BI453 Research Project

Academic Year 2020/2021

An investigation into the phylogeny and domain conservation of the DNMT protein family

Student name: Kathryn Ruddle

Student ID: 16305036

Course: Undenominated Science

<u>Supervisor: Professor Uri Frank</u>

College of Science and Engineering

School of Natural Sciences

National University of Ireland, Galway

Abstract

DNA methylation is an essential epigenetic mechanism in mammalian cells. The cellular machines that carry out DNA methylation are known as DNA methyltransferases or DNMTs and this protein family also shows a high level of conservation among animals. Many previous studies have focused mainly on vertebrates and much of the findings on DNA methylation and its processes have been rooted in this group. Recently, the perceived uniqueness of the hypermethylated genome to vertebrates has been challenged by studies into different animal lineages such as the porifera and cnidarians. As our knowledge of DNA methylation across the tree of life grows, so does our knowledge of its evolutionary history and relationships. Using phylogenetic analysis, this project aims to investigate DNMT protein conservation across multiple animal lineages. In this project, the DNMT sequences from various species were compared, particularly invertebrates, to study to relationships between lineages and to compare the level of domain conservation between them. The findings in this study show that DNMT1 and DNMT3 were both present and well conserved in non-bilaterian animals. However, some species, including hydrozoans, possess DNMT1 that showed a loss of the DMAP1 (DNA methylase associated protein) binding domain while hydrozoans species specifically showed an additional PWWP domain in their DNMT3 sequences. These findings are beginning to show that the genomes of these species are much more complex than once thought.

Table of Contents

Abbreviations

Abbreviation	Full
CpG	Cytosine-phospate-Guanine
TpG	Thymine-phosphate-Guanine
DNMT	DNA methyltransferase
DMAP1	DNA methyltransferase associated protein
BAH	Bromo adjacent homology
PWWP	Proline-trypotophan-tryptophan-proline
ADD	Atrx-DNMT3-DNMT3l
MSA	Multiple sequence alignment
ML	Maximum likelihood
CDD	Conserved domain database
WGD	Whole genome duplication

Introduction

DNA Methylation

DNA methylation is the biological process where a methyl group is added to a DNA base pair. Through this biological modification, changes occur to how the DNA is interacted with within the cell albeit without altering the DNA sequence itself, making it an integral epigenetic feature (Kyger *et al*, 2020*).* Both adenine and cytosine can undergo DNA methylation. Adenine methylation (6mA) was, until recently, thought to be abundant only in prokaryotes but recent studies now suggest that it may be more widespread in eukaryotes than previously believed (Iyer *et al*, 2016). Alternatively, cytosine methylation can be divided into two categories; those that produce N4-methylcytosine (4mC) and those that produce C5-methylcytosine (5mC). The former is found mainly in prokaryotes, but the latter is widespread throughout eukaryotes, has been well documented and is the focus of this study.

During 5mC methylation, the methylation mechanism affects gene expression, with methylated regions of DNA being more transcriptionally ‘’silent’’. This in turn affects biological pathways such as embryonic development, cellular differentiation, repression of transposable elements and X-chromosome inactivation (review in Kyger *et al*, 2021). Cytosine methylation mainly occurs at areas rich in CpG (cytosine-guanine) dinucleotides in DNA, commonly referred to as CpG islands. Studies have shown that a large majority of gene promoters exist within these CpG rich regions in humans and other mammals (Moore *et al*, 2013). 5mC methylation also has an important evolutionary role, as methylated cytosines can transition to a deaminated state spontaneously. This causes cytosine-guanine regions to convert to thymine-guanine (TpG) regions over a period of evolution. Meaning researchers had to employ methods of calculating CpG to TpG conversion rates in order to envision what historic methylation patterns may have looked like (Kyger *et al*, 2020).

Despite this type of methylation being the most common DNA base modification in animals, the level of 5mC deposition varies across the animal kingdom. 5mC methylation is widespread in vertebrates, with their genomes being referred to as hypermethylated. Conversely, invertebrates generally do not display hypermethylated genomes, and are often referred to as a displaying a mosaic methylation pattern (de Mendoza *et al*, 2019). Hypermethylation in vertebrates is thought to be a major innovation in gene regulation yet studies on DNA methylation in an invertebrate like the sponge *Amphimedon queenslandica,* showed a

hypermethylated genome also. This has challenged the longstanding theory that the hypermethylated state is unique only to vertebrates (de Mendoza *et al*, 2019).

DNA methylation amongst animal species

Animals are classically grouped according to Linnaean taxonomy following the order of kingdom, phyla, class, order, family, genus and species. In modern classification, there is no single official way to classify organisms, there are also several supergroups and subgroups. One such rank is the superphylum. A superphylum groups together many separate phyla that most likely descended from a common ancestor (Ruggiero *et al*, 2015). Most of our knowledge of cytosine methylation derives from studies on bilaterians, a broad superphylum that classifies species based on bilateral symmetry at the embryo stage (Brusca, 2016). Under the umbrella of the bilateria superphylum, several sub-superphylum are included: Lophotrocozoa (includes Annelida and Mollusca), Ecdysozoa (includes Arthropoda and Nematoda) and Deuterostomia (includes Chordata, thus vertebrates, and Echinodermata). Existing as a sister group to bilaterians, there are Cnidaria, Ctenophora and Porifera. Cnidarians can be further divided into several classes: Anthozoa, Hydrozoa, Polypodiozoa and Staurozoa (Figure 1).

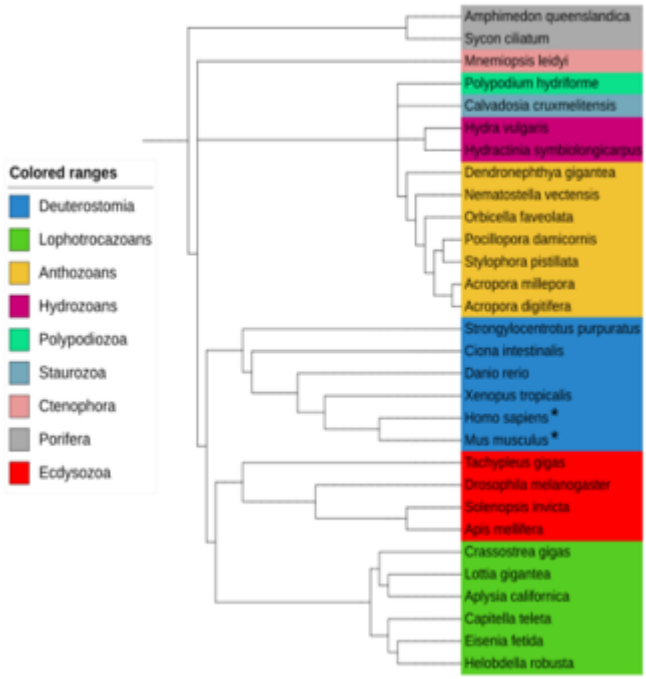


Figure 1: Cladogram depicting the species used in this study with the various colours representing their superphylum, phylum or class groupings (see legend on the left). *This is a cladogram created with phyloT, intended to display species used in the study as they are thought to relate to one another evolutionarily.* You can clearly see the various animal groups studied, grouped by superphylum (in the case of the lophotrochozoans, ecdysozoans, and deuterstomes), phylum (Porifera, Ctenophora) or simply class (anthozoans, hydrozoans, polypodiozoans and staurozoans)

It is widely accepted that some level of DNA methylation existed in the common ancestor of all animals, though whether this ancestor displayed a hypermethylated genome or a mosaic pattern more akin to invertebrates is not known (Bhattacharyya *et al*, 2020). Despite the fact that DNA methylation has been shown to play a key role in epigenetics, some species have completely lost this modification such as the nematode *Caenorhabditis elegans*. This would imply that DNA methylation is not a crucial mechanism in development of this particular species, regardless of its importance in others. Furthermore, the presence or absence of DNA methylation is still being disputed for some organisms such as *Drosophila melanogaster*. Interestingly, adenine methylation still occurs in *C.elegans* and *D. melanogaster*, but whether it carries out the same role as 5mC methylation in these species is not known, as the functions of adenine methylation are still not well understood (Kyger *et al*, 2020).

DNA methyltransferases (DNMTs)

DNA methylation is carried out via C-5 cytosine-specific DNA methyltransferases (DNMTs). Three distinct DNMT proteins have been identified in animals; DNMT1, DNMT3a and DNMT3b (Bhattacharyya *et al*, 2020). TRDMT1 (referred to as DNMT2 in this report) has also been identified and shows strong similarities in structure to DNA methyltransferases. However, DNMT2 methylates RNA and not DNA (Kyger *et al*, 2021) and was included in the study as an outgroup to DNMT1 and DNMT3.

Both DNMT1 and DNMT3 have distinct roles with the DNA methylation mechanism. DNMT3a and DNMT3b are *de novo* DNA methyltransferases and are vital in establishing methylation patterns in early development. (Ren *et al*, 2018). DNMT1 is the more abundant of the DNA methyltransferases in mammalian cells and is responsible for maintaining the DNA methylation patterns as cells undergo division (Lyko F, 2017). DNMTs are well conserved across many species regardless of the methylation levels in the genome and several conserved domains have been identified (Figure 2).

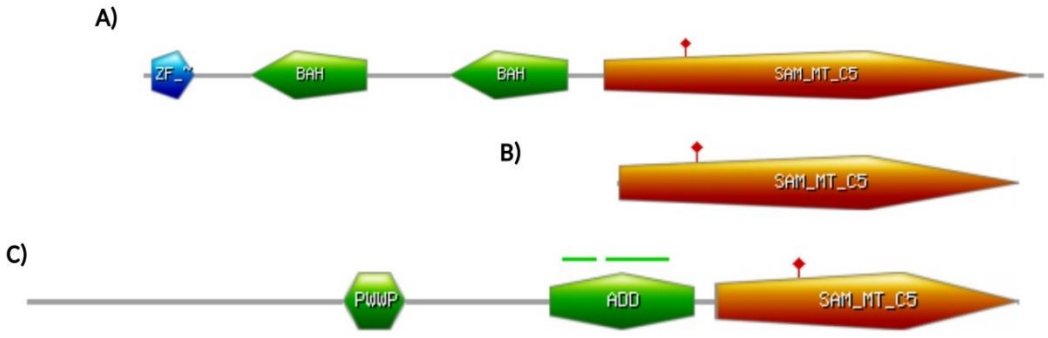


Figure 2: DNMT1 and DNMT3 domains. A) DNMT1 domains; DMAP_BI = DMAP binding domain, facilitates binding to DMAP which acts as a transcriptional co-repressor through its interaction with HDAC2. ZF = zinc-finger domain, a DNA binding motif. BAH = bromo-adjacent homology domain, though to facilitate protein-protein interactions. SAM_MT_C5 = C-5 cytosine specific DNA methylase domain, the red marker indicates the active site. **B)** DNMT2 domains: SAM_MT_C5 is as described above with the red marker indicating the active site. **C)** DNMT3 domains; PWWP = PWWP (Pro-Trp-Trp-Pro) domain, binds to H4K20me (histone-4 methylated at lysine 20). ADD = ADD domain (Atrx-DNMT3-DNMT3l), binds to H3K4me0 (histone-4 unmethylated at lysine 4). SAM_MT_C5 is as described above with the red marker indicating the active site.

In DNMT1, the DMAP1 binding domain is essential for the interaction between the DNMT and the DMAP1 (DNMT1-associated protein) which has been shown to act as a transcriptional co-repressor (Lyko, 2017). In mice embryos, the reaction between DMAP1 and DNMT1 has been shown to be essential to development (Naga Mohan *et al*, 2011). Also present in DNMT1 is the ZF domain, two BAH domains and the methyltransferase domain (Figure 2). In DNMT3, the PWWP domain has been shown to be essential in both DNMT3a and DNMT3b in guiding DNMTs to heterochromatin and has also shown the ability to bind to histones, specifically H4K20me (Chen *et al*, 2004). Similarly, the ADD domain also acts as a mechanism by which DNMT3 is guided to unmethylated histone tails such as H3K4me0 (Figure 2, Lyko *et al*, 2017). DNMT3 also has a well conserved methyltransferase domain (Figure 2).

Molecular Phylogeny

Although DNA methylation is considered to be a widespread epigenetic feature in animals, the level of CpG methylation varies amongst animal lineages. We do not yet fully understand the evolution of this epigenetic feature nor the machinery that facilitates DNA methylation (Liu *et al*, 2020). Molecular phylogenetics is the method of analysing genetic differences at the molecular level (within DNA or protein sequences) to gain insight into an organism's evolutionary history and relationships (Yang & Rannala, 2012). Molecular evolution occurs

through mutations that arise in DNA or protein sequences. Information on evolutionary relationships can be elucidated through the study of phylogenetic trees.

The creation of a phylogenetic tree begins with the generation of a multiple sequence alignment (MSA). This is typically a large-scale alignment of sequences that are assumed to have some form of evolutionary relationship (Thompson *et al,* 2011). Many algorithms exist to align sequences; two main categories being progressive and iterative methods. Progressive methods typically include two stages, the formation of a *guide tree*, and the consequent alignments of the sequences that is dictated by the guide tree (Yang & Rannala, 2012). This includes the ClustalO and MAFFT methods. Iterative alignment methods work in a similar way to progressive methods but are thought to be an improvement upon the latter as they repeatedly re-align sequences to one another to lend more support to the alignment (Yang & Rannala, 2012). An example of one such method is MUSCLE.

Phylogenetic trees can be inferred from the data generated from sequence alignments. The methods used to construct these trees can be either distance or character-based. Distance based methods involve the construction of a distance matrix based on the observed distances between each sequence. Whereas distance-based methods look at sequences as a whole, character-based methods compare each sequence simultaneously by comparing one character (a nucleotide base or an amino acid, for instance) at a time. Distance based methods such as maximum likelihood and Bayesian inference all use a substitution model; a computational means of describing evolutionary changes over time (Yang & Rannala, 2012). Several substitution models exist, with each assigning more or less importance to certain character substitutions depending on the type of data being used in the study. Examples of commonly used substitution models include the WAG and LG models, with the LG model being the more recent addition (Le & Gascuel, 2008).

Phylogenetic relationships of DNMT proteins in animals have previously described in multiple studies by de Mendoza *et al*, Kyger *et al*, Bhattacharyya *et al* and Liu *et al*. Figure 1 represents the list of species used in this study. Widely represented are the cnidarians (grouped by class). The hydrozoans such as *Hydractinia symbiolongicarpus* and *Hydra vulgaris* have not been included in previous studies and so they have been used here in an effort to understand the relationship of DNMTs in these species with other animals.

The aim of this study was to use various phylogenetic methods to conduct a study on the DNMT protein family. They are known to contain a well conserved methyltransferase domain and a

phylogenetic tree was created by focusing on this domain. It was also aimed to closely study the sequence conservation of the methyltransferase domain and the additional conserved domains found in DNMT1 and DNMT3.

Materials and Methods

Sequence Retrieval

Species to be used in this study are listed in Supplementary table 1. A selection of vertebrate species was included as DNA methylation is perhaps best understood in vertebrates at the present time and DNMT proteins are known to be well conserved within this group. In addition to these species, the aim was to include species from a broad range of lineages, though the scope of the study was small. Lophotrochozoan's and ecdysozoans were also included. Many cnidarian species (referred to here as anthozoans, hydrozoans and polypodiozoans) were also included. Though many studies have focused specifically on certain lineages, this project aimed to investigate as broad a range of lineages as possible.

DNMT sequences were obtained from various sources (Supplementary table 1) in preparation for sequence alignment. The DNMT sequences of *A.queenslandica*, *M. leidyi*, *S.ciliatum*, several cnidarians and myxozoans were retrieved from the supplementary datasets of previous reports (de Mendoza *et al*, 2019; Kyger *et al*, 2020). Further DNMTs were retrieved through a tBLASTn search using human DNMTs (retrieved from uniprot: P26358, O14717, Q9Y6K1).

The tBLASTn searches were performed with Blossum62, the E value was cut off at 0.05, and the maximum target sequence set at 50. The top hit sequence was brought to the Expasy website to translate the nucleotide sequences into amino acid sequences with the longest continuous amino acid sequence chosen for the alignment.

Multiple Sequence Alignment

Using the techniques outline above, sequences of DNMT1, DNMT2 and DNMT3 proteins were gathered from various species. DNMT1 and DNMT3 proteins were of interest for the study with DNMT2 being a suitable outgroup and means of rooting the tree. The sequences were compiled into a FASTA file with species name abbreviations as labels (H.sapi for *Homo sapiens*, Table 1). These sequences were imported into the SeaView program. In the ''Align'' menu of the SeaView program and under ''Alignment options'', the Muscle algorithm was selected. The ''Align all'' function was selected to run the alignment algorithm. An alignment using MAFFT was also generated and was later favored over MUSCLE for phylogenetic tree generation. The MAFFT alignment was created using MAFFT software, version 7 which is available online (https://mafft.cbrc.jp/alignment/software/). The FASTA file was uploaded to

the site. The E-INS-i method was used (recommended for less than 200 sequences with multiple conserved domains) and the BLOSUM62 matrix was used. The output file from MAFFT alignment was then saved and opened within the SeaView program. The alignment was saved and imported into the BioEdit program. With the ‘’edit’’ function enabled, the sequences were cut down to approximately 300 amino acids in length by highlighting large gaps and regions that were not conserved in all DNMT sequences. This was to cut the alignment down to the DNA methylase domain only and to avoid a computationally exhaustive dataset. This was then saved as a FASTA file to be used in the creation of a phylogenetic tree.

Phylogenetic Tree Generation

The multiple sequence alignment, now at approx. 300 amino acids was imported back into SeaView as described above. Under the ‘’Trees’’ menu, ‘’PhyML’’ was selected. The LG model was used, and bootstrapping was set at 1000.

Conserved Domain Analysis

Further study of the additional domains within DNMT1 and DNMT3 was completed by creating separate sequence alignments for both groups for easy analysis. This was completed using MAFFT as described above. These smaller alignments were easier to view and to observe amino acid substitutions. All DNMT1 and DNMT3 sequences were subjected to searches in the Pfam database, to note any loss of domains or additional domains. Searches were also performed on the NCBI CDD (conserved domain database).

To better observe the conserved motifs of the DNA methyltransferase domain, a sequence logo was created using a multiple sequence alignment of both DNMT1 and DNMT3. This was achieved by uploading the FASTA file of a multiple sequence alignment of DNMT1 and DNMT3 to the WebLogo website (www.weblogo.berkeley.edu). The only parameter adjusted was the number of residues displayed on each line to for visualization in this report.

Results

Phylogenetic Tree

The topology of the phylogenetic tree of DNMT shows the distinct clades of the three DNMT proteins with DNMT2 as an outgroup (Supplementary Figure 1). The support for these three groupings was strong with nodes for the DNMT1, DNMT2 and DNMT3 groupings having a maximum likelihood bootstra figure of >90%. (Supplementary Figure 1).

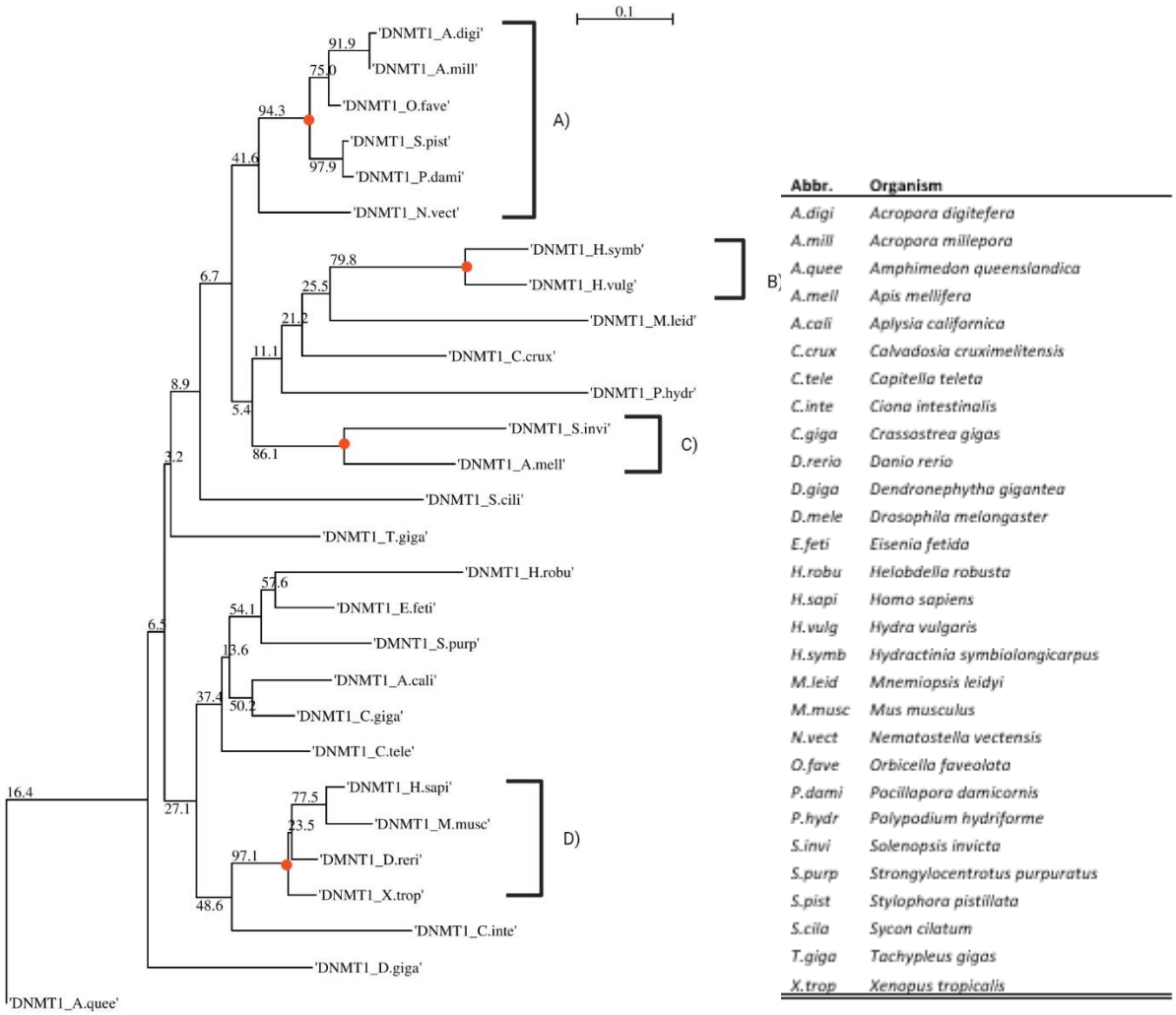


Figure 3: Phylogenetic subtree of DNMT1 proteins created using the maximum likelihood (ML) method. Red dots indicate nodes that were very well supported (>90%, with the exception of grouping B, which has a support figure of 79.8). **A)** Anthozoan DNMT1 homologs apart from *Dendronephytha gigantea*. **B)** Hydrozoan DNMT1 homologs. **C)** Ecdysozoan DNMT homologs apart from *Tachypleus giga* **D)** Vertebrate DNMT1 homologs. Species legend displayed on the right.

The DNMT1 clade was well supported (>95%) and was represented in a group that was distinct from DNMT2 and DNMT3. Within this clade, vertebrate homologs are well supported and

form one cluster (Fig 2, D). Hydrozoans DNMT1 sequences are also well supported within the tree (Fig 2, B). The majority of anthozoan DNMT1 homologs were also strongly supported in one group with *Nematostella vectensis* presenting as a sister group (Fig 2, A). However, *Dendronephytha gigantea* was not clustered with its fellow anthozoans. Another well supported cluster are the arthropods *Solenopsis invicta* and *Apis mellifera* (Fig 2, C).

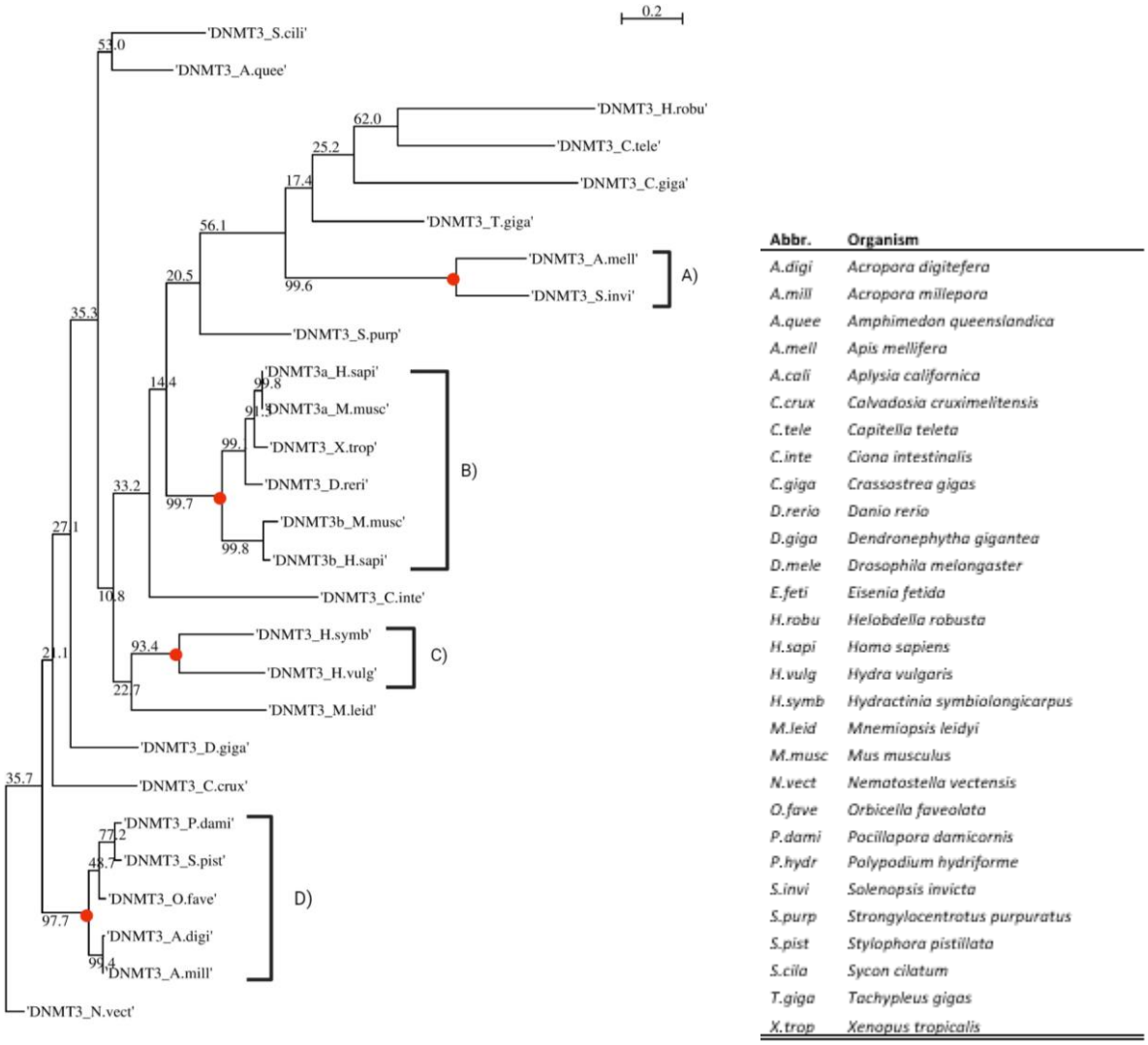


Figure 4: Phylogenetic subtree of DNMT3 proteins created using the maximum likelihood method (ML). Red dots indicate nodes that were very well supported (>90%). **A)** Ecdysozoan DNMT3 homologs apart from *Tachypleus gigas.* **B)** Vertebrate DNMT3 homologs (also DNMT3 paralogs DNMT3a and DNMT3b in *Mus musculus* and *Homo sapiens*). **C)** Hydrozoan DNMT3 homologs. **D)** Anthozoan DNMT3 homologs apart from *Dendronephytha gigantea.* Species legend displayed on the right.

Similary to the DNMT1 cluster, the DNMT3 cluster is well supported (>90%) and is distinct from DNMT1 and DNMT2 (Supplementary Figure 1). DNMT3a and DNMT3b for *H. sapiens* and *M. musculus* are displayed in two separate clades (Fig 3, B). DNMT3a and DNMT3b are

paralogs, meaning they both arise from a copy of a duplicated gene. The majority of anthozoan DNMT3 homologs are clustered together with good support (Fig 3, D). The ecdysozoan DNMT3 homologs of *S. invicta* and *A. mellifera* are also in a well supported cluster (Fig 3, A). The hydrozoan homologs have also clustered together and are well supported and show an increased support figure to the DNMT1 homologs (Fig 3, C, Fig 2, B).

Domain Analysis

DNA methyltransferase domain

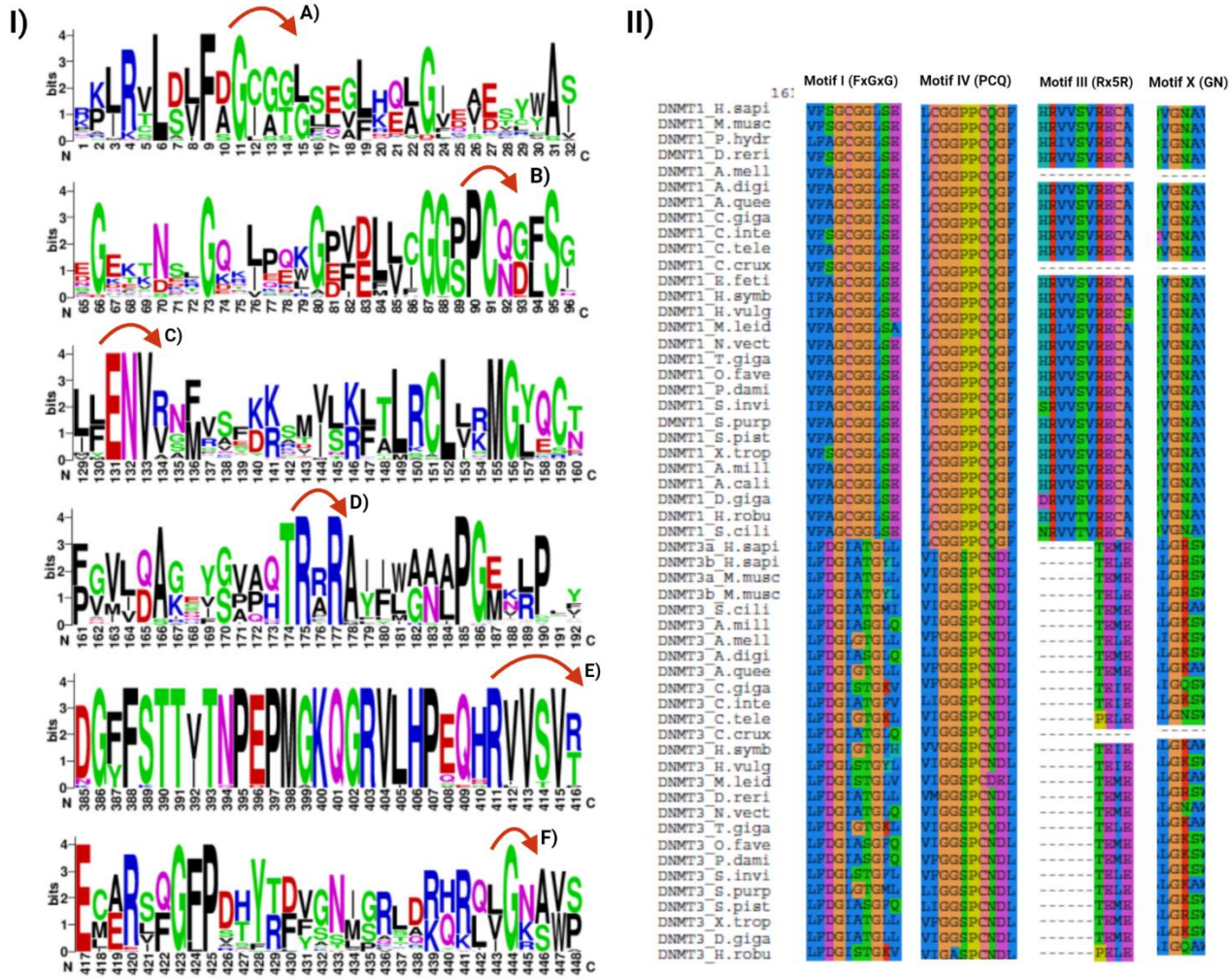


Figure 5: Conserved motifs in DNA methylase domain of DNMT1 and DNMT3. These motifs are known to be well conserved within the DNA methyltransferase domain (Bhattacharyya *et al*, 2020) **I)** Sequence logo focusing on the DNA methylase domain of DNMT1 and DNMT3 alignment. **a)** Motif I (FxGxG) **b)** Motif IV (PCQ) **c)** Motif VI (ENV) **d)** Newly observed motif (RxR).**e)** Motif III (Rx5R) **f)** Motif X (GN) **II)** Motif I, IV, III and X as viewed in the multiple sequence alignment created with MAFFT.

Along with generation of the phylogenetic trees, in depth domain analysis was completed on the DNA methylase domain. The sequence logo generated for this domain in DNMT1 and

DNMT3 can be seen above (Figure 5). All DNMT proteins are known to have a highly conserved catalytic domain and within this domain there are several well catagorised motifs (Bhattacharrya *et al*, 2020). The sequence logo was used as a means to observe these motifs within the species used in this study. Although there are ten motifs that have been catagorised in the catalytic domain of DNMT proteins, there are some that are only present in archaea and bacteria (Bhattacharrya *et al*, 2020). 6 motifs were clearly identified from the MSA in this study and have been labelled on the sequence logo. Motif I (FxGxG)(phenylalinine x glycine x glycine) (Figure 5, a)) was fully conserved among all DNMT1 sequences in the study, however this was not the case for DNMT3. While the FxG amino acid sequence was conserved among all DNMT3 sequences, the third amino acid in the motif varies (Figure 5, II)). The majority of the species had an alanine in this position. Five species had a serine residue in this position and six species retained the FxGxG motif that was observed in the DNMT1 proteins, although this was not observed in any one lineage and was scattered throughout. Motif IV (PCQ) (proline, cysteine, glutamine) was fully conserved among DNMT1 proteins but the glutamine in this motif has been substituted for an asparagine (N) residue in all DNMT3 proteins in the study (Figure 5, b)) (Figure 5 II)). Motif VI (ENV)(glutamate, asparagine, valine) was fully conserved across all protein sequences in the study, and this is clearly shown in the sequence logo (Figure 5, c)). Motif VIII (Rx_5E)(Arginine, glutamate)(Figure 5, d)) was fully conserved across all DNMT1 sequences but was completely absent from *Apis mellifera* and *Calvadosia cruxmelitensis* due to their DNMT1 sequences being considerably shorter than other DNMT1 protein (Figure 5, II)) This motif is not conserved among DNMT3 proteins with the exception of the glutamate residue which remained conserved in all sequences with the exception of *Calvadosia cruxmelitensis*. Motif X (GN) (Figure 5, e)) was fully conserved in DNMT1 sequences with the exception of *Calvadosia crusmelitensis* and *Apis mellifera*. In DNMT3 sequences however, the glycine residue was replaced by either a lysine or arginine residue in the majority of sequences (Figure 5, II)) Only *Nematostella vectensis* and *Capitella teleta* retained the GN motif as observed in DNMT1 sequences. *Calvadosia cruxmelitensis* DNMT3 lacked this motif altogether. A new motif as identified in Bhattacharyya *et al's* publication was also identified within this study (Bhattacharyya *et al*, 2020). This RxR motif was fully conserved in all sequences (Figure 5, f)

Domain Analysis

Further domain analysis, aside from the DNA methyltransferase domain, was completed after the phylogenetic trees had been generated. Although the trees were created by focusing on the DNA methylase domain, DNMT proteins contain many other conserved domains that are important for their function (See Fig 1-introductory figure of DNMT domains). As outlined in the Methods section of the report, each sequence was searched using the Pfam database and also the NCBI CDD. A chart visualising these findings is shown below (Figure 6).

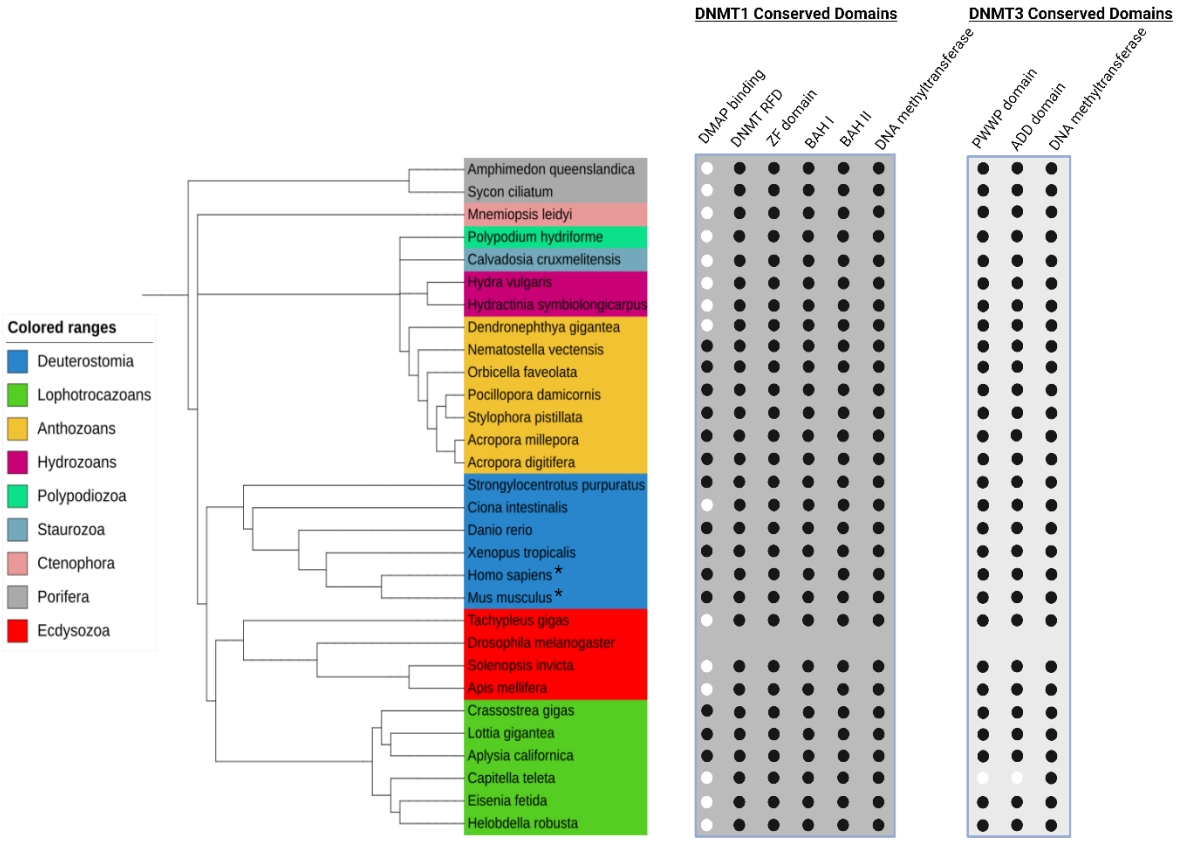


Figure 6: DNMT1 and DNMT3 domain conservation. *The cladogram shown here has been created using phyloT (www.phylot.biobyte.de) an online tree visualisation program that generates trees based on NCBI taxonomy. It is not a representation of the work carried out in this study and is used here simply for display purposes.* Both DNMT1 and DNMT3 are well conserved overall. However there is a loss of some domains. *Both Homo sapiens and Mus musculus included the DNMT3a paralog which is not displayed in the graph but was not missing any conserved domains.

The cladogram on the left represents the species used in the study and clearly displays the different lineages with color coded sections. Both DNMT1 and DNMT3 are well conserved overall. However, in DNMT1 there is a loss of the DMAP binding domain in some species

although there is no pattern for this loss except in the ecdysozoan lineage. The DMAP binding domain binds DMAP1 (DNMT associated protein) which acts as a transcriptional co-repressor through its interaction with histone deacetylase 2 (HDAC2). It is lost in approximately 50% of the species used in the study, and even in those species in which it remains, the sequence is not particularly well conserved with little, if any, residues fully conserved among all species. DNMT1 has retained conservation of all other domains with the RFD (replication foci domain) conserved among all species in the study and the same can be seen for both BAH domains and the zinc finger domain. Additionally, a region of Glycine-Lysine (GK) repeats that is not part of any noted conserved domain is shown to be extremely well conserved in DNMT1 sequences (Supplementary figure 2). DNMT3 shows much better conservation than DNMT1 although there is a loss of the ADD domain in some arthropods.

What is not displayed in the chart above is the additional discovery of some DNMT3 sequences that contained an additional PWWP domain. This double PWWP domain was only observed in *Hydractinia symbiolongicarpus, Hydra vulgaris* and *Solenopsis invicta*. Both *Hydractinia* and *Hydra* are hydrozoan cnidarians, however, *Solenopsis* is from the ecdysozoan lineage and this additional PWWP domain was not observed in its fellow ecdysozoans (Supplementary Figure 3)

Discussion

From the phylogenetic trees (Fig 3 and Fig 4) we can see that the relationships between vertebrate DNMT's was very well conserved and very well supported. This is in line with previous studies that have demonstrated that DNMT machinery and DNA methylation is highly conserved in vertebrates (Liu *et al*, 2020). *Hydractinia*, a species that has not been included in previous studies, was grouped with its fellow hydrozoan *Hydra vulgaris* for both DNMT1 and DNMT3. The anthozoans were also grouped together and well supported with the exception of *Dendronephytha* in both trees.

In the tree depicting DNMT3 (Fig 4), we can see that the paralogs for DNMT3a and DNMT3b for the human and mouse proteins have been grouped distinctly. It is a widely accepted theory that 2 rounds of whole genome duplication (WGD) occurred in vertebrates, although the timing of both rounds is debated. In a study performed by Liu *et al*, a large-scale phylogenetic analysis of chordates was performed. In this study, their proposed evolutionary history of DNMT3 suggests that the mutation resulting in the paralogs DNMT3a and DNMT3b occurred after the first round of WGD. It has also been hypothesized that the hypermethylated genome state observed in vertebrates was in response to this genome duplication event, as a compensation for the imbalance of genetic material. This theory has been somewhat disproven by the discovery of the hypermethylated genome of *Amphimedon queenslandica* as this species did not undergo a WGD event.

Looking at the DNMT1 and DNMT3 subtrees (Fig 3 and Fig 4), *Amphimedon* is shown to branch off much earlier than the vertebrate cluster, first shown as the rooted outgroup for DNMT1s and then clustered with its fellow porifera *Sycon ciliatum* in the DNMT3. In the de Mendoza study on a similar group of species, *Amphimedon* was also rooted as an outgroup for both DNMT1 and DNMT3 and much earlier branching that vertebrate species. This would suggest that *Amphimedons'* hypermethylated genome has no correlation to the evolutionary relationships of the methyltransferase domains in DNMT1 and DNMT3 between *Amphimedon* and vertebrates. Instead, this could demonstrate an example of convergent evolution. A means in which two unrelated species may develop similar phenotypic or genotypic traits. These shared traits do not arise as a result of diverging genes, as most evolutionary changes are thought to occur, but instead can arise because of similar environmental pressures on the organism (Stern, 2013).

Aside from the well supported clusters of vertebrates, anthozoans, ecdysozoans and hydrozoans, many other branches are not well supported enough to be able to derive any significant hypotheses but this is to be expected as vertebrates are evolutionarily quite distant from the early branching lineages such as the cnidarians. Despite the varying support levels in both the DNMT1 and DNMT3 trees, the domain analysis of the catalytic methyltransferase domain showed a high level of conservation in the various motifs throughout the domain. Of particular interest was the ENV domain which was completely conserved (this motif is located within the active site of the enzyme) and the newly observed motif RxR which has only recently been characterised in literature (Bhattacharyya, 2020).

Further domain analysis of the DNMT1 and DNMT3 proteins gave a closer insight to the additional conserved domains found in the proteins (Fig 1). Most notably was the apparent loss of the DMAP1 binding domain in some species. However, rather than these being incomplete DNMT1 sequences, it is likely that they are actually an isoform of DNMT1 that is expressed solely in oocytes, referred to as DNMTo (Hirasawa *et al*, 2008). This protein is around 118 residues shorter than its somatic counterpart. This oocyte-specific form of DNMT1 maintains methylation marks in pre-implantation embryos. DNMT1o is of particular interest as methylation patterns in the early development stages of an organism likely have a profound effect on the future methylation patterns of the cell (Hirasawa *et al,* 2008). Methylation levels during embryonic development are known to rise and fall. Typically, after fertilisation, a major reprogramming even occurs in mammals to erase gametic methylation pattern. This is essential to restore cells to a pluripotent state as the organism undergoes development (Yang & Chen, 2019). DMAP1 is a member of the TIP60-p400 complex which plays an essential role in embryonic stem cells and also interacts with HDAC2, a histone modification protein. Studies in mice have shown that embryos lacking the DMAP1 protein are not viable, showing that DMAP1 is essential to early development (Naga Mohan *et al,* 2011). This raises the question as to how DNMT1o interacts with DMAP1 if there is no DMAP1 binding domain present on the protein. The same study conducted on the mice embryos suggested a novel interaction between the two proteins that has yet to be elucidated (Naga Mohan *et al*, 2011).

Given that the BLAST searches were carried out using DNMT1 from *H. sapiens* as the query sequence, which contains a DMAP1 domain, it is possible that the shotgun assemblies found during the search are merely displaying the oocyte-specific isoform of DNMT1. Aside from the DMAP1 domain being absent in some sequences, all other domains were relatively well conserved. In DNMT3, all domains were well conserved apart from *Capitella teleta*, however

for this species, sequences were retrieved from Ensembl Metazoa site (ensembl.metazoa.org) and the lack of PWWP and ADD domains may be due to incomplete sequencing data rather than an evolutionary loss.

Finally, three species were shown to have an additional PWWP domain when subjected to a CDD search. When the sequences were searched through PFAM, which was the initial website of choice for domain analysis, these additional sites did not register. The additional sites were found in *Hydractinia*, *Hydra* and *Solenopsis* and have not been discussed in many of the recent studies on this subject. However, whether these additional sites have a purpose or whether one site is catalytically inactive is unknown.

Conclusion

To summarise the report, a multiple sequence alignment of multiple DNMT protein sequences from various animal lineages was completed followed by the generation of a phylogenetic tree and domain analysis. The findings in this study confirm that regardless of the genomic size of a species or its methylation levels, DNA methylation machinery has been well conserved. To infer further information on the evolutionary relationships between species and to increase the support of the phylogenetic trees, a larger scale study would have to be completed, with an increased number of species and perhaps even the inclusion of bacteria, fungi and plants as seen in some studies. Currently, the full function of DNA methylation in both invertebrates and vertebrates remains to be understood.

Although the scope of this study was small in comparison to previous work, we can take the species used in the study to be representatives of their animal groups and thus our findings further confirm that DNMT machinery is well conserved across multiple animal lineages. Consequently, the motifs of the methyltransferase domain also displayed a good level of conservation, even between vertebrates and invertebrates. During the domain analysis, some DNMT1 sequences were shown to be missing the DMAP1 binding domain which may be explained by the presence of a DNMT1 isoform DNMT1o that is abundant in oocytes. Furthermore, the relationship of DNMTs from the hydrozoan *Hydractinia symbiolongicarpus* with other animal species was elucidated, with additional PWWP sites being found in *Hydractinia* and *Hydra*, suggesting that DNA methylation in these species may be more complex than originally thought and is worthy of further study.

References

Bhattacharyya M, De S, Chakrabarti S (2020) Origin and Evolution of DNA methyltransferases (DNMT) along the tree of life: A multi-genome survey. *bioRxiv*: 2020.2004.2009.033167

Brusca, Richard C (2016) Introduction to the Bilateria and the Phylum Xenacoelomorpha: Triploblasty and Bilateral Symmetry Provide New Avenues for Animal Radiation . *Invertebrates Sinauer Associates* : 345–372

Chen T, Tsujimoto N, Li E (2004) The PWWP domain of Dnmt3a and Dnmt3b is required for directing DNA methylation to the major satellite repeats at pericentric heterochromatin. *Mol Cell Biol* 24: 9048-9058

de Mendoza A, Hatleberg WL, Pang K, Leininger S, Bogdanovic O, Pflueger J, Buckberry S, Technau U, Hejnol A, Adamska M *et al* (2019) Convergent evolution of a vertebrate-like methylome in a marine sponge. *Nat Ecol Evol* 3: 1464-1473

Hirasawa R, Chiba H, Kaneda M, Tajima S, Li E, Jaenisch R, Sasaki H (2008) Maternal and zygotic Dnmt1 are necessary and sufficient for the maintenance of DNA methylation imprints during preimplantation development. *Genes Dev* 22: 1607-1616

Iyer LM, Zhang D, Aravind L (2016) Adenine methylation in eukaryotes: Apprehending the complex evolutionary history and functional potential of an epigenetic modification. *Bioessays* 38: 27-40

Kyger R, Luzuriaga-Neira A, Layman T, Milkewitz Sandberg TO, Singh D, Huchon D, Peri S, Atkinson SD, Bartholomew JL, Yi SV *et al* (2020) Myxosporea (Myxozoa, Cnidaria) Lack DNA Cytosine Methylation. *Molecular Biology and Evolution* 38: 393-404

Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. *Mol Biol Evol* 25: 1307-1320

Liu J, Hu H, Panserat S, Marandel L (2020) Evolutionary history of DNA methylation related genes in chordates: new insights from multiple whole genome duplications. *Scientific Reports* 10: 970

Lyko F (2018) The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nature Reviews Genetics* 19: 81

Mohan KN, Ding F, Chaillet JR (2011) Distinct roles of DMAP1 in mouse development. *Mol Cell Biol* 31: 1861-1869

Moore LD, Le T, Fan G (2013) DNA methylation and its basic function. *Neuropsychopharmacology* 38: 23-38

Ren W, Gao L, Song J (2018) Structural Basis of DNMT1 and DNMT3A-Mediated DNA Methylation. *Genes (Basel)* 9

Ruggiero MA, Gordon DP, Orrell TM, Bailly N, Bourgoin T, Brusca RC, Cavalier-Smith T, Guiry MD, Kirk PM (2015) A Higher Level Classification of All Living Organisms. *PLOS ONE* 10: e0119248

Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. *Nature Reviews Genetics* 14: 204-220

Stern DL (2013) The genetic causes of convergent evolution. *Nature Reviews Genetics* 14: 751-764

Thompson JD, Linard B, Lecompte O, Poch O (2011) A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One* 6: e18093

Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* 13: 303-314

Zeng Y, Chen T (2019) DNA Methylation Reprogramming during Mammalian Development. *Genes (Basel)* 10

All figures in this report were created or edited with BioRender at BioRender.com

Acknowledgements

I would like to acknowledge colleagues and staff at NUI Galway for their assistance and contributions for the duration of the research project. I offer sincere thanks to my lab supervisor, Febri Marsa, whose patience and guidance was vital in the completion of this project. A further thank you to my supervisor, Professor Uri Frank, for his valued input and support. I would also like to extend gratitude to further members of the Frank Lab, Gabriel Krasovec and Helen Horkan for lending me their time and expertise.

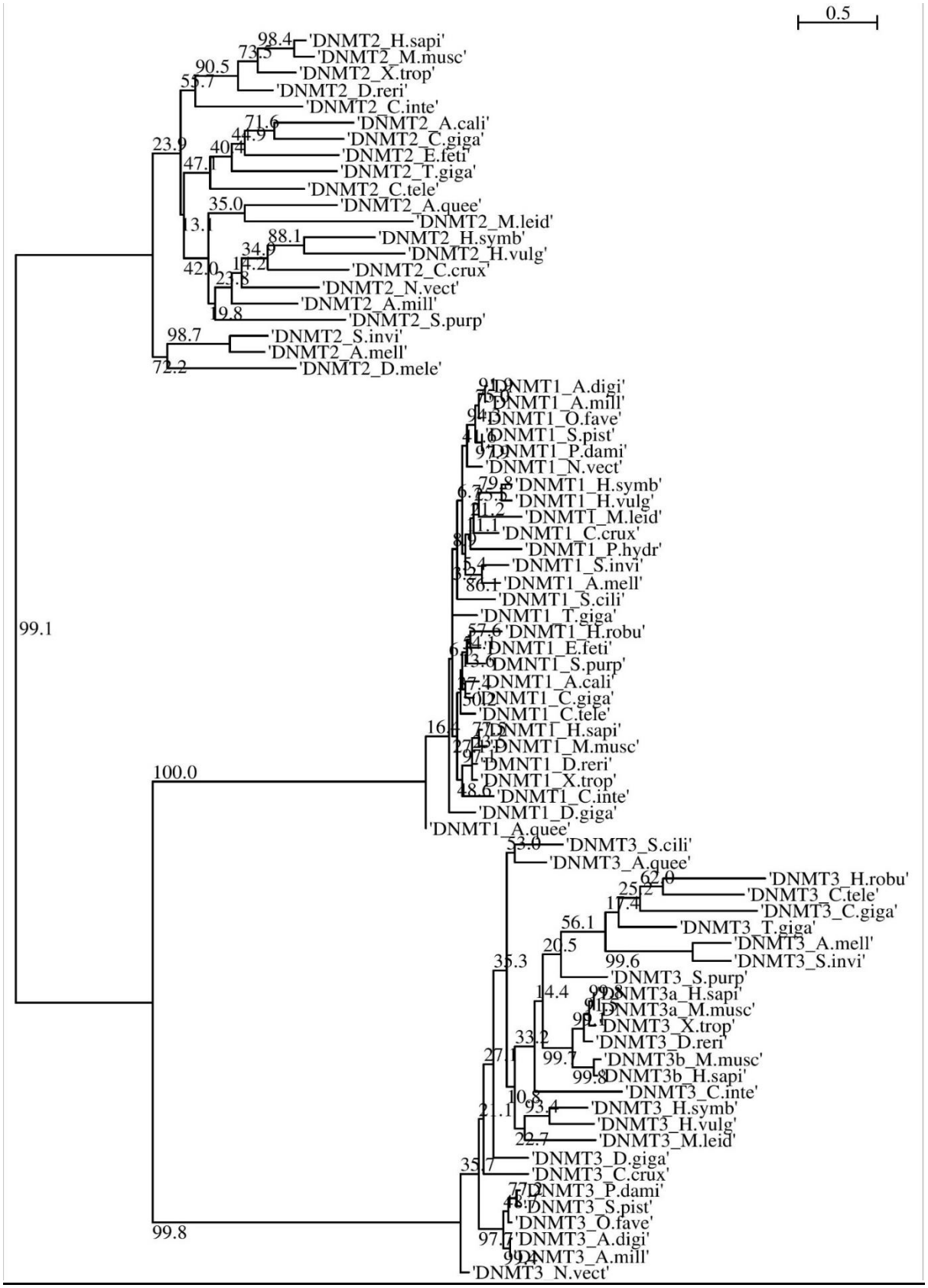
Supplementary Material

FASTA files and multiple sequence alignments are available on GitHub:
https://github.com/KathrynRuddle/NUIG-Research-Project.git

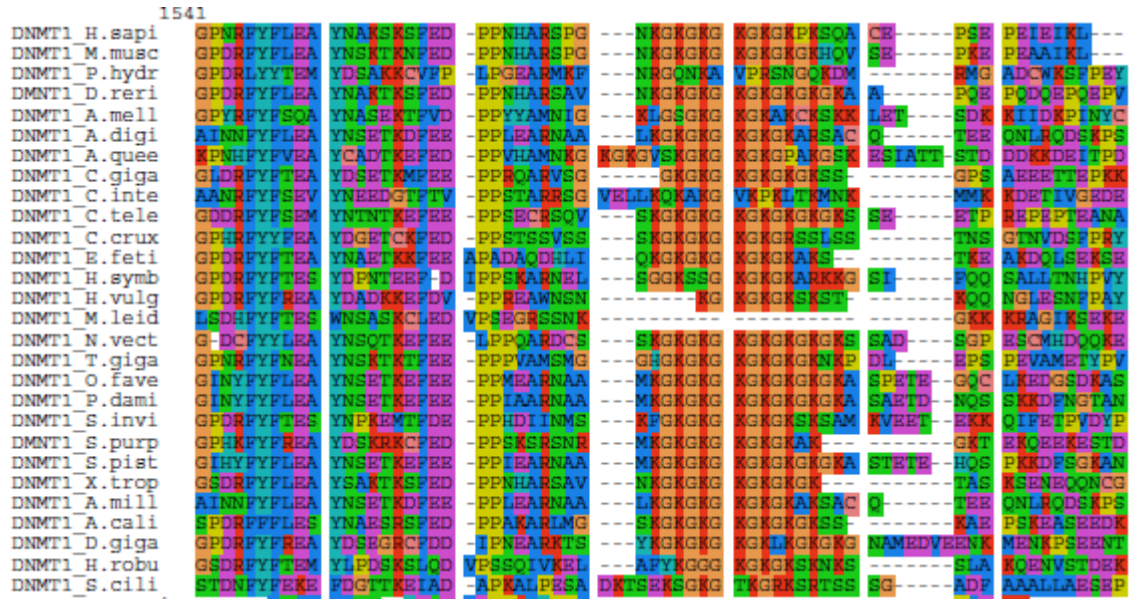
Supplementary table 1: Accession numbers and names of species used in this study.

		GenBank Accession Numbers			
Abbr.	**Organism**	**DNMT1**	**DNMT2**	**DNMT3**	
A.digi	*Acropora digitefera*	XP_015768114.1		XP_015768114.1	
A.mill	*Acropora millepora*	XP_029197815.1	XP_029199407	XP_029213858	
A.quee	*Amphimedon queenslandica*	XP_011402705.2		XP_019851113	
A.mell	***Apis mellifera***	GALO01024511.1	GALO01043524.1	HP486145.1	
A.cali	*Aplysia californica*	XP_005095276.1	XP_012941472		
C.crux	***Calvadosia cruximentensis***	HAHC01039510.1	HAHC01025640.1	HAHC01051751.1	
C.tele	***Capitella teleta***	CapteT160905**		CapteT1762**	
C.inte	***Ciona intestinalis***	GBKV01004999.1	GBKV01013876.1	XM_026838821.1	
C.giga	***Crassotea gigas***	GECI01034654.1	GECI01003928.1	GECI01029944.1	
D.rerio	*Danio rerio*	NP_571264.2	NP_001018153	NP_001018150	
D.giga	*Dendronephytha gigantea*	XP_028404797.1		XP_028409439	
D.mele	*Drosophila melongaster*		XP_015768114.1		
E.feti	***Eisenia fetida***	GIUK01119997.1	GIKG01013164.1		
H.robu	*Helobdella robusta*	XP_009029810		XP_009023046	
H.sapi	*Homo sapiens*	NP_001370.1	XP_005252431.1	ENSG00000119772	AAL57040.1
H.vulg	*Hydra vulgaris*	XP_012557244.1	XP_002166687.2	XP_012561137.1	
H.symb	***Hydractinia symbiolongicarpus***	GCHW01016739.1	GAWH01036605.1	GAWH01011216.1	
M.leid	***Mnemiopsis leidyi***	GSE124016*	GFAT01058962.1	GSE124016*	
M.musc	*Mus musculus*	AAH53047.1	NP_034197.3	NP_001258682.1	NP_001258676
N.vect	***Nematostella vectensis***	GSE124016*	XP__001624654	GSE124016*	
O.fave	*Orbicella faveolata*	XP_020612302.1		XP_020616882.1	
P.dami	*Pocillapora damicornis*	XP_027052738.1	XP_027057288	XP_027053693	
P.hydr	***Polypodium hydriforme***	LJ459796.1		GFUX01050926.1	
S.invi	***Solenopsis invicta***	LJ459796.1	LJ632846.1	LI689523.1	
S.purp	***Strongylocentrotus purpuratus***	XP_011667182.1	GHFM01010822.1	XP_030840512	
S.pist	***Stylophora pistillata***	XP_022779249.1	GARY01014085.1	XP_022794499	
S.cila	*Sycon cilatrum*	GSE124016*		GSE124016*	
T.giga	***Tachypleus gigas***	GILM01006715.1	GILM01009238.1	GILM01010037.1	
X.trop	*Xenopus tropicalis*	XP_002934384	NP_001004959	XP_004919668,	

Bold text represents species where one or more sequences were estimated using a tBLASTn search
***These sequences were accessed via GeoBank from the de Mendoza paper**
****As capitella teleta DNMT3 could not be found with a BLAST search, they were accessed via the Ensembl Metazoa site (http://metazoa.ensembl.org/)**

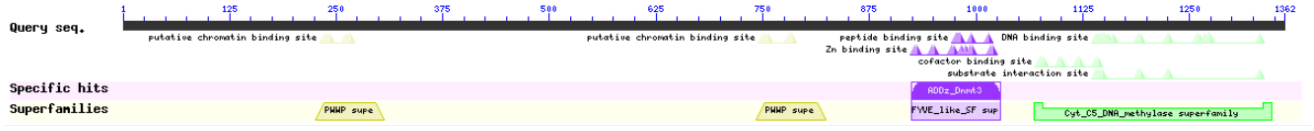


Supplementary figure 1: Full phylogenetic tree generated using the maximum likelihood method with support values shown. DNMT proteins were shown to group into three distinct clades; DNMT1, DNMT2 and DNMT3. These are all well supported with support figures of >99%. This highlights the high level of conservation of DNMT1, DNMT2 and DNMT3 homologs as we do not see clustering by species but rather by protein type.

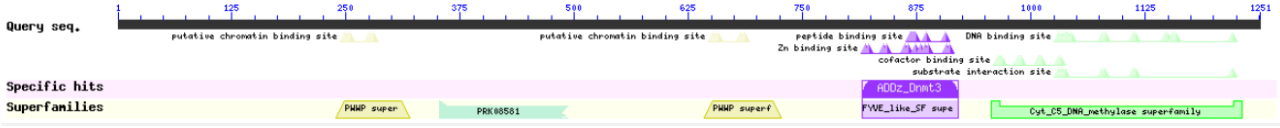


Supplementary figure 2: Region of GK repeats in DNMT1 sequences, absent only in *Mnemiopsis leidyi*. This repeat section is not mentioned in the literature but does not reside in the DNA methylase domain and may simply be a structural element.

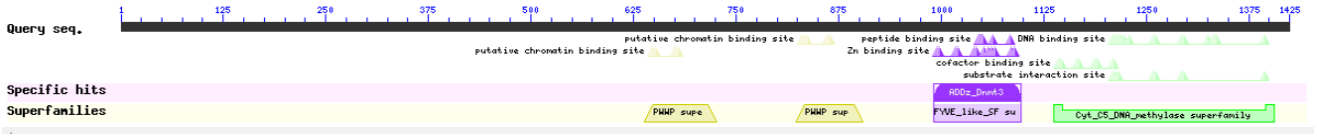
A



B



C



Supplementary Figure 3: Additional PWWP domains as viewed in NCBI CDD search. A) Additional PWWP domain seen in *Hydractinia symbiolongicarpus*. **B)** Additional PWWP domain as seen in *Hydra vulgaris*. **C)** Additional PWWP domain as seen in *Solenopsis invicta*. PWWP domains facilitate binding to methylated histone tails. Whether there is a purpose for an additional domain or one is simply inactive is unknown.

BI453 Research Report

Plagiarism Declaration

This form must be signed by the student, and must be included in the final submitted research report.

Student name: Kathryn Ruddle

ID number: 16305036

Report title: An investigation into the phylogeny and domain conservation of the DNMT protein family.

Supervisor: Professor Uri Frank

I hereby certify that this report is entirely my own work, and that the contents of this essay have not been published elsewhere in either paper or electronic form unless indicated otherwise through referencing.

Kathryn Ruddle 24/04/2021

Signature Date

The NUI Galway official Student Code of Conduct and the Code of Practice for dealing with plagiarism is available at http://www.nuigalway.ie/plagiarism/