Evolution of DNA methylation across Metazoa

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischens Grades

Doctor rerum naturalium
(Dr. rer. nat.)

im Fachgebiet
Informatik
Vorgelegt von

Master of Science in Informatik Jan Engelhardt
geboren am 12.05.1987 in Borna

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Peter F. Stadler, Universität Leipzig, Deutschland
2. Prof. Dr. Daniel Gautheret, CNRS, CEA, Université Paris-Sud, Frankreich

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 16.04.2021 mit dem Gesamtprädikat magna cum laude.

Contents

Abstract

DNA methylation is a crucial, abundant mechanism of gene regulation in vertebrates. It is less prevalent in many other metazoan organisms and completely absent in some key model species, such as *D. melanogaster* and *C. elegans*. In this thesis we report on a comprehensive study of the presence and absence of DNA methyltransferases (DNMTs) in 138 Ecdysozoa covering Arthropoda, Nematoda, Priapulida, Onychophora, and Tardigrada. We observe that loss of individual DNMTs independently occured multiple times across ecdysozoan phyla. In several cases, this resulted in a loss of DNA methylation.

In vertebrates, however, there is no single species known which lost DNA methylation. Actually, DNA methylation was greatly expanded after the 1R/2R whole genome duplication (WGD) and became a genome-wide phenomena. In our study of vertebrates we are not looking for losses of DNA methyltransferases and DNA methylation but are rather interested in the gain of additional DNA methyltransferase genes. In vertebrates there were a number of WGD. Most vertebrates only underwent two WGD but in the teleost lineage a third round of WGD occured and in some groups, e.g. Salmoniformes and some Cypriniformes even a forth WGD occured. The Carp-specific WGD (4R) is one of the most recent vertebrate WGD and is estimated to have occured 12.4 mya. We performed the most comprehensive analysis of the evolution of DNA methyltransferases after vertebrate whole-genome duplications (WGD) so far. We were able to show that the conservation of duplicated DNMT3 genes in Salmoniformes is more diverse than previously believed. We were also able to identify DNA methyltransferases in Cypriniformes which have, due to their recent WGD, quite complex genomes. Our results show that the patterns of retained and lost DNA methyltransferases after a forth round of WGD differ between Cypriniformes and Salmoniformes. We also proposed a

new nomenclature for teleost DNMT genes which correctly represents the orthology of DNMT genes for all teleost species.

Next to these purely computational projects we collaborated with the Aluru lab to investigate the effects of different disturbances on zebrafish DNA methylation. One disturbance is the inactivation of DNMT3aa and DNMT3ab as single knockouts as well as a double knockout. This was the first double knockout of DNMT genes in zebrafish which was ever generated. It allows us to study the subfunctionalization of the two DNMT3a genes their effect on genome-wide DNA methylation. Given our results we hypothesize that DNMT3aa and DNMT3ab can compensate for each other to a high degree. DNMT3a genes have likely been subfuntionalized but their loss can be compensated by DNMT3b genes. This compensation by DNMT3b genes works well enough that no notable phenotype can be observed in double knockout zebrafish but a difference is notable on the epigenome level. The second disturbance we studied is the exposure of zebrafish to the toxic chemical PCB126. We detected a moderate level of DNA methylation changes and a much larger effect on gene expression. Similar to previous reports we find little correlation between DNA methylation and gene expression changes. Therefore, while PCB126 exposure has a negative effect on DNA methylation it is likely that other gene regulatory mechanisms play a role as well, possibly even a greater one.

How do genes evolve and how are genes regulated are two of the main questions of modern molecular biology. In this thesis we have tried to shed more light on both questions. we have broadly expanded the phylogenetic range of species with a manually curated set of DNA methyltransferases. We have done this for ecdysozoan species which have lost all DNA methylating enzymes as well as for teleost fish which acquired more than ten copies of the, originally, two genes. We were also able to generate new insight into the subfunctionalization of the DNA methylation machinery in zebrafish and how it reacts to environmental effects.

Acknowledgments

First of all I would like to thank Peter Stadler for supervising me during this thesis.
Sonja Prohaska and all my other collaborators for working together with me on many different research projects.
Prof. Shoji Tajima, Prof. Mansi Srivastava, Neel Aluru, PhD and Josh Rosenthal, PhD for inviting me to join their laboratories as a guest during my PhD.
A large number of patient colleagues for keeping me motivated during the writing of this thesis. Special thanks goes to Petra Pregel and Jens Steuck for never ending administrative support.
My whole family for supporting me during my PhD.
The Joachim Herz foundation for granting me an Add-On fellowship.

Bibliographic description

Title:	Evolution of DNA methylation across Metazoa
Type:	Dissertation
Author:	Jan Engelhardt
Year:	2020
Professional discipline:	Computer Science
Language:	English
Pages in the main part:	94
Chapter in the main part:	6
Number of Figures:	24
Number of Tables:	29
Number of Appendices:	1
Number of Citations:	128
Key words:	DNA methylation, ecdysozoa, vertebrata, toxicology

This thesis is based on the following publications.

- Takahashi, S., Suetake, I., **Engelhardt, J.**, & Tajima, S. (2015). A novel method to analyze 5-hydroxymethylcytosine in CpG sequences using maintenance DNA methyltransferase, DNMT1. FEBS open bio, 5, 741-747.

- Akay, A., Di Domenico, T., Suen, K. M., Nabih, A., Parada, G. E., Larance, M., ..., **Engelhardt, J.**, ... & Rudolph, K. L. (2017). The helicase Aquarius/EMB-

4 is required to overcome intronic barriers to allow nuclear RNAi pathways to heritably silence transcription. Developmental cell, 42(3), 241-255.

- Fallmann, J., Will, S., **Engelhardt, J.**, Grüning, B., Backofen, R., & Stadler, P. F. (2017). Recent advances in RNA folding. Journal of biotechnology, 261, 97-104.

- Richards, C. L., Alonso, C., Becker, C., Bossdorf, O., Bucher, E., Colomé-Tatché, M., Durka, W., **Engelhardt, J.**, ... & Grosse, I. (2017). Ecological plant epigenetics: Evidence from model and non-model species, and the way forward. Ecology letters, 20(12), 1576-1590.

- **Engelhardt, J.**, Scheer O., Stadler P.F., & Prohaska S.J. Evolution of DNA methylation across Ecdysozoa. (*Currently in transfer to "Genome Biology and Evolution"*)

- **Engelhardt, J.**, Prohaska S.J., & Stadler P.F. Evolution of DNA methyltransferases after vertebrate of whole genome duplications. (*in preparation*)

- **Engelhardt, J.**, Karchner S.I., & Aluru N. Role of DNA methylation in altered testis gene expression patterns in zebrafish exposed to Pentachlorobiphenyl (PCB 126). (*in preparation*)

- **Engelhardt, J.**, Karchner S.I., & Aluru N. Alterations of DNA methylation after knockout of DNMT3aa and DNMT3ab in zebrafish. (*in preparation*)

Chapter 1

Introduction

1.1 Biological introduction

Gene regulation and epigenetics

In animal genomes thousands of genes are stored, see Figure 1.1 for a small overview. They contain the information which is necessary to produce RNA and, subsequently, protein molecules. RNA polymerases perform the act of transcription during which the information stored in the DNA is converted into the respective RNA molecule. Some of these RNA molecules, e.g. tRNAs, rRNAs and many more, have a function on their own. Messenger RNAs (mRNAs) on the other transports the information stored in the DNA to the cellular compartments which produce proteins, this process is called translation. The pathway from DNA to proteins via transcription and translation is long known in molecular biology and even called its central dogma [1]. The function of RNA and protein genes is currently the focus of many research projects. For many genes it is still unknown how its expression impacts a cell. Nevertheless, an equally important research topic is the regulation of these genes. The main protagonists of transcription are RNA polymerases. They bind to the DNA in front of the gene, the promoter region and subsequently read through the DNA. While they are doing so each "letter" in the DNA is transcribed to a "letter" in the RNA. However, there must be some kind of regulation which gene should be transcribed and which not. Otherwise, if, for example all genes would be transcribed or a random subset, a proper functioning of the cell would not be possible. A gene-specific mechanism to regulate the expression of a gene is the fact that many genes require additional factors to start their transcription. These factors are DNA-binding proteins, so called, transcription factors. They bind a specific region on the DNA, which is therefore called regulatory element, and if they are present they take part in regulating the respective gene. If they are increasing the expression of the gene, the regulatory element is called enhancer. If they decrease

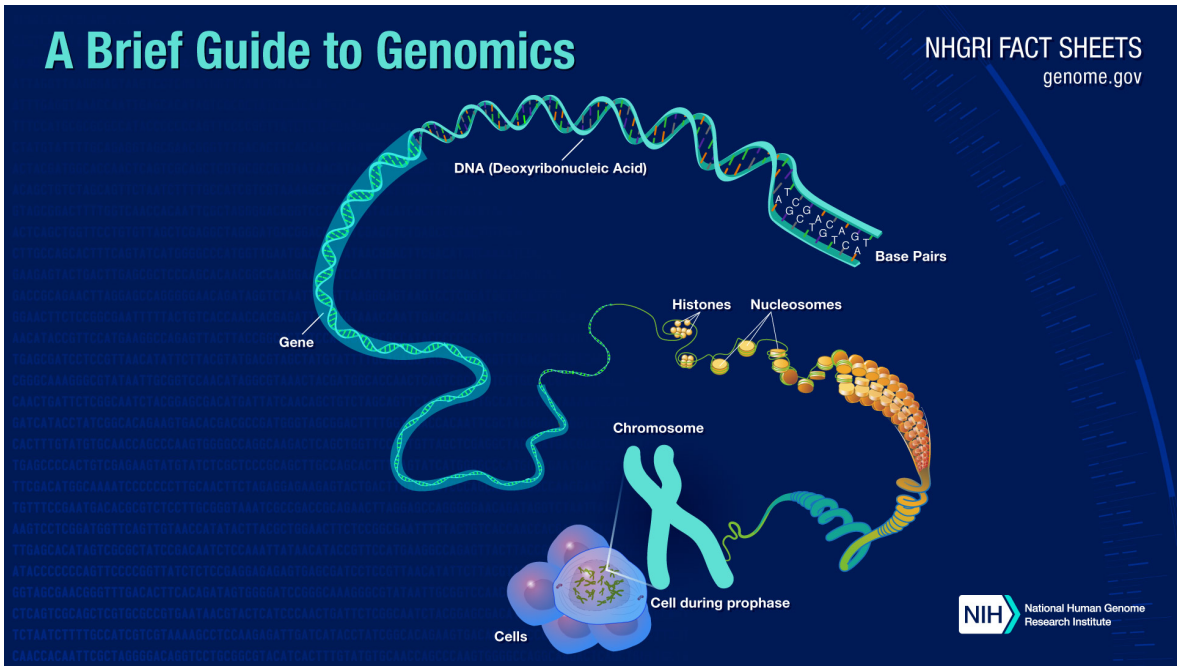

Figure 1.1: *An overview over the organization of an eukaryotic genome. (Source:* `https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics` *)*

the expression the are called silencer. Most of the known transcription factors bind relatively close to the gene which they are regulating and therefore can be called proximal enhancer/silencer. There are also reports about regulatory elements regulating genes from long distances [2] which are therefore called distal enhancer/silencer. RNA polymerases can bind to any gene but transcription factors bind in a sequence-specific manner. Therefore, there are transcription factors which regulate groups of genes that share certain characteristics. Since transcription factors can also regulate other transcription factor genes. In addition there can be several different transcription factors regulating a gene which leads to quite complicated regulatory networks. Researcher try to uncover these interactions by investigating "gene regulatory networks" (GRNs) [3].

The beforementioned mechanisms act in a gene-specific way or on a group of genes. There, are also gene regulatory mechanisms which impact large parts of the genome at once. They change if the DNA is accessible for DNA-binding proteins, like polymerases or transcription factors at all. DNA does not freely lie around in a cell like a string but is packed around a histone octamers which consists of eight histone proteins. The

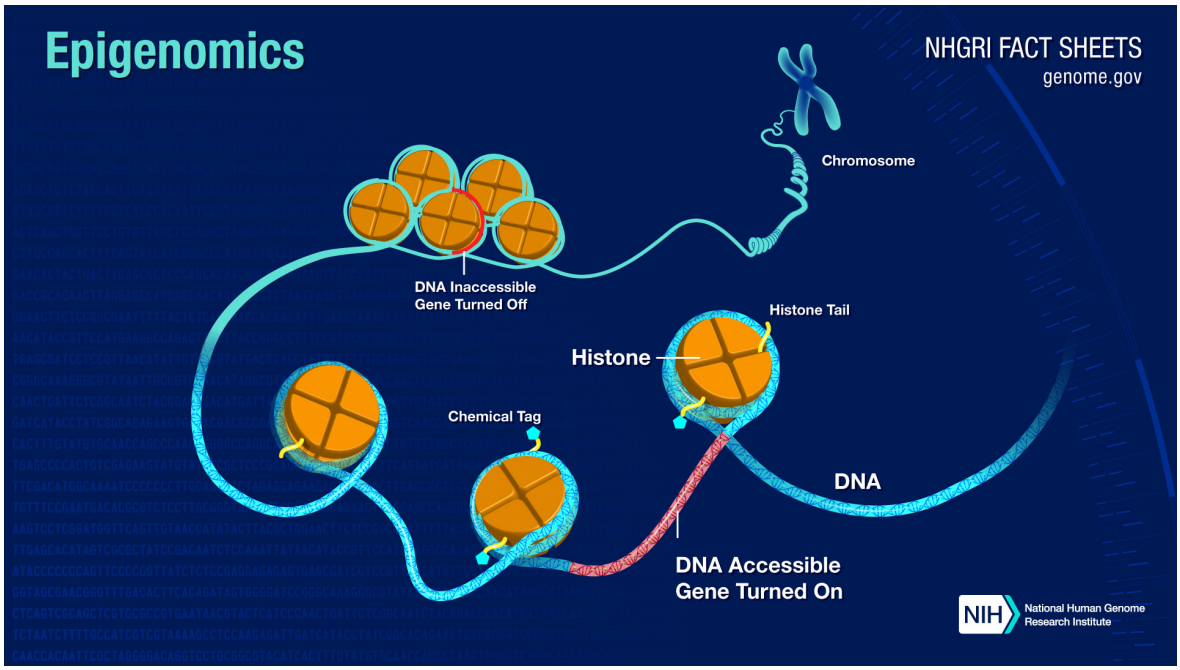

Figure 1.2: *An overview over the chromatin organization in an eukaryotic genome. (Source:* `https://www.genome.gov/about-genomics/fact-sheets/Epigenomics-Fact-Sheet`

DNA packed around histones is together called the chromatin. Depending on how tight this packaging is, the attached DNA is accessible or not. Only accesible DNA can be transcribed because otherwise RNA polymerases and transcription factors can not bind. The concept is visualized in Figure 1.2.

The DNA itself always of the four nucleic acids adenine, cytosine, guanine, thymine. The sequence of nucleic acid is the only information which is transcribed to RNA and subsequently translated into proteins. However, nucleic acids within the DNA can slightly changed by adding a chemical modification to them. In vertebrates the most common DNA modification is modification of a cytosine. This modification does not change the resulting RNA sequence. An methylated cytosine and an unmethylated cytosine both lead to a cytosine in the RNA sequence. But the fact if a cytosine is methylated or not can have an effect on gene regulation. Several methylated cytosines in the promoter region can prevent binding of RNA polymerases and therefore prevent the transcription of a gene. Similarly, methylation at an regulatory element can prevent the binding of transcription factors to it.

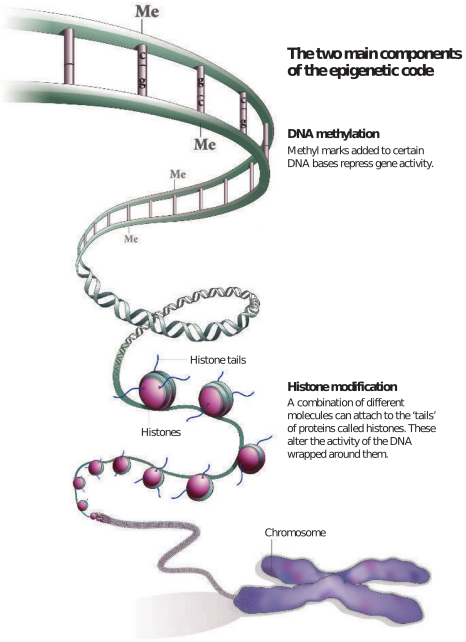

Figure 1.3: *The two kinds of epigenetic modifications: DNA methylation and histon modifications and where they are located in the chromatin. (Source: [4])*

DNA modifications are not the only modifications impacting gene regulation. The histones of the histone octamers have histone tails which is a part of the histone protein that is accessible from outside. The amino acids of these histone modifications are frequently modified. During this process different chemical modifications are added at specific amino acids of the histone tail, frequent modifications are for example methylation and acethylation. DNA methylation and histone modifications are commonly called epigenetic modifications and play a large role in making the chromatin accessible or inaccesible and therby regulating the expression of genes. The modifications on the different parts of the chromatin are visualized in Figure 1.3

Phylogeny of Metazoa

In this thesis we are focusing on metazoan animals and more specifically Ecdysozoa and Vertebrata. The respective groups are highlighted in Figure 1.4. As one can see there are two main groups of Bilateria: Protostomia and Deuterostomia. Ecdysozoa belong to the Protostomia together with their sister group Lophotrochozoa. Vertebrata on the other hand can be found with the other Cordata in the figure. They belong to

the Deuterostomia.

The shown phylogeny is already outdate in a few spots but none of them involve species we are working with. Acoela together with Xenoturbellida, for example, are currently often place as a sister group to the other Bilateria [5].

Metazoa

Evolution of DNA methylation across Metazoa

DNA methylation is prominent in vertebrates, where it is considered a fundamental part of vertebrate epigenetic programming [7]. In human, about 70-80% of CpGs are methylated. Several non-vertebrate model organisms, such as *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae* [8, 9] lack DNA methylation. It was discovered early-on, however, that some insects must have a DNA methylation mechanism [10]. Since then, several studies have investigated the heterogenous distribution of DNA methylation in insects [11, 12, 13] and other arthropods [14, 15]. These showed that most insect orders have kept some amount of DNA methylation. The most prominent counterexample are Diptera which include the genus Drosophila. In nematodes, DNA methylation has only been identified in a few species. The highest levels are found in *Romanomermis cuicivorax* and low amounts in *Trichinella spiralis*, *Trichuris muris* and *Plectus sambesii* [16, 17] suggesting an early loss during nematode evolution, prior to the separation of the nematod clades III, IV, and V.

In animals, DNA methylation predominantly occurs at CG sites [18, 7]. Two different sub-classes of enzymes are responsible for establishing DNA methylation. DNA methyltransferase 1 (DNMT1) reestablishes methylation on both DNA strands after a cell division. It preferentially targets hemi-methylated site. DNA methyltransferase 3 (DNMT3) can perform *de novo* methylation of unmethylated CpGs in the DNA. In vertebrates, DNMT3 is mainly active during embryonic development. However, the view of a clear separation of tasks has has been challenged [19, 7]. Not only does DNMT3 contribute to the maintenance of DNA methylation, DNMT1 has a notable *de novo* activity, as well. In addition DNMT1 might have other functions outside of DNA methylation [20, 21] but they have not been studied extensively. Mainly because DNMT1 or DNMT3 knock-outs in human embryonic stem cells or mouse embryos have catastrophic consequences, e.g. cell death or embryonic lethality [22].

DNMT2 has been believed to be a DNA methyltransferase as well until it was discov-

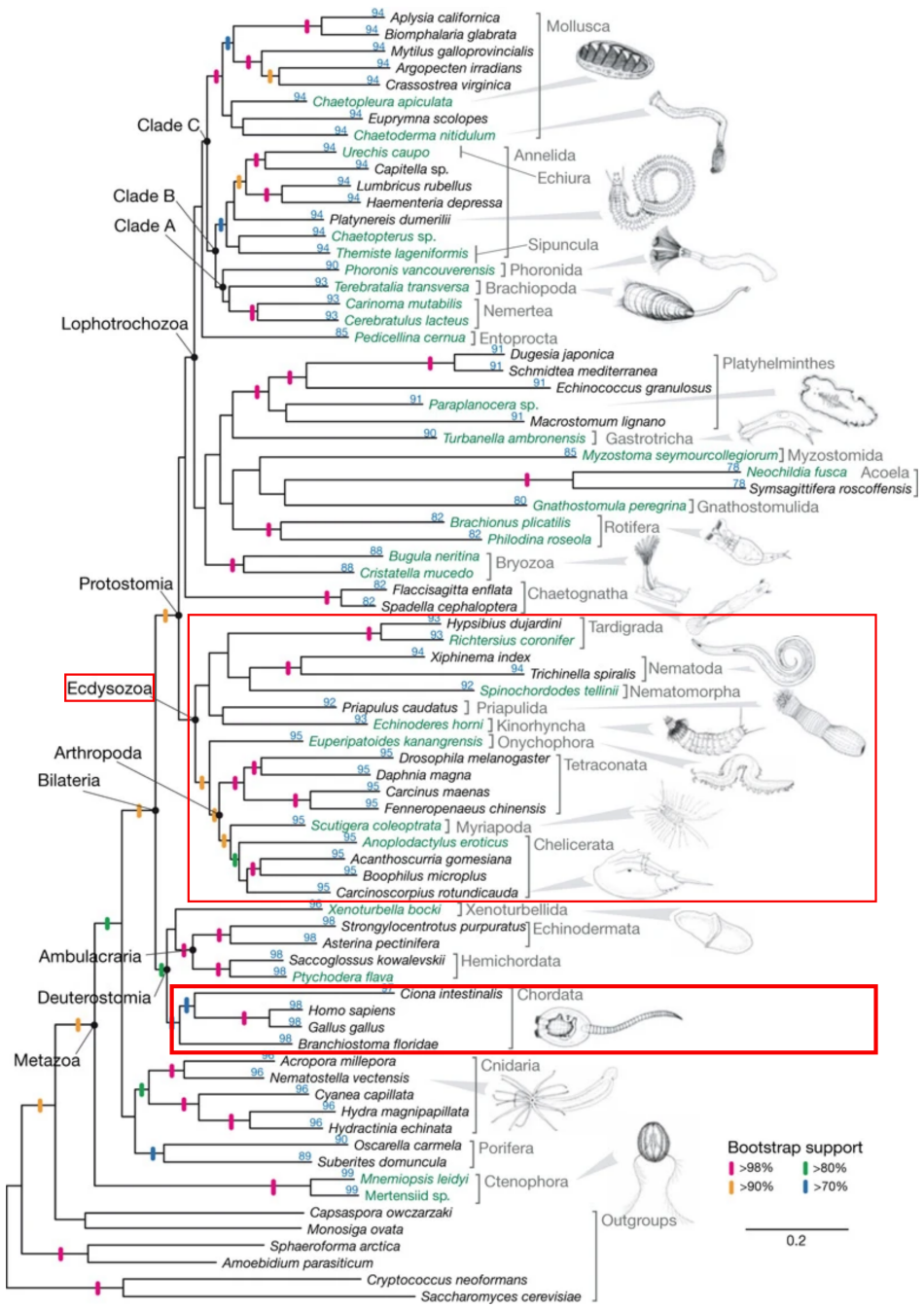

Figure 1.4: *The large red box highlights Ecdysozoa one of the two main groups we study. The smaller box highlights Chordata. We are actually studying only Vertebrata which are a subgroup of Chordata. As one can see given the whole metazoan phylogeny vertebrates are only a small part.(Source : after Dunn et al. [6])*

ered that it recognizes tRNAs as a substrate. It methylates cytosine C38 of tRNA(Asp) in human and therefore is actually an RNA methyltransferase [23].

DNA methyltransferases are believed to have emerged in bacterial systems from "ancient RNA-modifying enzymes" [24]. Subsequently, six distinct clade of DNA methyltransferases have been acquired by eukaryotic organisms through independent lateral transfer [24]. The DNMT clades thus do not have a common ancestor within the eukaryotes. DNMT1 and DNMT2 can be detected in most major eukaryotic groups, including animals, fungi and plants. Fungi lack DNMT3 but retained DNMT4 and DNMT5 similar to some, but not all, Chlorophyta (green algae). Embryophyta (land plants) lack DNMT4 and DNMT5 but harbor chromomethylase (Cmt), an additional DNA methytransferase related to DNMT1 [25]. In Eumetazoa only DNMT1, DNMT2 and DNMT3 can be found. Although DNA methylation clearly is an ancestral process, it is not very well conserved among Protostomia.

All DNA methyltransferases (DNMTs) have a catalytic domain at their C-terminus. It transfers a methyl group from the substrate S-AdoMet to the C5 atom of an unmethylated cytosine [7]. However, the different families of DNMTs can be distinguished by their regulatory domains and conserved motifs in the catalytic domain [26]. With five domains, DNMT1 has the most regulatory domains. The DMAP-binding domain binds DMAP1, a transcriptional co-repressor. Also HDAC2, a histone deacethylase, establishes contact to the N-terminal region of DNMT1 [27]. The RFTS domain (or RFD) targets the replication foci and directs DMAP1 and HDAC2 to the sites of DNA sythesis during S phase [27]. The CXXC domain is a zinc-finger domain that can be found in several chromatin-associated proteins and binds to unnmethylated CpC dinucleotides [28]. The two BAH (bromo-adjacent homology) domains have been proposed to act as modules for protein-protein interaction [29, 20].

DNMT3 has only two regulatory domains, a PWWP domain, named after the conserved Pro-Trp-Trp-Pro motif, and an ADD domain. Both mediate binding to chromatin. For the PWWP domain of (murine and human) DNMT3A, recognition of histone modifications H3K36me3 and recently also H3K36me2 has been reported [30, 31]. The ADD domain, is an atypical PHD finger domain, shared between ATRX, DNMT3, and DNMT3L, and has been shown to interact with histone H3 tails that are unmethylated at lysine 4 [32, 33].

DNMT2 has no regulatory domains [7].

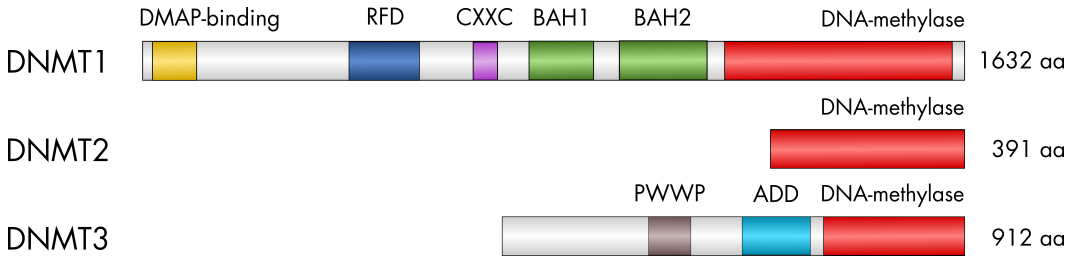

Figure 1.5: *Conserved domains of animal DNA methyltransferases. Scaling and numbers refer to the human homologs.*

1.2 Detecting DNA methylation

Bisulfite-sequencing

The development of high-throughput sequencing techniques lead to the beginning of the big data era in molecular biology. Fortunately, with a slight modification standard high-throughput DNA sequencing can be used to detect DNA methylation, so called Bisulfite-sequencing [34]. If normal DNA is treated with bisulfite all unmethylated cytosines are converted to uracil, after an additional Polyermase Chain Reaction (PCR), it will be converted to thymine and subsequently sequenced as such. Methylated cytosines on the other hand will not be converted and therefore will still be sequenced as cytonsines. If one has a reference genome available these conversion allows to detect DNA methylation. If there is a cytosine in the genome which is completly unmethylated there all reads mapping to it will contain a thymin at that position. If the cytosine was fully methylated there would be only cytosines mapping to it. If there is a mix of cytosines and thymines mapping to that position the methylation level can be calculated using the relative amount of cytosines in all reads. While a single nucleotide can only be methylated or not, mostly we do not sequence single cells but a set of many, even several thousands or more, cells. Since DNA methylation can be different in every cell the methylation level describes in how many of the cells which we sequenced DNA methylation is present. If the cells had a very homogenous distribution of DNA methylation, for example because they share the same cell type, then we would expect to mainly detect methylation levels close to either 0% or 100%. The more heterogenous the cells which we sequence are, the more difficult it becomes to interpret the results of the DNA methylation level.

Computational prediction of DNA methylation

Methylated DNA is subject to spontaneous deamination of 5-methylcytosine, which leads to the formation of thymine and, consequently, to T·G mismatches. Over time, this results in C to T transition mutations predominantly in the context of CpG sites and CpG depletion in frequently methylated regions of the DNA. This changes the the number observed CpGs observed relative to the number expected from the C/G content of the genome. The observed/expected CpG distribution has been used in several studies to infer the presence of DNA methylation [12, 13, 35].

In *Apis mellifera* it has been show that its genes can be divided in two classes, depending on whether they exhibit a low or a high amount of CpG dinucleotides. This was explained by the depletion of CpG dinucleotides if DNA methylation is present. The highly methylated (low CpG) genes were associated with basic biological processes while lowly methylated (high CpG) genes were enriched with functions associated with developmental processes [36]. This "bimodal distribution" of CpG dinucleotides can be used to predict the presence of DNA methylation.

In invertebrates, gene bodies are methylated more heavily than other parts of the genome. Higher methylation levels should lead to a stronger statistical signal and therefore make it easier to decide if DNA methylation is present or not. Therefore, gene bodies have recently been in the focus of studies investigating DNA methylation in invertebrates. Several different criteria have been developed to distinguish the patterns of methylated and unmethylated DNA.

Bewick *et al.* [12] use Gaussian mixture modeling (GMM) modeling with two components. Subsequently, they compare the 95% confidence intervals (CI) of the means. If they are overlapping they assumed a unimodal distribution, otherwise a bimodal one. In case of a bimodal distribution the presence of DNA methylation are assumed. Provataris *et al.* [13] use the same GMM modelling. They define three different modes: "Bimodal depleted", if the difference between both means is > 0.25 and the distribution with the lower O/E CpG ratio has a mean < 0.7, and the smaller component contains a proportion of the data > 0.1; "unimodal, indicative of DNA methylation", if they do not fall in the first category but the portion of data which falls in the distribution with the lower O/E CpG ratio is ≥ 0.36 (this cutoff represents the corresponding value in *Bombyx mori*). All other cases are classified as "unimodal, not indicative of DNA methylation". Aliaga *et al.* [35] use a method based on kernel density estimations.

They define four clusters based on the mode number (n), mean of the modes, skewness (sk) and standard deviation (sd). Three of the clusters are defined, among other parameters, as having one mode: "Ultra-low gene body methylation", "Low gene body methylation" and "Gene body methylation". Cluster with two modes (or 1 mode with skewness < -0.04) are defined as "Mosaic DNA methylation type".

The predictions of the different methods are largely consistent although they may differ in individual cases and do not always match the the observed presence or absence of DNMTs, see chapter 2.

Chapter 2

Evolution of DNA methylation across Ecdysozoa

2.1 Introduction

DNA methylation is a crucial mechanism in vertebrate gene regulation that plays a major role in cell fate decision making but their role in invertebrate gene regulation is much less clear. It appears that its function might differ significantly in different invertebrate groups. In the last years several experimental methods for detecting genomic DNA methylation have been developed. Nevertheless, they are still more expensive compared to sequencing the unmodified genome only. This can be problematic if one wants to widen the phylogenetic range of DNA methylation studies and include a large number of species. Another problem is that some of the lesser studied taxa are difficult to collect and culture which makes them less available for extensive experimental work. Bioinformatic studies such as the present one can help design such experimental studies. Relying on available public data we can make detailed predictions about the presence or absence of DNA methylation and the respective enzymes. Using these computational results one can decide more efficiently which taxa are most valuable to study to gain a new insight into the evolution of DNA methylation in invertebrates.

In this chapter, we present a detailed investigation of the presence and absence of DNA methyltransferases (DNMTs) across five ecdysozoan phyla, see Figure 2.1. Most of the 138 species analyzed here are from the phyla Arthropda and Nematoda. However, we also include less commonly studied groups such as Tardigrada, Onychophora and Priapulida. We identify at which points of the ecdysozoan evolution DNMTs were lost and investigate whethere there are common patterns between the phyla. In addition, we present an easy-to-use statistical approach for predicting the presence of genomic DNA methylation based on coding sequence data and apply it to our species of interest. The results of the predictions are compared with available experimental data.

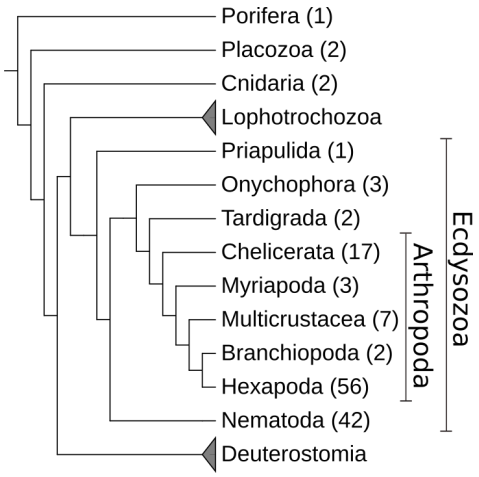

Figure 2.1: *Overview of the metazoan phylogeny with a focus on Ecdysozoa. The number of species per group used in this study is given in brackets. Lophotrochozoa and Deuterostomia are shown for orientation only.*

2.2 Methods

Identification of DNA methyltransferases

Proteome-based search The predicted proteins of the species analyzed were downloaded from different sources, see supplementary Table 1. For 82 and 42 species data was taken from NCBI [37] and Wormbase [38], respectively. Data for seven species each were retrieved from ENSEMBL [39] and Laumer *et al.* [40].

The protein domain models for DNA_methylase (PF00145), ADD_DNMT3 (PF17980), CH (PF00307), PWWP (PF00855), BAH (PF01426), DMAP_binding (PF06464), DNMT1-RFD (PF12047) and zf-CXXC (PF02008) were downloaded from the "Pfam protein families database" [41]. Initially, only the DNA_methylase model was used to identify DNA methyltransferase (DNMT) candidates in the set of proteins predicted using `hmmsearch` from the HMMER software `http://hmmer.org/` version 3.2.1. Proteins with a predicted DNA_methylase domain and a full sequence e-value < 0.001 were further considered as candidates. For these, all before mentioned protein domains were annotated. Finally, each DNMT candidate was classified into one of three classes using custom perl scripts. A **DNMT1 candidate** was required not to have a PWWP or ADD_DNMT3 domain. In addition, having a DNMT1_RFD, zf-CXXC and BAH domain it was considered a *full* DNMT1 candidate, with only one of them a *partial* DNMT1 candidate. A **DNMT3 candidate** was required not to have a DNMT1_RFD,

zf-CXXC or BAH domain. With both, a PWWP and a ADD_DNMT3 domain, it was considered a *full* DNMT3 candidate, with only one of them a *partial* DNMT3 candidate. A **DNMT2 candidate**, was required to have only a DNA_methylase domain and none of the other domains mentioned above. In a last step, the classification of the DNMT candidates was checked manually. The sequences of the DNA methylase domain of each candidate was extracted and aligned using `Clustal Omega` [42] version 1.2.4. A phylogenetic network was computed with *SplitStree4* [43] version 4.10 and inspected manually for phylogenetic congruence of gene and species phylogeny. In case of contradicting results the specific conserved sequence motifs of the methylase domain were inspected manually and the candidate reassigned to a different class or discarded if it did not contain the proper sequence motifs [26].

	DNMT1-RFD	zf-CXXC	BAH	PWWP	ADD_DNMT3	DNA_methylase
DNMT1 full	≥ 1	≥ 1	≥ 1	0	0	1
DNMT1 partial	$\geq 1^*$	$\geq 1^*$	$\geq 1^*$	0	0	1
DNMT3 full	0	0	0	1	1	1
DNMT3 partial	0	0	0	$\geq 1^*$	$\geq 1^*$	1
DNMT2	0	0	0	0	0	1

Table 2.1: *Classification of DNMT candidates according to the detected domains. If the numbers in multiple columns of one line are marked with an asterisk (*) the condition of only one of the columns has to be fullfilled.*

Genome-based search For selected subgroups an additional genome-based search for DNA methyltransferase (DNMT) candidates was performed. This was the case when the previously described workflow showed an unexpected absence of DNMTs in individual species. For example, a DNMT enzymes is detected in most species of a subgroup but is missing in one or two species. The groups that have been analyzed in addition were: Coleoptera for DNMT1 and DNMT3, Hymenoptera for DNMT3, Hemiptera for DNMT3, Chelicerate for all three DNMTs and Nematoda for all DNMTs. For each group, the DNMTs detected in the group, were used as queries. The programm `BLAT` [44] was used to search the query proteins against the species genome whenever the respective DNMT could not be found in the proteome. The script `pslScore.pl` (`https://genome-source.gi.ucsc.edu/gitlist/kent.git/raw/master/src/utils/pslScore/pslScore.pl`) available from the UCSC genome browser was used to assign a score to each genomic hit. The resulting bed-file was post-processed with the tools of the suite `bedtools` [45]. All hits were clustered

using `bedtools cluster`. If there were overlapping hits, only the best-scoring one was kept. Using blast-type output files from *BLAT* the genomic sequence to which the query was aligned could be extracted to get the full amino acid sequence corresponding to the hit. The full-length protein candidates were aligned using *Clustal Omega*. A phylogenetic network was computed with *SplitStree4* and inspected manually for phylogenetic congruence of gene and species phylogeny. Candidate proteins were discarded if they did not contain the methylase domain-specific, conserved sequence motifs. Otherwise they were kept as DNMT candidates.

This method allowed us to identify six additional DNMT enzymes in five species: *Asbolus verrucosus* DNMT1, *Soboliphyme baturini* DNMT2, *Acromyrmex echinatior* DNMT3, *Laodelphax striatellus* DNMT3, *Trichonephila clavipes* DNMT1 and DNMT3.

Inference of DNA methylation from CpG O/E value distributions

Coding sequences (CDS) for all species were downloaded from NCBI, Wormbase and ENSEMBL according to Supplementary Table 1. For the 7 species from *Laumer et al.* [40] this data was not available. For each CDS the Observed-Expected CpG ratio was calculated using the formula:

$$O/E_{CpG} = \frac{CG \times l}{C \times G} \tag{2.1}$$

with C, G, and CG being the number of the respective mono- an dinucleotids in the given CDS and l being the length of the CDS. CDS shorter than 100 nucleotides or with more than 5% of N's in the sequence were excluded.

We used a Gaussian Mixture Model (GMM) to identify possible subpopulations in the O/E CpG distribution. The Expectation Maximization algorithm in the python module 'sklearn' from the library *scikit-learn* [46] version 0.23.1 was used to estimate the parameters. The GMM was modeled with one or two components. For the GMM with one component, we calculated the Akaike information criterion (AIC). For the GMM with two components, we calculated the AIC and in addition the mean of each component, the distance d of the component means and the relative amount of data points in each component, see supplementary Table 2 and 3. For the distribution of O/E CpG values, the distribution mean, the sample standard deviation, and the skewness

were calculated as well. All pairs of parameters were analyzed using two-dimensional scatterplots generated with R.

We used the distance between the component means as an indicator for DNA methylation. If the distance is greater or equal to 0.25, we assume DNA methylation is present, otherwise it is absent.

Ecdysozoan Phylogeny

The topology of the ecdysozoan phylogeny, used for display only, is a composite of phylogenetic information compiled from several studies. The topology of Arthropoda was based on [47] and combined with phylogentic information for the taxa Coleoptera [48], Lepidoptera [49], Hymenoptera [50], Hemiptera [51], Aphididae [52, 53, 54], Crustacea [55], Copepoda [56], Chelicerata [57], Aranea [58], and Acari [59]. The topology of the nematode phylogeny was based on [60] and combined with phylogenetic information for the genera Plectus [17], Trichinella [61], Caenorhabditis [62], and Diploscapter [63].

2.3 Results

Presence and absence of DNA methyltransferases in Ecdysozoa species

We investigated the presence of DNMTs in the genomes of 138 species using a carefully designed homology search strategy (see Materials and Methods) aiming at minimizing false negatives. Candidate sequences were then curated carefully to avoid overprediction. Most of the available genomes belong to the Nematoda (42) and Arthropoda (85). Of the arthropod species, 56 are Hexapoda (insects) and 29 belong to other subphyla. Only 6 species are from Ecdysozoa groups outside of Nematoda or Arthropoda. In addition 5 species from groups outside of Bilateria have been included. Our findings are summarized in Figures 2.2, 2.3, and supplementary Figure 1. Potential losses of DNMT1, DNMT2, and DNMT3 are marked with stars in the respective colors. In the following paragraphs we discuss the results of our annotation efforts in more detail.

Arthropoda Arthropoda are an extremely species-rich and frequently studied group of invertebrates. The most prominent subphylum is Hexapoda, which contains, among

others, all insects. Several (emerging) model organism belong to insects, e.g. the fruit fly *Drosophila melanogaster* (Diptera), the silk moth *Bombyx mori* (Lepidoptera), the red flour beetle *Tribolium castaneum* (Coleoptera) or the honey bee *Apis mellifera* (Hymenoptera). The group of Crustacea (crabs, shrimp, lobster) is currently believed to be paraphyletic [55]. Multicrustacea consists of most of the "crustacean" species, e.g. the white leg shrimp *Penaeus vannamei* (Decapoda) or the amphipod *Hyalella azteca* (Amphipoda). Branchipoda with the frequently studied water flea *Daphnia pulex* (Cladocera) are currently placed more closely related to Hexapoda. The sister group to all of the beforementiond groups are Myriapoda (millipedes, centipedes). The earliest-branching group of Arthropoda are the Chelicerate. A diverse subgroup of Chelicerata are Arachnida (e.g. spiders, scorpions, ticks) but they also contain the Atlantic horseshoe crab *Limulus polyphemus* (Xiphosura) and sea spiders (Pantopoda). We analyzed 85 species of the phylum Arthropoda. They belong to 28 different taxonomic orders. An overview of the results can be found in Figure 2.2.

The subphylum Hexapoda was the largest group analyzed with 11 different orders. Two had a full set of DNMTs: Blattodea (3 species) and Thysanoptera (1). In four orders only DNMT1 and DNMT2 are present: Siphonaptera (1), Trichoptera (1), Lepidoptera (8) and Phthiraptera (1). In two only DNMT2 could be identified: Diptera (3) and Entomobryomorpha (2). In the remaining three orders the occurence of DNMT enzymes is heterogenous suggesting secondary losses within the order. Coleopetera (11 species) have all DNMTs, DNMT1 and DNMT2 or only DNMT2. Hymenoptera (12) mostly have all DNMTs but in two species of the genus Polistes DNMT3 could not be detected. In three species of Hemiptera (14) we also did not find DNMT3.

The subphylum Crustacea is currently believed to be paraphyletic [55] but the following species are considered part of it. In two species of the Daphnia genus all DNMTs have been found. They belong to the order Cladocera in the class Branchiopoda, formerly part of the subphylum Crustacea. Six additional orders of the former subphylum, belonging to the group of Multicrustacea have been studied. In Amphipoda (1) and Decapoda (1) all three DNMTs have been found, as well. In the orders Calanoida (2 species), Harpacticoida (1) and Siphonostomatoida (1) DNMT3 was not identified. In the calanoida *Lepeophtheirus salmonis* DNMT2 could not be identified as well. For the Calanoida *Calanus finmarchicus* only transcriptomic data was available. In Isopoda (1) DNMT1 and DNMT3 could not be detected.

In the subphylum Myriapoda three different orders have been analyzed with one

species each. All of them showed a full set of three DNMT enzymes. For the two species *Eudigraphis taiwaniensis* and *Glomeris marginata* only transcriptomic data was available.

17 species of the subphylum Chelicerata were analyzed. They belong to 8 different orders. We detected all three DNMTs in Xiphosura (1 species), Scorpiones (1), Aranea (3) and Ixodida (1). The same was the case for Trombidiformes (3) with the exception of *Tetranychus urticae* for which DNMT2 could not be found. In Sarcoptiformes (3) only DNMT3 was not detectable. In Mesostigmata (4) this was the case for DNMT1 and DNMT3. In the one species of Pantopoda (1) *Anoplodactylus insignis* DNMT1 could not be found but only transcriptomic data was available.

Nematoda Nematoda are, next to Arthropoda, the best-studied group of Ecdysozoa. Developing a complete nematode systematics is still an ongoing process. Most available genome data comes from the clades I, III, IV and V. Clade V contains the most well-known nematod species *Caenorhabditis elegans*.

42 nematodes species of five clades were analyzed. Of the 17 species in clade V most had no DNMTs, in 5 species DNMT2 could be detected. In clade III for 8 out of 10 species DNMT2 was present but not the other DNMTs. Clade IV with six species showed no signs of DNMT at all. In *Plectus sambesii*, the only representative of its clade, DNMT3 could not be found. In clade I, in 6 of the 8 species only DNMT2 and DNMT3 were detected. For one species all three DNMTs have been identified. In another one species only DNMT3 is present but DNMT2 could not be found. An overview of the results can be found in Figure 2.3.

Other Ecdysozoa These groups are not often in the focus of scientific studies. At least Tardigrada, commonly known as water bears, gained some interest because they can survive in very harsh conditions, such as extreme temperature, radiation, pressure, dehydration and even in outer space [64]. Onychophora or velvet worms are the sister taxon to Arthropoda+Tardigrada. Some species can bear live offsprings [65]. Priapulida (penis worms) are believed to be among the earliest branching Ecdysozoa and therefore are of great interest for comparative studies. Unfortunately, genomic data so far is only available for one species.

For Onychophora (3) only transcriptomic data was available. In two species DNMT1 and DNMT2 was detected in the third DNMT2 and DNMT3. In Tardigrada (2) only

DNMT2 could be identified. In the single member of the Priapulida all DNMTs were detected.

Early-branching Metazoa The systematics of early-branching Metazoa is difficult to resolve and currently still heavily discussed. The Cnidaria (jellyfish, sea anemones, corals) are believed to be the closest relatives to bilateral animals. Placozoa are a more distant taxa with Trichoplax as the most prominent genus. They are tiny and delicate marine animals and therefore difficult to study. For a long time only one species *Trichoplax adhaerens* was known along with a number of haplotypes. Only recently two more species have been described. Porifera, or sponges, are (together with Ctenophora) a contender for beeing the earliest branching phylum of Metazoa. They mainly occur in marine environment but due to their reproductive behaviour they are difficult to include in molecular biology studies. In the outgroup Placozoa (2) only DNMT2 was detected while in Cnidaria (2) and Porifera (1) all DNMT enzymes were found.

DNA methylation inferred from CpG O/E value distributions

The ratio of observed and expected CpGs serves as an indicator for the the presence of DNA methylation. Since in invertebrates often only a subset of genes is subject to CpG methylation we assume that the observed distribution is a mixture of two gaussian distributions. Similar to previous work, we use an expectation–maximization (EM) algorithm to estimate the parameters of this GMM [12, 13]. The results outlined below were used to revise the parameters reliably indicating bimodality and thus the presence of DNA methylation.

Coding sequence (CDS) data was available for all species except the seven whose data was from *Laumer et al.* [40]. For five species (*C. sinica*, *C. tropicalis*, *S. flava*, *M. sacchari*, *A. verrucosus*) the genome was not published, yet, therefore they have been excluded from this genome-wide analysis. Hence we were able to analyze O/E CpG ratios for the CDS of 126 species. In 94 species a model with two components was favored using the Akaike Information Criterion (AIC) and in 32 species the distribution was unimodal. Surprisingly, in the mononucleotide shuffled data still for 94 species a model with 2 components is favored and in the other 32 cases a model with 1 component. In 72 cases for two components and 10 for one component both datasets

favor a model with the the same amount of components. In 22 cases the real data suggests a two component model and the shuffled data one component, in 22 cases its the other way around. In total this means for 82 of 126 species shuffling of the CDS data does not change the model suggested by the Akaike information criterion (AIC).

Arthropoda						
		Real data			Shuffled data	
Range	Min.	Mean	Max.	Min.	Mean	Max.
meanLow	0.30	0.72	1.17	0.95	0.99	1.00
meanHigh	0.58	1.00	1.46	1.00	1.02	1.05
distance *d*	0.01	0.28	0.63	0.00	0.03	0.11
%low	0.14	0.46	0.87	0.37	0.72	0.81
Nematoda						
		Real data			Shuffled data	
Range	Min.	Mean	Max.	Min.	Mean	Max.
meanLow	0.34	0.94	1.16	0.93	0.98	1.00
meanHigh	0.59	1.10	1.48	1.00	1.02	1.07
distance *d*	0.00	0.15	0.58	0.00	0.04	0.14
%low	0.13	0.59	0.96	0.49	0.74	0.82

Table 2.2: *Summary of the Gaussian Mixture Modelling for real and shuffled data. "meanLow" and "meanHigh" are the component means corresponding to the components with lower and higher O/E CpG ratios (first and second row). The distance d between the means is given in the third row. "%low" gives the relative amount of data points (transcripts) in the component with the lower O/E CpG ratio, "%low" + "%high" equals to 1. Due to its extreme values the nematode Loa loa was excluded from this table. Its values are: "meanLow" 1/1, "meanHigh" 4.53/1.18, d 3.55/0.18 and "%low" 0.99/0.98 for the real/shuffled data.*

Although the AIC is generally accepted for GMMs, in our case comparing real and randomized data mostly the same number of components is suggested, This indicates that the CDS may also fall into two classes distinguished by overall GC content, not only by relative CpG abundance. In this case we expect that species without DNA methylation and randomized data should exhibit a smaller AIC and smaller separation between the two components of the distribution. Empirically, we find that the AIC is a poor decision criterion for our purposes. Table 2.2 shows that the mean distance between the two components is much larger in the real data compared to the shuffled data. Hence we use the difference between the means of the two Gaussians as indicator.

This requires a user-determined threshold above which the difference of two means is interpreted as indicative of DNA methylation. Naively, species having neither DNMT1 or DNMT3 should be less likely to contain DNA methylation, while species in which one or both of the enzymes are present should be more likely to have kept genomic DNA methylation. Of the 126 species analyzed, in 45 the DNMT1 and DNMT3 enzymes have

been found while in 46 neither was found. In 28 species only DNMT1 was detected and in 7 species only DNMT3, see Table 2.3. Figure 2.4 shows the means of both GMM components for all analyzed species, marked by different colors and symbols according to their set of DNMT1/3 enzymes and their taxonomic group. The diagonal line indicating a difference between the means d of 0.25 is able to separate almost all of the species with no DNMT1/3 from the others. Trying to avoid false positive predictions we choose this value as a conservative threshold. In our data, 55 of 126 species had a distance greater or equal 0.25 indicative of DNA methylation. The other 71 species have a distance smaller than 0.25.

		methylation	
		present	absent
enzymes present	total	$d \geq 0.25$	$d < 0.25$
DNMT1 & DNMT3	45	**36**	9
DNMT1 only	28	16	12
DNMT3 only	7	0	7
none	46	3	**43**
	131	58	73

Table 2.3: *Relationship between the combination of DNMT candidates and the predicted methylation level. Shown is the amount of species for which DNA methylation is predicted to be present or absent classified by the presence of DNMT enzyme combinations.*

2.4 Discussion

To our knowledge this study is the phylogenetically most diverse analysis of DNA methylation in Ecdysozoa, to-date. While Arthropoda and Nematoda are its two most studied phyla we also include species from Priapulida, Onychophora and Tardigrada. We therefore analyze five out of seven Ecdysozoa phyla.

Presence and Absence of DNA methyltransferases

Overall, our data show that both individual DNMTs and DNA methylation as a process have been lost independently in multiple lineages. Since the absence of an enzyme is difficult to prove conclusively, we rely on data from related species and invoke parsimonious patterns to identify loss events with confidence: the lack of evidence for a DNMT in an entire clade of related species makes a loss event a very plausible explanation.

There are several reasons why a DNMT may escape detection. The most prominent cause is a low quality, fragmented genome assembly. Not finding a homolog in a species with a high quality, completed genome assembly, in particular in model organisms such as *Caenorhabditis elegans* and *Drosophila melanogaster* makes a negative search result more reliable. It is also possible that a protein has diverged so far that it is no longer recognizable as a homolog in the target organism by the search method used. This explantion becomes more likely as the phylogenetic distance of the target to the closest species with a known homolog increases.

The predicted phyletic pattern of DNMT losses is quite different in Arthropoda and Nematoda. DNMT1 is found in most arthropod species analyzed in our study. Three independent loss events of DNMT1 are suggested by our data (2.2). In Nematoda only two events of DNMT1 loss are suggested but they occur earlier in the evolution of the studied nematod species. Therefore, only in two species DNMT1 can still be detected.

DNMT2 is most likely present in all Arthropoda. The absence in two individual species is probably a technical artifact since DNMT2 enzymes are present in closely related species in both cses. In Nematoda, absence of DNMT2 enzymes is fare more frequent. Given the near perfect conservation of DNMT2 in other metazoan species, this is rather unexpected. Interestingly, the candidate DNMT2 sequences are clearly more divergent compared to those in Arthropoda, which may hint at false positive predictions of 13 DNMT2 enzymes. In this case, a single loss event either after divergence of clade I or both, clade I and clade P, is plausible.

DNMT3 seems to be the most dispensable member of the DNMT family. According to our data, it was lost eight times in Arthropoda. It only occures in combination with DNMT1 and is lost prior to or simultaneously with loss of DNMT1. In Nematoda, DNMT3 is present in all members of clade I and absent in all other clades. Interestingly, in all but one species of clade I, we detected a DNMT3 in the absence of DNMT1.

Absence of DNMT3 in the presence of DNMT1 is frequently associated with low levels of CpG depletion. The week bimodality of the CpG ratio distribution may be the consequence of a return to an unbiased, unimodal distribution caused by decaying methylation levels due to failure to (re-)establish and maintain methylation. Under certain conditions, DNMT1 may have weak *de novo* activity [66]. The molecular mechanism involves binding to unmethylated CpGs via the CXXC domain and auto-inhibition of *de novo* methylation [29]. Via its regulatory domains DNMT1 interacts with epigenetic factors which may be involved in regulalting DNMT1 *de novo*

activity.

The loss events as defined in this study are well supported by the absence of the enzymes in related species, see the colored stars in Figures 2.2, 2.3 and supplementary Figure 1. More precisely, a loss is only inferred if the respective DNMT could not be found in all species of the respective subtree and if it contains at least 2 species. Considering the problems in gene detection, these rules remove cases where the poor quality of single genomes may prevent the detection of DNMTs. In Arthropoda all members of the DNMT family can be identified in several species of each subphylum. Therefore it is unlikely that the negative predictions are caused by extreme divergence of protein sequences that might have rendered them undetectable by homology search methods. The N50 value (that is, 50% of the genome is covered by contigs with a length of at least N50) serves a good measure of assembly quality for our purposes. In Arthropoda, five species are missing DNMT1 or DNMT3 and are not covered by the loss events we propose. The genomes of *Diaphorina citri* (Hemiptera), *Armadillidium vulgare* (Multicrustacea) and *Oryctes borbonicus* (Coleoptera) are the 13th, 8th and 7th worst assemblies in Arthropoda according to the N50 value, see supplementary Table 1. The N50 for *D. ponderosae* (Coleoptera) is around average and for *Anoplodactylus insignis* (Chelicerata) only a transcriptome is available. It is difficult therefore, to interpret these potential loss events. A more reliable prediction will be possible when better genomes or data from more closely related species become available.

The DNMT1/DNMT3 losses in Nematoda are more difficult to evaluate since there are so few positive findings. Their absence in clade III, IV and V is supported by the findings of [17]. These groups contain several high quality genomes, such as the one from the model organism *C. elegans*. The most likely reason for missing existing proteins would therefore be that they are already too diverged. However, DNA methylation has been verified to be absent in several of them and no findings of DNMT enzymes have ever been reported. Therefore, it seems reasonable to conclude that DNA methylation and both DNA metyltransferases are absent fron Nematoda of clade III, IV, and V.

In clade I, DNMT3 is evidently present. However, it seems that DNMT1 is absent in all but a single species examined. This pattern cannot be seen in any other ecdysozoan group. The exception is the earliest branching nematode *Romanomermis cuicivorax*, which posesses both, DNMT1 and DNMT3, as well as DNMT2. The case of *Plectus sambesii*, the sole member of clade P, is quite interesting because DNMT1 is present

while DNMT3 is absent. However, the genome of *P. sambesii* is the 3rd worst of all nematods putting the loss of DNMT3 into question. We can therefore suggest two possible scenarios, either DNMT3 was lost in the stem lineage of clade P and the clades III, IV and V, i.e. before the loss of DNMT1 or after branching of clade and simultaenously with loss of DNMT1.

The two missing DNMT2 in Arthropoda are likely to to be false negatives since homologs of DNMT2 were detected in all other arthropods. Likely, this is also the case in the nematode *Trichuris trichiura* since in the two other species of its genus DNMT2 was found. In clade III, IV, and IV the pattern seems not very parsimonious and our analysis reports three independent DNMT2 loss events. In addition, we did not detect DNMT2 candidate in two more species in clade III. Visual inspection of the DNMT2 alignment revealed that DNMT2 candidates of clades III and V are highly divergent. In conclusion, it remains questionable whether these enzymes are still functional DNA methyl transferases.

Species	Engelhardt et al.			Rovsic et al.		Exp. data
	DNMT1	DNMT3	Methyl.	DNMT1	DNMT3	
Nematoda						
R. culicivorax	X	X	X	X	X	X [17]
T. spiralis	O	X	**O**	O	X	X [17]
T. muris	O	X	**O**	O	X	X [17]
P. sambesii	X	O	**O**	X	O	X [17]
P. redivivus	O	O	O	O	O	n/a
B. xylophilus	O	O	O	O	O	n/a
M. hapla	O	O	O	O	O	n/a
G. pallida	O	O	O	O	O	n/a
A. suum	O	O	O	O	O	n/a
D. immitis	O	O	O	O	O	n/a
O. volvulus	O	O	O	O	O	n/a
B. malayi	O	O	O	O	O	n/a
N. brasiliensis	O	O	O	O	O	O [17]
C. briggsae	O	O	O	O	O	O [17]

Table 2.4: *The table contains all species analyzed in this study which have been analyzed as well in either Bewick et al., Provataris et al., Rovsic et al. or if experimental verification of DNA methylation is available. X - indicates presence; O - indicates absence; DNMT1/DNMT3 means the occurence of at least one paralog of the respective enzyme. Methyl. means if the respective study defines the genome as containing DNA methylation or not according to the O/E CpG content (In case of Provataris et al. 'X' is 'Unimodal, indicative of methylation' and 'XX' is 'bimodal depleted'). If the species name is bold there is a contradiction in DNMT occurences or methylation status between our data and another study. The column of our study which is contradicting is bold as well.*

Table 2.4 and 2.5 summarizes our results and provides a comparison with two recent studies. We analyzed 138 species in total, of which 35 and 34 have been previously

examined by Bewick *et al.* [12] and Provataris *et al.* [13], respectively. To the largest part, the results are is concordance. We were able to identify six DNMTs, i.e one DNMT1 (*P. vannamei*) and five DNMT3 candidates (*P. vannamei*, *I. scapularis*, *B. germanica*, *N. lugens* and *H. halys*), respectively, which have been missed in at least one other study. We on the other hand, only miss to identify the DNMT3 enzyme in *L. salmonis* reported by *et al.* [12]. Of the 42 Nematoda analyzed in our study, Rovsic *et al.* [17] investigated a subset of 14. The results for the presence/absence of DNMT enzymes in these 14 species are identical.

DNA methylation inferred from CpG O/E value distributions

Over evolutionary time, the distribution of CpG dinucleotides is influenced by DNA methylation, which gives rise to an increased rate of C to T mutations and, consequently, Cpg depletion. In case of genome-wide DNA methylation, as in vertebrates, the signal is easy to detect. The situation is more challenging in invertebrates, where methylation is often concentrated to a subset of coding regions. A two-component Gaussian Mixture modelling (GMM) approach is used to model the populations of methylated and unmethylated coding sequences. As we could show, the distance d between the component means is a reasonable measure for the level of DNA methylation in Ecdysozoa. Using d and a threshold of 0.25 we could confirm the previously reported *absence* of notable DNA methylation in several species, such as the fruit fly *Drosophila melanogaster* ($d = 0.01$), the red flour beetle *Tribolium castaneum* ($d = 0.08$) or the nematode *Caenorhabditis elegans* ($d = 0.20$). Furthermore, we predicted the *presence* of DNA methylation in a number of species such as, the insects *Bombyx mori* ($d = 0.39$), *Nicrophorus vespilloides* ($d = 0.37$), *Apis mellifera* ($d = 0.58$), *Acyrthosiphon pisum* ($d = 0.49$), *Blatella germanica* ($d = 0.30$), the water flea *Daphnia pulex* ($d = 0.32$) or the nematod *Romanomermis culicivorax* ($d = 0.58$), which is in concordance with the literature.

Unfortunately, the number of studies which used experimental methods to verify the presence of DNA methylation in Ecdysozoa is quite limited, in particular outside of Hexapoda. Our data suggests several loses of DNA methylation which can not be supported by evidence other than the computationally calculated O/E CpG ratio. Due to the predicted presence of DNA methylation in closely related species some "species-specific" losses seem questionable, e.g. *Danaus plexippus* ($d = 0.11$) and *Acromyrmex*

Species	Engelhardt et al.			Bewick et al.			Provataris et al.			Exp. data
	D1	D3	M	D1	D3	M	D1	D3	M	
Arthropoda										
L. polyphemus	X	X	X							X [67]
P. tepidariorum	X	X	X							X [67]
Ixodes scapularis	X	**X**	**O**				X	O	O	X [67]
Strigamia maritima	X	X	**X**				X	X	O	X [14]
P. vannamei	**X**	**X**	X				O	O	XX	n/a
A. vulgare	O	O	**O**							X [67]
L. salmonis	X	O	O				X	O	O	n/a
D. pulex	X	X	X	X	X	X	X	X	X	X [68]
D. magna	X	X	X							X [68]
F. candida	O	O	O				O	O	O	n/a
O. cincta	O	O	O				O	O	O	n/a
Z. nevadensis	**X**	X	X	O	X	X	X	X	XX	n/a
B. germanica	X	**X**	X	X	O	X				X [12]
N. lugens	X	**X**	O				X	O	O	n/a
H. halys	X	**X**	X	X	O	X				n/a
R. prolixus	X	O	**O**	X	O	X	X	O	O	n/a
C. lectularius	X	O	X	X	O	X				n/a
B. tabaci	X	X	X				X	X	XX	n/a
D. citri	X	O	X	X	O	X				n/a
A. pisum	X	X	X	X	X	X	X	X	XX	X [67]
A. gossypii	X	X	X				X	X	XX	n/a
P. humanus	X	O	X	X	O	X	X	O	XX	n/a
A. rosae	X	X	X	X	X	X				n/a
O. abietinus	X	X	**O**	X	X	X	X	X	O	n/a
N. vitripennis	X	X	X	X	X	X	X	X	XX	X [12]
P. dominula	X	O	X							X [69]
P. canadensis	X	O	X	X	O	X				X [69]
A. mellifera	X	X	X	X	X	X	X	X	XX	X [12]
B. impatiens	X	X	X	X	X	X	X	X	XX	n/a
H. saltator	X	X	**X**	X	X	X	X	X	O	X [70]
S. invicta	X	X	X	X	X	X	X	X	X	X [71]
A. echinatior	X	X	O	X	X	X	X	X	O	n/a
A. cephalotes	X	X	**X**	X	X	X	X	X	O	n/a
A. planipennis	X	X	X	X	X	X				n/a
N. vespilloides	X	X	X	X	X	X				X [12]
O. taurus	X	X	X	X	X	X				n/a
T. castaneum	X	O	O	X	O	O	X	O	O	O [12]
D. ponderosae	O	O	O	O	O	O	O	O	O	n/a
A. glabripennis	X	O	**O**	X	O	X				n/a
L. decemlineata	X	O	X	X	O	X				n/a
C. felis	X	O	X				X	O	O	n/a
A. aegypti	O	O	O	O	O	O	O	O	O	O [12]
A. gambiae	O	O	O	O	O	O	O	O	O	O [12]
D. melanogaster	O	O	O	O	O	O	O	O	O	O [12]
L. lunatus	X	O	X	X	O	X				n/a
P. xylostella	X	O	**X**	X	O	X	X	O	O	n/a
B. mori	X	O	X	X	O	X	X	O	X	X [12]
Operophtera brumata	X	O	X	X	O	X				n/a
P. xuthus	X	O	X	X	O	X	X	O	X	n/a
D. plexippus	X	O	**O**	X	O	O	X	O	X	n/a
H. melpomene	X	O	X	X	O	X	X	O	XX	X [67]
M. cinxia	X	O	**X**	X	O	O	X	O	O	n/a

Table 2.5: *For caption see Tab. 2.4. D1/3 stands for DNMT1/3; M for Methylation.*

echinatior ($d = 0.24$). Conversely, some of the positive findings are likely to be false

predictions, e.g. the nematods *Caenorhabditis angaria* ($d = 0.36$), *Loa loa* ($d = 3.55$) and *Strongyloides ratti* ($d = 0.25$). For many other species there is currently no experimental verification available. The reason for the incorrect predictions is currently not easy to explain. Mostly, there are other, presently unknown factors that influence the distribution in CpGs in the genome. Such effects are difficult to distinguish from the effects of DNA methylation.

Compuational predictions of methylation status have been performed with different methods by *Bewick et al.* [12] and *Provataris et al.* [13]. Supplementary Table 5 provides a summary of their findings and the respective results from our study. Compared to [12] there are three cases where we predict no DNA methylation while they predict DNA methylation: *N. lugens* ($d = 0.2$), *R. prolixus* ($d = 0.14$) and *D. plexippus* ($d = 0.11$). Compared to [13], there are five cases where we predict DNA methylation while they do not: *S. maritima* ($d = 0.35$), *H. saltator* ($d = 0.44$), *A. cephalotes* ($d = 0.27$), *P. xylostella* ($d = 0.28$) and *M. cinxia* ($d = 0.27$). In one case, *D. plexippus* again, we predict DNA methylation while they do not.

In total these are 8 species in which our methylation prediction disagree with at least one of the other two papers. In the case of *S. maritima* and *H. saltator* there is experimental evidence for DNA methylation so our prediction is backed up by that. For the other species no such data is available. *D. plexippus* is the only case where both other studies agree on contradicting our prediction. This species would be the only exception in Lepidoptera without DNA methylation, therefore it appears to be a likely false negative. The other 5 species are part of all three studies and in all cases our prediction is supported (three times [12], two times [13]) by one study and contradicted by the other. Our prediction is worse than the those of competing methods only in the single case of *D. plexippus*.

For 28 of the species examined, experimental data on the presence (22) and absence (6) of DNA methylation is available. We correctly predict the presence and absence of DNA methylation for 17 and 6 species, respectively, totaling to 23 out of 28. The remaining five predictions are false negatives. Note that there are no false positive predictions given the experimental data set at hand. Among the species corresponding to the false negative predictions are two arthropod species, *I. scapularis* ($d = 0.2$) and *A. vulgare* ($d = 0.21$), and three nematode species *T. spiralis* ($d = 0.24$), *T. muris* ($d = 0.08$) and *P. sambesi* ($d = 0.15$), see also supplementary Table 4 and 5. According to Lewis *et al.* [67], the level of DNA methylation in *A. vulgare* is very low which is

likely the reason why our prediction method fails. There is no obvious explanation why we miss DNA methylation in *I. scapularis*. In the there nematodes, notable levels of DNA methylation are mostly present at repeats, which cannot be captured by our method. According to Rovsic *et al.* [17] only the nematod *R. culicivorax* shows a bimodal distribution for DNA methylation across genes.

Conclusions

The amount of genomics and transcriptomics data from a wide range of species is constantly increasing. Often only a relatively small phylogenetic range is analyzed simultaneously. The analysis of "universal" evolutionary patterns, however, requires that the same analysis is applied to widely different groups of species. With this study we provided the largest and most diverse analysis of DNA methyltransferases enzymes in Ecdysozoa, to date. Previous studies have focussed on specific subgroupsm in particular Arthropoda [12, 13] and Nematoda [17] and covered only selected phyla. We combined data for five Ecdysozoan phyla (Priapulida, Nematoda, Onychophora, Tardigrada and Arthropoda) and identified DNMT1, DNMT2 and DNMT3 in four out of these phyla. The only exception are Tardigrada, where neither DNMT1 and DNMT3 was detected, suggesting the absence of DNA methylation in, at least the currently sequenced, tardigrade species. Our data show that DNA methyltransferases evolved independently and differently in the studied phyla of Ecdysozoa.

We proposed an adapted method to predict the DNA methylation status in a given species based on coding sequence (CDS) data. It was optimized over a wide phylogenetic range and requires only a single decisive parameter (the distance between the component means of a Gaussian Mixture Modelling) to achieve high specificity. Naturally, the method is limited if changes in the methylome have not yet altered the underlying genome significantly or if methylation is only present in small amounts. Our method can be easily applied to emerging model organisms since only coding sequence data is required.

The data presented here will help to guide future projects to experimentally study DNA methylation in non-model Ecdysozoa species. The proposed analysis should also be a worthwile addition to newly sequenced genomes. It allows to expand their scope from the genomic to the epigenomic level.

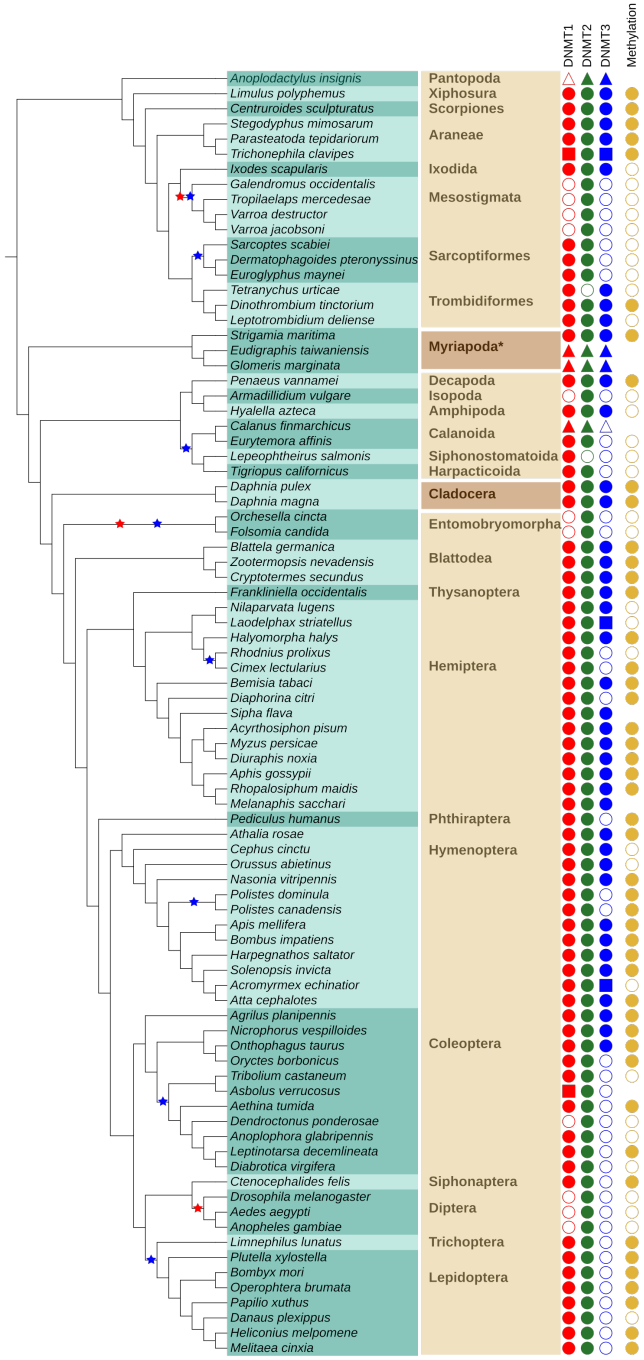

Figure 2.2: *Presence and absence of DNMT family members in Arthropoda indicated by filled and open symbols, respectively for DNMT1 (red), DNMT2 (green), and DNMT3 (blue). Data sources are indicated by symbol shape: proteome ◯, genome □, transcriptome △. The rightmost column (golden circles) shows the presence and absence of DNA methylation as predicted from the O/E CpG ratio. Absence of golden circle indicates missing data. The species list is given on turquoise background with alternating shades indicating the order membership. The name of the order (or suitable higher group marked with an asterisk *) is given in bold. Alternating shades of brown indicate (from top to bottom) Chelicerata, Myriapoda, Multicrustacea, Branchiopoda, and Hexapoda. Stars in the species tree denote proposed loss events inferred from absence of a DNMT in all species of a subtree comprising at least two leaves, disregarding absences in species with transcriptomic data only.*

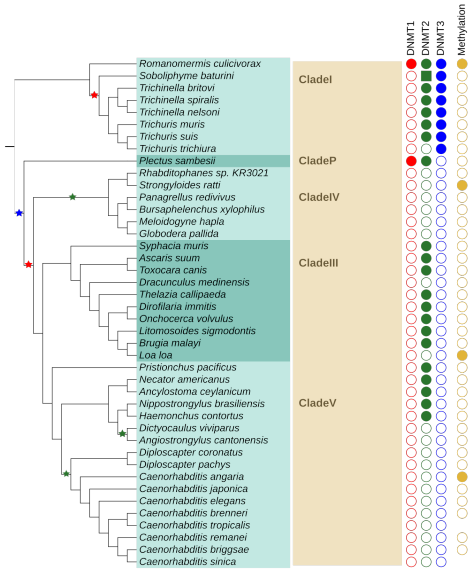

Figure 2.3: *Presence and absence of DNMT family members in Nematoda. See Fig. 2.2 for detailed legend. Instead of order names, clade names are given (in bold).*

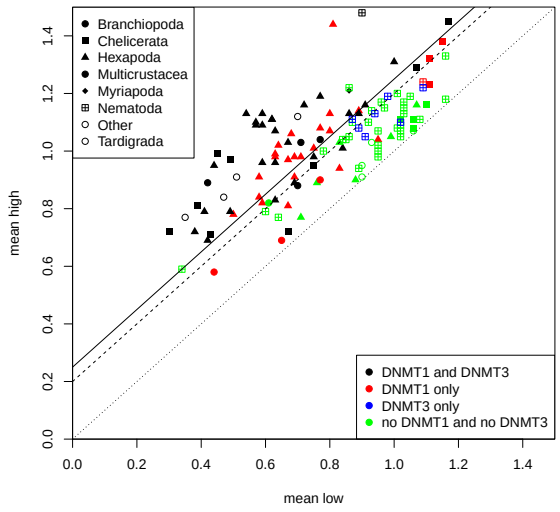

Figure 2.4: *Each point shows one species analyzed by Gaussian Mixture Modelling (GMM). The axes are the means of the two components. The taxonomic group is indicated by the style of the point. The color represents if both, DNMT1 and DNMT3 (green), have been found in the species, only DNMT1 (red), only DNMT3 (black) or neither one nor the other (blue). The diagonal lines indicate the distance between the mean of both GMM components. The dotted line indicates a distance of $d = 0$, the dashed one $d = 0.2$ and the solid line $d = 0.25$ (selected threshold).*

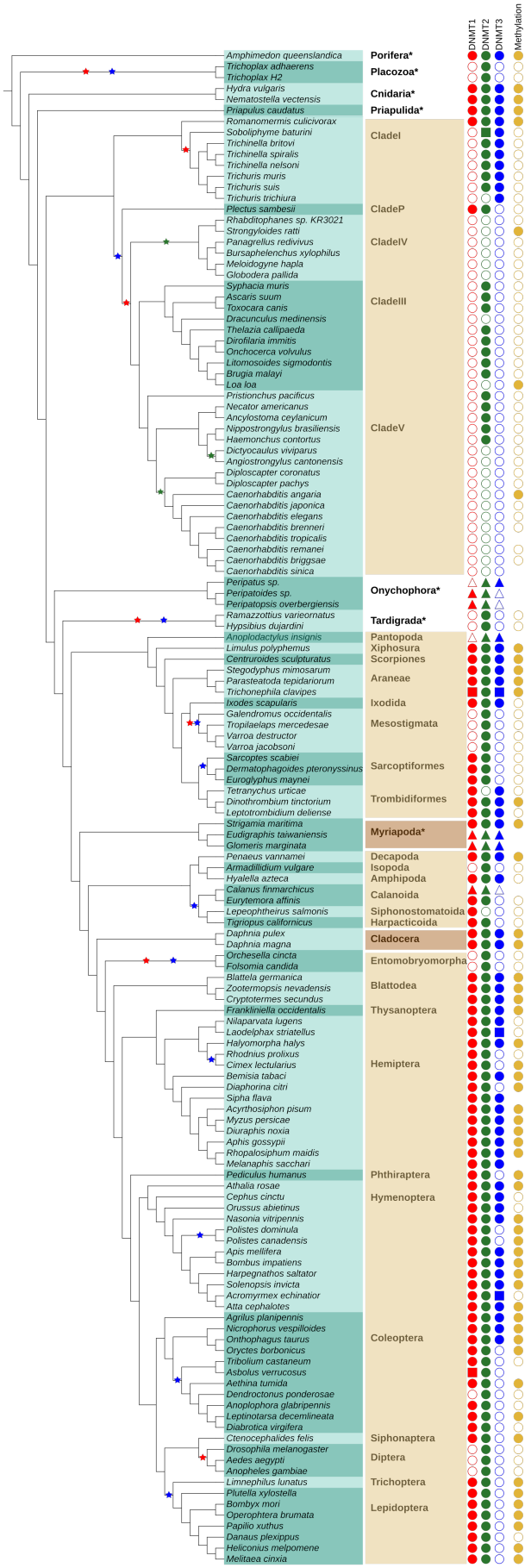

Figure 2.5: *Presence and absence of DNMT family members in Metazoa indicated by filled and open symbols, respectively for DNMT1 (red), DNMT2 (green), and DNMT3 (blue). Data sources are indicated by symbol shape: proteome ◯, genome □, transcriptome △. The rightmost column (golden circles) shows the presence and absence of DNA methylation as predicted from the O/E CpG ratio. Absence of golden circle indicates missing data. The species list is given on turquoise background with alternating shades indicating the order membership. The name of the order (or suitable higher group marked with an asterisk *) is given in bold. Alternating shades of brown indicate (from top to bottom) Nematoda, Chelicerata, Myriapoda, Multicrustacea, Branchiopoda, and Hexapoda. Stars in the species tree denote proposed loss events inferred from absence of a DNMT in all species of a subtree comprising at least two leaves, disregarding absences in species with transcriptomic data only.*

Chapter 3

Evolution of DNA methyltransferases after vertebrate whole genome duplications

3.1 Introduction

In the previous chapter we investigated the evolution of DNA methylation in Ecdysozoa. In this chapter we focus on vertebrates. Genome-wide DNA methylation co-occured with the 1R/2R whole genome duplication (WGD) and there is no known report of a loss of DNA methylation in vertebrates. Therefore, the focus of this chapter is different. We are not looking for losses of DNA methyltransferases and DNA methylation but are rather interested in the gain of additional DNA methyltransferase genes. While many additional genes after a whole-genome duplication are lost again over time it also happens that some of them are retained. Occasionally, the subfunctionalize if each of the copy only performs a part of the original function or even neofunctionalize and acquire a new function. In vertebrates there were a number of WGD. Most prominently the first and second round (1R/2R) of whole genome duplication. It happened after the split of vertebrates from tunicates [72]. Most vertebrates only underwent these two WGD but in the teleost lineage a third round of WGD occured and in some groups, e.g. Salmoniformes and some Cypriniformes even a forth WGD occured. While the Teleost-specific WGD (3R) occured already 320 million years ago (Mya) the one in Salmonid-specific WGD (4R) happened 100 mya [73]. The Carp-specific WGD (4R) is one of the most recent vertebrate WGD and is estimated to have occured 12.4 mya [74]. One example outside of teleosts is allotetraploid frog *Xenopus laevis* whose WGD occured approximately 17–18 mya [75].

In this chapter we are identifiying DNA methyltransferase genes in vertebrate species which underwent WGD. By doing so we are able to identify which of the duplicated

copies are lost or retained and can discuss if sub- or neofunctionalization is happening.

3.2 Methods

Proteome-based search The predicted proteins, CDS and gff data of the species analyzed were downloaded from different sources, see Table 6.3. For 24 species data was taken from NCBI [37]. Data for four species were retrieved from ENSEMBL [39] and for one species, *Thymallus thymallus* from the supplemental material of Varadharajan *et al.* [76]. All data was readily available to download for all species but *Oxygymnocypris stewartii*. For that species only the genome sequence was available in NCBI. Since the genomic locations of predicted CDS and protein sequences was provided in gff format in the supplemental material of the respective publication Liu *et al.* [77] it was used to extract the sequences from the genome.

The protein domain models for DNA_methylase (PF00145), ADD_DNMT3 (PF17980), CH (PF00307), PWWP (PF00855), BAH (PF01426), DMAP_binding (PF06464), DNMT1-RFD (PF12047) and zf-CXXC (PF02008) were downloaded from the "Pfam protein families database" [41]. Initially, only the DNA_methylase model was used to identify DNA methyltransferase (DNMT) candidates in the set of proteins predicted using `hmmsearch` from the HMMER software `http://hmmer.org/` version 3.2.1. Proteins with a predicted DNA_methylase domain and a full sequence e-value < 0.001 were further considered as candidates. For these, all before mentioned protein domains were annotated. Finally, each DNMT candidate was classified into one of three classes using custom perl scripts. A **DNMT1 candidate** was required not to have a PWWP or ADD_DNMT3 domain. In addition, having a DNMT1_RFD, zf-CXXC and BAH domain it was considered a *full* DNMT1 candidate, with only one of them a *partial* DNMT1 candidate. A **DNMT3 candidate** was required not to have a DNMT1_RFD, zf-CXXC or BAH domain. With both, a PWWP and a ADD_DNMT3 domain, it was considered a *full* DNMT3 candidate, with only one of them a *partial* DNMT3 candidate. In addition if a CH domain was detected the candidate was considered a *full*/*partial* DNMT3-CH candidate. A **DNMT2 candidate**, was required to have only a DNA_methylase domain and none of the other domains mentioned above.

To check if two or more DNMT candidates originate from the same genomic loci we used the existing gene anotation via the gff files. Given the protein id the corresponding gene and therefore the genomic locus was identified. Only one DNMT candidate per

locus was kept.

In a last step, the classification of the DNMT candidates was checked manually. The sequences of the DNA methylase domain of each candidate was extracted and aligned using `Clustal Omega` [42] version 1.2.4. A phylogenetic network was computed with *SplitStree4* [43] version 4.10 and inspected manually for phylogenetic congruence of gene and species phylogeny. In case of contradicting results the specific conserved sequence motifs of the methylase domain were inspected manually and the candidate reassigned to a different class or discarded if it did not contain the proper sequence motifs [26].

	CH	DNMT1-RFD	zf-CXXC	BAH	PWWP	ADD_DNMT3	DNA_methylase
DNMT1 full	0	≥ 1	≥ 1	≥ 1	0	0	1
DNMT1 partial	0	$\geq 1^*$	$\geq 1^*$	$\geq 1^*$	0	0	1
DNMT3 full	≥ 0	0	0	0	1	1	1
DNMT3 partial	≥ 0	0	0	0	$\geq 1^*$	$\geq 1^*$	1
DNMT2	0	0	0	0	0	0	1

Table 3.1: *Classification of DNMT candidates according to the detected domains. If the numbers in multiple columns of one line are marked with an asterisk (*) the condition of only one of the columns has to be fullfilled. If a DNMT3 candidate had a CH domain it is classified as DNMT3-CH full/partial.*

To get a measure for the similarity of the DNMT candidates to each other we performed a pairwise percent identity (ppi) check on thir coding sequences. The CDS sequences was downloaded for the same assembly version mentioned above. If two sequences had a ppi of more than 95% they were considered as potentially identical.

Classification of DNMT3 To distinguish the large number of DNMT3 candidates we perfomed a *SplitStree4* [43] analysis in several steps. From the Splitstree containing all DNMT candidates the split which only contained DNMT3 candidates was chosen. The Splitstree resulting from these sequences was seperated into DNMT3a and DNMT3b candidates. Based on the existing annotation of zebrafish DNMT3s the individual splits were named accordingly.

Vertebrate Phylogeny

The topology of the vertebrate phylogeny, used for display only, is based on Betancur-R *et al.* [78].

3.3 Results

In the following section the results of our analysis are presentented. A tabularized version of results of the prediction of DNA methyltransferases is shown in Table 3.2 and a graphical representation in Figure 3.1. A similar summary for the classification of DNMT3 candidates is shown in Table 3.3 and Figure 3.2.

DNA methyltransferases after the 1R/2R whole genome duplication

We analyzed 7 species which underwent the 1R/2R whole genome duplication (WGD) but no further genome duplication.

They belong to the groups Hyperoartia (lampreys), Chondrichthyes (cartilaginous fishes), Mammalia and Amphibia. Two species are from the group Actinopterygii (ray-finned fish) but outside of Teleostei, they belong to Polypteriformes (reedfish) and Lepisosteiformes (spotted gar).

In all seven species exactly one DNMT1 was detected. Only in the lamprey *Petromyzon marinus* no DNMT2 enzyme was detected. In the other six, one DNMT2 enzyme was found. In the five non-Actinopterygii species two DNMT3 enzymes were detected. In the two Actinopterygii, *Erpetoichthys calabaricus* and *Lepisosteus oculatus* two DNMT3 enzymes without a "CH" domain were detected but in addition one DNMT3 enzyme with a "CH" domain, as well.

DNA methyltransferases after the 3R whole genome duplication

We analyzed six species which underwent the 3R WGD but no additional one. *Danio rerio* and *Esox lucius* are close relatives to Cyprinidae and Salmonidae, respectively. The other four belong to the group Percomorphaceae. The two pufferfish *Tetraodon nigroviridis* and *Takifugu rubripes*, medaka *Oryzias latips* and nile tilapia *Oreochromis niloticus*.

In all six species one DNMT1 was detected. In four of them one DNMT2 was detected, in *Takifugu rubripes* and *Oreochromis niloticus* two DNMT2 enzymes were detected.

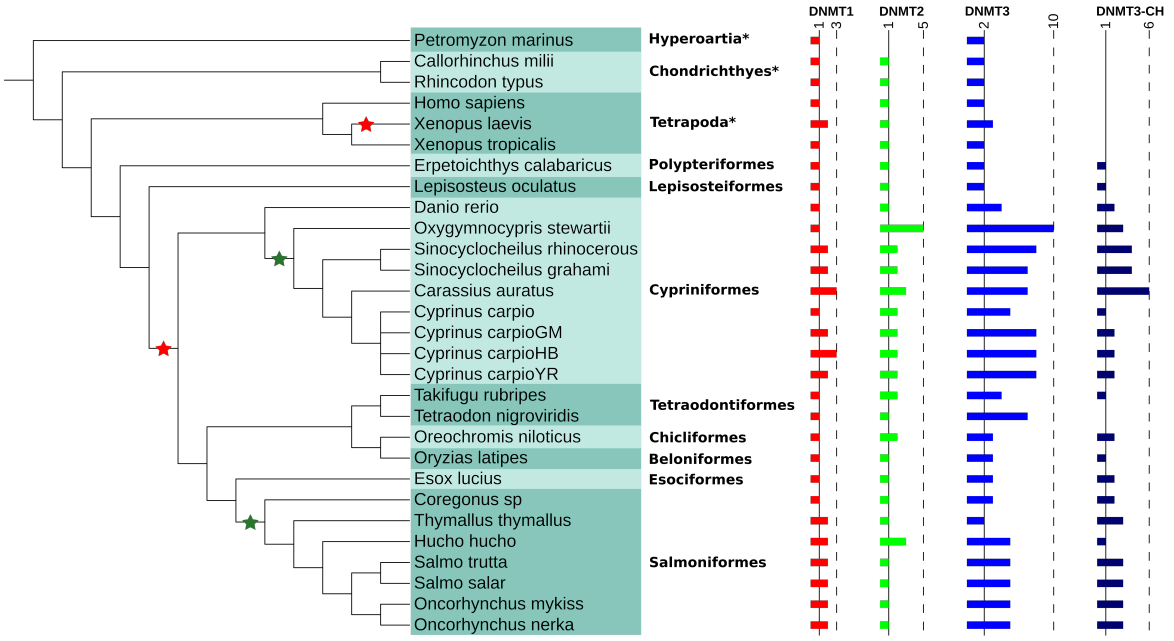

Figure 3.1: *Amount of DNMT family members in Vertebrata indicated by horizontal bar charts for DNMT1 (red), DNMT2 (green), DNMT3 (light blue) and DNMT3-CH (dark blue) which are DNMT3 candidates which contain a CH domain. The vertical scales show the amount of DNMTs at this point of the bar chart. The stars in the phylogenetic tree indicate a third (red) and forth (green) round of whole-genome duplication.*

In *Tetraodon nigroviridis* we detected seven DNMT3 enzymes without a "CH" domain, and in *Danio rerio* and *Takifugu rubripes* four. In the other three species only three of these enzymes were detected.

Two DNMT3 enzymes with a "CH" domain were detected in *Danio rerio*, *Esox lucius* and *Oreochromis niloticus*. In *Oryzias latips* and *Takifugu rubripes* only one was detected an in *Tetraodon nigroviridis* none.

DNA methyltransferases after the the 4R carp-specific whole genome duplication

Cypriniformes (carps, minnows, loaches) are one of the two orders which underwent a forth whole genome duplication (4R WGD). We analyzed eight species which underwent the carp-specific whole genome duplication (Cs4R). The common carp *Cyprinus carpio* and three subspecies hebao red carp (HR), yellow river carp (YR) and german mirror

carp (GM). The goldfish *Carassius auratus* is a close relative of them. Two species are from the genus *Sinocyclocheilus* which are cave fish only found in China. The earliest-branching relative to the beforementioned species is *Oxygymnocypris stewartii* which is found on high altitudes in Tibet.

In *Carassius auratus* and *Cyprinus carpio HB* three DNMT1 candidates were detected. In the two *Sinocyclocheilus* species, *Cyprinus carpio GM* and *Cyprinus carpio YR* we found two candidates and in *Cyprinus carpio* and *Oxygymnocypris stewartii* only one.

In *Oxygymnocypris stewartii* we detected five and in *Carassius auratus* three DNMT2 candidates, in the other six species only two.

Most DNMT3 candidates without a CH domain were detected in *Oxygymnocypris stewartii* with ten. In *Sinocyclocheilus rhinocerous* and the three carp subspecies we found eight. In *Sinocyclocheilus grahami* and *Carassius auratus* seven and in *Cyprinus carpio* five.

DNMT3 candidates with a CH domain were most abundant in *Carassius auratus* with six candidates in our analysis. In the two *Sinocyclocheilus* species we detected four candidates and in *Oxygymnocypris stewartii* three. In the *Cyprinus carpio* subspecies we could find two candidates and in the common carp itself only one.

DNA methyltransferases after the 4R salmonid-specific whole genome duplication

Salmoniformes (salmon, trout, whitefish, grayling) are the second order which underwent a forth genome duplication, the salmonid-specific whole genome duplication (Ss4R). We analyzed seven species with the Ss4R. Two *Salmo* (atlantic salmon and brown trout) and two *Oncorhynchus* (rainbow trout and sockeye salmon species. In addition the huchen *Hucho hucho*, the grayling *Thymallus thymallus* and a species of whitefish *Coregonus sp* which has not been identified to the species level.

Only in the whitefish *Coregonus sp* only one DNMT1 candidate was detected, in the other six species we found two DNMT2 candidates.

In *Hucho hucho* three DNMT2 candidates were detected, in the other species only one.

We detected three DNMT3 candidates without a CH domain in *Coregonus sp* and two in *Thymallus thymallus*. In the other five species five were detected.

We only detected one DNMT3 candidates with a CH domain in *Hucho hucho* and two in *Coregonus sp.* In the other species three candidates were detected.

Effects of the Xenopus-laevis-specific whole genome duplication

Vertebrate whole genome duplications outside of fish are even more rare. We analyzed one amphibian species with an additional round of genome duplication, the Xenopus-laevis-specific whole genome duplication (Xts3R).

In *Xenopus laevis* we detected two DNMT1 candidate, one DNMT2 candidate and three DNMT3 candidates all without a CH domain.

Potential identical DNMT candidates

In eight different species DNMT candidates with a pairwise percent identity of more than 95% were detected. In the group of Cypriniformes we detected in *Carassius auratus* two DNMT1, two DNMT2 and, in total, eight DNMT3 candidates. In *Sinocyclocheilus grahami* and *Sinocyclocheilus rhinocerous* two DNMT3 each. In the Salmoniformes we detected two DNMT1 with high identity in *Salmo trutta*, *Salmo salar*, *Oncorhynchus mykiss* and *Thymallus thymallus*. Two of such DNMT2s were detected in *Hucho hucho*. Of species without a 3R whole genome-duplication we only detected two DNMT2 candidates in *Takifugu rubripes* and two DNMT3 candidates in *Tetraodon nigroviridis* with a high identity. In Table 3.2 the potential identical candidates are indicated, as well.

Classification of DNMT3

Over time several different naming systems for DNMT3s in teleosts have been proposed, see Table 3.4. According to our own results none of them considers the evolutionary history entirely correctly therefore we propose a modified one. Based on the currently used names in the zebrafish gene annotation of Ensembl [39] we propose the following changes. DNMT3aa and DNMT3ab remain unchanged. DNMT3bb1 is changed to DNMT3b1a. DNMT3bb.2, DNMT3bb.3 and DNMT3ba are changed to DNMT3b2a1, DNMT3b2a2 and DNMT3b2b, respectively.

Species	DNMT1	DNMT2	DNMT3	DNMT3-CH
Callorhinchus milii	1	1	2	0
Rhincodon typus	1	1	2	0
Erpetoichthys calabaricus	1	1	2	1
Lepisosteus oculatus	1	1	2	1
Danio rerio	1	1	4	2
Carassius auratus	2-3	2-3	5-7	4-6
Cyprinus carpio	1	2	2/3	1
Cyprinus carpioGM	1/1	2	6/2	2
Cyprinus carpioHB	2/1	2	6/2	2
Cyprinus carpioYR	2	2	4/4	2
Sinocyclocheilus grahami	2	2	6-7	3/1
Sinocyclocheilus rhinocerous	2	2	6-7/1	3/1
Oxygymnocypris stewartii	1	1-5	7/3	3
Esox lucius	1	1	3	2
Salmo trutta	1-2	1	5	3
Salmo salar	1-2	1	4/1	3
Oncorhynchus mykiss	1-2	1	5	3
Oncorhynchus nerka	2	1	4/1	3
Hucho hucho	2	2-3	3/2	1
Thymallus thymallus	0/1-2	1	2	2/1
Coregonus sp	1	1	1/2	2
Takifugu rubripes	1	1-2	4	1
Tetraodon nigroviridis	1	1	4/1-3	0
Oreochromis niloticus	1	2	3	2
Oryzias latipes	1	1	3	1
Xenopus laevis	2	1	2/1	0
Xenopus tropicalis	1	1	1/1	0
Homo sapiens	1	1	2	0
Petromyzon marinus	0/1	0	1/1	0

Table 3.2: *The number of different DNMT candidates detected per species. DNMT3-CH indicates a DNMT3 candidate which contains a CH domain. If there are two numbers separated with a slash the first number indicates DNMT candidates with a full set of protein domains and the second number partial DNMT candidates where some domains are missing. If the amount is given as an interval there are DNMT candidates which have a pairwise percent identity of more than 95%. The lower number counts two similar candidats only once while the higher number counts all candidates, see methods section 3.2 for more details.*

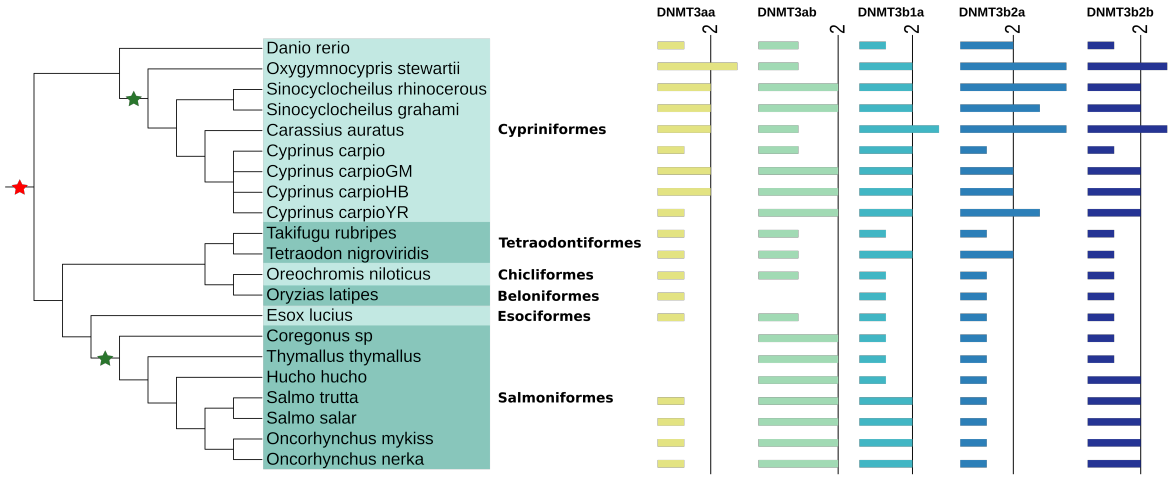

Figure 3.2: *Amount of DNMT3 family members in Teleostei indicated by horizontal bar charts for DNMT3aa (yellow), DNMT3ab (green), DNMT3b1a (turquoise), DNMT3b2a (light blue) and DNMT3b2b (dark blue). The vertical scales show the amount of DNMTs at this point of the bar chart. The stars in the phylogenetic tree indicate a third (red) and forth (green) round of whole-genome duplication.*

DNMT3 after the 1R/2R whole genome duplication In *Petromyzon marinus*, *Callorhinchus milii*, *Rhincodon typus*, *Homo sapiens* and *Xenopus tropicalis* we detected one DNMT3a and one DNMT3b candidate. In the other species which underwent additonal gene or genome duplications the pattern is more complex.

DNMT3 after the 3R whole genome duplication In the six species which underwent the 3R WGD but no additional one the prevalent pattern are five DNMT3s. Two of these (DNMT3aa and DNMT3ab) originate from DNMT3a. The other three (DNMT3b1a, DNMT3b2a and DNMT3b2b) originate from DNMT3b. There are a few exceptions. In *Danio rerio* two copies of DNMT3b2a were detected, in *Tetraodon nigroviridis* two copies of DNMT3b1a and DNMT3b2a, each and in *Oryzias latipes* no copy of DNMT3ab was detected.

DNMT3 after the 4R carp-specific whole genome duplication In *Cyprinus carpio GM* and *Cyprinus carpio HB* there are two DNMT3 copies from each of the five groups. In *Cyprinus carpio YR* the pattern is similar but only one copy of DNMT3aa was detected while for DNMT3b2a there were three. In *Cyprinus carpio* there was one copy for each of the five DNMT3s only of DNMT3b1a two copies were detected.

In *Carassius auratus* two DNMT3aa candidates and one DNMT3ab candidate were detected. In addition three copies of DNMT3b1a and DNMT3b2b, each, and four of DNMT3b2a. Two copies of each DNMT3 were also detected in *Sinocyclocheilus grahami* and *Sinocyclocheilus rhinocerous* with the exception of DNMT3b2a of which three copies were identified in the first mentioned species and four copies in the latter. In *Oxygymnocypris stewartii* we identified three copies of DNMT3aa and one of DNMT3ab. In the DNMT3b group we detected two copies of DNMT3b1a, four of DNMT3b2a and three of DNMT3b2b.

DNMT3 after the 4R salmonid-specific whole genome duplication In *Salmo trutta*, *Salmo salar*, *Oncorhynchus mykiss* and *Oncorhynchus nerka* we detected the same distribution of DNMT3 candidates with one DNMT3aa, two copies of DNMT3ab, DNMT3b1a and DNMT3b2b and one DNMT3b2a. *Hucho hucho*, *Thymallus thymallus* and *Coregonus sp.* shared the same pattern of DNMT3s as well. We detected in these species no DNMT3aa, two copies of DNMT3ab and only one DNMT3b1a, DNMT3b2a and DNMT3b2b. The only exception was an additional copy of DNMT3b2b in *Hucho hucho*.

DNMT3 after the Xenopus-laevis-specific whole genome duplication In *Xenopus laevis* two DNMT3a candidates and one DNMT3b candidate were detected.

3.4 Discussion

The evolution of DNA methyltransferases in fish has been investigated in several studies until now [79, 80, 81]. Some of the species we analyzed were part of this studies. However, only two species which underwent a 4R whole genome-duplication (WGD) have been in the focus until now: *Oncorhynchus mykiss* and *Salmo salar* in Liu *et al.* [81]. We included, in total, 15 species which underwent a 4R WGD. Most importantly they are from the groups of Salmoniformes as well as Cypriniformes which both underwent a 4R WGD independently. Aside from a detailed study of these groups we also included the earliest branching Actinopterygii investigated for DNA methyltransferases so far, the reedfish *Erpetoichthys calabaricus*. Therefore, this study is the most comprehensive analysis of the evolution of DNA methyltransferases after different WGD in vertebrates so far.

	Gene name				
Species	DNMT3aa	DNMT3ab	DNMT3bb.1	DNMT3bb.2/3	DNMT3ba
	DNMT3aa	DNMT3ab	DNMT3b1a	DNMT3b2a	DNMT3b2b
Danio rerio	1	1	1	2	1
Carassius auratus	2	1	3	4	3
Cyprinus carpio	1	1	2	1	1
Cyprinus carpio GM	2	2	2	2	2
Cyprinus carpio HB	2	2	2	2	2
Cyprinus carpio YR	1	2	2	3	2
Sinocyclocheilus grahami	2	2	2	3	2
Sinocyclocheilus rhinocerous	2	2	2	4	2
Oxygymnocypris stewartii	3	1	2	4	3
Esox lucius	1	1	1	1	1
Salmo trutta	1	2	2	1	2
Salmo salar	1	2	2	1	2
Oncorhynchus mykiss	1	2	2	1	2
Oncorhynchus nerka	1	2	2	1	2
Hucho hucho	0	2	1	1	2
Thymallus thymallus	0	2	1	1	1
Coregonus sp	0	2	1	1	1
Takifugu rubripes	1	1	1	1	1
Tetraodon nigroviridis	1	1	2	2	1
Oreochromis niloticus	1	1	1	1	1
Oryzias latipes	1	0	1	1	1

Table 3.3: *The number of different DNMT3 candidates detected per species. DNMT3bb.2/3 combines the locally duplicated zebrafish genes DNMT3bb.2 and DNMT3bb.3. DNMT3ba and DNMT3bb.2/3 contain a CH domain in most species.*

Evolution of DNA methyltransferases

Our results of the general evolution of DNA methyltransferases mainly support previous publications. After the 1R/2R whole-genome duplication (WGD) there is a local duplication of DNMT3b in the Actinopterygii lineage. Subsequently, one of the copies acquires a Calponine homology (CH) domain. After the 3R WGD only one copy is lost universally, DNMT3b1b, the others are kept in most species with lineage-specific losses in some. After the fourth WGD (4R) there are lineage-specific losses of some of the additonal copies but the number of different DNMT3 genes stays higher compared to teleost species without a fourth WGD.

Carp-specific 4R whole-genome duplication (Cs4R)

Our study provided the first analysis of the evolution of DNA methyltransferases after the carp-specific WGD. It is one of the most recent WGD in vertebrates and occured only 12.4 million years ago (mya). The one in Salmoniformes occured almost 90 million years earlier. As one would expect there are more genes still retained compared to

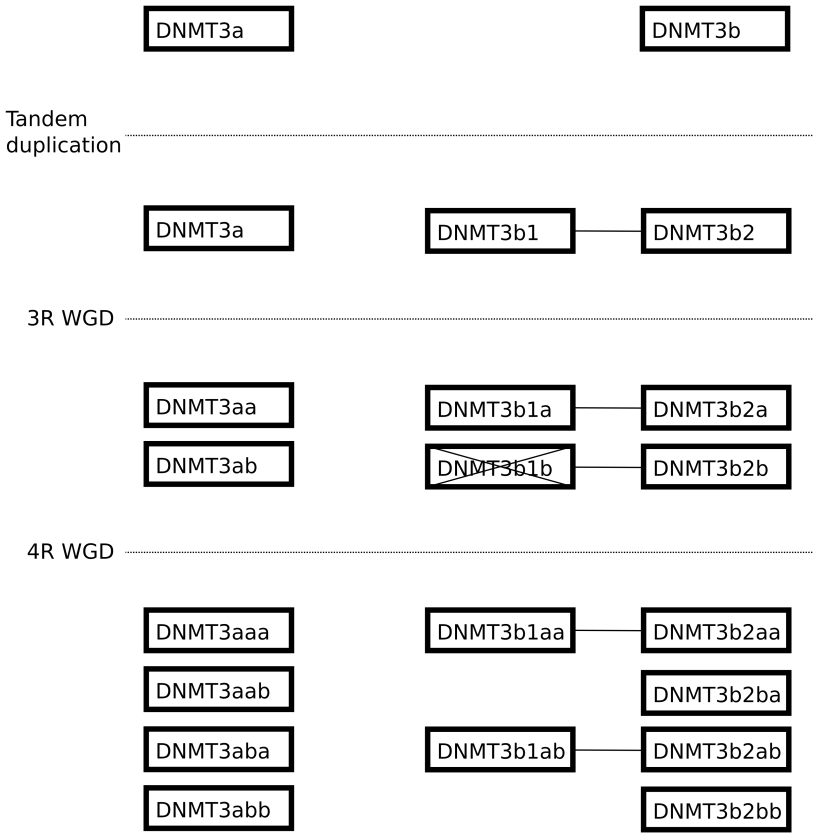

Figure 3.3: *The evolutionary history of DNMT3 in Actinopterygii. It starts with DNMT3a and DNMT3b which resulted from the 1R/2R whole genome duplication. After the split from Sarcopterygii a local duplication of DNMT3b happend. DNMT3b2 gained an additional CH domain before the 3R whole genome duplication (WGD). After the 3R WGD only one DNMT3 (DNMT3b1b) is lost universally. After the independent 4R WGDs paralogs were lost differently in different groups.*

Salmoniformes. We can see in several species that all copies which originated from WGD are still retained. Opposite to Salmoniformes there also seems to be no clear pattern which copies in the DNMT3a or DNMT3b groups is preferentially lost. The most heterogenous pattern is detected for DNMT3b2a (ex. DNMT3bb.2/.3). This is a complicated case since it is known that in the zebrafish *Danio rerior* a local duplication of DNMT3b2a is present. *Danio rerior* belongs to the Cypriniformes but did not underwent the Carp-specific 4R WGD. It is currently not clear at which point in the evolution of Cypriniformes this local gene duplication happened. Five of the eight species with the 4R WGD have three or more copies of this gene which would indicate that before the WGD already more than one copy was present. In addition

five of the eight species have a "full" DNMT3b2a candidate without a CH domain. A likely scenario would be that the local duplication and successive loss of the CH domain in one of the orthologs happened before the carp-specific WGD. To resolve this it should be informative to investigating more early-branching Cypriniformes.

The Cs4R WGD happened and the respective species are therefore still tetraploid. In *Cyprinus carpio* for example 50 chromosomes have been detected. While sequencing technologies have improved significantly in the last years assembling such genomes is still a challenge. Therefore, it is more difficult to prevent assembly errors compared to diploid species. We noticed that some of the detected DNMT candidates showed a very high pairwise identity. Our prediction of DNMT candidates is based on predicted proteins, therefore we already included in the prediction pipeline that two candidates can not originate from the same genomic loci. If they are overlapping only one candidate is kept. But since assembly errors might be more common in teleost genomes we calculated the pairwise percent identity between the coding sequence of the DNMT candidates of each group. Due to synonymous mutations in coding sequences they can have quite some differences without any changes to the amino acid sequence. We nevertheless considered only an identity of more than 95% to be almost identical. Such a high similiarity was found most frequently in cypriniformes. Unfortunately, even if the identity is very high it is still difficult to decide if this is a technical artifact or the biological reality. For example even in three Salmoniformes species two DNMT1 genes have a pairwise percent identity of more than 95%. Therefore, we decided to not remove candidates which originate from different genomic loci but rather inform about their high identity. If these findings were cause bad assembly errors they should be resolved if genome assemblies with better quality become available.

Salmoniformes-specific 4R whole-genome duplication (Ss4R)

Our results are concordant with previous studies but we are able to extend their findings significantly. The genus *Salmo* and *Oncorhyncus* show a very homogenous pattern of DNMT3 distribution. Outside of these groups there is a higher number of lineage-specific losses of DNMT3 copies. In the other studied Salmoniformes, *Hucho hucho*, *Thymallus thymallus* and *Coregonus sp.*, we could not detect any DNMT3aa indicating that there was no strong subfunctionalization since DNMT3ab, or possibly DNMT3b's as well, can compensate the loss. For the DNMT3b genes only in *Hucho*

huch a second copy of DNMT3b2b was detected the other duplications originating from the Ss4R seem to have been lost. Given the current phylogeny of Salmoniformes all of these losses happened independent from each other in their respective lineage. This secenario is not very parsimonious but nevertheless possible. Lien *et al.* [82] use a different Salmoniformes phylogeny in their publication about the atlantic salmon genome. It groups together the genus *Coregonus+Thymallus* as well as *Hucho+Salmo+Oncorhynchus*. This would make it likely to have shared losses in the lineage leading to *Coregonus+Thymallus*.

Renaming of Teleost DNA methyltransferases

Given the evolutionary history of DNA methyltransferases in teleosts we believe the DNMT3 gene names should be slightly altered to correctly represent the evolutionary history. There is a "ZFIN Zebrafish Nomenclature Convention" `https://wiki.zfin.org/display/general/ZFIN+Zebrafish+Nomenclature+Conventions` representing a community standard for gene names in zebrafish. Since zebrafish is the most studied teleost these conventions are ofter used for other teleost species as well. In this standard it is recommended to distinguish if a gene duplication has occured from a genome-wide duplication or from a tandem duplication. In the first case the letters "a","b" should be added to the gene name and in the latter case the symbols ".1",".2". However, tandem duplications are only defined if "a single mammalian orthologue" is present, which is not the case for the DNMT3s in question. Therefore, we opted for using digits without a dot, e.g. "1","2". In the history of zebrafish DNMT research the gene names have changed quiet often. The most commonly used names from the Ensembl gene annotation do not reflect the evolutionary history very well. The most recently suggested nomenclature by Liu *et al.* [81] captures the evolutionary history in the species they analyzed rather well. They also state: "We thus hypothesised that the ancestral dnmt3b duplicated at VGD2, both duplicates were fixed at least in holeostei, whereas one copy was lost in gnathostomes. Although we cannot exclude the possibility that the dnmt3ba/bb might arise from a punctual duplication occurred in ancestral holeostei, we are much in favour of the former hypothesis." [81]. However, throughout the manuscript it does not become clear why exactly the are in favor of the hypothesis that DNMT3ba/bb occured during a whole genome duplication. DNMT3b1a and DNMT3b2a frequently occour on the same scaffold in close vicinity. This is also the

case in the earliest branching Actinopterygii analyzed, i.e. *Erpetoichthys calabaricus Lepisosteus oculatus*. This can easily be explained by a local gene duplication after which the genes are located next to each after. After a whole-genome duplication, however, the genes are located on different chromsomes and during chromosome rearrangement would have to be reordered next to each other. Without further evidence for the hypotesis of Liu *et al.* [81] we believe it to be a more parsimonious hypothesis that DNMT3ba and DNMT3bb occoured from a local gene duplication. Consequently, the nomenclature we propose differs from theirs in this aspect, see Table 3.4

Gain of Calponin homology (CH) domain

The presence of a CH domain in teleost DNA methyltransferase has been known since more than 15 years [79] but their function is still unknown. The calponin homology (CH) domain is associated with actin binding. In recent years actin-binding has been found to be associated with several gene regulatory mechanisms like chromatin remodelling and transcription. But actin is correlated with activation of gene regulation [83, 84] instead of deactivation like DNA methylation. Therefore, which role DNA methyltransferases which acquired a CH domain play in these mechanisms still remains unclear. However, we were able to clarify at which point of Actinopterygii evolution calponin homology domains were first introduced into DNA methyltransferases.

We detected the earliest occurence of a DNMT3 with a calponin homology (CH) domain domain currently known. It was already reported in the spotted gar *Lepisosteus oculatus* [81]. We found it in the reedfish *Erpetoichthys calabaricus*, as well. The spotted gar belongs to the group Holostei, together with Teleostei they form the Neopterygii. If only these groups contained a CH domain it would be likely that it occured for the first time between 350 and 325 million years ago (mya). Approximately, 350 mya ago the split between Neopterygii and Chondrostei (sturgeon, paddlefish) took place according to Betancur-R *et al.* [78]. Reedfish belong to the group Cladistia. By detecting a CH domain in a Cladistia the origin of the CH domain is at the base of Actinopterygii. It most likely occured after the split from Sarcopterygii (coelacanth, tetrapods) before the diversification of Actinopterygii, This dates back to appr. 425 to 380 mya. Given our hypothesis that DNMT3ba and DNMT3bb originate from a local gene duplication it would have occured in the same time frame.

We did not analyze a genome from the group Chondrostei but our results suggest

Species	Ensembl	Shimoda [79], 2005	Campos [80], 2012	Liu [81], 2020	Engelhardt, 2020
L. oculatus	DNMT3aa DNMT3bb.1 DNMT3ba			DNMT3a DNMT3ba DNMT3bb	DNMT3a DNMT3b1 DNMT3b2
D. rerio	DNMT3aa DNMT3ab DNMT3bb.1 DNMT3bb.2 DNMT3bb.3 DNMT3ba	DNMT8 DNMT6 DNMT4 DNMT3 DNMT5 DNMT7	DNMT3a2 DNMT3a1 DNMT3b1 DNMT3b3 DNMT3b4 DNMT3b2	DNMT3aa DNMT3ab DNMT3ba DNMT3bbb1 DNMT3bbb2 DNMT3bba	DNMT3aa DNMT3ab DNMT3b1a DNMT3b2a1 DNMT3b2a2 DNMT3b2b
O. mykiss				DNMT3aa DNMT3ab1 DNMT3ab2 DNMT3ba1 DNMT3ba2 DNMT3bba1 DNMT3bba2 DNMT3bbb	DNMT3aa DNMT3aba DNMT3abb DNMT3b1aa DNMT3b1ab DNMT3b2ba DNMT3b2bb DNMT3b2aa
C. carpio GM					DNMT3aaa DNMT3aab DNMT3aba DNMT3abb DNMT3b1aa DNMT3b1ab DNMT3b2aa DNMT3b2ab DNMT3b2ba DNMT3b2bb

Table 3.4: *An overview of the different naming systems proposed for DNA methyltransferases 3 over time. The Ensembl gene names from column 1 are most frequently used at the moment. The names shown in the last colum is the most correct nomenclature according to this study.*

that its species should contain at least two DNMT3b genes, as well. One of them with a CH domain.

Conclusion

We performed the most comprehensive analysis of the evolution of DNA methyltransferases after vertebrate whole-genome duplications (WGD) so far. We were able to show that the conservation of duplicated DNMT3 genes in Salmoniformes is more diverse than previously believed. We were also able to identify DNA methyltransferases in Cypriniformes which have, due to their recent WGD, quite complex genomes. Our results show that the patterns of retained and lost DNA methyltransferases after a forth round of WGD differ between Cypriniformes and Salmoniformes.

An urging question still remains, why are so many additional copies of DNMT3 genes are kept. However, as we will see in the next chapter studying the functions of the individual DNA methyltransferases in the presence of so many copies is quite difficult even in zebrafish which only underwent the 3R WGD.

We hope that our results can help future projects pursuing similar projects in other teleost species.

Chapter 4

The effect of DNMT3aa and DNMT3ab knockout on DNA methylation in zebrafish

4.1 Introduction

In the previous chapter we investigated the evolutionary history of DNA methyltransferases (DNMTs) in vertebrates. As one could see in the teleost lineage especially duplicated DNMT3s were often retained in the genome. In the case of the zebrafish *Danio rerio*, for example, there are 5 copies of the gene DNMT3. If both copies of a duplicated gene are retained a common question is if there are alterations to their function over time. If this is the case one distinguishes between subfunctionalization and neofunctionalization. Subfunctionalization describes that a gene changes its function a subset of what it was before, e.g. it is only expressed in a subset of the cell types where it was present before or targets only a subset of the genes it was targeting before. Different from that is neofunctionalization in which a duplicated gene acquires a completely new function. Sometimes it is not easy decidable at which point a function should be considered "new". There is also an argument that subfunctionalization is only a transition state which is always followed by neofunctionalization [85]. The matter is complicated by the fact that "biological function" can be difficult to define since a general theoretical framework for it does not exist [86].

Together with the Aluru lab we have been trying to investigate possible functional differences of DNMT3 genes in zebrafish.

4.2 Methods

DNA methyltransferase knockout

Zebrafish which did not express a functional version of a specific DNA methyltransferase (DNMT) gene were produced using the transcription activator-like effector nucleases (TALEN) method in the Aluru lab in Woods Hole, USA. Zebrafish with a knockout for either DNMT3aa or DNMT3ab were generated. Subsequently, these knockouts were crossed to produce double knockout specimen.

Generating the DNA methylation data

Zebrafish of the three different knockouts as well as normal specimen (wildtype) were processed. We used whole embryos ten hours post fertilization and sequenced five individuals per condition resulting in twenty sequencing experiments in total. The sequencing was performed following the enhanced reduced representation bisulfite sequencing (eRRBS) library preparation on genomic DNA. Subsequently, a 50 bp paired-end sequencing on an Illumina HiSeq2500 platform was performed.

Processing the DNA methylation data The quality of the data was checked with `FastQC` [87]. It visualizes several quality parameter of high-throughput data, e.g. length and quality score distribution of the reads. Subsequently, the reads were trimmed using the "–rrbs" mode of `Trim galore` [88]. It removes adapter sequences which are artificially introduced during the sequencing procedure and occassionaly become part of the resulting reads.

We used the `Bisulfite Analysis Toolkit (BAT)` [89] to further process the eRRBS-seq data. It is a data analysis pipeline which mainly combines `segemehl` [90, 91] for mapping bisulfite-treated sequence data to the genome and `metilene` [92] to detect differentially methylated regions (DMRs).

The mapping was done using the bisulfite methyl-C seq mode of `segemehl` included in `BAT`. As a reference genome the zebrafish (*Danio rerio*) genome GRCz10/danRer10 was used. The methylated cytosines were called with `BAT_calling`. It calculates the methylation rate for each cytosine by calculating the relative amount of unmodified cytosines: $\#C/(\#C + \#T)$. These cytosines were filtered with `BAT_filter_vcf` to keep only the ones occuring in a CpG context and with a minimum coverage of 10

and a maximum of 100 reads. `BAT_summarize` prepared the input data for `metilene`. In this step the individual experiments are assigned to two different groups, between which the differential methylation will be called. In our case the two groups were the wildtype as the background and one of the knockouts as the experiment. Finally, the script `BAT_DMRcalling` was used. It executes `metilene` to call differentially methylated regions (DMRs). Standard settings were used which require the DMRs to contain at least 10 CpGs and a minimum difference of 0.1 between the mean methylation rates per group. DMRs with a q-value below 0.05 were considered to be significant.

Corresponding genes and GO annotation (GREAT) To associate DMRs with a gene we used the `Genomic Regions Enrichment of Annotations Tool (GREAT)` version 2.0.2 [93]. It predicts which gene is influenced by a cis regulatory element, in our case a DMR. This is done as a two step process. In the first step a regulatory domain of 5 kilobases (kb) upstream and 1 kb downstream of its TSS is assigned to each gene. In the second step each genomic region, e.g. DMR, is associated with all genes whose regulatory domains it overlaps. There are different methods to define the regulatory domain. We used the default method "basal plus extension" with the default values of 5 kb/1 kb for the basal regulatory domain and 1000 kb for the extension. It assigns a "basal regulatory domain" and ignores the presence of other genes. This regulatory domain is subsequently extended upstream and downstream by up to 1000 kb in each direction or up to the "basal regulatory domain" of the neighboring genes. Therfore, regulatory domains can overlap but only if the TSS of two neighboring genes are closer than 5 kb upstream or 1 kb downstream of each other. Since `GREAT` only supports the *Danio rerio genome* version Zv9/danRer7 we used the UCSC genome browser `liftOver` utility (`https://genome.ucsc.edu/cgi-bin/hgLiftOver`) to convert the coordinates of the DMRs from GRCz10/danRer10 to Zv9/danRer7. Once we had corresponding genes assigned to the DMRs we used the gene ontology (GO) anntation (biological processes, molecular function and KEGG pathway) of these genes to perform an enrichment analysis.

4.3 Results

DNA methylation profiling

The eRRBS-seq data of the genomic DNA was processed as described above. The sequencing was successful for all twenty specimen. Therefore, for each of the four conditions (wildtype and 3 knockouts) five replicates were available. On average appr. 22 million paired-end reads were generated per experiment. 16.3 million reads were the lowest and 29.7 million reads the highest outcome. The mapping rate was quite good with 91%, see Table 4.1 for more details.

Condition	Sample ID	Number of reads	Number of mapped reads
Wildtype	WT1	21,160,574	19,717,083 (93.18%)
Wildtype	WT2	21,175,283	19,780,188 (93.41%)
Wildtype	WT3	23,872,403	22,426,592 (93.94%)
Wildtype	WT4	29,683,913	28,176,471 (94.92%)
Wildtype	WT5	23,909,903	22,552,118 (94.32%)
DNMT3aa KO	3aa-1	24,198,956	22,204,611 (91.76%)
DNMT3aa KO	3aa-2	25,098,450	23,509,693 (93.67%)
DNMT3aa KO	3aa-3	20,870,473	19,750,963 (94.64%)
DNMT3aa KO	3aa-4	25,150,651	23,776,847 (94.54%)
DNMT3aa KO	3aa-5	17,237,368	16,186,735 (93.90%)
DNMT3ab KO	3ab-1	22,609,979	21,474,960 (94.98%)
DNMT3ab KO	3ab-2	19,567,134	18,442,493 (94.25%)
DNMT3ab KO	3ab-3	28,895,434	27,291,066 (94.45%)
DNMT3ab KO	3ab-4	21,548,010	19,869,057 (92.21%)
DNMT3ab KO	3ab-5	21,548,010	20,169,030 (93.60%)
DNMT3aa/ab KO	3aa-3ab-1	21,225,774	19,893,830 (93.72%)
DNMT3aa/ab KO	3aa-3ab-2	18,356,167	14,102,840 (76.83%)
DNMT3aa/ab KO	3aa-3ab-3	16,335,695	10,642,477 (65.15%)
DNMT3aa/ab KO	3aa-3ab-4	17,715,731	15,280,551 (86.25%)
DNMT3aa/ab KO	3aa-3ab-5	20,564,602	17,685,953 (86.00%)
Average		22,036,226	20,146,678 (91.43%)

Table 4.1: *#reads - one read consists of two pairs or fragments;*

Due to the bisulfite sequencing we can count, for each cytosine position in the genome, the methylation rate by determining the fraction of methylated cytosines (sequenced as cytosines) among all sequenced cytosines or thymines at this position. Of all cytosines in a CpG context on average 89% had a methylation level larger than

Sample ID	#un-mCpG	#mCpG	mCpG level
WT1	278,214 (17%)	1,407,748 (83%)	78%
WT2	249,772 (11%)	1,926,294 (89%)	79%
WT3	239,860 (10%)	2,274,156 (90%)	83%
WT4	224,516 (09%)	2,183,770 (91%)	82%
WT5	169,760 (09%)	1,693,395 (91%)	83%
3aa-1	252,751 (16%)	1,321,380 (84%)	78%
3aa-2	251,953 (16%)	1,356,384 (84%)	79%
3aa-3	230,066 (10%)	2,129,666 (90%)	82%
3aa-4	224,109 (09%)	2,167,617 (91%)	82%
3aa-5	196,231 (09%)	2,098,945 (91%)	84%
3ab-1	220,629 (10%)	2,054,925 (90%)	82%
3ab-2	218,519 (10%)	2,035,202 (90%)	82%
3ab-3	232,484 (09%)	2,228,808 (91%)	83%
3ab-4	263,002 (17%)	1,329,127 (83%)	78%
3ab-5	210,102 (10%)	1,957,408 (90%)	82%
3aa-3ab-1	267,073 (15%)	1,458,157 (85%)	80%
3aa-3ab-2	226,386 (11%)	1,757,310 (89%)	80%
3aa-3ab-3	244,615 (16%)	1,332,924 (84%)	74%
3aa-3ab-4	243,546 (12%)	1,774,514 (88%)	79%
3aa-3ab-5	251,996 (12%)	1,819,586 (88%)	79%
Average	234,779 (11%)	1,815,370 (89%)	80%

Table 4.2: *#reads - one read consists of two pairs or fragments;*

zero. The average genome-wide methylation level was 80%, see 4.2 for more details. The global methylation level in the wildtype and the single knockouts were very similar only in the double knockout it was slighyl lower, see Figure 4.1 for a visualization.

Differentially methylated regions

In the DNMT3aa knockout we detected a total of 103 differentially methylated regions (DMR). The largest hypermethylation was by 48% while the largest hypomethylation was a loss of 66%.

23 of the 103 DMRs had a q-value smaller than 0.05 and therefore have been considered as significant DMRs. 18 DMRs were hypermethylated with a methylation difference between +36% and +48%. The other five were hypomethylated with a methylation loss between -23% and -66%. Interestingly, seven of the DMRs are located on chromosome 4 (four hyper- and one hypomethylated). On the other 10 chromosomes

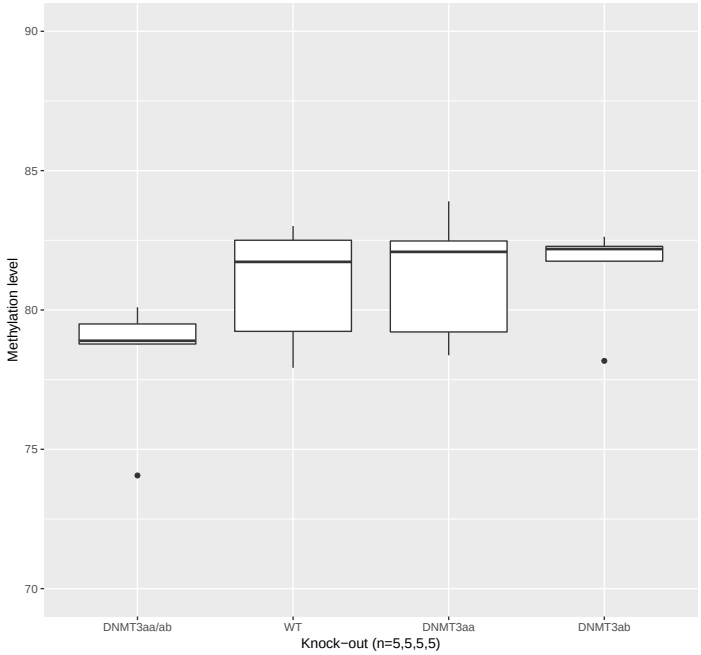

Figure 4.1: *Boxplots of the global CpG methylation level of each experiment (see Tab. 4.1 4th column). The experiments are sorted according to their median from left to right: DNMT3aa/ab double KO, Wildtype, DNMT3aa KO, DNMT3ab KO.*

and one scaffold the highest amount of detected DMRs was three.

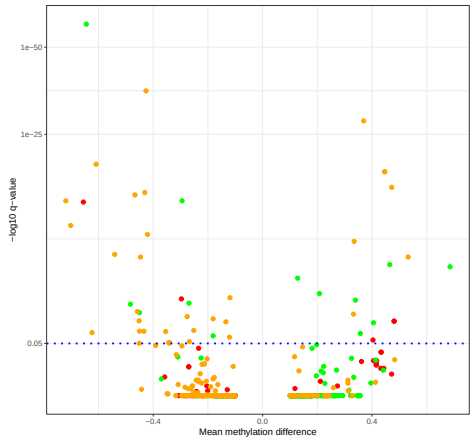

Figure 4.2: *The volcano plot shows on the x-axis the methylation difference and on the y-axis the q-value. Each circle is a DMR predicted by metilene in the DNMT3aa KO (red), DNMT3ab KO (green) or DNMT3aa/ab double KO (orange). The blue dashed line indicates a q-value of 0.05, DMRs below this line were considered significant. DMRs with a methylation difference smaller than 0 are hypomethylated and otherwise hypermethylated.*

Using the `Genomic Regions Enrichment of Annotations Tool (GREAT)` [93] we were able to assign 39 genes to 21 of the 23 DMRs. None of the DMRs had more than two genes assigned. The genes corresoponding to hypermethylated DMRs were assigned to a number of GO terms, e.g. chloride transmembrane transport, RNA polyadenylation and iron-sulfur cluster assembly. For the hypomethylated DMRs the gene crebl2 was associated to a number of "postive regulation" terms, e.g. fat cell differentiation, lipid biosynthetic process or glucose import, see Supplemental table 6.12 for a list of GO terms. Hypermethylated DMRs were assigned to a number of molecular function terms, the ones with the highest number of genes invovled were cation binding (13 genes), ion binding (18 genes) and metal ion binding (12 genes). Hypomethylated DMRs were only associated with hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, see supplemental table 6.13 for more details.

In the DNMT3ab knockout we detected 208 DMRs. The largest hypermethylation was +69% methylation and the largest hypomethylation -64% methylation.

24 of the 208 DMRs had a significant q-value. 16 DMRs were hypermethylated with a methylation difference between +69% and +13%. The other eight were hypomethylated with a methylation loss between -18% and -64%. The DMRs were evenly

distributed among 18 chromosomes.

For 22 of the 24 DMRs a corresponding gene was detected. Most of them corresponded to two genes in total there were 41 different genes. Hypermethylated DMRs were enriched for biological process GO terms like detection of gravity, defense response to fungus or pigment granule dispersal. The highest amount of genes was associated to the, not very specific, term cellular process (20 genes). Hypomethylated DMRs were associated to the positive regulation of fat cell differentiation, similar to the ones in the DNMT3aa knockout. Five genes were also associated to a number of terms related to different compound metabolic process, see supplemental table 6.14 for more details. In the molecular function terms palmitoyltransferase activity, carboxylic acid binding and glycine binding were among the highst scoring ones for hypermethylated DMRs. Hypomethylated DMRs were, as in the DNMT3aa knockout associated to hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds but also to hydrolase activity and receptor binding, see supplemental table 6.15 for more details.

In the DNMT3aa and DNMT3ab double knockout we detected 214 DMRs. The largest hypermethylation was +53% methylation and the largest hypomethylation -72% methylation.

47 of the 214 DMRs had a significant q-value. 11 DMRs were hypermethylated with a methylation difference between +53% and +12%. The other 36 were hypomethylated with a methylation loss between -11% and -72%. The DMRs were distributed among 18 chromosomes and three scaffolds. Nine DMRs were located on chromosome 4 (8 hypo- and 1 hypermethylated).

42 of the 47 DMRs were corresponding to mostly two genes each and 77 genes in total. Hypermethylated DMRs were associated with rhythmic process, RNA metabolic process or female somatic sex determination. Hypomethylated DMRs were, once again, associated to positive regulation of fat cell differentiation but also to DNA methylation involved in gamete generation and parasympathetic nervous system development, see supplemental table 6.16 for more details. Molecular function terms of hypermethylated DMRs were for example serine C-palmitoyltransferase activity and ubiquitin conjugating enzyme binding (GO:0031624). Of the genes corresponding to hypomethylated DMRs a large number was associated with heterocyclic compound binding (29 genes) and nucleic acid binding (21) or more specifically piRNA binding (1 gene) and tRNA binding (1 gene), see supplemental table 6.17 for more details.

The overlap between the significant DMRs of all three datasets is relatively scarce,

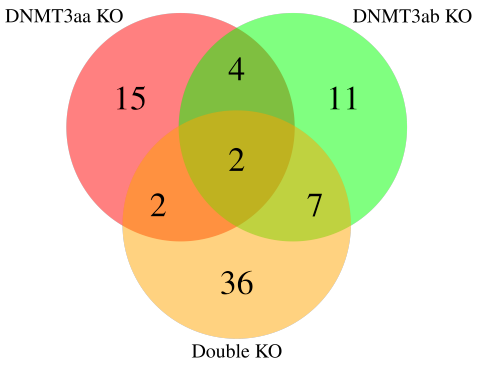

Figure 4.3: *Overlap of the the differentially methylated regions (DMRs) in the DNMT3aa knockout (red), the DNMT3ab knockout (green) and the DNMT3aa/DNMT3ab double knockout.*

see Figure 4.3 for a graphical representation.

Four DMRs reported in the DNMT3aa and DNMT3ab knockout are overlapping. In both knockouts they show the same methylation difference, two are hypermethylated and two are hypomethylated.

The DNMT3aa single knockout and the double knockout share two DMRs. They are hypomethylated in both conditions.

The DNMT3ab single knockout and the double knockout share seven DMRs. In both knockouts they show the same methylation difference, three are hypermethylated and four are hypomethylated.

Only two DMRs are detected in all three datasets, both are hypomethylated. One is located on chromosome 12 and corresponds to the genes bnip3 and dpysl4. The other is on chromosome 4 and corresponds to the gene crebl2. Only in the DNMT3aa knockout it corresponds to gpr19, as well. Appr. 500 nt upstream of that DMR there is a second one which is only detected in the DNMT3ab KO and the DNMT3aa/ab double KO. The second DMR corresponds to gpr19, as well.

The overlap of the corresponding genes between the three knockout conditions was correlated strongly to the overlap of the DMRs, see Figure 4.4 for a graphical representation.

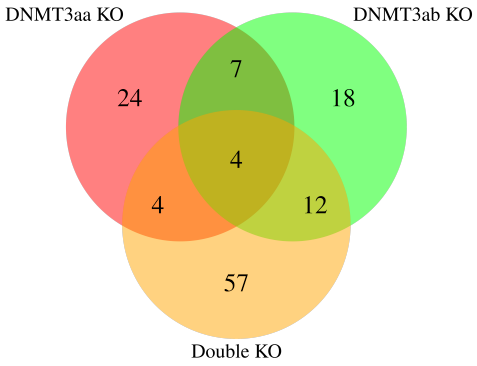

Figure 4.4: *Overlap of the the genes corresponding to differentially methylated regions (DMRs) in the DNMT3aa knockout (red), the DNMT3ab knockout (green) and the DNMT3aa/DNMT3ab double knockout.*

4.4 Discussion

If one would knockout all of its six *de novo* DNA methyltransferases the genome of a zebrafish should be almost devoid of DNA methylation. But this state would most likely not be viable. In this project we deactivated two of the six *de novo* methyltransferases, DNMT3aa and DNMT3ab. We performed single knockouts as well as the double knockout of both enzymes. Interestingly, even in the double-knockout the average genome-wide methylation level is still at appr. 78% across five replicates. Therefore, it is likely that the other three *de novo* methyltransferases compensate the knockout to a high degree.

We have not been able to identify a notable phenotype in zebrafish with either single or double knockouts of DNMT3a. Very recently, November 2020, a study has been published which also performs single knockouts of DNMT3aa and DNMT3ab [94]. They performed extensive behavioral analysis and report that "DNMT3aa KO fish possessed abnormal exploratory behaviors and less fear response to the predator" while "dnmt3ab KO fish displayed less aggression, fear response to the predator, and interests to interact with their conspecifics, loosen shoaling formation, and dysregulated color preference index ranking" [94]. Both knockouts have in common that they "showed higher locomotion activity during the night cycle, which is a sign of anxiety" [94]. It would be interesting to compare our genomic DNA methylation result to theirs as well, especially since they detect very large amounts of differentially methylated regions, 15,962 DMRs in the DNMT3aa KO and 9543 DMRs in the DNMT3ab KO.

Knockout	Hypermethylated DMR	Hypomethylated DMR
DNMT3aa KO	18	5
DNMT3ab KO	16	8
DNMT3aa/ab KO	11	36

Table 4.3: *The distribution of hyper- and hypomethylated differentially methylated regions (DMRs) in the different knockout conditions.*

Unfortunately, their description of used methods is rather short "using a standardized computational mapping approach to analyze the methylome" [94] and the data is not yet stored publicly.

Differential methylation

If DNA methyltransferase are inactivated one would expect the level of DNA methylation to decrease. On a genome-wide scale this is only slightly the case in the double knockout but not in the single knockouts. In the double knockout the amount of dected DMRs is equal to the amount of the single knockouts combined but as one can see in Figure 4.3 they are mostly located on different genomic loci. Therefore the double knockout does not combine the alterations of the single knockouts but leads to differential DNA methylation mainly independent from the single knockouts.

It is most interesting to notice how the ratio of hypomethyalted DMRs differs between the single and the double knockouts, see Table 4.3. While in DNMT3aa and DNMT3ab knockouts 22% of 23 DMRs and 33% of 24 DMRs, respectively are hypomethylated this number is at 77% in 47 DMRs of the double knockout. The Go terms associated to the genes corresponding to hypermethylated DMRs are quite different between the individual experiments. In hypomethylated DMRs "positive regulation of fat cell differentiation", for example is enriched in all three knockout conditions. The mechanism how knockout of a DNA methyltransferase can cause hypermethylation is not directly obvious. It could be an indirect relation where a loss of DNA methylation causes a gain of DNA methylation at another loci as a secondary effect, e.g. by changing the chromatin confirmation and therefore making a region more accesible for, one of the remaining, DNA methyltransferase. Concerning, the expected effect of a loss of DNA methylation after a DNMT3 knockout one can conclude that single knockouts of DNMT3aa or DNMT3ab only have a minor effect. But if both enzymes are inactivated the amount of hypermethylation is notably increased. Therefore, it is a likely scenario

that DNMT3aa and DNMT3ab can compensate for each other and have only a limited amount of specific activity.

There are four genes which corresponded to hypomethylated DMRs in all thre knock-out conditions. The genes are bnip3, dpysl4, crebl2 and gpr19. Apparently, DNMT3b enzymes fail to rescue the DNA methylation in the respective DMRs. Therefore it is possible that these regions are specifically methylated by DNMT3a enzymes.

Bnip3 which is associated with "positive regulation of apoptotic process" can induce cell death by "opening the mitochondrial permeability transition pore" [95]. A failure to do so can result in resistance of cells to cell death which is a key characteristic of cancer. Consequently, bnip3 has been found to be invoved in several cancer, e.g small cell lung cancer [95] or pancreatic cancer [96]. It would be interessting if a a gene whose misregulation can have such drastic consequences is specifically methylated by DNMT3a.

Dpysl4 is an "p53-inducible regulator of energy metabolism in both cancer cells and normal cells, such as adipocytes" [97]. p53 is the most prominent tumor suppressor gene in human. Low expression of dpysl4 is "significantly associated with poor survival of breast and ovarian cancers" [97].

Crebl2 is associated with "positive regulation of fat cell differentiation". It is a transcription factor and acts as a metabolic regulator [98]. Its knockdown in mouse embryonic fibroblasts "leads to elevated glucose uptake, elevated glycolysis as observed by lactate secretion, and elevated triglyceride biosynthesis." [98]

Gpr19 is a g-protein coupled receptor which is also invoved in cancer development. "GPR19 plays a potential role in metastasis by promoting the mesenchymal-epithelial transition (MET) through the ERK/MAPK pathway, thus facilitating colonization of metastatic breast tumor cells" [99].

All of the four genes are known to be involved in cancer-related pathways. Huma DNMT3a is already in the focus of cancer research. Especially in acute myeloid leukemia (AML) mutations of DNMT3a are frequently found [100]. According to `mycancergenome.org`, a cancer medicine knowledge resource, DNMT3a is altered in 3.61% of all cancers, most frequently in lung adenocarcinoma and acute myeloid leukemia. Such a high impact on cancer development is not known for DNMT3b.

Subfunctionalization

In the previous chapter we showcased the evolution of DNA methyltransferase enzymes during vertebrate genome duplications. Starting from a single DNMT3 enzyme two copies, DNMT3a and DNMT3b, were retained after the 1R/2R whole genome duplication. In zebrafish the 3R whole genome duplication and a local gene duplication resulted in a total of six enzymes, two DNMT3a and four DNMT3b genes. Therefore, a sub- or even an neofunctionalization could have happened first between DNMT3a and DNMT3b and subsequently in the additional orthologs as well.

Currently, most results on the different function of DNMT3a and DNMT3b are from mouse experiments. Both enzymes show stage and cell-specific expression but it has been shown that if one of the enzymes is deactivate the other can partly compensate for it [101]. Nevertheless, mice with a homozygous knockout of DNMT3a die four weeks after birth. Homozygous knockouts of DNMT3b seem to haven an even more severe effect and result in death before birth [102].

The expression of DNMT3 genes in zebrafish in different life stages and cell types has been analyzed before [103, 80]. Both studies agree that the two DNMT3a enzymes are more ubiquitously expressed while the three DNMT3b enzymes are more cell-type specific. In addition DNMT3a is still expressed after embryonic development, with the highest levels in brain.

Our investigation was the first which analyzed a homozygous double knock-out of two DNMT3 genes. Given our results it appears that there is almost no subfunctionalization between DNMT3aa and DNMT3ab in regard to establishing DNA methylation since the amount of hypomethylation is very minor in the single knockouts. Interestingly, there seems to be a higher subfunctionalization for preventing DNA methylation since the difference between the single knockouts is mainly in hypermethylated DMRs. However, it is not clear how the mechanism for a DNMT3 preventing DNA methylation exactly works.

Both zebrafish DNMT3a genes taken together alredy show a stronger subfunctionalization compared to the DNMT3b genes. This is shown by the fact that for 36 hypomethylated DMRs the remaining DNMT3b genes were not able to compensate the inactivation of the two DNMT3a genes.

Evolutionary evidence

Our results from the previous chapter further support our hypothesis. In four teleost species either DNMT3aa or DNMT3ab could not have been detected. In *Oryzias latipes* no DNMT3ab was identified and in the Salmoniformes *Coregonus sp.*, *Thymallus thymallus* and *Hucho hucho* no DNMT3aa was detected, see Table 3.2. In none of the analzed species both DNMT3aa and DNMT3ab are missing. Therefore, it seems that, while in several cases duplicated copies of DNMT3a or DNMT3b are lost in teleosts there is negative selection against loosing all DNMT3a genes.

Studying DNMT3b

While there is a notable effect of the DNMT3aa/ab double knockout it is rather moderate with only 47 differentially methylated region. It is likely that DNMT3b genes compensate for most of the knockout effects. As a next step for studying subfunctionalization of DNMT3 genes in zebrafish it would be quite interesting to study the function of the individual DNMT3b genes. In cooperation with the Aluru lab we actually started to perform experiments using the Crispr/Cas9 system to generate single knockouts for all DNMT3b genes. Unfortunately, the experiments were not succesfull in the given amount of time. It is possible that generating individual knockouts is quite challenging since they are four very similar proteins. Three of which are even located closely to each other on the same genomic loci. This makes it challenging to generate single knockouts without off-target effects.

Further experimental data

In the current experimental setup we do not have gene expression data. Therefore we have no information about the effects of DNA methylation changes on gene expression. While we detected a number of DNA methylation changes it is difficult to say how strong their functional effect is. In principle it is possible that while a significant amount of DNA methylation is lost at a regulatory element it is still inaccessible to the binding of transcription factors. To better study the DMRs specific to the individual DNMT3a genes it would be advantageous to generate this data as well in the future. If such experiments were performed again it might be worthwile to consider including a method able to capture chromosome conformation, like ATAC-seq [104]. DNMT3

proteins have been reported to interact with the histone H3 tail through their PWWP [105] and ADD [32] domain. Therefore, information about the accesibility of genomic regions with or without changes in DNA methylation might give additional insight into the mechanistic relationships.

Conclusion

We have performed the first analysis of the effects of DNMT3aa and DNMT3ab double knockouts on genome-wide DNA methylation in zebrafish. Given our results we hypothesize that DNMT3aa and DNMT3ab can compensate for each other to a high degree. DNMT3a genes have likely been subfuntionalized but their loss can be compensated by DNMT3b genes. This compensation by DNMT3b genes works well enough that no notable phenotype can be observed in double knockout zebrafish but a difference is notable on the epigenome level.

The genes which are hypomethylated in all three knockout conditions are known to be related to cancer development. Zebrafish is already used as a model organism for the research on the effects of DNA methylation on cancer [106] but according to that publication DNMTs are currently not used in a zebrafish cancer model. Our results indicate that the involvement of DNMT3a in cancer could be conserved in zebrafish and therefore opening the possibility to develop additional epigenetic disease models. The more zebrafish is used to study epigenetic mechanisms the more important it becomes to study its DNA methylation machinery as detailed as possible.

Chapter 5

Role of DNA methylation in altered testis gene expression patterns in adult zebrafish exposed to Pentachlorobiphenyl

5.1 Introduction

In the last three chapters we have focused on the evolution and function of DNA methyltransferases in different metazoan lineages. In this chapter we are investigating an actual example of the effects DNA methylation may have on gene expression after the exposure of zebrafish to a chemical.

Most heritable information of metazoan organisms is stored in the DNA. It is a very stable way to save information and does not change over the lifetime of an individual. What can change is the way how this information is processed through gene regulation. DNA methylation is known to be an important gene regulatory mechanism. It is also believed that environmental conditions may have an effect on DNA methylation and therefore, indirectly, on gene regulation [107]. This might allow an organism to be more flexible in changing environmental conditions by beeing able to response with changes in gene regulation. On the other hand, this information flow from environmental conditions to gene regulatory processes harbors the risk of detrimental alterations. This can be caused, for example, by environmental pollutants. Zebrafish is a popular vertebrate model organism to study development and model human diseases but it has also emerged as a model to study the effects toxicants might have on an organism [107].

Most of the studies performed so far, have been focusing on DNA methylation changes in specific parts of the genome, e.g. single genes or regulatory elements. With the advent of hight-throughput sequencing technologies in the last years it became

possible to study these effects on a genome-wide scale, as well.

PCB126 (3, 3',4, 4', 5-pentachlorobiphenyl) is a polychlorinated biphenyl (PCB) which is ubiquitously distributed in the environment. PCBs have been widely used in electrical equipment and industrial processes until the 1980's. They have been one of twelve pollutants, the so called "dirty dozen", whose production was banned globally by the "Stockholm Convention on Persistent Organic Pollutants" in 2001. Unfortunately, they are still widely distributed in the environment and therefore continue to impact public health.

Dioxin-like PCBs such as PCB126 have been studied intensely in the last decades and their mode of activation is well understood. It involves the activation of the transcription factor aryl hydrocarbon receptor (AHR) [108]. The target genes of AHR have been extensively studied as well [109] but if DNA methylation plays a role in the activation is less well understood. Also an association between altered DNA methylation and an exposure to PCB in humans has been demonstrated [110, 111, 112, 113]. The genome-wide changes of DNA methylation and gene expression in brain and liver after PCB exposure in zebrafish has been investigated recently [114]. In this work we wanted to study the effects PCB exposure on zebrafish testis. We therefore investigated DNA methylation and gene expression changes on a genome-wide level and correlated them to each other.

5.2 Methods

Generating the DNA methylation data

Male zebrafish were exposed to either 0.3 nM PCB126, 10 nM PCB126 or solvent carrier (0.01% DMSO) for 24 hours. Each treatment had six biological replicates. After the treatment the fish were kept in normal conditions for seven days before they were euthanized and testis tissue was dissected. From these samples total RNA and genomic DNA was extracted for sequencing.

Enhanced reduced representation bisulfite sequencing (eRRBS) library preparation was performed on the genomic DNA. Subsequently, a 50 bp paired-end sequencing on an Illumina HiSeq2500 platform was performed. For the total RNA 50 bp single-end sequencing on the Illumina HiSeq2000 platform was performed.

Processing the DNA methylation data The quality of the data was checked with `FastQC` [87]. It visualizes several quality parameter of high-throughput data, e.g. length and quality score distribution of the reads. Subsequently, the reads were trimmed using the "–rrbs" mode of `Trim galore` [88]. It removes adapter sequences which are artificially introduced during the sequencing procedure and occassionaly become part of the resulting reads.

We used the `Bisulfite Analysis Toolkit (BAT)` [89] to further process the eRRBS-seq data. It is a data analysis pipeline which mainly combines `segemehl` [90, 91] for mapping bisulfite-treated sequence data to the genome and `metilene` [92] to detect differentially methylated regions (DMRs).

The mapping was done using the bisulfite methyl-C seq mode of `segemehl` included in `BAT`. As a reference genome the zebrafish (*Danio rerio*) genome GRCz10/danRer10 was used. The methylated cytosines were called with `BAT_calling`. It calculates the methylation rate for each cytosine by calculating the relative amount of unmodified cytosines: $\#C/(\#C+\#T)$. These cytosines were filtered with `BAT_filter_vcf` to keep only the ones occuring in a CpG context and with a minimum coverage of 10 and a maximum of 100 reads. `BAT_summarize` prepared the input data for `metilene`. In this step the individual experiments are assigned to two different groups, between which the differential methylation will be called. In our case the two groups were 0.01% DMSO as the background and one of the PCB126 exposures as the experiment. Finally, the script `BAT_DMRcalling` was used. It executes `metilene` to call differentially methylated regions (DMRs). Standard settings were used which require the DMRs to contain at least 10 CpGs and a minimum difference of 0.1 between the mean methylation rates per group. DMRs with a q-value below 0.05 were considered to be significant.

Corresponding genes and GO annotation (GREAT) To associate DMRs with a gene we used the `Genomic Regions Enrichment of Annotations Tool (GREAT)` version 2.0.2 [93]. It predicts which gene is influenced by a cis regulatory element, in our case a DMR. This is done as a two step process. In the first step a regulatory domain of 5 kilobases (kb) upstream and 1 kb downstream of its TSS is assigned to each gene. In the second step each genomic region, e.g. DMR, is associated with all genes whose regulatory domains it overlaps. There are different methods to define the regulatory domain. We used the default method "basal plus extension" with the default values of 5 kb/1 kb for the basal regulatory domain and 1000 kb for the extension. It assigns

a “basal regulatory domain” and ignores the presence of other genes. This regulatory domain is subsequently extended upstream and downstream by up to 1000 kb in each direction or up to the “basal regulatory domain” of the neighboring genes. Therfore, regulatory domains can overlap but only if the TSS of two neighboring genes are closer than 5 kb upstream or 1 kb downstream of each other. Since `GREAT` only supports the *Danio rerio genome* version Zv9/danRer7 we used the UCSC genome browser `liftOver` utility (`https://genome.ucsc.edu/cgi-bin/hgLiftOver`) to convert the coordinates of the DMRs from GRCz10/danRer10 to Zv9/danRer7. Once we had corresponding genes assigned to the DMRs we used the gene ontology (GO) anntation (biological processes, molecular function and KEGG pathway) of these genes to perform an enrichment analysis.

Processing the RNA sequencing data RNA was extracted from the same specimen which were used for the DNA extraction. The RNA sequencing data was processed as described in [115]. In short, the reads were quality checked using `FastQC` [87] and trimmed with `Trimmomatic` [116]. The mapping to the *Danio rerio* genome GRCz10/danRer10 was done using `STAR` [117]. With the mapped reads the FPKM (fragments per kilobase of transcript per million mapped read) for each gene was calculated. The Ensembl [39] gene annotation was used.

To detect differentially expressed genes we used the `DESeq2` package [118] with standard parameters. `DESeq2` uses negative binomial generalized linear models to test for differential expression of genes. Differentially expressed genes with an adjusted p-value smaller than 0.05 were considered significant.

For significantly differentially expressed genes with a log2 fold change of at least 2 a gene ontology (GO) enrichment analysis was performed. This was done using the DAVID Bioinformatics Resources [119] (`https://david.ncifcrf.gov/home.jsp`). An enrichment for biological processes, molecular function and KEGG pathways was analysed.

5.3 Results

DNA methylation profiling The eRRBS-seq data of the genomic DNA was processed as described above. Each of the three treatments had six biological replicates. Unfortunately, for one replicate of the DMSO and one of the 10 nM PCB126 treat-

ment the sequencing process was not succesfull, leaving only five replicates for these treatments. On average each sequencing produced 11.9 million pared-end reads. The mapping rate with appr. 95% on average was quite good, see Table 5.1 for more details. By counting the fraction of reads indicating DNA methylation for a certain position among all reads covering this position we can calculate the DNA methylation rate for this position. Of all cytosines in a CpG context 1,816,810 (93%) had a methylation level > 0 (at least one read indicating DNA methylation on this position). The average genome-wide methylation level was 84%, see Table 5.2 for more details. The global CpG methylation level was slightly higher for the DMSO treatment (84.8%) and slightly lower for the 0.3 nM (83.17%) and 10 nM (83.2%) PCB126 treatment, see Figure 5.1 for a visualization.

Condition	Sample ID	Number of reads	Number of mapped reads
DMSO	D1	8,168,404	7,632,045 (93.43%)
DMSO	D2	9,957,843	9,412,813 (94.53%)
DMSO	D3	10,786,595	10,233,729 (94.87%)
DMSO	D4	11,799,398	11,191,119 (94.84%)
DMSO	D5	14,501,290	13,578,935 (93.64%)
PCB 0.3nM	P0.3-9	13,635,225	12,757,987 (93.57%)
PCB 0.3nM	P0.3-11	13,674,864	12,977,478 (94.90%)
PCB 0.3nM	P0.3-12	11,830,503	11,269,978 (95.26%)
PCB 0.3nM	P0.3-13	13,968,933	13,279,248 (95.06%)
PCB 0.3nM	P0.3-14	12,536,363	11,890,055 (94.84%)
PCB 0.3nM	P0.3-16	12,210,513	11,558,361 (94.66%)
PCB 10nM	P10-17	11,824,432	11,224,550 (94.93%)
PCB 10nM	P10-18	9,949,807	9,424,059 (94.71%)
PCB 10nM	P10-21	10,983,355	10,403,421 (94.72%)
PCB 10nM	P10-23	12,880,370	12,242,973 (95.05%)
PCB 10nM	P10-24	11,786,444	11,224,951 (95.24%)
Average		11,905,896	11,268,856 (94.65%)

Table 5.1: *Number of sequenced and mapped paired-end reads for all eRRBS-seq libraries.*

PCB126-induced changes in DNA methylation in testis PCB 126 0.3 nM treatment

In total 308 differentially methylated regions (DMRs) were predicted of which 37 had an adjusted p-value (or q-value) < 0.05, see Figure 5.2 for an overview. Among the 37 DMRs there were 10 hypermethylated and 27 hypomethylated regions. None of them

Sample ID	#unmethyl. CpGs	#methyl. mCpGs	Global CpG methyl.level
D1	74,718 (5%)	1,430,201 (95%)	86%
D2	103,721 (6%)	1,646,267 (94%)	85%
D3	122,507 (7%)	1,733,399 (93%)	85%
D4	144,119 (7%)	1,816,488 (93%)	84%
D5	168,425 (7%)	2,096,685 (93%)	84%
P0.3-9	148,244 (7%)	1,979,684 (93%)	84%
P0.3-11	152,740 (7%)	1,983,935 (93%)	84%
P0.3-12	112,014 (6%)	1,769,691 (94%)	81%
P0.3-13	153,408 (7%)	1,999,865 (93%)	84%
P0.3-14	144,600 (7%)	1,923,376 (93%)	84%
P0.3-16	165,813 (8%)	1,797,084 (92%)	82%
P10-17	145,967 (8%)	1,757,976 (92%)	82%
P10-18	117,585 (7%)	1,681,977 (93%)	84%
P10-21	123,288 (6%)	1,805,228 (94%)	84%
P10-23	149,059 (7%)	1,882,976 (93%)	83%
P10-24	131,856 (7%)	1,764,152 (93%)	83%
Average	134,879 (7%)	1,816,810 (93%)	84%

Table 5.2: *The table shows the amount of cytosines in a CpG context in each sequencing experiment. The second column shows unmethylated (methylation leven = 0) and the third column methylated (methylation level > 0) CpG's. The 4th colum shows the the average methylation level of all CpG's.*

showed a percent methylation difference of larger than 40%. Three hypomethylated DMRs had a methylation difference larger than 30%. The highest concentration of DMRs was on chromosome 4 with 9 DMRs, 24% of the total amount. 7 of these 9 DMRs were hypomethylated.

Hypermethylated DMRs were significantly enriched for 46 Gene ontology (Go) biological process terms, e.g. pigment granule dispersal and pigment granule aggregation in cell center. 17 Go molecular function terms were enriched, among them melatonin receptor activity and inward rectifier potassium channel activity.

Hypomethylated DMRs are enriched in 29 process terms. The most significant ones were RNA polyadenylation and RNA 3'-end processing. 42 Go molecular function terms were enriched, e.g. polynucleotide adenylyltransferase activity and adenylyltransferase activity, see Table 5.3 for the top five Go terms and supplement section 6.3 for the top twenty terms.

PCB 126 10 nM treatment 460 differentially methylated regions (DMRs) were

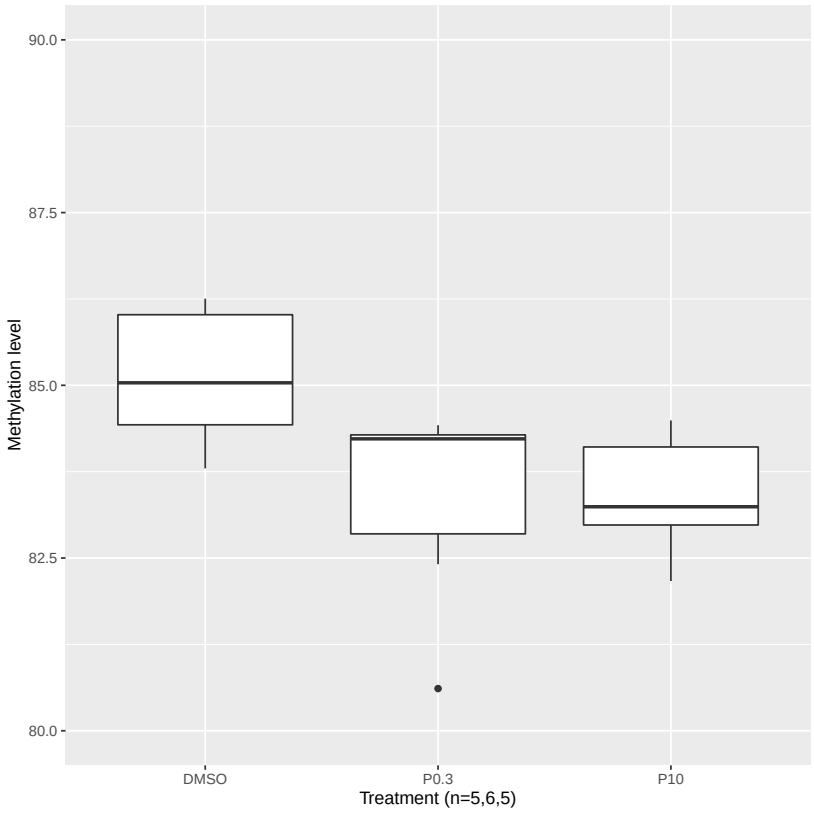

Figure 5.1: *Boxplots of the global CpG methylation level of each experiment (see Tab. 5.2 4th column). Each boxplot represents a different condition, from left to right: DMSO, 0.3nM PCB126 and 10nM PCB 126.*

predicted by `metilene` of which 92 had an adjusted p-value (or q-value) < 0.05, see Figure 5.3 for an overview. Of the 92 DMRs 80 were hypomethylated and 12 hypermethylated. Two hypomethylated and 10 hypermethylated DMRs showed a percent methylation difference larger than 40%. The highest concentration of DMRs is on chromosome 4 with 34 DMRs, 37% of the total amount. 31 of these 34 DMRs are hypomethylated.

Hypermethylated DMRs in the testis showed significant enrichment of a number of Gene ontology (Go) biological process terms. There were 79 in total, among them monovalent inorganic cation transport and pigment granule dispersal. For molecular functions there were 24 terms, among them monovalent inorganic cation transmembrane transporter activity and inorganic cation transmembrane transporter activity, see Table 5.4.

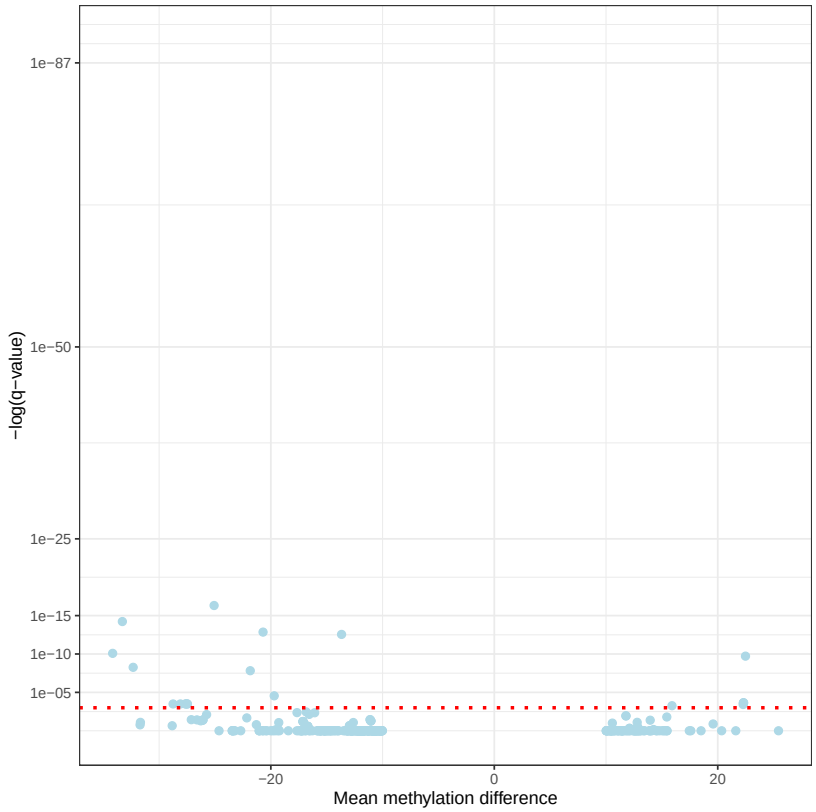

Figure 5.2: *0.3 nM PCB126-induced changes of DNA methylation. The volcano plot shows on the x-axis the methylation difference and on the y-axis the q-value. Each point is a DMR predicted by metilene. The red dashed line indicates a q-value of 0.05, DMRs below this line were considered significant. DMRs with a methylation difference smaller than 0 are hypomethylated and otherwise hypermethylated.*

Hypomethylated DMRs were enriched in 34 Go biological process terms, among them RNA polyadenylation and iron-sulfur cluster assembly. Molecular functions were enriched in 36 terms, e.g. ion binding and nucleic acid binding, see Table 5.4 for the top five Go terms and supplement section 6.3 for the top twenty terms.

Transcriptional changes We obtained an average of 26.1 million reads mapping to ENSEMBL genes in the DMSO-treated control sample. The libraries of individuals treated with PCB126 0.3 nM or 10 nM resulted in 26.6 and 23.1 million reads mapping to ENSEMBL genes.

The gene Cyp1a is a known target of AHR after exposure to PCB126. Therefore it should be significantly upregulated in the replicates with a PCB treatment. The gene

Biological Process - Hypermethylated DMRs	
Go term	Adj. p-value
pigment granule dispersal (GO:0051876)	< 0.0001
pigment granule aggregation in cell center (GO:0051877)	< 0.0001
establishment of pigment granule localization (GO:0051905)	0.0002
pigment granule localization (GO:0051875)	0.0003
cellular pigmentation (GO:0033059)	0.0003
Hypomethylated DMRs	
RNA polyadenylation (GO:0043631)	0.0002
RNA 3'-end processing (GO:0031123)	0.0004
dADP catabolic process (GO:0046057)	0.0014
dGDP catabolic process (GO:0046067)	0.0014
GDP catabolic process (GO:0046712)	0.0014
Molecular function - Hypermethylated DMRs	
Go term	p-value
melatonin receptor activity (GO:0008502)	0.0001
inward rectifier potassium channel activity (GO:0005242)	0.0003
voltage-gated potassium channel activity (GO:0005249)	0.0056
potassium channel activity (GO:0005267)	0.0087
potassium ion transmembrane transporter activity (GO:0015079)	0.0088
Hypomethylated DMRs	
polynucleotide adenylyltransferase activity (GO:0004652)	0.0001
adenylyltransferase activity (GO:0070566)	0.0004
8-oxo-dGDP phosphatase activity (GO:0044715)	0.0011
8-oxo-GDP phosphatase activity (GO:0044716)	0.0011
8-hydroxy-dADP phosphatase activity (GO:0044717)	0.0011

Table 5.3: *GO terms of the genes corresponding to differentially methylated regions (DMRs) in the PCB126 0.3 nM treatment. The five termes with the lowest p-value are shown.*

Cyp1a is upregulated in the PCB126 0.3 nM and 10 nM treatment by a fold change of appr. 190 and 480 respectively. In both cases the adjusted p-value is below 0.05.

On a genome-wide scale there were a total of 767 and 4,708 differentially expressed genes (DEGs) in the 0.3 nM and 10 nM PCB126 treatment with an adjusted p-value of 0.5 or smaller.

Among the 767 DEGs in the 0.3 nM treatment, 458 were upregulated and 309 were downregulated. Among the upregulated genes 214 (46.7%) and the downregulated genes 144 (46.6%), had a fold change of more than 2.

The upregulated genes were enriched in GO terms such as response to external

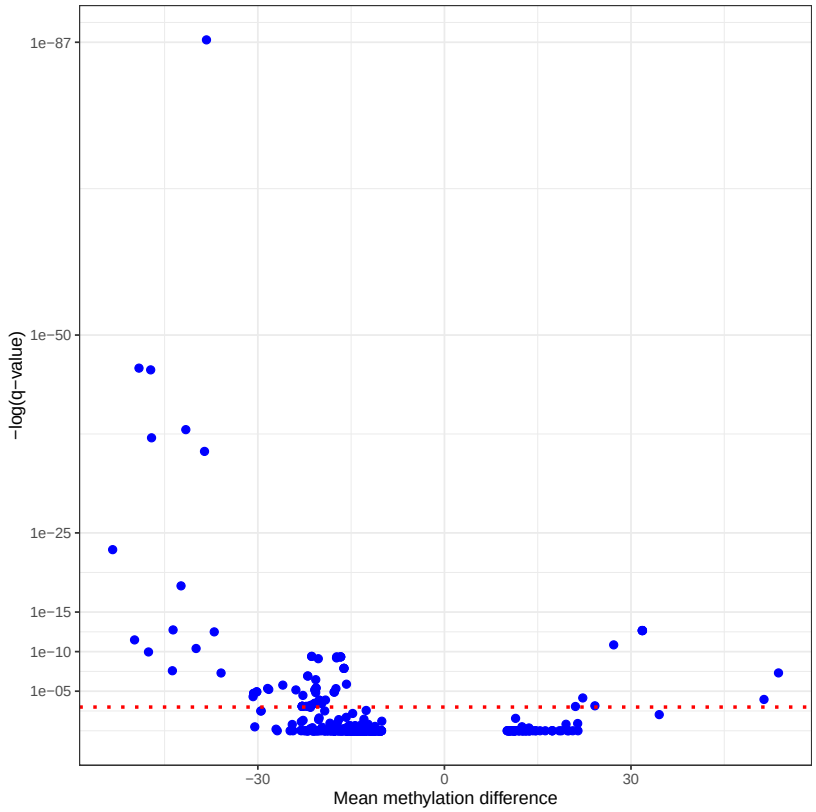

Figure 5.3: *10 nM PCB126-induced changes of DNA methylation. The volcano plot shows on the x-axis the methylation difference and on the y-axis the q-value. Each point is a DMR predicted by metilene. The red dashed line indicates a q-value of 0.05, DMRs below this line were considered significant. DMRs with a methylation difference smaller than 0 are hypomethylated and otherwise hypermethylated.*

stimulus and response to chemical (both biological process). The downregulated genes were only enriched in two KEGG pathways ECM-receptor interaction and TGF-beta signaling pathway, see Table 5.5 for details.

The PCB126 10 nM exposure resulted in the differential expression of 4,708 genes. Among these 2,822 genes were upregulated and 1,886 genes were downregulated. Among the upregulated genes 1,534 (54.4%) and in the downregulated genes 324 (17.2%) had a fold change of more than 2.

The upregulated genes of the PCB126 10 nM treatment are enriched in similar GO terms compared to the lighter PCB126 treatment, For example immune response and response to external stimulus. But opposite to the 0.3 nM treatment they are also enriched in molecular function GO terms, e.g. oxidoreductase activity and cytochrome-

Biological Process - Hypermethylated DMRs	
Go term	p-value
monovalent inorganic cation transport (GO:0015672)	< 0.0001
pigment granule dispersal (GO:0051876)	< 0.0001
pigment granule aggregation in cell center (GO:0051877)	< 0.0001
cation transport (GO:0006812)	0.0002
ATP biosynthetic process (GO:0006754)	0.0003
Hypomethylated DMRs	
RNA polyadenylation (GO:0043631)	0.0015
iron-sulfur cluster assembly (GO:0016226)	0.0023
calcium-independent cell-cell adhesion (GO:0016338)	0.0030
RNA 3'-end processing (GO:0031123)	0.0032
gluconeogenesis (GO:0006094)	0.0048
Molecular function - Hypermethylated DMRs	
Go term	p-value
monovalent inorganic cation transmembrane transporter activity (GO:0015077)	< 0.0001
inorganic cation transmembrane transporter activity (GO:0022890)	0.0001
melatonin receptor activity (GO:0008502)	0.0002
cation transmembrane transporter activity (GO:0008324)	0.0003
inward rectifier potassium channel activity (GO:0005242)	0.0005
Hypomethylated DMRs	
ion binding (GO:0043167)	< 0.0001
nucleic acid binding (GO:0003676)	< 0.0001
metal ion binding (GO:0046872)	< 0.0001
cation binding (GO:0043169)	< 0.0001
organic cyclic compound binding (GO:0097159)	< 0.0001

Table 5.4: *GO terms of the genes corresponding to differentially methylated regions (DMRs) in the PCB126 10 nM treatment. The five termes with the lowest p-value are shown.*

c oxidase activity but not in KEGG pathways. Downregulated genes of the 10 nM treatment were only enriched in three KEGG pathways Oxidative phosphorylation, Cytokine-cytokine receptor interaction and Jak-STAT signaling pathway, see Table 5.6 for details.

Relationship between methylation and transcriptional changes In the PCB126 0.3 nM treatment one hypermethylated differentially methylated region (DMR) corresponds to an upregulated differentially expressed gene (DEG) while three hypomethylated DMRs correspond to two up and one down-regulated DEG. Using

Biological Process - Upregulated	
Term	p-value
response to external stimulus (GO:0009605)	0.0002
taxis (GO:0042330)	0.0028
response to chemical (GO:0042221)	0.0024
response to stress (GO:0006950)	0.0155
immune response (GO:0006955)	0.0367
KEGG - Downregulated	
Term	p-value
ECM-receptor interaction (dre04512)	0.0085
TGF-beta signaling pathway (dre04350)	0.0050

Table 5.5: *GO terms of the PCB126 0.3 nM treatment. GOTERM_BP_2, GOTERM_MF_2 and KEGG pathway. Only the five best significant ones (adj. p-value <0.05) are shown. MF and KEGG for upregulated as well as BP and MF for downregulated genes had no significant enrichments.*

Biological Process - Upregulated	
Term	p-value
immune response (GO:0006955)	< 0.0001
response to external stimulus (GO:0009605)	< 0.0001
leukocyte migration (GO:0050900)	< 0.0001
taxis (GO:0042330)	< 0.0001
response to chemical (GO:0042221)	< 0.0001
Molecular Function - Upregulated	
Term	p-value
oxidoreductase activity (GO:0016491)	< 0.0001
cytochrome-c oxidase activity (GO:0004129)	0.0015
protein binding (GO:0005515)	0.0052
carbohydrate binding (GO:0030246)	0.0469
enzyme regulator activity (GO:0030234)	0.0488
KEGG - Downregulated	
Term	p-value
Oxidative phosphorylation (dre00190)	< 0.0001
Cytokine-cytokine receptor interaction (dre04060)	< 0.0001
Jak-STAT signaling pathway (dre04630)	0.0008

Table 5.6: *GO terms of the PCB126 10 nM treatment. GOTERM_BP_2, GOTERM_MF_2 and KEGG pathway. Only the five best significant (adj. p-value <0.5) ones are shown. Downregulated genes had no significant enrichments.*

`BAT_correlate` we performed a Sperman's rank correlation test between the DNA methylation and the gene expression change and calculated the adjusted p-value. In the 0.3 nM treatment we detected 58 correlations in total, four of which had an adjusted p-value smaller than 0.05. None of the significantly correlated genes had a fold change of 2 or more. DEGs with a corresponding DMR are shown in Table 5.7.

DMR ID	Gene ID	Mean methyl. difference	log2 fold expr. change	Correlation adj. p-value
DMR_3	ENSDARG00000028661	0.14	0.78	**0.0279**
DMR_14	ENSDARG00000103318	-0.34	0.39	0.0694
DMR_15	ENSDARG00000103318	-0.22	0.39	0.115
DMR_2	ENSDARG00000052037	-0.20	-3.56	0.1323

Table 5.7: *All significantly differentially expressed genes (DEG) with a corresponding differentially methylated region (DMR) of the 0.3 nM PCB126 treatment. Mean methylation difference (3rd colum) corresponds to the DMR, while log2 fold expression change (4th column) corresponds to the DEG. The 5th colum shows the p-value of Sperman's rank correlation test betweeen the DMR and the DEG. A fold change* > 2 *and a adj. p-value* < 0.05 *are highlighted in bold.*

PCB126 10 nM treatment

Two hypermethylated DMRs have a corresponding differentially expressed gene. In both cases the gene is downregulated. 15 hypomethylated DMRs correspond to a DEG. 12 of these genes are upregulated while 3 are downregulated. The fold change ranges from 1.3 up to 8.6, for details see Table 5.8. In the 10 nM treatment the total number of correlations was 138. 29 of these correlations were considered to be significant ($p - value < 0.05$). Five significantly correlated genes had a fold change of 2 or more. DEGs with a corresponding DMR are shown in Table 5.8. One example of a strong correlation between the DEG ENSDARG00000089382 (zgc:158463) and DMR_69 is shown in Figure 5.4, unfortunately the function of the gene is unknown..

5.4 Discussion

The effects of PCB126 exposure on DNA methylation and gene expression in zebrafish tissues has been previously studied in liver and brain [114]. We successfully investigated these effects in a new tissue, testis. Opposite to the previous study we have analyzed the effect of PCB126 in two different concentrations, 0.3 nM and 10nM for 24 h. This

DMR ID	Gene ID	Mean methyl. difference	log2 fold expr. change	Correlation adj. p-value
DMR_2	ENSDARG00000030289	0.54	-0.45	0.171
DMR_12	ENSDARG00000005482	0.22	-0.47	0.2162
DMR_6	ENSDARG00000005185	-0.24	**3.10**	0.2162
DMR_10	ENSDARG00000015472	-0.23	**1.16**	**0.0154**
DMR_35	ENSDARG00000069311	-0.20	**1.51**	**0.0022**
DMR_34	ENSDARG00000070845	-0.18	**1.89**	**0.0022**
DMR_70*	ENSDARG00000070845	-0.18	**1.89**	**0.022**
DMR_69	ENSDARG00000089382	-0.20	**1.21**	**0.0072**
DMR_1	ENSDARG00000069996	-0.40	0.90	**0.0107**
DMR_3	ENSDARG00000052361	-0.13	0.87	0.7033
DMR_11	ENSDARG00000102824	-0.21	0.75	0.7033
DMR_73	ENSDARG00000036567	-0.21	0.61	0.752
DMR_17	ENSDARG00000103318	-0.16	0.41	0.0831
DMR_18	ENSDARG00000103318	-0.47	0.41	**0.0011**
DMR_22	ENSDARG00000020730	-0.20	-0.55	0.0218
DMR_22	ENSDARG00000044718	-0.20	-0.58	**0.0046**
DMR_30	ENSDARG00000013312	-0.20	-0.60	**0.0004**

Table 5.8: *All significantly differentially expressed genes (DEG) with a corresponding differentially methylated region (DMR) of the 10 nM PCB126 treatment. Mean methylation difference (3rd colum) corresponds to the DMR, while log2 fold expression change (4th column) corresponds to the DEG. The 5th colum shows the p-value of Sperman's rank correlation test betweeen the DMR and the DEG. A fold change* > 2 *and a adj. p-value* < 0.05 *are highlighted in bold.*
(- DMR_70's location on chromsome 5 in danRer10 is converted to the exact location of DMR_34 in danRer7 and therefore they correspond to the same gene.)*

Condition	DMRs		DEGs		DMR w. DEG	Correlations
	Hyper	Hypo	Up	Down		
0.3 nM PCB126	10	27	458	309	4	1
10 nM PCB126	12	80	2,822	1,886	17	9

Table 5.9: *Summary of the detected differentially methylated regions (DMRs) and differentially expressed genes (DEGs). The 4th column shows the amount of DMRs which corresponding to a DEG according to the Genomic Regions Enrichment of Annotations Tool (GREAT) [93]. The 5th column shows the amount of DMRs and DEGs which have a significant correlation of their methylation and gene expression change according to a Sperman's rank correlation test.*

does not only allow us to see the effect PCB126 has on genomic DNA methylation and gene expression but also how these effects differ depending on the concentration.

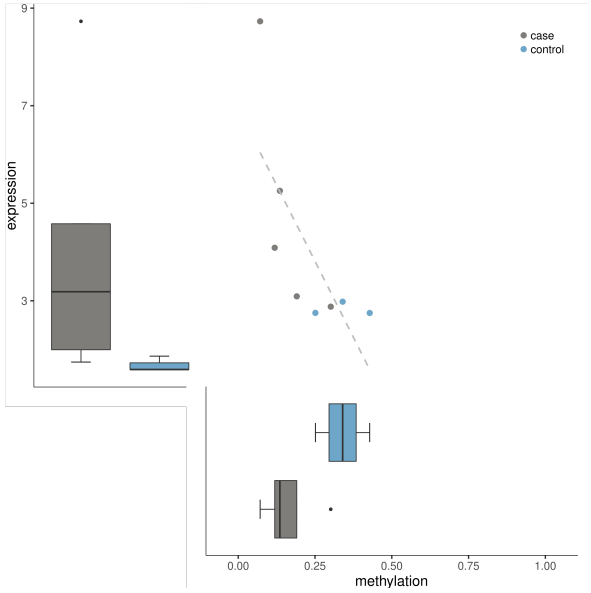

Figure 5.4: *Correlation plot between the differentially methylated region DMR_69 and the differentially expressed gene ENSDARG00000089382. The adj. p-value of the Sperman's rank correlation test was 0.0072, the fold change 2.3 and the methylation difference -20%.*

Differential DNA methylation

Genomic DNA methylation is notably reduced after exposure to PCB126. An exposure to a higher concentration of PCB126 leads to a higher loss of DNA methylation. This was also shown by the number of differentially methylated regions (DMRs), 37 in the lower concentration vs 92 in the higher concentration, 73% and 87% of them are hypomethylated.

Genes corresponding to the detected DMRs were enrichment in a couple of GO terms. However, it is difficult to identify a common function for the genes corresponding to the DMRs since mostly only a small number of genes was associated with each term. It is possible that the method used for assigning genes to DMRs does not work perfectly. We assume that the DMR regulates a corresponding gene. If the DMR does not overlap with the promoter region, which is mostly the case, it would most likely act as an enhancer [120]. However, enhancer can be quite distance from the gene they are regulating and they are not necessarily in the direct vicinity of the gene they are regulating. In human it has been shown that DNA loops can span several hundred thousand kilobases [121]. Therefore, it is difficult to correctly predict a corresponding

gene computationally if no further information is available.

Differential gene expression

The differential effects on gene expression seem notably higher than the ones on DNA methylation. In the 0.3 nM and 10 nM PCB126 exposure we detected 767 and 4,708 differentially expressed genes (DEG). We see the same correlation as before that the higher PCB126 exposure causes a larger difference in gene expression. The upregulated genes, in both conditions, are enriched for Go terms like "response to chemical" and "response to stress" which is quite expected given the experimental setup. Overall, the amount of enriched Go terms was very moderate. If we compare the results from the 10 nM PCB126 exposure in testis to the results by [114] we see that the number of DEGs in testis is similar to the one in brain. One difference is that the the amount of up- and downregulated genes in liver and brain is almost evenly distributed while in our data 83% of the DEGs (with fold change > 2) are upregulated. In liver and brain many of the upregulated genes have been enriched for Go terms "response to external stimulus" which is a reaction to the chemical treatment. Downregulated genes were enriched in more diverse Go terms in liver and brain but not in testis. Therefore, it seems that in testis the reaction to PCB126 exposure mainly consists of the upregulation of genes responding to the chemical/stress.

Transposable elements

In Neel *et al.* [114] it was found that trdi1 a key player in piRNA biogenesis was upregulated in zebrafish brain tissue after exposure to PCB126. In our data of testis neither tdrd1 nor henmt1, another important protein for piRNA production, was differentially methylated. It was hypothesized that the upregulation was a reaction to an increased re-activation of transposable elements (TE). Since we do not see an upregulation of piRNA pathway genes it is likely that there is no high amount of TE re-activation. Therefore, in this regard the results of testis tissue are more similar to liver tissue analyzed in [114], as well.

DNA methylation machinery Key proteins for regulating genomic DNA methylation are the DNA methyltransferases which were extensively discussed in the previous chapters. On the other hand proteins of the family ten-eleven translocation (Tet) can

remove DNA methylation by oxidizing DNA methylation and therefore causing its removal. Of the DNMTs only DNMT1 is slightly downregulated, 1.4 fold, only in the 10 nM concentration. In the Tet family only Tet2 is upregulated by 1.6 fold Tet1 and Tet3 are not differentially expressed. This indicates that there is no drastic change in the general DNA methylation machinery and therefore changes in DNA methylation are more likely caused by mechanisms like chromatin remodelling.

Removal of DNA methylation

DNA methylation is believed to be a quite stable regulatory mechanism. It is mainly established during embryonic development therefore changes of DNA methylation are most frequently a loss of DNA methylation instead of a gain. It can be removed passively by cell divisions or actively by oxidization of methylcytosine to hydroxymethylcytosine from Tet enzymes.

We performed the DNA methylation seven days after exposure to PCB126. We are not aware of a recent study measuring the turnover time of cells in zebrafish testis. For rat testis it has been reported that Leydig cells and peritubular cells have turnover times of at least 142 and 85 days [122]. If this is similar in zebrafish then it is unlikely that passive removal of DNA methylation through cell divisions had enought time to make a large impact after seven days.

Active demethylation by Tet enzymes has been reported most frequently in embryonic stem cells or brain cells. Most of these studies were performed in mammals but a similar distribution was shown in zebrafish [123]. It is noteworthy that, in this study, the lowest levels of hydroxymethylation have been reported in testis (0.01%) [123].

Taken together this leaves the possibility that testis is not a tissue where large DNA methylation changes can be seen after a relatively short amount of days.

Correlation between methylation and expression

The number of genes were we are able to detect a correlation between differential DNA methylation and gene expression is relatively small. There are four in the 0.3 nM PCB126 treatment and 17 in the 10 nM treatment. As discussed above, one problem might be that we inncorectly associated DMRs and genes with each other due to complex regulatory interactions. However, a low correlation between DNA methylation and gene expression changes has been reported by Neel *et al.* [114] and

several other studies [124, 125, 126], as well. On the bright side, 14 of the 17 correlations we find in the 10 nM treatment show the indirect correlation one would expect between DNA methylation and gene expression. This means in the case of hypermethylated DMRs the gene is upregulated and for hypomethylated DMRs it is downregulated. Such a behaviour is traditionally reported from promoter regions. In the case of other regulatory elements (RE) it would mean that the methylated RE performs a positive regulation on the corresponding gene. An example would be that the RE contains binding sites for a transcription factor which activates the expression of the gene. Such RE are also called enhancers. Therefore, one can say that 14/17 of the DMRs which correlate with gene expression are likely to be enhancers instead of silencers.

Conclusion

It has been shown in several studies that DNA methylation and gene expression changes can have very little correlation [114, 124, 125, 126]. This observation has been confirmed by our analysis. There are other gene regulatory mechanisms which are more dynamic than DNA methylation. Chromatin remodelling for example has been shown to play a much greater role in the gene regulation of zebrafish fin regeneration than DNA methylation [127]. It should be interesting to complement our results on gene expression changes after PCB126 exposure with genome-wide chromatin profiling.

Chapter 6

Conclusions

How do genes evolve and how are genes regulated are two of the main questions of modern molecular biology. In this thesis we have tried to shed more light on both questions. The gene family we investigated, DNA methyltransferases (DNMTs), is a great model to do this. It has two members DNMT1, which main function is the maintenance of DNA methylation after cell division and DNMT3 which establishes *de novo* DNA methylation during embryonic development. We also included DNMT2 in our study which actually is a RNA methyltransferase and not a DNA methyltransferase. But it was previously included in the DNMT family and contains the same catalytic domain as DNMT1 and DNMT3.

If one takes into account the whole group of metazoan animals there are many lineage-specific gains or losses of DNMTs. The function of DNMTs is equally interesting since they are the only proteins in Metazoa which can add methylation to DNA. Especially in vertebrates, DNA methylation is one of the most universal gene regulatory mechanisms. Incorrect establishment or maintenance of DNA methylation often has catastrophic consequences from embryonic lethality to the development of a diverse range of cancers.

We have focused on two different groups in the metazoan tree. One was Ecdysozoa (insects, spiders, crustaceans, roundworms) which is one of the two large subdivisions of Protostomia. The other group we studied were Vertebrata (fish, amphibians, mammals) which belong to Deuterostomia. Protostomia are the sister group of Deuterostomia therefore Ecdysozoa and Vertebrata are quite distant from each other within the bilaterian animals. Therefore, it is highly interesting to study the independent evolution of DNA methyltransferases in both groups.

Our study in Ecdysozoa is the phylogenetically most diverse analysis of DNA methylation in this group, to-date. While Arthropoda and Nematoda are its two most studied phyla we also include species from Priapulida, Onychophora and Tardigrada. We therefore analyze five out of seven Ecdysozoa phyla and identified DNMT1 and DNMT3 in

four out of these phyla. In two phyla, Arthropoda and Nematoda, there are lineage-specific losses of DNA methylation but it is not lost in the whole phyla. In Priapulida and Oncychophora the available data was much more limited but we did not detect any species which clearly lost DNA methylation. The only phyla without any detected DNMT1 or DNMT3 genes are Tardigrada. Suggesting the absence of DNA methylation in, at least the currently sequenced, tardigrade species. Our data shows that DNA methyltransferases evolved independently and differently in the studied phyla of Ecdysozoa.

In Vertebrata the picture is very different. There were two rounds of whole genome duplication (1R/2R WGD) in an ancestor of all vertebrates. They most likely occurred after the split from the tunicates which are the closest extant sister group of vertebrates. All species for which DNA methylation has been analyzed show a pattern of genome-wide DNA methylation which means that almost every cytosine in a CpG context (the target motif of DNMTs) is methylated. Outside of vertebrates DNA methylation is present in much lower levels. Naturally, there are no known cases of a vertebrate who lost DNA methylation. We have studied the effects additional rounds of whole genome duplication (WGD) had on the evolution of DNMT genes. A third round (3R) happened in all teleost fish and *Xenopus laevis*. In some groups of teleosts, Salmoniformes and a subgroup of Cypriniformes even a forth round (4R) of whole-genome duplication happened. Immediately after a whole-genome duplication the entire genome and therefore all of its genes are present twice. Most of the additional genes are subsequently lost again and the genome undergoes so-called rediploidization. We can see this pattern very clearly for DNMT1. No species which only underwent the first and second WGD retained more than copy of DNMT1. Even, the species which underwent the teleost-specific third WGD, but no forth one, lost the additional copy already. Only in species with a more recent WGD, up to appr. 100 million years ago, have still kept an additional copy of DNMT1. Keeping in mind that such behavior is the common one it is very interesting to note that the evolution of DNMT3 genes happens quite differently. After each WGD there are additional copies of DNMT3 which are kept. This begins with the 1R/2R WGD after which two copies of DNMT3 are kept. It continues with, on average, five copies of DNMT3 in teleost species after the 3R. In species with an additional forth WGD we even detected ten and more copies of DNMT3 genes.

According to the evolutionary theory and its applications on genome evolution one

assumes that whatever is kept in the genome is likely to have a beneficial function. Given the stark contrast between DNMT1 and DNMT3 evolution it is likely that the retained additional copies of DNMT3 underwent a subfunctionalization. This would mean that they either subfunctionalized their targeting on the DNA or that they subfunctionalized at which stage or in which cell they are expressed. For the latter option there is some evidence available [80]. Our study of DNA methyltransferases included, for the first time, species covering all known whole-genome duplication in vertebrates. By generating such an atlas of DNA methyltransferases after different WGD events we were able to propose a new nomenclature for DNMT genes. It can be used in all vertebrate species and is compatible with the currently used gene names in tetrapods without the need to change them. Most importantly the nomenclature correctly reflects the evolutionary history and therefore the orthology between DNMT genes in species with a 3R and 4R WGD. Since the nomenclature, most frequently, used at the moment does not correctly reflect the orthology of DNMT genes in teleosts, changes would be required. While this is not a welcome process for most researchers used to the current names it would nevertheless simplify future research projects. In zebrafish, for example, there currently are DNMT3bb.1, DNMT3bb.2 and DNMT3bb.3 and DNMT3ba. But confusingly DNMT3bb.1 is more similar to DNMT3ba than to the other DNMT3bb genes since they originate from the same ancestral gene. DNMT3bb.2 and DNMT3bb.3 originate from a local gene duplication specific to Cypriniformes, the other copy of its ancestral gene after the 3R was lost. Therefore, DNMT3bb.1 and DNMT3ba evolve independently since a much longer time, appr. 320 million years while DNMT3bb.2 and DNMT3bb.3 only evolved independently for less than appr. 200 million years. While representing such information might not be important for some research projects, especially if one is already very familiar with DNMTs we believe that biology itself is already complex enough and we should take every chance to simplify its description. Especially, if one wants to investigate DNMT enzymes in species with a 4R WGD the naming becomes even more complicated and it would be very practical if the names represent actual orthology between genes.

Investigating the subfunctionalization of DNMT3 genes in teleosts could give very valuable insights into general the process of gene evolution as well as specifically into the functions of DNMT3. We have started to perform such an project as well by studying the subfunctionalization between DNMT3aa and DNMT3ab in zebrafish. We studied the effect single knockouts and a double knockout of both genes has on genome-wide

DNA methylation. We observed most alterations of DNA methylation in the double knockout and very few hypomethylated regions in the two single knockouts. Therefore we hypothesize that DNMT3aa and DNMT3ab can compensate for each other to a high degree. However, if both are inactivated the four remaining DNMT3b proteins can not full compensate the loss of DNA methylation at certain genomic locations. This was the first investigation of a DNMT3 double knockout in zebrafish ever. While the results are quite interesting and novel the effects of the double knockout on DNA methylation is relatively moderate and we could not observe a notable phenotype. In addition a difficulty of studying DNA methylation comes into play. While the DNA is the same in every cell, DNA methylation is not and it can even dynamically change within a cell. We have only studied one time point and one tissue, actually a whole embryo (10 hours post fertilization). It is quite likely that the changes of DNA methylation are different at other time points or specific body parts. The NIH Roadmap Epigenomics Consortium for example, analyzed, among others, DNA methylation in 111 different human cell lines and found plenty of differences [128]. Therefore, while our analysis is a great starting point for investigating the subfunctionalization of DNMT3 genes in zebrafish there is plenty of room for additional work. Next to the difficulty of studying DNA methylation in general the fact that four different DNMT3 genes remain active makes it also difficult entangle the subfunctionalizations. Ideally, one would have transgenic lines where five of the six DNMT3 genes are inactivated and only one remains active, assuming this state is viable. This would allow to specifically study the function of each DNMT3 gene individually. We have already tried to generate additional DNMT3b knockout lines using Crispr/Cas9 but the similarity and close vicinity of the DNMT3b genes makes it a difficult task. If one would succeed to generate additional single knockouts the cross breeding of the resulting specimen would take several months in every crossing step since zebrafish reach maturity after at least 3 month. Therefore, while promising more in-depth studies would be time-intensive due to the cross breeding and expensive, if different cell types or time points are included.

Aside from the curious case of subfunctionalization of DNMT3s in zebrafish it is also a great model species for studying environmental effects. We exposed zebrafish to different levels of the chemical PCB126 and subsequently analyzed the changes in DNA methylation and gene expression in the testis. While there was a notable effect on the gene expression level the alterations of DNA methylation were less pronounced. There have not been many genome-wide studies analyzing environmental effects on

DNA methylation in zebrafish. In a similar experimental setup which analyzed brain and liver tissue the results were comparable [114]. While we certainly have been able to detect an effect on DNA methylation it might not be the primary driver of gene expression changes. Therefore, one should try to study the impact of other gene regulatory mechanisms, e.g. histone modifications as well, to learn which mechanism has the greatest impact. Aside from that there is still the possibility that there is a stronger effect on DNA methylation at a different time point or cell type. It would be interesting to investigate several time points to see at which pace DNA methylation is changed and therefore estimate if more alterations can be expected after additional time.

In this thesis we have broadly expanded the phylogenetic range of species with a manually curated set of DNA methyltransferases. We have done this for ecdysozoan species which have lost all DNA methylating enzymes as well as for teleost fish which acquired more than ten copies of the, originally, two genes. We hope that our systematic approach for annotating and classifying DNA methylating enzymes in such a large range of species can be helpful to future comparative projects. We would be especially delighted if our effort to systematize the nomenclature of DNMT genes would prove to be useful. We were able to generate new insight into the subfunctionalization of the DNA methylation machinery in zebrafish and how it reacts to environmental effects. There is still much less knowledge available about how DNA methylation is regulated in zebrafish compared to mouse or human systems. Nevertheless we hope our work can inspire continuing effort to study DNA methylation outside of mammalian model organisms.

Supplement

6.1 Evolution of DNA methylation across Ecdysozoa

Species	DB	Download link
Drosophila melanogaster	N	../GCF/000/001/215/GCF_000001215.4_Release_6_plus_ISO1_MT/GCF_000001215.4_Release_6_plus_ISO1_MT_protein.faa.gz
Aedes aegypti	N	../GCF/002/204/515/GCF_002204515.2_AaegL5.0/GCF_002204515.2_AaegL5.0_protein.faa.gz
Anopheles gambiae	N	../GCF/000/005/575/GCF_000005575.2_AgamP3/GCF_000005575.2_AgamP3_protein.faa.gz
Ctenocephalides felis	N	../GCF/003/426/905/GCF_003426905.1_ASM342690v1/GCF_003426905.1_ASM342690v1_protein.faa.gz
Bombyx mori	N	../GCF/000/151/625/GCF_000151625.1_ASM15162v1/GCF_000151625.1_ASM15162v1_protein.faa.gz
Danaus plexippus	N	../GCA/000/235/995/GCA_000235995.2_Dpv3/GCA_000235995.2_Dpv3_protein.faa.gz
Operophtera brumata	N	../GCA/001/266/575/GCA_001266575.1_ASM126657v1/GCA_001266575.1_ASM126657v1_protein.faa.gz
Heliconius melpomene	E	../heliconius_melpomene/pep/Heliconius_melpomene.Hmel1.pep.all.fa.gz
Melitaea cinxia	E	../melitaea_cinxia/pep/Melitaea_cinxia.MelCinx1.0.pep.all.fa.gz
Papilio xuthus	N	../GCF/000/836/235/GCF_000836235.1_Pxut_1.0/GCF_000836235.1_Pxut_1.0_protein.faa.gz
Plutella xylostella	N	../GCF/000/330/985/GCF_000330985.1_DBM_FJ_V1.1/GCF_000330985.1_DBM_FJ_V1.1_protein.faa.gz
Limnephilus lunatus	O	http://download.lepbase.org/v4/sequence/Limnephilus_lunatus_v1_-_proteins.fa.gz
Agrilus planipennis	N	../GCF/000/699/045/GCF_000699045.2_Apla_2.0/GCF_000699045.2_Apla_2.0_protein.faa.gz
Nicrophorus vespilloides	N	../GCF/001/412/225/GCF_001412225.1_Nicve_v1.0/GCF_001412225.1_Nicve_v1.0_protein.faa.gz
Onthophagus taurus	N	../GCF/000/648/695/GCF_000648695.1_Otau_2.0/GCF_000648695.1_Otau_2.0_protein.faa.gz
Oryctes borbonicus	N	../GCA/001/443/705/GCA_001443705.1_ASM144370v1/GCA_001443705.1_ASM144370v1_protein.faa.gz
Anoplophora glabripennis	N	../GCF/000/390/285/GCF_000390285.2_Agla_2.0/GCF_000390285.2_Agla_2.0_protein.faa.gz
Leptinotarsa decemlineata	N	../GCF/000/500/325/GCF_000500325.1_Ldec_2.0/GCF_000500325.1_Ldec_2.0_protein.faa.gz
Diabrotica virgifera	N	../GCF/003/013/835/GCF_003013835.1_Dvir_v2.0/GCF_003013835.1_Dvir_v2.0_protein.faa.gz
Dendroctonus ponderosae	N	../GCF/000/355/655/GCF_000355655.1_DendPond_male_1.0/GCF_000355655.1_DendPond_male_1.0_protein.faa.gz
Aethina tumida	N	../GCF/001/937/115/GCF_001937115.1_Atum_1.0/GCF_001937115.1_Atum_1.0_protein.faa.gz
Tribolium castaneum	N	../GCF/000/002/335/GCF_000002335.3_Tcas5.2/GCF_000002335.3_Tcas5.2_protein.faa.gz
Asbolus verrucosus	N	../GCA/004/193/795/GCA_004193795.1_BDFB_1.0/GCA_004193795.1_BDFB_1.0_protein.faa.gz
Apis mellifera	N	../GCF/003/254/395/GCF_003254395.2_Amel_HAv3.1/GCF_003254395.2_Amel_HAv3.1_protein.faa.gz
Bombus impatiens	N	../GCF/000/188/095/GCF_000188095.3_BIMP_2.2/GCF_000188095.3_BIMP_2.2_protein.faa.gz
Atta cephalotes	N	../GCF/000/143/395/GCF_000143395.1_Attacep1.0/GCF_000143395.1_Attacep1.0_protein.faa.gz
Acromyrmex echinatior	N	../GCF/000/204/515/GCF_000204515.1_Aech_3.9/GCF_000204515.1_Aech_3.9_protein.faa.gz
Harpegnathos saltator	N	../GCF/003/227/715/GCF_003227715.1_Hsal_v8.5/GCF_003227715.1_Hsal_v8.5_protein.faa.gz
Solenopsis invicta	N	../GCF/000/188/075/GCF_000188075.2_Si_gnH/GCF_000188075.2_Si_gnH_protein.faa.gz
Polistes dominula	N	../GCF/001/465/965/GCF_001465965.1_Pdom_r1.2/GCF_001465965.1_Pdom_r1.2_protein.faa.gz
Polistes canadensis	N	../GCF/001/313/835/GCF_001313835.1_ASM131383v1/GCF_001313835.1_ASM131383v1_protein.faa.gz
Nasonia vitripennis	N	../GCF/000/002/325/GCF_000002325.3_Nvit_2.1/GCF_000002325.3_Nvit_2.1_protein.faa.gz
Cephus cinctu	N	../GCF/000/341/935/GCF_000341935.1_Ccin1/GCF_000341935.1_Ccin1_protein.faa.gz
Orussus abietinus	N	../GCF/000/612/105/GCF_000612105.2_Oabi_2.0/GCF_000612105.2_Oabi_2.0_protein.faa.gz

Athalia rosae	N	../GCF/000/344/095/GCF_000344095.2_Aros_2.0/GCF_000344095.2_Aros_2.0_protein.faa.gz
Pediculus humanus	N	../GCF/000/006/295/GCF_000006295.1_JCVI_LOUSE_1.0/GCF_000006295.1_JCVI_LOUSE_1.0_protein.faa.gz
Nilaparvata lugens	N	../GCF/000/757/685/GCF_000757685.1_NilLug1.0/GCF_000757685.1_NilLug1.0_protein.faa.gz
Laodelphax striatellus	N	../GCA/003/335/185/GCA_003335185.2_ASM333518v2/GCA_003335185.2_ASM333518v2_protein.faa.gz
Rhodnius prolixus	E	../rhodnius_prolixus/pep/Rhodnius_prolixus.RproC3.pep.all.fa.gz
Cimex lectularius	N	../GCF/000/648/675/GCF_000648675.2_Clec_2.1/GCF_000648675.2_Clec_2.1_protein.faa.gz
Halyomorpha halys	N	../GCF/000/696/795/GCF_000696795.2_Hhal_2.0/GCF_000696795.2_Hhal_2.0_protein.faa.gz
Bemisia tabaci	N	../GCF/001/854/935/GCF_001854935.1_ASM185493v1/GCF_001854935.1_ASM185493v1_protein.faa.gz
Melanaphis sacchari	N	../GCF/002/803/265/GCF_002803265.2_SCAv2.0/GCF_002803265.2_SCAv2.0_protein.faa.gz
Aphis gossypii	N	../GCF/004/010/815/GCF_004010815.1_ASM401081v1/GCF_004010815.1_ASM401081v1_protein.faa.gz
Rhopalosiphum maidis	N	../GCF/003/676/215/GCF_003676215.2_ASM367621v3/GCF_003676215.2_ASM367621v3_protein.faa.gz
Acyrthosiphon pisum	N	../GCF/005/508/785/GCF_005508785.1_pea_aphid_22Mar2018_4r6ur/GCF_005508785.1_pea_aphid_22Mar2018_4r6ur_protein.faa.gz
Diuraphis noxia	N	../GCF/001/186/385/GCF_001186385.1_Dnoxia_1.0/GCF_001186385.1_Dnoxia_1.0_protein.faa.gz
Myzus persicae	N	../GCF/001/856/785/GCF_001856785.1_MPER_G0061.0/GCF_001856785.1_MPER_G0061.0_protein.faa.gz
Sipha flava	N	../GCF/003/268/045/GCF_003268045.1_YSA_version1/GCF_003268045.1_YSA_version1_protein.faa.gz
Diaphorina citri	N	../GCF/000/475/195/GCF_000475195.1_Diaci_psyllid_genome_assembly_version_1.1/GCF_000475195.1_Diaci_psyllid_genome_assembly_version_1.1_protein.faa.gz
Frankliniella occidentalis	N	../GCF/000/697/945/GCF_000697945.2_Focc_2.1/GCF_000697945.2_Focc_2.1_protein.faa.gz
Zootermopsis nevadensis	N	../GCF/000/696/155/GCF_000696155.1_ZooNev1.0/GCF_000696155.1_ZooNev1.0_protein.faa.gz
Cryptotermes secundus	N	../GCF/002/891/405/GCF_002891405.2_Csec_1.0/GCF_002891405.2_Csec_1.0_protein.faa.gz
Blattela germanica	N	../GCA/003/018/175/GCA_003018175.1_Bger_1.1/GCA_003018175.1_Bger_1.1_protein.faa.gz
Orchesella cincta	N	../GCA/001/718/145/GCA_001718145.1_ASM171814v1/GCA_001718145.1_ASM171814v1_protein.faa.gz
Folsomia candida	N	../GCF/002/217/175/GCF_002217175.1_ASM221717v1/GCF_002217175.1_ASM221717v1_protein.faa.gz
Daphnia pulex	N	../GCA/000/187/875/GCA_000187875.1_V1.0/GCA_000187875.1_V1.0_protein.faa.gz
Daphnia magna	E	../daphnia_magna/pep/Daphnia_magna.daphmag2.4.pep.all.fa.gz
Tigriopus californicus	N	../GCA/007/210/705/GCA_007210705.1_Tcal_SD_v2.1/GCA_007210705.1_Tcal_SD_v2.1_protein.faa.gz
Lepeophtheirus salmonis	E	ftp://ftp.ensemblgenomes.org/pub/metazoa/release-43/fasta/lepeophtheirus_salmonis/pep/Lepeophtheirus_salmonis.LSalAtl2s.pep.all.fa.gz
Eurytemora affinis	N	../GCF/000/591/075/GCF_000591075.1_Eaff_2.0/GCF_000591075.1_Eaff_2.0_protein.faa.gz
Calanus finmarchicus	O	https://datadryad.org/stash/dataset/doi:10.5061/dryad.293kp3d
Hyalella azteca	N	../GCF/000/764/305/GCF_000764305.1_Hazt_2.0/GCF_000764305.1_Hazt_2.0_protein.faa.gz
Armadillidium vulgare	N	../GCA/004/104/545/GCA_004104545.1_Arma_vul_BF2787/GCA_004104545.1_Arma_vul_BF2787_protein.faa.gz
Penaeus vannamei	N	../GCF/003/789/085/GCF_003789085.1_ASM378908v1/GCF_003789085.1_ASM378908v1_protein.faa.gz
Glomeris marginata	O	https://datadryad.org/stash/dataset/doi:10.5061/dryad.293kp3d
Eudigraphis taiwaniensis	O	https://datadryad.org/stash/dataset/doi:10.5061/dryad.293kp3d
Strigamia maritima	E	ftp://ftp.ensemblgenomes.org/pub/release-44/metazoa/fasta/strigamia_maritima/pep/

		Strigamia_maritima.Smar1.pep.all.fa.gz
Leptotrombidium deliense	N	../GCA/003/675/905/GCA_003675905.1_ASM367590v1/GCA_003675905.1_ASM367590v1_protein.faa.gz
Dinothrombium tinctorium	N	../GCA/003/675/995/GCA_003675995.1_ASM367599v1/GCA_003675995.1_ASM367599v1_protein.faa.gz
Tetranychus urticae	N	../GCF/000/239/435/GCF_000239435.1_ASM23943v1/GCF_000239435.1_ASM23943v1_protein.faa.gz
Euroglyphus maynei	N	../GCA/002/135/145/GCA_002135145.1_EurM1.0/GCA_002135145.1_EurM1.0_protein.faa.gz
Dermatophagoides pteronyssinus	N	../GCF/001/901/225/GCF_001901225.1_ASM190122v2/GCF_001901225.1_ASM190122v2_protein.faa.gz
D. pteronyssinus	N	../GCF/001/901/225/GCF_001901225.1_ASM190122v2/GCF_001901225.1_ASM190122v2_protein.faa.gz
Sarcoptes scabiei	N	../GCA/000/828/355/GCA_000828355.1_SarSca1.0/GCA_000828355.1_SarSca1.0_protein.faa.gz
Varroa destructor	N	../GCF/002/443/255/GCF_002443255.1_Vdes_3.0/GCF_002443255.1_Vdes_3.0_protein.faa.gz
Varroa jacobsoni	N	../GCF/002/532/875/GCF_002532875.1_vjacob_1.0/GCF_002532875.1_vjacob_1.0_protein.faa.gz
Tropilaelaps mercedesae	N	../GCA/002/081/605/GCA_002081605.1_T._mercedesae_v01/GCA_002081605.1_T._mercedesae_v01_protein.faa.gz
Galendromus occidentalis	N	../GCF/000/255/335/GCF_000255335.1_Mocc_1.0/GCF_000255335.1_Mocc_1.0_protein.faa.gz
Ixodes scapularis	N	../GCF/002/892/825/GCF_002892825.2_ISE6_asm2.2_deduplicated/ GCF_002892825.2_ISE6_asm2.2_deduplicated_protein.faa.gz
Trichonephila clavipes	N	../GCA/002/102/615/GCA_002102615.1_NepCla1.0/GCA_002102615.1_NepCla1.0_protein.faa.gz
Parasteatoda tepidariorum	N	../GCF/000/365/465/GCF_000365465.2_Ptep_2.0/GCF_000365465.2_Ptep_2.0_protein.faa.gz
Stegodyphus mimosarum	N	../GCA/000/611/955/GCA_000611955.2_Stegodyphus_mimosarum_v1/ GCA_000611955.2_Stegodyphus_mimosarum_v1_protein.faa.gz
Centruroides sculpturatus	N	../GCF/000/671/375/GCF_000671375.1_Cexi_2.0/GCF_000671375.1_Cexi_2.0_protein.faa.gz
Limulus polyphemus	N	../GCF/000/517/525/GCF_000517525.1_Limulus_polyphemus-2.1.2/ GCF_000517525.1_Limulus_polyphemus-2.1.2_protein.faa.gz
Anoplodactylus insignis	O	https://datadryad.org/stash/dataset/doi:10.5061/dryad.293kp3d
Hypsibius dujardini	N	../GCA/002/082/055/GCA_002082055.1_nHd_3.1/GCA_002082055.1_nHd_3.1_protein.faa.gz
Ramazzottius varieornatus	N	../GCA/001/949/185/GCA_001949185.1_Rvar_4.0/GCA_001949185.1_Rvar_4.0_protein.faa.gz
Peripatopsis overbergiensis	O	https://datadryad.org/stash/dataset/doi:10.5061/dryad.293kp3d
Peripatoides sp	O	https://datadryad.org/stash/dataset/doi:10.5061/dryad.293kp3d
Peripatus sp	O	https://datadryad.org/stash/dataset/doi:10.5061/dryad.293kp3d
Caenorhabditis sinica	W	../caenorhabditis_sinica/PRJNA194557/caenorhabditis_sinica.PRJNA194557.WBPS14.protein.fa.gz
Caenorhabditis briggsae	O	ftp://ftp.wormbase.org/pub/wormbase/releases/WS271/species/c_briggsae/PRJNA10731/ c_briggsae.PRJNA10731.WS271.protein.fa.gz
Caenorhabditis remanei	W	../caenorhabditis_remanei/PRJNA248909/caenorhabditis_remanei.PRJNA248909.WBPS14.protein.fa.gz
Caenorhabditis tropicalis	W	../caenorhabditis_tropicalis/PRJNA53597/caenorhabditis_tropicalis.PRJNA53597.WBPS14.protein.fa.gz
Caenorhabditis brenneri	W	../caenorhabditis_brenneri/PRJNA20035/caenorhabditis_brenneri.PRJNA20035.WBPS14.protein.fa.gz
Caenorhabditis elegans	O	ftp://ftp.wormbase.org/pub/wormbase/releases/WS271/species/c_elegans/PRJNA13758/ c_elegans.PRJNA13758.WS271.protein.fa.gz
Caenorhabditis japonica	O	ftp://ftp.wormbase.org/pub/wormbase/releases/WS271/species/c_japonica/PRJNA12591/ c_japonica.PRJNA12591.WS271.protein.fa.gz

Caenorhabditis angaria	W	../caenorhabditis_angaria/PRJNA51225/caenorhabditis_angaria.PRJNA51225.WBPS14.protein.fa.gz
Diploscapter pachys	W	../diploscapter_pachys/PRJNA280107/diploscapter_pachys.PRJNA280107.WBPS14.protein.fa.gz
Diploscapter coronatus	W	../diploscapter_coronatus/PRJDB3143/diploscapter_coronatus.PRJDB3143.WBPS14.protein.fa.gz
Angiostrongylus cantonensis	W	../angiostrongylus_cantonensis/PRJEB493/angiostrongylus_cantonensis.PRJEB493.WBPS14.protein.fa.gz
Dictyocaulus viviparus	W	../dictyocaulus_viviparus/PRJNA72587/dictyocaulus_viviparus.PRJNA72587.WBPS14.protein.fa.gz
Haemonchus contortus	W	../haemonchus_contortus/PRJEB506/haemonchus_contortus.PRJEB506.WBPS14.protein.fa.gz
Nippostrongylus brasiliensis	W	../nippostrongylus_brasiliensis/PRJEB511/nippostrongylus_brasiliensis.PRJEB511.WBPS14.protein.fa.gz
Ancylostoma ceylanicum	W	../ancylostoma_ceylanicum/PRJNA231479/ancylostoma_ceylanicum.PRJNA231479.WBPS14.protein.fa.gz
Necator americanus	W	../necator_americanus/PRJNA72135/necator_americanus.PRJNA72135.WBPS14.protein.fa.gz
Pristionchus pacificus	O	ftp://ftp.wormbase.org/pub/wormbase/releases/WS271/species/p_pacificus/PRJNA12644/p_pacificus.PRJNA12644.WS271.protein.fa.gz
Loa loa	W	../loa_loa/PRJNA246086/loa_loa.PRJNA246086.WBPS14.protein.fa.gz
Brugia malayi	O	ftp://ftp.wormbase.org/pub/wormbase/releases/WS271/species/b_malayi/PRJNA10729/b_malayi.PRJNA10729.WS271.protein.fa.gz
Litomosoides sigmodontis	W	../litomosoides_sigmodontis/PRJEB3075/litomosoides_sigmodontis.PRJEB3075.WBPS14.protein.fa.gz
Onchocerca volvulus	O	ftp://ftp.wormbase.org/pub/wormbase/releases/WS271/species/o_volvulus/PRJEB513/o_volvulus.PRJEB513.WS271.protein.fa.gz
Dirofilaria immitis	W	../dirofilaria_immitis/PRJEB1797/dirofilaria_immitis.PRJEB1797.WBPS14.protein.fa.gz
Thelazia callipaeda	W	../thelazia_callipaeda/PRJEB1205/thelazia_callipaeda.PRJEB1205.WBPS14.protein.fa.gz
Dracunculus medinensis	W	../dracunculus_medinensis/PRJEB500/dracunculus_medinensis.PRJEB500.WBPS14.protein.fa.gz
Toxocara canis	W	../toxocara_canis/PRJEB533/toxocara_canis.PRJEB533.WBPS14.protein.fa.gz
Ascaris suum	W	../ascaris_suum/PRJNA62057/ascaris_suum.PRJNA62057.WBPS14.protein.fa.gz
Syphacia muris	W	../syphacia_muris/PRJEB524/syphacia_muris.PRJEB524.WBPS14.protein.fa.gz
Globodera pallida	W	../globodera_pallida/PRJEB123/globodera_pallida.PRJEB123.WBPS14.protein.fa.gz
Meloidogyne hapla	W	../meloidogyne_hapla/PRJNA29083/meloidogyne_hapla.PRJNA29083.WBPS14.protein.fa.gz
Bursaphelenchus xylophilus	W	../bursaphelenchus_xylophilus/PRJEA64437/bursaphelenchus_xylophilus.PRJEA64437.WBPS14.protein.fa.gz
Panagrellus redivivus	W	../panagrellus_redivivus/PRJNA186477/panagrellus_redivivus.PRJNA186477.WBPS14.protein.fa.gz
Strongyloides ratti	O	ftp://ftp.wormbase.org/pub/wormbase/releases/WS271/species/s_ratti/PRJEB125/s_ratti.PRJEB125.WS271.protein.fa.gz
Rhabditophanes sp.	W	../rhabditophanes_kr3021/PRJEB1297/rhabditophanes_kr3021.PRJEB1297.WBPS14.protein.fa.gz
Plectus sambesii	W	../plectus_sambesii/PRJNA390260/plectus_sambesii.PRJNA390260.WBPS14.protein.fa.gz
Trichuris trichiura	W	../trichuris_trichiura/PRJEB535/trichuris_trichiura.PRJEB535.WBPS14.protein.fa.gz
Trichuris suis	W	../trichuris_suis/PRJNA179528/trichuris_suis.PRJNA179528.WBPS14.protein.fa.gz
Trichuris muris	O	ftp://ftp.wormbase.org/pub/wormbase/releases/WS271/species/t_muris/PRJEB126/t_muris.PRJEB126.WS271.protein.fa.gz
Trichinella nelsoni	W	../trichinella_nelsoni/PRJNA257433/trichinella_nelsoni.PRJNA257433.WBPS14.protein.fa.gz
Trichinella spiralis	W	../trichinella_spiralis/PRJNA12603/trichinella_spiralis.PRJNA12603.WBPS14.protein.fa.gz

Trichinella britovi	W	../trichinella_britovi/PRJNA257433/trichinella_britovi.PRJNA257433.WBPS14.protein.fa.gz
Soboliphyme baturini	W	../soboliphyme_baturini/PRJEB516/soboliphyme_baturini.PRJEB516.WBPS14.protein.fa.gz
Romanomermis culicivorax	W	../romanomermis_culicivorax/PRJEB1358/romanomermis_culicivorax.PRJEB1358.WBPS14.protein.fa.gz
Priapulus caudatus	N	../GCF/000/485/595/GCF_000485595.1_Priapulus_caudatus-5.0.1/ GCF_000485595.1_Priapulus_caudatus-5.0.1_protein.faa.gz
Nematostella vectensis	N	../GCF/000/209/225/GCF_000209225.1_ASM20922v1/GCF_000209225.1_ASM20922v1_protein.faa.gz
Hydra vulgaris	N	../GCF/000/004/095/GCF_000004095.1_Hydra_RP_1.0/GCF_000004095.1_Hydra_RP_1.0_protein.faa.gz
Trichoplax adhaerens	N	../GCF/000/150/275/GCF_000150275.1_v1.0/GCF_000150275.1_v1.0_protein.faa.gz
Trichoplax H2	N	../GCA/003/344/405/GCA_003344405.1_TrispH2_1.0/GCA_003344405.1_TrispH2_1.0_protein.faa.gz
Amphimedon queenslandica	N	../GCF/000/090/795/GCF_000090795.1_v1.0/GCF_000090795.1_v1.0_protein.faa.gz

Table 6.1: *The given url is the link to the predicted protein data. The abbreviations in the second colum stand for the following databases: N - NCBI, W - Wormbase, E - ENSEMBL, O - Other. In case of datadryad.org no additional data was available. For the NCBI, Wormbase and ENSEMBL, genome and CDS data was downloaded for the same version of the respective species. Due to their length some URL are shortened. If they are from one of the following database they start with the given address: NCBI - ftp://ftp.ncbi.nlm.nih.gov/genomes/all/..; ENSEMBL - ftp://ftp.ensemblgenomes.org/pub/metazoa/release-44/fasta/..; Wormbase ParaSite - ftp://ftp.ebi.ac.uk/pub/databases/wormbase/parasite/releases/WBPS14/species/..;*

Species	N50
Drosophila melanogaster	25286936
Aedes aegypti	409777670
Anopheles gambiae	49364325
Ctenocephalides felis	71713785
Bombyx mori	4008358
Danaus plexippus	715714
Operophtera brumata	65630
Heliconius melpomene	194302
Melitaea cinxia	119328
Papilio xuthus	6198915
Plutella xylostella	737182
Limnephilus lunatus	54650
Agrilus planipennis	1113421
Nicrophorus vespilloides	122407
Onthophagus taurus	337157
Oryctes borbonicus	33367
Anoplophora glabripennis	678234
Leptinotarsa decemlineata	139046
Diabrotica virgifera	489108
Dendroctonus ponderosae	628732
Aethina tumida	298879
Tribolium castaneum	15265516
Asbolus verrucosus	5726
Apis mellifera	13619445
Bombus impatiens	1399493
Atta cephalotes	5154485
Acromyrmex echinatior	1110580
Harpegnathos saltator	1078644
Solenopsis invicta	621039
Polistes dominula	1625592
Polistes canadensis	521566
Nasonia vitripennis	897131
Cephus cinctu	622163
Orussus abietinus	612083
Athalia rosae	943070

Pediculus humanus	497057
Nilaparvata lugens	356597
Laodelphax striatellus	1084798
Rhodnius prolixus	1088772
Cimex lectularius	1637644
Halyomorpha halys	393089
Bemisia tabaci	3232964
Melanaphis sacchari	3012626
Aphis gossypii	437960
Rhopalosiphum maidis	93298903
Acyrthosiphon pisum	132544852
Diuraphis noxia	397774
Myzus persicae	435781
Sipha flava	1686648
Diaphorina citri	109898
Frankliniella occidentalis	438040
Zootermopsis nevadensis	751105
Cryptotermes secundus	1184893
Blattela germanica	1056071
Orchesella cincta	65879
Folsomia candida	6519406
Daphnia pulex	642089
Daphnia magna	397658
Tigriopus californicus	15806032
Lepeophtheirus salmonis	478276
Eurytemora affinis	252275
Calanus finmarchicus	n/a
Hyalella azteca	215427
Armadillidium vulgare	51088
Penaeus vannamei	605555
Glomeris marginata	n/a
Eudigraphis taiwaniensis	n/a
Strigamia maritima	139451
Leptotrombidium deliense	2941
Dinothrombium tinctorium	16512
Tetranychus urticae	2993488
Euroglyphus maynei	788

Dermatophagoides pteronyssinus	450436
Sarcoptes scabiei	11557
Varroa destructor	58536683
Varroa jacobsoni	233810
Tropilaelaps mercedesae	28859
Galendromus occidentalis	896831
Ixodes scapularis	835681
Trichonephila clavipes	62959
Parasteatoda tepidariorum	4055356
Stegodyphus mimosarum	480636
Centruroides sculpturatus	537465
Limulus polyphemus	254089
Anoplodactylus insignis	n/a
Hypsibius dujardini	342180
Ramazzottius varieornatus	4740345
Peripatopsis overbergiensis	n/a
Peripatoides sp	n/a
Peripatus sp	n/a
Caenorhabditis sinica	25228
Caenorhabditis briggsae	17485439
Caenorhabditis remanei	1522088
Caenorhabditis tropicalis	20921866
Caenorhabditis brenneri	381961
Caenorhabditis elegans	17493829
Caenorhabditis japonica	94149
Caenorhabditis angaria	79858
Diploscapter pachys	124241
Diploscapter coronatus	1007652
Angiostrongylus cantonensis	43900
Dictyocaulus viviparus	225748
Haemonchus contortus	47382676
Nippostrongylus brasiliensis	33527
Ancylostoma ceylanicum	668412
Necator americanus	211861
Pristionchus pacificus	23915096
Loa loa	180288
Brugia malayi	14214749

Litomosoides sigmodontis	45863
Onchocerca volvulus	25485961
Dirofilaria immitis	71281
Thelazia callipaeda	51228
Dracunculus medinensis	665026
Toxocara canis	31192
Ascaris suum	4646302
Syphacia muris	60730
Globodera pallida	120481
Meloidogyne hapla	37608
Bursaphelenchus xylophilus	949830
Panagrellus redivivus	262414
Strongyloides ratti	11693564
Rhabditophanes sp.	537195
Plectus sambesii	23450
Trichuris trichiura	70602
Trichuris suis	1322386
Trichuris muris	28941788
Trichinella nelsoni	293867
Trichinella spiralis	6373445
Trichinella britovi	147150
Soboliphyme baturini	19774
Romanomermis culicivorax	17632
Priapulus caudatus	209727
Nematostella vectensis	472588
Hydra vulgaris	96317
Trichoplax adhaerens	5978658
Trichoplax H2	376320
Amphimedon queenslandica	120365

Table 6.2: *The N50 value in the last column corresponds to the respective genome assembly.*

6.2 Evolution of DNA methyltransferases after vertebrate whole genome duplications

Species	Download link
Callorhinchus milii	../GCF/000/165/045/GCF_000165045.1_Callorhinchus_milii-6.1.3/GCF_000165045.1_Callorhinchus_milii-6.1.3_protein.faa.gz
Rhincodon typus	../GCF/001/642/345/GCF_001642345.1_ASM164234v2/GCF_001642345.1_ASM164234v2_protein.faa.gz
Erpetoichthys calabaricus	../GCF/900/747/795/GCF_900747795.1_fErpCal1.1/GCF_900747795.1_fErpCal1.1_protein.faa.gz
Lepisosteus oculatus	../GCF/000/242/695/GCF_000242695.1_LepOcu1/GCF_000242695.1_LepOcu1_protein.faa.gz
Danio rerio	../GCF/000/002/035/GCF_000002035.6_GRCz11/GCF_000002035.6_GRCz11_protein.faa.gz
Carassius auratus	../GCF/003/368/295/GCF_003368295.1_ASM336829v1/GCF_003368295.1_ASM336829v1_protein.faa.gz
Cyprinus carpio	../GCF/000/951/615/GCF_000951615.1_common_carp_genome/GCF_000951615.1_common_carp_genome_protein.faa.gz
Sinocyclocheilus grahami	../GCF/001/515/645/GCF_001515645.1_SAMN03320097.WGS_v1.1/GCF_001515645.1_SAMN03320097.WGS_v1.1_protein.faa.gz
Sinocyclocheilus rhinocerous	../GCF/001/515/625/GCF_001515625.1_SAMN03320098_v1.1/GCF_001515625.1_SAMN03320098_v1.1_protein.faa.gz
Esox lucius	../GCF/004/634/155/GCF_004634155.1_Eluc_v4/GCF_004634155.1_Eluc_v4_protein.faa.gz
Salmo trutta	../GCF/901/001/165/GCF_901001165.1_fSalTru1.1/GCF_901001165.1_fSalTru1.1_protein.faa.gz
Salmo salar	../GCF/000/233/375/GCF_000233375.1_ICSASG_v2/GCF_000233375.1_ICSASG_v2_protein.faa.gz
Oncorhynchus mykiss	../GCF/002/163/495/GCF_002163495.1_Omyk_1.0/GCF_002163495.1_Omyk_1.0_protein.faa.gz
Oncorhynchus nerka	../GCF/006/149/115/GCF_006149115.1_Oner_1.0/GCF_006149115.1_Oner_1.0_protein.faa.gz
Coregonus sp	../GCA/902/810/595/GCA_902810595.1_AWG_v2/GCA_902810595.1_AWG_v2_protein.faa.gz
Takifugu rubripes	../GCF/901/000/725/GCF_901000725.2_fTakRub1.2/GCF_901000725.2_fTakRub1.2_protein.faa.gz
Tetraodon nigroviridis	../GCA/000/180/735/GCA_000180735.1_ASM18073v1/GCA_000180735.1_ASM18073v1_protein.faa.gz
Oreochromis niloticus	../GCF/001/858/045/GCF_001858045.2_O_niloticus_UMD_NMBU/GCF_001858045.2_O_niloticus_UMD_NMBU_protein.faa.gz
Oryzias latipes	../GCF/002/234/675/GCF_002234675.1_ASM223467v1/GCF_002234675.1_ASM223467v1_protein.faa.gz
Xenopus laevis	../GCF/001/663/975/GCF_001663975.1_Xenopus_laevis_v2/GCF_001663975.1_Xenopus_laevis_v2_protein.faa.gz
Xenopus tropicalis	../GCF/000/004/195/GCF_000004195.4_UCB_Xtro_10.0/GCF_000004195.4_UCB_Xtro_10.0_protein.faa.gz
Homo sapiens	../GCF/000/001/405/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_protein.faa.gz
Cyprinus carpioGM^{E}	../cyprinus_carpio_germanmirror/pep/Cyprinus_carpio_germanmirror.German_Mirror_carp_1.0.pep.all.fa.gz
Cyprinus carpioHB^{E}	../cyprinus_carpio_hebaored/pep/Cyprinus_carpio_hebaored.Hebao_red_carp_1.0.pep.all.fa.gz
Cyprinus carpioYR^{E}	../cyprinus_carpio_huanghe/pep/Cyprinus_carpio_huanghe.Hunaghe_carp_2.0.pep.all.fa.gz
Oxygymnocypris stewartii	../GCA/003/573/665/GCA_003573665.1_Novo_Ost_1.0/GCA_003573665.1_Novo_Ost_1.0_genomic.fna.gz
Hucho hucho	../hucho_hucho/pep/Hucho_hucho.ASM331708v1.pep.all.fa.gz
Thymallus thymallus [76]	https://figshare.com/articles/dataset/Grayling_draft_genome_dataset/5135257
Petromyzon marinus^{E}	../petromyzon_marinus/pep/Petromyzon_marinus.Pmarinus_7.0.pep.all.fa.gz

Table 6.3: *The given url is the link to the used data. Data for species marked with ^{E} is from ENSEMBL otherwise from NCBI or the given citation. Most URL are shortened. They start with the following address: NCBI - ftp://ftp.ncbi.nlm.nih.gov/genomes/all/..; ENSEMBL - ftp://ftp.ensembl.org/pub/release-101/fasta/..;*

6.3 Role of DNA methylation in altered testis gene expression patterns in adult zebrafish (Danio rerio) exposed to Pentachlorobiphenyl (PCB 126)

Biological Process - Hypermethylated DMRs Go term description	p-value	Genes
pigment granule dispersal (GO:0051876)	< 0.0001	kcnj13
pigment granule aggregation in cell center (GO:0051877)	0.0001	kcnj13
establishment of pigment granule localization (GO:0051905)	0.0002	kcnj13
pigment granule localization (GO:0051875)	0.0003	kcnj13
cellular pigmentation (GO:0033059)	0.0003	kcnj13
establishment of vesicle localization (GO:0051650)	0.0004	kcnj13
rhythmic process (GO:0048511)	0.0005	mtnr1ba
vesicle localization (GO:0051648)	0.0006	kcnj13
establishment of organelle localization (GO:0051656)	0.0021	kcnj13
organelle localization (GO:0051640)	0.0024	kcnj13
pigmentation (GO:0043473)	0.0055	kcnj13
axonal defasciculation (GO:0007414)	0.0068	slit2
potassium ion transport (GO:0006813)	0.0094	kcnj13
regulation of ion transmembrane transport (GO:0034765)	0.0112	kcnj13
double-strand break repair via nonhomologous end joining (GO:0006303)	0.0113	xrcc5
regulation of transmembrane transport (GO:0034762)	0.0117	kcnj13
regulation of ion transport (GO:0043269)	0.0127	kcnj13
intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator (GO:0042771)	0.0133	xrcc5
endocardial progenitor cell migration to the midline involved in heart field formation (GO:0003262)	0.0173	slit2
intrinsic apoptotic signaling pathway by p53 class mediator (GO:0072332)	0.0182	xrcc5

Table 6.4: *Biological process GO terms of the genes corresponding to differentially methylated regions (DMRs) in the PCB126 0.3 nM treatment. The twenty termes with the lowest p-value are shown.*

Molecular Function - Hypermethylated DMRs Go term description	p-value	Genes
melatonin receptor activity (GO:0008502)	0.0001	mtnr1ba
inward rectifier potassium channel activity (GO:0005242)	0.0003	kcnj13
voltage-gated potassium channel activity (GO:0005249)	0.0056	kcnj13
potassium channel activity (GO:0005267)	0.0087	kcnj13
potassium ion transmembrane transporter activity (GO:0015079)	0.0088	kcnj13
telomeric DNA binding (GO:0042162)	0.0127	xrcc5
voltage-gated cation channel activity (GO:0022843)	0.0151	kcnj13
non-membrane spanning protein tyrosine kinase activity (GO:0004715)	0.0155	fer
ligand-gated ion channel activity (GO:0015276)	0.0158	kcnj13
ATP-dependent DNA helicase activity (GO:0004003)	0.0193	xrcc5
voltage-gated ion channel activity (GO:0005244)	0.0194	kcnj13
voltage-gated channel activity (GO:0022832)	0.0198	kcnj13
damaged DNA binding (GO:0003684)	0.0263	xrcc5
cation channel activity (GO:0005261)	0.0306	kcnj13
DNA helicase activity (GO:0003678)	0.0352	xrcc5
DNA-dependent ATPase activity (GO:0008094)	0.0394	xrcc5
monovalent inorganic cation transmembrane transporter activity (GO:0015077)	0.0406	kcnj13

Table 6.5: *Molecular function GO terms of the genes corresponding to differentially methylated regions (DMRs) in the PCB126 0.3 nM treatment. The twenty termes with the lowest p-value are shown.*

6.4 Knockout of DNMT3aa and DNMT3ab in zebrafish (Danio rerio)

Biological Process - Hypomethylated DMRs Go term description	p-value	Genes
RNA polyadenylation (GO:0043631)	0.0002	papd4
RNA 3'-end processing (GO:0031123)	0.0004	papd4
GDP catabolic process (GO:0046712)	0.0011	nudt18
dADP catabolic process (GO:0046057)	0.0011	nudt18
dGDP catabolic process (GO:0046067)	0.0011	nudt18
chemokine-mediated signaling pathway (GO:0070098)	0.0018	cxcr7b
deoxyribonucleotide catabolic process (GO:0009264)	0.0020	nudt18
cristae formation (GO:0042407)	0.0109	chchd6a
inner mitochondrial membrane organization (GO:0007007)	0.0136	chchd6a
carbohydrate biosynthetic process (GO:0016051)	0.0146	chst13,pc
mitochondrial membrane organization (GO:0007006)	0.0174	chchd6a
cytokine-mediated signaling pathway (GO:0019221)	0.0186	cxcr7b
deoxyribonucleoside diphosphate metabolic process (GO:0009186)	0.0199	nudt18
2'-deoxyribonucleotide metabolic process (GO:0009394)	0.0202	nudt18
deoxyribonucleotide metabolic process (GO:0009262)	0.0218	nudt18
cellular response to cytokine stimulus (GO:0071345)	0.0223	cxcr7b
microtubule-based movement (GO:0007018)	0.0229	dync1li1
iron-sulfur cluster assembly (GO:0016226)	0.0234	glrx5
gluconeogenesis (GO:0006094)	0.0341	pc
hexose biosynthetic process (GO:0019319)	0.0345	pc

Table 6.6: *Biological process GO terms of the genes corresponding to differentially methylated regions (DMRs) in the PCB126 0.3 nM treatment. The twenty termes with the lowest p-value are shown.*

Molecular Function - Hypomethylated DMRs Go term description	p-value	Genes
polynucleotide adenylyltransferase activity (GO:0004652)	0.0001	papd4
adenylyltransferase activity (GO:0070566)	0.0004	papd4
8-hydroxy-dADP phosphatase activity (GO:0044717)	0.0011	nudt18
8-oxo-GDP phosphatase activity (GO:0044716)	0.0011	nudt18
8-oxo-dGDP phosphatase activity (GO:0044715)	0.0011	nudt18
coreceptor activity (GO:0015026)	0.0018	cxcr7b
ATP binding (GO:0005524)	0.0023	dnaja2,dync1li1,papd4,pc*
adenyl ribonucleotide binding (GO:0032559)	0.0024	dnaja2,dync1li1,papd4,pc*
adenyl nucleotide binding (GO:0030554)	0.0027	dnaja2,dync1li1,papd4,pc*
nucleoside-diphosphatase activity (GO:0017110)	0.0036	nudt18
purine ribonucleoside triphosphate	0.0037	dnaja2,dync1li1,papd4,pc*
binding (GO:0035639)	0.0001	papd4
purine ribonucleoside binding (GO:0032550)	0.0038	dnaja2,dync1li1,papd4,pc*
ribonucleoside binding (GO:0032549)	0.0039	dnaja2,dync1li1,papd4,pc*
nucleoside binding (GO:0001882)	0.0039	dnaja2,dync1li1,papd4,pc*
purine ribonucleotide binding (GO:0032555)	0.0041	dnaja2,dync1li1,papd4,pc*
chemokine binding (GO:0019956)	0.0043	cxcr7b
ribonucleotide binding (GO:0032553)	0.0044	dnaja2,dync1li1,papd4,pc*
purine nucleotide binding (GO:0017076)	0.0045	dnaja2,dync1li1,papd4,pc*
carbohydrate derivative binding (GO:0097367)	0.0059	dnaja2,dync1li1,papd4,pc*
cytokine binding (GO:0019955)	0.0071	cxcr7b

Table 6.7: *Molecular function process GO terms of the genes corresponding to differentially methylated regions (DMRs) in the PCB126 0.3 nM treatment. The twenty termes with the lowest p-value are shown. * - Only the first four genes are shown.*

Biological Process - Hypermethylated DMRs Go term description	p-value	Genes
monovalent inorganic cation transport (GO:0015672)	< 0.0001	atp1a1b,atp5g3a,kcnj13,nnt
pigment granule dispersal (GO:0051876)	< 0.0001	kcnj13
pigment granule aggregation in cell center (GO:0051877)	0.0001	kcnj13
cation transport (GO:0006812)	0.0002	atp1a1b,atp5g3a,kcnj13,nnt
ATP biosynthetic process (GO:0006754)	0.0003	atp1a1b,atp5g3a
establishment of pigment granule localization (GO:0051905)	0.0003	kcnj13
pigment granule localization (GO:0051875)	0.0004	kcnj13
cellular pigmentation (GO:0033059)	0.0004	kcnj13
purine ribonucleoside triphosphate biosynthetic process (GO:0009206)	0.0006	atp1a1b,atp5g3a
proton transport (GO:0015992)	0.0006	atp5g3a,nnt
establishment of vesicle localization (GO:0051650)	0.0006	kcnj13
purine ribonucleoside monophosphate biosynthetic process (GO:0009168)	0.0008	atp1a1b,atp5g3a
rhythmic process (GO:0048511)	0.0008	mtnr1ba
vesicle localization (GO:0051648)	0.0008	kcnj13
ribonucleoside triphosphate biosynthetic process (GO:0009201)	0.0009	atp1a1b,atp5g3a
nucleoside triphosphate biosynthetic process (GO:0009142)	0.0010	atp1a1b,atp5g3a
ion transport (GO:0006811)	0.0013	atp1a1b,atp5g3a,kcnj13,nnt
ribonucleoside monophosphate biosynthetic process (GO:0009156)	0.0013	atp1a1b,atp5g3a
nucleoside monophosphate biosynthetic process (GO:0009124)	0.0014	atp1a1b,atp5g3a
ATP metabolic process (GO:0046034)	0.0015	atp1a1b,atp5g3a

Table 6.8: *Biological process GO terms of the genes corresponding to differentially methylated regions (DMRs) in the PCB126 10 nM treatment. The twenty termes with the lowest p-value are shown.*

Molecular Function - Hypermethylated DMRs Go term description	p-value	Genes
monovalent inorganic cation transmembrane transporter activity (GO:0015077)	< 0.0001	atp1a1b,atp5g3a,kcnj13,slc6a1l
inorganic cation transmembrane transporter activity (GO:0022890)	0.0001	atp1a1b,atp5g3a,kcnj13,slc6a1l
melatonin receptor activity (GO:0008502)	0.0002	mtnr1ba
cation transmembrane transporter activity (GO:0008324)	0.0003	atp1a1b,atp5g3a,kcnj13,slc6a1l
inward rectifier potassium channel activity (GO:0005242)	0.0005	kcnj13
ion transmembrane transporter activity (GO:0015075)	0.0012	atp1a1b,atp5g3a,kcnj13,slc6a1l
substrate-specific transmembrane transporter activity (GO:0022891)	0.0014	atp1a1b,atp5g3a,kcnj13,slc6a1l
3-oxo-5-alpha-steroid 4-dehydrogenase activity (GO:0003865)	0.0016	srd5a2a
transmembrane transporter activity (GO:0022857)	0.0021	atp1a1b,atp5g3a,kcnj13,slc6a1l
substrate-specific transporter activity (GO:0022892)	0.0021	atp1a1b,atp5g3a,kcnj13,slc6a1l
NAD(P)+ transhydrogenase (AB-specific) activity (GO:0008750)	0.0025	nnt
transporter activity (GO:0005215)	0.0050	atp1a1b,atp5g3a,kcnj13,slc6a1l
voltage-gated potassium channel activity (GO:0005249)	0.0081	kcnj13
metal ion transmembrane transporter activity (GO:0046873)	0.0090	kcnj13,slc6a1l
gamma-aminobutyric acid:sodium symporter activity (GO:0005332)	0.0093	slc6a1l
Notch binding (GO:0005112)	0.0099	jag1a
calcium ion binding (GO:0005509)	0.0102	creld2,fstl5,jag1a,lrp1bb
sodium:amino acid symporter activity (GO:0005283)	0.0103	slc6a1l
potassium channel activity (GO:0005267)	0.0125	kcnj13
potassium ion transmembrane transporter activity (GO:0015079)	0.0127	kcnj13

Table 6.9: *Molecular function GO terms of the genes corresponding to differentially methylated regions (DMRs) in the PCB126 10 nM treatment. The twenty termes with the lowest p-value are shown.*

Biological Process - Hypomethylated DMRs Go term description	p-value	Genes
RNA polyadenylation (GO:0043631)	0.0015	papd4
iron-sulfur cluster assembly (GO:0016226)	0.0023	glrx5
calcium-independent cell-cell adhesion (GO:0016338)	0.0030	cldnd
RNA 3'-end processing (GO:0031123)	0.0032	papd4
gluconeogenesis (GO:0006094)	0.0048	pc
hexose biosynthetic process (GO:0019319)	0.0050	pc
monosaccharide biosynthetic process (GO:0046364)	0.0051	pc
response to cadmium ion (GO:0046686)	0.0057	pc
smooth muscle contraction (GO:0006939)	0.0057	si:dkey-63b1.1
vasoconstriction (GO:0042310)	0.0078	si:dkey-63b1.1
muscle contraction (GO:0006936)	0.0199	si:dkey-63b1.1
regulation of blood vessel size (GO:0050880)	0.0206	si:dkey-63b1.1
formation of translation initiation complex (GO:0001732)	0.0235	eif3s10
regulation of GTPase activity (GO:0043087)	0.0238	iqsec3a,tbc1d5,tsc2
regulation of nucleoside metabolic process (GO:0009118)	0.0239	iqsec3a,tbc1d5,tsc2
regulation of purine nucleotide catabolic process (GO:0033121)	0.0239	iqsec3a,tbc1d5,tsc2
negative regulation of canonical Wnt receptor signaling pathway (GO:0090090)	0.0258	gpc3,lzts2a
regulation of melanocyte differentiation (GO:0045634)	0.0287	hipk2
response to metal ion (GO:0010038)	0.0292	pc
intestinal epithelial cell differentiation (GO:0060575)	0.0301	tsc2

Table 6.10: *Biological process GO terms of the genes corresponding to differentially methylated regions (DMRs) in the PCB126 10 nM treatment. The twenty termes with the lowest p-value are shown.*

Molecular Function - Hypomethylated DMRs Go term description	p-value	Genes
ion binding (GO:0043167)	< 0.0001	acox1,cpn1,cygb2,dync1li1*
nucleic acid binding (GO:0003676)	< 0.0001	eif3s10,esrra,hipk2,lhx1b*
metal ion binding (GO:0046872)	< 0.0001	cpn1,cygb2,esrra,glrx5,lhx1b*
cation binding (GO:0043169)	< 0.0001	cpn1,cygb2,esrra,glrx5,lhx1b*
organic cyclic compound binding (GO:0097159)	< 0.0001	acox1,cygb2,dync1li1,eif3s10*
binding (GO:0005488)	< 0.0001	acox1,cldnd,cpn1,cygb2*
heterocyclic compound binding (GO:1901363)	< 0.0001	acox1,cygb2,dync1li1,eif3s10*
bradykinin receptor activity (GO:0004947)	0.0005	si:dkey-63b1.1
polynucleotide adenylyltransferase activity (GO:0004652)	0.0009	papd4
monocarboxylic acid transmembrane transporter activity (GO:0008028)	0.0011	slc16a1,slc6a1l
pyruvate carboxylase activity (GO:0004736)	0.0014	pc
gamma-aminobutyric acid:sodium symporter activity (GO:0005332)	0.0014	slc6a1l
sodium:amino acid symporter activity (GO:0005283)	0.0018	slc6a1l
ligase activity, forming carbon-carbon bonds (GO:0016885)	0.0025	pc
biotin carboxylase activity (GO:0004075)	0.0033	pc
adenylyltransferase activity (GO:0070566)	0.0034	papd4
electron carrier activity (GO:0009055)	0.0050	glrx5
transferase activity (GO:0016740)	0.0069	aanat1,hipk2,mrm1,papd4*
carboxylic acid transmembrane transporter activity (GO:0046943)	0.0085	slc16a1,slc6a1l
microtubule motor activity (GO:0003777)	0.0114	dync1li1,kif21a

Table 6.11: *Molecular function GO terms of the genes corresponding to differentially methylated regions (DMRs) in the PCB126 10 nM treatment. The twenty termes with the lowest p-value are shown. * - Only the first four genes are shown.*

Biological Processes - Hypermethylated DMRs Go term description	p-value	Genes
chloride transmembrane transport (GO:1902476)	0.0120	glra1
RNA polyadenylation (GO:0043631)	0.0135	papd4
iron-sulfur cluster assembly (GO:0016226)	0.0167	glrx5
RNA 3'-end processing (GO:0031123)	0.0196	papd4
response to amino acid stimulus (GO:0043200)	0.0210	glra1
smooth muscle contraction (GO:0006939)	0.0264	si:dkey-63b1.1
vesicle docking involved in exocytosis (GO:0006904)	0.0290	stxbp1b
vasoconstriction (GO:0042310)	0.0309	si:dkey-63b1.1
vesicle docking (GO:0048278)	0.0312	stxbp1b
transforming growth factor beta receptor signaling pathway (GO:0007179)	0.0407	tgfbr2
Biological Processes - Hypomethylated DMRs		
positive regulation of fat cell differentiation (GO:0045600)	0.0002	crebl2
positive regulation of lipid biosynthetic process (GO:0046889)	0.0006	crebl2
positive regulation of glucose import (GO:0046326)	0.0007	crebl2
positive regulation of lipid metabolic process (GO:0045834)	0.0008	crebl2
regulation of filopodium assembly (GO:0051489)	0.0014	gpm6ab
regulation of fat cell differentiation (GO:0045598)	0.0016	crebl2
regulation of cell projection assembly (GO:0060491)	0.0023	gpm6ab
regulation of lipid biosynthetic process (GO:0046890)	0.0047	crebl2
pyrimidine nucleobase catabolic process (GO:0006208)	0.0056	dpysl4
nucleobase catabolic process (GO:0046113)	0.0057	dpysl4
pyrimidine-containing compound catabolic process (GO:0072529)	0.0063	dpysl4
pyrimidine nucleobase metabolic process (GO:0006206)	0.0063	dpysl4
positive regulation of transport (GO:0051050)	0.0064	crebl2
regulation of lipid metabolic process (GO:0019216)	0.0072	crebl2
nucleobase metabolic process (GO:0009112)	0.0147	dpysl4
positive regulation of apoptotic process (GO:0043065)	0.0206	bnip3
pyrimidine-containing compound metabolic process (GO:0072527)	0.0241	dpysl4
positive regulation of cellular process (GO:0048522)	0.0262	bnip3,crebl2
positive regulation of cell differentiation (GO:0045597)	0.0289	crebl2
positive regulation of biological process (GO:0048518)	0.0334	bnip3,crebl2

Table 6.12: *Biological processes GO terms of the genes corresponding to differentially methylated regions (DMRs) in the DNMT3aa knock-out. The twenty termes with the lowest p-value are shown.*

Molecular Function - Hypermethylated DMRs Go term description	p-value	Genes
transforming growth factor beta receptor activity, type II (GO:0005026)	0.0058	tgfbr2
glycine binding (GO:0016594)	0.0070	glra1
transmitter-gated ion channel activity (GO:0022824)	0.0070	glra1
bradykinin receptor activity (GO:0004947)	0.0074	si:dkey-63b1.1
cation binding (GO:0043169)	0.0080	glra1,glrx5,lnpep,neurl1b +9*
polynucleotide adenylyltransferase activity (GO:0004652)	0.0105	papd4
extracellular-glycine-gated chloride channel activity (GO:0016934)	0.0108	glra1
oligosaccharyl transferase activity (GO:0004576)	0.0119	stt3b
gamma-aminobutyric acid:sodium symporter activity (GO:0005332)	0.0131	slc6a1l
sodium:amino acid symporter activity (GO:0005283)	0.0146	slc6a1l
amino acid binding (GO:0016597)	0.0196	glra1
adenylyltransferase activity (GO:0070566)	0.0202	papd4
ion binding (GO:0043167)	0.0232	dync1li1,glra1,glrx5,lnpep +14*
electron carrier activity (GO:0009055)	0.0247	glrx5
metal ion binding (GO:0046872)	0.0258	glrx5,lnpep,neurl1b,papd4 +8*
anion transmembrane transporter activity (GO:0008509)	0.0321	glra1,slc6a1l
amino acid transmembrane transporter activity (GO:0015171)	0.0384	slc6a1l
2 iron, 2 sulfur cluster binding (GO:0051537)	0.0398	glrx5
protein disulfide oxidoreductase activity (GO:0015035)	0.0402	glrx5
disulfide oxidoreductase activity (GO:0015036)	0.0405	glrx5
Molecular Function - Hypomethylated DMRs		
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amides (GO:0016812)	0.0045	dpysl4
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds (GO:0016810)	0.0376	dpysl4

Table 6.13: *Molecular function GO terms of the genes corresponding to differentially methylated regions (DMRs) in the DNMT3aa knock-out. The twenty termes with the lowest p-value are shown. * - Only the first four genes are shown.*

Biological Processes - Hypermethylated DMRs Go term description	p-value	Genes
detection of gravity (GO:0009590)	0.0015	stm
defense response to fungus (GO:0050832)	0.0061	ncf1
respiratory burst (GO:0045730)	0.0084	ncf1
pigment granule dispersal (GO:0051876)	0.0095	kcnj13
chloride transmembrane transport (GO:1902476)	0.0106	glra1
inner ear morphogenesis (GO:0042472)	0.0132	irx1a,stm
ear morphogenesis (GO:0042471)	0.0136	irx1a,stm
pigment granule aggregation in cell center (GO:0051877)	0.0167	kcnj13
response to amino acid stimulus (GO:0043200)	0.0185	glra1
response to heat (GO:0009408)	0.0218	hsf1
cellular process (GO:0009987)	0.0256	dido1,ebf1b,glra1,grna +16*
inner ear development (GO:0048839)	0.0308	irx1a,stm
ear development (GO:0043583)	0.0314	irx1a,stm
establishment of pigment granule localization (GO:0051905)	0.0331	kcnj13
DNA damage checkpoint (GO:0000077)	0.0342	rad9b
erythrocyte development (GO:0048821)	0.0342	rps14
pigment granule localization (GO:0051875)	0.0359	kcnj13
myeloid cell development (GO:0061515)	0.0374	rps14
cellular pigmentation (GO:0033059)	0.0385	kcnj13
response to fungus (GO:0009620)	0.0424	ncf1
Biological Processes - Hypomethylated DMRs		
positive regulation of fat cell differentiation (GO:0045600)	< 0.0001	crebl2
positive regulation of lipid biosynthetic process (GO:0046889)	< 0.0001	crebl2
positive regulation of glucose import (GO:0046326)	< 0.0001	crebl2
positive regulation of lipid metabolic process (GO:0045834)	< 0.0001	crebl2
regulation of fat cell differentiation (GO:0045598)	< 0.0001	crebl2
regulation of lipid biosynthetic process (GO:0046890)	< 0.0001	crebl2
positive regulation of transport (GO:0051050)	< 0.0001	crebl2
regulation of lipid metabolic process (GO:0019216)	0.0001	crebl2
nucleobase-containing compound metabolic process (GO:0006139)	0.0006	atp5g3a,crebl2,dpysl4,gna11b,hel_dr4
heterocycle metabolic process (GO:0046483)	0.0007	atp5g3a,crebl2,dpysl4,gna11b,hel_dr4
cellular aromatic compound metabolic process (GO:0006725)	0.0007	atp5g3a,crebl2,dpysl4,gna11b,hel_dr4
cellular nitrogen compound metabolic process (GO:0034641)	0.0007	atp5g3a,crebl2,dpysl4,gna11b,hel_dr4
organic cyclic compound metabolic process (GO:1901360)	0.0008	atp5g3a,crebl2,dpysl4,gna11b,hel_dr4
positive regulation of cell differentiation (GO:0045597)	0.0009	crebl2
nitrogen compound metabolic process (GO:0006807)	0.0011	atp5g3a,crebl2,dpysl4,gna11b,hel_dr4
nucleobase-containing small molecule metabolic process (GO:0055086)	0.0017	atp5g3a,dpysl4,gna11b
positive regulation of developmental process (GO:0051094)	0.0031	crebl2
phospholipase C-activating dopamine receptor signaling pathway (GO:0060158)	0.0056	gna11b
ATP synthesis coupled proton transport (GO:0015986)	0.0072	atp5g3a
positive regulation of cellular process (GO:0048522)	0.0072	bnip3,crebl2

Table 6.14: *Biological processes GO terms of the genes corresponding to differentially methylated regions (DMRs) in the DNMT3ab knock-out. The twenty termes with the lowest p-value are shown. * - Only the first four genes are shown.*

Molecular Function - Hypermethylated DMRs Go term description	p-value	Genes
serine C-palmitoyltransferase activity (GO:0004758)	0.0014	sptssa
palmitoyltransferase activity (GO:0016409)	0.0016	sptssa
carboxylic acid binding (GO:0031406)	0.0028	egln3,glra1
C-acyltransferase activity (GO:0016408)	0.0028	sptssa
glycine binding (GO:0016594)	0.0062	glra1
transmitter-gated ion channel activity (GO:0022824)	0.0062	glra1
extracellular-glycine-gated chloride channel activity (GO:0016934)	0.0095	glra1
superoxide-generating NADPH oxidase activity (GO:0016175)	0.0113	ncf1
structural constituent of ribosome (GO:0003735)	0.0131	rpl27,rps14
amino acid binding (GO:0016597)	0.0173	glra1
melatonin receptor activity (GO:0008502)	0.0255	mtnr1ba
ligand-gated ion channel activity (GO:0015276)	0.0344	glra1,kcnj13
phosphoprotein phosphatase activity (GO:0004721)	0.0405	pptc7a,ptprb
inward rectifier potassium channel activity (GO:0005242)	0.0405	kcnj13
L-ascorbic acid binding (GO:0031418)	0.0497	egln3
Molecular Function - Hypomethylated DMRs		
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amides (GO:0016812)	0.0072	dpysl4
G-protein beta/gamma-subunit complex binding (GO:0031683)	0.0221	gna11b
DNA helicase activity (GO:0003678)	0.0282	hel_dr4
hydrolase activity (GO:0016787)	0.0447	dpysl4,gna11b,hel_dr4,irbp
receptor binding (GO:0005102)	0.0482	gdf2,gna11b

Table 6.15: *Molecular function GO terms of the genes corresponding to differentially methylated regions (DMRs) in the DNMT3ab knock-out. The twenty termes with the lowest p-value are shown.*

Biological Processes - Hypermethylated DMRs Go term description	p-value	Genes
rhythmic process (GO:0048511)	0.0005	foxl2,mtnr1ba
detection of gravity (GO:0009590)	0.0010	stm
female somatic sex determination (GO:0019101)	0.0013	foxl2
apoptotic DNA fragmentation (GO:0006309)	0.0014	foxl2
extraocular skeletal muscle development (GO:0002074)	0.0028	foxl2
DNA catabolic process (GO:0006308)	0.0031	foxl2
RNA metabolic process (GO:0016070)	0.0032	foxl2,hsf1,pou3f1,ptbp2a
triglyceride biosynthetic process (GO:0019432)	0.0034	agpat9l
CDP-diacylglycerol biosynthetic process (GO:0016024)	0.0047	agpat9l
ovarian follicle development (GO:0001541)	0.0052	foxl2
sex determination (GO:0007530)	0.0054	foxl2
gene expression (GO:0010467)	0.0055	foxl2,hsf1,pou3f1,ptbp2a
nitrogen compound metabolic process (GO:0006807)	0.0057	foxl2,hsf1,pou3f1,ptbp2a,sptssa
pigment granule dispersal (GO:0051876)	0.0064	kcnj13
mRNA processing (GO:0006397)	0.0078	ptbp2a
nucleic acid metabolic process (GO:0090304)	0.0084	foxl2,hsf1,pou3f1,ptbp2a
mRNA metabolic process (GO:0016071)	0.0095	ptbp2a
female gonad development (GO:0008585)	0.0103	foxl2
female sex differentiation (GO:0046660)	0.0105	foxl2
pigment granule aggregation in cell center (GO:0051877)	0.0111	kcnj13
Biological Processes - Hypomethylated DMRs		
positive regulation of fat cell differentiation (GO:0045600)	< 0.0001	crebl2
positive regulation of lipid biosynthetic process (GO:0046889)	< 0.0001	crebl2
positive regulation of glucose import (GO:0046326)	< 0.0001	crebl2
positive regulation of lipid metabolic process (GO:0045834)	< 0.0001	crebl2
regulation of fat cell differentiation (GO:0045598)	0.0001	crebl2
regulation of lipid biosynthetic process (GO:0046890)	0.0005	crebl2
positive regulation of transport (GO:0051050)	0.0009	crebl2
regulation of lipid metabolic process (GO:0019216)	0.0011	crebl2
selenocysteinyl-tRNA(Sec) biosynthetic process (GO:0097056)	0.0039	sepsecs
facial nerve development (GO:0021561)	0.0082	hoxb1a
RNA 5'-end processing (GO:0000966)	0.0083	piwil2
negative regulation of SMAD protein complex assembly (GO:0010991)	0.0083	piwil2
germ-line stem cell maintenance (GO:0030718)	0.0096	piwil2
preganglionic parasympathetic nervous system development (GO:0021783)	0.0116	hoxb1a
positive regulation of cell differentiation (GO:0045597)	0.0169	crebl2
parasympathetic nervous system development (GO:0048486)	0.0174	hoxb1a
glutamine biosynthetic process (GO:0006542)	0.0194	glula
DNA methylation involved in gamete generation (GO:0043046)	0.0202	piwil2
negative regulation of protein complex assembly (GO:0031333)	0.0204	piwil2
olfactory bulb development (GO:0021772)	0.0207	ptprsa

Table 6.16: *Biological processes GO terms of the genes corresponding to differentially methylated regions (DMRs) in the DNMT3aa/ab double knock-out. The twenty termes with the lowest p-value are shown.*

Molecular Function - Hypermethylated DMRs Go term description	p-value	Genes
serine C-palmitoyltransferase activity (GO:0004758)	0.0010	sptssa
palmitoyltransferase activity (GO:0016409)	0.0011	sptssa
ubiquitin conjugating enzyme binding (GO:0031624)	0.0013	foxl2
iron-sulfur cluster binding (GO:0051536)	0.0015	dpyda
C-acyltransferase activity (GO:0016408)	0.0019	sptssa
transferase activity, transferring acyl groups other than amino-acyl groups (GO:0016747)	0.0024	agpat9l,sptssa
glycerol-3-phosphate O-acyltransferase activity (GO:0004366)	0.0027	agpat9l
1-acylglycerol-3-phosphate O-acyltransferase activity (GO:0003841)	0.0033	agpat9l
acylglycerol O-acyltransferase activity (GO:0016411)	0.0040	agpat9l
estrogen receptor binding (GO:0030331)	0.0060	foxl2
transferase activity, transferring acyl groups (GO:0016746)	0.0067	agpat9l,sptssa
cysteine-type endopeptidase regulator activity involved in apoptotic process (GO:0043028)	0.0083	foxl2
steroid hormone receptor binding (GO:0035258)	0.0091	foxl2
nuclear hormone receptor binding (GO:0035257)	0.0144	foxl2
hormone receptor binding (GO:0051427)	0.0162	foxl2
melatonin receptor activity (GO:0008502)	0.0170	mtnr1ba
O-acyltransferase activity (GO:0008374)	0.0208	agpat9l
fibroblast growth factor receptor binding (GO:0005104)	0.0240	fgf10b
inward rectifier potassium channel activity (GO:0005242)	0.0272	kcnj13
L-ascorbic acid binding (GO:0031418)	0.0334	egln3
Molecular Function - Hypomethylated DMRs		
heterocyclic compound binding (GO:1901363)	0.0033	arl8,bbs12,crebl2,dnajc27 +25*
organic cyclic compound binding (GO:0097159)	0.0035	arl8,bbs12,crebl2,dnajc27 +25*
transferase activity, transferring selenium-containing groups (GO:0016785)	0.0039	sepsecs
pyruvate dehydrogenase (acetyl-transferring) activity (GO:0004739)	0.0057	pdha1b
pyruvate dehydrogenase activity (GO:0004738)	0.0063	pdha1b
ATPase activator activity (GO:0001671)	0.0063	ahsa1
nucleic acid binding (GO:0003676)	0.0135	crebl2,gb:am422109,hoxb1a,lbx2 +1
ATPase regulator activity (GO:0060590)	0.0157	ahsa1
oxidoreductase activity, acting on the aldehyde or oxo group of donors, disulfide as acceptor (GO:0016624)	0.0158	pdha1b
piRNA binding (GO:0034584)	0.0179	piwil2
neuropeptide hormone activity (GO:0005184)	0.0192	pomca
glutamate-ammonia ligase activity (GO:0004356)	0.0194	glula
bile acid:sodium symporter activity (GO:0008508)	0.0275	slc10a2
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amides (GO:0016812)	0.0303	dpysl4
transmembrane signaling receptor activity (GO:0004888)	0.0304	gpr19,grid2,ntrk2a,tmtopsb +2*
chaperone binding (GO:0051087)	0.0354	ahsa1
tRNA binding (GO:0000049)	0.0403	sepsecs

Table 6.17: *Molecular function GO terms of the genes corresponding to differentially methylated regions (DMRs) in the DNMT3aa/ab double knock-out. The twenty termes with the lowest p-value are shown. * - Only the first four genes are shown.*

Bibliography

[1] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[2] Lan TM Dao, Ariel O Galindo-Albarrán, Jaime A Castro-Mondragon, Charlotte Andrieu-Soler, Alejandra Medina-Rivera, Charbel Souaid, Guillaume Charbonnier, Aurélien Griffon, Laurent Vanhille, Tharshana Stephen, et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature genetics*, 49(7):1073, 2017.

[3] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008.

[4] Jane Qiu. Unfinished symphony, 2006.

[5] Johanna Taylor Cannon, Bruno Cossermelli Vellutini, Julian Smith, Fredrik Ronquist, Ulf Jondelius, and Andreas Hejnol. Xenacoelomorpha is the sister group to nephrozoa. *Nature*, 530(7588):89–93, 2016.

[6] Casey W Dunn, Andreas Hejnol, David Q Matus, Kevin Pang, William E Browne, Stephen A Smith, Elaine Seaver, Greg W Rouse, Matthias Obst, Gregory D Edgecombe, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188):745–749, 2008.

[7] Frank Lyko. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nature Reviews Genetics*, 19(2):81, 2018.

[8] Assaf Zemach, Ivy E McDaniel, Pedro Silva, and Daniel Zilberman. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, 328(5980):916–919, 2010.

[9] G"unter Raddatz, Paloma M Guzzardo, Nelly Olova, Marcelo Rosado Fantappié, Markus Rampp, Matthias Schaefer, Wolf Reik, Gregory J Hannon, and Frank Lyko. Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proceedings of the National Academy of Sciences*, 110(21):8627–8631, 2013.

[10] Chalasani Devajyothi and Vani Brahmachari. Detection of a CpA methylase in an insect system: characterization and substrate specificity. *Molecular and cellular biochemistry*, 110(2):103–111, 1992.

[11] LM Field, Frank Lyko, Mauro Mandrioli, and G Prantera. DNA methylation in insects. *Insect molecular biology*, 13(2):109–115, 2004.

[12] Adam J Bewick, Kevin J Vogel, Allen J Moore, and Robert J Schmitz. Evolution of DNA methylation across insects. *Molecular biology and evolution*, 34(3):654–665, 2017.

[13] Panagiotis Provataris, Karen Meusemann, Oliver Niehuis, Sonja Grath, and Bernhard Misof. Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola. *Genome biology and evolution*, 10(4):1185–1197, 2018.

[14] Alex de Mendoza, Jahnvi Pflueger, and Ryan Lister. Capture of a functionally active methyl-CpG binding domain by an arthropod retrotransposon family. *Genome research*, 29(8):1277–1286, 2019.

[15] Fanny Gatzmann, Cassandra Falckenhayn, Julian Gutekunst, Katharina Hanna, Günter Raddatz, Vitor Coutinho Carneiro, and Frank Lyko. The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. *Epigenetics & chromatin*, 11(1):57, 2018.

[16] Fei Gao, Xiaolei Liu, Xiu-Ping Wu, Xue-Lin Wang, Desheng Gong, Hanlin Lu, Yudong Xia, Yanxia Song, Junwen Wang, Jing Du, et al. Differential DNA methylation in discrete developmental stages of the parasitic nematode Trichinella spiralis. *Genome biology*, 13(10):R100, 2012.

[17] Silvana Rošić, Rachel Amouroux, Cristina E Requena, Ana Gomes, Max Emperle, Toni Beltran, Jayant K Rane, Sarah Linnett, Murray E Selkirk, Philipp H Schiffer, et al. Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity. *Nature genetics*, 50(3):452–459, 2018.

[18] Mary Grace Goll and Timothy H Bestor. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.*, 74:481–514, 2005.

[19] Albert Jeltsch and Renata Z Jurkowska. New concepts in DNA methylation. *Trends in biochemical sciences*, 39(7):310–318, 2014.

[20] Olya Yarychkivska, Zoha Shahabuddin, Nicole Comfort, Mathieu Boulard, and Timothy H Bestor. BAH domains and a histone-like motif in DNA methyltransferase 1 (DNMT1) regulate de novo and maintenance methylation in vivo. *Journal of Biological Chemistry*, 293(50):19466–19475, 2018.

[21] Nora KE Schulz, C Isabel Wagner, Julia Ebeling, G"unter Raddatz, Maike F Diddens-de Buhr, Frank Lyko, and Joachim Kurtz. Dnmt1 has an essential function despite the absence of CpG DNA methylation in the red flour beetle Tribolium castaneum. *Scientific reports*, 8(1):1–10, 2018.

[22] Jing Liao, Rahul Karnik, Hongcang Gu, Michael J Ziller, Kendell Clement, Alexander M Tsankov, Veronika Akopian, Casey A Gifford, Julie Donaghey, Christina Galonska, et al. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nature genetics*, 47(5):469–478, 2015.

[23] Mary Grace Goll, Finn Kirpekar, Keith A Maggert, Jeffrey A Yoder, Chih-Lin Hsieh, Xiaoyu Zhang, Kent G Golic, Steven E Jacobsen, and Timothy H Bestor. Methylation of trNAasp by the DNA methyltransferase homolog Dnmt2. *Science*, 311(5759):395–398, 2006.

[24] Lakshminarayan M Iyer, Saraswathi Abhiman, and L Aravind. Natural history of eukaryotic DNA methylation systems. In *Progress in molecular biology and translational science*, volume 101, pages 25–104. Elsevier, 2011.

[25] Jason T Huff and Daniel Zilberman. Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell*, 156(6):1286–1297, 2014.

[26] Tomasz P Jurkowski and Albert Jeltsch. On the evolutionary origin of eukaryotic DNA methyltransferases and Dnmt2. *PloS one*, 6(11), 2011.

[27] Michael R Rountree, Kurtis E Bachman, and Stephen B Baylin. DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. *Nature genetics*, 25(3):269–277, 2000.

[28] Timothy H Bestor. Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. *The EMBO journal*, 11(7):2611–2617, 1992.

[29] Jikui Song, Olga Rechkoblit, Timothy H Bestor, and Dinshaw J Patel. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science*, 331(6020):1036–1040, 2011.

[30] Arunkumar Dhayalan, Arumugam Rajavelu, Philipp Rathert, Raluca Tamas, Renata Z Jurkowska, Sergey Ragozin, and Albert Jeltsch. The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *Journal of Biological Chemistry*, 285(34):26114–26120, 2010.

[31] Daniel N Weinberg, Simon Papillon-Cavanagh, Haifen Chen, Yuan Yue, Xiao Chen, Kartik N Rajagopalan, Cynthia Horth, John T McGuire, Xinjing Xu, Hamid Nikbakht, et al. The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature*, 573(7773):281–286, 2019.

[32] Yingying Zhang, Renata Jurkowska, Szabolcs Soeroes, Arumugam Rajavelu, Arunkumar Dhayalan, Ina Bock, Philipp Rathert, Ole Brandt, Richard Reinhardt, Wolfgang Fischle, et al. Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic acids research*, 38(13):4246–4253, 2010.

[33] Steen KT Ooi, Chen Qiu, Emily Bernstein, Keqin Li, Da Jia, Zhe Yang, Hediye Erdjument-Bromage, Paul Tempst, Shau-Ping Lin, C David Allis, et al. Dnmt3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, 448(7154):714–717, 2007.

[34] Zachary D Smith, Hongcang Gu, Christoph Bock, Andreas Gnirke, and Alexander Meissner. High-throughput bisulfite sequencing in mammalian genomes. *Methods*, 48(3):226–232, 2009.

[35] Benoît Aliaga, Ingo Bulla, Gabriel Mouahid, David Duval, and Christoph Grunau. Universality of the DNA methylation codes in Eucaryotes. *Scientific reports*, 9(1):1–11, 2019.

[36] Navin Elango, Brendan G Hunt, Michael AD Goodisman, and V Yi Soojin. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. *Proceedings of the National Academy of Sciences*, 106(27):11206–11211, 2009.

[37] Eric W Sayers, Richa Agarwala, Evan E Bolton, J Rodney Brister, Kathi Canese, Karen Clark, Ryan Connor, Nicolas Fiorini, Kathryn Funk, Timothy Hefferon, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 47(Database issue):D23, 2019.

[38] Todd W Harris, Valerio Arnaboldi, Scott Cain, Juancarlos Chan, Wen J Chen, Jaehyoung Cho, Paul Davis, Sibyl Gao, Christian A Grove, Ranjana Kishore, et al. WormBase: a modern model organism information resource. *Nucleic acids research*, 48(D1):D762–D767, 2020.

[39] Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, et al. Ensembl 2020. *Nucleic acids research*, 48(D1):D682–D688, 2020.

[40] Christopher E Laumer, Rosa Fernández, Sarah Lemer, David Combosch, Kevin M Kocot, Ana Riesgo, Sónia CS Andrade, Wolfgang Sterrer, Martin V Sørensen, and Gonzalo Giribet. Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proceedings of the royal society B*, 286(1906):20190831, 2019.

[41] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, et al. The Pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432, 2019.

[42] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7(1):539, 2011.

[43] Daniel H Huson and David Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, 23(2):254–267, 2006.

[44] W James Kent. BLAT-the BLAST-like alignment tool. *Genome research*, 12(4):656–664, 2002.

[45] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[47] Bernhard Misof, Shanlin Liu, Karen Meusemann, Ralph S Peters, Alexander Donath, Christoph Mayer, Paul B Frandsen, Jessica Ware, Tomáš Flouri, Rolf G Beutel, et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210):763–767, 2014.

[48] Shao-Qian Zhang, Li-Heng Che, Yun Li, Dan Liang, Hong Pang, Adam Ślipiński, and Peng Zhang. Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. *Nature Communications*, 9(1):1–11, 2018.

[49] Akito Y Kawahara, David Plotkin, Marianne Espeland, Karen Meusemann, Emmanuel FA Toussaint, Alexander Donath, France Gimnich, Paul B Frandsen, Andreas Zwick, Mario dos Reis, et al. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proceedings of the National Academy of Sciences*, 116(45):22657–22663, 2019.

[50] Ralph S Peters, Lars Krogmann, Christoph Mayer, Alexander Donath, Simon Gunkel, Karen Meusemann, Alexey Kozlov, Lars Podsiadlowski, Malte Petersen, Robert Lanfear, et al. Evolutionary history of the Hymenoptera. *Current Biology*, 27(7):1013–1018, 2017.

[51] Kevin P Johnson, Christopher H Dietrich, Frank Friedrich, Rolf G Beutel, Benjamin Wipfler, Ralph S Peters, Julie M Allen, Malte Petersen, Alexander Donath, Kimberly KO Walden, et al. Phylogenomics and the evolution of hemipteroid insects. *Proceedings of the National Academy of Sciences*, 115(50):12775–12780, 2018.

[52] Carol D von Dohlen, Carol A Rowe, and Ole E Heie. A test of morphological hypotheses for tribal and subtribal relationships of Aphidinae (Insecta: Hemiptera: Aphididae) using DNA sequences. *Molecular Phylogenetics and Evolution*, 38(2):316–329, 2006.

[53] Hyojoong Kim, Seunghwan Lee, and Yikweon Jang. Macroevolutionary patterns in the Aphidini aphids (Hemiptera: Aphididae): diversification, host association, and biogeographic origins. *PloS one*, 6(9), 2011.

[54] Eva Nováková, Václav Hypša, Joanne Klein, Robert G Foottit, Carol D von Dohlen, and Nancy A Moran. Reconstructing the phylogeny of aphids (Hemiptera: Aphididae) using DNA of the obligate symbiont Buchnera aphidicola. *Molecular Phylogenetics and Evolution*, 68(1):42–54, 2013.

[55] Martin Schwentner, David J Combosch, Joey Pakes Nelson, and Gonzalo Giribet. A phylogenomic solution to the origin of insects by resolving crustacean-hexapod relationships. *Current Biology*, 27(12):1818–1824, 2017.

[56] Sahar Khodami, J Vaun McArthur, Leocadio Blanco-Bercial, and Pedro Martinez Arbizu. Molecular phylogeny and revision of copepod orders (Crustacea: Copepoda). *Scientific reports*, 7(1):1–11, 2017.

[57] Prashant P Sharma, Evelyn E Schwager, Cassandra G Extavour, and Gonzalo Giribet. Evolution of the chelicera: a dachshund domain is retained in the deutocerebral appendage of Opiliones (Arthropoda, Chelicerata). *Evolution & development*, 14(6):522–533, 2012.

[58] Rosa Fernández, Robert J Kallal, Dimitar Dimitrov, Jesús A Ballesteros, Miquel A Arnedo, Gonzalo Giribet, and Gustavo Hormiga. Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. *Current Biology*, 28(9):1489–1497, 2018.

[59] Paula Arribas, Carmelo Andújar, María Lourdes Moraza, Benjamin Linard, Brent C Emerson, and Alfried P Vogler. Mitochondrial metagenomics reveals the ancient origin and phylodiversity of soil mites and provides a phylogeny of the Acari. *Molecular Biology and Evolution*, 37(3):683–694, 2020.

[60] International Helminth Genomes Consortium et al. Comparative genomics of the major parasitic worms. *Nature genetics*, 51(1):163, 2019.

[61] Pasi K Korhonen, Edoardo Pozio, Giuseppe La Rosa, Bill CH Chang, Anson V Koehler, Eric P Hoberg, Peter R Boag, Patrick Tan, Aaron R Jex, Andreas Hofmann, et al. Phylogenomic and biogeographic reconstruction of the Trichinella complex. *Nature communications*, 7(1):1–8, 2016.

[62] Lewis Stevens, Marie-Anne Félix, Toni Beltran, Christian Braendle, Carlos Caurcel, Sarah Fausett, David Fitch, Lise Frézal, Charlie Gosse, Taniya Kaur, et al. Comparative genomics of 10 new Caenorhabditis species. *Evolution Letters*, 3(2):217–236, 2019.

[63] Hél'ene Fradin, Karin Kiontke, Charles Zegar, Michelle Gutwein, Jessica Lucas, Mikhail Kovtun, David L Corcoran, L Ryan Baugh, David HA Fitch, Fabio Piano, et al. Genome architecture and evolution of a unichromosomal asexual nematode. *Current Biology*, 27(19):2928–2939, 2017.

[64] K Ingemar J"onsson, Elke Rabbow, Ralph O Schill, Mats Harms-Ringdahl, and Petra Rettberg. Tardigrades survive exposure to space in low Earth orbit. *Current biology*, 18(17):R729–R731, 2008.

[65] Andrew N Ostrovsky, Scott Lidgard, Dennis P Gordon, Thomas Schwaha, Grigory Genikhovich, and Alexander V Ereskovsky. Matrotrophy and placentation in invertebrates: a new paradigm. *Biological Reviews*, 91(3):673–711, 2016.

[66] Thomas Dahlet, Andrea Arg"ueso Lleida, Hala Al Adhami, Michael Dumas, Ambre Bender, Richard P Ngondo, Manon Tanguy, Judith Vallet, Ghislain Auclair, Ana"ıs F Bardet, et al. Genome-wide analysis in the mouse embryo reveals the importance of DNA methylation for transcription integrity. *Nature communications*, 11(1):1–14, 2020.

[67] Samuel Lewis, Laura Ross, SA Bain, E Pahita, SA Smith, R Cordaux, EM Miska, B Lenhard, FM Jiggins, and Peter Sarkies. Widespread conservation and lineage-specific diversification of genome-wide DNA methylation patterns across arthropods. *PLoS genetics*, 16(6):e1008864, 2020.

[68] Jouni Kvist, Camila Gonçalves Athan'asio, Omid Shams Solari, James B Brown, John K Colbourne, Michael E Pfrender, and Leda Mirbahai. Pattern of DNA methylation in Daphnia: evolutionary perspective. *Genome biology and evolution*, 10(8):1988–2007, 2018.

[69] Daniel S Standage, Ali J Berens, Karl M Glastad, Andrew J Severin, Volker P Brendel, and Amy L Toth. Genome, transcriptome and methylome sequencing of a primitively eusocial wasp reveal a greatly reduced DNA methylation system in a social insect. *Molecular Ecology*, 25(8):1769–1784, 2016.

[70] Roberto Bonasio, Qiye Li, Jinmin Lian, Navdeep S Mutti, Lijun Jin, Hongmei Zhao, Pei Zhang, Ping Wen, Hui Xiang, Yun Ding, et al. Genome-wide and caste-specific DNA methylomes of the ants Camponotus floridanus and Harpegnathos saltator. *Current Biology*, 22(19):1755–1764, 2012.

[71] Brendan G Hunt, Karl M Glastad, Soojin V Yi, and Michael AD Goodisman. Patterning and regulatory associations of DNA methylation are mirrored by histone modifications in insects. *Genome biology and evolution*, 5(3):591–598, 2013.

[72] Linda Z Holland and Daniel Ocampo Daza. A new look at an old question: when did the second whole genome duplication occur in vertebrate evolution? *Genome biology*, 19(1):1–4, 2018.

[73] Jeremy Pasquier, Cédric Cabau, Thaovi Nguyen, Elodie Jouanno, Dany Severac, Ingo Braasch, Laurent Journot, Pierre Pontarotti, Christophe Klopp, John H Postlethwait, et al. Gene evolution and gene expression after whole genome duplication in fish: the phylofish database. *BMC genomics*, 17(1):1–10, 2016.

[74] Peng Xu, Jian Xu, Guangjian Liu, Lin Chen, Zhixiong Zhou, Wenzhu Peng, Yanliang Jiang, Zixia Zhao, Zhiying Jia, Yonghua Sun, et al. The allotetraploid origin and asymmetrical genome evolution of the common carp cyprinus carpio. *Nature communications*, 10(1):1–11, 2019.

[75] Adam M Session, Yoshinobu Uno, Taejoon Kwon, Jarrod A Chapman, Atsushi Toyoda, Shuji Takahashi, Akimasa Fukui, Akira Hikosaka, Atsushi Suzuki, Mariko Kondo, et al. Genome evolution in the allotetraploid frog xenopus laevis. *Nature*, 538(7625):336–343, 2016.

[76] Srinidhi Varadharajan, Simen R Sandve, Gareth B Gillard, Ole K Tørresen, Teshome D Mulugeta, Torgeir R Hvidsten, Sigbjørn Lien, Leif Asbjørn Vøllestad, Sissel Jentoft, Alexander J Nederbragt, et al. The grayling genome reveals selection on gene expression regulation after whole-genome duplication. *Genome Biology and Evolution*, 10(10):2785–2800, 2018.

[77] Hai-Ping Liu, Shi-Jun Xiao, Nan Wu, Di Wang, Yan-Chao Liu, Chao-Wei Zhou, Qi-Yong Liu, Rui-Bin Yang, Wen-Kai Jiang, Qi-Qi Liang, et al. The sequence and de novo assembly of oxygymnocypris stewartii genome. *Scientific data*, 6:190009, 2019.

[78] Ricardo Betancur-R, Edward O Wiley, Gloria Arratia, Arturo Acero, Nicolas Bailly, Masaki Miya, Guillaume Lecointre, and Guillermo Orti. Phylogenetic classification of bony fishes. *BMC evolutionary biology*, 17(1):162, 2017.

[79] Nobuyoshi Shimoda, Kimi Yamakoshi, Akimitsu Miyake, and Hiroyuki Takeda. Identification of a gene required for de novo dna methylation of the zebrafish no tail gene. *Developmental dynamics: an official publication of the American Association of Anatomists*, 233(4):1509–1516, 2005.

[80] Catarina Campos, Luisa MP Valente, and Jorge MO Fernandes. Molecular evolution of zebrafish dnmt3 genes and thermal plasticity of their expression during embryonic development. *Gene*, 500(1):93–100, 2012.

[81] Jingwei Liu, Huihua Hu, Stéphane Panserat, and Lucie Marandel. Evolutionary history of dna methylation related genes in chordates: new insights from multiple whole genome duplications. *Scientific reports*, 10(1):1–14, 2020.

[82] Sigbjørn Lien, Ben F Koop, Simen R Sandve, Jason R Miller, Matthew P Kent, Torfinn Nome, Torgeir R Hvidsten, Jong S Leong, David R Minkley, Aleksey Zimin, et al. The atlantic salmon genome provides insights into rediploidization. *Nature*, 533(7602):200–205, 2016.

[83] Verena Hurst, Kenji Shimada, and Susan M Gasser. Nuclear actin and actin-binding proteins in dna repair. *Trends in cell biology*, 29(6):462–476, 2019.

[84] Kei Miyamoto and JB Gurdon. Transcriptional regulation and nuclear reprogramming: roles of nuclear actin and actin-binding proteins. *Cellular and Molecular Life Sciences*, 70(18):3289–3302, 2013.

[85] Shruti Rastogi and David A Liberles. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC evolutionary biology*, 5(1):28, 2005.

[86] Manfred D Laubichler, Peter F Stadler, Sonja J Prohaska, and Katja Nowick. The relativity of biological function. *Theory in Biosciences*, 134(3-4):143–147, 2015.

[87] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

[88] Felix Krueger. Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*, 516:517, 2015.

[89] Helene Kretzmer, Christian Otto, and Steve Hoffmann. Bat: Bisulfite analysis toolkit. *F1000Research*, 6(1490):1490, 2017.

[90] Christian Otto, Peter F Stadler, and Steve Hoffmann. Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinformatics*, 28(13):1698–1704, 2012.

[91] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M Sharma, Philipp Khaitovich, Jörg Vogel, Peter F Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9):e1000502, 2009.

[92] Frank Jühling, Helene Kretzmer, Stephan H Bernhart, Christian Otto, Peter F Stadler, and Steve Hoffmann. metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome research*, 26(2):256–262, 2016.

[93] Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. Great improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5):495–501, 2010.

[94] Yu-Heng Lai, Gilbert Audira, Sung-Tzu Liang, Petrus Siregar, Michael Edbert Suryanto, Huan-Chau Lin, Omar Villalobos, Oliver B Villaflores, Erwei Hao, Ken-Hong Lim, et al. Duplicated dnmt3aa and dnmt3ab dna methyltransferase genes play essential and non-overlapped functions on modulating behavioral control in zebrafish. *Genes*, 11(11):1322, 2020.

[95] Alexandra Giatromanolaki, Michael I Koukourakis, Heidi M Sowter, Efthimios Sivridis, Spencer Gibson, Kevin C Gatter, and Adrian L Harris. Bnip3 expression is linked with hypoxia-regulated protein expression and with poor prognosis in non–small cell lung cancer. *Clinical Cancer Research*, 10(16):5566–5571, 2004.

[96] Jiro Okami, Diane M Simeone, and Craig D Logsdon. Silencing of the hypoxia-inducible cell death protein bnip3 in pancreatic cancer. *Cancer research*, 64(15):5338–5346, 2004.

[97] Hidekazu Nagano, Naoko Hashimoto, Akitoshi Nakayama, Sawako Suzuki, Yui Miyabayashi, Azusa Yamato, Seiichiro Higuchi, Masanori Fujimoto, Ikki Sakuma, Minako Beppu, et al. p53-inducible dpysl4 associates with mitochondrial supercomplexes and regulates energy metabolism in adipocytes and cancer cells. *Proceedings of the National Academy of Sciences*, 115(33):8370–8375, 2018.

[98] Marcel Tiebe, Marilena Lutz, Deniz Senyilmaz Tiebe, and Aurelio A Teleman. Crebl2 regulates cell metabolism in muscle and liver cells. *Scientific reports*, 9(1):1–12, 2019.

[99] Angad Rao and Deron R Herr. G protein-coupled receptor gpr19 regulates e-cadherin expression and invasion of breast cancer cells. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1864(7):1318–1327, 2017.

[100] Wu Zhang and Jie Xu. Dna methyltransferases and their roles in tumorigenesis. *Biomarker research*, 5(1):1–8, 2017.

[101] Daisuke Watanabe, Isao Suetake, Takashi Tada, and Shoji Tajima. Stage-and cell-specific expression of dnmt3a and dnmt3b during embryogenesis. *Mechanisms of development*, 118(1-2):187–190, 2002.

[102] Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[103] Tamara HL Smith, Terry Mark Collins, and Ross A McGowan. Expression of the dnmt3 genes in zebrafish development: similarity to dnmt3a and dnmt3b. *Development genes and evolution*, 220(11-12):347–353, 2011.

[104] David C Klein and Sarah J Hainer. Genomic methods in profiling dna accessibility and factor localization. *Chromosome Research*, 28(1):69–85, 2020.

[105] Michael Dukatz, Katharina Holzer, Michel Choudalakis, Max Emperle, Cristiana Lungu, Pavel Bashtrykov, and Albert Jeltsch. H3k36me2/3 binding and dna binding of the dna methyltransferase dnmt3a pwwp domain both contribute to its chromatin interaction. *Journal of Molecular Biology*, 431(24):5063–5074, 2019.

[106] Yelena Chernyavskaya, Brandon Kent, and Kirsten C Sadler. Zebrafish discoveries in cancer epigenetics. In *Cancer and Zebrafish*, pages 169–197. Springer, 2016.

[107] Neelakanteswar Aluru. Epigenetic effects of environmental chemicals: Insights from zebrafish. *Current opinion in toxicology*, 6:26–33, 2017.

[108] Stephen Safe. Molecular biology of the ah receptor and its role in carcinogenesis. *Toxicology letters*, 120(1-3):1–7, 2001.

[109] Sonia Mulero-Navarro and Pedro M Fernandez-Salguero. New trends in aryl hydrocarbon receptor biology. *Frontiers in cell and developmental biology*, 4:45, 2016.

[110] Claudia Consales, Gunnar Toft, Giorgio Leter, Jens Peter E Bonde, Raffaella Uccelli, Francesca Pacchierotti, Patrizia Eleuteri, Bo AG Jönsson, Aleksander Giwercman, Henning S Pedersen, et al. Exposure to persistent organic pollutants and sperm dna methylation changes in arctic and european populations. *Environmental and Molecular Mutagenesis*, 57(3):200–209, 2016.

[111] Hiroaki Itoh, Motoki Iwasaki, Yoshio Kasuga, Shiro Yokoyama, Hiroshi Onuma, Hideki Nishimura, Ritsu Kusama, Teruhiko Yoshida, Kazuhito Yokoyama, and Shoichiro Tsugane. Association between serum organochlorines and global methylation level of leukocyte dna among japanese women: a cross-sectional study. *Science of the total environment*, 490:603–609, 2014.

[112] Mi Hwa Lee, Eo Rin Cho, Jung-eun Lim, and Sun Ha Jee. Association between serum persistent organic pollutants and dna methylation in korean adults. *Environmental research*, 158:333–341, 2017.

[113] Lars Lind, Johanna Penell, Karin Luttropp, Louise Nordfors, Anne-Christine Syvänen, Tomas Axelsson, Samira Salihovic, Bert van Bavel, Tove Fall, Erik Ingelsson, et al. Global dna hypermethylation is associated with high serum levels of persistent organic pollutants in an elderly population. *Environment international*, 59:456–461, 2013.

[114] Neelakanteswar Aluru, Sibel I Karchner, Keegan S Krick, Wei Zhu, and Jiang Liu. Role of dna methylation in altered gene expression patterns in adult zebrafish (danio rerio) exposed to 3, 3', 4, 4', 5-pentachlorobiphenyl (pcb 126). *Environmental epigenetics*, 4(1):dvy005, 2018.

[115] Neelakanteswar Aluru, Sibel I Karchner, and Lilah Glazer. Early life exposure to low levels of ahr agonist pcb126 (3, 3', 4, 4', 5-pentachlorobiphenyl) reprograms gene expression in adult brain. *Toxicological Sciences*, 160(2):386–397, 2017.

[116] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[117] Alexander Dobin and Thomas R Gingeras. Mapping rna-seq reads with star. *Current protocols in bioinformatics*, 51(1):11–14, 2015.

[118] Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the deseq2 package. *Genome Biol*, 15(550):10–1186, 2014.

[119] Brad T Sherman, Richard A Lempicki, et al. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44, 2009.

[120] Allegra Angeloni and Ozren Bogdanovic. Enhancer dna methylation: implications for gene regulation. *Essays in biochemistry*, 63(6):707–715, 2019.

[121] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[122] KJ Teerds, DG De Rooij, FFG Rommerts, I Van Der Tweel, and CJG Wensing. Turnover time of leydig cells and other interstitial cells in testes of adult rats. *Archives of andrology*, 23(2):105–111, 1989.

[123] Jorke H Kamstra, Marianne Løken, Peter Aleström, and Juliette Legler. Dynamics of dna hydroxymethylation in zebrafish. *Zebrafish*, 12(3):230–237, 2015.

[124] Yen Ching Lim, Jie Li, Yiyun Ni, Qi Liang, Junjiao Zhang, George SH Yeo, Jianxin Lyu, Shengnan Jin, and Chunming Ding. A complex association between dna methylation and gene expression in human placenta at first and third trimesters. *PLoS One*, 12(7):e0181155, 2017.

[125] Meike Gölzenleuchter, Rahul Kanwar, Manal Zaibak, Fadi Al Saiegh, Theresa Hartung, Jana Klukas, Regenia L Smalley, Julie M Cunningham, Maria E Figueroa, Gary P Schroth, et al. Plasticity of dna methylation in a nerve injury model of pain. *Epigenetics*, 10(3):200–212, 2015.

[126] Eric L Fritz, Brad R Rosenberg, Kenneth Lay, Aleksandra Mihailović, Thomas Tuschl, and F Nina Papavasiliou. A comprehensive analysis of the effects of the deaminase aid on the transcriptome and methylome of activated b cells. *Nature immunology*, 14(7):749–755, 2013.

[127] Hyung Joo Lee, Yiran Hou, Yujie Chen, Zea Z Dailey, Aiyana Riddihough, Hyo Sik Jang, Ting Wang, and Stephen L Johnson. Regenerating zebrafish fin epigenome is characterized by stable lineage-specific dna methylation and dynamic chromatin accessibility. *Genome biology*, 21(1):1–17, 2020.

[128] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäss aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

(Ort, Datum)

(Unterschrift)