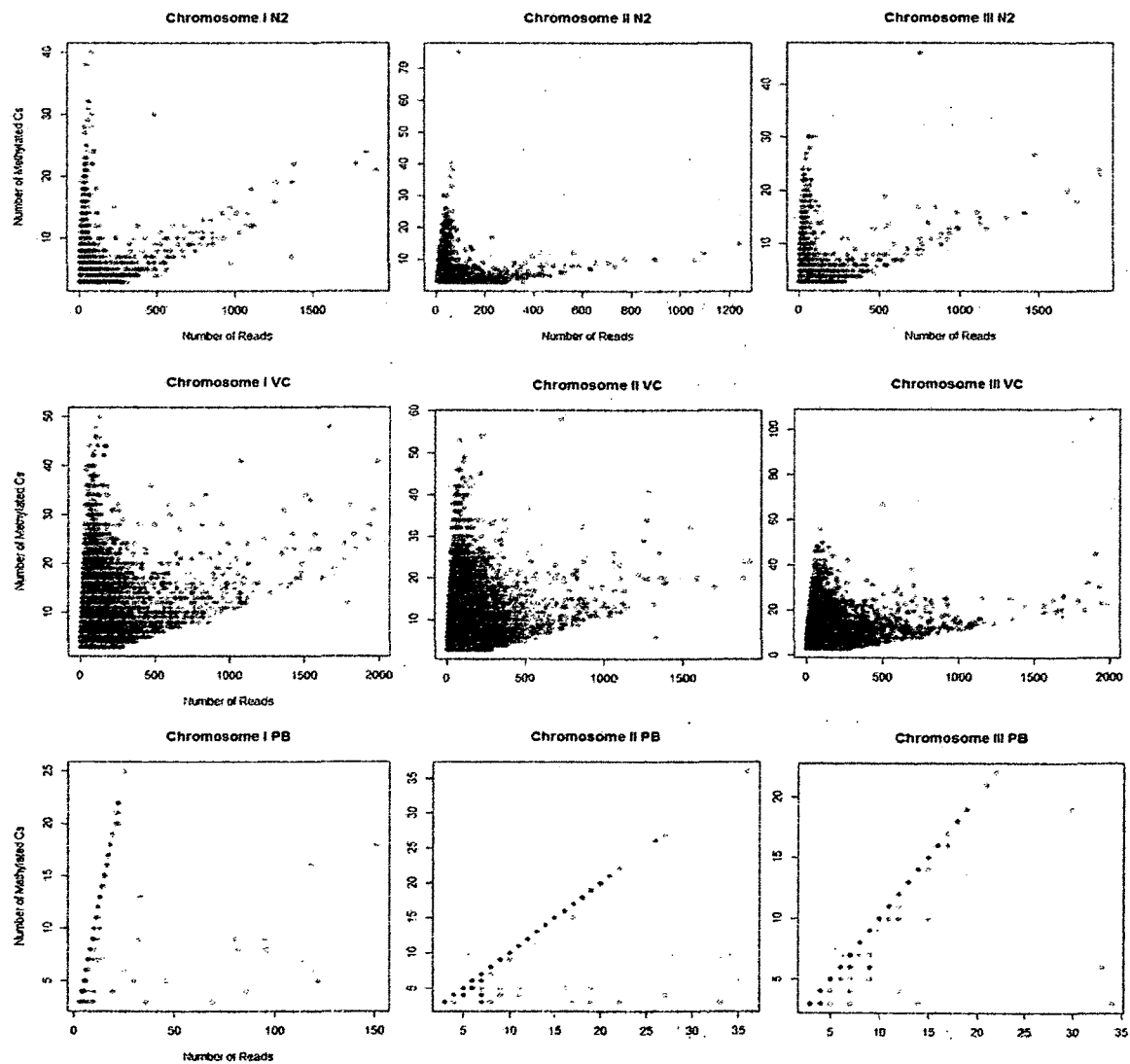
A comparison between categories and chromosomes reveals the different modes of methylation in the enriched N2 (A) versus GWBS (B)N2, (C) VC2684 and (D) PB306. For the enriched dataset constitutively methylated Cs were called at a cut off of %50 or more of the total reads for the loci that show evidence for methylation (3 or more putative MeC). Facultative or differentially methylated loci were called at a cut off of %50 at those sites. For the GBWS data maximum coverage was set to 2000 and the ratio per methylated C site to be included in the analysis was at least 2% of the total coverage must be methylated C and must have at least 3 MeC containing reads. Constitutively methylated Cs were called at a cut off of 80% or more of the total reads per loci providing evidence for methylation. Facultative or differentially methylated loci were called at a cut off of 80% or less of the total reads for the loci provide evidence for methylation. The Y axis is the total methylated sites and the percentage of facultative methylation (red) and the percentage of constitutive sites (blue) of that total.

Consistent with the hypothesis that the primary mode of DNA methylation in the natural isolate PB306 is constitutive methylation, we observe most categories to be dominated by constitutive methylation (Fig.). The only exception to this is the ribosomal DNA category, however, the same pattern was shared with the methylation enriched dataset in Chapter 2, where there is strong evidence to support that the enrichment step biases the sequencing of constitutively methylated sites. Furthermore, since there are multiple copies of ribosomal DNA in the form of ribosomal repeats it is not surprising that the ribosomal DNA category is not consistent with the rest of the categories that contain unique sequence. The multicopy nature of the rRNA repeat precludes us from knowing if this results from a subset of nuclei with a position methylated in all cases or a subset of the repeats methylated in all nuclei.

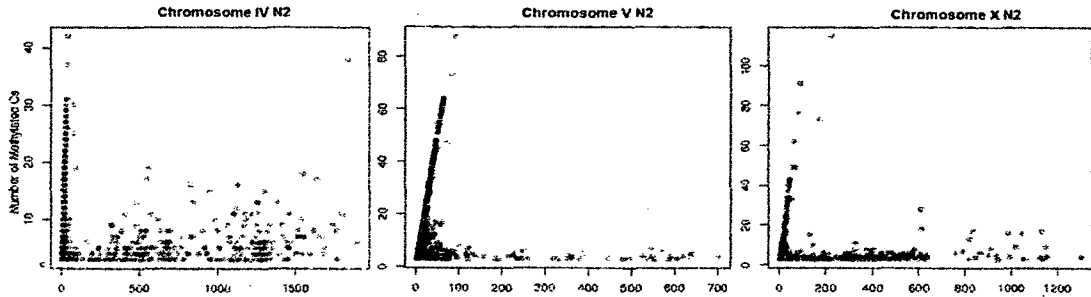
Methylation Pulldown



Genome Wide Bisulfite Sequencing



Methylation Pulldown



Genome Wide Bisulfite Sequencing

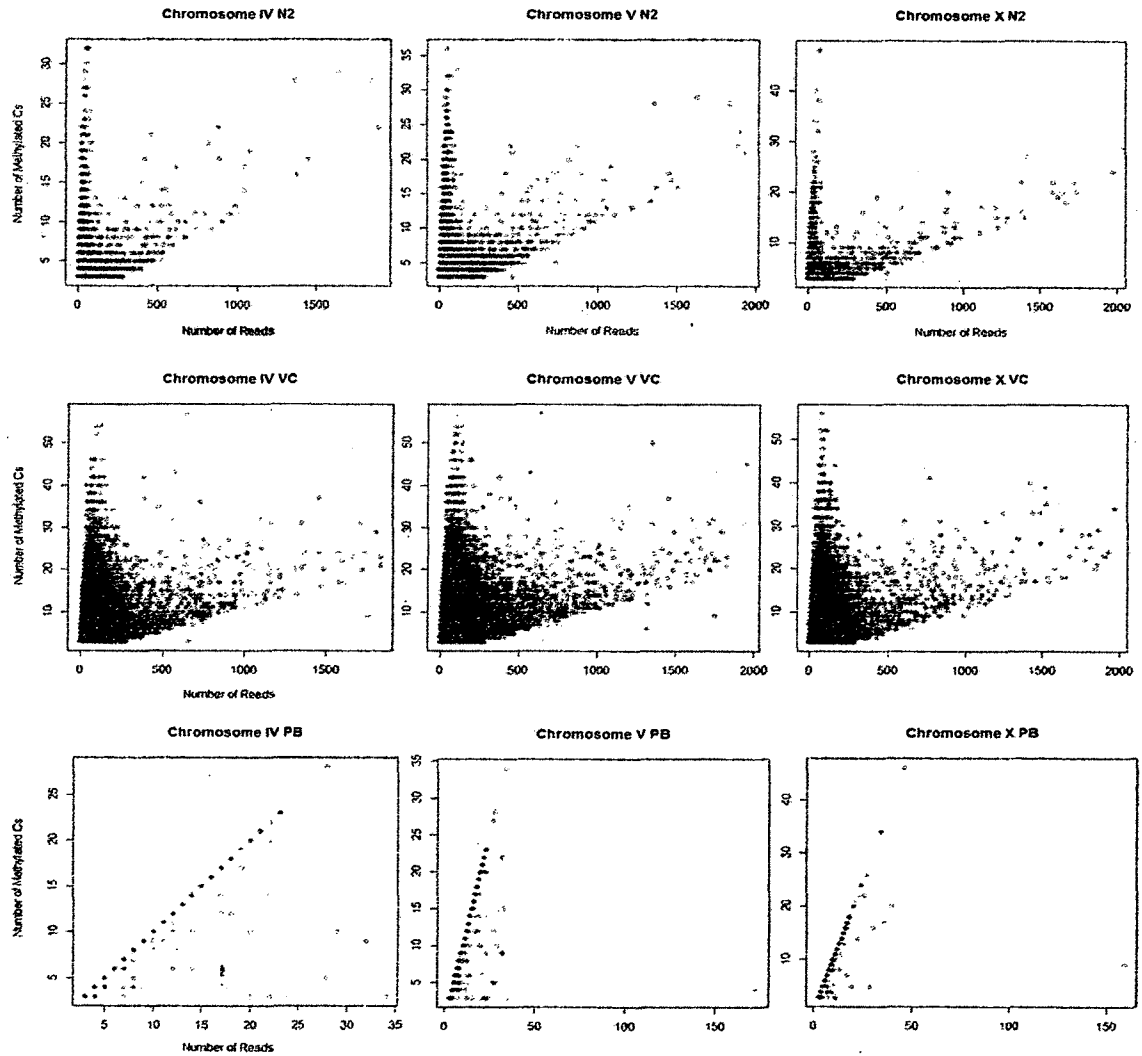


Fig. 14 Comparison of The Distribution of Constitutive vs. Facultative Methylation per Chromosome

A comparison between chromosomes reveals that the mode of DNA methylation observed has consistent distribution between chromosomes. The scatter plots are divided in two, the first division being Chromosome I – III and the second being Chromosome IV-X. The top rows are methylated C enriched data from Chapter 2, the bottom three rows are GWBS data from this chapter.

The distribution of methylated sites per chromosome differed between the natural isolate and laboratory strains as well (Fig.20-22). MeC counts were divided into 1Mb bins along each chromosome by position and the frequency was plotted on the same scale to show the relative levels of methylation across each chromosome. For the GWBS analysis of N2 and VC2684 the pattern observed in the enriched data is lost. By contrast the same pattern observed in the enrich sample (biased toward the gene rich chromosome cores) was observed in the PB306 GWBS analysis. Together these observations are similar in pattern to the shared biased toward constitutive sites found in the enriched N2 datasets and PB306.

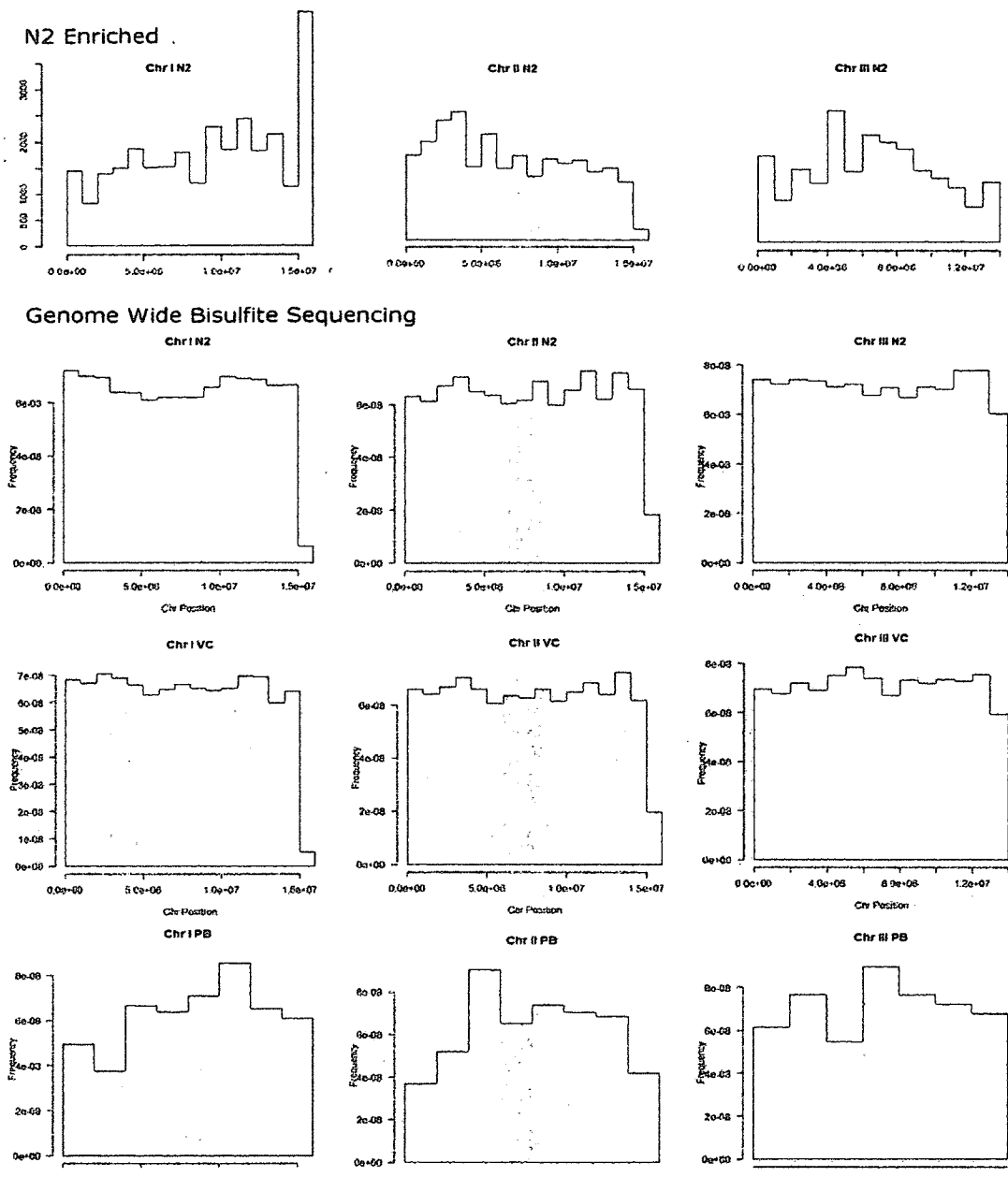


Fig. 15 Distribution of MeC per Chromosome I-III N2 Enriched vs. GWBS

MeC counts were divided into 1Mb bins along each chromosome by position and the frequency was plotted on the same scale per strain to show the relative levels of methylation across each chromosome.

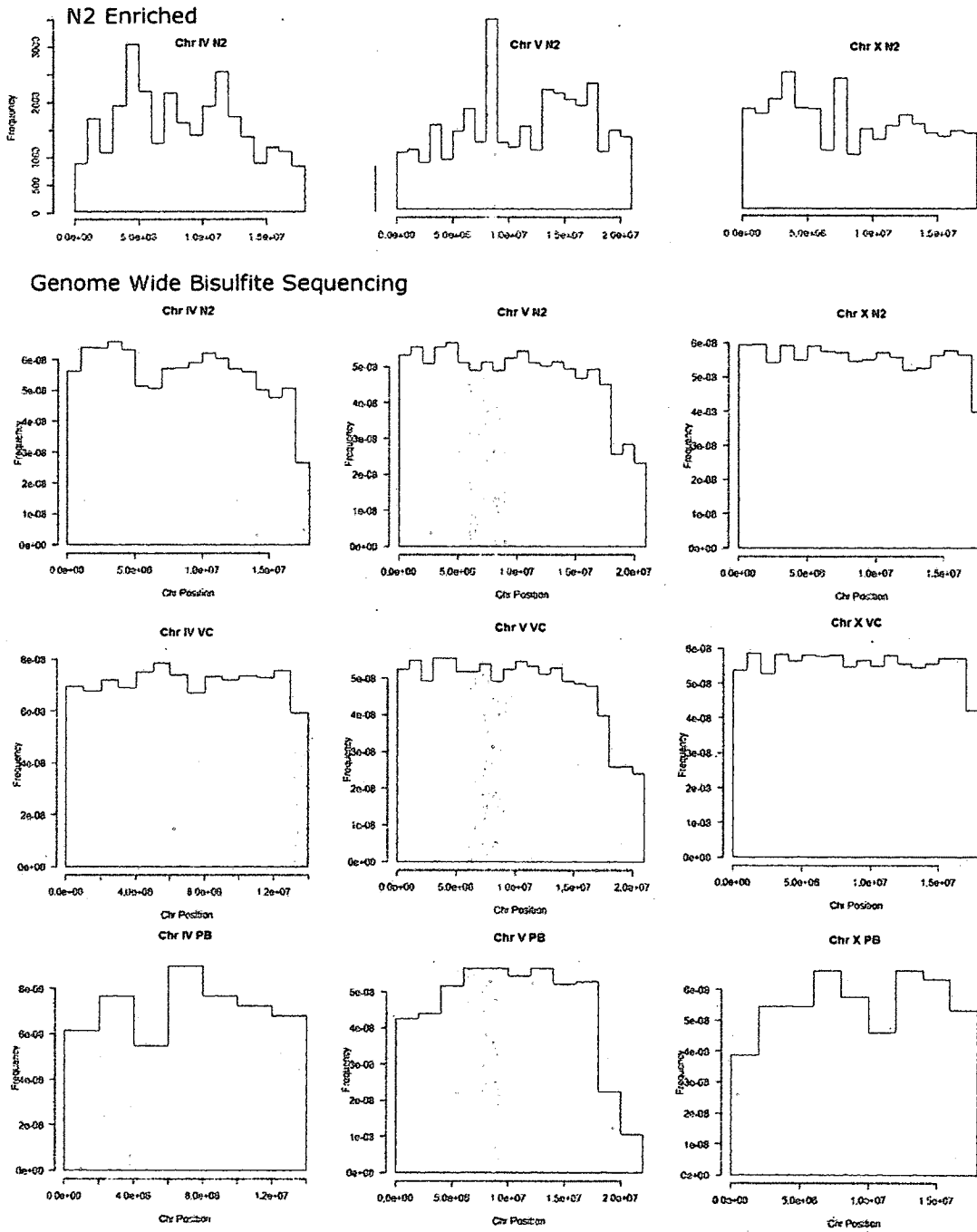


Fig. 16 Distribution of MeC per Chromosome IV-X N2 Enriched vs. GWBS

MeC counts were divided into 1Mb bins along each chromosome by position and the frequency was plotted on the same scale per strain to show the relative levels of methylation across each chromosome.

Analysis of the functional distribution of putative MeC containing sites from the GWBS analysis

In a comparison between the original data from Chapter 2 and the GWBS data of N2 in this chapter (figure 23), the core patterns of bias toward methylation of gene bodies and transposable elements are not only reproduced but more extreme in the GWBS datasets.

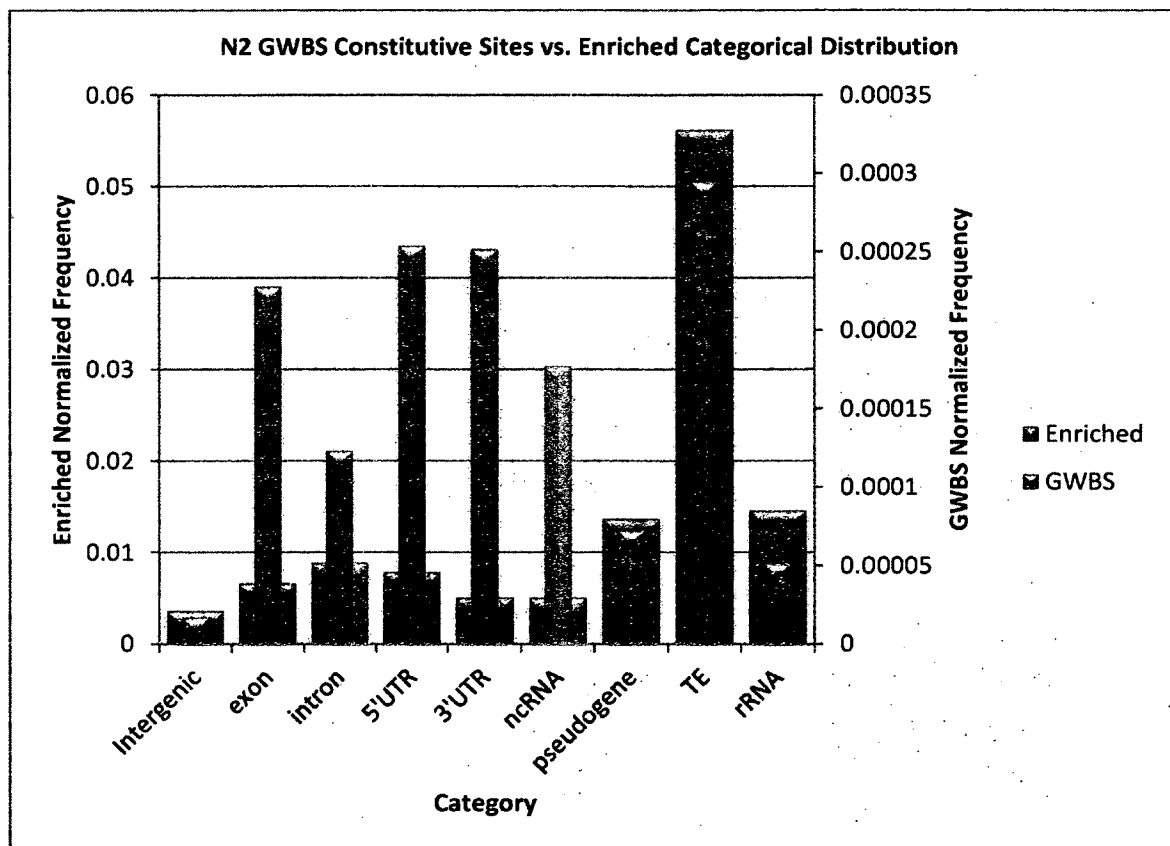


Fig. 17 Comparison of Constitutively Methylated Sites per Category N2 GWBS vs. N2 Enriched

C. elegans reference sequences were used for alignment and categorization of all putative methylated sites. Data were normalized based on all Cs in the *C. elegans* genome. Categories included; Intergenic (IG), exons(coding), 5'UTR s, introns, non-coding RNAs (ncRNA), pseudogenes (pseudo), small nuclear RNAs (snRNA), 3'UTRs, transposable elements (TE), transfer RNAs (tRNA) and ribosomal RNAs (rRNA). For the enriched dataset constitutively methylated Cs where called at a cut off of %50 or more

of the total reads for the loci that show evidence for methylation (3 or more putative MeC). Facultative or differentially methylated loci were called at a cut off of %50 at those sites. For GWBS data the maximum coverage was set to 2000 and the ratio per methylated C site to be included in the analysis was at least 2% of the total coverage must be methylated C and must have at least 3 MeC containing reads. Constitutively methylated Cs were called at a cut off of 80% or more of the total reads per loci providing evidence for methylation.

In a comparison across all taxa, the only significant difference between the lines was a shift in the proportion of constitutive MeC sites in the 5' UTRs. While both enriched and GWBS datasets for N2 had comparable levels of MeC in both 5' and 3' UTRs VC2864 showed a much reduced proportion of MeC in the 5' UTR. Interestingly a pattern similar to that observed in the PB306 datasets

Applying the same comparison to the natural isolate PB306, we find correlation in UTR methylation patterns between VC2864 constitutive sites when all sites are considered in PB306 (Fig.18), as well as, when only constitutive sites are considered (Fig.18). Despite UTR methylation patterns being shared between VC2864 and PB306, we find all other categories to be significantly different in comparison to N2 and VC2864. Another observation is that when comparing all methylated sites to only constitutive sites in PB306 the pattern does not change with the single exception being the disappearance of rRNA methylation in the constitutive only analysis. This suggests that in natural isolates, constitutive methylation is the dominant mode while laboratory strains have developed an abundance of facultatively methylated sites.

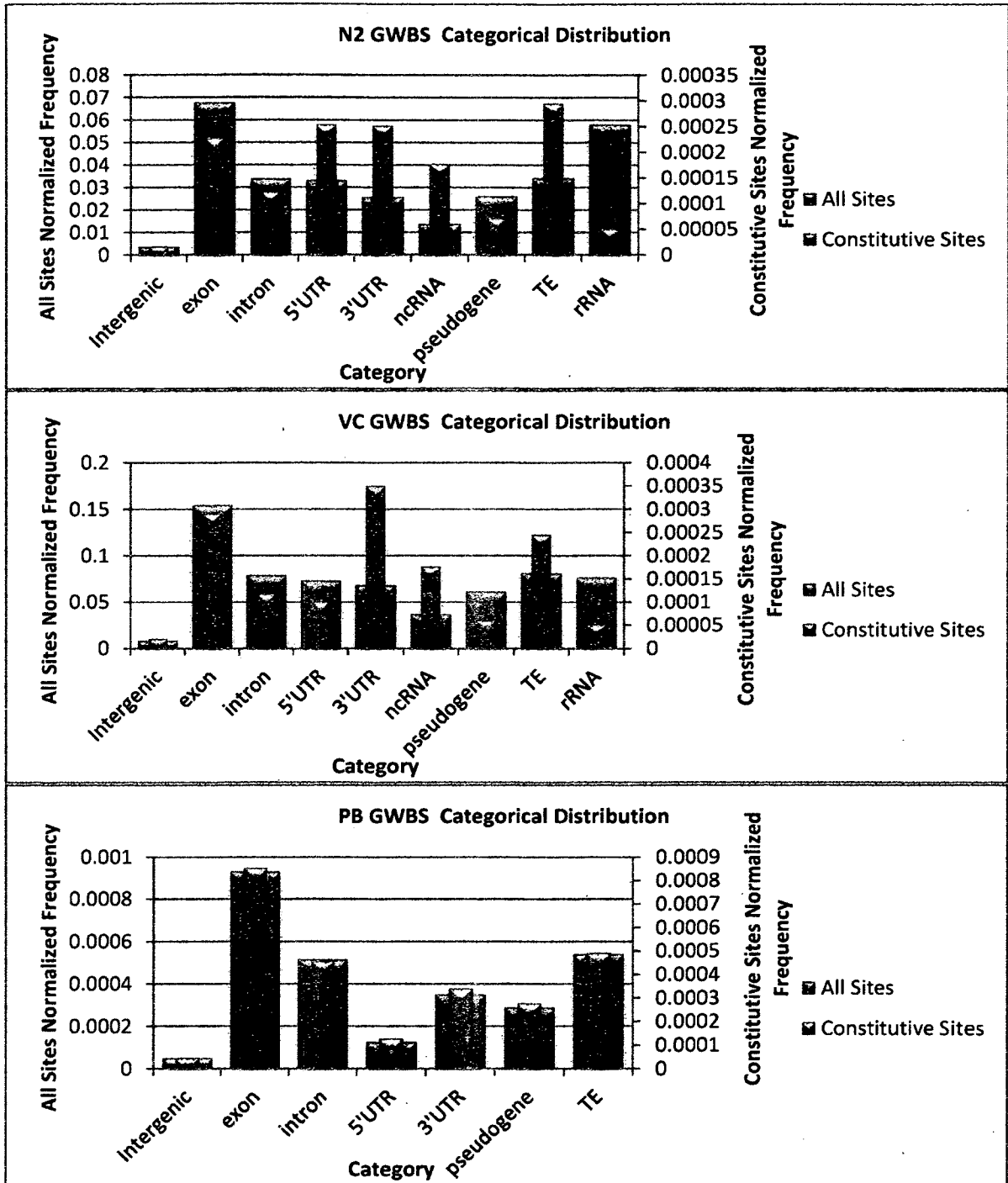


Fig. 18 Comparison of All MeC Sites vs. Constitutive Sites per Category N2 GWBS vs. PB306 vs. VC2864

C. elegans reference sequences were used for alignment and categorization of all

putative methylated sites. Data were normalized based on all Cs in the *C. elegans* genome. Categories included; Intergenic (IG), exons(coding), 5'UTR s, introns, non-coding RNAs (ncRNA), pseudogenes (pseudo), small nuclear RNAs (snRNA), 3'UTRs, transposable elements (TE), transfer RNAs (tRNA) and ribosomal RNAs (rRNA). Maximum coverage was set to 2000 and the ratio per methylated C site to be included in the analysis was at least 2% of the total coverage must be methylated C and must have at least 3 MeC containing reads. Constitutively methylated Cs where called at a cut off of 80% or more of the total reads per loci providing evidence for methylation. In PB306 (last frame) the ncRNA and rRNA categories were ignored due to lack of significant data.

Reproducibility and conservation of Methylated sites

Expected Shared Sites of Methylation Between Data Sets

	N2 Enriched	N2 GWBS	VC GWBS	PB GWBS
N2 Enriched	160,988			
N2 GWBS	3	606		
VC GWBS	4	0	863	
PB GWBS	26	0	0	5,812

Table.4 Expected Shared Sites of MeC Between Datasets

In a final comparison among strains we set out to compare the specific sites that were methylated in each analysis. To focus this comparison we limited our analysis to the constitutively methylated positions. Table 4 shows the total number of constitutively methylated sites in each dataset (above diagonal) and the expected number of overlapping sites is we assume a random distribution (below diagonal).

Focused PCR and direct sequencing of Bisulfite treated DNA.

As a method to confirm the existence of constitutively methylated sites in these genomes we designed PCR primers that flank putative constitutive MeC sites and amplified two regions using bisulfite treated DNA as a template. The PCR products were sequenced directly using traditional Sanger sequencing. The sequences were aligned with Clustal W (Li 2003) to the region of the reference that the primers were targeted to. The region of the reference that corresponds to the PCR product was also confirmed by BLAST as being the best hit. We have confirmed 33 out of 36 MeC sites tested (28 in N2 and 5 in VC2864). As can be seen in Figure 19 complete conversion of non-methylated C to T is observed and several positions are clearly dominated by MeC in the direct PCR sequencing experiments consistent with our Illumina data.

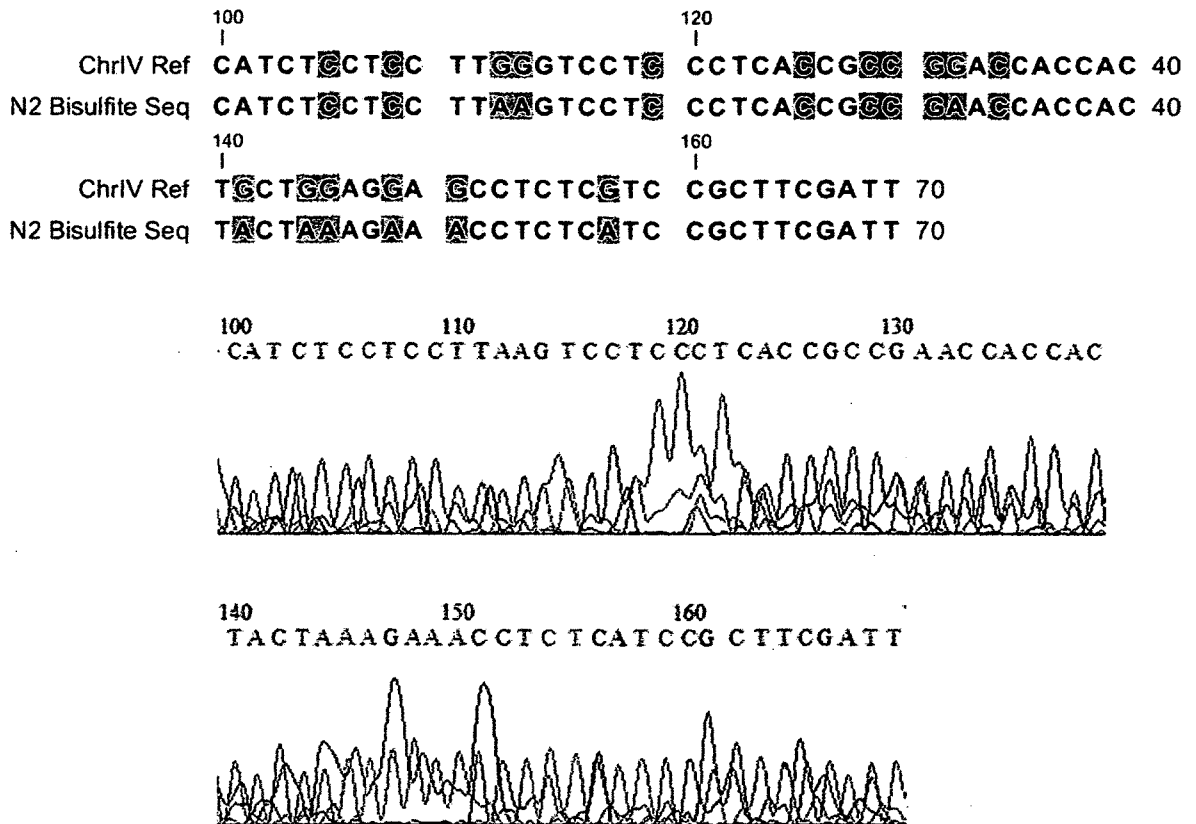


Fig.19 Clustal W Alignment MeC Conformation PCR Product Against the Reference

Alignment and corresponding chromatogram of PCR product sequenced in the reverse direction (N2 Bisulfite Seq Product) targeted to a region in Chromosome IV (ChrIV Reference) where, in N2, there is a high concentration of constitutively methylated C. Bases shaded blue in the alignment represent confirmed MeC sites and the bases shade in red are sites that were converted during the bisulfite treatment process.

Conclusion

In the first genome wide analysis of DNA methylation in three strains of *C. elegans* we have shown that the genomes of these strains are indeed methylated. After confirming the strains validity, we have shown that the modes of DNA methylation differ between the natural isolate PB306 and the laboratory strain. Initially, the laboratory strain N2 was shown to primarily methylate constitutively, however, we now have strong evidence to suggest that this was an artifact of enrichment bias. Through the use of genome wide bisulfite sequencing we have discovered that the primary mode of methylation in the laboratory strain is in fact facultative since both laboratory strains primarily methylate this way. Also when filtering the GWBS data for exclusively constitutively methylated sites we find a shift from completely different categorical methylation patterns to almost identical categorical methylation patterns. Furthermore, we have found that the natural isolate is dominated by constitutively methylated sites and in comparison to the laboratory strain, has over three times more constitutively methylated sites. There are differences in distribution of methylation along the chromosomes, however, the targeting of contexts are similar when comparing all strains studied, but differ in levels with all strains having a preponderance for non-symmetric methylation yet PB306 having a more even non-symmetric to symmetric ratio and N2 and VC having 4-5 times more non-symmetric methylation than symmetric.

It is clear that in all strains surveyed, the ancestral pattern of methylation

is shared when filtered for constitutive sites and when all sites are considered. Even though gene body methylation is consistent, surprisingly the mode of methylation differs tremendously between the natural isolate and the laboratory strain leaving one to question if these modes and patterns of methylation are a derivative of a long life in the laboratory or if the natural isolate is an exception.

Another striking observation is that only one methylated site is shared between data sets. One would expect that if the methylated sites are in fact constitutive and these constitutive sites are heritable then at least the constitutive sites would be conserved in at least the comparison between the enriched N2 and GWBS N2 data sets. However, we only find one shared site between VC2864 and N2 GWBS. Even though there is only one shared site, all share the same preponderance for gene body methylation and when looking at constitutive categorical methylation levels and patterns, they are conserved between the two N2 datasets. Therefore it seems that while methylation targets remain heritable, “hitting the bull’s eye” or exact site of methylation may not be as important. This may explain the random distribution of symmetric and asymmetric methylation and may shed light on a possible mechanism to overcome the deleterious effects of DNA methylation by not consistently methylating and destabilizing the same C. On the other hand, this may be the result of the mutagenic effects of methylation. While in some nuclei, the target C remains unmethylated, in other nuclei it is methylated and the mutagenic effects of methylation cause the deamination of the MeC and make heritability of this mark no longer possible.

Therefore the intergenerational conservation of the exact site of methylation is not inherited and the methylation pattern is reset and the re-targeting of gene bodies is initiated.

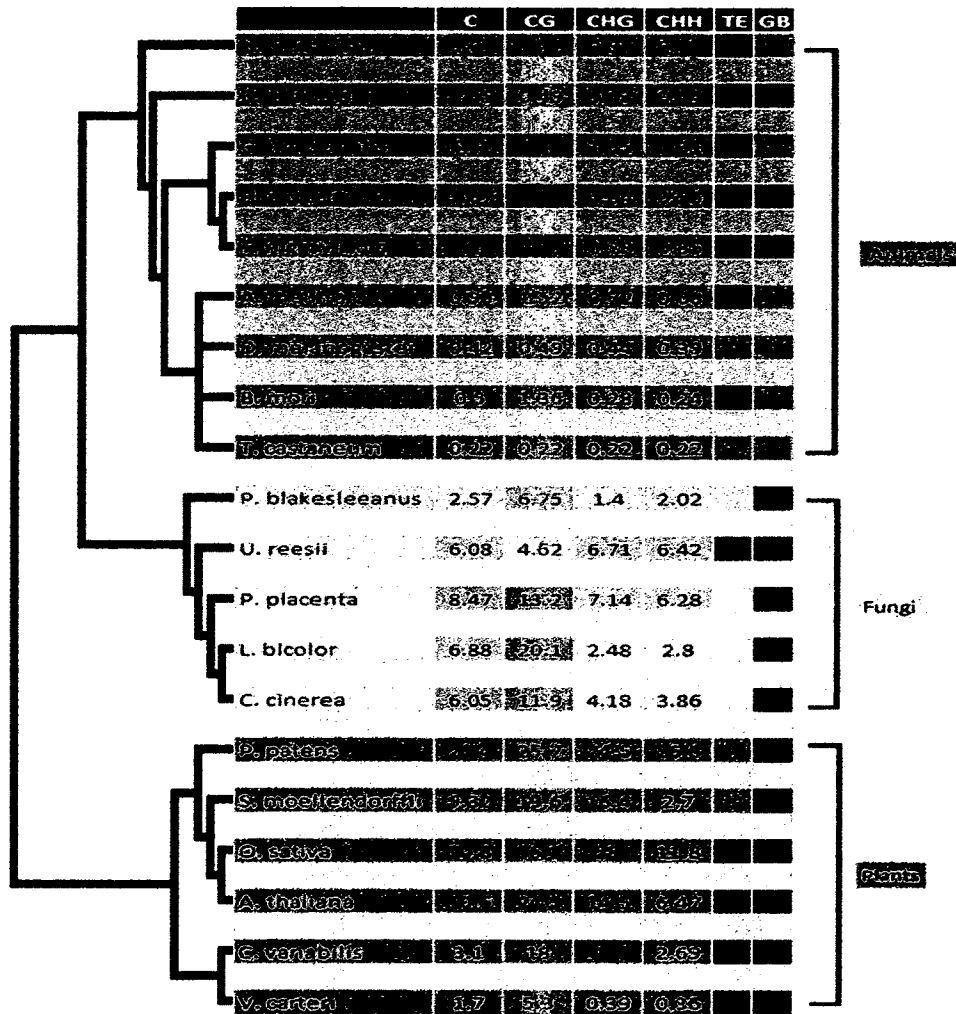


Fig. 20 Normalized frequencies of DNA methylation context and evolution

The phylogenetic tree was based on the NCBI Taxonomy Browser. The values represent the normalized fraction in percent of methylated Cs per motif. The filled boxes on the right indicate high methylation of gene bodies (GB) and transposable elements (TE). Data obtained from; Lister et al., 2009, Zemack et al., 2010, Su et al., 2011 and this study.

CHAPTER – 4

THE MUTAGENIC CONSEQUENCES OF DNA METHYLATION

Background

Methylation of cytosine residues was first demonstrated to be mutagenic in *E. coli* (Coulondre, Miller et al. 1978). These initial studies identified methylated cytosines (MeC) as hotspots for spontaneous base substitutions. Mutations which occur at CpG dinucleotides are easily recognized because of the nature of base substitutions. Deamination of MeC at CpG dinucleotides results in the formation of TpG. Alternatively, if deamination occurs on the complementary DNA strand CpA is generated.

Methylation of cytosine at a CpG dinucleotide has been shown in mammalian cells to increase the probability of a C→T or corresponding G→A transition mutation between 12- and 42-fold (Cooper and Youssoufian 1988). The increased deamination rate of MeC relative to C, however, still does not account for the high frequency of mutagenesis observed at CpG sites (Gonzalzo and Jones 1997). The G→T mismatches resulting from deamination of MeC are also believed to be more difficult for the cell to repair than G→U mismatches which can result from the deamination of cytosine, since thymine unlike uracil is a normal component of DNA. The high efficiency of repair of G→U but not G→T mismatches by the well characterized uracil-DNA glycosylase (UDG) repair pathway may also contribute to the increased relative frequency of mutagenesis

caused by MeC deamination (Gonzalvo and Jones 1997). Excision of U has been found to be as much as 6000-fold more efficient than excision of T at identical template sites using extracts from human colonic mucosa (Schmutte, Yang et al. 1995).

In the context of genome wide MeC there remains a conundrum. If DNA methylation is mutagenic or causes increases in cytosine deamination, we would expect to see a higher occurrence of G/C to A/T base substitutions subjecting MeC containing positions to a directional mutation pressure. In fact, this is inherent in the logic used to explain the existence of CpG islands where hypomethylation near promoters results in a presumed lack of said directional mutation pressure and the accumulation of CpG Islands. To test the hypothesis that sites containing MeC are mutagenic in *C elegans*, we set out to compare the methylation landscape (those sites identified as MeC containing) to the spontaneous mutation landscape in *C. elegans* (Denver, Dolan et al. 2009). We are in fortuitous position having a single nucleotide resolution map of positions that have undergone spontaneous mutation in Mutation Accumulation (MA) lines of *C. elegans* strain N2.

Methods

Mutation positions were provided by Denver et al., (Denver, Dolan et al. 2009) using two different *C. elegans* reference versions WS170 and WS185. Since the methylation analysis was conducted using the *C. elegans* reference version WS187, the mapping of the methylation datasets were all redone using *C. elegans* reference versions WS170 and WS185 for direct comparison between positions of MeC sites and sites of high mutation rates. Mapping parameters and filters were set identically to the analysis in Chapter 2 for the enriched dataset and Chapter 3 for the GWBS datasets.

Results and Discussion

Recently, Denver et al. (Denver, Dolan et al. 2009) performed a genome wide mutation study on *C. elegans* using Mutation Accumulation (MA) lines. By bottlenecking and reducing the effective population size, this study has characterized the mutational landscape of *C. elegans*. Interestingly, a strong mutational bias from G/C to A/T nucleotides was detected in the MA lines. By comparing the positions of these mutations to the positions found to be methylated, the effects of DNA methylation on single nucleotide polymorphisms can be explored.

In their work Denver and colleagues identified mutations at 393 sites across 12 different lines. Of these sites 220 were G and C positions reflecting a very strongly biased pattern of spontaneous mutation toward G/C to A/T transitions. By comparison, sites of methylation as presented in Chapters 2 and 3 were confined predominately to genic regions or regions of high gene density and numbered 160,988 for the enriched N2 dataset, 1,010,585 for the N2 GWBS dataset, 1,585,465 for VC2864 GWBS, and 8243 for PB306. When we compare the specific positions containing MeC in the methylated DNA enriched dataset from Chapter 2 with the MA line mutation, we find no base substitution mutations that share the same position with methylated positions, at random we would expect 1 site to be shared when only G/C is considered (Table2). When analyzing the N2 and VC2864 GWBS data from Chapter 3 we also find that none

of the sites of methylation overlap with sites of mutation even though we expect 6 and 10 respectively.

An implication of this is that DNA methylation may not contribute to an increase in base substitution mutation rate or that the pattern of methylation is not conserved in the MA lines.

Ratios of Expected Shared Sites Methylation and Mutation

	G/C Sites	Total Bases	Expected Shared Sites	Observed Shared Sites
N2 Enriched	160,988	35,539,203	1.00	0
N2 GWBS	1,010,585	35,539,203	6.26	0
VC2864 GWBS	1,585,465	35,539,203	9.81	0
PB306 GWBS	8,243	35,539,203	0.05	0

Table. 5 Expected vs. Observed Frequency of Shared Sites Enriched and GWBS Datasets.

Expected ratios were calculated as the product of the ratios of occurring sites in mutations and methylation. G/C sites only include the total sites occurring at G and C. Total bases is every G and C in the reference. Expected ratios were calculated as the product of the ratios of occurring sites in mutation sites, methylation sites and total G/C sites. We find that there is no obvious correlation between methylated sites and observed mutations.

The lack of correlation between sites of DNA methylation and mutation may reflect an absence of DNA methylation in the germline or different pattern of methylation restricted to the germline and not reflected in our methylation assays. Another explanation could be that DNA methylation may not have a large effect on mutation or may target repair mechanisms to counteract the mutational effects (Cuozzo, Porcellini et al. 2007). It should also be noted that while we might expect MeC

sites to be hotspots for mutation and among the first to appear in MA experiments the mechanisms giving rise to the preponderance to G/C to A/T mutations in the *C. elegans* MA lines is just not likely to be MeC residues. As pointed out in Denver and Colleagues (with the “knowledge” that *C. elegans* does not have MeC) the likely mechanisms based on spontaneous damage would be oxidative resulting in 5-hydroxyuracil (resulting from the oxidative deamination of cytosine) and 8-oxoguanine. Similarly, as pointed out in chapter 3, while the patterns of methylation are strongly reproducible across lines the positions defined as MeC in each analysis show no overlap. This is not surprising given these are epigenetically inherited but also dramatically reduces the effectiveness of a directional mutation model where if sites were consistently methylated over many generations their existence would be short lived. When this is taken into account, given the low level of methylation in *C. elegans* and the shifting positions containing MeC it is unlikely that MeC is a significant mutagenic force in the *C. elegans* genome. However, since we find no intergenerational conservation of methylated sites, there remains a possibility that not inheriting the specific site of methylation is due to depletion of C/G nucleotides. Additionally, if loss of C/G sites due to mutation is tied to methylation we would observe higher rates of mutation in areas of higher rates of methylation and not specifically the site of methylation. With this reasoning we would expect to see mutation rates increase in areas of gene bodies. In fact, in MA experiments Denver et al., did observe a higher rate of mutations in coding regions (Fig.20).

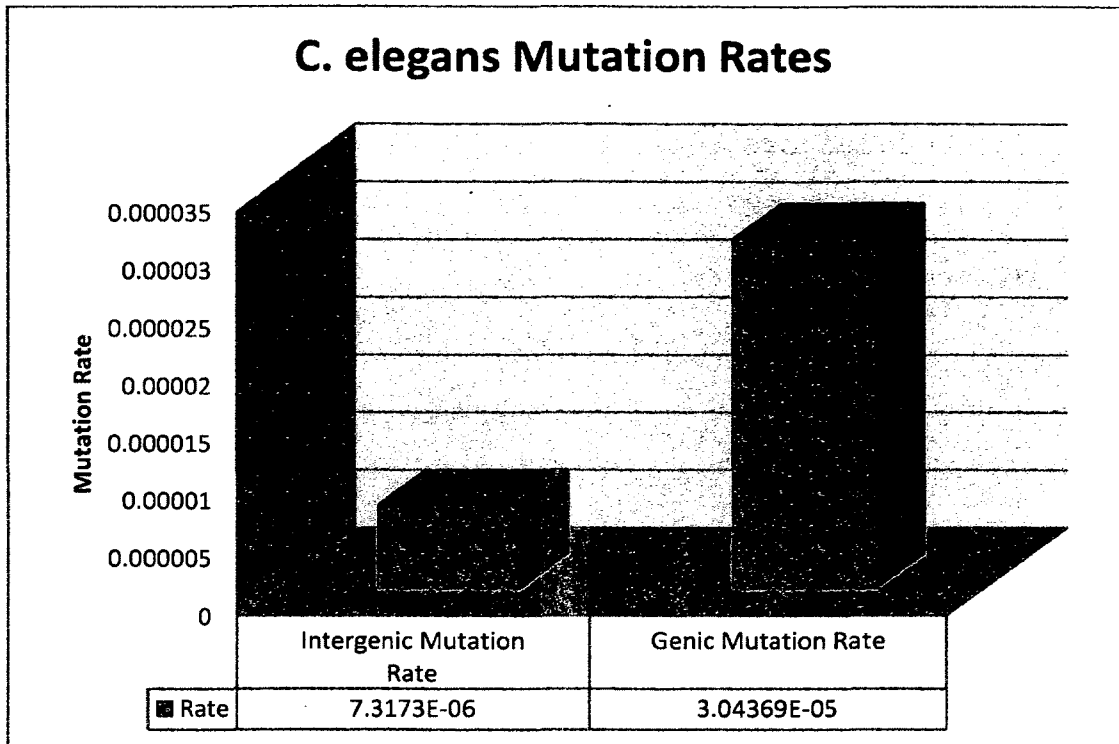


Fig. 21 *C.elegans* Intergenic vs. Genic Mutation rates

Sites of mutation were categorized and divided into counts of intergenic and genic sites of mutation. These positions were then filtered for only C or G as the original base in the reference. The rates were calculated as the number of intergenic or genic C/G mutation divided by the total number C/G sites in the reference in intergenic or genic regions. A test for equality of proportions reveals that the mutation rates for intergenic regions differ significantly from the rate of genic mutation; p-value < 2.2e-16.

Conclusion

Since patterns of methylation differ from generation to generation, we find no correlation between methylated sites and sites of high mutation rate. However, we find strong correlation in genomic regions of high methylation rates and high mutation rates. Moreover, the mutational bias within those regions of C/G to T/A also suggests the involvement of methylation mediated deamination and ultimately depletion of C/G sites.

While we may not expect a site specific directional mutation effect of MeC, the strong spatial bias observed in chapters 2 and 3 toward methylation of genes bodies could result in higher mutation rates in genes compared to intergenic regions and a potential directional mutation pressure that could lower GC content within gene bodies.

REFERENCES

- Bailey, T. L., M. Boden, et al. (2009). "MEME Suite: tools for motif discovery and searching." Nucleic Acids Research **37**(suppl 2): W202-W208.
- Bestor, T., A. Laudano, et al. (1988). "Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases." J Mol Biol **203**(3210246): 971-983.
- Bestor, T. H. (2000). "The DNA methyltransferases of mammals." Human Molecular Genetics **9**(16): 2395-2402.
- Bird, A. and S. Tweedie (1995). "Transcriptional Noise and the Evolution of Gene Number." Philosophical Transactions: Biological Sciences **349**(1329): 249-253.
- Bird, A. P. and A. P. Wolffe (1999). "Methylation-induced repression--belts, braces, and chromatin." Cell **99**(10589672): 451-454.
- Boyes, J. and A. Bird (1991). "DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein." Cell **64**(2004419): 1123-1134.
- Brenner, S. (1974). "The genetics of *Caenorhabditis elegans*." Genetics **77**(1): 71-94.
- Cao, X., W. Aufsatz, et al. (2003). "Role of the DRM and CMT3 Methyltransferases in RNA-Directed DNA Methylation." Current Biology **13**(24): 2212-2217.
- Cao, X. and S. E. Jacobsen (2002). "Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes." Proc Natl Acad Sci U S A **99** Suppl 4(12151602): 16491-16498.
- Cao, X. and S. E. Jacobsen (2002). "Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing." Curr Biol **12**(12121623): 1138-1144.
- Cao, X., N. M. Springer, et al. (2000). "Conserved plant genes with similarity to mammalian de novo DNA methyltransferases." Proc Natl Acad Sci U S A **97**(10781108): 4979-4984.
- Chandler, V. L. and V. Walbot (1986). "DNA modification of a maize transposable element correlates with loss of activity." Proc. Natl Acad. Sci. USA **83**: 1767-1771.
- Colot, V. and J. L. Rossignol (1999). "Eukaryotic DNA methylation as an evolutionary device." Bioessays **21**(10376011): 402-411.
- Cooper, D. N. and H. Youssoufian (1988). "The CpG dinucleotide and human genetic disease." Hum Genet **78**(3338800): 151-155.
- Coulondre, C., J. H. Miller, et al. (1978). "Molecular basis of base substitution hotspots in *Escherichia coli*." Nature **274**(355893): 775-780.
- Cuozzo, C., A. Porcellini, et al. (2007). "DNA Damage, Homology-Directed Repair, and DNA Methylation." PLoS Genet **3**(7): e110.
- Cutter, A. D., A. Dey, et al. (2009). "Evolution of the *Caenorhabditis elegans* Genome."

- Molecular Biology and Evolution **26**(6): 1199-1234.
- Denver, D. R., P. C. Dolan, et al. (2009). "A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes." Proceedings of the National Academy of Sciences **106**(38): 16310-16314.
- Denver, D. R., L. J. Wilhelm, et al. (2012). "Variation in Base-Substitution Mutation in Experimental and Natural Lineages of *Caenorhabditis* Nematodes." Genome Biology and Evolution **4**(4): 513-522.
- Emmons, S. W. and L. Yesner (1984). "High-frequency excision of transposable element Tc1 in the nematode *Caenorhabditis elegans* is limited to somatic cells." Cell **36**(3): 599-605.
- Feng, S., S. J. Cokus, et al. (2010). "Conservation and divergence of methylation patterning in plants and animals." Proceedings of the National Academy of Sciences **107**(19): 8689-8694.
- Finnegan, E. J. and K. A. Kovac (2000). "Plant DNA methyltransferases." Plant Mol Biol **43**(10999404): 189-201.
- Finnegan, E. J., W. J. Peacock, et al. (1996). "Reduced DNA methylation in *Arabidopsis thaliana* results in abnormal plant development." Proc Natl Acad Sci U S A **93**(8710891): 8449-8454.
- Gao, F., X. Liu, et al. (2012). "Differential DNA methylation in discrete developmental stages of the parasitic nematode *Trichinella spiralis*." Genome Biology **13**(10): R100.
- Genereux, D. P., W. C. Johnson, et al. (2008). "Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies." Nucleic Acids Research **36**(22): e150.
- Goll, M. and T. Bestor (2005). "Eukaryotic cytosine methyltransferases." Annu Rev Biochem **74**: 481 - 514.
- Gonzalzo, M. L. and P. A. Jones (1997). "Mutagenic and epigenetic effects of DNA methylation." Mutat Res **386**(9113112): 107-118.
- Goyon, C., T. I. Nogueira, et al. (1994). "Perpetuation of cytosine methylation in *Ascobolus immersus* implies a novel type of maintenance methylase." J Mol Biol **240**(8021939): 42-51.
- Gruenbaum, Y., H. Cedar, et al. (1982). "Substrate and sequence specificity of a eukaryotic DNA methylase." Nature **295**(7057921): 620-622.
- Gruenbaum, Y., T. Naveh-Manly, et al. (1981). "Sequence specificity of methylation in higher plant DNA." Nature **292**(6267477): 860-862.
- Gutierrez, A. and R. J. Sommer (2004). "Evolution of dnmt-2 and mbd-2-like genes in the free-living nematodes *Pristionchus pacificus*, *Caenorhabditis elegans* and *Caenorhabditis briggsae*." Nucleic Acids Research **32**(21): 6388-6396.
- Harvey, S. C. and M. E. Viney (2007). "Thermal variation reveals natural variation between isolates of *Caenorhabditis elegans*." Journal of Experimental Zoology Part B: Molecular and Developmental Evolution **308B**(4): 409-416.

- Hata, K., M. Okano, et al. (2002). "Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice." Development **129**(8): 1983-1993.
- Hendrich, B. and A. Bird (1998). "Identification and characterization of a family of mammalian methyl-CpG binding proteins." Mol Cell Biol **18**(9774669): 6538-6547.
- Henikoff, S. and L. Comai (1998). "A DNA methyltransferase homolog with a chromodomain exists in multiple polymorphic forms in Arabidopsis." Genetics **149**(9584105): 307-318.
- Hermann, A., S. Schmitt, et al. (2003). "The Human Dnmt2 Has Residual DNA-(Cytosine-C5) Methyltransferase Activity." Journal of Biological Chemistry **278**(34): 31717-31721.
- Hung, M. S., N. Karthikeyan, et al. (1999). "Drosophila proteins related to vertebrate DNA (5-cytosine) methyltransferases." Proc Natl Acad Sci U S A **96**(10518555): 11940-11945.
- Jeltsch, A. (2010). "Phylogeny of Methylomes." Science **328**(5980): 837-838.
- Jones, A. L., C. L. Thomas, et al. (1998). "De novo methylation and co-suppression induced by a cytoplasmically replicating plant RNA virus." EMBO J **17**(9799246): 6385-6393.
- Jones, L., F. Ratcliff, et al. (2001). "RNA-directed transcriptional gene silencing in plants can be inherited independently of the RNA trigger and requires Met1 for maintenance." Curr Biol **11**(11378384): 747-757.
- Kishimoto, N., H. Sakai, et al. (2001). "Site specificity of the Arabidopsis MET1 DNA methyltransferase demonstrated through hypermethylation of the superman locus." Plant Mol Biol **46**(11442057): 171-183.
- Klose, R. J. and A. P. Bird (2006). "Genomic DNA methylation: the mark and its mediators." Trends in Biochemical Sciences **31**(2): 89-97.
- Kunert, N., J. Marhold, et al. (2003). "A Dnmt2-like protein mediates DNA methylation in Drosophila." Development **130**(21): 5083-5090.
- Lawrence, R. J., K. Earley, et al. (2004). "A Concerted DNA Methylation/Histone Methylation Switch Regulates rRNA Gene Dosage Control and Nucleolar Dominance." Molecular Cell **13**(4): 599-609.
- Lee, P. P., D. R. Fitzpatrick, et al. (2001). "A Critical Role for Dnmt1 and DNA Methylation in T Cell Development, Function, and Survival." Immunity **15**(5): 763-774.
- Leonhardt, H., A. W. Page, et al. (1992). "A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei." Cell **71**(1423634): 865-873.
- Li, E., T. H. Bestor, et al. (1992). "Targeted mutation of the DNA methyltransferase gene results in embryonic lethality." Cell **69**(6): 915-926.
- Li, K.-B. (2003). "ClustalW-MPI: ClustalW analysis using distributed and parallel computing." Bioinformatics **19**(12): 1585-1586.

- Li, L., C. J. Stoeckert, et al. (2003). "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." Genome Research **13**(9): 2178-2189.
- Li, R., Y. Li, et al. (2008). "SOAP: short oligonucleotide alignment program." Bioinformatics **24**(18227114): 713-714.
- Lindroth, A. M., X. Cao, et al. (2001). "Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation." Science **292**(11349138): 2077-2080.
- Lister, R., M. Pelizzola, et al. (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences." Nature **462**(19829295): 315-322.
- Liu, S., L. Lin, et al. (2011). "A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species." Nucleic Acids Research **39**(2): 578-588.
- Lyko, F. (2001). "DNA methylation learns to fly." Trends Genet **17**(4): 169-172.
- Lyko, F., B. H. Ramsahoye, et al. (2000). "DNA methylation in *Drosophila melanogaster*." Nature **408**(11117732): 538-540.
- Lynch, M. and J. S. Conery (2003). "The origins of genome complexity." Science **302**(5649): 1401-1404.
- Matzke, M. A., M. F. Mette, et al. (2000). "Transgene silencing by the host genome defense: implications for the evolution of epigenetic control mechanisms in plants and vertebrates." Plant Mol Biol **43**(10999419): 401-415.
- Maunakea, A. K., R. P. Nagarajan, et al. (2010). "Conserved role of intragenic DNA methylation in regulating alternative promoters." Nature **466**(7303): 253-257.
- Okano, M., D. W. Bell, et al. (1999). "DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development." Cell **99**(3): 247-257.
- Okano, M., S. Xie, et al. (1998). "Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases." Nat Genet **19**(3): 219-220.
- Parfrey, L. W., D. J. G. Lahr, et al. (2011). "Estimating the timing of early eukaryotic diversification with multigene molecular clocks." Proceedings of the National Academy of Sciences **108**(33): 13624-13629.
- Pélissier, T., S. Tutois, et al. (1996). "DNA regions flanking the major *Arabidopsis thaliana* satellite are principally enriched in Athila retroelement sequences." Genetica **97**(2): 141-151.
- Pfeifer, G. P. (2006). "Mutagenesis at methylated CpG sequences." Curr Top Microbiol Immunol **301**: 259-281.
- Pomraning, K. R., K. M. Smith, et al. (2009). "Genome-wide high throughput analysis of DNA methylation in eukaryotes." Methods **47**(18950712): 142-150.
- Quail, M. A., I. Kozarewa, et al. (2008). "A large genome center's improvements to the Illumina sequencing system." Nat Meth **5**(12): 1005-1010.
- Rakyan, V. K., T. Hildmann, et al. (2004). "DNA Methylation Profiling of the Human Major Histocompatibility Complex: A Pilot Study for the Human Epigenome Project." PLoS Biol **2**(12): e405.

- Ramsahoye, B. H., D. Biniszkiwicz, et al. (2000). "Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a." Proceedings of the National Academy of Sciences **97**(10): 5237-5242.
- Regev, A., M. J. Lamb, et al. (1998). "The role of DNA methylation in invertebrates: Developmental regulation or genome defense?" Molecular Biology and Evolution **15**(7): 880-891.
- Rideout, W. M., G. A. Coetzee, et al. (1990). "5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes." Science **249**(1697983): 1288-1290.
- Robert, V. J. P., T. Sijen, et al. (2005). "Chromatin and RNAi factors protect the *C. elegans* germline against repetitive sequences." Genes & Development **19**(7): 782-787.
- Rockman, M. V. and L. Kruglyak (2009). "Recombinational Landscape and Population Genomics of *Caenorhabditis elegans*." PLoS Genet **5**(3): e1000419.
- Rollins, R. A., F. Haghghi, et al. (2006). "Large-scale structure of genomic methylation patterns." Genome Research **16**(2): 157-163.
- Ronemus, M. J., M. Galbiati, et al. (1996). "Demethylation-induced developmental pleiotropy in *Arabidopsis*." Science **273**(8662558): 654-657.
- Rountree, M. R., K. E. Bachman, et al. (2000). "DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci." Nat Genet **25**(3): 269-277.
- Schaefer, M. and F. Lyko (2007). "DNA methylation with a sting: An active DNA methylation system in the honeybee." Bioessays **29**(3): 208-211.
- Schmutte, C., A. S. Yang, et al. (1995). "Base excision repair of U:G mismatches at a mutational hotspot in the p53 gene is more efficient than base excision repair of T:G mismatches in extracts of human colon tumors." Cancer Res **55**(7641186): 3742-3746.
- Selker, E. U., D. Y. Fritz, et al. (1993). "Dense nonsymmetrical DNA methylation resulting from repeat-induced point mutation in *Neurospora*." Science **262**(8259516): 1724-1728.
- Selker, E. U., N. A. Tountas, et al. (2003). "The methylated component of the *Neurospora crassa* genome." Nature **422**(12712205): 893-897.
- Simmen, M. W., S. Leitgeb, et al. (1999). "Nonmethylated transposable elements and methylated genes in a chordate genome." Science **283**(10024242): 1164-1167.
- Simpson, V., T. Johnson, et al. (1986). "*Caenorhabditis elegans* DNA does not contain 5-methylcytosine at any time during development or aging." Nucleic Acids Res **14**: 9.
- Simpson, V. J., T. E. Johnson, et al. (1986). "*Caenorhabditis elegans* DNA does not contain 5-methylcytosine at any time during development or aging." Nucleic Acids Research **14**(16): 6711-6719.
- Suetake, I., F. Shinozaki, et al. (2004). "DNMT3L stimulates the DNA methylation activity

- of Dnmt3a and Dnmt3b through a direct interaction." J Biol Chem **279**(26): 27816-27823.
- Tycowski, K., M. Shu, et al. (1994). "Requirement for intron-encoded U22 small nucleolar RNA in 18S ribosomal RNA maturation." Science **266**(5190): 1558-1561.
- Vandegheuchte, M. B., F. Lemière, et al. (2009). "Quantitative DNA-methylation in *Daphnia magna* and effects of multigeneration Zn exposure." Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology **150**(3): 343-348.
- Watt, F. and P. L. Molloy (1988). "Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter." Genes Dev **2**(3192075): 1136-1143.
- Wilson, G. G. and N. E. Murray (1991). "Restriction and modification systems." Annu Rev Genet **25**(1812816): 585-627.
- Xi, Y. and W. Li (2009). "BSMAP: whole genome bisulfite sequence MAPPING program." BMC Bioinformatics **10**(1): 232.
- Yoder, J. A., C. P. Walsh, et al. (1997). "Cytosine methylation and the ecology of intragenomic parasites." Trends Genet **13**(9260521): 335-340.
- Zemach, A., I. McDaniel, et al. (2010). "Genome-wide evolutionary analysis of eukaryotic DNA methylation." Science **328**: 916 - 919.
- Zemach, A., I. E. McDaniel, et al. (2010). "Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation." Science **328**(5980): 916-919.
- Zhang, X., J. Yazaki, et al. (2006). "Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in *Arabidopsis*." Cell **126**(6): 1189-1201.