Cumulative dissertation submitted in partial fulfillment of the

of the requirements for the doctoral degree

(Dr. rer. nat.)

Evolution of DNA Methylation in Insects

by

Panagiotis Provataris

from Cholargos, Greece

Bonn, August, 2020

Faculty of Mathematics and Natural Sciences

Rheinische Friedrich-Wilhelms University of Bonn

Carried out at the Zoological Research Museum A. Koenig, Bonn

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen

Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

Erstgutachter: Prof. Dr. Bernhard Misof

Zweitgutachter: Prof. Dr. Frank Lyko

Fachnahes Kommissionsmitglied: Prof. Dr. Martin Sander

Fachfremdes Kommissionsmitglied: Prof. Dr. Gabrielle Koenig

Tag der Promotion: 31 März 2021

Jahr der Veröffentlichung: 2021

To my father, who could not wait.

Summary

DNA methylation, the addition of a methyl group (-CH3) to a genomic cytosine by a DNA methyltransferase, is the most thoroughly studied epigenetic modification across eukaryotes. Functionally, it is associated with the regulation of gene expression and the silencing of transposable elements. Most of our knowledge about DNA methylation comes from model organisms that taxonomically belong to vertebrate animals and plants. In invertebrate animals, the most well-studied group is insects. Studied insects possess sparsely methylated genomes in comparison to vertebrate and other invertebrate animals, an observation that has led to the assumption that DNA methylation has been ancestrally reduced in insects. However, DNA methylation has only been studied in a taxonomically restricted group of insect species. Here I present a comprehensive comparative study of DNA methylation that covers all major extant insect groups. By performing computational and experimental comparative analyses I discovered that DNA methylation was reduced in the last common ancestor of holometabolous insects, while hemimetabolous insects show similar patterns of DNA methylation with other invertebrate animals. Additionally, I identified that DNA methyltransferase 1 is necessary and can be sufficient for the insect DNA methylation machinery to remain functional. By utilizing a novel approach for the classification of DNA methylation patterns in insects, I propose that transposable elements have played a major role in the evolutionary transitions among insect DNA methylation patterns. Furthermore, by creating a mechanistic model to explain the evolutionary trajectories of DNA methylation in insects, I recognized that DNA methylation has

followed opposite evolutionary trajectories in insects and vertebrates that are shaped in both groups by the interplay between transposable elements and DNA methylation. I anticipate that this work will contribute towards understanding how the molecular biodiversity of insects, the most speciose group of animals, emerged.

Contents

List of Figures and Tables

Figures

Tables

Chapter 1

General Introduction

Eukaryotic genomes possess heritable chemical modifications that can be found either on DNA, or on chromatin proteins. These modifications are not ubiquitously placed across a genome, but tend to be concentrated in certain regions where they can modify the properties of affected elements, for example alter the expression profiles of genes or restrict the activity of transposable elements (Suzuki and Bird 2008; Schuebeler 2015). Most intriguing is the fact that the underlying DNA sequence is not altered by these modifications. The sum of all such modifications of a genome is called the epigenome[1] and its study is called epigenomics. Thus, to comprehend the impact that the epigenome has on the genome one needs to first create a map of the epigenome and subsequently characterize the effect that epigenetic modifications have on different types of genomic elements. However, epigenomes may considerably vary among species, among individuals belonging to a species, among cell types, or among cells of a specific cell type during the lifetime of an organism (Xiao et al. 2014). This is why creating maps that describe single epigenomic states can only provide limited information on the biological impact of epigenomic phenomena, especially from an evolutionary perspective. The answer to this problem is an emerging field called *comparative epigenomics*, which utilizes big data science approaches to compare a plethora of epigenomic states in order to understand how epigenomic change affects or is affected by other biological or environmental processes. In this thesis, I utilize comparative epigenomics to study the most well-known epigenetic modification, DNA methylation, from an evolutionary perspective.

1 The ancient Greek adverb epi (*επί*) which is added as a prefix to the word genome and literally translates to “over” or “above” signifies features that are “on top of” or “in addition to” the traditional genome-based inheritance.

DNA methylation is the addition of a methyl group (-CH3) to specific nucleobases, most commonly cytosines, by a DNA methyltransferase (DNMT). In animal genomes, DNA methylation is predominantly found at CG sites, while in plant and fungal genomes it additionally occurs in CHG and CHH context (H stands for A, C, or T; Suzuki and Bird 2008). Currently, the most comprehensive method for characterizing genomic patterns of DNA methylation is Whole Genome Bisulfite Sequencing (WGBS) (Lister and Ecker 2009). WGBS is similar to whole genome sequencing, except for bisulfite conversion. In brief, DNA is treated with sodium bisulfite which converts only non-methylated cytosines to uracils. Due to the following PCR amplification, non-methylated cytosines eventually appear as thymines, while methylated cytosines are unchanged. Therefore, the methylation state of the original DNA sequence can be inferred by comparing it to bisulfite-treated DNA, with single-base resolution (Clark et al. 1994). The application of WGBS on a comparative scale has provided invaluable insights towards understanding the evolutionary patterns of DNA methylation across eukaryotes, summarized in the following two points:

i) The function of DNA methylation varies depending on its location in the genome.

In well-studied mammalian systems, methylation of promoter regions is typically linked to the transcriptional repression of downstream genes, although this effect is not observable in all studied instances (Bestor et al. 2014; Lou et al. 2014; Schübeler 2015). Methylation of transposable and other repetitive elements is also repressive and is thought to limit the expression and, consequently, the genomic expansion of

these elements (Yoder et al. 1997; Schübeler 2015). In contrast, DNA methylation located in exons and introns (termed gene body methylation) is associated with actively transcribed genes (Feng et al. 2010; Zemach et al. 2010). However, the manner in which gene body methylation interacts with active transcription is still poorly understood (Schübeler 2015). One long-standing hypothesis posits that intragenic methylation functions to reduce transcriptional noise (Bird 1995; Suzuki et al. 2007), and has recently received support by multiple studies (Neri et al. 2017; Gatzmann et al. 2018; Li et al. 2018). Thus, to comprehend the functions of DNA methylation, its distribution across the genome needs to be considered (Jones 2012).

ii) The genomic patterns of DNA methylation are phylogenetically highly variable.

In vertebrates, and particularly in the well-studied mammals, DNA methylation occurs throughout the genome (global methylation pattern), with gene bodies, intergenic regions, and repetitive elements being rather consistently methylated (although gene body methylation levels tend to be higher compared to flanking regions) (Suzuki and Bird 2008). Short, CG-rich regions (termed CpG islands) that typically overlap with promoters constitute an exception (Schultz et al. 2015; Mendizabal and Yi 2016). The global pattern of DNA methylation is thought to have evolved early during vertebrate evolution (Tweedie et al. 1997). In invertebrate animals and plants, DNA methylation typically occurs in subsets of genomic elements (mosaic methylation pattern) (Suzuki and Bird 2008; Zemach et al. 2010). Gene bodies are the prime targets of DNA methylation in invertebrate genomes, but the proportion of highly methylated gene bodies significantly varies across different taxonomic groups (Feng et al. 2010;

Zemach et al. 2010). Apart from gene bodies, DNA methylation is specifically targeted to transposable elements (TEs) in plant species and is correlated with TE silencing (Slotkin and Martienssen 2007). Despite its obvious essential functions in many studied organisms, DNA methylation has also been lost or significantly reduced in various invertebrate and plant species (Raddatz et al. 2013; Niederhuth et al. 2016; Takuno et al. 2016). In conclusion, the prevailing idea is that gene body methylation is considered an ancestral property of eukaryotes, while methylation of TEs is thought to have evolved independently in plants and vertebrates (Suzuki and Bird 2008; Zemach et al. 2010). However, the recent identification of substantial TE methylation in a number of invertebrate species (Gao et al. 2012; Falckenhayn et al. 2013; Wang et al. 2014; Kao et al. 2016; Rosic et al. 2018) suggests that preferential methylation of TEs may have evolved multiple times during metazoan evolution (Yi 2012).

Due to its evolutionary lability, general conclusions on the patterns and functions of DNA methylation across clades can only be substantiated in the presence of extensive comparative DNA methylation data. Since the accurate characterization of DNA methylation patterns via WGBS requires the presence of reference genomes, a dense and phylogenetically representative sampling of genomes is a prerequisite to comparatively study DNA methylation within a group of interest. Furthermore, to make associations between the patterns and the functions of DNA methylation in studied genomes, accurate annotations of genomic elements (i.e. protein-coding genes, repetitive elements) are indispensable. Finally, in order to accurately infer the evolution of DNA methylation, not only representative sets of sequenced genomes

and their annotations are required, but also statistically robust inferences of species relationships within the group of interest.

As of today, over 138 insect genome assemblies have been published[2] (according to InsectBase; Yin et al. 2016) and recently a dense sampling of phylogenetically diverse insect genomes became accessible via the i5K initiative, which aims to sequence more than 5,000 arthropod genomes (i5K consortium 2013). Through this initiative not only genome assemblies that can be used as reference for WGBS studies are provided, but also high quality annotations of protein-coding genes have been generated for many non-model insects (Poelchau et al. 2015; Thomas et al. 2020). Additionally, an insect-specific workflow for the annotation of repetitive elements have been recently released (Petersen et al. 2019). Finally, Misof and colleagues through the 1KITE initiative (http://www.1kite.org) have published the most comprehensive phylogenetic reconstruction of insects to date (Misof et al. 2014; Kjer et al. 2015), which was followed by numerous publications that attempted to resolve phylogenetic relationships within major insect groups (Peters et al 2017; Johnson et al. 2018; McKenna et al. 2019; Kawahara et al. 2019; Wipfler et al. 2019; Vasilikopoulos et al. 2020). Thus, the necessary prerequisites to comparatively study the evolution of DNA methylation in insects have become available.

1.1 Evolutionary Patterns of DNA Methylation in Insects

Although the presence of DNA methylation in insects was first documented more than 25 years ago (Sarkar et al. 1992), it was not until recently that insects emerged as a

2 http://www.insect-genome.com/ -Last accessed August 12, 2020.

model group for studying DNA methylation (Lyko and Maleszka 2011). The key finding that nutritionally regulated levels of DNA methylation contribute to the development of alternative castes in the honeybee, *Apis mellifera* (Kucharski et al. 2008), led to the hypothesis posing that DNA methylation may be strongly associated with phenotypic plasticity in insects (Lyko and Maleszka 2011; Bonasio 2014). This hypothesis induced a significant number of studies exploring the suggested association, especially in social insects (Elango et al. 2009; Lyko et al. 2010; Bonasio et al. 2012; Foret et al. 2012; Patalano et al. 2015; Glastad et al. 2016; Libbrecht et al. 2016; Standage et al. 2016; Bewick et al 2017; Glastad et al. 2017). Nevertheless, significant advances have only been made on a restricted taxonomic range.

Levels of DNA Methylation in Insects

Currently, our knowledge regarding DNA methylation in insects is largely derived from species belonging to insects that go through distinctive larval, pupal, and adult stages, called Holometabola[3]. DNA methylation has been found at appreciable levels in species belonging to Hymenoptera (sawflies, wasps, ants, and bees), Lepidoptera (butterflies and moths), and Coleoptera (beetles) (Lyko et al. 2010; Xiang et al. 2010; Bonasio et al. 2012; Hunt et al. 2013; Wang et al. 2013; Cunningham et al. 2015; Patalano et al. 2015; Libbrecht et al. 2016; Rehan et al. 2016; Glastad et al. 2017), where it is sparsely targeted across genes, primarily found in exons (Glastad et al.

3 Holometabola comprise approximately 80% of insect and 60% of animal species that have been described. This is mainly due to the four megadiverse groups: sawflies, wasps, bees, and ants (Hymenoptera), beetles (Coleoptera), butterflies and moths (Lepidoptera), and true flies (Diptera) (McMahon and Hayword 2016; Stork 2018).

2014). In Diptera (flies and allies), DNA methylation is found at extremely low levels (Lyko et al. 2000; Boffelli et al. 2014; Bewick et al. 2016; Falckenhayn et al. 2016), lacking defined patterning across the genome (Raddatz et al. 2013; Falckenhayn et al. 2016; Bewick et al. 2016). Additionally, DNA methylation has experienced extreme reductions or may be even lost in some species within the orders of Hymenoptera and Coleoptera (Zemach et al. 2010; Standage et al. 2016; Schulz et al. 2018). Therefore, lineage-specific reductions on the levels of DNA methylation seem common among holometabolous insects.

Comparisons between certain holometabolous insect species and other animals suggested that insects possess substantially lower levels of DNA methylation compared to both vertebrate and other invertebrate animals (Feng et al. 2010; Zemach et al. 2010; Sarda et al. 2012). Such comparisons have corroborated the hypothesis that DNA methylation was reduced in the last common ancestor of insects, eventually leading to its loss (or extreme reduction) in certain lineages (Glastad et al. 2014). However, evidence from hemimetabolous insects[4] does not seem to further support this hypothesis. Single-species studies using experimental measurements of DNA methylation in locusts (Orthoptera), a stick insect (Phasmatodea), and termites and cockroaches (Blattodea) have shown that these species possess significantly elevated levels of DNA methylation compared to studied holometabolous species (Krauss et al. 2009; Falckenhayn et al. 2013; Wang et al. 2014; Glastad et al. 2016; Bewick et al. 2017). For example, DNA methylation is estimated to be found in 1.3% to 1.6% of all

4 Insects that, in contrast to holometabolous ones, do not go through a pupal stage during their development. They develop from nymph to imago and usually the nymph already resembles the adult form.

genomic cytosines in the two locust species, whereas the corresponding values for species of ants, the honey bee, and the silk moth are 0.1% or lower (Falckenhayn et al. 2013; Wang et al. 2014). Furthermore, 12% of CG sites are methylated in the genome of *Zootermopsis nevadensis* (Glastad et al. 2016), whereas 1–2% of genomic CpGs are methylated in holometabolous social insects (Glastad et al. 2011). Thus, assumptions based on phylogenetically restricted evidence are not reliable enough to infer the evolution of DNA methylation in insects and comparative epigenomics approaches are required to improve our understanding on the patterns of DNA methylation across insects.

The Distribution of DNA Methylation within Insect Genomes

In species belonging to Hymenoptera, Coleoptera and Lepidoptera, DNA methylation is typically confined to a subset of gene bodies (exons and introns). Exons are the prime targets of DNA methylation within genes and typically show much higher levels of DNA methylation compared to introns (Glastad et al. 2014). In particular, DNA methylation is preferentially found in exons located near the 5' end of genes and tends to drop towards the 3' end, with a significant decrease outside gene boundaries (Zemach et al. 2010; Bonasio et al. 2012; Hunt et al. 2013; Wang et al. 2013; Glastad et al. 2014). Thus, not all exons along a gene body exhibit strong methylation. TEs are not preferentially methylated even in transposon-rich species, like *B. mori* (Xiang et al. 2010; Zemach et al. 2010; Glastad et al. 2014; Glastad et al 2019). The described patterns of DNA methylation in holometabolous insects starkly contrast with what is observed in (i) vertebrates (mostly mammals), where the majority of the genome is

heavily methylated and (ii) in plants where methylation apart from gene bodies, is heavily targeted to TEs and contributes to their transcriptional repression via a unique RNAi–directed mechanism that is not present in animals (Suzuki and Bird 2008).

Sparse evidence on the distribution of DNA methylation in hemimetabolous insect genomes exists, but notable differences on the patterns of DNA methylation compared to Holometabola are already visible. The comprehensively analyzed methylome of *Zootermopsis nevadensis* revealed a number of them: 1) A significantly larger subset of genes was highly methylated in the termite in comparison to the ants *Apis mellifera* and *Camponotus floridanus* (75% to ~35% in both Hymenoptera) 2) Apart from exons, introns exhibit considerable methylation, with only slightly lower methylation levels compared to exons. 3) Within genes, DNA methylation levels increase towards the 3' end of the gene and do not drop outside the 3' gene boundary. 4) TEs are substantially methylated, but only inside gene bodies. Since gene bodies are the primary targets of DNA methylation in *Z. nevadensis*, the lack of TE methylation outside genes does not clearly point to preferential targeting of these elements (Glastad et al. 2016).

Additional Putative Targets of DNA Methylation in Insects

Preliminary evidence from three polyneopteran species (two locusts and a stick insect) suggests diverse patterns of DNA methylation. In all three species, genes, TEs and other repeats are in general methylated (Krauss et al. 2009; Falckenhayn et al. 2013; Wang et al. 2014). Particularly, in the migratory locust, *Locusta migratoria*, introns (which tend to be TE-rich) were reported as the prime targets of DNA

methylation within the genome. Additionally, TEs and other repeats were found to be more highly methylated compared to exons (Wang et al. 2014). However, due to the lack of reference genomes (Krauss et al. 2009; Falckenhayn et al. 2013) or due to the poor assembly and annotation of the *L. migratoria* genome, the extent of TE methylation or the position of methylated TE sequences with respect to genes in these species remains unknown (for example, we do not know if repeat-poor introns are also highly methylated, or if repeat sequences outside genes are also highly methylated in the *L. migratoria* genome and thus, inferring the patterns of repeat methylation is not possible). Likewise, a low coverage methylome of a crustacean, *Parhyale hawaiensis*, which shows high TE abundance similar to *L. migratoria*, suggested that various types of TEs are highly methylated (Kao et al. 2016). These reports hint towards the presence of preferential TE methylation in hemimetabolous insects, with a potential role for regulation similar to plants or vertebrates. However, more systematic and more detailed studies need to be conducted in order to comprehensively investigate the interaction between DNA methylation and repetitive DNA sequences in insects.

1.2 The DNA Methyltransferase Toolkit of Insects

In mammals, DNA methyltransferase 3 (DNMT3) has been considered responsible for the establishment of DNA methylation patterns *de novo* and DNA methyltransferase 1 (DNMT1) for maintaining methylation patterns across cell generations. However, recent evidence supports that strictly separating the DNMTs between *de novo* and maintenance enzymes is no longer valid, as their functional roles overlap (Jeltsch and

Jurkowska 2014; Maleszka 2016). These enzymes are thought to possess similar functions in insects (Wang et al. 2006), while their absence is associated with negligible methylation levels and a lack of detectable methylation patterning (Raddatz et al. 2013; Falckenhayn et al. 2016). However, it seems that methylation systems of certain insects may be able to remain functional in absence of DNMT3. For example, in the genome of the silkmoth, *Bombyx mori*, low but not insignificant levels of DNA methylation are mediated by a single copy of DNMT1 (Xiang et al. 2010; Zemach et al. 2010). However due to a lack of comparative studies prior to the initiation of this thesis, the distribution of DNMTs in insects, and thus, an accurate characterization of the insect DNMT toolkit have remained elusive[5].

1.3 Research Questions

Comparative studies on the evolution of DNA methylation in insects have been scarce. Additionally, single-species investigations suffered from various obstacles, such as lack of available reference genomes or poor assembly and annotation of sequenced genomes. These factors hampered accurate descriptions of the distribution and the levels of DNA methylation of insect genomes, particularly of hemimetabolous insects. The increasing availability of transcriptomes and reference genomes covering a phylogenetically diverse groups of hemimetabolous and holometabolous insect species, combined with the decreasing cost of WGBS allowed for an exhaustive

5 Upon the initiation of this thesis two comparative studies that, among other things, characterized the distribution of DNA methyltransferases across insects were published by Bewick and colleagues in 2017, and myself and colleagues in 2018. The latter is presented in the third chapter of this thesis.

comparative investigation of DNA methylation across insects, which will be presented in this thesis.

In the second chapter, I used computational methods to describe the patterns of DNA methylation and characterize the DNMT toolkit of two basal Hymenoptera: the primarily phytophagous sawfly *Athalia rosae*, and the parasitoid sawfly *Orussus abietinus.* Although Hymenoptera are the most-well studied insect taxon regarding DNA methylation, most studies have focused on eusocial taxa like bees, ants, and paper wasps (Elango et al. 2009; Lyko et al. 2010; Bonasio et al. 2012; Patalano et al. 2015; Libbrecht et al. 2016). My indirect inference on the levels of DNA methylation of the two sawflies demonstrates that basal, asocial Hymenoptera show signatures of substantial DNA methylation levels not only in exons, but also in introns, a trait that is mostly absent from the eusocial species of the group. Additionally, I identified a rare duplication of DNMT3 in *Athalia rosae*, the only duplication of the enzyme described in this group. In conclusion, I suggest that some basal Hymenoptera may display different patterns of DNA methylation and possess a differentiated DNMT toolkit compared with aculeate[6] species.

In the third chapter, I use a combination of computational and experimental evidence to infer the presence and estimate the levels of DNA methyation in protein-coding sequences across insects by exploiting publicly available genomic data and transcriptomic data. Based on my findings I propose that DNA methylation levels

6 Aculeata or commonly referred to as “stinging wasps” is a subgroup of Hymenoptera that includes jewel wasps, vespid wasps, ants, and bees among other taxa. The defining trait of aculeates is the modification of the ovipositor, from an egg-laying structure into a stinger, used to inject venom to threats or prey. However, not all aculeates have stingers (e.g. ants or parasitoid wasps; Definition inspired by Peters et al. 2017).

were reduced in the stem lineage leading to Holometabola, contradicting previous hypotheses posing that DNA methylation levels were reduced in the last common ancestor of insects (Glastad et al. 2014). Additionally, I developed a computational workflow that allowed for an accurate characterization of the DNMT toolkit of most major insect groups and suggest that in insects, in contrast to vertebrates, DNMT1 enzymes are necessary and sufficient for the establishment and maintenance of genomic methylation.

The fourth chapter is the very essence of this thesis. In contrast to the previous chapters, which are mostly limited to inferences on the patterns of DNA methylation due to a lack of experimental data, the availability of an extensive and taxonomically representative DNA methylation dataset allowed for an accurate characterization of the genomic patterns of DNA methylation across insects. I found that in many hemimetabolous insects not only exons, but introns, and repetitive DNA sequences are targeted by DNA methylation. Ancestral state reconstructions showed that the last common ancestor of insects shared all of the aforementioned traits, whereas in the last common ancestor of holometabolous insects DNA methylation became confined to the exons of certain genes. Additionally, by classifying DNA methylation patterns into four main groups I was able to create a model that identified TE activity as the major force for the evolutionary transition from a moderately methylated ancestral insect genome to a sparsely methylated holometabolous genome. This finding has important implications on the evolvability of holometabolous genomes and, in consequence, to the molecular mechanisms underlying the extreme biodiversity of Holometabola.

1.4 References

Bewick AJ, Vogel KJ, Moore AJ, & Schmitz RJ. 2017. Evolution of DNA methylation across insects. *Molecular Biology and Evolution*, *34*(3), 654–665.

Bird AP. 1995. Gene number, noise reduction and biological complexity. *Trends in Genetics*, *11*(3), 94–100.

Boffelli D, Takayama S, & Martin DIK. 2014. Now you see it: Genome methylation makes a comeback in *Drosophila*. *BioEssays*, *36*(12), 1138–1144.

Bonasio R. 2014. The role of chromatin and epigenetics in the polyphenisms of ant castes. *Briefings in Functional Genomics*, *13*(3), 235–245.

Bonasio R, et al. 2012. Genome-wide and caste-specific DNA methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Current Biology*, *22*(19), 1755–1764.

De Mendoza A, et al. 2018. Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons. *Nature Communications*, *9*(1), 1–11.

Elango N, Hunt BG, Goodisman MaD, & Yi SV. 2009. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(27), 11206–11211.

Falckenhayn C, et al. 2013. Characterization of genome methylation patterns in the desert locust Schistocerca gregaria. *Journal of Experimental Biology*, *216*(8), 1423–1429.

Falckenhayn C, et al. 2016. Comprehensive DNA methylation analysis of the Aedes aegypti genome. *Scientific Reports*, *6*, 1–8.

Feng S, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, *107*(19), 8689–8694.

Foret S, et al. 2012. DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. *Proceedings of the National Academy of Sciences*, *109*(13), 4968–4973.

Gao F, et al. 2012. Differential DNA methylation in discrete developmental stages of the parasitic nematode Trichinella spiralis. *Genome Biology*, *13*(10), R100.

Gatzmann F, et al. 2018. The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. *Epigenetics & Chromatin*, *11*(1), 57.

Glastad KM, Hunt BG, Yi SV, & Goodisman MAD. 2011. DNA methylation in insects: On the brink of the epigenomic era. *Insect Molecular Biology*, *20*(5), 553–565.

Glastad KM, et al. 2017. Variation in DNA Methylation Is Not Consistently Reflected by Sociality in Hymenoptera. *Genome Biology and Evolution*, *9*(6), 1687–1698.

Glastad KM, Gokhale K, Liebig J, & Goodisman MAD. 2016. The caste- and sex-specific DNA methylome of the termite Zootermopsis nevadensis. *Scientific Reports*, *6*(October), 1–14.

Glastad KM, Hunt BG, & Goodisman MA. 2014. Evolutionary insights into DNA methylation in insects. *Current Opinion in Insect Science*, *1*, 25–30.

Glastad KM, Hunt BG, & Goodisman MAD. 2019. Epigenetics in insects: Genome regulation and the generation of phenotypic diversity. *Annual Review of Entomology*, *64*, 185–203.

Jeltsch A, & Jurkowska RZ. 2014. New concepts in DNA methylation. *Trends in Biochemical Sciences*, *39*(7), 310–318.

Johnson KP, et al. 2018. Phylogenomics and the evolution of hemipteroid insects. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(50), 12775–12780.

Jones PA. 2012. Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, *13*(7), 484–492.

Kao D, et al. 2016. The genome of the crustacean *Parhyale hawaiensis*, a model for animal development, regeneration, immunity and lignocellulose digestion. *ELife*, *5*, 1–45.

Kawahara AY, et al. 2019. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(45), 22657–22663.

Kjer KM, et al. 2015. Response to Comment on “phylogenomics resolves the timing and pattern of insect evolution.” *Science*, *349*(6247), 487.

Krauss V, Eisenhardt C, & Unger T. 2009. The genome of the stick insect Medauroidea extradentata is strongly methylated within genes and repetitive DNA. *PLoS ONE*, *4*(9).

Li Y, et al. 2018. DNA methylation regulates transcriptional homeostasis of algal endosymbiosis in the coral model Aiptasia. *Science Advances*, *4*(8).

Libbrecht R, Oxley PR, Keller L, & Kronauer DJC. 2016. Robust DNA methylation in the clonal raider ant brain. *Current Biology*, *26*(3), 391–395.

Lister R, & Ecker JR. 2009. Finding the fifth base: Genome-wide sequencing of cytosine methylation. *Genome Research*, *19*(6), 959–966.

Lyko F, Ramsahoye BH, & Jaenisch R. 2000. DNA methylation in Drosophila melanogaster. *Nature*, *408*(November), 538–540.

Lyko F, et al. 2010. The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. *PLoS Biology*, *8*(11).

Lyko F, & Maleszka R. 2011. Insects as innovative models for functional studies of DNA methylation. *Trends in Genetics*, *27*(4), 127–131.

Macleod D, Clark VH, & Bird A. 1999. Absence of genome-wide changes in DNA methylation during development of the zebrafish [1]. *Nature Genetics*, *23*(2), 139–140.

Maleszka R. 2016. Epigenetic code and insect behavioural plasticity. *Current Opinion in Insect Science*, *15*, 45–52.

McKenna DD, et al. 2019. The evolution and genomic basis of beetle diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(49), 24729–24737.

Mendizabal I, & Yi SV. 2016. Whole-genome bisulfite sequencing maps from multiple human tissues reveal novel CpG islands associated with tissue-specific regulation. *Human Molecular Genetics*, *25*(1), 69–82.

Misof B, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, *346*(6210), 763–767.

Neri F, et al. 2017. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, *543*(7643), 72–77.

Niederhuth CE, et al. 2016. Widespread natural variation of DNA methylation within angiosperms. *Genome Biology*, *17*(1), 1–19.

Patalano S, et al. 2015. Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. *Proceedings of the National Academy of Sciences*, *112*(45), 13970–13975.

Peters RS, et al. 2017. Evolutionary History of the Hymenoptera. *Current Biology*, *27*(7), 1013–1018.

Petersen M, et al. 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evolutionary Biology*, *19*(1), 1–15.

Poelchau M, et al. 2015. The i5k Workspace@NAL-enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Research*, *43*(D1), D714–D719.

Raddatz G, et al. 2013. Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proceedings of the National Academy of Sciences*, *110*(21), 8627–8631.

Rošić S, et al. 2018. Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity. *Nature Genetics*.

Sarda S, Zeng J, Hunt BG, & Yi SV. 2012. The evolution of invertebrate gene body methylation. *Molecular Biology and Evolution*, *29*(8), 1907–1916.

Sarkar S, Rao SRV, Gupta VS, & Hendre RR. 1992. 5-Methylcytosine content in Gryllotalpa fossor (Orthoptera). *Genome*, *35*(1), 163–166.

Schübeler D. 2015. Function and information content of DNA methylation. *Nature*, *517*(7534), 321–326.

Schultz MD, et al. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, *523*(7559), 212–216.

Schulz NKE, et al. 2018. 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.

Standage DS, et al. 2016. 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.

Susan JC, Harrison J, Paul CL, & Frommer M. 1994. High sensitivity mapping of methylated cytosines. *Nucleic Acids Research*, *22*(15), 2990–2997.

Suzuki MM, & Bird A. 2008. DNA methylation landscapes: Provocative insights from epigenomics. *Nature Reviews Genetics*, *9*(6), 465–476.

Suzuki MM, Kerr ARW, De Sousa D, & Bird A. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Research*, *17*(5), 625–631.

Takuno S, Ran JH, & Gaut BS. 2016. Evolutionary patterns of genic DNA methylation vary across land plants. *Nature Plants*, *2*(2), 1–7.

Thomas GWC, et al. 2020. Gene content evolution in the arthropods. *Genome Biology*, *21*(1), 1–14.

Tweedie S, Charlton J, Clark V, & Bird A. 1997. Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Molecular and Cellular Biology*, *17*(3), 1469–1475.

Vasilikopoulos A, et al. 2020. An integrative phylogenomic approach to elucidate the evolutionary history and divergence times of Neuropterida (Insecta: Holometabola). *BMC Evolutionary Biology*, *20*(1), 1–24.

Wang X, et al. 2013. Function and Evolution of DNA Methylation in *Nasonia vitripennis*. *PLoS Genetics*, *9*(10).

Wang X, et al. 2014. The locust genome provides insight into swarm formation and long-distance flight. *Nature Communications*, *5*, 2957.

Wang Y, et al. 2006. Functional CpG Methylation System in a Social Insect. *Science*, *314*(5799), 645–647.

Wipfler B, et al. 2019. Evolutionary history of Polyneoptera and its implications for our understanding of early winged insects. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(8), 3024–3029.

Xiao S, Cao X, & Zhong S. 2014. Comparative epigenomics: Defining and utilizing epigenomic variations across species, time-course, and individuals. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *6*(5), 345–352.

Yi S. 2012. Birds do it, bees do it, worms and ciliates do it too: DNA methylation from unexpected corners of the tree of life. *Genome Biology*, *13*(10), 5–7.

Yin C, et al. (2016). InsectBase: A resource for insect genomes and transcriptomes. *Nucleic Acids Research*, *44*(D1), D801–D807.

Yoder JA, Walsh CP, & Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, *13*(8), 335–340.

Chapter 2

DNA Methylation Patterns in the Sawflies *Athalia rosae* and *Orussus abietinus*

This chapter has been published as part of the supplementary material in the following publication: Oeyen et al. 2020. Sawfly genomes reveal evolutionary acquisitions that fostered the mega-radiation of parasitoid and eusocial Hymenoptera. *Genome Biology and Evolution*. 12(7), 1099-1188. doi: 10.1093/gbe/evaa106

Authors of only this part of the original article: Provataris P., Grath S., Misof B.

2.1 Introduction

DNA methylation, the addition of a methyl group to a genomic cytosine, is a common epigenetic modification among eukaryotes (Zemach et al., 2010). Functionally, it is associated with the regulation of gene expression and the silencing of transposable elements (Schübeler, 2015). In metazoa, it primarily targets CpG dinucleotides (Feng et al., 2010) and this addition is catalyzed by two types of DNA methyltransferases (DNMTs), DNMT1 and DNMT3. DNMT1 is considered a maintenance methyltransferase, faithfully copying methylation patterns among cell generations, while DNMT3 is responsible for the *de novo* establishment of methylation patterns (Goll and Bestor, 2005). Another enzyme of the DNMT family, TRDMT1 (also known as DNMT2), previously considered a DNA methyltransferase, has been shown to methylate tRNA (Goll et al. 2006).

Within insects, the patterns, levels, and functional importance of DNA methylation are variable (Glastad et al. 2011). For example, the fruit fly, *Drosophila melanogaster* has extremely low levels of DNA methylation and lacks functional DNMTs (Raddatz et al., 2013). In contrast to *D. melanogaster*, patterns of DNA methylation play a crucial role for the development of alternative castes in the honey bee, *Apis mellifera* (Kucharski et al., 2008; Foret et al., 2012), which possess a full complement of DNMTs (Wang et al., 2006). The potential role of DNA methylation in regulating social behavior, caste development or phenotypic plasticity in general, motivated studies in insects displaying varying levels of sociality, mostly Hymenoptera (Lyko et al., 2010; Bonasio et al., 2012; Glastad et al., 2013; Simola et al., 2013; Wang et al., 2013; Cunningham et al., 2015; Patalano et al., 2015).

However, nothing is known about the presence of DNMTs or methylation in the basal Hymenoptera lineages.

Methylation studies so far, have shown that the primary targets of DNA methylation in Hymenoptera and other holometabolous lineages are the gene bodies (exons + introns), but mostly the exons, of highly conserved housekeeping genes (Glastad et al., 2014). In species where methylation persists, like *Apis mellifera*, genes are typically categorized in two classes based on their methylation content (lowly and highly methylated genes) (Sarda et al., 2012). In the absence of direct measurements, patterns and levels of DNA methylation can be estimated through CpG depletion, that is, the underrepresentation of CpG dinucleotides in regions affected by DNA methylation, due to the accelerated mutational decay of methylated cytosines (Yi and Goodisman, 2009; Sarda et al., 2012). Therefore, in order to identify the occurrence of DNA methylation we calculated the normalized CpG content [CpG observed /expected (o/e)] and searched for the DNMT toolkit genes in the genomes of *Orussus abietinus* and *Athalia rosae*.

2.2 Material and Methods

We calculated the normalized CpG dinucleotide content using the following equation:

$$CpG\ o/e = \frac{P_{CpG}}{P_C * P_G}$$

where PCpG, PC and PG are the frequencies of CpG dinucleotides, C nucleotides and G nucleotides, respectively. CpG o/e is a metric of depletion of CpG dinucleotides, normalized by G and C nucleotide content (GC content) and length of the specific

region of interest (Elango et al., 2009). In addition, we determined the distribution of GpC o/e, that is, the observed/expected ratio of 5'-GpC-3' dinucleotides, to control for potential artifacts related to GC content (Fryxell and Moon, 2005).

We calculated the CpG o/e distributions for different genomic regions of the sawflies Athalia rosae and Orussus abietinus. To calculate the CpG o/e of the genomic background we separated scaffolds into 1000-long nucleotide fragments. The term “CDS” represents coding sequences of each gene combined. Individual sequences that contained more than 5% 'N' characters and had a CpG o/e equal to zero were excluded from the graphs and the calculations of mean values. To test for bimodality of CpG o/e distributions, we used model-based clustering using Gaussian Mixture Model (GMM). We used the MCLUST package (Fraley and Raftery 1999; Fraley and Raftery 2003) version 5.2 as implemented in R to estimate the heterogeneous variance model for two components (k = 2, bimodality). All analyses were performed under R version 3.2.5 on a x86_64-pc-linux-gnu (64-bit) platform under Ubuntu precise (12.04.5 LTS).

For the identification of the DNMTs (DNMT1, 2, and 3) in the sawflies, DNMT (DNMT1, 2, 3) protein sequences of *Apis mellifera* and *Nasonia vitripennis* were downloaded from OrthoDBv8 (Kriventseva et al., 2015). Subsequently, each of the downloaded DNMT sequences was scanned against the Pfam-A profile HMM library (Finn et al., 2014). The latter step was made to confirm that the enzymes to be used as candidates possessed the expected domain structure. In the next step, we used the DNMTs of *N. vitripennis* and *A. mellifera* to query the draft genome of each sawfly with the i5K Blast application. Finally, gene models predicted by MAKER

were visually evaluated taking into account the available RNAseq data and if necessary modified, according to the i5K annotation guidelines. To confirm the validity of each prediction, the corresponding protein sequences for every gene prediction were scanned against both InterPro (Mitchell et al., 2014) and UniProt databases (The Uniprot Consortium, 2014), without observing any signs of controversy.

2.3 Results

Full sets of DNMTs were identified in both genomes. For *Orussus abietinus*, the DNMT1 was split between Scaffolds 33 and 538 (Table 2.1). Furthermore, the modification of the predicted models was required in two cases. First, for the *O. abietinus* DNMT2 model, the 5' end was aligned with the one predicted by SNAP, while the 3' end was aligned with the one predicted by Augustus gene prediction tool. This model was then confirmed by Pfam scan as the best one available (Table 2.1). Second, the A. rosae DNMT3 gene model was split in two paralogs, as the predicted model contained the anticipated domains twice and in sequential order, a fact dictating the presence of two distinct enzymes (Table 2.2).

We relied on CpG o/e distributions in order to assess the levels and patterns of germline methylation in different genomic regions of the sawflies. First, we found that CpG dinucleotides are overrepresented in both genomes, with the mean CpG o/e value for the whole genome being approximately 1.41 for *A. rosae* and 2.27 for *O. abietinus* (Figure 2.1). Second, in both species, gene coding sequences (CDS) and to a

lesser extent introns, display lower mean CpG o/e values than the genomic background (Figure 2.1).

In contrast to the similarities observed in the levels of CpG depletion, *A. rosae* displays a distinct CpG o/e pattern to *O. abietinus*. All genomic regions in *A. rosae* inspected resemble a “bimodal” pattern. The fitting of two normal distributions to the CpG o/e distributions (Figure 2.2) reveal two distinct classes of introns, coding regions, and genomic regions, with one class presenting lower than expected CpG frequencies (high methylation) and the other class higher (methylation is lower or absent). Additionally, the GpC o/e distributions of introns, CDS and the whole genome are unimodal with a mean frequency around 1 (Figure 2.2). In contrast to *A. rosae, O. abietinus* does not show “bimodal” CpG o/e distributions for any of the inspected genomic elements. Fitting two normal distributions in the CpG o/e distributions of *O. abietinus* does not point to the existence of two separate classes of introns, gene coding regions or other genomic elements as the frequencies of the assigned components are close to 1 or higher, or the proportions of the second component are negligible and do not represent a distinct class (Figure 2.2).

2.4 Discussion

DNMT duplication is a common phenomenon within insects and especially within Hymenoptera. Paralogs of the maintenance methylation enzyme, DNMT1, were identified in *Apis mellifera*, *Bombus terrestris* and *Bombus impatiens*, *Nasonia vitripennis*, *Acyrthosiphon pisum* and *Pediculus humanus* (Werren et al., 2010; Sadd et al., 2015). Regarding DNMT3, the ants *Harpegnathos saltator* and *Camponotus*

floridanus both possess single copy enzymes similar to human DNMT3B and a single copy of DNMT3L, a non-catalytic factor required for germ cell methylation in mammals (Bonasio et al., 2010). However, unlike DNMT3L which lacks a PWWP domain (Pfam id: PF00855) (Jurkowska et al., 2011), both of the DNMT3 paralogs identified in this project for *Athalia rosae* possess such a region, alongside an active catalytic DNA methylase domain (Pfam id: PF00145). The same is true for the DNMT3 identified in the genome of *Orussus abietinus*. Consequently, DNMT3 copies identified for both sawflies possess the domain structure of an active *de novo* DNA methyltransferase eznyme.

The overrepresentation of CpG dinucelotides constitutes a common pattern among various hymenopteran genomes (*Apis mellifera*, *Nasonia vitripennis*, *Cerapachys biroi*, *Harpegnathos saltator* and *Camponotus floridanus*) (Honeybee Genome Sequencing Consortium, 2006; Bonasio et al., 2010; Xiao et al., 2013) despite the presence of genomic methylation in all of them (Lyko et al., 2010; Bonasio et al., 2012; Wang et al., 2013; Libbrecht et al., 2016). The lower CpG o/e value in CDS and introns, when compared to the genomic background, is consistent with the notion that gene bodies, and primarily exons, are the main targets of DNA methylation in Holometabola (Glastad et al., 2014). Furthermore, the “bimodal” CpG o/e pattern in *Athalia rosae* is typical for species that present DNA methylation, like *A. mellifera* (Elango et al., 2009). Due to the separation of CDS, introns and genomic regions into two distinct classes of lower and higher than expected CpG frequencies, as well as the unimodal GpC o/e distributions of the same features, we can attribute the observed

CpG o/e bimodality to DNA methylation and not other processes related to GC content (Fryxell and Moon, 2005).

Overall, CpG o/e distributions are indicative for the presence of germline methylation in *A. rosae*, but not for *O. abietinus*. However, "unimodal" CpG o/e distributions have been observed in species with empirically verified DNA methylation, a fact dictating that CpG o/e bimodality is not always a consequence of DNA methylation (Glastad et al., 2011). Furthermore, the conspicuously lower mean for the CDS of *O. abietinus* compared to other genomic regions may be also a hint for higher CpG depletion of exons in comparison to other regions, but direct measurements of DNA methylation should be applied in order to resolve the matter.

Within Hymenoptera, bimodality of CpG o/e distributions in genic sequences, similar to the pattern we present for *A. rosae* has been observed in species belonging to the family Apidae (*Apis mellifera*, *Bombus terrestris* and *Bombus impatiens*) (Elango et al., 2009; Sadd et al., 2015). The halicitid bee *Lasioglossum albipes*, paper wasps of the genus *Polistes* (*P. dominula, P. canadensis*), various ant species (*Solenopsis invicta*, *Harpegnathos saltator, Camponotus floridanus, Cerapachys biroi*) and chalcid wasps (*Nasonia vitripennis, Ceratosolen solmsi*) lack striking bimodality as observed in bees (Park et al., 2011; Wurm et al., 2011; Bonasio et al., 2012; Kocher et al., 2013; Xiao et al., 2013; Oxley et al., 2014; Patalano et al., 2015; Standage et al., 2016). However, DNA methylation is widely present in ants (Bonasio et al., 2012; Hunt et al., 2013; Patalano et al., 2015; Libbrecht et al., 2016), has been thoroughly documented in *Nasonia vitripennis* (Wang et al. 2013) and is lower than other Hymenoptera, but present in paper wasps (lower than ants and *A. mellifera* in

both *Polistes* species, greatly reduced in *P. dominula*) (Patalano et al., 2015; Standage et al., 2016).

In conclusion, the full set of DNMTs is present in these two basal hymenopteran species, with a duplication of DNMT3 in the genome of *Athalia rosae*. Germ line methylation, with a pattern similar to *Apis mellifera* and other derived Hymenoptera, is present and pronounced in the genome *Athalia rosae*. Though no clear evidence of germ line methylation was found in the genome of *Orussus abietinus*, it is still possible to be present and direct measurements of methylation should be applied to resolve the question.

2.5 Figures and Tables

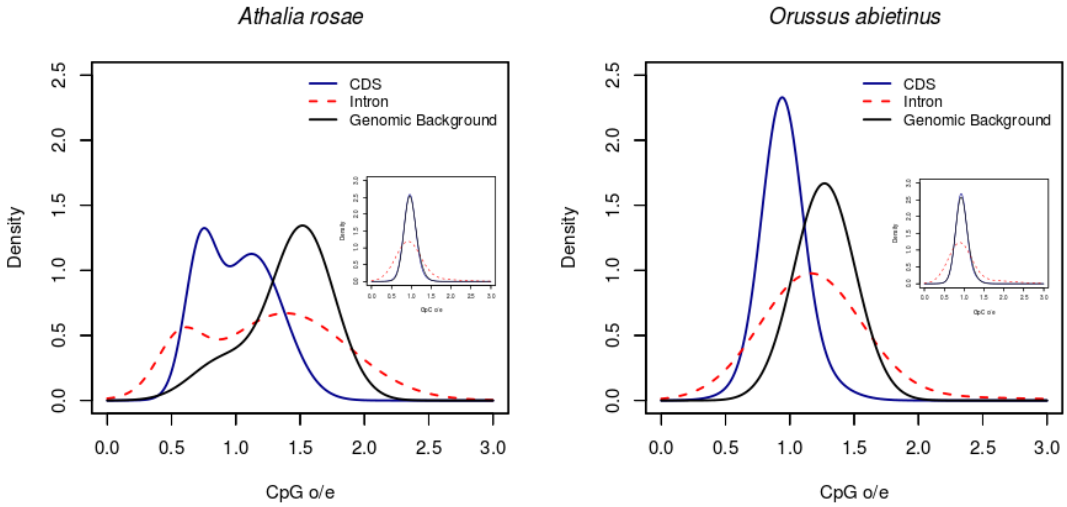
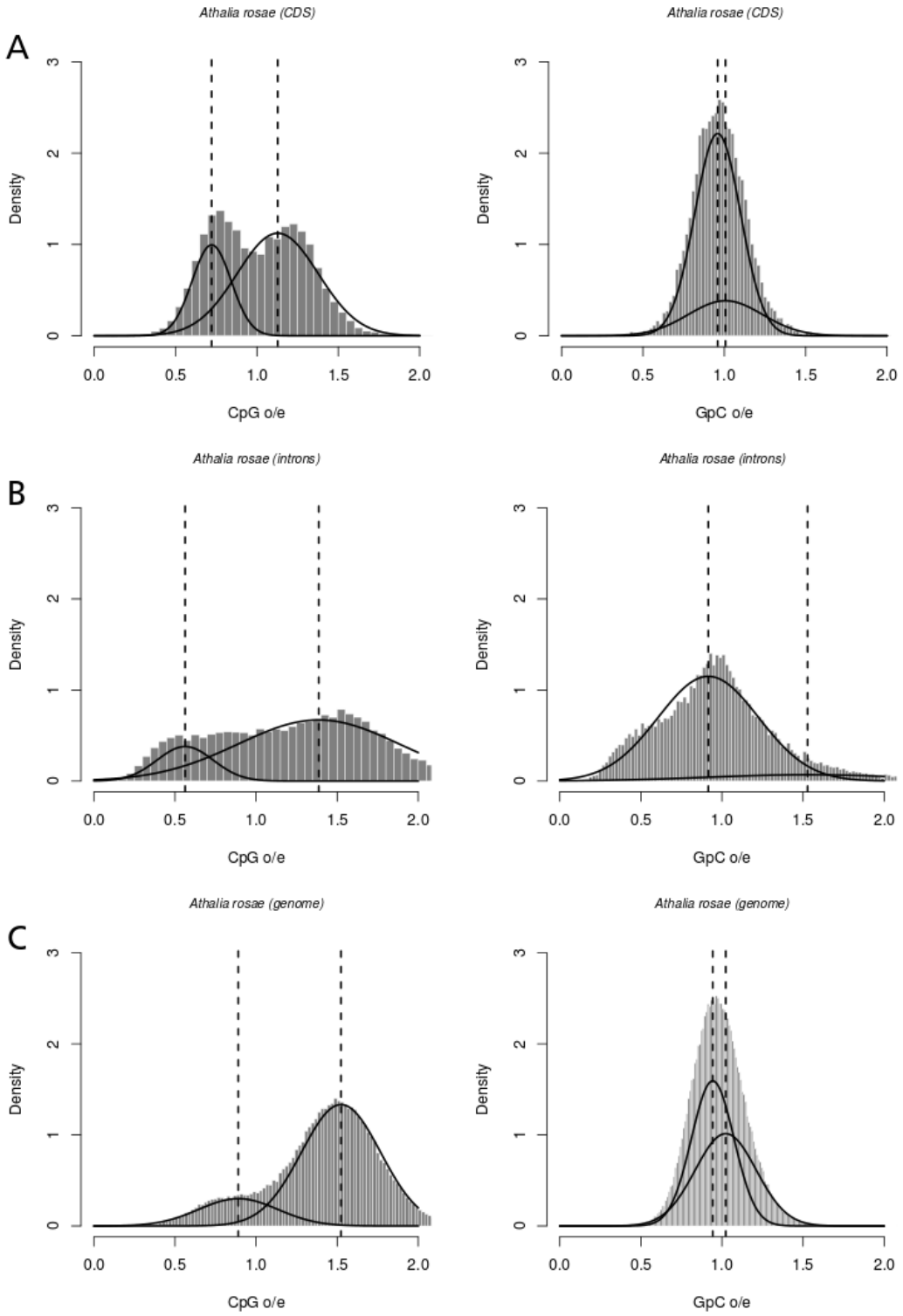


Figure 2.1: Distributions of CpG o/e and GpC o/e of *Athalia rosae* and *Orussus abietinus* for different genomic elements. The densities of CpG o/e were plotted for different genomic elements. They show stronger CpG depletion, and therefore higher DNA methylation, in coding sequences (CDS) and introns in comparison to the genomic background (whole genome, 1 kB windows), in particular for *Athalia rosae*. GpC o/e (inset) was used to control for effects unrelated to DNA methylation.


A
Athalia rosae (CDS)
Density
CpG o/e
Athalia rosae (CDS)
Density
GpC o/e
B
Athalia rosae (introns)
Density
CpG o/e
Athalia rosae (introns)
Density
GpC o/e
C
Athalia rosae (genome)
Density
CpG o/e
Athalia rosae (genome)
Density
GpC o/e

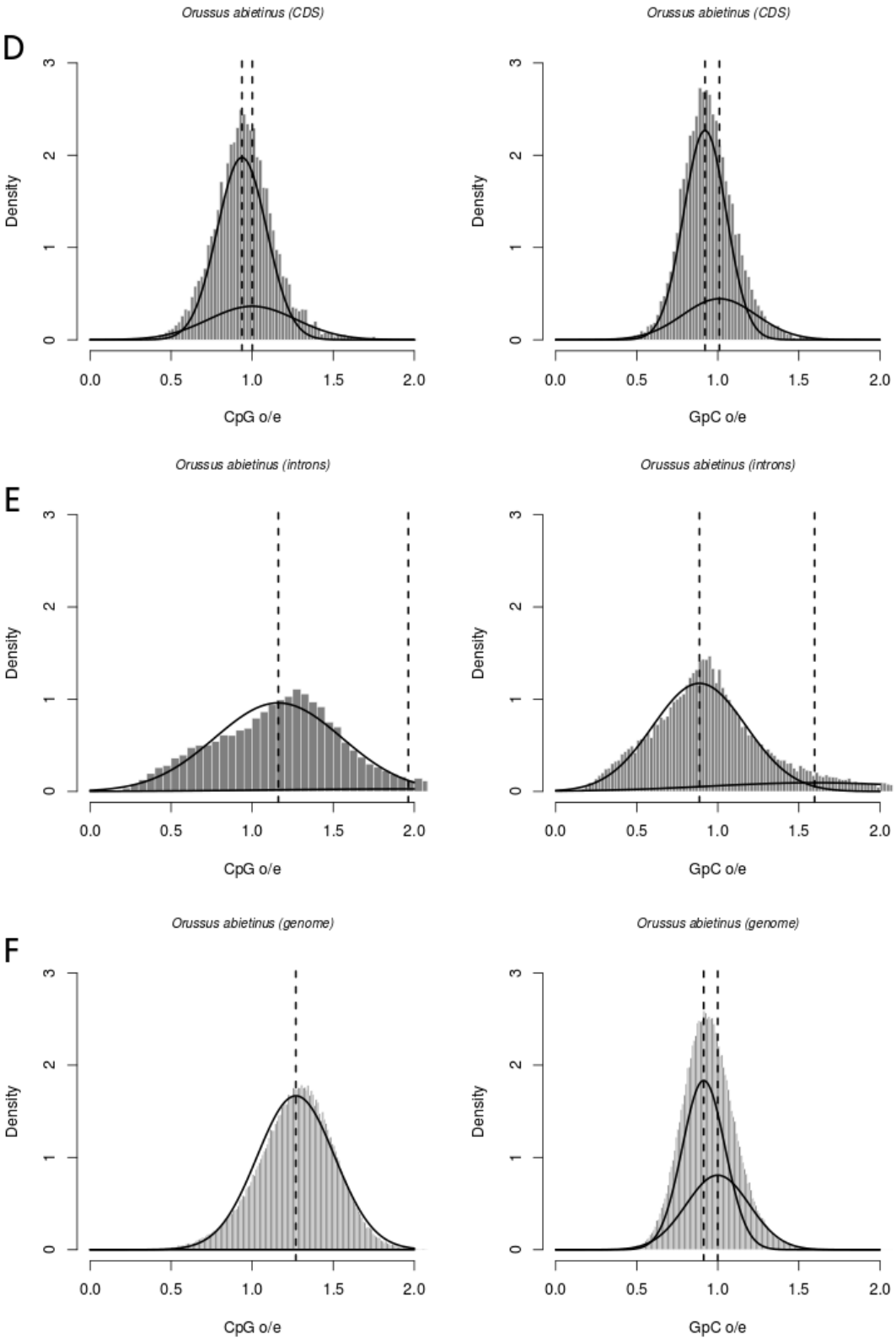

D
Orussus abietinus (CDS)
Density
CpG o/e
Orussus abietinus (CDS)
Density
GpC o/e
E
Orussus abietinus (introns)
Density
CpG o/e
Orussus abietinus (introns)
Density
GpC o/e
F
Orussus abietinus (genome)
Density
CpG o/e
Orussus abietinus (genome)
Density
GpC o/e

Figure 2.2: CpG and GpC o/e distributions for different genomic regions of *Athalia rosae* and *Orussus abietinus* for the bimodal model. The densities for the heterogeneous variance models for a bimodal (k = 2) case for CpG o/e as estimated by model-based clustering using Gaussian Mixture Model (GMM) as implemented in mclust 5.2 were plotted for the coding sequences (A and D), introns (B and E) and whole genomes (C and F) for both species. Black curves indicate the density distributions of each fitted component for the bimodal case. Black vertical dashed lines indicate the means for each fitted component models.

Table 2.1: Annotation details for the DNMTs and TRDMT1 of *Orussus abietinus*. The double presence of DNMT1 does not indicate the presence of paralogs, but a potential problem with scaffolding.

Enzyme name	Domains identified by Pfam scan	Scaffold number
DNMT1	DNMT1-RFD BAH BAH zf-CXXC	33
DNMT1	DNA methylase	538
TRDMT1	DNA methylase	21
DNMT3	PWWP DNA methylase	120

Table 2.2: Annotation details for the DNMTs and TRDMT1 of *Athalia rosae*. The presence of two paralogs of DNMT3 is indicated by “a” and “b”.

Enzyme name	Domains identified by Pfam scan	Scaffold number
DNMT1	DNMT1-RFD BAH BAH zf-CXXC DNA methylase	226
TRDMT1	DNA methylase	5
DNMT3 – a	PWWP DNA methylase	33
DNMT3 – b	PWWP DNA methylase	33

2.6 References

Bonasio R, et al. 2012. Genome-wide and caste-specific DNA methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Current Biology*. *22*(19), 1755–1764.

Bonasio R, et al. 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science.* 329, 1068–1071.

Cunningham CB, et al. 2015. The genome and methylome of a beetle with complex social behavior, *Nicrophorus vespilloides* (Coleoptera: Silphidae). *Genome Biology and Evolution*. *7*(12), 3383-3396.

Elango N, Hunt BG, Goodisman MAD, Yi SV. 2009. 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.

Feng S, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*. 107, 8689–8694.

Finn RD, et al. 2014. Pfam: The protein families database. *Nucleic Acids Research*. 42, 222–230.

Foret S, Kucharski R, Pellegrini M. 2012. DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. *Proceedings of the National Academy of Sciences*, *109*(13), 4968-4973.

Fraley C, Raftery AE. 1999. MCLUST: Software for model-based cluster analysis. *Journal of Classification*. *16*(2), 297-306.

Fraley C, Raftery AE (2003): Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. *Journal of Classification*. *20*(2), 263-86.

Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. *Molecular Biology and Evolution*. *22*(3), 650–658.

Glastad KM, Hunt BG, Goodisman MAD (2014): Evolutionary insights into DNA methylation in insects. *Current Opinion in Insect Science*. 1, 25-30.

Glastad KM, Hunt BG, Goodisman MAD. 2013. Evidence of a conserved functional role for DNA methylation in termites. *Insect Molecular Biology*. *22*(2), 143–154.

Glastad KM, Hunt BG, Yi SV, Goodisman MAD. 2011. DNA methylation in insects: On the brink of the epigenomic era. *Insect Molecular Biology*. 20, 553–565.

Goll MG, Bestor TH. 2005. Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry*. 74, 481–514.

Goll MG, et al. 2006. Methylation of tRNA. *Science*. 311, 395–398.

Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*. *43*(7114), 931.

Hunt BG, Glastad KM, Yi SV, Goodisman MAD. 2013. Patterning and regulatory associations of DNA methylation are mirrored by histone modifications in insects. *Genome Biology and Evolution*. *5*(3), 591-8.

Jurkowska RZ, Jurkowski TP, Jeltsch A. 2011. Structure and function of mammalian DNA methyltransferases. *ChemBiochem*. *12*(2), 206-22.

Kocher SD, et al. 2013. The draft genome of a socially polymorphic halictid bee, *Lasioglossum albipes*. *Genome Biology*. *14*(12):1.

Kriventseva EV, et al. 2014. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research*. *43*(D1), D250-6.

Kucharski R, Maleszka J, Foret S, Maleszka R. 2008. Nutritional control of reproductive status in honeybees via DNA methylation. *Science.* 319, 1827–1830.

Libbrecht R, Oxley PR, Keller L, Kronauer DJC (2016): Robust DNA Methylation in the Clonal Raider Ant Brain. *Current Biology*. *26*(3), 391-5.

Lyko F, et al. 2010. The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. *PLoS Biology*. *8*(11).

Mitchell A, et al. 2014. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*. Gku1243.

Oxley PR, et al. 2014. The genome of the Clonal raider ant *Cerapachys biroi*. *Current Biology*. *24*(4), 451-8.

Park J, et al. 2011. Comparative analyses of DNA methylation and sequence evolution using *Nasonia* genomes. *Molecular Biology and Evolution*. *28*(12), 3345-54.

Patalano S, et al. 2015. Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. *Proceedings of the National Academy of Sciences*. *112*(45), 13970-13975.

Raddatz G, et al. 2013. Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proceedings of the National Academy of Sciences*. 110, 8627–8631.

Sadd BM, et al. 2015. The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biology*. 16, 76. Available from: http://genomebiology.com /2015/ 16/1/76

Sarda S, Zeng J, Hunt BG, Yi SV. 2012. The Evolution of Invertebrate Gene Body Methylation. *Molecular Biology and Evolution*. *29*(8), 1907–1916.

Schübeler D (2015): Function and information content of DNA methylation. *Nature* 517, 321–326.

Simola DF, et al. 2013. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Research*. *23*(8):1235–1247.

Standage DS, et al. 2016. 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.

The Uniprot Consortium. 2014. UniProt: a hub for protein information. *Nucleic Acids Research*. 43, D204–D212. Gku989.

Wang X, et al. 2013. Function and evolution of DNA methylation in *Nasonia vitripennis*. *PLoS Genetics*. *9*(10), e1003872.

Wang Y, et al. 2006. Functional CpG methylation system in a social insect. *Science*. *314*(5799), 645–647.

Werren JH, et al. 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science*. 327, 343–348.

Wurm Y, et al. 2011. The genome of the fire ant *Solenopsis invicta*. *Proceedings of the National Academy of Sciences*. 108, 5679–5684.

Xiao JH, et al. 2013. Obligate mutualism within a host drives the extreme specialization of a fig wasp genome. *Genome Biology*. 14**,** R141.

Yi SV, Goodisman MAD. 2009. Computational approaches for understanding the evolution of DNA methylation in animals. *Epigenetics*. *4*(8), 551–556.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science.* 328, 916–919.

Chapter 3

Signatures of DNA Methylation across Insects Suggest Reduced DNA Methylation Levels in Holometabola

This chapter has been published in: Provataris P., Meusemann K., Niehuis O., Grath S., Misof B. 2018. Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola. *Genome Biology and Evolution*. 10(4), 1185–1197. doi:10.1093/gbe/evy066

3.1 Introduction

Methylation of cytosine residues constitutes a common epigenetic modification among eukaryotes. It is functionally associated with the regulation of expression of genomic elements (Zemach et al. 2010). For example, promoter-proximate methylation is linked to the transcriptional repression of associated genes (Jones 2012; Schübeler 2015). Methylation of repetitive non-coding DNA elements also has a repressive effect, limiting the expression, and thus, the genomic expansion of these elements (Schübeler 2015). In contrast, intragenic methylation is associated with active transcription (Feng et al. 2010; Zemach et al. 2010; Jones 2012), but a "cause and effect" relationship has not been established in this context (Schübeler 2015). Although DNA methylation is widely present among eukaryotes, the levels (i.e. the proportion of methylated cytosines or CpG sites in a given genome), patterns, and genomic targets of DNA methylation are not evolutionarily conserved. In vertebrates, and especially in mammals, CpG dinucleotides are heavily methylated genome-wide, with the exception of CpG islands. CpG islands typically overlap with promoter regions and remain mostly unmethylated (Schübeler et al. 2015). In contrast, invertebrates show intermediate or even negligible levels of DNA methylation at CpG sites, which is typically targeted to a subset of gene bodies (the term gene body refers to the transcribed part of a gene, comprised of exons and introns) (Suzuki and Bird 2008; Feng et al. 2010; Zemach et al. 2010; but see Wang et al. 2014; Kao et al. 2016).

In insects, the levels of gene body methylation vary considerably (Glastad et al. 2011). On one hand, model organisms like the fruit fly, *Drosophila melanogaster*,

and the red flour beetle, *Tribolium castaneum*, do not display notable DNA methylation levels in their genomes (Zemach et al. 2010; Bewick et al. 2017). On the other hand, nutritionally regulated levels of DNA methylation contribute to the ontogenetic establishment of alternative castes in the honey bee, *Apis mellifera* (Kucharski et al. 2008; Foret et al. 2012). This observation has supported the hypothesis that DNA methylation in insects is associated with caste development and the evolution of (eu)sociality, but recent empirical evidence from research on eusocial Hymenoptera (wasps, ants, and bees) suggests that this association is not universal (Bonasio et al. 2012; Patalano et al. 2015; Kapheim et al. 2015; Libbrecht et al. 2016; Standage et al. 2016). Obviously, a taxonomically representative description of the levels and patterns of DNA methylation is one major prerequisite to improve our understanding of the evolution and, eventually, the function of DNA methylation in insects. Therefore, we conducted a comparative analysis of DNA methylation patterns in insects by making use of recently published, extensive transcriptomic (Misof et al. 2014) and publicly available genomic sequence data covering all extant insect orders.

Two types of DNA methyltransferases (DNMTs), DNMT1 and DNMT3, are responsible for DNA methylation in animals (Goll and Bestor 2005). In mammals, DNMTs carry out *de novo* (DNMT3) and maintenance (DNMT1) methylation, with functional overlap (Jeltsch and Jurkowska 2014). Another, non-canonical member of the DNMT family, TRDMT1 (tRNA aspartic acid methyltransferase 1, most commonly known as DNMT2), long considered a DNA methyltransferase, has seemingly shifted substrate and is now known to methylate tRNA, not DNA (Goll et al.2006; Lyko 2017). It is generally assumed that these functions are conserved in

insects (Wang et al. 2006). This assumption is supported by the observation that the absence of DNMT1 and DNMT3 is associated with the loss or extreme reduction of DNA methylation in *D. melanogaster* (Raddatz et al. 2013). However, in contrast to mammals, the DNMT toolkit of insects that show substantial levels of DNA methylation in their genomes is not conserved. For example, the silk moth, *Bombyx mori*, has empirically determined DNA methylation, but lacks copies of DNMT3 homologs from its genome (Xiang et al. 2010; Bewick et al 2016). Thus, functional methylation systems in insects can be realized in the absence of DNMT3. The frequency of DNMT3 loss in different lineages is, however, unknown due to a lack of extensive comparative data.

While DNMTs are responsible for generating methylcytosine (5mC) residues, Tet dioxygenases are shown to convert 5mC to hydroxymethylcytosine (5hmC) in many animal species (Pastor et al 2013). In contrast to mammals which harbor three Tet paralogs, invertebrate species, including some insects, seem to encode a single Tet homolg without a characterized function in the majority of cases (Pastor et al 2013; Wojciechowski et al 2014). In the honey bee, it was recently shown that the single Tet enzyme is capable of converting 5mC to 5hmC (Wojciechowski et al 2014). However, Tet enzymes may display functional promiscuity in insects, since a Tet homolg seems to mediate N6-methyladenine demethylation and 5mC demethylation in the *D. melanogaster* DNA and mRNA, respectively (Zhang et al. 2015; Delatte et al. 2016). The distribution of Tet enzymes in insects and its relationship to the presence of 5mC is currently not known.

Comparative analyses using experimental data have shown that the levels of gene body methylation of the silk moth and the honey bee are substantially lower compared to other invertebrates (sea squirt, *Ciona intestinallis*, and sea anemone, *Nematostella vectensis*; Sarda et al. 2012). These results have fueled the hypothesis that DNA methylation was reduced in the ancestors of insects (Glastad et al. 2014). However, experimental and computational evidence from research on hemimetabolous lineages point to significantly elevated DNA methylation levels in species belonging to Orthoptera, Phasmatodea, and Isoptera compared to certain Hymenoptera and *Bombyx mori* (Krauss et al. 2009; Falckenhayn et al. 2013; Glastad et al. 2013; Terrapon et al. 2014; Glastad et al. 2016). Thus, the validity of the proposed hypothesis on the ancestral state of DNA methylation in insects is questionable.

The methylomic profiling of insects, mostly representing Hymenoptera, and to a lesser extent Lepidoptera and Coleoptera, revealed largely similar patterns of DNA methylation, primarily targeted to exons of protein-coding genes (Lyko et al. 2010; Xiang et al. 2010; Bonasio et al. 2012; Wang et al. 2013; Cunningham et al. 2015; Patalano et al. 2015; Libbrecht et al. 2016; Rehan et al. 2016; Standage et al. 2016). Additionally, genes targeted by DNA methylation were ubiquitously expressed among various tissue types (Foret et al. 2009; Xiang et al. 2010), among different morphs in ants (Bonasio et al. 2012; Libbrecht et al. 2016), and among developmental stages in the parasitoid wasp, *Nasonia vitripennis* (Wang et al. 2013). Gene ontology annotations showed that the majority of these genes mostly serve basic cellular functions, exhibit a highly methylated state among species, and are highly conserved

at sequence level (Elango et al. 2009; Lyko et al. 2010; Hunt et al. 2013; Wang et al. 2013; Cunningham et al. 2015; Rehan et al. 2016). These patterns are even found when comparing orthologous genes among distantly related invertebrates (Sarda et al. 2012). These findings strongly imply that the targeting of DNA methylation in insect genomes is non-random, but a solid explanation for this observation remains elusive.

The aim of the present study is to improve our understanding of the evolution of DNA methylation in insects. Specifically, we focused on the hypothesis stating that DNA methylation has been reduced in the ancestors of insects (Glastad et al. 2014). For this purpose, we analyzed whole-body transcriptomes and genomic data (protein-coding sequences and predicted proteins) of 143 insect species, representing all 32 currently recognized insect orders, and eleven outgroup species. First, we document the presence or absence of DNA methyltransferases (DNMT1, DNMT3). Second, we use the normalized CpG dinucleotide content (CpG observed/expected or simply CpG o/e) to predict the occurrence and estimate the levels of DNA methylation in protein-coding sequences. The last approach provided the means to assess the relationship between DNA methylation and the evolutionary conservation of genes across insects.

We found that, unlike in vertebrates, the phylogenetic distribution of DNMT1 in insects is much wider compared to DNMT3. Based on the patterns of CpG o/e distributions, our data suggest that DNA methylation is widespread among insect orders. More importantly, we estimate DNA methylation levels of protein-coding sequences to be significantly higher in hemimetabolous insects than in Holometabola. Finally, we show that single-copy genes present across insects tend to display signs of heavy DNA methylation compared to the genomic background. Our analyses point to

a complex DNA methylation landscape in insects and set the basis for large scale comparative analyses using direct measurements of DNA methylation.

3.2 Material and Methods

Data Acquisition

We identified DNMTs, Tet dioxygenases and calculated CpG o/e ratios of 102 transcript assemblies from the 1KITE project (www.1kite.org) representing species of all extant insect orders (Misof et al. 2014); we used the latest version of all 1KITE assemblies (Table S1). Details concerning sequencing and assembly are described by Misof et al. (2014) and Mayer et al. (2016). We appended the 1KITE data with additional transcriptomic and genomic (CDS and predicted proteins) data of 53 arthropod species obtained from public and other resources (Table S2). For orthology assessment (see below), we used the 1KITE species, the aforementioned published transcriptomes, and 14 arthropod official gene sets previously used by Misof et al. (2014, Tables S2, S4; in this study, Table S3).

Identification of DNA Methyltransferases and Tet Dioxygenases

To search for DNMT1, DNMT3, TRDMT1, and Tet homologs in the transcriptomes and genomes presented previously, we constructed profile Hidden Markov Models (pHMMs) for the proteins in question (all pHMMS are available at: doi: 10.17632/8y5wm8887b.3). Amino-acid sequences of arthropod DNMTs and Tet proteins were downloaded from OrthoDB using the text-based search option (Kriventseva et al. 2014; Zdobnov et al 2017). Subsequently, we aligned each group

of orthologous sequences using MAFFT L-INSI (Katoh and Standley 2013) and generated pHMMs from each alignment using HMMER 3.1b1 (www.hmmer.org). We translated transcript sequences into all six possible reading frames with Exonerate, version 2.2.0 (Slater and Birney 2005) and searched with each pHMM the translated transcriptome and genome (predicted proteins) data using hmmsearch with default options (HMMER 3.1b1).

Since DNMT1, DNMT3, and TRDMT1 share a homologous DNA methylase domain (Pfam-accession no. PF00145), some sequences were identified as common candidates among these three proteins. Consequently, we removed redundant candidate sequences by keeping the ones with lowest e-value. Furthermore, we excluded all candidate sequences with an e-value higher than 10^{-5} from downstream analyses. To determine whether or not the candidate sequences were properly annotated as DNMT1, DNMT3, or TRDMT1, we introduced the following levels of control. First, we used blastp (BLAST+ v 2.2.28, Camacho et al. 2009) to search candidate sequences against *Nasonia vitripennis* OGS v 2.0 (Munoz-Torres et al. 2010). We selected *Nasonia vitripennis* as reference since it possesses a well-characterized DNMT toolkit (Werren et al. 2010). We excluded all candidate sequences that did not match a corresponding *Nasonia* DNMT as a best hit. Second, we scanned all remaining candidate sequences with a *Nasonia* match against Pfam-A pHMM library (version 27, Finn et al. 2014) and kept only the ones that did contain a characteristic DNA methylase or DNMT1-RFD domain (PF12047 which is a unique DNMT1 domain).

To search for Tet proteins, we compared candidate amino-acid sequences with the Pfam-A pHMM library (version 27, Finn et al. 2014) and retained only the ones that contained an annotated Tet-JBP domain (Pfam-accession no. PF12851) (for a detailed process on the identification of DNMTs and Tet, see Supplementary Material sections 1 and 2).

Calculation of Normalized CpG Dinucleotide Content (CpG o/e)

The normalized CpG dinucleotide content can serve as a proxy for the presence of DNA methylation, because cytosines targeted by DNA methylation are prone to spontaneous deamination into thymines, leading to a gradual reduction of CpG dinucleotides, termed CpG depletion. Therefore, in genomic regions that are subject to intense germline methylation over evolutionary time, CpGs are underrepresented. In contrast, regions with limited germline methylation maintain a high CpG content (Bird 1980). In insects and other invertebrates with considerable levels of DNA methylation in their genomes, two classes of genes are present, one with low CpG o/e (high germline DNA methylation) and another with high CpG o/e (low germline DNA methylation). Thus, a bimodal CpG o/e distribution typically occurs in such cases. In contrast, in species with very low or no DNA methylation, only one class of genes is expected, signified by a unimodal CpG o/e distribution and lack of CpG depletion.

We calculated the normalized CpG dinucleotide content using the following equation:

$$CpGo/e = \frac{P_{CpG}}{P_C * P_G}$$

where P_{CpG}, P_C and P_G are the frequencies of 5'-CpG-3 dinucleotides, C nucleotides and G nucleotides, respectively, estimated from each sequence. In addition, we plotted distributions of the normalized GpC content, to control for causative factors unrelated to DNA methylation, like GC content (Fryxell and Moon 2005). We excluded sequences containing less than 200 nucleotides or containing more than 5% ambiguous nucleotides (N) from the calculation of normalized dinucleotide content. Furthermore, we excluded all nucleotide sequences with a normalized dinucleotide content equal to zero from any downstream analyses. All analyses were carried out using custom-made Perl and R (R Core Team, 2016) scripts.

Inferring the Presence of DNA Methylation Based on CpG o/e Distributions

Species like the honeybee, *Apis mellifera,* and the pea aphid, *Acyrthosiphon pisum,* in which DNA methylation has been experimentally verified, display clear bimodal CpG o/e distributions in protein-coding sequences with two distinct components, one with low CpG o/e and one with high CpG o/e values. A bimodal CpG o/e distribution may thus serve as an indication for the presence of DNA methylation. However, species with experimentally verified DNA methylation, like the branchiopod *Daphnia pulex,* the silk moth *Bombyx mori,* and the beetle *Nicrophorus vespilloides,* lack clearly defined bimodality in protein-coding sequences, but the presence of DNA methylation is indicated due to an extensive tail spanning towards the low CpG o/e part of their distributions (Glastad et al. 2011; Sarda et al. 2012; Cunningham et al. 2015). In contrast, species like *Drosophila melanogaster* and *Tribolium castaneum* in which

DNA methylation in protein-coding sequences is extremely reduced or absent, display a unimodal, almost normal CpG o/e distribution, with a mean around one (*Drosophila melanogaster* ~0.89, *Tribolium castaneum* ~1.1) (Elango et al. 2009). Using these empirically well-documented cases, we defined a set of criteria to infer the presence of DNA methylation based on the modality of CpG o/e distributions. To test the modality of CpG o/e distributions, we used the Gaussian mixture modeling software package mclust (v 5.2) similar to Park and colleagues (2011) and fitted two Gaussian distributions in the CpG o/e and GpC o/e distributions of each species in question. We consider the following criteria as sufficient evidence for the presence of germline DNA methylation in protein-coding sequences of a species:

1) a CpG o/e distribution is bimodal, with one class of genes showing signs of CpG depletion. To identify bimodality, we expect the absolute difference of the means of the two fitted Gaussian distributions to be 0.25 or higher, while one of the fitted means is lower than 0.7. Furthermore, the proportion of data belonging to the smallest of the fitted components should be higher than 0.1. These criteria of bimodality should not be fulfilled by the GpC o/e distribution, which is unaffected by DNA methylation. A CpG o/e distribution fulfilling this set of criteria is described as "bimodal depleted" (Figure 3.1.A).

2) In the absence of clearly defined bimodality, as observed in *Bombyx mori and Daphnia pulex*, we do not expect the criteria of bimodality to apply. However, in both these species a large proportion of data belongs to the smallest of the two fitted distributions (0.36 in *B. mori* and 0.43 in *D. pulex*). If we apply such criteria, we can identify species with similar CpG o/e distributions which, based on empirical

evidence, should indicate the presence of DNA methylation. Therefore, we set the threshold for the proportion of smallest of the fitted normal distributions to 0.36 or higher (equal to that of *B. mori* or higher). This should not apply to the corresponding GpC o/e distribution. The CpG o/e distributions of these species are described as "unimodal, indicative of DNA methylation" (Figure 3.1.B, C).

If the above criteria did not apply, we considered the evidence as insufficient to infer the presence of DNA methylation. The CpG o/e distributions of these species are described as "unimodal, not indicative of DNA methylation" (Figure 3.1.D). We acknowledge that these criteria are conservative. However, we think that missing true positives is likely less misleading than building conclusions based on false positives.

Phylogenetic Generalized Least Squares Analysis

We used Phylogenetic Generalized Least Squares (PGLS) to correlate estimations of DNA methylation in protein-coding sequences (continuous dependent variable; obtained from Bewick et al. 2017) to the mode of development (categorical independent variable, binary coded as hemimetabolism or holometabolism) in 26 holometabolous and 14 hemimetabolous insect species (Table S1 in Bewick et al. 2017). The multilocus coalescent tree estimated by Bewick and colleagues (Figure 1 in Bewick et al. 2017) was used to control for statistical nonindependence between species traits. To perform PGLS, we used the R packages ape (Paradis et al. 2004) and nlme (Pinheiro et al. 2017).

Orthology Assessment

We used an ortholog set of 1,478 protein-coding genes that are single-copy in twelve reference species (Misof et al. 2014). We used Orthograph version 0.5.4 (Petersen et al. 2017) to identify the protein-coding sequences of orthologs of the 1,478 single-copy genes in 129 additional species (see Tables S1 and S3 in this study; Tables S1, S2, and S4 in Misof et al. 2014). We applied a relaxed setting for the reciprocal best hit search to any of the reference species included in the ortholog set. In all identified orthologs (see Table S4), we subsequently masked stop codons and Seleocysteine with X in the predicted amino-acid sequences and with NNN in the coding nucleotide sequences (CDS). We then aligned all orthologous amino-acid sequences as outlined by Misof et al (2014), including check for suspiciously aligned outlier sequences, alignment-refinement of identified outliers, and exclusion of persistent outliers. Subsequently, we generated corresponding multiple sequence alignments (MSAs) on nucleotide level with the software pal2nal (Suyama et al. 2006), using the amino-acid MSAs as blueprint. Finally, the 1,478 MSAs on nucleotide level served as basis for CpG o/e calculations (see Supplementary Material, section 3).

3.3 Results

DNMT1 Homologs are Likely Indispensable for Maintaining a Functional Methylation System in Insects

We characterized the occurrence of DNMTs and Tet proteins in the investigated insect and outgroup species by using profile Hidden Markov Models (pHMMs) constructed from orthologous protein sequences of arthropods for each of the proteins in question.

With these pHMMs at hand, we searched transcriptomes representing all insect orders, crustaceans and myriapods. Transcriptomic data were complemented by genomic data (protein predictions) of species belonging to nine insect orders (Collembola, Isoptera, Hemiptera, Psocodea, Hymenoptera, Strepsiptera, Coleoptera, Lepidoptera, and Diptera) plus crustaceans, myriapods, and a chelicerate (see Materials and Methods).

We identified homologous sequences of DNMT1 in species belonging to all insect orders and outgroups, except Collembola (seven species, including three with sequenced genomes), Diptera (13 species, including three with sequenced genomes), and Strepsiptera (two species, including one with a sequenced genome) (Figure 3.2; Table S5). DNMT3 homologs were not identified in species belonging to these three orders either, which apparently lack all currently known cytosine-specific DNA methyltransferases. In contrast to DNMT1, DNMT3 was sparsely found in insects, being present in species belonging to only seven out of 32 insect orders (Hemimetabola: Diplura, Orthoptera, Isoptera, Hemiptera, and Thysanoptera; Holometabola: Hymenoptera and Coleoptera), plus species of crustaceans and myriapods (Figure 3.2; Table S5). Within hemimetabolous insects, DNMT3 was absent from Palaeoptera (seven species) and the polyneopteran clade formed by Mantophasmatodea, Grylloblattodea, Embioptera, and Phasmatodea (eight species). Within Holometabola, DNMT3 was lacking from Neuropterida (eight species) and Mecopterida (40 species) (Figure 3.2).

The tRNA methyltransferase TRDMT1 was the most commonly found enzyme in our dataset being present in species belonging to 31 out of 32 insect orders

(140/154 species possessed putative TRDMT1 homologs). TRDMT1 was absent from the transcriptome of the only representative of Zoraptera in our dataset, *Zorotypus caudeli* (Table S5).

We identified homologous sequences of Tet dioxygenases in species belonging to 25 out of 32 insect orders, plus species belonging to all three outgroups (Table S6). Within hemimetabolous insects, Tet homologs are apparently missing in Arachaeognatha (2 species), in the polyneopteran clade formed by Mantophasmatodea, Grylloblattodea, Embioptera, and Phasmatodea (eight species), and in Mantodea (3 species). Within Holometabola, only Strepsiptera lack Tet homologs (2 species). We have to note that Tet homologs were consistently identified in genomes (28/30), but not in transcriptomes (52/124).

CpG o/e Patterns Suggest DNA Methylation Being Taxonomically Widespread in Winged Insects

In order to infer the occurrence of DNA methylation in insects, we calculated CpG o/e ratios of protein-coding sequences in 143 species covering all insect orders and eleven additional outgroup species (see Materials and Methods). CpG o/e has been widely used as a proxy for estimating the patterns and levels of DNA methylation in various species of invertebrates (Suzuki et al. 2007; Elango et al. 2009; Glastad et al. 2013) with high concordance to empirical measurements (Glastad et al. 2011; Sarda et al. 2012).

Applying a set of stringent criteria (see Materials and Methods), we identified CpG o/e distributions pointing to the presence of DNA methylation in species

belonging to 24 out of 32 total insect orders (Figure 3.2; Figure S1; Table S7). Furthermore, our data suggest that DNA methylation is applied by close relatives of insects, as we found signatures of DNA methylation in crustaceans (four out of seven species), including the only representative of remipedes (the proposed sister group of insects; Misof et al. 2014) *Xibalbanus tulumensis*, and in the diplopod, *Glomeris pustulata* (Figure 3.2; Figure S1; Table S7). Interestingly, however, CpG o/e distributions pointing to the presence of DNA methylation were not consistently observed in apterygote insect orders, as only species of Diplura (one out of two species) and Zygentoma (two out of three species), but not Protura (one species), Collembolla (seven species), or Archaeognatha (two species) showed signs strongly suggesting the occurrence of DNA methylation. In contrast, we found consistent evidence for the occurrence of DNA methylation in winged hemimetabolous insects, including all representatives of Palaeoptera (all seven species), all polyneopteran orders, except Dermaptera (24 out of 27 species), and many representatives of Condylognatha (Hemiptera (ten out of 16 species), Thysanoptera (all three species)) (Figure 3.2; Figure S1; Table S7). CpG o/e distributions strongly suggesting the presence of DNA methylation are comparatively sparse in Holometabola (17 out of 70 species in total). Representatives of Diptera (15 species), Neuroptera (four species), Raphidioptera (two species) and Strepsiptera (two species) showed no signs of DNA methylation. These results show that CpG o/e distributions pointing to the presence of germline DNA methylation in protein-coding sequences can be easily tracked in the majority of hemimetabolous insects, but are largely absent from holometabolous species.

We did not identify CpG o/e distributions pointing to the presence of DNA methylation in any of the species belonging to eight insect orders (i.e., Protura, Collembola, Archaeognatha, Dermaptera, Neuroptera, Raphidioptera, Strepsiptera, and Diptera). However, in certain species belonging to Archaeognatha, Collembola, Diptera, and Protura, unimodal CpG o/e distributions displayed low mean values (below 0.9 and as low as ~0.7) while corresponding GpC o/e distributions displayed mean values close to the expected ones under random chance (mean ~ 0.9 or higher) (Table S7). These mean CpG o/e values are lower than the ones observed in species with extremely reduced or no DNA methylation (*Aedes aegypti* ~1.1, *Anopheles gambiae* ~1.0, *Drosophila melanogaster* ~0.9, *Tribolium castaneum* ~1.1).

Normalized CpG Content Points to Lower Levels of DNA Methylation in Holometabola

Normalized CpG content constitutes a powerful means for drawing conclusions not only for the patterns, but also for the levels of genomic DNA methylation (Yi and Goodisman 2009). Thus, we calculated the mean CpG o/e value of each transcriptome included in our analysis. First, we compared mean CpG o/e values of holometabolous insects (52 species), to those of hemimetabolous insects (67 species) and outgroup species (six crustacean and two myriapod species) (Figure 3.3.A). Holometabolous insect species exhibited higher overall mean CpG o/e values (lower mean germline DNA methylation) in protein-coding sequences compared to both hemimetabolous insects and outgroups (Kruskal-Wallis H test, $P < 0.001$; ignoring phylogenetic relatedness). Subsequently, we compared mean CpG o/e values of insect species

separated by order (Figure 3.3.C). The majority of species belonging to hemimetabolous insect orders show lower mean CpG o/e values than species belonging to holometabolous orders, except Dermaptera and Psocodea. Species belonging to Zygentoma, Odonata, and most polyneopteran orders (excluding Dermaptera) consistently display very low mean CpG o/e values, with Mantodea representing the most extreme example. Condylognathan species (i.e., Hemiptera and Thysanoptera) tend to display higher mean CpG o/e values than most Polyneoptera, but still clearly lower values than species belonging to orders of Holometabola. Proturan, collembolan, and dipluran species exhibit higher mean values than species of Palaeoptera and Polyneoptera, with the exception of Dermaptera. In conclusion, mean CpG o/e values suggest lower levels of germline DNA methylation in the protein-coding sequences of Holometabola and their closest relatives, Psocodea (Misof et al. 2014).

Despite offering a decent first approximation on the levels of DNA methylation within genes (Sarda et al. 2012), CpG o/e is also suggested to be influenced by other factors, such as local GC content (Fryxell and Moon 2005) and recombination or gene conversion (Kent et al. 2012), for which we cannot currently control. Furthermore, certain insect lineages (most commonly Hymenoptera) are known to possess high mean CpG o/e values, genome-wide (Simola et al. 2013). Thus, we tested whether our observation that levels of DNA methylation are lower in protein-coding sequences of Holometabola compared to hemimetabolous insects still holds when using experimental DNA methylation data. To do that, we exploited the recently published and most comprehensive to date insect DNA methylation dataset,

encompassing holometabolous species from four orders (Hymenoptera, Coleoptera, Lepidoptera, and Diptera) and hemimetabolous species from three orders (Isoptera, Blattodea, and Hemiptera) published by Bewick et al. (2016).

We performed a PGLS analysis, to measure the strength of phylogenetic signal (following the definition by Revell et al 2011) between DNA methylation in protein-coding sequences and the mode of insect development (hemimetabolism or holometabolism). To measure phylogenetic signal we used Pagel's lambda (λ; Pagel 1999). In brief, a λ equal to one ($\boldsymbol{\lambda_1}$) corresponds to traits being as similar among species as expected from the phylogenetic tree, assuming a Brownian motion model of evolution. In contrast, a λ equal to zero ($\boldsymbol{\lambda_0}$) suggests species traits evolving independently from the phylogenetic tree. We estimated weak phylogenetic signal between DNA methylation and the mode of insect development ($\boldsymbol{\lambda_{ml}}$= 0.047). Most importantly, $\boldsymbol{\lambda_{ml}}$ was significantly different from $\boldsymbol{\lambda_1}$, but not significantly different from $\boldsymbol{\lambda_0}$ (Table S8). Thus, we can directly compare DNA methylation values between holometabolous and hemimetabolous insects as the traits in this dataset are independent from the given phylogeny. Similar to our CpG o/e comparisons, we found that holometabolous insects tend to display significantly lower DNA methylation levels in protein-coding sequences compared to hemimetabolous insects (Mann Whitney U test, $P < 0.001$;Figure 3.3.B).

Single-copy Genes Across all Insect Orders Show Signs of High DNA Methylation

Sarda and colleagues showed that most evolutionarily conserved genes tend to be highly methylated among four distantly related invertebrates (Sarda et al. 2012). We investigated whether there is a congruent pattern among insects. For this purpose, we analyzed a set of 1,478 clusters of nuclear-encoded protein-coding genes that have been retained in single-copy across insects and whose DNA sequences we obtained from the genomes and transcriptomes of 141 species representing all insect orders and other arthropods (Misof et al. 2014). For each transcriptome/official gene set we compared the CpG o/e distribution of all transcripts/genes with the CpG o/e distribution of the corresponding set of single-copy genes. We found that in species that possess methylation-indicative CpG o/e distributions, these single-copy genes tend to be overrepresented among low CpG o/e genes (Figure S2). To clearly display this relationship, we compared the median CpG o/e value of all transcripts/genes to the median CpG o/e value of the single-copy gene set of each species. Specifically, we selected a conservative set of taxa that according to our analysis and/or empirical evidence do not display signs of DNA methylation (i.e., lack of DNMT1 and DNMT3 accompanied by a CpG o/e distribution that does not indicate the presence of DNA methylation, or experimentally verified lack of CG DNA methylation from protein-coding sequences), namely Collembola, Strepsiptera, and Diptera (see Discussion), plus two beetles (Coleoptera), *Tribolium castaneum* and *Dendroctonus ponderosae*, and calculated a linear regression between the median CpG o/e values of all transcriptomes/official gene sets and the corresponding set of single-copy genes.

Using these taxa as reference, we found that in a number of species the calculated median CpG o/e value of the set of single-copy genes is significantly lower than the median CpG o/e value of the corresponding transcriptome/official gene set (Figure 3.4; Table S9). Overall, we found that genes that are consistently present across diverse insect lineages and possess highly conserved amino-acid sequences tend to exhibit low CpG o/e values, thus, high historical levels of germline DNA methylation.

3.4 Discussion

The Taxonomic Distribution of DNMTs in Insects

Our results suggest that DNMT1 was present in the last common ancestor of all insects and the last common ancestor of each extant insect order, except Collembola, Diptera, and Strepsiptera. Furthermore, our results are in agreement with previously published work on species of Diptera (reviewed by Glastad et al. 2011; Falckenhayn et al. 2016; Bewick et al. 2017) and Strepsiptera (Niehuis et al. 2012). The losses of DNMT1 in Collembola, Strepsiptera, and Diptera are certainly evolutionarily independent phenomena, since phylogenetic reconstructions rule out a close relationship among these lineages (Misof et al. 2014). We conclude that the loss of DNMT1 in insects is an evolutionarily rare event. In contrast, DNMT3 has been possibly lost numerous times during the evolutionary history of insects. Independent DNMT3 gains constitute an unlikely scenario for insects (Bewick et al. 2017). We did not identify DNMT3 in major insect groups such as Mecopterida, Palaeoptera, Neuropterida and most Polyneoptera (except Orthoptera and Isoptera). However, the absence of DNMT3 from the inspected transcriptomes could be attributed to low or

no expression of the corresponding gene. For example, we did not find DNMT3 in the transcriptome of the brown planthopper, *Nilaparvata lugens*, although it was shown that DNMT3 is weakly expressed in all life stages, but the mated and gravid females of this species (Zhang et al. 2013). In Mecopterida, our dense taxonomic sampling (40 species) combined with the availability of sequenced genomes provide congruent evidence for the loss of DNMT3 in this clade (Misof et al. 2014). Furthermore, Bewick and colleagues (2016) did not identify DNMT3 in the genomes of two palaeopteran species, in congruence with our results. The case is less clear in Neuropterida and Polyneoptera (excluding Isoptera and Orthoptera). In these clades, our species sampling per order is comparatively low and sequenced genomes were not yet published. To conclude, DNMT1 and DNMT3 do not constitute an indispensable functional pair in insects (in contrast to vertebrates), since the insect DNMT toolkit seems to be mainly comprised of DNMT1 homologs.

CpG o/e Patterns when DNMTs are Present

Based on CpG o/e distributions, it is reasonable to assume that species belonging to Trichoptera, Siphonaptera, Lepidoptera, Mecoptera (all belong to Mecopterida), Odonata, and Ephemeroptera (together form Palaeoptera) possess functional methylation systems despite the apparent loss of DNMT3. DNA methylation occurs in species belonging to 20 additional insect orders based on indicative CpG o/e distributions and DNMT3 complemented DNMT1 in just seven of them. Thus, our data indicate that DNA methylation is established and maintained without DNMT3 homologs in a possibly wide range of insect taxa. In Protura, Archaeognatha,

Dermaptera, Raphidioptera, and Neuroptera, only copies of DNMT1 were found in at least one species per order, but the corresponding CpG o/e distributions are not unequivocally pointing to the presence of DNA methylation. However, some insect species with experimentally verified DNA methylation at protein-coding sequences lack bimodal CpG o/e distributions despite the presence of either DNMT1 or both DNMT1 and DNMT3 (Glastad et al. 2011; Oxley et al. 2014; Libbrecht et al. 2016). Therefore, DNA methylation probably occurs at an even higher number of insect orders than the ones specified here.

The likely presence of CG methylation in protein-coding sequences of multiple insect taxa despite the absence of DNMT3 homologs, shows that the definition of a functional methylation toolkit needs to be redefined in insects. In certain species, like *B. mori* or the paper wasp *Polistes canadensis*, which possess a single DNMT1 homolog as their only identified DNA methyltransferase (Xiang et al. 2010; Patalano et al. 2015), it is possible that DNA methylation is introduced and maintained by this one enzyme (Maleszka 2016). However, in some insects, including multiple Hymenoptera and the human body louse, *P. humanus*, which also lacks DNMT3, more than one DNMT1 homologs are present (Glastad et al. 2011; Lyko and Maleszka 2011). Thus, certain DNMT1 paralogs may have shifted their function and are able to methylate *de nov*o and/or in contexts other than CG, similar to vertebrate DNMT3 enzymes. Another scenario is that a novel and currently unknown enzymatic machinery may be able to carry out DNA methylation in insects (Glastad et al. 2011; Maleszka 2016).

CpG o/e Patterns when DNMT1 and DNMT3 are Absent

It has been shown that the absence of DNMT1 and DNMT3 from the genomes of invertebrate species, including the dipteran insects, *Aedes aegypti, Aedes albopictus, Anopheles gambiae,* and *Drosophila melanogaster*, the nematode, *Caenorhabditis elegans*, and the trematode *Schistostoma mansonii,* is correlated with the absence or extreme reduction of DNA methylation (Simpson et al. 1986; Raddatz et al. 2013; Falckenhayn et al. 2016; Bewick et al. 2017). In line with this observation, we did not identify DNMT1, DNMT3, or methylation-indicative CpG o/e distributions in protein-coding sequences of species belonging to Collembola, Diptera, and Strepsiptera. Thus, since DNA methylation is predominantly found in CG context at protein-coding sequences across insects (Bewick et al. 2017), it is highly probable that species in these three orders lack or show extremely low levels of DNA methylation. Only TRDMT1 homolgs were identified in these species, reflecting the predicted absence of DNA methylation. The potential losses or extreme reductions of DNA methylation and its accompanying machinery in species belonging to three phylogenetically distinct insect lineages support the notion that DNA methylation might not be vital for the proper ontogenetic development of various insect species (Lyko and Maleszka 2011; Raddatz et al. 2013).

The Taxonomic Distribution of Tet Dioxygenases in Insects

Our results show that Tet dioxygenases are widely distributed across insects, since we identified homologs in species belonging to most insect orders. The underrepresentation of putative Tet homologs in transcriptomes compared to genomes

can be attributed to low or no expression of the Tet gene and hence its absence from the analyzed transcriptomes. The identification of Tet homologs in the genome, but not the transcriptome of the springtail *Folsomia candida* or the mountain pine beetle *Dendroctonus ponderosae* substantiate this idea. The presence of Tet homologs in species belonging to Collembola and Diptera, in which according to our analyses and/ or experimental evidence (Bewick et al .2016) DNA methylation is extremely reduced or absent is in line with the proposed multifunctional role of Tet enzymes in insect genomes (Maleszka 2016). In Collembola, Diptera, or other insects in which DNA methylation is extremely reduced or absent, Tet homologs may act as 6mA DNA demethylases and/or 5mC mRNA demethylases, similar to their roles in *D. melanogaster* (Zhang et al. 2015; Delatte et al. 2016). Thus, the presence of Tet enzymes in insects may not be strictly correlated to its most designated function, that is, 5mC DNA demethylation.

The Presence of DNA Methylation is Ancestral to Insects

The identification of a complete DNMT toolkit and the presence of methylation-indicative CpG o/e distributions in crustaceans show that DNA methylation is probably ancestral to insects. The absence of DNMTs from the transcriptome of the remipede *Xibalbanus tulumensis* should be considered a limitation of this specific transcriptomic dataset, since the species shows signs of heavy CpG depletion of protein-coding sequences, while no other remipede species was examined. Thus, the potential losses or extreme reductions of DNA methylation and its machinery from insect groups are secondary, lineage-specific events (Glastad et al. 2011). This pattern

shows that DNA methylation is a dispensable epigenetic mechanism for insects and its function may be compensated by other molecular mechanisms (Glastad et al. 2011; Raddatz et al. 2013).

DNA Methylation has Been Reduced in Holometabola

The sparse presence of DNA methylation observed in holometabolous species and comparative analyses between two holometabolous insects (*Apis mellifera* and *Bombyx mori*) and two other invertebrates (*Nematostella vectensis* and *Ciona intestinallis*) (Sarda et al. 2012) led to the hypothesis that the levels of DNA methylation may have been reduced in the ancestors of insects (Glastad et al. 2014). However, our comparative analysis, combined with experimental evidence from single-species studies, point to a different scenario: the heavy CpG depletion of protein-coding sequences observed in the majority of species belonging to Zygentoma, Palaeoptera, Polyneoptera, and to a lesser extent Condylognatha, suggest that DNA methylation levels have been reduced in the ancestors of Holometabola, while there is no indication that DNA methylation levels were already reduced in the ancestors of insects. Our analysis of published empirical methylation data (Bewick et al. 2017) backs this hypothesis. Furthermore, empirical evidence obtained from direct measurements of DNA methylation in Orthoptera (*Schistocerca gregaria, Locusta migratoria*), Phasmatodea (*Medauroidea extradetata*), and Isoptera (*Zootermospis nevadensis*) and computational evidence from analyzing Isoptera (*Zootermopsis nevadensis, Coptotermes lacteus, Reticulitermes flavipes*) support this conclusion. These polyneopteran species are, in comparison to holometabolous insects,

characterized by significantly elevated levels of DNA methylation (Krauss et al. 2009; Falckenhayn et al. 2013; Glastad et al. 2013; Terrapon et al. 2014; Wang et al. 2014; Glastad et al. 2016). Alternatively, the high mean CpG o/e values of Psocodea, the proposed sister group of Holometabola (Misof et al. 2014), suggest a reduction in the levels of DNA methylation that already occurred in the last common ancestor of Psocodea and Holometabola.

Evolutionary Conservation of Genes is Strongly Associated with DNA Methylation in Insects

We showed that a set of single-copy genes that are associated with housekeeping functions (Misof et al. 2014) and have orthologs in all insects tend to display signatures of heavy DNA methylation in species with evident historical germline methylation. Our result is in line with those of previous investigations showing that the majority of orthologs among four distantly related invertebrates are extensively methylated (Sarda et al. 2012) and reveals that most evolutionarily conserved housekeeping genes have been strongly methylated throughout insect evolution.

The evolutionary interconnection between DNA methylation and housekeeping genes may have a functional explanation. Bird (1995) conjectured that intragenic methylation may reduce transcriptional noise (high transcript variability) by suppressing spurious transcription initiation in vertebrate genomes. Both points of this hypothesis have recently received support by studies on mammalian systems. First, Huh and colleagues found that transcriptional noise is reduced in heavily methylated human genes (Huh et al. 2013). Second, Neri and colleagues showed that

DNMT3-dependent intragenic DNA methylation acts to prevent spurious transcription initiation in mouse cells (Neri et al. 2017). Reducing transcriptional noise could be especially beneficial for constitutively expressed housekeeping genes (Suzuki et al. 2007). Thus, it is likely that intragenic DNA methylation acts to reduce transcriptional noise on evolutionarily conserved housekeeping genes in insects, perhaps with a mechanism similar to the one described by Neri et al. (2017). However, since many insect species that show signs of intragenic DNA methylation seem to lack DNMT3 homologs, a DNMT3-independent enzymatic machinery would contribute to a noise reduction mechanism in certain insects.

3.5 Conclusions

Our results provide an invaluable resource for experimental studies designed towards continuing this line of work. Experimental tests designed for investigating the functional role of DNMT1 homologs should be applied, by employing, for example, RNAi and/or CRISPR/Cas based methods, especially in DNMT3-deficient species. Additionally, large scale comparative studies using direct measurements of DNA methylation, such as whole genome bisulfite sequencing, should be conducted. Applying such approaches will not only aid in estimating the levels of DNA methylation in certain lineages, but also in determining the genomic targets of DNA methylation with accuracy, which in turn may provide important insights towards understanding its function in insects.

3.6 Figures

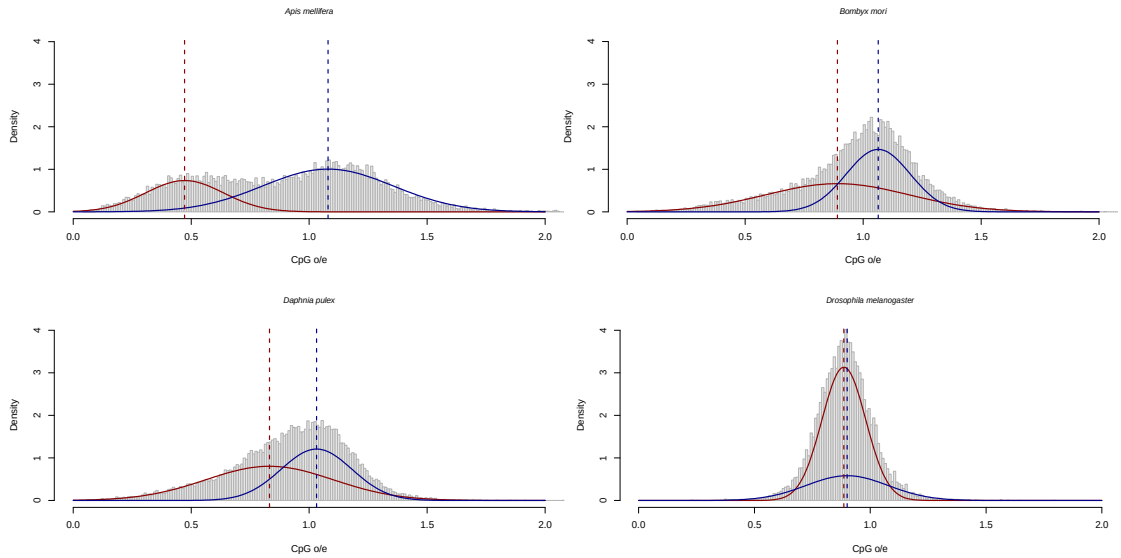


Figure 3.1 Distinct types of CpG o/e distributions in protein-coding sequences of four insect species. A mixture of two Gaussian distributions was fitted to the data using mclust (v. 5.2). Dark red and dark blue dashed lines correspond to the means of each fitted distribution ($mean_{low}$ and $mean_{high}$, respectively). (A) *Apis mellifera* displays a clearly bimodal CpG o/e distribution, with one component displaying low CpG o/e values (sequences mostly affected by CpG depletion) and the other one high (sequences less affected by CpG depletion). We describe the CpG o/e distribution of *A.mellifera* as "bimodal depleted" since difference between the component means is higher than 0.25 ($mean_{high}$-$mean_{low}$= 0.61), while the low CpG o/e component has a mean lower than 0.7 ($mean_{low}$= 0.47). (B, C) *Bombyx mori* and *Daphnia pulex* lack clearly defined bimodality ($mean_{high}$-$mean_{low}$< 0.25 and $mean_{low}$> 0.7 in both cases), but their low CpG o/e component displays a characteristic extensive tail, which contains a significant proportion of data (0.36 and 0.43, respectively). We describe distributions that lack clearly defined bimodality similar to *B.mori* and *D.pulex*, but their smallest component contains a significant proportion of data ($proportion_{low}$=0.36 or higher) as "unimodal, indicative of DNA methylation". (D) Finally, *Drosophila melanogaster,* which is almost devoid of DNA methylation from protein-coding sequences, displays a clearly unimodal CpG o/e distribution with two component means being almost identical ($mean_{high}$-$mean_{low}$= 0.004), show no signs of significant CpG depletion ($mean_{low}$=

0.886), and the proportion of data belonging to the smallest component is very low ($proportion_{low}$=0.087 < 0.36). We describe the CpG o/e distribution of *D. melanogaster* as “unimodal, not indicative of DNA methylation”.

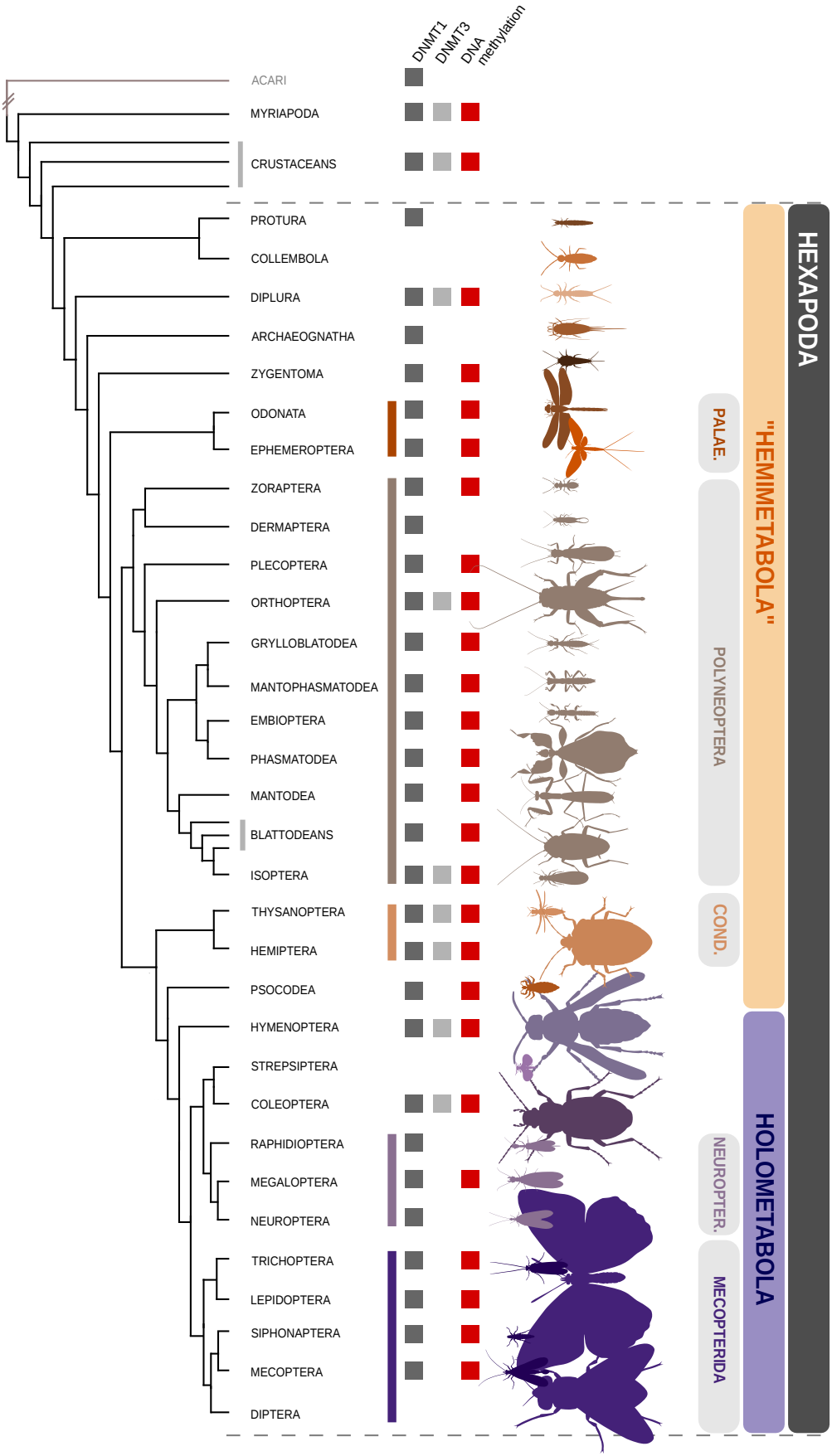

DNMT1
DNMT3
DNA methylation
ACARI
MYRIAPODA
CRUSTACEANS
PROTURA
COLLEMBOLA
DIPLURA
ARCHAEOGNATHA
ZYGENTOMA
ODONATA
EPHEMEROPTERA
ZORAPTERA
DERMAPTERA
PLECOPTERA
ORTHOPTERA
GRYLLOBLATODEA
MANTOPHASMATODEA
EMBIOPTERA
PHASMATODEA
MANTODEA
BLATTODEANS
ISOPTERA
THYSANOPTERA
HEMIPTERA
PSOCODEA
HYMENOPTERA
STREPSIPTERA
COLEOPTERA
RAPHIDIOPTERA
MEGALOPTERA
NEUROPTERA
TRICHOPTERA
LEPIDOPTERA
SIPHONAPTERA
MECOPTERA
DIPTERA
PALAE.
POLYNEOPTERA
COND.
NEUROPTER.
MECOPTERIDA
"HEMIMETABOLA"
HOLOMETABOLA
HEXAPODA

Figure 3.2: Occurrence of DNA methyltransferases and DNA methylation in investigated species. We plotted the presence of DNA methyltransferases (DNMT1, DNMT3) on a phylogram representing the phylogenetic relationships among all investigated species. Additionally, we plotted the presence of DNA methylation as inferred by the CpG o/e distributions of investigated species on this phylogram (DNMT1: dark gray, DNMT3: light gray, DNA methylation: red). The phylogenetic relationships of depicted insect orders and outgroups are congruent with the proposed relationships in Misof et al. (2014). DNMT1 is found in species belonging to all insect orders except in Collembola, Strepsiptera and Diptera. DNMT3 was only identified in seven insect orders. Methylation-indicative CpG o/e distributions were identified in species belonging to 24 insect orders plus crustacean and myriapod species. PALAE. Palaeoptera; COND. Condylognatha; NEUROPTER. Neuropterida.

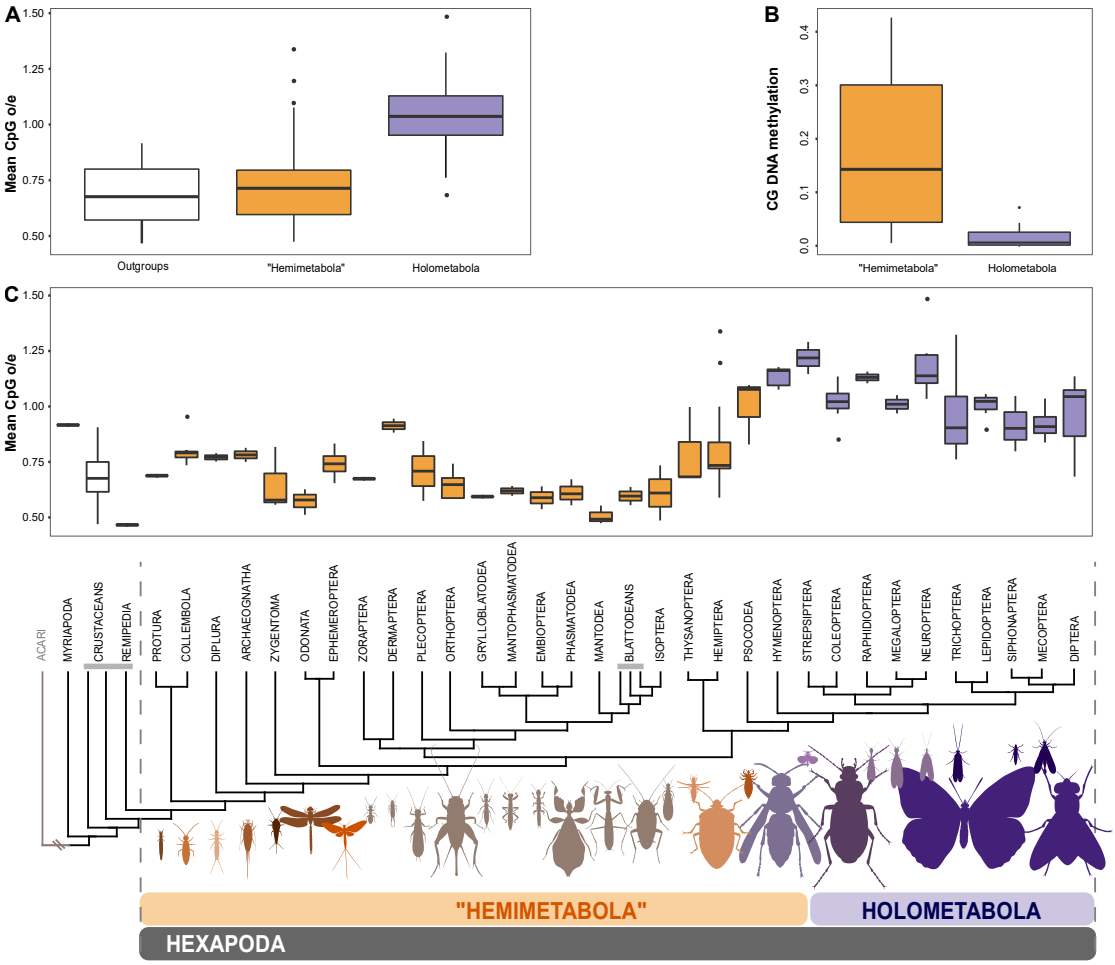


Figure 3.3: Comparison of inferred DNA methylation levels across insects. (A) Comparison of mean normalized CpG dinucleotide content (CpG o/e) among species belonging to Holometabola (52 species, violet box plot), Hemimetabola (67 species, orange box plot) and other arthropod outgroups (8 species, white box plot) based on investigated transcriptomes (127 species in total, representing all currently recognized insect orders plus crustacean and myriapod outgroups). We tested whether the difference of mean CpG o/e values among Hemimetabola, Holometabola, and outgroups was significant with a Kruskal-Wallis H test ($P < 0.001$). (B) Comparison of CG DNA methylation levels between the protein-coding sequences of 14 hemimetabolous and 26 holometabolous insects species. Holometabolous species display lower levels of DNA methylation in protein-coding sequences compared to hemimetabolous species (Mann-Whitney U test P< 001). (C) Comparison of mean CpG o/ e values of species described in (A) separated by insect order. The CpG o/e levels strongly vary among

insect orders, but orders of Holometabola show higher overall mean CpG o/e values than orders belonging to hemimetabolous insects.

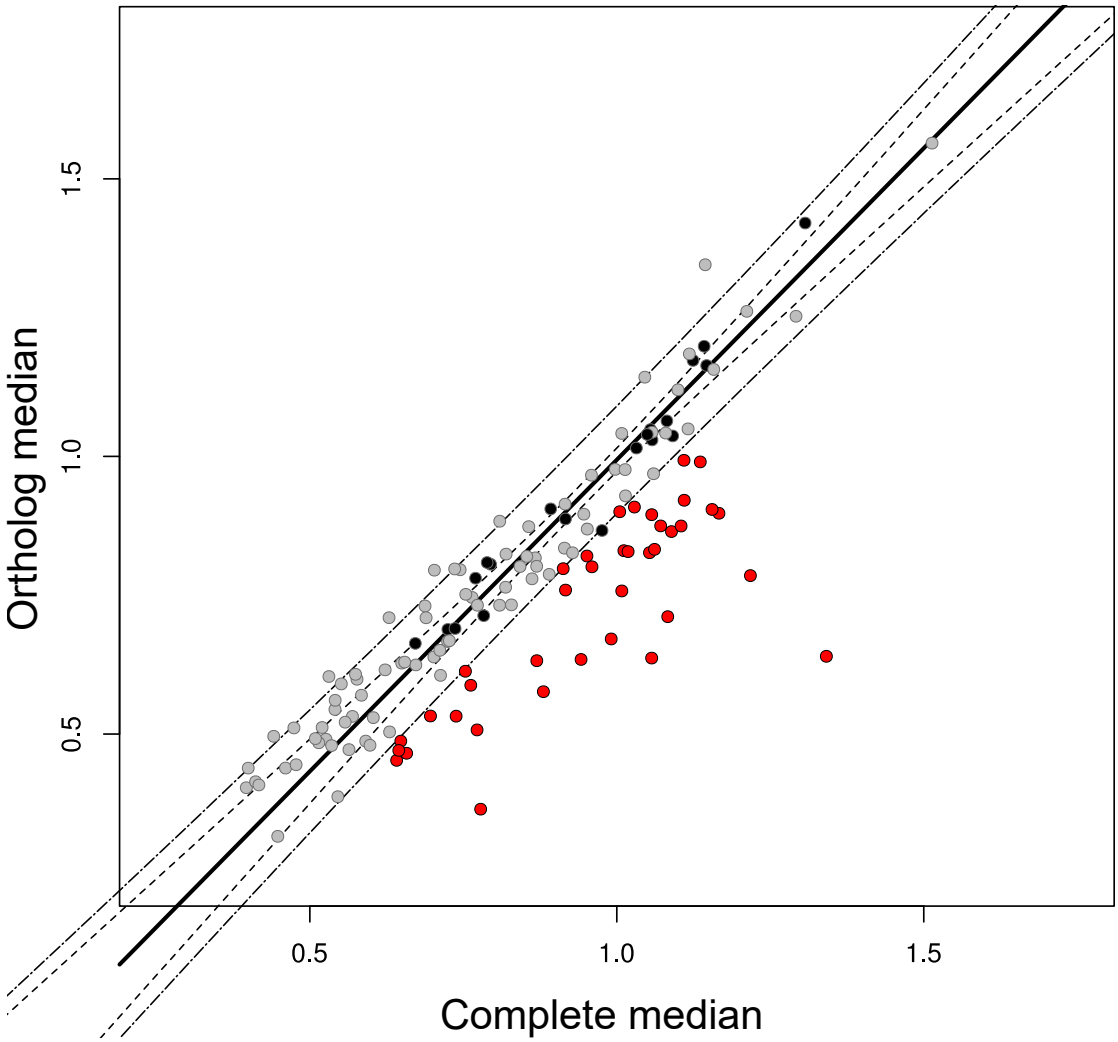


Figure 3.4: Comparison of the median CpG o/e value of all transcripts/genes of a transcriptome/official gene set (complete median) with the median CpG o/e value of a subset of 1,478 single-copy genes with orthologs across 141 insect and other arthropod species (ortholog median). Black dots indicate species with no signs of DNA methylation according to our analysis and/or experimental evidence (species from the orders Collembola, Strepsiptera, Diptera, plus two beetles, *Dendroctonus ponderosae* and *Tribolium castaneum*). Based on the median CpG o/e values of these species, we calculated a linear regression (black solid line). The black dashed lines indicate the confidence intervals and the black

dash-dotted lines indicate the prediction intervals that were calculated based on this regression. Species in which the median CpG o/e of single-copy genes is significantly lower than the median CpG o/e of the transcriptomic or genomic background are colored red (dots below the lower dash-dotted line). The remaining species are shown in gray.

3.7 References

Bewick AJ, Vogel KJ, Moore AJ, Schmitz RJ. 2017. Evolution of DNA methylation across insects. *Molecular Biology and Evolution*. 34, 654-665.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*. 8, 1499–1504.

Bird AP. 1995. Gene number, noise reduction and biological complexity. *Trends Genetics*. 11, 94–100.

Bonasio R et al. 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science*. 329, 1068–1071.

Bonasio R et al. 2012. Genome-wide and caste-specific DNA methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Current Biology. 22, 1755–1764.

Camacho C et al. 2009. BLAST plus: architecture and applications. *BMC Bioinformatics*. 10:1.

Cunningham CB et al. 2015. The genome and methylome of a beetle with complex social behavior, *Nicrophorus vespilloides* (Coleoptera: Silphidae). *Genome Biology and Evolution*. 7, 3383-3396

Delatte B et al. 2016. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science*. 351, 282–285.

Elango N, Hunt BG, Goodisman MAD, Yi SV. 2009. 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, 11206–11211.

Falckenhayn C et al. 2013. Characterization of genome methylation patterns in the desert locust *Schistocerca gregaria*. *Journal of Experimental Biology*. 216, 1423–1429.

Falckenhayn C et al. 2016. Comprehensive DNA methylation analysis of the *Aedes aegypti* genome. *Scientific Reports*. 6, 36444.

Feng S et al. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*. 107, 8689–8694.

Finn RD et al. 2014. Pfam: The protein families database. Nucleic Acids Res. 42:222–230.

Foret S, Kucharski R, Pellegrini M. 2012. DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. *Proceedings of the National Academy of Sciences*. 109, 4968–4973.

Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. *Molecular Biology and Evolution*. 22, 650–658.

Glastad KM, Gokhale K, Liebig J, Goodisman MAD. 2016. The caste- and sex-specific DNA methylome of the termite *Zootermopsis nevadensis*. *Scientific Reports*. 6, 37110.

Glastad KM, Hunt BG, Goodisman MAD. 2014. Evolutionary insights into DNA methylation in insects. *Current Opinion in Insect Science*. 1, 25–30.

Glastad KM, Hunt BG, Goodisman MAD. 2013. Evidence of a conserved functional role for DNA methylation in termites. *Insect Molecular Biology*. 22, 143–154.

Glastad KM, Hunt BG, Yi SV, Goodisman MAD. 2011. DNA methylation in insects: On the brink of the epigenomic era. *Insect Molecular Biology*. 20, 553–565.

Goll MG, Bestor TH. 2005. Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry*. 74, 481–514.

Goll MG et al. 2006. Methylation of tRNA. *Science*. 311, 395–398.

Huh I, Zeng J, Park T, Yi SV. 2013. DNA methylation and transcriptional noise. *Epigenetics & Chromatin.* 6, 9.

Hunt BG, Glastad KM, Yi SV, Goodisman MAD. 2013. Patterning and regulatory associations of DNA methylation are mirrored by histone modifications in insects. *Genome Biology and Evolution*. 5, 591–598.

Jeltsch A, Jurkowska RZ. 2014. New concepts in DNA methylation. *Trends in Biochemichal Sciences*. 39, 310–318.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Natural Review Genetics*. 13, 484–492.

Kao D et al. 2016. The genome of the crustacean *Parhyale hawaiensis*, a model for animal development, regeneration, immunity and lignocellulose digestion. e*LIFE*. 5, e20062.

Kapheim KM et al. 2015. Genomic signatures of evolutionary transitions from solitary to group living. *Science*. 348, 1139–1143.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*. 30, 772–780.

Krauss V, Eisenhardt C, Unger T. 2009. The Genome of the Stick Insect *Medauroidea extradentata* Is Strongly Methylated within Genes and Repetitive DNA. *PLoS One*. 4, e7223.

Kriventseva EV et al. 2014. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research*. 43, D250–D256.

Kucharski R, Maleszka J, Foret S, Maleszka R. 2008. Nutritional control of reproductive status in honeybees via DNA methylation. *Science.* 319, 1827–1830.

Libbrecht R, Oxley PR, Keller L, Kronauer DJC. 2016. Robust DNA Methylation in the Clonal Raider Ant Brain. *Current Biology*. 26, 391-395.

Lyko F. 2017. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nature Reviews Genetics*. doi:10.1038/nrg.2017.80

Lyko F, et al. 2010. The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. *PLoS Biology*. 8, e1000506.

Lyko F, Maleszka R. 2011. Insects as innovative models for functional studies of DNA methylation. *Current Opinion in Insect Science*. 27, 127-31.

Maleszka R. 2016. Epigenetic code and insect behavioural plasticity. *Current Opinion in Insect Science*. 15, 45–52.

Misof B et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science.* 346, 763–767.

Munoz-Torres MC et al. 2010. Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Research*. 39, D658–D662.

Neri F, Rapelli S. 2017. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*. 543, 72–77.

Niehuis O et al. 2012. Genomic and morphological evidence converge to resolve the enigma of Strepsiptera. *Current Biology*. 22, 1309–1313.

Oxley PR et al. 2014. The genome of the Clonal raider ant *Cerapachys biroi*. *Current Biology*. 24, 451–458.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature.* 401, 877–884.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics.* 20, 289–290.

Park J et al. 2011. Comparative analyses of DNA methylation and sequence evolution using *Nasonia* genomes. *Molecular Biology Evolution*. 28, 3345–3354.

Pastor WA, Aravind L, Rao A. 2013. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature Reviews Molecular Cell Biology*. 14, 341–356.

Patalano S et al. 2015. Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. *Proceedings of the National Academy of Sciences*. 112, 13970–13975.

Petersen M et al. 2017. Orthograph: A versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics*. 18, 1–10.

Pinheiro JC, Bates DJ, DebRoy SD, Sarkar D and R Core Team. 2017. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131, https://CRAN.R-project.org/package=nlme.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raddatz G et al. 2013. Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proceedings of the National Academy of Sciences*. 110, 8627–8631.

Rehan SM, Glastad KM, Lawson SP, Hunt BG. 2016. The genome and methylome of a subsocial small carpenter bee, *Ceratina calcarata*. Genome Biology and Evolution. 8, 1401-1410.

Revell LJ, Harmon LJ, Collar DC, Oakley T. 2008. Phylogenetic Signal, Evolutionary Process, and Rate. *Systematic Biology*. 57, 591–601.

Sarda S, Zeng J, Hunt BG, Yi SV. 2012. The evolution of invertebrate gene body methylation. *Molecular Biology and Evolution*. 29, 1907–1916.

Schübeler D. 2015. Function and information content of DNA methylation. *Nature*. 517, 321–326.

Simola DF et al. 2013. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Research*. 23, 1235–1247.

Simpson VJ, Johnson TE, Hammen RF. 1986. *Caenorhabditis elegans* DNA does not contain 5-methylcytosine at any time during development or aging. *Nucleic Acids Research*. 14, 6711–6719.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 6, 31.

Standage DS et al. 2016. Genome, transcriptome, and methylome sequencing of a primitively eusocial wasp reveal a greatly reduced DNA methylation system in a social insect. *Molecular Ecology*. 25, 1769-1784.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*. 34, 609–612.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. Nature Reviews Genetics. 9, 465–476.

Suzuki MM, Kerr ARW, De Sousa D, Bird A. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Research*. 17, 625–631.

Terrapon N et al. 2014. Molecular traces of alternative social organization in a termite genome. *Nature Communications*. 5, 3636.

Wang X et al. 2014. The locust genome provides insight into swarm formation and long-distance flight. *Nature Communications*. 5, 2957.

Wang X et al. 2013. Function and Evolution of DNA Methylation in *Nasonia vitripennis*. *PLoS Genetics*. 9, e1003872.

Wang Y et al. 2006. Functional CpG methylation system in a social insect. *Science*. 314, 645–647.

Werren JH et al. 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science*. 327, 343–348.

Wojciechowski M et al. 2014. Insights into DNA hydroxymethylation in the honeybee from in-depth analyses of TET dioxygenase. *Open Biology*. 4, 140110.

Xiang H et al. 2010. Single base-resolution methylome of the silk moth reveals a sparse epigenomic map. *Nature Biotechnology*. 28, 516–520.

Yi SV, Goodisman MAD. 2009. Computational approaches for understanding the evolution of DNA methylation in animals. *Epigenetics.* 4, 551–556.

Zdobnov EM et al. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research.* 45, 744–749.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science.* 328, 916–919.

Zhang G et al. 2015. N6-methyladenine DNA modification in *Drosophila*. *Cell.* 161, 893–906.

Chapter 4

Evolutionary Trajectories of DNA Methylation in Insects

This chapter is intended for publication in *Nature Ecology and Evolution*.

Authors: Provataris P., Gatzmann F., Petersen M., Oeyen J.P., Vasilikopoulos A., Donath A., Niehuis O., Nosil P., de Carvalho C., Zdobnov E., Legrand C., Raddatz G., Lyko F., Misof B.

Author contributions: Project Design: PP, FL, BM. Analyses: PP, FG, MP, GR, JPO, AV. Manuscript design and writing: PP, FL, BM. Supplementary material: PP, ON, AD. Statistical analyses: PP, CL. Genome data: ON, AD, PN, CC, EZ, BM.

4.1 Introduction

DNA methylation, the addition of a methyl (-CH3) group to a genomic cytosine by a DNA methyltransferase (DNMT), is the most well-studied epigenetic modification across the tree of life. Large-scale comparative studies across eukaryotes (Feng et al 2010; Zemach et al. 2010) and across large eukaryotic groups like plants (Takuno et al. 2016; Niederhuth et al. 2016; Bewick et al. 2017a) or fungi (Bewick et al. 2019) have been facilitated by the increasing availability of efficient whole genome sequencing approaches. A common conclusion among these studies is that DNA methylation is evolutionarily dynamic, as its distribution across genomes (Feng et al 2010; Zemach et al. 2010; Bewick et al. 2019) or the amount of methylation targeted to specific genomic elements (Takuno et al. 2016; Niederhuth et al. 2016; Bewick et al. 2017) varies. However, the mechanisms that drive DNA methylation change between species, and thus, the evolutionary trajectories of DNA methylation remain poorly understood.

Our detailed knowledge about evolutionary relationships in insects (Misof et al. 2014) provides a unique opportunity to reconstitute evolutionary trajectories. Glastad and colleagues (2014) concluded that insect genomes are generally lowly methylated compared to other animals. Additionally, other authors pointed towards the observation that studied insect genomes are characterized by sparse DNA methylation that is targeted to a subset of exons, whereas introns and repetitive elements are hypomethylated (Lyko et al. 2010; Xiang et al. 2010; Bonasio et al. 2012; Wang et al. 2013; Cunningham et al. 2015; Patalano et al. 2015; Libbrecht et al. 2016; Rehan et al. 2016; Standage et al. 2016; Glastad et al. 2019). In contrast, we previously

suggested that sparse methylation of protein-coding sequences is a trait of Holometabola, but cannot be observed in all insects, as hemimetabolous insects[1] were predicted to possess significantly higher levels of DNA methylation in protein-coding sequences (Provataris et al. 2018). These *in silico* inferences can be corroborated by limited experimental evidence (Falckenhayn et al. 2013; Glastad et al. 2016; Bewick et al. 2017b; Provataris et al. 2018). Thus, the evolutionary history of DNA methylation in insects might substantially differ from the currently prevailing views.

In animals, DNA methylation is established by DNMT3 and maintained by DNMT1, although their roles often overlap (Jeltsch and Jurkowska 2016). The functionality of DNMTs is important for normal mammalian development and DNMT mutations are linked to a number of human diseases, including cancer (Lyko 2017). In contrast to mammals, it was previously shown that several insects maintain functional methylation systems with DNMT1 homologs as the only identifiable DNA methyltransferases, while DNMT3 homologs have been frequently lost across insects (Bewick et al. 2017b; Provataris et al. 2018). Based on these findings, we hypothesized that the exclusive presence of DNMT1 is necessary and can be sufficient for a functional DNA methylation system in insects (Provataris et al. 2018). However, the manner in which DNMT1 homologs are able to perform both *de novo* and maintenance methylation is currently unknown.

Comparative analyses on the evolution of DNA methylation in insects performed to date either focused on a taxonomically restricted set of species (species from 6

1 We will be referring to hemimetabolous insects also as Hemimetabola, despite them not being a monophyletic group.

orders, mostly Holometabola) without exploring the distribution of DNA methylation across the studied genomes (Bewick et al. 2017b) or lacked experimental data (Provataris et al. 2018). To overcome these obstacles, we compiled a dataset that covers the taxonomic breadth of insects. More specifically, we generated deeply-covered methylomes of 25 species, mostly focusing on understudied hemimetabolous insect groups. Pairing the generated data with publicly available ones, we compiled a dataset comprising methylomes of 46 species from 18 insect orders, thus massively expanding pre-existing taxonomic coverage. Our results show that hemimetabolous insects possess higher levels of DNA methylation compared to holometabolous insects, particularly in introns and repetitive elements. Furthermore, by adopting a novel in-depth approach for analyzing gene body methylation patterns, we found that virtually all studied insect methylomes can be classified in four distinct patterns and that the evolutionary transition between these patterns is likely driven by methylation of intragenic repeat sequences. Finally, we suggest that duplication and subsequent modification of functional domains of DNMT1 homologs may be the key towards establishing and maintaining DNA methylation patterns in absence of DNMT3 for many insects. By combining all our findings, we infer a mechanistic model for the evolutionary transitions of DNA methylation across insects.

4.2 Results

DNA Methylation Levels are Lower and Less Variable in Holometabola Compared to Other Insects

To systematically analyze DNA methylation across insects we generated or acquired WGBS data of 46 species from 18 out of the 31 existing insect orders in an effort to maximize taxonomic representation. More specifically, our dataset consisted of 19 hemimetabolous insect species from 11 orders and 27 holometabolous insect species from 7 orders (table 4.S1). We compared the average methylation levels among species belonging to these two insect groups and found that holometabolous insects show significantly lower levels of DNA methylation compared with hemimetabolous insects, both genome-wide and within gene bodies (figure 4.1.A; figure 4.S1; table 4.S2). Furthermore, Holometabola showed significantly lower variation in the levels of DNA methylation compared with hemimetabolous insects (figure 4.S1). The highest genome-wide levels of DNA methylation in Hemimetabola were found in the two-pronged bristletail *Campodea augens* (>20%). Genome-wide methylation levels were consistently high in grasshoppers and crickets (Orthoptera), stick insects (Phasmatodea), ice crawlers (Grylloblattodea), cockroaches and termites (Blattodea) (ranging between 7% and 14%). The lowest genome-wide levels of DNA methylation in Hemimetabola were observed in lice (Psocodea), as predicted by our previous computational assessment (Provataris et al. 2018). Conversely, global levels of DNA methylation were consistently low in Holometabola (~1% or less), with the exception of three distantly related species; the sawfly *Neodiprion lecontei* (3.4%; Hymenoptera), the Asian long-horned beetle *Anoplophora glabripennis* (8.4%;

Coleoptera), and the cat flea *Ctenocephalides felis* (2.6%; Siphonaptera). Finally, DNA methylation was lost or became extremely reduced in several distantly related taxa including springtails (Collembola), the twisted-wing parasite *Stylops ovinae*, and the scorpionfly *Panorpa germanica*, in addition to the previously reported losses/extreme reductions in the paper wasp *Polistes dominula*, the red flour beetle *Tribolium castaneum*, and dipterans (figure 4.1.A; Zemach et al. 2010; Standage et al. 2016; Bewick et al. 2017b). Thus, in spite of the apparent variability in the levels of DNA methylation across insects, Holometabola display consistently lower levels of DNA methylation compared with other insects.

The Four Main Patterns of Gene Body Methylation in Insects

It has been proposed that gene bodies are the primary targets of DNA methylation in insects, with exons displaying considerably higher levels of DNA methylation compared to introns (Glastad et al. 2014). Our analysis yielded a more differentiated picture: Although gene body methylation was enriched in the exons of holometabolous species with appreciable methylation levels, a number of distantly related hemimetabolous species showed higher methylation levels in introns (figure 4.1.A; table 4.S2). This suggests that the enrichment of DNA methylation in exons is not a common trait across all insects as previously considered (Glastad et al. 2014), but rather a characteristic trait of Holometabola.

For a more detailed analysis of gene body methylation patterns, we compared gene body methylation levels among the genes of each species. When comparing the average methylation levels of CG dinucleotides located in the exons of a gene to the

average methylation levels of CGs located in the introns of the same gene we identified four main patterns of gene body methylation (figure 4.1.B; figure 4.S2). The first pattern is characterized by bimodal gene body methylation, with high levels of exon and intron methylation in one gene group and low levels of exon and intron methylation in the other (figure 4.1.B, dark green color). This pattern, which we will be referring to as Body Methylation (BM), is restricted to hemimetabolous insects. The second pattern is defined by low intron methylation and variable exon methylation (figure 4.1.B, dark blue color). It is the dominant gene body methylation pattern in Holometabola, also being present in Psocodea, the proposed sister group of Holometabola (figure 4.1.A, dark blue color; Misof et al. 2014; Johnson et al. 2018). This is the pattern that has often been considered as prototypical for insects and we will be referring to it as Exon Methylation (EM) pattern. The third pattern combines features of BM and EM (figure 4.1.B, light green color): one group of genes shows high gene body methylation levels in both exons and introns (similar to BM), whereas the second group of genes is characterized by lowly methylated introns and variably methylated exons (similar to EM). We will be calling this the Mixed Methylation (MM) pattern. MM was mostly present in hemimetabolous insects that show exon-enriched DNA methylation (figure 4.1.A), with the exception of lice. It was also found in the three distantly related holometabolous species with the highest levels of gene body methylation, namely *N. lecontei*, *A. glabripennis*, and *C. felis*. Finally, insects in which DNA methylation was extremely reduced or absent were characterized by a single cluster of extremely lowly methylated or unmethylated gene bodies, which we named UM (lowly methylated or Unmethylated). This pattern

appears in springtails (Collembolla), primarily unwinged hexapods branching off at the base of the insect tree, and in four different groups within Holometabola (Hymenoptera, Coleoptera, Strepsiptera, and Diptera). A comparison of the patterns revealed that genome-wide, gene body, and intron methylation levels were high in BM, intermediate in MM, and low in EM (figure 4.S.3), whereas exon methylation levels where similarly high in BM and MM and lower in EM species. This shows that increased methylation levels in introns and to a lesser extent in exons contribute to high global methylation levels.

DNA Methylation is Targeted to Repetitive Elements in Hemimetabola

In contrast to gene bodies, transposable and other repetitive DNA elements have not been identified as consensus targets of DNA methylation in insects (Glastad et al. 2014). However, previous studies lacked systematic taxonomic coverage and mostly focused on Holometabola. We annotated repetitive elements *de novo* and profiled their methylation levels across insects. We found that repeat sequences were much more highly methylated in Hemimetabola compared to Holometabola (figure 4.S.4). Furthermore, repeat sequences, similar to gene bodies, were on average more highly methylated than the genomic average (positive DNA methylation enrichment) in Hemimetabola, but unlike gene bodies, they were more lowly methylated than the genomic average in Holometabola (negative DNA methylation enrichment) (figure 4.2.A, figure 4.S5). Since we observed that intragenic repeat sequences showed much higher methylation levels compared with intergenic repeat sequences across insects (figure 4.S6), we compared methylation at intergenic and intragenic repeat sequences

separately between the two insect groups. Our results showed that both intragenic and intergenic repeat sequences showed positive and significantly higher DNA methylation enrichment in hemimetabolous insects compared to Holometabola (figure 4.S7). In holometabolous insects only intragenic repeat sequences were negatively enriched for DNA methylation, as intergenic repeat sequences were similarly methylated to the genomic average. These findings suggest that repeat sequences are targeted by DNA methylation in many hemimetabolous insects, but tend to be hypomethylated in the gene bodies of holometabolous insects.

The Impact of Intragenic Repeat Sequences on the Patterns of Gene Body Methylation

We compared DNA methylation levels at intragenic and intergenic repeat sequences among species with different patterns of gene body methylation. We found that insects with a bimodal gene body methylation pattern (BM) showed positive and significantly higher DNA methylation enrichment at both intragenic and intergenic repeat sequences compared with insects with an exon-enriched (EM) or a mixed methylation (MM) pattern (figure 4.S8.A). Furthermore, in BM insects, intergenic repeat sequences showed higher median DNA methylation enrichment compared to intragenic repeat sequences. In EM and MM species, intergenic repeat sequences were similarly methylated to the genomic average, but intragenic repeat sequences showed negative DNA methylation enrichment, particularly in EM species (figure 4.S8.A). In summary, insects with the highest levels of gene body methylation (BM insects) also

exhibited the highest DNA methylation enrichment at repeat sequences (figure 4.2.B; figure 4.S8.B).

To further explore how the presence of repeat sequences is associated with the patterns of gene body methylation, we compared methylation levels between exons that contain repeat sequences and exons that are free of repeat sequences and repeated the same comparison for introns. We found that in BM insects, repeat-containing exons and introns were similarly methylated to repeat-free exons and introns, respectively (figure 4.2.C; figure 4.S9.A and 4.S9.B). In EM insects, however, the absence of repeat sequences was associated with significantly higher levels of DNA methylation in both exons and introns (figure 4.2.C; figure 4.S9.A and 4.S9.B). Despite the fact that exons, in general, display higher average levels of DNA methylation than introns in EM species, we found that repeat-containing exons were on average more lowly methylated than repeat-free introns (figure 4.S9.C). This is possible because the vast majority of intronic CG dinucleotides of EM species are located in virtually unmethylated, repeat-containing introns (figure 4.S9.D) Finally, MM insects once again display a mixed phenotype as the presence of intragenic repeat sequences did not affect the methylation levels of exons, but was associated with significantly lower levels of intron methylation (figure 4.2.C). Thus, the increasing levels of gene body methylation from EM to MM to BM insects are associated with the increasing presence of methylated intragenic repeats.

We previously observed that the highly methylated gene group of MM species mirrors the highly methylated gene group of BM species, whereas the variably methylated gene group of MM species mirrors the variably methylated gene group of

EM species. Based on this, we hypothesized that exons and introns in each of the gene groups of MM species would be affected in a similar fashion by the presence of intragenic repeats. To test this hypothesis, we analyzed methylation levels of repeat-containing and repeat-free exons/introns in each gene group. In the high gene body methylation group of MM species, repeat-containing exons and introns were similarly methylated to repeat-free exons and introns, respectively (figures 4.S10.A. and 4.S10.B). The same was true for BM species. Thus, similar to BM species, the methylation status of high gene body methylation genes was not affected by the presence of intragenic repeat sequences in MM species. When performing the same comparison for the variable exon/low intron methylation gene group of MM species, we found that, one one hand, repeat-containing introns had significantly lower methylation levels than their repeat-free counterparts, similar to EM species. On the other hand, repeat-containing exons were similarly methylated to repeat-free exons, as in BM species (Figure 4.S10.A and 4.S10.B). In summary, the main difference identified between BM and MM species is that in the latter lowly methylated introns tend to possess substantially lower methylation levels when they contain repeats. In addition, we observed that these extremely lowly methylated, repeat-containing introns host around three quarters of the total genic CG content in MM species. Comparatively, lowly methylated, repeat-containing introns host only half of the genic CGs in BM species, a quarter less than the introns of MM species. The remaining quarter of CGs is located in highly methylated repeat-containing introns in BM species (Figure 4.S10.C). Thus, the main difference identified when shifting from a

BM to a MM pattern is that a number of repeat-rich genes have vastly decreased levels of intron methylation.

Gene Body Methylation Allows for Higher Incorporation of Repeats in Introns

We previously showed that gene body methylation is strongly and positively correlated with repeat methylation in insects. BM species, which have the highest levels of gene body methylation, also show the highest intragenic methylation of repeats, whereas EM species have the lowest amounts of gene body methylation and intragenic repeats are strongly hypomethylated. By studying mostly vertebrate animals and a few invertebrate outgroups, Zhou and colleagues (2020) recently suggested that increased amounts of genomic methylation allows for efficient incorporation of TE-derived DNA. Based on the previous, we hypothesized that a larger proportion of gene bodies in BM species will contain repeats, compared to MM and EM species. In UM species, in which gene bodies are practically unmethylated, the majority of gene bodies should be repeat free. However, because repeats are much more likely to be found in introns rather than exons, we inspected exons and introns separately for each pattern type. When inspecting exons we found that approximately 90% of all exons regardless pattern type were repeat-free. In contrast, the majority of introns in BM species contained repeats (~80%, median), whereas in UM species only a small proportion of introns contained repeats (~25%, median). We found a strong negative correlation between average levels of gene body methylation and the proportion of introns that contain repeats across all four gene body methylation

patterns (Figure 4.S11). Consequently, we suggest that higher levels of gene body methylation allow for more frequent incorporation of repeats in intronic regions.

The Ancestral State of Insect DNA Methylation

We used ancestral state reconstruction methods in order track the evolutionary history of the examined traits across the phylogeny of insects. Our analysis revealed that the Last Insect Common Ancestor (LICA) possessed high global levels of DNA methylation similar to most hemimetabolous insects, whereas a substantial reduction occurred in the last common ancestor of Holometabola (Figure 4.3.A). Genome-wide levels of DNA methylation have been further reduced in all major holometabolous lineages, with only few secondary increases in the three holometabolous species that exhibit a MM gene body methylation pattern (*N. lecontei*, *A. glabripennis*, and *C.felis*). Gene body methylation levels followed an overall similar evolutionary history compared to genome-wide levels (figure 4.S12). Exon methylation levels were relatively high in the LICA and were further increased in most Hemimetabola with a BM or a MM pattern (Figure 4.3.A). Exon methylation followed an opposite trajectory in Holometabola. While exon methylation levels remained similar between the last common ancestor of Holometabola and the LICA, they were substantially reduced in each major holometabolous insect group examined, with further extreme reductions towards the terminal taxa. The only exceptions to this pattern were the three holometabolous MM species. The methylation levels of introns and repeat sequences were characterized by relatively high levels of DNA methylation in the LICA, with subsequent further increases of DNA methylation in certain BM species

and subsequent decreases in MM species, in Hemimetabola. In Holometabola, introns and repeat sequences followed a nearly identical evolutionary history to each other and to the genome-wide levels of DNA methylation (Figure 4.3.A).

Further analysis also revealed that the LICA likely possessed a BM pattern of gene body methylation, which is characterized by high levels of DNA methylation in gene bodies and repeats, whereas the last common ancestor of Holometabola showed an EM pattern, which is characterized by sparse, exon-targeted DNA methylation and repeat hypomethylation. The MM pattern has independently emerged from or reverted to the BM pattern in Hemimetabola and from the EM pattern in Holometabola, multiple times. No direct transition from the BM to the EM pattern or vice versa was revealed by our analyses. Thus, the MM pattern of gene body methylation likely constitutes a transitional state between the BM and EM patterns (Figure 4.3.B, Figure 4.S13). Finally, the UM pattern has appeared 5 times independently across insects, and four of these times it has emerged from the EM pattern in Holometabola. These findings illustrate the evolutionary trajectories of DNA methylation in insects.

Dynamic Evolution of the DNMT Toolkit in Insects

We previously found that DNMT1 is widely conserved across insects, whereas DNMT3 has been frequently lost (Provataris et al. 2018). Indeed, 41 out of 46 species in our dataset contained DNMT1 homologs, whereas only 21 of these species also contained DNMT3 homologs (table 4.S3). Furthermore, we and others suggested that DNMT1 is necessary and can be sufficient for a functional methylation system in insects, since DNA methylation was present in groups that appeared to have lost

DNMT3 (Bewick et al. 2017b; Provataris et al. 2018). In our current analysis, DNMT1 was the only detectable DNA methyltransferase in 16 out of the 37 species that showed a defined pattern of DNA methylation and we identified nine independent emergences of such DNMT1-only methylation systems (figure 4.4.B). DNMT1 duplication events accompany DNMT3 losses in most of these lineages (table 4.S3).

Analyzing the domain composition of DNMT1 enzymes across insects, we found in several species that one of the DNMT1 paralogs lacks the zf-CXXC autoinhibitory domain (Figure 4.4A). This domain prevents DNMT1 from performing *de novo* methylation and ensures that only hemimethylated CpG sites undergo methylation (Song et al., 2011). Thus, a loss of this domain could unlock the *de novo* methylation potential of DNMT1 (Song et al., 2011). Based on our gene tree reconstruction, an activated DNMT1 (aDNMT1) emerged independently in seven distantly related insect groups from a duplication of the zf-CXXC-containing DNMT1 enzyme (Figure 4.4B). The emergence of aDNMT1 has likely occurred even more frequently within insects, considering that it is present in species belonging to seven out of 14 insect orders with defined DNA methylation patterns. In five of these instances aDNMT1 emerged after DNMT3 loss, but it has coexisted with DNMT3 in both lice and the parasitoid wasp *N. vitripennis*. Our findings point to an evolutionary scenario under which a duplication and subsequent reconfiguration of DNMT1 has compensated for the loss of DNMT3.

4.3 Discussion

Redefining Methylation Targets of Insects

Based on the assumption that DNA methylation is primarily targeted to exons in insects (Glastad et al. 2014), we previously hypothesized that holometabolous insects possess lower levels of DNA methylation compared to other insects, as we inferred comparatively higher methylation levels in protein-coding sequences of hemimetabolous insects (Provataris et al. 2018). We showed that our hypothesis was accurate, however our assumption was not. The vast majority of previously characterized insect methylomes that led to this assumption were holometabolous species that display an EM pattern according to our classification (Lyko et al. 2010; Xiang et al. 2010; Bonasio et al. 2012; Wang et al. 2013; Cunningham et al. 2015; Patalano et al. 2015; Rehan et al. 2016; Glastad et al. 2017). When analyzing such methylomes in more detail, we observed that DNA methylation was primarily targeted to exons and introns that do not contain repeats. However, the most substantial differences regarding the targeting of DNA methylation were observed in insects with BM or MM pattern, which comprise most Hemimetabola. In these species a number of gene bodies and not just exons, are the main targets of DNA methylation. Finally, although sporadic evidence for high levels of intron methylation has been documented previously in three hemimetabolous species (Glastad et al. 2016; Bewick et al. 2019), evidence for consistent methylation of transposable elements (TEs) and other repeats in insects was lacking (Glastad et al. 2019). We showed that in hemimetabolous species with a BM pattern, although repeats were on average more lowly methylated than genes, intragenic repeats were more highly methylated than genes and intergenic

repeats were more highly methylated than intergenic regions. This finding places repeats as major targets of DNA methylation in hemimetabolous insects and shows that it makes more sense to assess repeat methylation in a context-dependent manner for more accurate characterization of the phenomenon in organisms with mosaic methylation.

Transitional States of Gene Body Methylation in Insects

In order to understand how DNA methylation expands and contracts during the evolutionary history of insects we first need to establish transitional states among the different patterns of DNA methylation we presented. Based on our ancestral state reconstruction analysis, we identified the MM pattern as a transitional state between the BM and EM patterns. However, there exist additional observations that support this assumption. First, MM species possess intermediate genome-wide, gene body, and repeat methylation levels compared BM and EM insects. Second, MM insects show high levels of gene body methylation in one group of genes similar to BM insects, and exon-enriched DNA methylation in another, similar to EM insects. Depending on the relative size of each gene group, certain MM species in our dataset resemble EM or BM species more than others. For example, in the thrips *Frankliniella occidentallis* and the cat flea *C. felis*, the group of high gene body methylation genes is substantially smaller compared to the long-horned beetle *A. glabripennis* and the bed bug *C. lectularius*. In a similar fashion, the termites are more similar to BM species compared to the long-horned beetle and the bed bug. Additionally, EM-like MM species have lower genome-wide and gene body

methylation levels and show weaker targeting of DNA methylation to repeats compared to BM-like MM species. Last but not least, MM species show a mixed pattern of intragenic repeat methylation, as repeat-rich exons are similarly methylated to repeat-free exons, similar to BM species, whereas repeat-rich introns are on average much more lowly methylated than repeat-free introns, similar to EM species. Thus, we suggest that starting from an ancestral hexapod with high levels of gene body and repeat methylation insects had to go through an intermediate phase during which the targeting of DNA methylation primarily to introns and repeats weakened, before attaining the typical, sparse, exon-targeted DNA methylation pattern found in majority of extant Holometabola.

After establishing why MM is an intermediate state between BM and EM, we need to place UM among the other three patterns. Our ancestral state reconstructions show that the UM pattern emerged once from the BM pattern in the base of the insect tree and another four times from the EM pattern in Hymenoptera and Mecopterida. However, a transition from the EM model is much more likely, since it requires a small amount of methylation loss in a restricted number of exons. For example, from the two paper wasps in our dataset (*P.canadensis* and *P. dominula*), *P. canadensis* shows a very slight methylation enrichment in exons compared to other Hymenoptera, while in *P. dominula* exon-enrichment is practically non-detectable (Patalano et al 2015; Standage et al. 2016). These two species, which are the only Hymenoptera that lack DNMT3 in our dataset, are examples of a transition from an EM pattern to a UM pattern. Additionally, both Hymneoptera and Mecopterida are represented by a substantial amount of species in our dataset, whereas basal hexapods are only

represented by three species whose last common ancestor is dated to more than 450 million years ago (Misof et al. 2014), and it is likely that with additional sampling a direct transition from BM to UM would disappear. Thus, we suggest that a loss of exon-targeted methylation coinciding with a disruption of a functional DNMT toolkit results in extreme methylation reduction or even complete loss in insects.

A Mechanistic Model for the Expansion and Contraction of Gene Body Methylation in Insects

As we have now established the order of transition among methylation patterns, we can propose a mechanism on how these transitions occur. We suggest that shifting from a BM to a MM pattern involves the insertion of a large amount of repeats in one or more introns of a gene belonging to the high gene body methylation group. These newly inserted intronic repeats are not methylated and carry a large amount of repeat-associated CG dinucleotides. Thus, the average level of intron methylation for that gene substantially lowers, causing the gene to move to the low intron methylation gene group. This phenomenon also contributes to the decrease of the average methylation levels at intragenic repeats observed in MM species. At the same time, since we showed that exons tend to be repeat-free across insects regardless of gene body methylation pattern, the exons of that gene will remain largely repeat-free and will thus retain their original, high levels of exon methylation, creating the high exon-low intron part of the variable exon/ low intron methylation gene group of MM species. The repetition of this process for many genes would result to a substantial reduction in the levels of gene body methylation. Finally, the shrinkage of the high

gene body methylation gene group of MM species limits their ability to assimilate TE-derived DNA compared to BM species, gradually leading a larger proportion of their introns to remain repeat-free. Thus, as intron methylation contracts, the ability to retain TE-derived DNA declines.

The transition from the MM pattern to the EM pattern occurs through three main processes. First, is the continuation of the process we described when transitioning from the BM to the MM pattern. The high gene body methylation gene group of MM species continues to shrink through the insertion of repeats in introns until it is completely devoid of genes. Second, is a process that becomes apparent in the EM state, but is likely initiated during the MM state. DNA methylation levels are substantially lowered in exons that contain repeats. This happens as newly inserted repeat sequences in exons, though rare, tend to not be methylated as they generally do in BM and MM species. The result is an increase in the CG content of these exons and a reduction their average methylation levels. Thus, intragenic repeats in EM species tend to reside in regions with very low levels of DNA methylation, resulting in the strong hypomethylation we observed in EM species. Finally, EM species also experience a substantial drop in genome-wide levels of DNA methylation that also severely affects gene bodies. This general decrease of DNA methylation affects the ability of EM gene bodies to incorporate repeat-derived DNA, leading to a further increase in the proportion of introns that are of free repeats compared to MM species.

The final transition from the EM pattern to the UM pattern happens through a complete loss or extreme reduction of DNA methylation, even from repeat-free exons and introns and is accompanied by a complete loss of DNMTs that retain DNA

methylation-related functions. The complete loss of DNA methylation means that DNA methylation can no longer facilitate the incorporation of repeat sequences in introns and thus introns tend to remain repeat-free in UM species.

DNA Methylation Follows Contrasting Evolutionary Trajectories in Insects and Vertebrates

Zhou and colleagues (2020) recently proposed that the integration of TEs led to increased methylation levels and genomic expansion in higher order vertebrates. This happened as the methylation of newly inserted TEs restricted their harmful effects, thus facilitating their genomic integration and also induced local hypermethylation of adjacent host DNA. However, it has been shown that TE insertions may have the opposite effect. Unmethylated, CG-rich, repeat sequences may induce local hypomethylation of flanking host DNA (Grandi et al. 2015). Thus, it is plausible that during the evolutionary history of insects, novel TE insertions that were not methylated, led to further local hypomethylation of host DNA, a phenomenon that may have substantially added to the TE-assisted reduction of DNA methylation described in our model. Our findings point to contrasting evolutionary trajectories of DNA methylation between insects and vertebrates that are shaped by the same force: the interplay between TEs and DNA methylation.

DNMT1 Establishes and Maintains DNA Methylation in Absence of DNMT3

The paradoxical absence of DNMT3 from functional DNA methylation systems of insects has been a puzzling question for long (Glastad et al. 2011; Lyko and Maleszka 2011). The most common hypotheses attempting to explain this paradox posit that an unknown enzyme unrelated to DNA methyltransferases carries out *de novo* methylation or that a DNMT1 gene is compensating for DNMT3 loss (Lyko and Maleszka 2011; Glastad et al. 2014; Maleszka 2016; Bewick et al. 2017b; Provataris et al. 2018). Recent comparative studies revealed that such DNMT1-only methylation systems are common in insects (Bewick et al. 2017b; Provataris et al. 2018) and our current study adds to this finding, because we found that a DNMT1-only toolkit was almost as common as a DNMT1-DNMT3 toolkit. However, what was lacking from previous studies was the frequency of DNMT1 duplication and evidence for a plausible mechanism under which DNMT1 could perform *de novo* methylation. We filled this gap by showing that DNMT1 duplication and DNMT3 loss are equally frequent and typically coincide, and that most of the DNMT1 paralogs in species lacking DNMT3 have been modified in a way that likely unlocks their capacity to methylate *de novo*. This modified version of DNMT1 was recently hypothesized to be associated with increased TE methylation, since it was found in two arthropod species in which TEs are the prime targets of DNA methylation (Lewis et al. 2020). However, we recovered aDNMT1 in seven distantly related insect species and in none of these TE methylation is dominant. On the contrary, in the lice (*P. humanus* and *L. bostrychophila*), the jewel wasp, *N. vitripennis*, and the cat flea, *C. felis*, we found

that TEs and other repeats are hypomethylated. In summary, since there exist no other enzymes known to date that are able to establish and maintain 5mC other than DNMTs, it is more plausible that DNA methylation is established and maintained by DNMT1 homologs in insects that have lost DNMT3.

4.4 Conclusions

The expansion of DNA methylation at the vertebrate-invertebrate boundary has been linked with multiple evolutionary innovations that characterize vertebrate systems. DNA methylation-mediated transcriptional noise reduction (Huh et al. 2013; Neri et al. 2017; Gatzmann et al. 2018; Liew et al. 2018) has been linked to the increased gene number of vertebrates (Bird 1995; Prachumwat and Li 2008). The emergence of promoter methylation has been deemed a novel mechanism of gene expression regulation (Tweedie et al 1997; Keller et al. 2016). Finally, TE repression via DNA methylation was proposed to have facilitated genomic expansion and the generation of novel regulatory sequences in vertebrates (Zhou et al. 2020). Our work suggests that DNA methylation has undergone a contraction at the hemimetabolous-holometabolous insect boundary. Thus, insects constitute an exemplary model group of animals for identifying the effects that DNA methylation contraction has on genome evolution.

4.5 Material and Methods

Specimen Acquisition and DNA Extraction

Specimens of 25 insect species destined for WGBS were obtained from established laboratory colonies with few exceptions. Alternatively, they were field collected or commercially purchased (table 4.S4). All specimens were stored in ethanol and kept in minus 20 degrees Celsius prior to DNA extraction. For large enough specimens heads and/or thoraxes, but not guts were used for extracting DNA to decrease the possibility of contamination (table 4.S4). Genomic DNA was isolated using the Blood and Cell Culture DNA Kit (Qiagen, Hilden, Germany).

Whole-Genome Bisulfite Sequencing

Genomic DNA was isolated as described above. As a spike-in control, unmethylated bacteriophage lambda DNA was used (table 4.S3). The TruSeq PCR-Free Library Prep Kit (LT; Illumina, San Diego, US) was used for library preparation and the Epitect Kit (Qiagen) for bisulfite conversion. Library amplification was performed using the Kapa HiFi HotStart Uracil + ReadyMix (2 ×; Kapa Biosystems). Samples were then sequenced on an Illumina HiSeq platform.

Data Acquisition

In addition to the in-project sequenced methylomes, raw data for 21 additional insect methylomes, genome assemblies used for the mapping of all WGBS data, and corresponding annotations of protein-coding genes were obtained from various

sources (table 4.S1). Further information on genome sequencing, assembly, annotation of protein-coding genes, and annotation of repeat sequences done for this projects are available at the supplementary material and methods.

Whole-Genome Bisulfite Sequencing Data Analysis

Read pairs were quality trimmed (minimum quality value ≥ 15 and minimum length ≥ 36bp) using Trimmomatic version 0.35 (Bolger et al. 2014). Data for all species were mapped using BSMAP version 2.73 (Xi and Li 2009). Only read pairs whose both reads mapped uniquely and with appropriate orientation and distance between each other were used for downstream analyses. Reads identified as sequencing duplicates were excluded from the calculation of methylation ratios. Methylation ratios were determined using the Python script distributed with BSMAP. Read counts were merged between strands so that each covered CG dinucleotide in the genome assembly was represented by a single methylation ratio. CG sites covered by less than four reads were excluded from all methylation level calculations in an effort to reduce the effect of errors related to WGBS. DNA methylation levels for each CG site were calculated by dividing the number of methylated reads to the total number of methylated plus unmethylated reads. For the analysis of the patterns of gene body methylation, we filtered for deeply-covered protein-coding genes. Only genes that contained a minimum of 5 CG dinucleotides per 1Kb of DNA and for which 50% of their CG dinucleotides fulfilled our coverage criteria were included in this analysis. All graphs were produced in R (R Core team 2019) with the exception of the ancestral

state reconstruction graphs for discrete characters which were produced by PASTML (Ishikawa et al. 2019).

Identification of DNA Methyltransferases

To identify DNA methyltransferases in the genomes of the insect species included in this study we followed the pipeline described by Provataris et al. 2018. In brief, we used previously constructed hidden Markov Models (pHMMs) of DNMT1 and DNMT3 (available at doi: 10.17632/8y5wm8887b.3) and searched the set of predicted proteins of each species using hmmsearch with default options (HMMER 3.2.1; www.hmmer.org). To determine whether candidate sequences were correctly identified as DNMT1 or DNMT3 we used blastp (BLAST+ 2.8.0) and searched with them against *Nasonia vitripennis* OGS v 2.0 (Munoz-Torres et al. 2011). Subsequently, all candidate sequences that did not match a corresponding *Nasonia* DNMT as best hit where excluded. Finally, we searched remaining candidate sequences against the non-redundant NCBI database and excluded candidate sequences whose best matches were of bacterial origin. To distinguish between isoforms and paralogs for each DNMT1 and DNMT3 candidate sequence, we made sure that each enzyme we classified as a paralog, originated from a different genomic location.

To identify aDNMT1 enzymes, we scanned all DNMT1 candidate sequences against the PFAM-A database v.31 (El-Gebali et al. 2018) using hmmscan (HMMER 3.2.1; www.hmmer.org). For three species (*Ephemera danica*, *Galloisiana yuasai*, and *Timema cristinae*) the exact domain content of one of the DNMT1 copies could not

identified due to gaps in the genome assembly. Consequently, we could not identify whether these enzymes were aDNMT1, and we classified them as DNMT1.

Phylogenetic Analyses

We used the BUSCO *v2* ortholog set comprised of 1,066 groups of orthologous genes that are single-copy across Arthropoda (Simao et al. 2015) as input for Orthograph. We used Orthograph version 0.6.3 (Petersen et al. 2017) to identify orthologs of the 1,066 single-copy genes in the 46 species used in our DNA methylation analyses. To speed up the best reciprocal hit step of Orthograph, we used a subset of 16 species included in the BUSCO v2 set, which are representative of all major arthropod lineages (Supplementary file 1). For downstream analyses, we only used the amino acid sequence output of Orthograph. Orthologous sequences were aligned with MAFFT version 7.310 using the option *l-insi* (Katoh and Stanley 2013). All multiple sequence alignments (MSAs) were assessed for quality (and modified if necessary) and masked using the procedure described by Misof et al. 2014 (although all identified outliers were removed without performing refinement and realignment steps). Subsequently, MSAs were concatenated into a supermatrix using FasConCat (Kueck and Meusemann 2010). The phylogenetic information content of each gene partition in the supermatrix was assessed using MARE version 0.1.2-rc (Misof et al. 2013) and all uninformative partitions were removed. We constrained MARE by forcing it to retain all taxa while removing partitions. Optimal models of sequence evolution were selected with ModelFinder (Kalyaanamoorthy et al. 2017). Phylogenetic tree reconstruction was carried out using IQ-TREE version 1.6.11

(Nguyen et al. 2014). We used an “edge-proportional” partition scheme allowing each gene partition its own specific evolutionary rate (Chernomor et al. 2016). We conducted 10 independent tree searches and for five of them we used a random starting tree. Statistical support for each node was assessed using the bootstrap method (Felsenstein et al. 1985). The best tree was chosen based on the likelihood score as outputted by IQ-TREE. We repeated the tree search step by constraining the phylogenetic relationships among Odonata, Ephemeroptera, and Neoptera to reconstruct the Palaeoptera hypothesis on the origin of winged insects (Pterygora) and the relationships among Hemiptera, Thysanoptera, Psocodea, and Holometabola to reconstruct Condylognatha (Hemiptera + Thysanoptera) being a sister group to the clade formed by Psocodea plus Holometabola (Misof et al. 2014; Johnson et al. 2018). We did this to present more comprehensive evolutionary scenarios when reconstructing the ancestral state of insect DNA methylation (see **ancestral state reconstruction analyses).**

To reconstruct a gene tree of identified insect DNMT1 (Supplementary file 2) sequences we first made sure that only one isoform of each enzyme remains in our dataset. Prior to alignment, we used PREQUAL (Whelan et al. 2018) to mask regions with non-homologous adjacent characters. The partial DNMT1 sequences of *Timema cristinae* and *Panorpa germanica* were removed after the PREQUAL step because almost the whole length of their aminoacid sequence was masked and did not therefore bear any phylogenetic information content. Homologous aminoacid sequences were aligned using FSA (Bradley et al. 2009). We opted for FSA for this analysis because its increased accuracy compared to mafft would not come at the cost

of speed, as we were only reconstructing a single-gene phylogeny. Randomly similar sections were identified using AliScore (version 2.2; Misof and Misof, 2009) and removed using AliCut (version 2.3; https://github.com/PatrickKueck/AliCUT/). Model selection, phylogenetic reconstruction, node statistical support, and tree selection were all carried out exactly as described for the species tree.

Ancestral State Reconstruction Analyses

To carry out ancestral state reconstruction analyses, two different software packages were utilized. For discretely coded characters we used PASTML and applied all three available maximum likelihood methods under the F81 model (Ishikawa et al. 2019). More specifically, for the ancestral state reconstruction of the insect DNMT toolkit, DNMT1 and DNMT3 were simply coded as present or absent at each terminal node. The four patterns of gene body methylation were numerically coded (zero to three) at each terminal node. To reconstruct the global levels of DNA methylation, the levels of gene body methylation, and the enrichment of DNA methylation in repeat sequences (continuous characters) we used the R package phytools (Revel et al 2011). Each ancestral state reconstruction analysis was carried out twice using both the constrained and the unconstrained phylogenetic trees in order to reflect alternate evolutionary scenarios for groups whose phylogenetic relationships are not fully resolved to date (the origin of Pterygora and the placement of Psocodea).

4.6 Figures

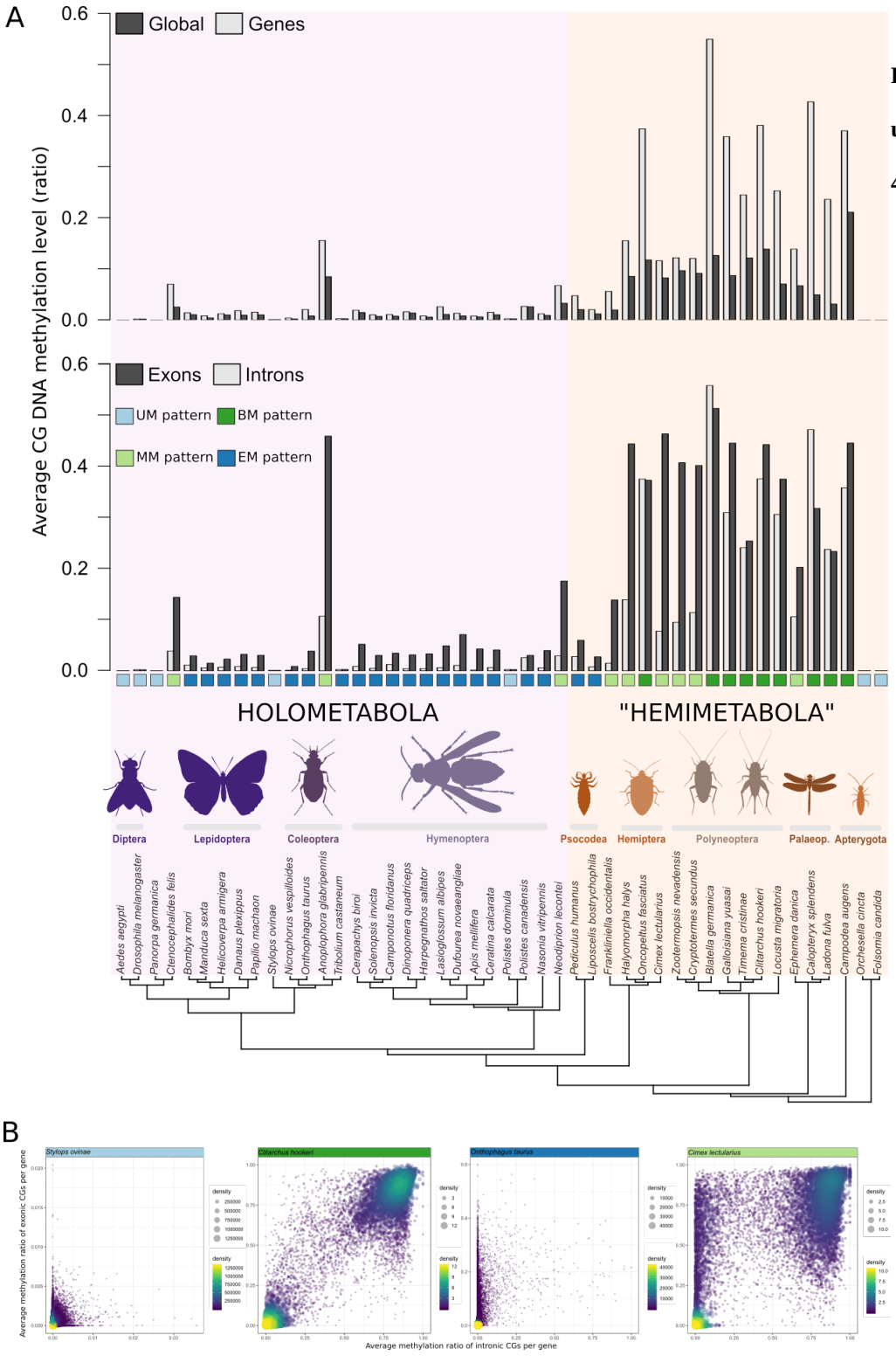


Figure 4.1: Genomic levels and patterns of DNA methylation across insects. A) Hemimetabolous insects show

substantially higher levels of DNA methylation compared to Holometabola both within genes and genome-wide. Gene body methylation is characterized by increased methylation of introns in hemimetabolous insects. B) By comparing the average methylation of exons to the average methylation of introns for each gene we identified four main patterns of gene body methylation across insects (from left to right: Lowly methylated or Unmethylated (UM) represented by *Stylops ovinae*, Body Methylation (BM) represented by *Clitarchus hookeri*, Exon Methylation (EM) represented by *Onthophagus taurus*, and Mixed Methylation represented by *Cimex lectularius*). The BM and MM patterns are overrepresented in Hemimetabola, whereas the EM pattern is dominant in Holometabola.

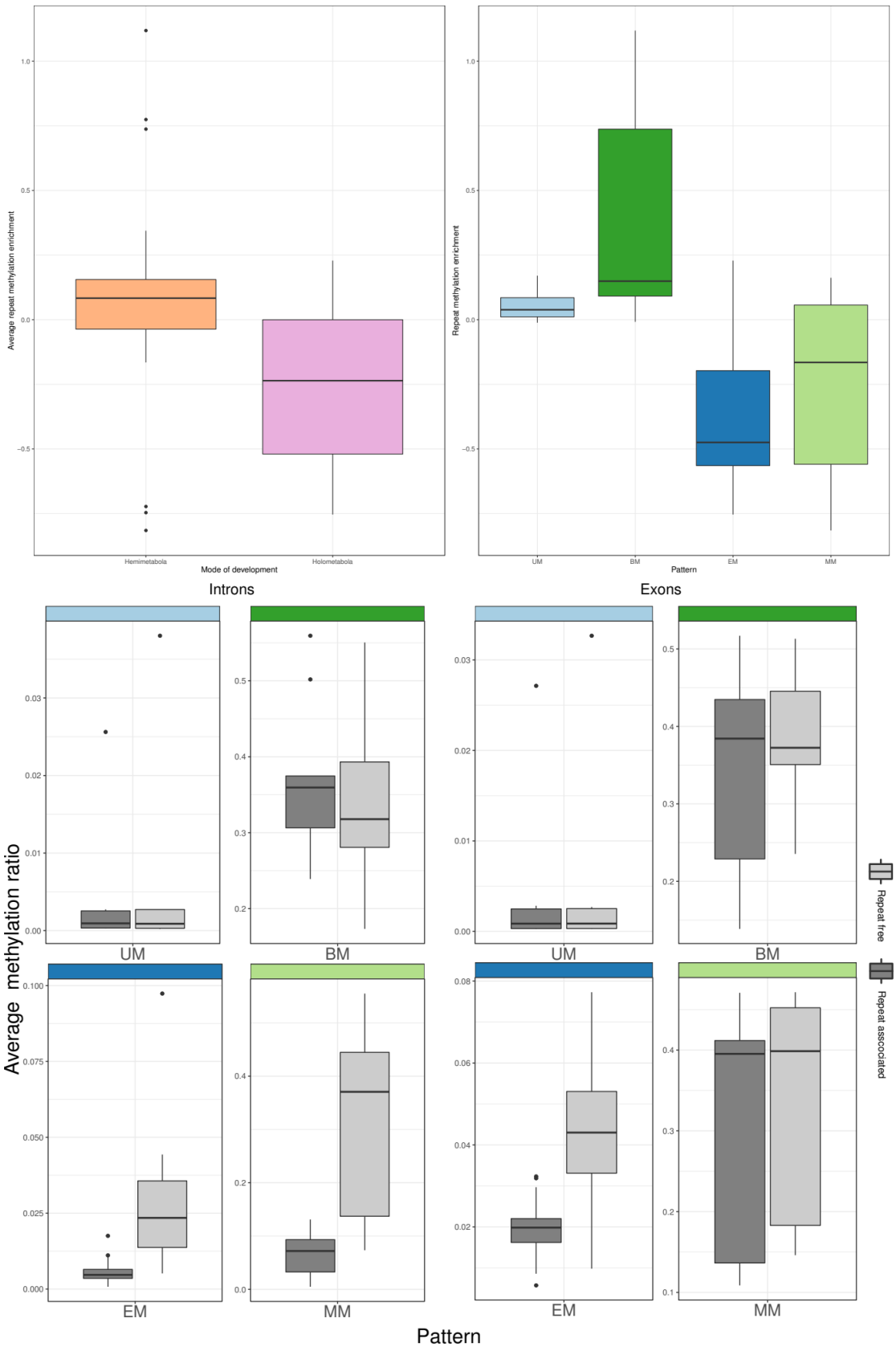

Average repeat methylation enrichment
1.0
0.5
0.0
−0.5
Hemimetabola
Holometabola
Mode of development
Repeat methylation enrichment
UM
BM
EM
MM
Pattern
Introns
Exons
Average methylation ratio
Repeat free
Repeat associated
Pattern

Figure 4.2: DNA methylation at repetitive elements. A) In hemimetabolous insects repetitive elements show higher levels of DNA methylation compared to the genome average, whereas in Holometabola repetitive elements tend to be hypomethylated (Wilcoxon rank sum test, p-value=0.00225). B) In insects displaying a BM gene body methylation pattern repetitive elements tend to be more highly methylated than the genome average, whereas in MM and EM insects repetitive elements tend to be hypomethylated (Kruskal Wallis *H* test, p-value =0.0002319). C) Comparison between the average methylation ratios of repeat-free and repeat-associated exons/introns among the four patterns of gene body methylation (BM: Wilcoxon signed-rank test, p-value exons= 0.4258, p-value introns=0.6523; EM: Wilcoxon signed-rank test, p-value exons= p-value<0.005, p-value introns<0.005; MM: Wilcoxon signed-rank test, p-value exons=0.123, p-value introns<0.005).

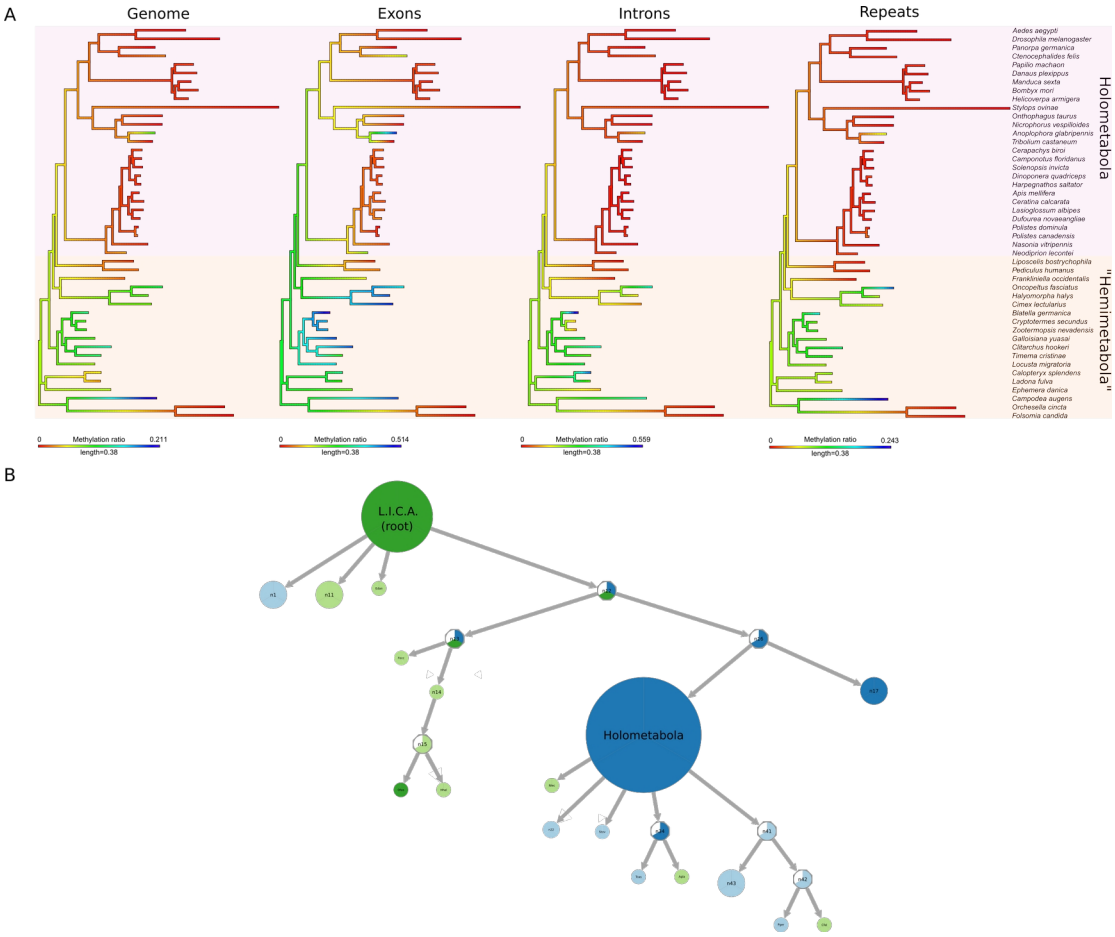


Figure 4.3: Ancestral state reconstruction of insect DNA methylation levels and patterns. A) Reconstructed states of DNA methylation levels of different genomic features mapped onto internal edges and nodes of the insect tree using a color gradient. Warmer colors represent lower levels of DNA methylation. B) Diagrammatic representation of reconstructed states of insect gene body methylation patterns (BM: dark green, UM: light blue, MM: light green, EM: dark blue, Unknown: white). Only nodes that have a different pattern compared to the parent node are displayed. For the position of each node on the insect tree please look at figure S12. L.I.C.A.: Last Insect Common Ancestor; Edan: *Ephemera danica;* Ofas: *Oncopeltus fasciatus;* Hhal: *Halyomorpha halys;* Clec: *Cimex lectularius;* Pdom: *Polistes dominula;* Stov: *Stylops ovinae;* Nlec: *Neodiprion lecontei;* Tcas: *Tribolium castaneum;* Agla: *Anoplophora glabripennis*; Pger: *Panorpa germinca;* Cfel: *Ctenocephalides felis*.

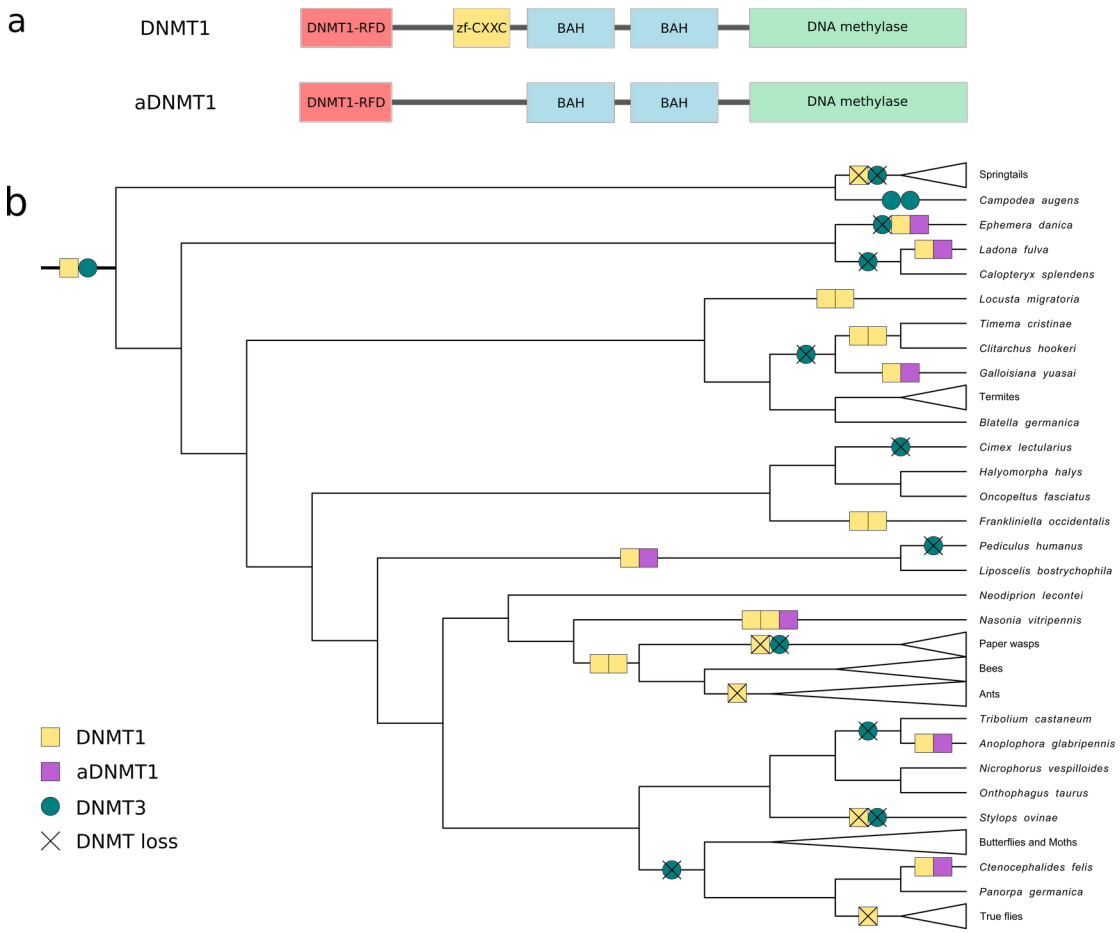


Figure 4.4: A) Comparison of the domain content between a typical insect DNMT1 and the activated DNMT1 (aDNMT1). ADNMT1 lacks the *de novo* methylation inhibiting zf-CXXC domain. B) The evolutionary history of insect DNMTs. ADNMT1 has emerged seven times independently in the 46 insect taxa examined and coexists with DNMT1 in absence of DNMT3 in six out of eight taxa that posses it. The presence of aDNMT1 and the loss of DNMT3 could be ancestral to Palaeoptera, as the genome assembly of *Calopteryx splendens* is very fragmented and a second aDNMT1 paralog could not be identified, something that likely affected our ancestral state reconstruction. From the two DNMT1 paralogs present in the last common ancestor of paper wasps, ants, and bees, one gene was lost in the paper wasp lineage and the other in the ant lineage. Only modifications (duplications or losses) relative to the ancestral toolkit are mapped on the tree.

4.7 References

Bewick AJ, et al. 2019. Diversity of cytosine methylation across the fungal tree of life. *Nature Ecology & Evolution*, *3*(3), 479–490.

Bewick AJ, et al. 2017a. The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. *Genome Biology*, *18*(1), 1–13.

Bewick AJ, Vogel KJ, Moore AJ, & Schmitz RJ. 2017b. Evolution of DNA methylation across insects. *Molecular Biology and Evolution*, *34*(3), 654–665.

Bird AP. 1995. Gene number, noise reduction and biological complexity. *Trends in Genetics*, *11*(3), 94–100.

Bolger AM, Lohse M, & Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bonasio R, et al. 2012. Genome-wide and caste-specific DNA methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Current Biology*, *22*(19), 1755–1764.

Bradley RK, et al. 2009. Fast Statistical Alignment. *PLOS Computational Biology*, *5*(5), 1–15.

Chernomor O, von Haeseler A, & Minh BQ. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, *65*(6), 997–1008.

Cunningham CB, et al. 2015. The genome and methylome of a beetle with complex social behavior, *Nicrophorus vespilloides* (Coleoptera: Silphidae). *Genome Biology and Evolution*, *7*(12), 3383–3396.

El-Gebali S, et al. 2018. The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432.

Falckenhayn C, et al. 2013. Characterization of genome methylation patterns in the desert locust *Schistocerca gregaria*. *Journal of Experimental Biology*, *216*(8), 1423–1429.

Felsenstein J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, *39*(4), 783.

Feng S, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, *107*(19), 8689–8694.

Gatzmann F, et al. 2018. The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. *Epigenetics & Chromatin*, *11*(1), 57.

Glastad KM, Hunt BG, Yi SV, & Goodisman MAD. 2011. DNA methylation in insects: On the brink of the epigenomic era. *Insect Molecular Biology*, *20*(5), 553–565.

Glastad KM, et al. 2017. Variation in DNA Methylation Is Not Consistently Reflected by Sociality in Hymenoptera. *Genome Biology and Evolution*, *9*(6), 1687–1698.

Glastad KM, Gokhale K, Liebig J, & Goodisman MAD. 2016. The caste- and sex-specific DNA methylome of the termite *Zootermopsis nevadensis*. *Scientific Reports*, *6*(October), 1–14.

Glastad KM, Hunt BG, & Goodisman MAD. 2014. Evolutionary insights into DNA methylation in insects. *Current Opinion in Insect Science*, *1*, 25–30.

Glastad KM, Hunt BG, & Goodisman MAD. 2019. Epigenetics in insects: Genome regulation and the generation of phenotypic diversity. *Annual Review of Entomology*, *64*, 185–203.

Grandi FC, et al. 2015. Retrotransposition creates sloping shores: a graded influence of hypomethylated CpG islands on flanking CpG sites. *Genome Research*, *25*(8), 1135–1146.

Huh I, Zeng J, Park T, & Yi SV. 2013. DNA methylation and transcriptional noise. *Epigenetics and Chromatin*, *6*(1), 1–10.

Ishikawa SA, Zhukova A, Iwasaki W, & Gascuel O. 2019. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution*, *36*(9), 2069–2085.

Jeltsch A, & Jurkowska RZ. 2016. Allosteric control of mammalian DNA methyltransferases - a new regulatory paradigm. *Nucleic Acids Research*, *44*(18), 8556–8575.

Johnson KP, et al. 2018. Phylogenomics and the evolution of hemipteroid insects. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(50), 12775–12780.

Kalyaanamoorthy S, et al. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589.

Katoh K, & Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Keller TE, Han P, & Yi SV. 2016. Evolutionary transition of promoter and gene body DNA methylation across invertebrate-vertebrate boundary. *Molecular Biology and Evolution*, *33*(4), 1019–1028.

Kück P, & Meusemann K. 2010. FASconCAT: Convenient handling of data matrices. *Molecular Phylogenetics and Evolution*, *56*(3), 1115–1118.

Lewis SH, et al. 2020. Widespread conservation and lineage-specific diversification of genome-wide DNA methylation patterns across arthropods. *PLOS Genetics*, *16*(6), 1–24.

Libbrecht R, Oxley PR, Keller L, & Kronauer DJC. 2016. Robust DNA methylation in the clonal raider ant brain. *Current Biology*, *26*(3), 391–395.

Liew YJ, et al. 2018. Epigenome-associated phenotypic acclimatization to ocean acidification in a reef-building coral. *Science Advances*, *4*(6).

Lyko F. 2017. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nature Reviews Genetics*.

Lyko F, et al. 2010. The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. *PLoS Biology*, *8*(11).

Lyko F, & Maleszka R. 2011. Insects as innovative models for functional studies of DNA methylation. *Trends in Genetics*, *27*(4), 127–131.

Misof B, et al. 2013. Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. *BMC Bioinformatics*, *14*, 348.

Misof B, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, *346*(6210), 763–767.

Misof B, & Misof K. 2009. A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments : A More Objective Means of Data Exclusion. *Systematic Biology*, *58*(1), 21–34.

Munoz-Torres MC, et al. 2011. Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Research*, *39*(Database issue), D658-62.

Neri F, et al. 2017. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, *543*(7643), 72–77.

Nguyen LT, Schmidt HA, von Haeseler A, & Minh BQ. 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274.

Niederhuth CE, et al. 2016. Widespread natural variation of DNA methylation within angiosperms. *Genome Biology*, *17*(1), 1–19.

Patalano S, et al. 2015. Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. *Proceedings of the National Academy of Sciences*, *112*(45), 13970–13975.

Petersen M, et al. 2017. Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics*, *18*(1), 111.

Prachumwat A, & Li WH. 2008. Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes. *Genome Research*, *18*(2), 221–232.

Provataris P, et al. 2018. Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola. *Genome Biology and Evolution*, *10*(March), 1185–1197.

R Core Team (2019). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rehan SM, Glastad KM, Lawson SP, & Hunt BG. 2016. The Genome and Methylome of a Subsocial Small Carpenter Bee, *Ceratina calcarata*. *Genome Biology and Evolution*, *8*(5), 1401–1410.

Revell LJ. 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*(2), 217–223.

Simão FA, et al. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

Standage DS, et al. 2016. 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.

Takuno S, Ran JH, & Gaut BS. 2016. Evolutionary patterns of genic DNA methylation vary across land plants. *Nature Plants*, *2*(2), 1–7.

Takuno S, Seymour DK, & Gaut BS. 2017. The evolutionary dynamics of orthologs that shift in gene body methylation between Arabidopsis species. *Molecular Biology and Evolution*, *34*(6), 1479–1491.

Tweedie S, Charlton J, Clark V, & Bird A. 1997. Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Molecular and Cellular Biology*, *17*(3), 1469–1475.

Wang X, et al. 2013. Function and Evolution of DNA Methylation in *Nasonia vitripennis*. *PLoS Genetics*, *9*(10).

Whelan S, Irisarri I, & Burki F. 2018. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics*, *34*(22), 3929–3930.

Xi Y, & Li W. 2009. BSMAP: Whole genome BiSulfite sequence MAPping program. *BMC Bioinformatics*, *10*, 1–9.

Xiang H, et al. 2010. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nature Biotechnology*, *28*(5), 516–520.

Zemach A, McDaniel IE, Silva P, & Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, *328*(5980), 916–919.

Zhou W, Liang G, Molloy PL, & Jones PA. 2020. DNA methylation enables transposable element-driven genome expansion. *Proceedings of the National Academy of Sciences*, *117*(32), 201921719.

Chapter 5

General Discussion

This chapter will focus on discussing hypotheses for which analyses carried out in chapters three and four provided evidence, but could not be addressed in these chapters as they would overextend the focus of the original article intended for publication. Finally, the impact of this thesis in the field of animal comparative epigenomics will be evaluated and future research directions will be provided.

5.1 Unaddressed Hypotheses

Pleiotropy Underlies the Evolutionary Conservation of DNMT1 in Insects

In chapters three and four we provided strong evidence for the existence of DNMT1-only methylomes in many insects and found that DNMT1 is more evolutionarily conserved compared to DNMT3. In this paragraph we will use our results and evidence from the literature to better understand the evolutionary success of DNMT1. To begin with, as we previously suggested in chapters three and four, DNMT1 is likely able to methylate *de novo*. *In vitro* studies have shown that DNMT1 is a very capable *de novo* DNA methyltransferase and several studies support that this activity can also be detected *in vivo* (reviewed by Jeltsch and Jurkowska 2014). Furthermore, a single copy of DNMT1 can seemingly sustain functional methylation in absence of DNMT3, with the most characteristic example being butterflies and moths (Lepidoptera). In such cases, one scenario is that certain DNMT1 isoforms are better suited to perform *de novo* methylation, compensating for the loss of DNMT3. Alternatively, this single DNMT1 copy might be enough for maintaining methylation patterns over long evolutionary periods (Lyko 2017). For example, it has been proposed that the yeast *Cryptococcus neoformans* has maintained genomic

methylation for over 50 million years without a *de novo* methyltransferase (Catania et al. 2020). Finally, recent studies in the red flour beetle *Tribolium castaneum* (Song et al. 2017; Schulz et al. 2018) and the large milkweed bug *Oncopeltus fasciatus* (Amukamara et al. 2020) point to functions unrelated to DNA methylation for DNMT1. These additional functions have likely played a major role for its evolutionary conservation, especially in species with extremely reduced DNA methylation, like *T. castaneum*. In consequence, pleiotropy likely underlies the evolutionary conservation of DNMT1 and its success over DNMT3 in insects.

ADNMT1 May Facilitate Secondary Expansions of DNA Methylation

In chapter four we suggested that DNMT1-only methylomes are often mediated by two DNMT1 copies, one of which has been reconfigured towards *de novo* methylation, which we called aDNMT1. There exist two evolutionary scenarios under which aDNMT1 could have emerged. The first scenario is that aDNMT1 emerged while DNMT3 was still part of the methylation toolkit of a species. Under this scenario it is likely that a newly emerged aDNMT1 performed a similar function to DNMT3, thus creating a degenerate[1] methylation toolkit. ADNMT1, either through performing better than DNMT3 at their designated task and/or due to its pleiotropic

1 Degeneracy in biological systems describes the ability of elements that are structurally different to perform the same function or yield the same output (Edelman and Gally2001). Despite aDNMT1 and DNMT3 sharing a similar catalytic domain, they are structurally different and have diverged before the appearance of eukaryotes (Jurkowski and Jeltsch 2011). Thus, we could characterize a methylation toolkit that contains both aDNMT1 and DNMT3 as degenerate under the evolutionary scenario described here.

properties described in the previous paragraph, could have rendered DNMT3 less essential, reducing selective pressure to maintain it. In chapter four, we confirmed the coexistence of aDNMT1 and DNMT3 in lice and the jewel wasp, *Nasonia vitripennis*. Interestingly, DNMT3 was lost from the miniaturized genome of the human body louse (Kirkness et al. 2010), *Pediculus humanus*, but is still present in the substantially larger genome of the booklouse, *Liposcelis bostrychophila*, and the two species possess a similar pattern and comparable levels of DNA methylation. Thus, the replacement of DNMT3 by aDNMT1 did not result in a disruption of DNA methylation in *P. humanus*. The second scenario is that the loss of DNMT3 preceded DNMT1 duplication and the emergence of aDNMT1. In this scenario a single copy of DNMT1 could be sufficient for keeping methylation patterns intact over long evolutionary periods (Lyko 2017), especially if methylation levels remain very low. A characteristic example for this scenario are Mecopterida, a group from which DNMT3 was ancestrally lost (Provataris et al. 2018). In this group, DNA methylation has been maintained at very low levels by a single copy of DNMT1 in all butterfly and moth species (Lepidoptera) and has been extremely reduced in true flies as well as the German scorpionfly, *Panorpa germanica*. However, in the cat flea, *Ctenocephalides felis*, DNA methylation has undergone a secondary expansion according to the ancestral state reconstruction analyses presented in chapter four, being mediated by a DNMT1-aDNMT1 toolkit. Another such example comes from beetles, where the Asian long horn beetle, *Anoplophora glabripennis*, has the same toolkit as *C. felis* and vastly increased methylation levels compared to *T. castaneum*. Thus, I can hypothesize that aDNMT1 has contributed to a secondary increase of DNA

methylation in *C. felis* and *A. glabripennis.* In conclusion, I propose that the emergence of aDNMT1 may successfully replace DNMT3 without affecting the patterning of DNA methylation and may enable an expansion of genomic methylation in DNMT1-only species.

The Possible Benefits of DNA Methylation Reduction

To identify benefits of methylation reduction or loss, I first need to revisit the designated functions of DNA methylation. The most widely recognized function of DNA methylation, known as the "genome defence model", posits that DNA methylation contributes to the repression of transposable elements (TEs) and consequently prevents DNA damage that would be caused by their unhindered activity (Yoder et al. 1997; Bird 2002). This model was fueled by comparisons between mammalian genomes, which are ubiquitously methylated and experience a rather limited amount of TE-derived mutations relative to the abundance of TEs they accommodate, and the fruit fly, *Drosophila melanogaster,* an insect species that practically lacks DNA methylation and experiences a substantially higher amount of TE-associated mutations (Yoder 1997). The second model posits that the primary function of DNA methylation is the reduction of transcriptional noise, because the production of a large amount of non-essential transcripts from irrelevant promoters, which may be TE-derived, would burden cellular gene expression (Bird 1995; Bird 2002). These models are not mutually exclusive and have both recently received additional support. Zhou and colleagues (2020) provided evidence supporting that the expansion of DNA methylation during the evolutionary history of vertebrates was a

result of the interplay between TEs and DNA methylation and did not only function to suppress TE activity, but also enabled the integration of TEs in the host genome. Additionally, by focusing on mammalian systems they provided evidence that the intergated TEs are commonly repurposed into regulatory elements, primarily through an accumulation of DNA methylation-induced mutations. Bird's (1995) conjecture has received support from both mammalian and invertebrate taxa (including insects) with the consensus being that methylation of genes contributes to gene expression regulation and to the reduction of transcriptional noise (Maunakea et al. 2010; Huh et al. 2013; Wang et al. 2013; Glastad et al. 2016; Neri et al. 2017; Gatzmann et al. 2018; Li et al. 2018; Liew et al. 2018). In conclusion, DNA methylation may be beneficial in multiple ways, from offering protection against intragenomic parasites, to facilitating the expansion of existing regulatory networks, and to safeguarding the integrity of gene expression.

In chapter four I proposed that the trajectory of insect DNA methylation has been the exact opposite of the one observed in vertebrates. The last common ancestor of insects possessed substantial methylation levels at gene bodies and TEs, whereas methylation levels massively contracted in the last common ancestor of Holometabola and were completely lost from many extant species, most notably true flies (Diptera). This poses the question on how the reduction/loss of DNA methylation affected holometabolous genomes, as by abolishing the mark they would also abolish its benefits.

Starting from the genome defence model, it is long known that in *D. melanogaster,* a species that lacks DNA methylation, small RNAs induce a repressive

chromatin state near newly inserted TEs (Chung et al. 2008; Choi and Lee, 2020). The adaptation to repress novel TEs is so fast in *D. melanogaster* that it can even occur during the lifetime of a single fly (Khurana et al. 2011). Additionally, expansions of gene families belonging to these small RNA pathways were recently identified in Diptera (Lewis et al. 2018). Further studies on *D. melanogaster* have shown that TE integration may lead to the generation of novel regulatory sequences (reviewed by Chuong et al. 2017), without the assistance of DNA methylation-induced mutations (Zhou et al. 2020). Thus, in absence of DNA methylation other mechanisms may contribute to genome defence and the generation of novel TE-derived regulatory sequences.

In the last two paragraphs, I will focus on how the contraction of DNA methylation in insects may potentially affect gene expression. In chapter four I proposed that the reduction of DNA methylation occurs as newly inserted TEs are not preferentially methylated, and that their insertion induces further hypomethylation of adjacent host DNA. It is also now understood that repressive marks that act to silence TE activity, including DNA methylation, may also suppress the expression of adjacent genes (Choi and Lee 2020). Thus, it is possible the lack of DNA methylation at gene-proximate TEs may be beneficial, as the expression of these genes would not be altered by newly added methylation or genes that were previously repressed by DNA methylation might be activated due to TE-induced hypomethylation.

Furthermore, in chapter four, I showed that a substantially larger proportion of genes tends to be methylated in hemimetabolous insects compared to Holometabola. Methylated genes in Holometabola tend to be evolutionarily conserved, are enriched

in housekeeping functions (Elango et al. 2009; Lyko et al. 2010; Hunt et al. 2013; Wang et al. 2013; Cunningham et al. 2015; Rehan et al. 2016; Provataris et al. 2018), and their expression profile is tightly regulated as they show intermediate levels of expression that is stable across tissues, morphs, or developmental stages (Xiang et al 2010; Bonasio et al. 2012; Foret et al. 2012; Wang et al. 2013; Libbrecht et al. 2016). It is unlikely that all methylated genes in the genomes of hemimetabolous insects are housekeeping and evolutionarily conserved. Indeed, Glastad and colleagues (2016) found that methylated genes in the termite, *Zootermopsis nevadensis*, include genes that are tissue-specific and do not show as strong evolutionary conservation as the methylated genes of Holometabola (see also Sarda et al. 2012). However, they also found that methylated genes tend to exhibit less transcriptional noise. Thus, I propose that as gene body methylation contracted during the evolutionary history of insects, an increasing proportion of their gene repertoire was less tightly regulated and thus exhibited higher levels of transcriptional noise. According to Bird's theory (1995), this could burden gene expression programs, but could also provide holometabolous genomes with higher evolvability potential. Barosso and colleagues (2018), by studying mammalian gene expression, showed that transcritpional noise is positively correlated with the evolutionary rate of proteins, as highly constrained genes displayed less transcriptional noise than fast evolving ones. Therefore, I suggest that the gene repertoire of lowly methylated holometabolous genomes is comparably more evolvable than the one of strongly methylated hemimetabolous genomes due to lower constraint at the protein sequence level. Finally, in chapter four I found that the contraction of gene body methylation has also kept genes of Holometabola rather free

of TEs as DNA methylation was not there to facilitate their genomic integration (Zhou et al. 2020). Because a loss of repressive marks may lead to the reactivation of old TEs (reviewed by Chuong et al 2020), the absence of TEs from the majority of holometabolous insect genes could have potentially acted to balance out the harmful effects of methylation reduction and transcriptional noise increase. Overall, I propose that holometabolous insects, the most species-diverse group of animals (Stork 2018), have reaped the benefits of the contraction of DNA methylation, a fact that may have ultimately contributed to increased molecular biodiversity.

5.2 Conclusions

This thesis constitutes the first comprehensive study on insect comparative epigenomics to date. Our results show that the most well-studied insects, concentrated in a small region of the insect phylogenetic tree, are not representative of the entire group. This highlights the importance of macroevolutionary studies in understanding patterns that could not be identified by focusing on restricted sets of taxa. At the same time, this thesis sets the ground for experimental and detailed comparative analyses that can deepen our understanding on the evolution and function of DNA methylation in insects. More specifically, the function of DNMT1 in DNMT3-deficient insect methylation systems should be comprehended and compared to the function of the vertebrate DNMT toolkit. The target species for such studies have now been provided. Most importantly, the evolutionary mechanisms that underlie the change in epigenomic regulation between hemimetabolous and holometabolous insects should be thoroughly studied. An initial mechanistic model on how DNA methylation

contracted during the evolutionary history of insects has also been provided and is now ready for thorough evaluation. I sincerely hope I made a valuable contribution.

5.3 References

Amukamara AU, et al. 2020. More Than DNA Methylation: Does Pleiotropy Drive the Complex Pattern of Evolution of Dnmt1? Frontiers in Ecology and Evolution, 8, 4.

Barroso GV, Puzovic N, & Dutheil JY. 2018. The Evolution of Gene-Specific Transcriptional Noise Is Driven by Selection at the Pathway Level. *Genetics*, *208*(1), 173–189.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev*, *16*, 6–21.

Bird AP. 1995. Gene number, noise reduction and biological complexity. *Trends in Genetics*, *11*(3), 94–100.

Bonasio R, et al. 2012. Genome-wide and caste-specific DNA methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Current Biology*, *22*(19), 1755–1764.

Catania S, et al. 2020. Evolutionary persistence of DNA methylation for millions of years after ancient loss of a *de novo* methyltransferase. *Cell*, *180*(2), 263-277.

Choi JY, & Lee YCG. 2020. Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. *PLoS Genetics*, *16*(7), e1008872.

Chung WJ, Okamura K, Martin R, & Lai EC. 2008. Endogenous RNA interference provides a somatic defense against Drosophila transposons. *Current Biology*, *18*(11), 795–802.

Chuong EB, Elde NC, & Feschotte C. 2017. Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics*, *18*(2), 71–86.

Cunningham CB, et al. 2015. The genome and methylome of a beetle with complex social behavior, *Nicrophorus vespilloides* (Coleoptera: Silphidae). *Genome Biology and Evolution*, *7*(12), 3383–3396.

Edelman GM, & Gally JA. 2001. Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(24), 13763–13768.

Elango N, Hunt BG, Goodisman MAD, & Yi SV. 2009. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(27), 11206–11211.

Foret S, et al. 2012. DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. *Proceedings of the National Academy of Sciences*, *109*(13), 4968–4973.

Gatzmann F, et al. 2018. The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. *Epigenetics & Chromatin*, *11*(1), 57.

Glastad KM, Gokhale K, Liebig J, & Goodisman MAD. 2016. The caste- and sex-specific DNA methylome of the termite *Zootermopsis nevadensis*. *Scientific Reports*, *6*(October), 1–14.

Huh I, Zeng J, Park T, & Yi SV. 2013. DNA methylation and transcriptional noise. *Epigenetics and Chromatin*, *6*(1), 1–10.

Hunt BG, Glastad KM, & Goodisman MAD. 2013. Genome composition, caste, and molecular evolution in eusocial insects. *Proceedings of the National Academy of Sciences*, *110*(6), E445–E446.

Jeltsch A, & Jurkowska RZ. 2014. New concepts in DNA methylation. *Trends in Biochemical Sciences*, *39*(7), 310–318.

Jurkowski TP, & Jeltsch A. 2011. On the evolutionary origin of eukaryotic DNA methyltransferases and Dnmt2. *PLoS ONE*, *6*(11), 1–9.

Khurana JS, et al. 2011. Adaptation to P element transposon invasion in *Drosophila melanogaster*. *Cell*, *147*(7), 1551–1563.

Kirkness EF, et al. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences*, *107*(27), 12168–12173.

Lewis SH, et al. 2018. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. *Nature Ecology and Evolution*, *2*(1), 174–181.

Li Y, et al. 2018. DNA methylation regulates transcriptional homeostasis of algal endosymbiosis in the coral model Aiptasia. *Science Advances*, *4*(8).

Libbrecht R, Oxley PR, Keller L, & Kronauer DJC. 2016. Robust DNA methylation in the clonal raider ant brain. *Current Biology*, *26*(3), 391–395.

Lyko F. 2017. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nature Reviews Genetics*.

Lyko F, et al. 2010. The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. *PLoS Biology*, *8*(11).

Maunakea AK, et al. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, *466*(7303), 253–257.

Neri F, et al. 2017. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, *543*(7643), 72–77.

Provataris P, et al. 2018. Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola. *Genome Biology and Evolution*, *10*(March), 1185–1197.

Rehan SM, Glastad KM, Lawson SP, & Hunt BG. 2016. The Genome and Methylome of a Subsocial Small Carpenter Bee, *Ceratina calcarata*. *Genome Biology and Evolution*, *8*(5), 1401–1410.

Sarda S, Zeng J, Hunt BG, & Yi SV. 2012. The evolution of invertebrate gene body methylation. *Molecular Biology and Evolution*, *29*(8), 1907–1916.

Schulz NKE, et al. 2018. 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.

Song X, et al. 2017. Genome-wide DNA methylomes from discrete developmental stages reveal the predominance of non-CpG methylation in *Tribolium castaneum*. *DNA Research*, *24*(5), 445–458.

Stork NE. 2018. How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annual Review of Entomology*, *63*(September 2017), 31–45.

Wang X, et al. 2013. Function and Evolution of DNA Methylation in *Nasonia vitripennis*. *PLoS Genetics*, *9*(10).

Xiang H, et al. 2010. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nature Biotechnology*, *28*(5), 516–520.

Yoder JA, Walsh CP, & Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, *13*(8), 335–340.

Appendices

Appendix to Chapter 3

Supplementary material and methods, supplementary tables, and supplementary figures can be found at:

https://academic.oup.com/gbe/article/10/4/1185/4943971#supplementary-data

or by using the following doi: 10.1093/gbe/evy066

Appendix to Chapter 4

The supplementary text, supplementary tables (4.S1-4.S4), supplementary figures (4.S1-4.S13), and supplementary files 1 and 2 can be found in the provided CD/in the electronic supplement.

Acknowledgements

I am grateful to Bernhard Misof for the opportunity to carry out my doctoral thesis project in his group. I am also truly grateful for the unconditional support and guidance he offered in my times of need. Our conversations formed my view not only towards science, but also towards life in general. I will always carry part of his thoughts, ideas, and spirit with me from now on.

I am grateful to Frank Lyko for accepting to co-advise me and provide valuable time, thoughts, and resources for this project to realize its full potential.

I would also like to thank Martin Sander and Gabriele Koenig for their willingness to join my thesis committee and read my thesis.

I also consider myself extremely lucky to have shared an office with Alex Vasilikopoulos and Jan Philip Oeyen. I will really miss you guys.

I would also like to thank Sonja Grath, Oliver Niehuis, and especially Karen Meusemann for their cooperation, support, and advice through the first years of this project.

I appreciate the financial support I received during my doctoral thesis from the Bakalas brothers foundation, the Zoological Research Museum A. Koenig, the Alexander Koenig Gessellschaft, and the German Cancer Research Center.

I would also like to thank many members of the ZFMK in no particular order for being there one way or the other throughout the past five years: Matthias Geiger, Julia Schwarzer, Sebastian Martin, Jonas Astrin, Claudia Etzbauer, Ameli Kirse, Victoria Moris, Simon Käfer, Tanja Ziesmann, Hamideh Fard, Dirk Rohwedder, Jonas

Eberle, Annette LaRoche, Sandra Kukowka, Sandra Middelhof, Jeanne Wilbrandt, Sandra Meid, Thomas Pauli, Wolfgang Waegele, Heike Waegele, Elise Laetz, Gabby Nottenrock, Carola Greve and Ralf Peters. I apologize to anyone I forgot to mention here.

I would also like to thank the members of Frank Lyko's group in the German Cancer Research Center for aiding with my project and making sure I feel comfortable during my visit in Heidelberg: Katharina Hanna, Fanny Gatzmann, Vitor Coutinho, Julian Gutekunst, Manuel Rodriguez-Paredes, Franzeska Tuorto, Carine Legrand, Olena Maiakovska, Llorenç Solé Boldo, Sina Toenges, and especially my friend Guenter Raddatz.

I also want to thank my new family here in Germany, Juan, Elsa, and Steffi, for being who they are. I also thank my doggy, Shibo, for reminding me that there is life outside of work.

I consider myself very lucky to have met my girlfriend Yara, who took care of me during the wrap-up stages of this work and for sticking by me through ups and downs over the past three years. Yaris, I am in your debt.

Finally, I cannot express enough gratitude to my mother Thomi and my sister Anna for their unconditional love and support which I can never return. Mama I will never forget your sacrifices.

Declaration of Authorship

I herewith declare that I have written this thesis independently and myself. I did not use any other sources than those listed. All places where the exact words or analogous text were taken from sources are indicated. I assure that this thesis has not been submitted for examination elsewhere.

August 31, 2020