Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the

Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

M. Sc. Cassandra Falckenhayn

born in Berlin, Germany

Oral-examination:.....

The Methylome of the Marbled Crayfish *Procambarus virginalis*

Referees:

Prof. Dr. Frank Lyko

Prof. Dr. Ana Martin-Villalba

Zusammenfassung

Die Bedeutung der DNA-Methylierung in Invertebraten scheint unterschiedlich zu der in Säugetieren zu sein und ihre evolutionäre Konservierung innerhalb der Invertebraten ist unklar. Bisher geben nur zwei Studien einen groben Überblick über Krebstiermethylome. Der parthenogene Marmorkrebs weist trotz seiner genetischen Uniformität eine hohe Umweltadaptabilität auf und verfügt aus diesem Grund über die notwendigen Eigenschaften eines Modellorganismus.

Das Ziel dieser Arbeit war es, das Methylom des Mamorkrebses auf dem Auflösungsvermögen einzelner Basen mittels Whole-Genome Bisulfite Sequencing zu charakterisieren, um neue Einblicke in die Methylierung in Krebstieren und die evolutionäre Konservierung innerhalb der Invertebraten zu gewinnen.

Die Analyse der mitochondrialen DNA von verschiedenen Mamorkrebspopulationen belegt einen gemeinsamen Ursprung und legt die Betrachtung des Mamorkrebses als unabhängige asexuelle Art, *Procambarus virginalis,* nahe. Aufgrund des großen Genoms von *P. virginalis* wurde das Transkriptom assembliert. Der Vergleich zu anderen Arten zeigte, dass die erste Version eine gute Qualität hat und ein konserviertes DNA-Methylierungs-System beinhaltet. Die CpG-Depletion in Transkriptsequenzen und die massenspektrometrische Analyse bestätigten eine historische Keimbahn- sowie gegenwärtige DNA-Methylierung in verschiedenen Geweben von *P. virginalis*.

Das Methylom von *P. virginalis* wies die wichtigsten Merkmale von Tiermethylomen auf, wie z.B. die Genmethylierung. Die Genmethylierung war bimodal verteilt und hatte das typische Muster eines mosaik methyliert Invertebratengenoms. Primär waren die Housekeeping-Gene methyliert mit einer parabolischen Beziehung zu ihrer Expression, was darauf hindeutet, dass die DNA-Methylierung von Housekeeping-Genen ihre Expression feinabstimmt. Die Repeats waren generell hypomethyliert und ihre Methylierung war abhängig von ihrer Position zu den Genen. Der Vergleich der Genmethylierung zwischen Individuen und Geweben wies eine hohe Reproduzierbarkeit der Methylierungsmuster auf, während der Vergleich zwischen *P. fallax* und *P. virginalis* eine Genhypomethylierung in *P. virginalis* aufzeigte. Diese kann jedoch nicht die mittels Massenspektrometrie detektierte globale Hypomethylierung in *P. virginalis* erklären.

Zusammenfassend zeigen diese Ergebnisse, dass das *P. virginalis* Methylom durch gewebsinvariante Housekeeping-Gen-Methylierung gekennzeichnet ist und die bevorzugte Methylierung von Housekeeping-Genen in *P. virginalis* untermauert einen funktionellen Unterschied zur gewebespezifischen Methylierung in Säugetieren. Mit dieser Arbeit werden neue Einblicke in die evolutionäre Konservierung von Gen- und Repeatmethylierung in Invertebraten, insbesondere Krebstieren ermöglicht.

Abstract

DNA methylation in invertebrates seems to play a different role as in mammals and its evolutionary conservation among invertebrates is unclear. Only two studies describe crustacean methylomes giving just a small overview. The parthenogenetic reproducing marbled crayfish display a high environmental adaptability besides its genetic uniformity and thus, possess the necessary attributes of a laboratory model organism.

The aim of this work was to characterize the methylome of the marbled crayfish at single-base resolution using whole-genome bisulfite sequencing in an attempt to give new insights into DNA methylation in crustaceans and thus, in the evolutionary conservation among invertebrates.

Analysis of the mitochondrial DNA of different marbled crayfish strains revealed a single origin and suggests to consider the marbled crayfish as independent asexual species *Procambarus virginalis*. Furthermore, since the *P. virginalis* possess a large genome size, the transcriptome was assembled and comparison to other species revealed a relative good quality of the first draft transcriptome as well as the presence of a conserved DNA methylation system in *P. virginalis*. Analysis of the CpG depletion in protein-coding sequences and mass spectrometry confirmed historical germline and current DNA methylation in various tissues of *P. virginalis*.

The methylome was characterized by the key features of animal methylomes with methylation targeted to gene bodies. The gene bodies displayed the typical pattern of a mosaically methylated invertebrate genome and a bimodal distribution of their methylation levels. Targeted gene bodies were annotated as housekeeping genes and methylation showed a parabolic relationship to housekeeping gene expression suggesting that the DNA methylation of housekeeping genes might fine-tune their expression. Additionally, repeats were generally hypomethylated and the methylation of repeats depended on their position to gene bodies. Finally, inter-individual and inter-tissue comparison of gene body methylation revealed a high reproducibility of the methylation patterns, while inter-species comparison between *P. fallax* and *P. virginalis* displayed an overall hypomethylation in the *P. virginalis* genes which however, could not explain the by mass spectrometry detected global hypomethylation in *P. virginalis*. These findings uncovered that the *P. virginalis* methylome is characterized by tissue-invariant housekeeping gene methylation.

This thesis describes novel insights into the evolutionary conservation of gene body and repeat methylation in invertebrates, especially crustaceans, and the preferential methylation of housekeeping genes highlights a functional difference to the tissue-specific methylation in mammals.

Contents

L1 List of Figures

L2 List of Tables

L3 List of Abbreviations

°C	Degree Celsius
µg	Microgram
5hmC	5-hydoxymethylcytosine
5mC	5-methylcytosine
6mA	N6-methyladenine
µl	Microliter
µM	Micromolar
bp	Base Pair
C	Cytosine
cDNA	Complementary DNA
CDS	Protein-Coding Sequence
CHG	Cytosine-nonGuanosine-Guanosine Trinucleotide
CHH	Cytosine-nonGuanosine-nonGuanosine Trinucleotide
CpG	Cytosine Guanosine Dinucleotide
CpGo/e	amount of observed in relation to the amount of expected CpGs
Da	Dalton
DMSO	Dimethyl Sulfoxide
Dnmt	DNA Methyltransferase
dNTPs	Desoxynucleotides Mix
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic Acid
Gb	Giga Base Pairs (1,000,000,000 bp)
HBSS	Hank's Balanced Salt Solution
HD	Laboratory Marbled Crayfish Strain Heidelberg
HKG	Housekeeping Gene
kb	Kilo Base Pairs (1,000 bp)
LINEs	Long Interspersed Elements
LTRs	Long Terminal Repeats
mg	Milligram
ng	Nanogram
nt	nucleotide
PBS	Phosphate Buffered Solution
PCR	Polymerase chain reaction
PIC	Preinitiation complex
qRT-PCR	Quantitative real time polymerase chain reaction
SDS	Sodium Dodecyl Sulfate
SINEs	Short Interspersed Elements
SNP	Single Nucleotide Polymorphism
T	Thymine
TAE Buffer	Tris-Acetic Acid-EDTA Buffer
TBE Buffer	Tris-Borat-EDTA Buffer
TBP	TATA-box Binding Protein
TE Buffer	Tris-EDTA Buffer
Tet	Ten-Eleven Translocation Methylcytosine Dioxygenase
TPM	Transcripts per Kilobase Million
TSS	Transcription Start Site
TTS	Transcription Termination Site
UTR	Untranslated Region
WGBS	Whole-Genome Bisulfite Sequencing
WGS	Whole-Genome Sequencing

1 Introduction

1.1 Epigenetic Modifications

Epigenetics is the study of inherited changes in phenotypes (cellular and physiological) and consequently gene expression patterns that did not resulted from alterations in the base-pair nucleotide sequence of genes (A. Bird, 2007). Epigenetic is participating in cellular identity and lineage choice (Fisher, 2002). Moreover, it is widely accepted that epigenetic mechanisms are involved in environmentally controlled phenotypic plasticity and thus, in connecting the genome and the environment (Duncan, Gluckman, & Dearden, 2014; Lyko & Maleszka, 2011).

Epigenetic marks are mainly covalent modifications of histones and DNA (Bernstein, Meissner, & Lander, 2007; Goldberg, Allis, & Bernstein, 2007). Histone modifications can influence the chromatin structure via histone-histone and histone-DNA interactions (Tessarz & Kouzarides, 2014). Depending on the type of histone modification like acetylation, methylation and ubiquitylation the chromatin structure is either compact (heterochromatin) or open (euchromatin) (Bannister & Kouzarides, 2011). Therefore, histone modifications are involved in the regulation of replication, transcription and DNA repair (Tessarz & Kouzarides, 2014). DNA modifications are attachments of a functional group to an atom of the nucleobase (DNA base) and comprise cytosine, uracil and adenine (Breiling & Lyko, 2015). DNA methylation is a type of epigenetic DNA modification where a methyl group is attached either to the nitrogen atom of the amino group at the 6th carbon-atom of adenine (N6-methyladenine: 6mA) or to the 5th carbon-atom of cytosine (5-methylcytosine: 5mC), which is catalyzed by two different classes of enzymes (Breiling & Lyko, 2015). N6-methyladenine is the predominant DNA modification in prokaryotes and primarily functions in the host defense system (Luo et al., 2015). In contrast, the methylation of adenine in eukaryotes has remained largely uncharacterized and recent publications indicated a possible role of N6-methyladenin in transcription (Fu et al., 2015; Greer et al., 2015; Luo et al., 2015; Ratel et al., 2006; Zhang et al., 2015). Moreover, 5-methylcytosine is the most common DNA modification in eukaryotes and hence, methylated cytosines are in the focus of the majority of DNA methylation studies (Luo et al., 2015; Vanyushin, Tkacheva, & Belozersky, 1970).

5-methylcytosine is functionally involved in genomic imprinting, cell differentiation and silencing of repetitive DNA (P. A. Jones, 2012). Additionally, methylation patterns change during development, aging and diseases like cancer (Horvath, 2013; P. A. Jones, 2012; Smith & Meissner, 2013). The majority of DNA methylation studies was performed in mammals, but few analyses of insect methylomes already generated new ideas about the significance of DNA methylation as a regulatory mechanism to organismal biology (Lyko & Maleszka, 2011).

However, it is still surprisingly challenging to assign the function to the DNA methylation at a specific gene (Schübeler, 2015).

1.2 DNA Cytosine Methylation

The significance of DNA methylation for organismal vitality was demonstrated in 1992 by the knockout of the catalyzing enzyme which resulted in embryonic lethality in mice (E. Li, Bestor, & Jaenisch, 1992). In the same year bisulfite sequencing was performed for the first time to analyzes 5-methylcytosine at single bases of a human promoter sequence (Frommer et al., 1992). Since then the understanding of the evolutionary conservation of the catalyzing enzymes and methylation patterns in animals could be expanded.

1.2.1 The Animal Methylation Machinery

Methylation of cytosines in animals relies upon the family of DNA methyltransferases (Dnmts), which can be divided into three subfamilies: Dnmt1, Dnmt2 and Dnmt3 (Goll & Bestor, 2005; Law & Jacobsen, 2010). All three subfamilies show strong sequence conservation in their C-terminal catalytic motifs (Fig. **1.1A**) and can catalyze the methylation of cytosines (Goll & Bestor, 2005; Jurkowska, Jurkowski, & Jeltsch, 2011). However, they are distinct in their N-terminal regulatory domains and function (Fig. **1.1A**) (Goll & Bestor, 2005). Dnmt2 uses its DNA methyltransferase mechanism to methylate cytosines in tRNAs (Jurkowski et al., 2008). Dnmt3 known as *de novo* methyltransferase establishes new DNA methylation patterns (Fig. **1.1B**) and Dnmt1 known as maintenance methyltransferase copies methylation marks from the parental DNA strand to the new synthesized daughter strand (Fig. **1.1B**) (Goll & Bestor, 2005; Law & Jacobsen, 2010).

Since Dnmt2 methylates tRNAs (Goll et al., 2006; Schaefer et al., 2010), Dnmt2-only organisms lack DNA methylation (Raddatz et al., 2013). Interestingly, the gene copy numbers of Dnmt1 and Dnmt3 varies within the animal kingdom (Goll & Bestor, 2005). Mammals, for example, possess one copy of Dnmt1 and three copies of Dnmt3, while in some invertebrates the number of Dnmt1 expanded up to three copies in *Nasonia* (Fig. **1.1C**) (Goll & Bestor, 2005; Werren et al., 2010). *Bombyx mori* and *Tribolium castaneum* both possess only one copy of Dnmt1 and lack Dnmt3 (Fig. **1.1C**) (Richards et al., 2008; Werren et al., 2010). However, the genome of *T. castaneum* is unmethylated, while the *B. mori* genome is methylated (Xiang et al., 2010). As *B. mori* is the only known example for a Dnmt1-mediated methylome in animals, at least one copy of Dnmt1 and Dnmt3 are considered necessary for a functional genome-wide DNA methylation system (Goll & Bestor, 2005; Lyko & Maleszka, 2011; Yi & Goodisman, 2009).

DNA methylation is a stable epigenetic mark and methylation patterns can only become dynamic via demethylation mechanisms (Schübeler, 2015). Demethylation occurs either passively by replication in absence of maintenance methylation or actively by removing methylated cytosines (Law & Jacobsen, 2010). The ten-eleven translocation (Tet) family can oxidize 5-methylcytosine to 5-hydroxymethylcytosine (5hmC) (Fig. **1.1B**) and subsequently to the intermediates 5-formylcytosine and 5-carboxylcytosine, which are targeted by base excision repair mechanisms (He et al., 2011; Ito et al., 2011; Tahiliani et al., 2009). Therefore, the Tet family provides a potential pathway for active 5mC-demethylation (Tahiliani et al., 2009).

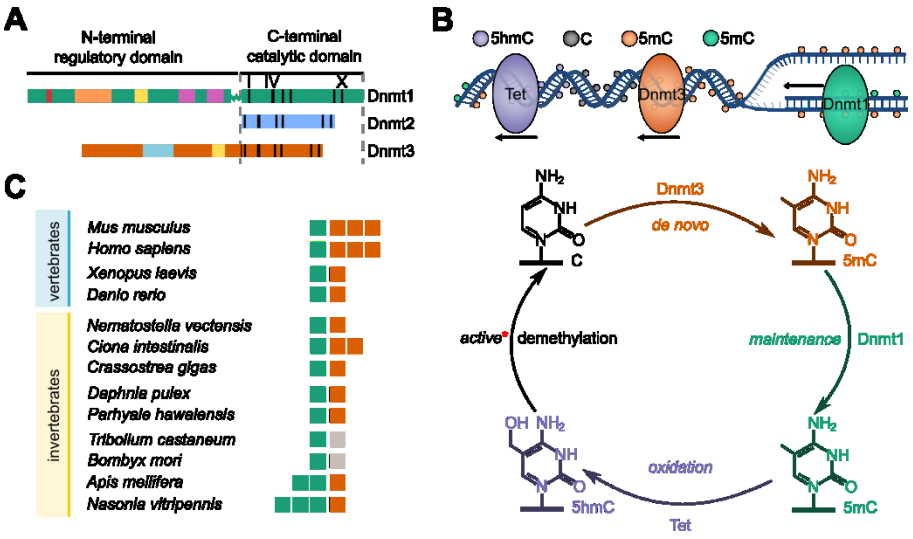


Figure 1.1 DNA methylation machinery.

(**A**) Overview of the structure of the mammalian DNA methyltransferases (Dnmts). All Dnmts share the 10 motifs of the C-terminal catalytic domain and differ in their N-terminal regulatory domain. Dnmt1: NLS (red), replication foci (orange), Cys-rich (yellow), BAH (purple); Dnmt3: PWWP (blue), Cys-rich (yellow) (adopted from Goll & Bestor, 2005). (**B**) Schematic illustration of a DNA methylation system for dynamic modification of methylation patterns. Displayed are the enzymes on the DNA strand (top) and the corresponding base modification (bottom). Dnmt3 (orange) establishes new methylation patterns (*de novo*). Dnmt1 (green) copies the methylation mark from the maternal to the daughter strand (*maintenance*). Tet (purple) oxidizes 5-methylcytosine (5mC: orange and green) to 5-hydroxymethylcytosine (5hmC: purple) and subsequently to higher oxidation stages (*oxydation*). The oxidized 5mC is replaced by cytosine (C: black) via base excision repair mechanisms (*active* demethylation). *Higher oxidation stages catalyzed by Tet and the subsequent excision repair mechanisms are not depicted. (**C**) Distribution of the DNA methyltransferase families Dnmt1 (green) and Dnmt3 (orange) in selected vertebrates and invertebrates. The number of boxes represents the number of gene copies found in each species. Missing gene copies are depicted in gray. Vertebrates: *Xenopus laevis* (frog), *Danio rerio* (fish), *Mus musculus* (mouse) and *Homo sapiens* (Goll & Bestor, 2005). Invertebrates: *Nematostella vectensis* (sea anemone) (Zemach et al., 2010), *Ciona instetinalis* (sea squirt) (Goll & Bestor, 2005), *Crassostrea gigas* (oyster) (Xiaotong Wang et al., 2014), *Parhyale hawaiensis* (sand flea) (Kao et al., 2016), *Daphnia pulex* (water flea), *Apis mellifera* (honeybee), *Nasonia vitripennis* (wasp), *Tribolium castaneum* (beetle) and *Bombyx mori* (silkworm) (Werren et al., 2010).

1.2.2 Methylation Patterns

CpG dinucleotides are symmetric on both strands and methylation of CpG dinucleotides ensures the faithful propagation of the methylation pattern from the maternal strand to the newly synthesized daughter strand (Goll & Bestor, 2005; Song, Rechkoblit, Bestor, & Patel, 2011). Consequently, in animals, methylation is CpG-specific and symmetric on both strands (A. P. Bird, 1980). Non-CG methylation (in the context of CHG and CHH, respectively) was observed in mammalian embryonic stem cells, mammalian oocytes and plants (Ramsahoye et al., 2000; Tomizawa et al., 2011; Zemach & Zilberman, 2010). Methylation levels of CpG dinucleotides in animals display a bimodal distribution as observed in *Apis mellifera*, *Crassotrea gigas* or *Homo sapiens* (Raddatz et al., 2013; Xiaotong Wang et al., 2014), only the methylation level of CpG dinucleotides in *Bombyx mori* is unimodal with a peak at around 50 % (Xiang et al., 2010). *B. mori* is may be an exception, since it is the only known example for a Dnmt1-mediated methylome (Fig. **1.1C**)(Xiang et al., 2010). Together, the basic features of Dnmt1-Dnmt3-dependent animal methylomes are CpG-specifc, symmetric methylation (Zemach & Zilberman, 2010).

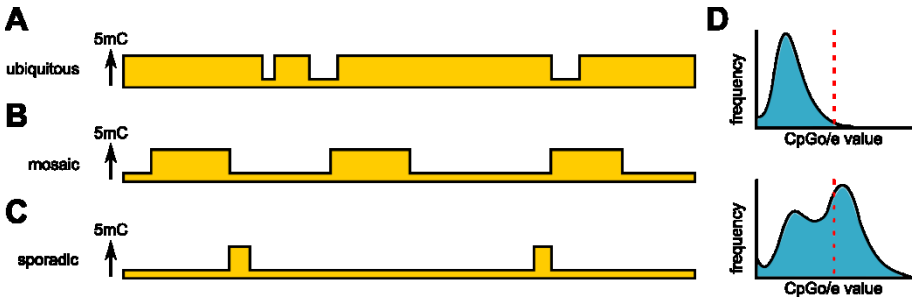


Figure 1.2 Major categories of animal methylomes.

Ubiquitously methylated genome (**A**) displaying a high CpG-depletion (**D** top) resultig in an unimodal distribution shifted towards very low CpGo/e values (below 1.0 red line) which is typical for mammalian methylomes like Homo sapiens. Mosaically (**B**) and sporadically (**C**) methylated genomes displaying moderate CpG-depletion (**D** bottom) resulting in a bimodal distribution shifted towards low CpGo/e values which is typical for invertebrates like Crassostrea gigas and Apis mellifera, respectively. Patterns and CpGo/e values were adopted from (Breiling & Lyko, 2015; Schübeler, 2015; Yi & Goodisman, 2009).

Even though animal methylomes display the same characteristics of CpG methylation, they differ in their methylation patterns. Vertebrate genomes are entirely methylated and thus, display an ubiquitous methylation pattern (Fig. **1.2A**) (Breiling & Lyko, 2015; Schübeler, 2015). In contrast, methylation in invertebrates are targeted to specific genomic elements and can be divided into mosaic and sporadic methylation patterns (Fig. **1.2B** and **1.2C**) depending on the overall amount of methlyation marks (Breiling & Lyko, 2015; Schübeler, 2015). Many insect

methylomes are defined by a small amount of methylated CpG dinucleotides (e.g. *A. mellifera*) and therefore show a sporadic methylation pattern (Breiling & Lyko, 2015).

Furthermore, methylated cytosines can spontaneously deaminate to thymines leading to a reduced amount of observed CpG dinucleotides than expected (calculated as CpGo/e value). When the C-to-T depletion occurs in the germline, it is inherited to the next generations and the fraction of depleted Cs accumulates over evolutionary time displaying the fraction of historically methylated cytosines (historical germline methylation) (Yi & Goodisman, 2009). Comparing distributions of CpG depletion in protein-coding sequences of different genomes can indicate the level of DNA methylation (Yi & Goodisman, 2009). Moreover, the CpGo/e distributions of various animals are either unimodal or bimodal and more or less shifted towards lower CpGo/e values (Fig. **1.2D**). These differences in the CpGo/e distribution also suggests different gene body methylation patterns between animals mainly vertebrates and invertebrates (Yi & Goodisman, 2009).

Gene Body Methylation

All eukaryotic methylomes, except of fungal, display methylation of gene bodies (Feng et al., 2010; Zemach et al., 2010). In plants which also methylate cytosines in the nonCG context, methylation of gene bodies is exclusively found at cytosines of CpG dinucleotides (Zemach & Zilberman, 2010). The methylation patterns of plant gene bodies is characterized by relatively high levels in the gene bodies as well as upstream and downstream of the genes with a sharp dip almost down to zero at the transcription start and termination site (TSS and TTS) (Fig. **1.3A**) (Feng et al., 2010; Zemach et al., 2010). In *Arabidopsis thaliana*, constitutively expressed genes are heavily methylated, while tissue-specific or inducible genes are less methylated (Zemach & Zilberman, 2010). Additionally, Zilberman et al. (2007) proposed a model for transcription-coupled gene body methylation in which the methylation level is the consequence of the transcription rate.

The gene body methylation patterns in vertebrates are similar to the patterns in plants, except the sharp decrease at the TSS is less distinct and the methylation levels around the TTS just decrease down to the background level downstream of the gene (Fig. **1.3B**) (Feng et al., 2010; Zemach et al., 2010). Gene body methylation in vertebrates is generally associated with gene expression, but the levels within the gene bodies only slightly correlate with the transcription rate (Zemach et al., 2010). Moreover, analysis of methylation differences in genomic features between several human cell and tissue types revealed that the variation was the lowest in the gene body and highest in enhancers and promoters (Ziller et al., 2013). Finally,

the expression of tissue-specific genes in mammals depends on methylation at regulatory regions like enhancers or promoters (Hon et al., 2013; Kundaje et al., 2015; Ziller et al., 2013).

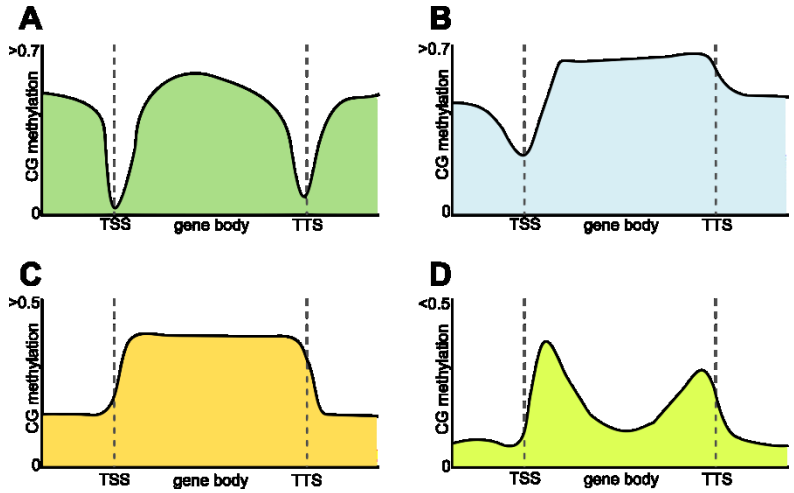


Figure 1.3 Gene body methylation patterns.

Schematic illustration of gene body methylation patterns in plants (**A**), vertebrates (**B**) and invertebrates with mosacially (**C**) and sporadically (**D**) methylated genome, respectively. The transcription start site (TSS) and transcription termination site (TTS) are depicted by dashed lines. Relative methylation levels are indicated as orientation for the order of magnitude. The figure is adopted from (Zemach et al., 2010).

Invertebrate gene body methylation patterns differ from those observed in plants and vertebrates. The methylation level upstream and downstream of the gene body is distinctly lower and increases within the gene body (Fig. **1.3C** and **1.3D**). Notably, the gene body methylation patterns in invertebrates seem to be distinguishable into two categories similar to the global methylation patterns. For example, the tunicate *Ciona intestinalis*, which possess a mosaically methylated genome, display a gene body methylation pattern resembling a plateau (Fig. **1.3C**). The methylation level increases around the TSS stays constant within the gene body and declines around the TTS down to the ground level (Feng et al., 2010; Zemach et al., 2010). In contrast, the gene body methylation in invertebrates with a sporadically methylated genome, like the honey bee *Apis mellifera* do not plateau in the gene body (Fig. **1.3D**). Furthermore, the methylation peaks shortly after the TSS and before the TTS with a minor peak around the TTS (Zemach et al., 2010). Similar to plants the methylation of gene bodies in invertebrates display an parabolic relationship to gene expression with highest methylation of moderate expressed genes (Zemach et al., 2010). Additionally, only a subset of genes is targeted by DNA methylation and several studies identified characteristics which seem to be shared among invertebrates, but their related biological function is unclear (Asselman, De Coninck, Pfrender, & De Schamphelaere, 2016; Cunningham et al., 2015; Cassandra Falckenhayn et al., 2012; Kao

et al., 2016; Lyko et al., 2010; Suzuki et al., 2013; Suzuki, Kerr, De Sousa, & Bird, 2007; Xianhui Wang et al., 2014; Xiaotong Wang et al., 2014; Xu Wang et al., 2013; Xiang et al., 2010; Zemach et al., 2010). Moreover, in *A. mellifera* methylation seems to be correlated with the outcome of alternative splicing as it was shown for one gene, but a similar correlation could not be found in *Nasonia vitripennis* (Lyko et al., 2010; Xu Wang et al., 2013). Furthermore, the majority of invertebrate methylomes were studied in insects (Cunningham et al., 2015; Cassandra Falckenhayn et al., 2012; Lyko et al., 2010; Xianhui Wang et al., 2014; Xu Wang et al., 2013; Xiang et al., 2010). After all, the functional role of gene body methylation among invertebrates remains elusive.

Repeat Methylation

While gene body methylation is a basal evolutionary feature of eukaryotic methylomes (Feng et al., 2010; Sarda et al., 2012; Zemach et al., 2010), the evolutionary conservation of repeat methylation is controversial. In plants, fungi and vertebrates methylation of transposable elements (TEs) is associated with TE silencing and thus, a key mechanism for the defense against transposable elements and maintenance of genomic stability (Zemach & Zilberman, 2010). In plants methylation in repetitive elements occurs at cytosines in each context (CG, CHG and CHH) (Feng et al., 2010; Zemach et al., 2010). The methylation level increases towards the repeat element and plateaus within the element (Fig. **1.4A**). The repeat methylation patterns in vertebrates are similar to the patterns observed in plants (Fig. **1.4B**) though, the methylation plateau is less distinct which might be due to the overall higher basal methylation level in vertebrates (Feng et al., 2010; Zemach et al., 2010).

Zemach et al. (2010) observed that repetitive elements in some invertebrates are hypomethylated displaying a methylation pattern inverse to the described patterns in plants and vertebrates (Fig. **1.4C**). They concluded that repeat methylation as TE defense was lost during early animal evolution and evolved independently in the vertebrate lineage, while in invertebrates TEs are silenced via other mechanisms (Zemach & Zilberman, 2010). Interestingly, Zemach et al. (2010) reported that repeat elements in *Ciona intestinalis* are unmethylated, whereas Feng et al. (2010) observed a moderate methyation. Moreover, several invertebrates with repeat methylation e.g. *Schistocerca gregaria* and *Crassostrea gigas* (Falckenhayn et al., 2012; Xiaotong Wang et al., 2014) and with hypomethylated repeats like *Nasonia vitripennis* were reported (Fig. **1.4D**) (Xu Wang et al., 2013). However, the relationship between repeat methylation and repeat expression was not analyzed and thus, it is unclear if the repeat methylation in those invertebrates plays a similar role in TE silencing as in vertebrates.

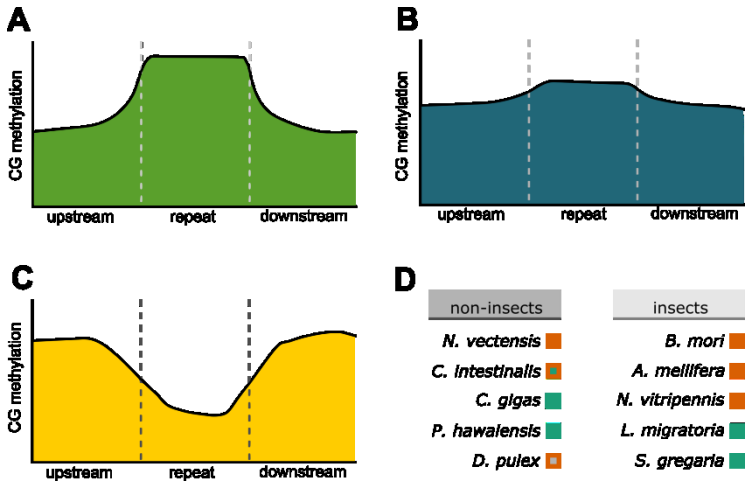


Figure 1.4 Eukaryotic repeat methylation.

Schematic illustration of repeat methylation pattern in plants (**A**) and vertebrates (**B**) as well as an example of hypomethylation in invertebrates (**C**). (**D**) Reported methylation (orange box) or hypomethylation (green box) of repetitive elements in invertebrates classified into insects (right) and non-insect species (left). Note, for *Ciona intestinalis* contradicting observations were reported about its repeat methylation (orange and green box) (Feng et al., 2010; Zemach et al., 2010). Gene body methylation, but not repeat methylation was reported for *Daphnia pulex*. Thus it is assumed that repeats in *D. pulex* are hypomethylated (orange and grey box) (Asselman et al., 2016). Figures are adopted from (Feng et al., 2010).

1.2.3 Analyzing DNA methylation

The analysis of 5-methylcytosine (5mC) started with its detection in DNA using different chromatographic procedures but also mass spectrometry (Hotchkiss, 1948; Kuo, McCune, & Gehrke, 1980; A Razin & Cedar, 1977; Aharon Razin & Sedate, 1977; Wyatt, 1950). Those methods can only detect the fraction of 5mC in the genome and thus, the discovery of the bisulfite reaction in 1970 revolutionized the analysis of DNA methylation (Fig. **1.5**).

Treatment of DNA with bisulfite leads to conversion of cytosine into uracil, while methylated cytosines remain unaffected (Hayatsu, Wataya, Kai, & Ida, 1970). When DNA is treated with bisulfite and sequenced after a PCR, methylated cytosines are still cytosines in the sequence, while unmethylated cytosines are sequenced as thymines. Thus, comparison of the bisulfite sequences to the reference sequence reveals the position of methylated and unmethylated cytosines, as unmethylated cytosines are displayed as mismatches between reference and bisulfite sequence. This principle was first applied in 1992 enabling the analysis of 5mC at single bases in several clones of a specific sequence (Fig. **1.5**) (Frommer et al., 1992). With the development of new sequencing technologies, high-throughput sequencing, the

sequencing depth of the analyzed sequence loci could be increased from several clones to hundreds of molecules in 2003 (Fig. **1.5**) (Colella et al., 2003; Tost, Dunker, & Gut, 2003). In 2008 additionally to the sequencing depth, the amount of sequenced loci increased by applying Whole-Genome Bisulfite Sequencing (WGBS) (Fig. **1.5**). The DNA is treated with bisulfite and instead selecting specific loci the whole genome is sequenced and thus, WGBS enables the analysis of the methylome at single-base resolution which is currently the gold standard (Cokus et al., 2008).

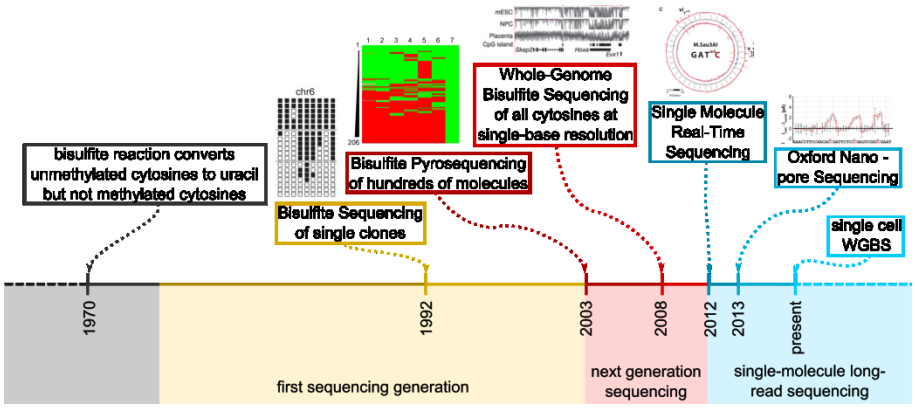


Figure 1.5 Time line of bisulfite sequencing methods.

The majority of methods are based on the bisulfite treatment of DNA, which was described in 1970 (Hayatsu et al., 1970). With the development of the first sequencing generation in 1975 (Sanger Sequencing) and 1977 (Maxam-Gilber Sequencing) the analysis of 5mC at single bases in several clones of a specific sequence started in 1992 (Frommer et al., 1992; Maxam & Gilbert, 1977; Sanger & Coulson, 1975). The next generation sequencing technologies like Pyrosequencing or Illumina sequencing enabled the analysis of hundreds of molecules first for single loci in 2003 and then for the whole genomes in 2008 (Cokus et al., 2008; Colella et al., 2003; Tost et al., 2003). The first bisulfite-free methylation analysis at single-base resolution was performed using single-molecule sequencing technologies in 2012 and 2013, respectively (Clark et al., 2012; Laszlo et al., 2013). Nowadays, the trend goes towards applying whole-genome bisulfite sequencing (WGBS) to single cells (Farlik et al., 2015; Gravina et al., 2016; Hu et al., 2016; Smallwood et al., 2014). Examples of analyzed methylation patterns above the time events are representatives for the typical visualization for the applied methods and are adopted from (Clark et al., 2012; Hon et al., 2013; Laszlo et al., 2013; Lyko et al., 2010; Meissner et al., 2005).

Nevertheless, other assays, combining bisulfite treatment with other methods like DNA methylation microarrays or RRBS (Reduced Representation Bisulfite Sequencing), have been developed and are currently used (Meissner et al., 2005; Weber et al., 2005). Nowadays, new sequencing technologies, Oxford Nano Pore (ONP) and Single Molecule Real-Time Sequencing (SMRT-Seq), are establishing which can sequence single molecules and parallel detect base modifications including 5mC and 6mA (Fig. **1.5**) (Clarke et al., 2009; Flusberg et al., 2010). However, these technologies still have some disadvantages, e.g. a relative high error rate (Laver

et al., 2015). Thus, they were applied in only few studies (Clark et al., 2012; Laszlo et al., 2013). Parallel to the single-molecule sequencing technologies, bisulfite sequencing of single cells becomes more popular (Farlik et al., 2015; Gravina et al., 2016; Hu et al., 2016; Smallwood et al., 2014). Though, the applicability of single cell sequencing is limited, as the used method is amplification biased and leads to a low genome coverage (Ning et al., 2014).

1.3 Marbled Crayfish

In 2003 Scholtz et al. described an all-female crayfish, which was first discovered in 1995 (Günter Vogt, Tolley, & Scholtz, 2004), reproducing by parthenogenesis (Fig. **1.6A**). Analysis of different microsatellite markers in various generations of this all-female crayfish revealed that it propagates apomictically (Martin, Kohlmann, & Scholtz, 2007). Apomixis is a form of thelytokous parthenogenesis in which meiosis is completely suppressed (Fig. **1.6B**) (Simon et al., 2003). Consequently, the offspring of the all-female crayfish is genetically uniform. Nevertheless, offspring of the same clutch display differences in their coloration, growth, lifespan, reproduction and behavior (Fig. **1.6C**) (Günter Vogt et al., 2008). This crayfish is the only known decapod crustacean that reproduces obligatorily parthenogenetic (Scholtz et al., 2003).

Since the taxonomic identity of the all-female crayfish was unknown, it was named marbled crayfish after its marbled carapace (Scholtz et al., 2003). Scholtz et al. (2003) could classify the marbled crayfish as member of the North American Cambaridae family. Since then, several authors considered the marbled crayfish as parthenogenetic *Procambarus alleni* (G. Vogt, 2008), while others suggested *Procambarus fallax* as its sexually reproducing ancestor (Scholtz et al., 2003). To clarify the taxonomic status of the marbled crayfish, Martin et al. (2010) compared morphological features and two mitochondrial loci (cytochrome c oxidase subunit I and 12S rRNA) of marbled crayfish with several *P. alleni* and *P. fallax* individuals from wild populations in Florida, USA. The marbled crayfish was morphologically indistinguishable from to *P. fallax* and the divergence in the mitochondrial loci between *P. alleni* and marbled crayfish was ten times higher than between *P. fallax* and marbled crayfish (Martin et al., 2010). Thus, Martin et al. (2010) concluded that the marbled crayfish is the parthenogenetic form of *P. fallax* and suggested *Procambarus fallax* f. *virginalis* as its preliminary taxonomic name.

Even though *P. fallax* is native to Florida and southern Georgia, USA (Crandall, 2010), wild populations of marbled crayfish developed from releases in Madagascar and various European countries like Germany and Sweden (Fig. **1.6D**) (Bohman et al., 2013; Chucholl & Pfeiffer, 2010; J. P. G. Jones et al., 2009; Lőkkös et al., 2016; Novitsky & Son, 2016). Notably, the annual temperature differences between Madagascar (19.5 ± 2.7°C, Antananarivo) and

Sweden (6.6 ± 7.2°C, Stockholm) are enormous (World Weather Online, 2012a, 2012b). Additionally, marbled crayfish occur in both lentic and lotic freshwater habitats including rivers, lakes, fish ponds, swamps, rice paddies, brick pits and drainage ditches (Heimer, 2010; J. P. G. Jones et al., 2009). Moreover, a marbled crayfish population was found in a pit mine lake which was a former soft coal opencast mining (Dümpelmann & Bonacker, 2012). The water of the lake has an increased level of sulfur of 640 - 740 mg/l (normal waters: 25 - 50 mg/l) and a decreased pH of 3.9 - 4.2 (*02.01 Gewässergüte Chemie*, 2004; Dümpelmann & Bonacker, 2012). Thus, the genetically uniform marbled crayfish seems to be capable to adapt to a broader variety of habitats than its sexually reproducing ancestor *P. fallax*.

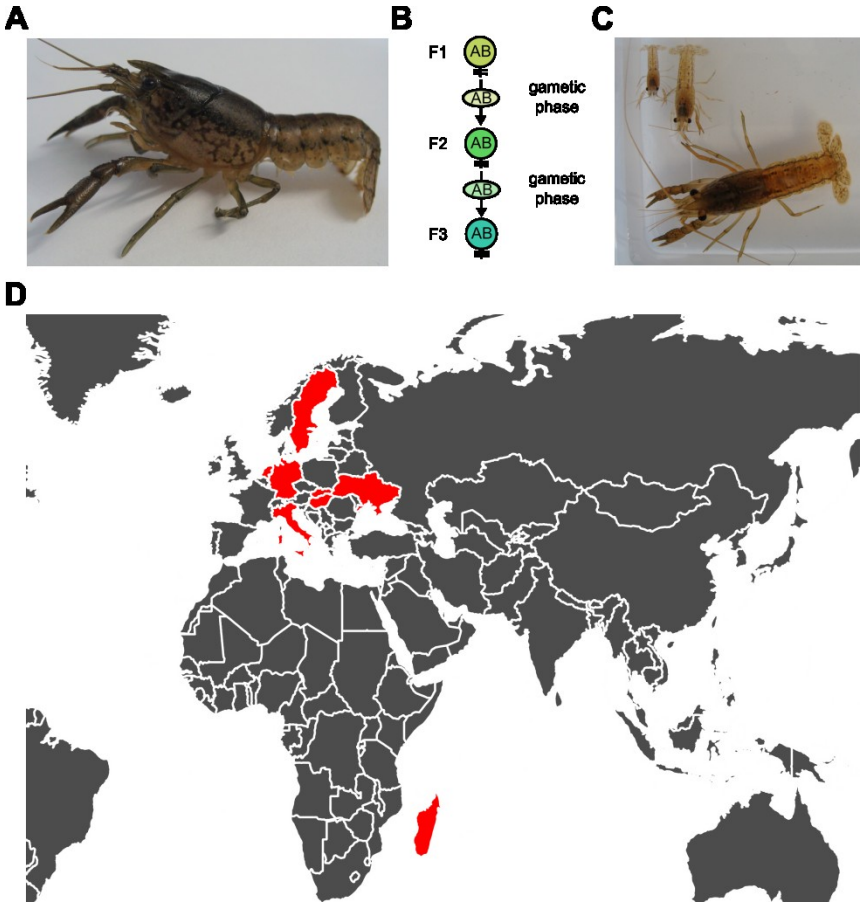


Figure 1.6 Size differences and distribution of the parthenogenetic marbled crayfish.
(**A**) Picture of an adult marbled crayfish specimen. (**B**) Two different alleles (A and B) of a gene during three generations (F1-F3) of apomixis, a mode of thelytokous parthenogenesis (adapted from Martin, Kohlmann & Scholtz, 2007). (**C**) Size differences between coeval offspring of the same clutch, reared together. (**D**) Global distribution of marbled crayfish. Countries with occurrences of marbled crayfish are highlighted in red (Bohman et al., 2013; Chucholl & Pfeiffer, 2010; Holdich & Pöckl, 2007; J. P. G. Jones et al., 2009; Kawai & Takahata, 2010; Liptak et al., 2016; Lőkkös et al., 2016; Marzano et al., 2009; Novitsky & Son, 2016).

1.4 Aims of the PhD Thesis

Vertebrates and invertebrates share key features of DNA methylation, but they differ in their methylation patterns indicating that the DNA methylation in invertebrates may has a different role as in mammals (Feng et al., 2010; Zemach et al., 2010). The only crustacean species among the analyzed non-insect invertebrate methylomes are the water flea *Daphnia pulex* and the sand flea *Parhyale hawaiensis* (Asselman et al., 2016; Kao et al., 2016). However, both studies give just a small insight into the methylome of crustaceans. Many crustaceans are keystone species with ecological and environmental relevance for their habitats (Colbourne et al., 2011; Günter Vogt, 2008). The marbled crayfish reproduces parthenogenetically with a high quantity of eggs per clutch and lives in a wide range of habitats demanding minor standards to the water quality compared to other crustaceans (G. Vogt, 2008). Thus, the marbled crayfish has the necessary attributes to be a laboratory model organism.

To broaden the knowledge about DNA methlyation in crustaceans, the main aim of this doctoral thesis was to characterize the methylome of the marbled crayfish at single-base resolution using whole-genome bisulfite sequencing. Performing a detailed analysis of the gene body and repeat methylation patterns will give new insights into the evolutionary conservation of DNA methylation among invertebrates. The findings will help to establish the marbled crayfish as new model organism for epigenetics. Besides the main aim, this work had two additional aims: first, to clarify the taxonomic status of the marbled crayfish and second, to assemble the marbled crayfish transcriptome as basis for molecular biological and bioinformatic analysis.

2 Materials and Methods

2.1 Equipment

- BD Accuri C6 Flow Cytometer (BD Biociences)
- BioPhotometer (Eppendorf)
- Centrifuge 5415D (Eppendorf)
- Centrifuge 5415R (Eppendorf)
- Centrifuge 5804R (Eppendorf)
- FLUOstar Optima (BMG Labtech)
- Genomic ScreenTape (Agilent)
- GS Junior 454 Sequencing (Roche)
- NanoDrop 2000 (Thermo Scientific)
- Needle 0.5 x 22 mm (Terumo)
- Real Time PCR System, LightCycler 480 (Roche)
- RNA ScreenTape (Agilent)
- Sterile filter 0.45 µm (Sarstedt)
- Syringe 1ml (Ersta)
- TapeStation 2200 (Agilent)
- Thermoxycler, DNA Engine (BioRad)
- TissueRuptor (Qiagen)
- 384-well plates (Steinbrenner)

2.2 Chemicals, Buffers and Reaction Kits

2.2.1 Chemicals

- Absolute QPCR SYBR Green Mix (Thermo Scientific)
- Acetic acid (Merck)
- Agarose (Roth)
- Boric acid (Sigma-Aldrich)
- Chloroform (VWR)
- Citric acid (Riedel-de Haen)
- Complete Mini Protease Inhibitor Cocktail (Roche)
- DMSO (Sigma-Aldrich)

- DNase-free, RNase-free Water (gibco Life Technologies)
- DTT (Gerbu)
- dNTPs (Fermentas Life Sciences)
- EDTA (Gerbu)
- Ethanol (Sigma-Aldrich)
- Glucose (Applican)
- Igepal / NP-40 (Sigma-Aldrich)
- Isopropanol (Sigma-Aldrich)
- Oligo$(dT)_{20}$ (Invitrogen)
- PBS 1x (gibco Life Technologies)
- PicoGreen (molecular probes Life Technologies)
- Propidium Iodide (PI) 1 mg/ml (Life Technologies Thermo Fisher Scientific)
- Protein Assay Dye Reagent Concentrate (Bio-Rad)
- Proteinase K (Ambion)
- ReadyMix PCR buffer (Thermo Fisher Scientific)
- RNase A 50 mg/ml (Sigma-Aldrich)
- SDS (Roth)
- Sodium Chloride (Sigma-Aldrich)
- Sodium Desoxycholate (Sigma-Aldrich)
- Taq-Polymerase ThermoPrime Plus DNA Polymerase (Thermo Fisher Scientific)
- TE 20x (molecular probes Life Technologies)
- Tris (Sigma-Aldrich)
- Trisodium Citrate Dihydrate (Sigma-Aldrich)
- Trizol (Ambion)

2.2.2 Buffers

- Crayfish Anticoagulant: 100 mM Glucose, 34 mM Trisodium Citrate, 26 mM Citric acid, 15.8 mM EDTA, pH 4.6
- Pre-Lyses Buffer: 10 mM Tris pH 7.5, 5 mM EDTA pH 8.0, 10 mM NaCl
- RIPA Buffer: 0.1 % SDS, 0.5 % Igepal (NP-40), 0.5 % Sodium Desoxycholate in 1x PBS, 1 mM DTT, 1 tablet Complete Mini Protease Inhibitor Cocktail (for 10 ml buffer)
- TAE 1x Buffer: 40 mM Tris pH 7.6, 20 mM acetic acid, 1 mM EDTA
- TBE 1x Buffer: 89 mM Tris pH 7.6, 89 mM boric acid, 2 mM EDTA

2.2.3 Reaction Kits

- Blood & Cell Culture Kit (Qiagen)
- DNeasy Blood & Tissue Kit (Qiagen)
- EpiTec Bisulfite Kit (Qiagen)
- QIAquick Gel Extraction Kit (Qiagen)
- QIAquick PCR Purification Kit (Qiagen)
- QuantiTect Reverse Transcription Kit (Qiagen)
- RNeasy Mini Kit (Qiagen)

2.3 Software

- BLAST (Camacho et al., 2009)
- Bowtie2 (Langmead & Salzberg, 2012)
- BSMAP version 2.73 (Xi & Li, 2009)
- BUSCO (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015)
- CAP3 (Huang & Madan, 1999)
- CD-Search (Marchler-Bauer & Bryant, 2004)
- CD-HIT-EST (W. Li & Godzik, 2006)
- ExPASy translate tool (Gasteiger et al., 2003)
- FastUniq (Xu et al., 2012)
- MAKER (Cantarel et al., 2008)
- MITObim1.6 (Hahn, Bachmann, & Chevreux, 2013)
- QuickGO (Dimmer et al., 2008)
- R (R Core Development Team, 2013)
- RepeatMasker (Smit, Hubley, & Green, 2013)
- RPSBLAST (Marchler-Bauer et al., 2002)
- RSEM (B. Li & Dewey, 2011)
- SAMtools (H. Li, 2011; H. Li et al., 2009)
- SOAPdenovo-Trans version 1.03 (Xie et al., 2014)
- Transcriptome Computational Workbench (Soderlund, Nelson, Willer, & Gang, 2013)
- Velvet 2.0 (Zerbino & Birney, 2008)

2.4 Marbled Crayfish Handling

2.4.1 Marbled Crayfish Strains and Culture Conditions

Two laboratory strains were established: Heidelberg founded in 2003 from a single female originated from the first described marbled crayfish population established by F. Steuerwald in 1995 (Günter Vogt et al., 2004) and Petshop founded by a female marbled crayfish purchased in the German pet shop “Kölle Zoo” in 2004. Additionally, individuals from two wild populations were caught: Moosweiher from the lake Moosweiher near Freiburg (provided by M. Pfeiffer) first described in 2009 (Chucholl & Pfeiffer, 2010) and Madagascar from Antananarivo, Madagascar, southeast Africa (provided by F. Glaw) first described in 2007 (J. P. G. Jones et al., 2009). Individuals of *Procambarus fallax* and *Procambarus alleni* used in this study were brought from German aquarium traders in 2013 and 2014, respectively. Crayfish were kept either communally or individually in 18.90 x 54.80 x 38.40 cm (H x D x W) plastic boxes. The boxes were filled with tap water, gravel and potsherd as shelters (Fig. **2.1A**). The room temperature was constant at 25 °C and the water temperature at around 20 °C. A natural light-dark cycle was applied. All juveniles and adult animals were daily fed with TetraWafer Mix pellets.

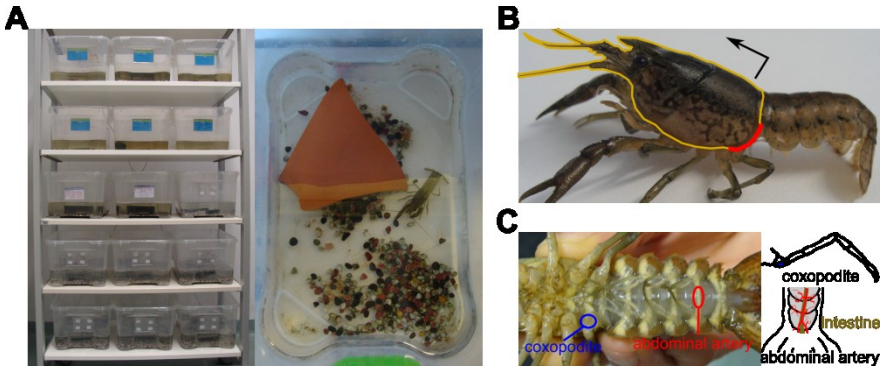


Figure 2.1 Marbled crayfish handling.

(**A**) Culture conditions. Left: overview of the marbled crayfish laboratory population. Right: Plastic box filled with tap water, gravel and potsherd as shelter. (**B**) Dissection of the marbled crayfish by lifting up (direction: black arrow) the carapace (orange). The area to position the finger underneath the carapace is indicated in red. (**C**) Extraction of hemocytes from two sides: coxopodite of the 4th pereopod (blue circle and right top schemata) and abdominal artery (red circle and right bottom schemata). Left : general overview of the positions at the crayfish body. Right top: schematic illustration of the crayfish leg segments. Right bottom: schematic illustration of the crayfish abdominal artery.

2.4.2 Tissue Dissection

First, crayfish were dapped with tissue to remove excessive water, then their body weight, the total length and the carapace length was recorded. Next, the crayfish were fixated by wrapping them in paper towel covering their eyes. Then, the carapace was lifted up separating the head from the thorax and abdomen (Fig. **2.1B**). The pereopods (walking legs) and the claws were cut with a scalpel to facilitate access to the organs. Sterile forceps were used to extract surgically the tissue. First of all, the weight of ovaries and hepatopancreas was recorded. Afterwards, the tissue was divided in equal parts and immediately frozen in liquid nitrogen. The abdominal musculature was extracted from the chitinous exoskeleton and the intestine was carefully removed, before it was divided and frozen. Tissue was stored at -80°C until extraction of DNA, RNA or proteins was performed.

2.5 Flowcytometric Analyses

2.5.1 Hemocytes Isolation

Hemocytes of *P. virginalis* were extracted from the ventral abdominal artery or from the coxopodite of the 4th pereopod (walking leg; Fig. **2.1C**) using a 0.5x25 mm needle and 1 ml syringe filled with 100 µl crayfish anticoagulant (100 mM Glucose, 34 mM Trisodium Citrate, 26 mM Citric acid, 15.8 mM EDTA, pH 4.6). After centrifugation for 5 min at 1,400 rpm the pellet was washed in 1x PBS and centrifugated again under the same condition. The pellet was resuspended in 1 x PBS with 10% DMSO and aliquoted for storage at -80°C.

2.5.2 Peripheral Blood Cell Isolation

Human and mouse whole-blood samples were mixed 1:1 with 1x PBS and then gently layered over a Ficoll-Hypaque solution (in a ratio 3 parts Ficoll : 10 parts blood mixture). The centrifugation was performed for 20 min at 400xg with the slowest acceleration rate and brake off. The upper aqueous plasma phase was removed and the underlying phase was transferred to a new reaction tube. After adding 3 volumes of Hank's balanced salt solution (HBSS), the samples were centrifuged for 10 min at 400xg. The washing step with HBSS was repeated and the pellets were resuspended in 1x PBS. Aliquots were either used immediately for flowcytometry or stored at -80°C in 10% DMSO.

2.5.3 Genome Size Estimation

Aliquotes of 100 µl prepared cells were gently thawed on ice and equilibrated to room temperature before adding 2 µl RNase A stock solution (50 mg/ml) and 5 µl Propidium Iodid stock solution (1 mg/ml). After incubation for 30 min, the samples were diluted with 100 µl 1x PBS and shortly mixed. Propidium Iodide stained cells were counted and fluorescence intensity per cell was measured using a BD Accuri C6 Flow Cytometer with a 488 nm laser and the standard 585 nm filter (detector FL2). After determining the cell density of each sample (cell counts / µl), the same amount of stained cells from different organisms were mixed together and analyzed again with the flow cytometer. The genome size (GS) was calculated by proportioning the median fluorescence signal (FS) of stained cells per haploid genome multiplied with the known genome size of the used standard (in bp; formula **2.1**).

Formula 2.1: $$\mathrm{GS_B} = \frac{\mathrm{median(FS_B)}}{\mathrm{median(FS_A)}} \cdot \frac{\mathrm{ploidy\ level_A}}{\mathrm{ploidy\ level_B}} \cdot \mathrm{GS_A\ [bp]}$$

A = used standard species; B = species with unknown gnome size;
GS = genome size; FS = fluorescence signal

2.6 Nucleic Acids Analyses

2.6.1 DNA Extraction and Quality Control

Genomic DNA was isolated from frozen tissue using either DNeasy Blood & Tissue Kit (Qiagen), Blood & Cell Culture Kit (Qiagen) or the Lyses Protocol. The tissue was homogenized in lyses buffer of the corresponding protocol using the TissueRuptor (Qiagen). DNA extraction with the DNeasy Blood & Tissue Kit (Qiagen) was performed according to the manufacturer's instructions.

The protocol QIAGEN Genomic DNA Handbook of the Blood & Cell Culture Kit (Qiagen) was slightly changed as follows: The tissue was homogenized in lyses buffer. RNase A and Proteinase K was added to the sample and the mix was incubated at 53 °C for 1 h. Precipitation was performed with Isoporpanol and centrifugation at 6,000rcf.

Following the Lyses Protocol, 4.5 ml pre-lyses buffer (10 mM Tris pH 7.5, 5 mM EDTA pH 8.0, 10 mM NaCl) was mixed with 25 µl 50 mg/ml RNase A, 25 µl 20 mg/ml Proteinase K and 500 µl 10% SDS. The homogenized tissue was incubated either at 37 °C over night or at 55 °C for 5 h. After adding 2.5 ml 5 M NaCl, the samples were centrifuged for 15 min at full speed and 4 °C. The aqueous phase was aliquoted to 1.5 ml tubes and centrifuged again for 15 min at higher speed to pelletise the remaining fine particles. The clear aqueous phase was pooled in a new

15 ml tube and 5.6 ml Isopropanol was added. After mixing, the samples were centrifuged for 10 min at full speed. Pellets were washed with 70 % Ethanol and transferred to a new 1.5 ml tube and subsequently centrifuged at maximum speed for 5 min. Pellets were resolved in 25-100 µl DNase-free water.

The quality of isolated genomic DNA was assessed via 8% TBE/TAE-Agarose Gel (1% (w/v) Agarose, 1x TBE) and/or via 2200 TapeStation (Agilent) following manufactuerer's instructions for Genomic ScreenTape. The concentration was determined either via NanoDrop 2000 (Thermo Scientific) following the manufacturer's instructions or PicoGreen. A DNA standard serial dilution (1.56 ng, 3.125 ng, 6.25 ng, 12.5 ng, 25 ng, 50 ng and 100 ng) in 1x TE was freshly prepared for each PicoGreen measurement. The DNA standard dilution and DNA samples (1 µl in 99 µl 1x TE) were measured in triplicates. To each 100 µl DNA solution, 100 µl freshly prepared PicoGreen (1:200 in 1x TE) was added and the fluorescence signals at 520 nm (extension 485 nm, emission 520 nm) were detected by FLUOstar Optima (BMG Labtech). The DNA concentration of the sample was determined relative to the DNA standard serial dilution.

2.6.2 RNA Extraction and Quality Control

Total RNA was isolated from frozen tissues with a sample size of 20 - 60 mg. Thawed tissues were homogenized in 1 ml Trizol and heavily shook after adding 200 µl Chloroform. The samples were incubated for 5 min at room temperature (RT), before centrifugation for 15 min at 12,000rcf and 4 °C. Then the upper aqueous phase was transferred into a new reaction tube and 1 volume of Isopropanol was added. After precipitation for 1 h on ice, the samples were centrifuged for 30 min at 16,000rcf and 4 °C. The pellets were washed with 70% Ethanol and resuspended in 20 - 100 µl RNase-free water. Total RNA was treated with DNase using the RNeasy Mini Kit (Qiagen) following the manufacturer's RNeasy Mini Protocol for RNA Cleanup in combination with the On-Column DNase Digestion Protocol. Quality of extracted RNA was assessed via 2200 TapeStation (Agilent) following manufacturer's instructions for RNA ScreenTape. The Concentration was determined via NanoDrop 2000 (Thermo Scientific) following manufacturer's instructions.

2.6.3 Quantitative Real Time Polymerase Chain Reaction (qRT-PCR)

Reverse transcription was performed using the QuantiTect Reverse Transcription Kit (Qiagen). In a first step, 1 µg of total RNA was mixed with 2 µl 7x gDNA Wipeout buffer and 14 µl DNase-free, RNase-free water and incubated for 5 min at 42 °C. In the second step, 4 µl 5x reverse transcriptase buffer, 1 µl 50 µM Oligo$(dT)_{20}$ primers and 1 µl reverse transcriptase

were added to the incubated RNA mixture and heated for 30 min at 42 °C followed by 15 min at 95 °C. The cDNA was then stored at -20 °C or immediately used for qRT-PCR analyses using the Absolute QPCR SYBR Green Mix. Shortly one qRT-PCR reaction consisted of 1 µl of cDNA, 5 µl 2x QPCR SYBR Green Mix, 3.6 µl water, 0.2 µl 10 µM forward primer and 0.2 µl 10 µM reverse primer (primer are listed in table **2.1** and corresponding amplicon sequences and location within the target enzyme are in Fig. **5.1**). The samples were measured on a LightCycler 480 (Roche) as triplicates in a 384-well plate. qRT-PCR conditions were as follows: denaturating for 15 min at 95 min, 40 cycles (10 sec at 95 °C followed by 30 sec at 60 °C), melting at 95 °C and cooling for 10 min at 40 °C. The data analyses were performed with the provided LigthCycler 480 software (Roche).

Table 2.1 Primer Sequences used for qRT-PCR.

Corresponding amplicon sequences and location within targeted enzyme are shown in the Appendix.

Primer			Amplicon		Targeted Enzyme	
ID	**type**	**5'-3' sequence**	**Name**	**length [bp]**	**Name**	**Domain**
CasF_027	forward	CCACAGCTACAGAACATCG	TBP2	122	TATAbox BP	TBP
CasF_028	reverse	CTCATGATGACGGCTGC				
CasF_007	forward	GGGAGAAGGCACTGATTGG	Dnmt1.2	150	Dnmt1	Dnmt1-RFD
CasF_008	reverse	CGATCATCGTTGTTCACCAG				
CasF_009	forward	GAATGGAACATCAGCACCTGC	Dnmt3.1	133	Dnmt3	PWWP
CasF_010	reverse	CGGTGCTCTCATTCCACAATC				
CasF_025	forward	CCAGTAGAAGTGATCAACAGTG	Tet3	100	Tet	Tet_JBP
CasF_026	reverse	CCTCCAATATCTGGATCGTGG				

2.6.4 High Throughput Sequencing

Library preparation and sequencing was performed either by the High Throughput Sequencing Unit of the Genomics and Proteomics Core Facility at the DKFZ or by Eurofins MWG GmbH (Ebersberg, Germany). The following sequencing approaches were performed and a detailed overview of the used tissues and individuals are listed in table **2.2**. Data sets which were not used for the analyses are listed in the appendix.

WGS^{DKFZ}: Core Facility R & D protocol for genomic DNA as starting material was used for library preparation. The selected fragment size was 300 bp. The library was sequenced on an Illumina HiSeq V3 platform in paired-end mode and 100 bp read length.

Bi-Seq^{DKFZ}: Whole-genome bisulfite sequencing was performed with the R & D protocol of the Core Facility for genomic DNA as starting material. The library fragment size of 300 bp was selected and sequenced on Illumina HiSeq V3 platform in paired-end mode and 100 bp read length. Corresponding base coverage was calculated as described in section 2.8.4 and are listed in table **5.2** in the appendix.

***:** The library was produced in the same way as for WGS^{DKFZ} and Bi-Seq^{DKFZ}, respectively. Sequencing platform was Illumina HiSeqX in paired-end mode and read length of 150 bp.

Table 2.2 Overview of sequenced samples.

Samples are listed per animal ID, sequencing approach and tissue. WGS: whole-genome sequencing. BiSeq: whole-genome bisulfite sequencing. RNA-Seq: whole transcriptome sequencing. Hepato: hepatopancreas. Haem: haematopoietic tissue. Antennal: antennal glands (green glands). abdM: abdominal muscle.

Species	strain/ sex	animal ID	tissue	seqtype
P. virginalis	Heidelberg	HD1	abdM	Bi-Seq^{DKFZ}
			hepato	Bi-Seq^{DKFZ}
		HD2	abdM	WGS^{DKFZ}
			hepato	Bi-Seq^{DKFZ}
				RNA-Seq^{DKFZ}
			hepato, haem, antennal, abdM	RNA-Seq^{MWG}
	Petshop	Pet1	abdM	WGS^{MWG}
	Moosweiher	MW1	abdM	WGS^{DKFZ}
			hepato	Bi-Seq^{DKFZ}
			gills	Bi-Seq^{DKFZ*}
	Madagascar	Mad1	abdM	WGS^{DKFZ}
P. fallax	female	PFF1	abdM, hepato	WGS^{DKFZ}
			hepato	Bi-Seq^{DKFZ}
		PFF4	abdM	Bi-Seq^{DKFZ}
			hepato	Bi-Seq^{DKFZ}
P. alleni	female	PAF1	abdM, hepato	WGS^{DKFZ}

RNA-Seq^{DKFZ}: The library preparation was performed with the Core Facility R & D protocol for totalRNA as starting material. Following platforms were used for the sequencing: Illumina HiSeq V3 in paired-end mode with a read length of 100 bp and Illumina HiSeq V3 in paired-end mode with a read length of 125 bp, respectively.

WGS^{MWG}: Sample, which was whole-genome sequenced by Eurofins MWG GmbH (Ebersberg, Germany), was part of the *P. virginalis* genome assembly project of Julian Gutekunst. However, reads were also used for the comparison of the mitochondrial DNA sequences. For this reason, the sample is listed in table **2.2** as well.

RNA-Seq^{MWG}: First, from totalRNA poly(A)+RNA was isolated, which was used for library preparation. Then, the cDNA library was normalized by one cycle of denaturation followed by re-association. After PCR amplification of the normalized ss-cDNA (single stranded cDNA), the library was size fractionated in the range of 500 to 1,200 bp. High throughput sequencing was performed on Illumina MiSeq in paired end mode and read length of 250 bp.

2.7 Protein Analyses

2.7.1 Protein Extraction

Tissue samples of hepatopancreas were used for whole protein extraction. The samples were homogenized in RIPA Buffer (0.1 % SDS, 0.5 % Igepal, 0.5 % Sodium Desoxycholate in 1x PBS, 1 mM DTT, Complete Mini Protease Inhibitor Cocktail) and incubated on ice for 30 min, followed by centrifugation at maximum speed for 30 min. Supernatant was filtrated through a 0.45 µm sterile filter and the protein concentration was determined by Bradford dye assay.

Protein samples were diluted 1:800 for the Bradford dye assay. To each prepared dilution, 1/4th volume of Protein Assay Dye Reagent Concentrate (Bio-Rad) was added and vortexed for 10 sec. After incubation at room temperature the absorption at 595 nm was measured with BioPhotometer and the concentration was determined in relation to a protein standard serial dilution.

2.7.2 Protein Mas-spectrometric Analyses

Protein extracts of hepatopancreas from nine individuals (three of each marbled crayfish strain Heidelberg, Petshop and Moosweiher) were treated in two different ways for mas-spectrometric analyses (performed by Oliver Popp). One half of the protein extract was fractionated by molecular weight into six fractions using a SDS-gel. Each fraction was then measured in a label-free quantification (LFQ) approach. The other half of the protein extract was labeled with dimethyl distinguishing three groups based on the marbled crayfish strain: light +28 Da (Moosweiher), medium +32 Da (Heidelberg) and heavy +36 Da (Pethop). Then, individuals with the same ID of each strain were mixed and measured together in a single run as dimethyl labeling (DML) approach.

For measurement of samples from both approaches, the proteins were treated as follows (performed by Oliver Popp). The disulfide bridges of the proteins were broken down by treatment with TCEP (Tris-2-carboxyethyl-phosphin) and the secondary structure by alkylation with Chloroacetamide. Finally the proteins were digested with trypsin. After this, the samples were used for hydrophobe reverse high pressure liquid chromatography coupled to a two-dimensional (LC-MS/MS) Q-Exactive Orbital-rap mass spectrometer (Thermo Fisher Scientific).
The MaxQuant and PEAKS software was used to analyze the recorded MS-files (performed by Oliver Popp). For analyses with MaxQuant following settings were used: carboamidomethyl as fixed modification, oxidation as variable modification, 1 % false positive rate and Orbitrap as used instrument adjustment. For identification of false positives and contaminants a database with reverted protein sequences and with typical contamination proteins was used by Oliver Popp.

2.8 Bioinformatical Analyses

2.8.1 Mitochondrial DNA Analyses

Genomic DNA was isolated from hepatopancreas and abdominal musculature (section 2.6.1) and sequenced on an Illumina HiSeq platform. Read pairs were trimmed according to their quality value (minimum quality value ≥ 30) and filtered by their length (minimum length ≥ 30 bp). The reference mitochondrial genome of the *P. virginalis* Heidelberg strain was assembled with Velvet 2.0 using the following settings for paired-end read libaries: kmer size 23, insert size 300, minimum coverage 5, expected coverage 10. Mitochondrial sequences of *P. fallax* and *P. alleni* were assembled by MITObim1.6. As seed sequences published mitochondrial DNA fragments from *P. fallax* (FJ619800) and *P. alleni* (HQ171462, FJ619802, HQ171451) were used for the assembly.
The assembled mitochondrial DNA of *P. virginalis* was annotated by BLASTx and BLASTn search against the protein and nucleotide sequences of the annotated *P. clarkii* mitochondrial genome (JX316743). To identify single nucleotide polymorphisms (SNPs) between the marbled crayfish populations, the sequences of Petshop, Moosweiher and Madagascar specimens were established by mapping the quality trimmed reads against the assembled mitochondrial DNA of the Heidelberg strain using Bowtie2. SNP calling was performed with mpileup and bcftools from SAMtools with a minimum quality value > 30. Mitochondrial sequences of *P. fallax* and *P. alleni* were compared to the sequence of *P. virignalis* identifying the mismatches by BLASTn alignments.

2.8.2 Transcriptome Analyses

Tanscriptome Assembly and Quality Control

Isolated total RNA from hepatopancreas, abdominal musculature, hematopoetic tissue and green glands (section 2.6.2) was mixed to equal parts and sequenced by Eurofins MWG GmbH (Ebersberg, Germany). Parallel, total RNA from hepatopancreas was isolated and sequenced on an Illumina HiSeq platform (table **2.2**). Both data sets were treaded separately and assembled as follows. Duplicated reads were removed using FastUniq and read pairs were quality trimmed (minimum quality value ≥ 20 and minimum length ≥ 50 bp). SOAPdenovo-Trans-127mer was used to assemble the transcriptome with kmer sizes in the range from 19 to 63 (Fig. **2.2**) and insert size 200. Firstly, all generated scaffold sequences without gaps were used for further filtering. Scaffolds with gap sequences were doubled checked and wrongly inserted gaps were removed. Wrongly inserted gaps were identified by the contigs used for the particular scaffold, which only perfectly matched the scaffold sequence without the gap region.

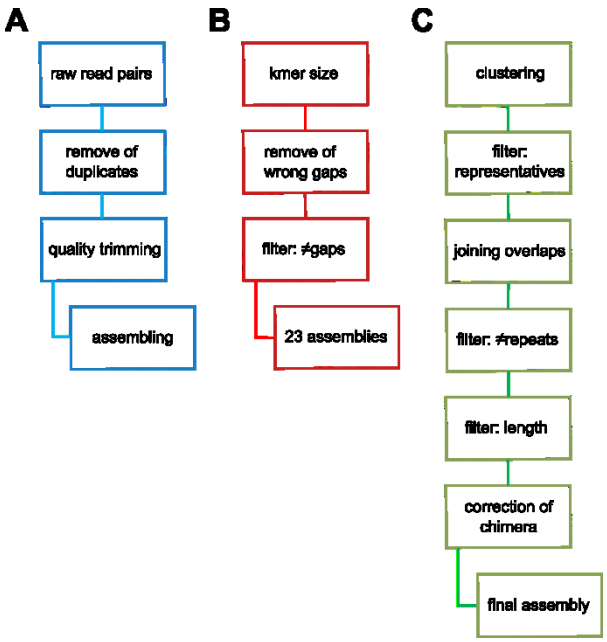


Figure 2.2 Work flow of the transcriptome assembly.

Raw reads generated by the DKFZ Core Facility and a company were treated separately. (**A**) Raw read processing. (**B**) Assembly of processed reads using kmer sizes from 19 to 63 and post-processing of each of the 23 assemblies. (**C**) Combining all the 46 generated assemblies and filtering of convincing transcripts.

Then, the 23 generated transcriptome assemblies from both data sets were merged together by clustering the transcript sequences with 97 % sequence identity using CD-HIT-EST. The longest sequence of the clusters were kept as cluster representative for further filtering. Overlapping

transcripts from different assemblies were joined into one single transcript using CAP3. Repetitive regions were identified with RepeatMasker and transcript sequences with repeats representing more than 10 % of the total sequence length were removed as source for possible miss-assembly. Finally, transcripts with a minimum length of 300 bp were used as the final assembly and analyzed for cis-self and trans-self chimera. Chimeras are produced during the process of *de novo* assembly of transcriptomes (Yang & Smith, 2013). While chimeric multi-genes are transcripts containing two different genes, cis-self (same strand orientation) and trans-self (opposite strand orientation) chimeras are transcript sequences repeating the same gene (Yang & Smith, 2013). Cis-self and trans-self chimera were identified by splitting the transcript sequences in two equal parts and aligning them to each other using BLASTn. Identified chimera were corrected when possible and incorporated to the final transcriptome assembly.

The quality of the assembled transcriptome was assessed by determining the completeness of 2,675 orthologous genes conserved among arthropods using BUSCO. Shortly, BUSCO performs a sequence comparison using either cDNA sequences or protein-coding sequences as input. The *P. virginalis* transcriptome was then ranked together with transcriptomes of 13 other species analyzed in the same way: *Drosophila melanogaster* (EnsemblMetazoa assembly version 6), *Anopheles gambiae* (EnsemblMetazoa assembly version 4), *Apis mellifera* (EnsemblMetazoa assembly version 4), *Tribolium castaneum* (EnsemblMetazoa assembly version 3), *Aedes aegypti* (EnsemblMetazoa assembly version 3), *Acyrthosiphon pisum* (EnsemblMetazoa assembly version 2), *Nasonia vitripennis* (EnsemblMetazoa assembly version 2), *Daphnia pulex* (EnsemblMetazoa assembly version 1), *Bombyx mori* (EnsemblMetazoa assembly version 1), *Ixodes scapularis* (EnsemblMetazoa assembly version 1), *Lepeophtheirus salmonis* (EnsemblMetazoa assembly version 1), *Litopenaeus vannamei* (TSA assembly version 1, accession numbers JP355723-JP376614, JP382831-JP435443) and *Astacus leptodactylus* (TSA assembly version 1, acession number GAFY00000000.1).

Mass-spectrometric analyses of protein extracts from hepatopancreas was used to assess the quality of the assembled transcriptome as a second approach (section 2.7). The measured MS-spectra were translated into peptide sequences using two different softwares MaxQuant and PEAKS (done by Oliver Popp). MaxQuant predicts the peptide sequences based on a provided protein database, while PEAKS performes a *de novo* peptide calling. The reported peptides were filtered by the quality value of the peptide call (PEP value ≤ 0.1 and ALC ≥ 50, respectively). The remaining peptides of the MaxQuant call were classified according to their matched protein hits into contaminants (with hits in the database listing contaminants), false positives (with hits in the reverted protein database of *P. virginalis*), proteins (peptides with a unique hit in the corresponding *P. virginalis* database) and paralogues/ splice variants (with

multiple hits in *P. virginalis* database). The remaining peptides of the PEAKS call were mapped to the *P. virginalis* transcriptome using BLASTp (e-value ≤ 0.001). The portion of identified proteins in the *P. virginalis* transcriptome was determined by the amount of unique protein hits extracted from the list of protein groups divided by the total amount of proteins in the *P. virginalis* transcriptome database.

Transcriptome Annotation

Using the automated annotation pipeline from Transcriptome Computational Workbench the *P. virginalis* transcript sequences were annotated with Universal Protein Rescource (UniProt) terms (performed by Julian Gutekunst). UniProt terms were then linked to their Gene Ontology (GO) terms by applying QuickGO. For annotation with Clusters of Orhtologous Groups (COG) the COG database was downloaded and *P. virginalis* sequences were annotated using RSPBLAST. Annotation with the Kyoto Encyclopedia of Genes and Genomes (KEGG) was performed using the tool provided on the official website.

For phylogenetic analysis of the assembled *P. virginalis* transcriptome one species was selected to generate a database representing one of the following phylostrata: bilateria (*Xenopus laevis*, mRNA sequences Xenbase version), pancrustacea (*Drosophila melanogaster*), crustacea (*Daphnia pulex*), decapoda (*Litopenaeus vannamei*) and astacoidea (*Pontastacus leptodactylus/ Astacus leptodactylus*). The *P. vriginalis* transcripts were then aligned against the generated databases using BLASTx (e-value 10^{-10}). Sequences with significant BLAST hits in all databases were classified as bilaterian, sequences with hits only in *D. melanogaster*, *D. pulex*, *L. vannamei* and *P. leptodactylus* as pancrustacean, and so forth. The remaining sequences without significant sequence similarity to one of the species were classified as unique.

As the closest relative with a publicly available genome sequence *Daphnia pulex* was used to identify the transcript sequences of Dnmt1, Dnmt3 and Tet. The protein sequences of *Daphnia pulex* (Dnmt1, Dnmt3 and Tet) were aligned to the transcriptome database of the *P. virginalis* assembly by tBLASTx (e-value 10^{-5}). Candidate sequences were then validated by searching with BLASTx against the non-redundant protein sequence database of NCBI. Additionally completeness of the enzymes was assessed by annotation of the conserved domains with NCBI's CD-search using the translated protein sequences produced with the ExPASy translate tool. The sequences of the *P. virginalis* Dnmt1, Dnmt3 and Tet have been deposited in GenBank (accession numbers KM453737, KM453738 and KM453739).

CpG depletion of Transcrptsequences

Protein-coding sequences (cds) of the assembled *P. virginalis* transcripts were predicted by the automated annotation pipeline from Transcriptome Computational Workbench (applied by Julian Gutekunst). The predicted coding sequences were used for analyses of the evolutionary CpG depletion in *P. virginalis*. The normalized CpG content [amount of observed CpGs to amount of expected CpGs (o/e)] was determined as the amount of CpGs in the coding sequence multiplied by the sequence length divided by the CpG probability of the protein sequence (formula **2.2**). As control the GpCo/e value of each protein-coding sequence was calculated to exclude possible sequence biases influencing the CpGo/e value. The GpCo/e distributions are depicted in the appendix. The distribution of CpGo/e values were plotted with superposition of two Gaussian distributions fitted to the data using normalmixEM of the R package mixtools. For comparison the CpGo/e values of protein-coding sequences of other species were analyzed in the same way: *Drosphila melanogaster* (genome version 6), *Apis mellifera* (genome version 4), *Daphnia pulex* (genome version 1), *Crassostrea gigas* (genome version 9) and *Homo sapiens* (genome version hg38) downloaded from Ensembl.

Formula 2.2: $$\mathrm{CpG}_{\mathrm{o/e}}(\mathrm{sequence}_{\mathrm{s}}) = \frac{\mathrm{n_S} \cdot \sum_{\mathrm{i=1}}^{\mathrm{n_S}} \mathrm{CpGs}}{\sum_{\mathrm{i=1}}^{\mathrm{n_S}} \mathrm{Gs} \cdot \sum_{\mathrm{i=1}}^{\mathrm{n_S}} \mathrm{Cs}}$$

s = current protein-coding sequence; n_S = length of current sequence

2.8.3 Gene Classification

In general, for classification of *P. virginalis* genes, genome assembly version 0.32 (minimum scaffold length ≥ 10 kb) was used (provided by Julian Gutekunst). Based on the provided General Feature Format (GFF) file containing predicted genes, the corresponding coding sequences were extracted from the genome assembly and translated into protein sequences.

For phylostratigraphic analyses, the protein sequences were divided into 9 phylostrata ranging from (1 to 9) cellular organism, Eukaryota, Opisthokonts, Metazoa, Eumetazoa, Bilateria, Protostomia, Arthropoda and the remaining set of genes. Shortly, the protein sequences were mapped to each node represented by several fully sequenced genomes using BLASTp (e-value 10^{-10}). Sequences with significant hits were categorized according to the oldest phylogenic node annotation of the hit gene. A complete list of organisms, which were used for the phylostratigraphic analyses, are listed in the appendix.

The *P. virginalis* genes were classified as housekeeping genes (HKGs) by mapping them to protein sequences of a set of human housekeeping genes (Eisenberg & Levanon, 2013) using BLASTp (e-value 10^{-10}).

2.8.4 DNA Methylation Analyses

Genomic DNA was isolated as described in section 2.6.1, bisulfite treated and sequenced on an Illumina HiSeq platform (table **2.2** and section 2.6.4). Read pairs were quality trimmed (minimum quality value ≥ 15 and minimum length ≥ 36 bp) and mapped to the *P. virginalis* genome assembly version 0.32 (minimum scaffold length ≥ 10 kb; provided by Julian Gutekunst) using BSMAP. Correctly mapped read pairs (appropriate orientation and distance to each other) with both reads mapping uniquely to the same scaffold were used for methylation calling. The methylation ratio (methylation calling) for each CpG was determined by the Python script distributed with the BSMAP package. The provided Python script was slightly changed to analyze only reads fulfilling the following additional criteria: i) minimum quality value of the base at C position ≥ 30 and ii) minimum quality value of the two bases before and after the C position ≥ 20. Only C-positions with a minimum coverage of three reads were used in further analyses.

In general, the mapping efficiency was defined as the portion of mapped reads from all reads used for the mapping (formula **2.3**). The strand-specific CpG-base coverage was determined by the sum of mapped reads over all CpG-positions divided by the amount of covered CpG-positions (coverage ≠ 0) in the genome (formula **2.4**). For the calculation of the genome coverage the positions with undetermined base (N) were removed. The genome coverage was defined as portion of covered positions (minimum coverage > 0) from all positions (formula **2.5**).

The conversion rate was determined by calculating the methylation level of the mitochondrial DNA as portion of deamination artifacts (formula **2.6**). For the methylation ratio the amount of methylated observations (reads with a C in their sequence) of a position was divided by the total amount of observation (reads with a C or T in their sequence) at this positions (formula **2.7**). The strand specific density of methylated CpGs across a scaffold was calculated by dividing the number of methylated CpGs (methylation ratio ≥ 0.8 and coverage ≥ 3) by the length of the used 1 kb non-overlapping sliding window. In a similar way, the average methylation of genomic features as predicted by the maker pipeline (performed by Julian Gutekunst) was calculated as the total amount of methylated CpGs (minimum methylation ratio ≥ 0.8 and minimum coverage ≥ 3) divided by the total amount of CpGs (minimum coverage ≥ 3).

The distribution of the average methylation ratio 4 kb upstream to 4 kb downstream of the predicted genes was calculated as the sum of methylation ratios at this position divided by the total amount of observed methylation ratios at this position (minimum coverage ≥3; formula **2.8**). As the gene length differs the position within the gene was determined by normalization to the gene length.

The methylation level of each gene body was calculated as the sum of methylation ratios within this gene divided by the total amount of observed methylation ratios within the gene (minimum coverage ≥ 3; formula **2.9**).

Analysis of repetitive elements was performed similar to the analysis of genes (formula **2.8** and **2.9**). Though, the length of the upstream and downstream region was only 3 kb instead of 4 kb.

For differential gene body methylation analyses the genes used for the calculations had to fulfill the following criteria: i) minimum coverage ≥ 3 per CpG-position in both samples and ii) minimum amount of covered positions ≥ 5 shared by both samples. Methylation level of each filtered gene was then calculated as described above and the methylation difference was determined by subtraction of the calculated methylation levels.

Formula 2.3: $$\text{mapping efficiency} = \frac{\sum \text{mapped reads}}{\sum \text{sequenced reads}} \cdot 100\ [\%]$$

Formula 2.4: $$\text{CpG base coverage} = \frac{\sum_{i=1}^{t} \text{observations(i)}}{\sum_{i=1}^{t} \partial(\text{observations(i)})}\ [\text{x fold}]$$

$$\partial(\text{observation(i)}) = \begin{cases} 1 & \text{observations} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Formula 2.5: $$\text{genome coverage} = \frac{\sum_{i=1}^{n} \partial(\text{observations(i)})}{n} \cdot 100\ [\%]$$

Formula 2.6: $$\text{conversion rate}_{\text{mitochondrion}} = \left(1 - \frac{\sum_{i=1}^{s} \text{methylated observations(i)}}{\sum_{i=1}^{s} \text{observations(i)}}\right) \cdot 100\ [\%]$$

Formula 2.7: $$\text{methylation ratio(i)} = \frac{\sum \text{methylated observations}}{\sum \text{observations}}$$

Formula 2.8: $$\text{average gene body methylation ratio (i)} = \frac{\sum_{j=1}^{m} \text{methylation ratio(j)}}{m - y}$$

Formula 2.9: $$\text{average gene body methylation level(j)} = \frac{\sum_{i=1}^{n} \text{methylation ratio(i)}}{n - x}$$

n = all sequence positions (without N-bases); i = current position; m = all gene bodies; j = current gene body; x = positions with coverage < 3; y = gene body with position coverage < 3; t = all CpG-positions; methylated observations = reads with a C in their sequence at the analyzed position; observations = reads at the analyzed position; s = all C-positions

2.8.5 Expression Analyses

Total RNA was isolated as described in section 2.6.2 and sequenced on an Illumina HiSeq platform (table **2.2**). Read pairs were quality trimmed (minimum quality value ≥ 15 and minimum length ≥ 36 bp) and mapped to the *P. virginalis* genome assembly version 0.32 (minimum scaffold length ≥ 10 kb; provided by Julian Gutekunst) using RSEM and bowtie2 as mapper. The calculated transcripts per million (TPM) value of each predicted transcript was used for expression analyses as it is more comparable across samples (B. Li, Ruotti, Stewart, Thomson, & Dewey, 2009). The $\log_{10}$(TPM) of each transcript was determined and divided into 8 equal bins ranging from lowly expressed (rank 1) to highly expressed genes (rank 5-8). Transcripts with a TPM value of zero were classified as unexpressed genes (rank 0).

3 Results

3.1 Initial Analyses

Initial analyses were performed, before studying the methylome, at single-base resolution to clarify the taxonomic status of the marbled crayfish and to provide a basis for molecular biological and bioinformatic analysis. Thus, the mitochondrial genome, the genome size and key features of the transpcriptome were analyzed in detail.

3.1.1 Sequencing and analysis of mitochondrial genomes

The taxonomic status of the marbled crayfish is discussed controversially (Martin et al., 2010). Martin et al. (2010) suggested a close relationship of marbled crayfish to *Procambarus fallax* by sequence comparison of two mitochondrial genes. To further elucidate its taxonomic status, the mitochondrial DNA of the marbled crayfish was analyzed in detail by sequence comparison to its suggested closest relatives *Procambarus fallax* and *Procambarus alleni*, and including marbled crayfish from 4 different strains. The mitochondrial genome sequences of marbled crayfish, *P. fallax* and *P. alleni* were assembled and annotated. All mitochondrial features were completely assembled, only the AT-rich sequence of the control region (D-loop) was partially assembled (Fig. **3.1A**). Sequence comparison between marbled crayfish and *P. fallax* revealed 144 single nucleotide polymorphisms (SNPs) and between marbled crayfish and *P. alleni* 1,165 SNPs (Fig. **3.1B**) suggesting a closer genetic relationship between marbled crayfish and *P. fallax*. These findings are consistent with observations of Martin et al. (2010) comparing the marbled crayfish 12S rRNA and cythocrome oxidase subunit I (COI; positions are depicted by purple bars in Fig. **3.1B**) to several *P. fallax* and *P. alleni* individuals. Martin et al. (2010) compared only two marbled crayfish individuals, one from their laboratory population in Berlin and one specimen found in Saxony. In this study four individuals were analyzed, two from distinct laboratory populations (Heidelberg and Petshop) and two from different, stable wild populations (Moosweiher and Madagascar) (Chucholl & Pfeiffer, 2010; J. P. G. Jones et al., 2009). Notably, analysis of the four marbled crayfish individuals revealed identical mitochondrial sequences (Fig. **3.1B**) indicating a single origin of the analyzed marbled crayfish populations. This provides a strong argument for the consideration of marbled crayfish as an independent species (Procambarus virginalis, see discussion for details).

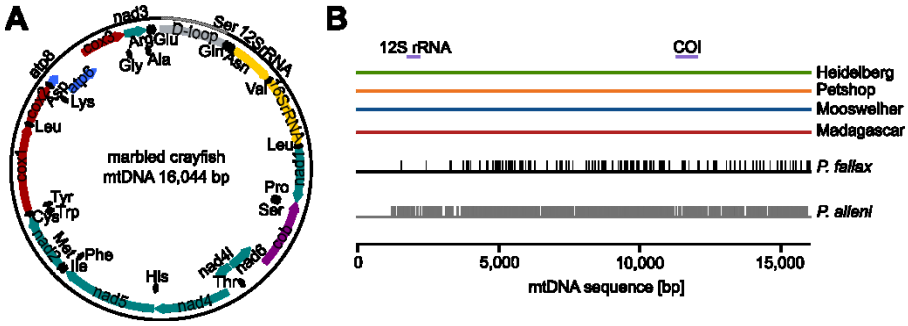


Figure 3.1 Annotation and comparison of the marbled crayfish mitochondrial DNA.

(**A**) Location of the genes annotated in the assembled mitochondrial DNA of marbled crayfish (created with SnapGene): tRNAs (black), rRNAs (yellow), cytochrome b (purple) and subunits of NADH dehydrogenase (nad: green), ATP synthase (atp: blue) and cytochrome c oxidase (cox: red). D-loop the control region is depicted in grey. (**B**) Comparison of marbled crayfish, *P. fallax* and *P. alleni* mitochondrial genomes: sequences of four marbled crayfish individuals, two from laboratory populations (Heidelberg and Petshop) and two from wild populations (Moosweiher and Madagascar), and sequences of *P. fallax* and *P. alleni*, repsectively. SNPs are indicated by vertical bars. The 12S rRNA and cythocrome oxidase subunit I (COI) genes were used for an earlier phylogenetic analysis (Martin et al., 2010) and are indicated by a purple bar.

3.1.2 Nuclear DNA Content of *P. virginalis* Haemocytes

Since a reference genome is required for methylation analysis at single-base resolution, the genome size of *P. virginalis* was estimated to determine the sequencing requirements for the genome assembly. The nuclear DNA content was analyzed by comparative flow cytometry of propidium iodide stained *P. virginalis* haemocytes. The measured fluorescence signal of the stained marbled crayfish cells was more intense than the measured fluorescence signal of the used standards (Fig. **3.2A**) indicating a genome size larger than the mouse and human genome. Considering the fact that the marbled crayfish genome is triplod (Martin, Thonagel, & Scholtz, 2016), the genome size of *P. virginalis* was estimated at 3.7 Gb (Fig. **3.2B**). Since the assembly of large genomes of polyploid organisms is particularly challenging (Claros et al., 2012; Iwasaki et al., 2016), the *P. virginalis* genome assembly was not pursued in this PhD project.

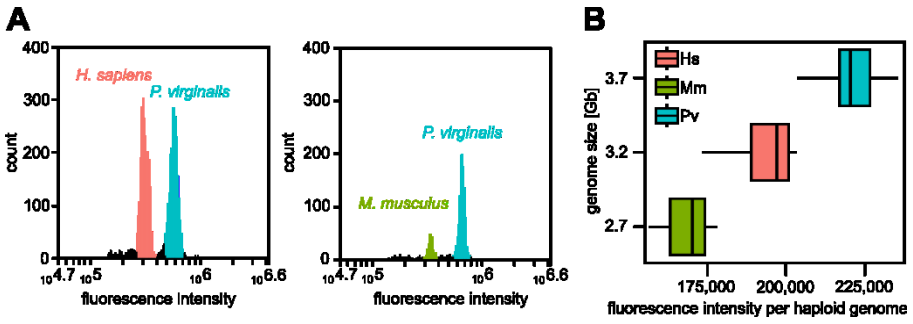


Figure 3.2 Size estimation of the *P. virginalis* genome.

(**A**) Flow cytometry of propidium iodide stained haemocytes of marbled crayfish (blue peak) mixed with stained peripheral blood mononuclear cells of *H. sapiens* (left, pink peak) and *M. musculus* (right, green peak) as standards. (**B**) The genome size of *P. virginalis* was determined by comparing the fluorescence intensity per haploid genome of *P. virginalis* (Pv: blue) to the standards *H. sapiens* (Hs: pink) and *M. musculus* (Mm: green) with known genome sizes. The plot shows the measurement of two biological and three technical replicates.

3.1.3 The Transcriptome of *P. virginalis*

Transcriptome assembly was performed using a normalized sequencing library prepared from four different tissues (for details see material and methods section 8.2). The sequencing resulted in 48.4 Gb of sequence information which was assembled into a final transcriptome consisting of 22,338 transcripts with an average sequence length of 1,525 bp. The quality of the transcriptome was assessed using computational benchmarking and mass spectrometry.

Quality assessment of the P. virginalis transcriptome assembly

A set of conserved genes from arthropod genomes (Simão et al., 2015) was used for the Benchmarking with Universal-Single Copy Orthologs tool (BUSCO; Waterhouse et al., 2013). The analysis showed that 65 % of the 2,675 orthologous genes were found as complete proteins in the assembly (Fig. **3.3A**). Notably, the percentage of orthologous genes as complete proteins was increased to 75% and higher, when transcriptome assemblies were used in an improved mode (assembly version ≥ 2; Fig. **3.3A**). Among the first transcriptome assemblies only the transcriptomes of *D. pulex* and *B. mori* contained a higher fraction of orthologous genes with complete protein sequence than the *P. virginalis* transcriptome (Fig. **3.3A**). These results confirmed that the quality of the assembled *P. virginalis* transcriptome is comparable to other recently published arthropod transcriptomes.

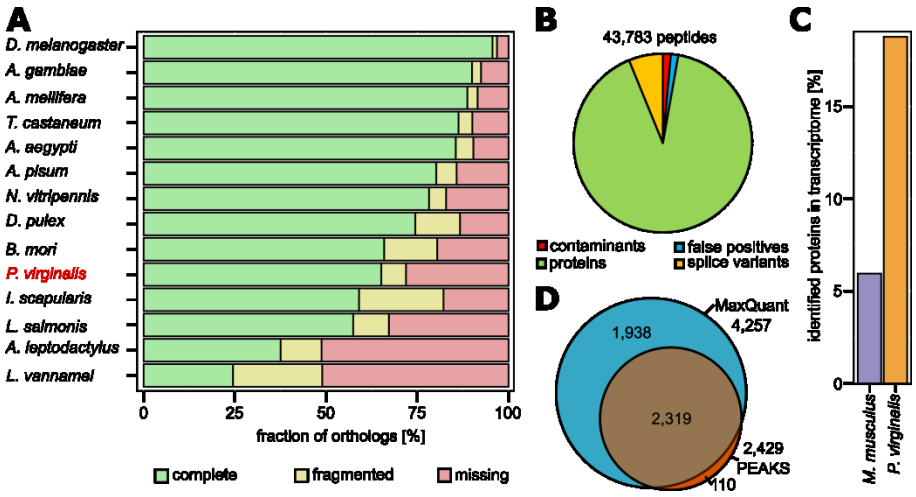


Figure 3.3 Quality control of the assembled *P. virginalis* transcriptome.

Quality assesment by (**A**) determining the completeness of 2675 orthologous genes in comparison to 13 different species using Benchmarking Universal Single-Copy Orthologs (BUSCO) and by (**B**, **C** and **D**) mass-spectrometric analyses. (**A**) BUSCO Analysis. Bars represent the percentage of complete (green), fragmented (yellow) and missing (red) orthologs. Transcript sequences were downloaded from EnsemblMetazoa for *D. melanogaster* (version 6), *A. gambiae* (version 4), *A. mellifera* (version 4), *T. castaneum* (version 3), *A. aegypti* (version 3), *A. pisum* (version 2), *N. vitripennis* (version 2), *D. pulex* (version 1), *B. mori* (version 1), *I. scapularis* (version 1) and *L. salmonis* (version 1). The first assembly version of *L. vannamei* and *A. leptodactylus* transcriptomes can be accessed via TSA accession numbers JP355723 - JP376614, JP382831 - JP435443 and GAFY00000000.1, respectively. (**B** and **C**) Mass-spectrometry (performed by Oliver Popp) using the MaxQuant software for peptide calling. (**B**) portion of contaminants (red), false positives (blue), paralogues/ splice variants (yellow) and proteins (green) of the 43,783 detected peptides in *P. virginalis* protein extracs. (**C**) fraction of proteins in the transcriptome validated by mass spectrometry. (**D**) Intersection of transcripts validated by mass-spectrometry using the PEAKS software and the MaxQuant software.

To further emphasize the quality of the transcriptome assembly with a different approach, mass-spectrometric analysis of protein extracts from marbled crayfish hepatopancreas was performed. Based on the detected MS-spectra the corresponding peptide sequences were predicted using bioinformatic software. Two different softwares were used for the peptide calling. While MaxQuant calls the peptides based on a given database, PEAKS performs *de novo* calling of peptides independently from the provided protein sequences. Using the MaxQuant application, 42,566 out of 43,783 (97.2 %) peptides matched to the *P. virginalis* transcriptome meaning that the analyzed sample had a minor fraction of contaminants and false positives (Fig. **3.3B**). These 42,566 peptides validated 4,185 of the 22,288 (18.8 %) predicted protein sequences in the *P. virginalis* transcriptome (Fig. **3.3C**). As orientation a mouse data set was provided by our cooperation partner who performed the mass-spectrometric analysis. The mouse data set was generated with the same procedure as the *P. virginalis* data set. Mouse peptides confirmed a smaller fraction of the mouse transcriptome (6 %) compared to the

P. virginalis analysis (Fig. **3.3C**) implying that an acceptable fraction of the assembled *P. virginalis* transcripts could be confirmed. The PEAKS software identified 1,713,864 peptides and 141,771 peptides could be mapped to the *P. virginalis* transcriptome. The mapped peptides validated 2,429 (10.89 %) of the predicted transcripts. To note, 95 % of the transcripts validated by PEAKS were also validated by the MaxQuant application (Fig. **3.3D**) and thus, the majority of confirmed *P. virginalis* transcripts was validated by both software applications. Taken together, an acceptable quality of the assembled transcriptome was confirmed by two different approaches.

Annotation of the P. virginalis transcripts

After assessing the quality of the *P. virginalis* transcriptome, the transcripts were annotated using four different databases (Fig. **3.4A** and **3.4B**). The database Cluster of Orthologous Groups of proteins (COG) contains protein sequences classified into groups of similar functions based on consistent patterns of sequence similarities and thus allows to functionally annotate newly sequenced genomes (Tatusov, Koonin, & Lipman, 2012). Additionally to the COG database, the Kyoto Encyclopedia of Genes and Genomes (KEGG) provide information about the corresponding interaction, reaction and relation networks of the functional annotated sequence (Kanehisa, 1996). In contrast, the Universal Protein Resource (UniProt) is the largest collection of protein sequences and their annotation (Bateman et al., 2015) and links the sequences to database records of Gene Ontology (GO) which provides functional annotation and information of parent and child processes (Blake et al., 2015). Since UniProt also contains unreviewed, annotated records in comparison to the other databases, the majority of transcripts was annotated with UniProt terms (Fig. **3.4A**). However, combining the results of all databases together 9,483 of the 22,338 (42.5 %) sequences remained unannotated (Fig. **3.4B**).

Additionally, the transcript sequences of *P. virginalis* were analyzed for sequence similarity to transcriptomes of other species, including *Xenopus laevis* (bilaterian core), *Drosophila melanogaster* (pancrustacean), *Daphnia pulex* (crustacean), *Litopenaeus vannamei (decapodan),* and *Pontastacus leptodactylus* (astacoidea), depicted in Fig. **3.4C**. The analysis revealed sequence similarities to the majority of transcripts with the largest fraction belonging to the bilaterian core (41.1 %; Fig. **3.4D**). Notably, 4,306 (19.3 %) were not homologous and thus classified as unique. This fraction is comparable to the fraction of unique genes reported for other genomes (Colbourne et al., 2011). Thus, the vast majority of assembled transcripts was found in transcriptomes of other species and a large fraction could be annotated.

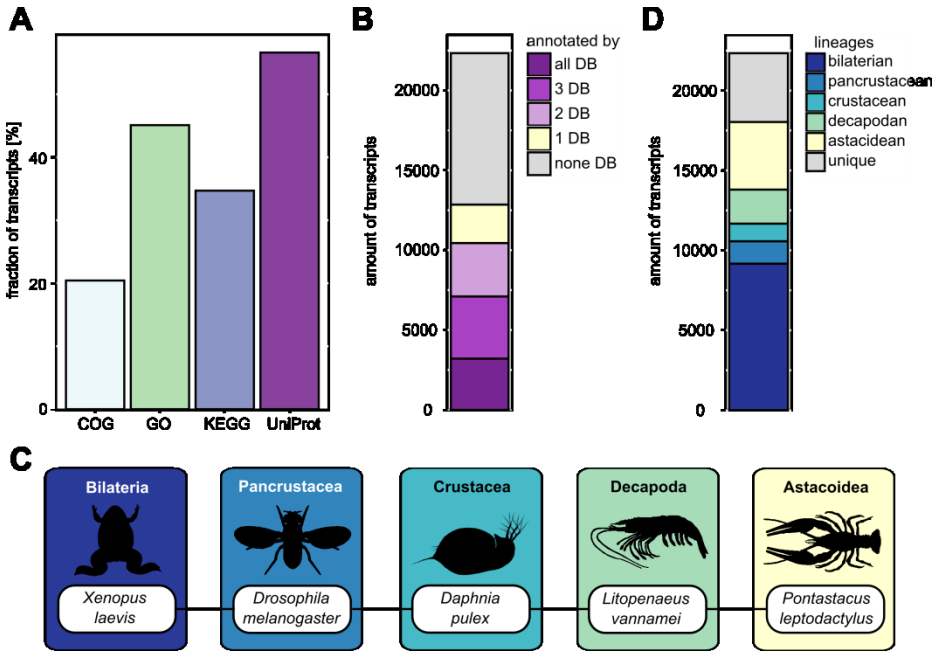


Figure 3.4 Annotation of the *P. virginalis* transcriptome.

(**A**) Fraction of transcripts annotated with database terms using four different databases: Cluster of Orthologous Groups of proteins (COG), Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Universal Protein Resource (UniProt). (**B**) Classification of the *P. virginalis* transcript sequences into groups annotated by the amount of databases (DB). (**C** and **D**) Comparison of *P. virginalis* transcripts with the transcriptomes of *Xenopus laevis*, *Drosophila melanogaster*, *Daphnia pulex*, *Litopenaeus vannamei* and *Pontastacus leptodactylus* representing the core bilaterian transcripts (dark blue), pancrustacean (blue), crustacean (light blue), decapodan (aqua marine) and astacidean transcripts (light yellow). (**D**) Remaining transcripts with no sequence similarity are coloured in grey.

3.1.4 Evidences of DNA Methylation in *P. virginalis*

Before studying the *P. virginalis* methylome at single-base resolution, solid evidences for the presence of DNA methylation in *P. virginalis* were collected by analyzing the historical DNA methylation and the DNA methylation machinery of *P. virginalis*.

Historical germline DNA methylation in P. virginalis

Methylated cytosines can spontaneously deaminate to thymines with a high frequency (Shen, Rideout, & Jones, 1994). When the hydrolytic deamination occurs in the germline, this C-to-T depletion is accumulated over time and leaves an evolutionary signature in the genome (Glastad et al., 2011). Thus, the fraction of the C-to-T depletion inherited to the next generations reflects the fraction of cytosines which were historically methylated in the germline. As DNA

methylation in animals is almost entirely targeted to CpG dinucleotides, the ratio of reduced CpG dinucleotides in a sequence (calculated as CpGo/e value) can be used to estimate levels of DNA methylation in comparison to other genomes (Yi & Goodisman, 2009). The distributions of CpG depletion in protein-coding sequences (cds) of *P. virginalis* and other species with known methylation levels were calculated to evaluate the presence of historical germline DNA methylation in the assembled *P. virginalis* transcriptome (Fig. **3.5**).

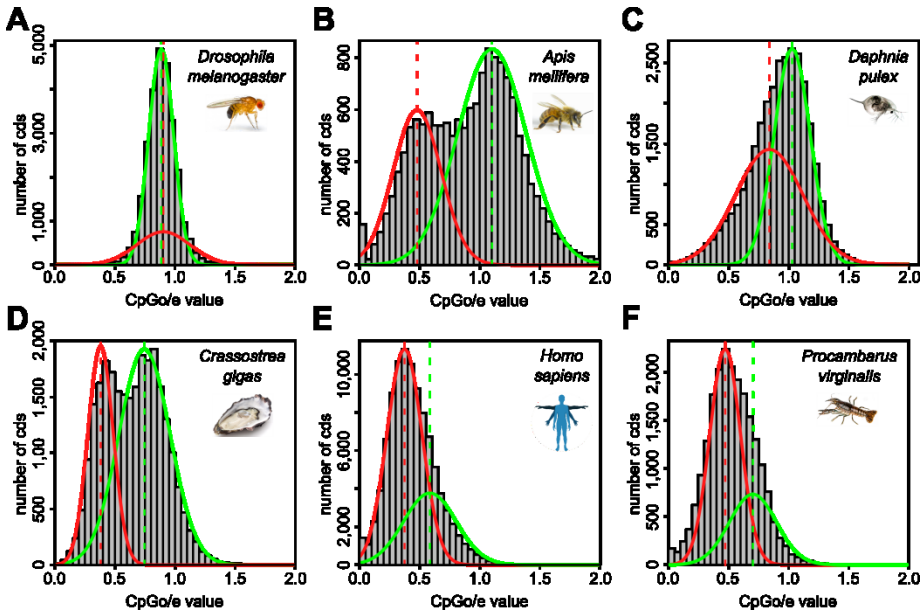


Figure 3.5 Evolutionary CpG depletion in protein-coding sequences (cds) of various species.
Distribution of normalized CpG content [amount of observed CpGs to amount of expected CpGs (o/e)] with superposition of two Gaussian distributions fitted to the data using normalmixEM of the R package mixtools. Dashed lines indicate means of the fitted curves. Plots A to E are ordered from the lowest to the highest genome-wide methylation level: (**A**) *Drosophila melanogaster* (lacking DNA methylation) (Raddatz et al., 2013), (**B**) *Apis mellifera* (0.11 %) (Lyko et al., 2010), (**C**) *Daphnia pulex* (0.25 %) (Asselman et al., 2016), (**D**) *Crassostrea gigas* (1.96 %) (Xiaotong Wang et al., 2014) and (**E**) *Homo sapiens* (3.93 %)(Lister et al., 2009). **(F)** Distribution of CpGo/e values in *P. virginalis* protein-coding sequences.

Since *Drosophila melanogaster* lacks DNA methylation (Raddatz et al., 2013), its protein-coding sequences showed almost no CpG depletion and thus the unimodal distribution centered around a CpGo/e value of 1.0 (Fig. **3.5A**). In comparison to *D. melanogaster*, the amount of observed CpGs to the amount of expected CpGs (CpGo/e) were decreased in the protein-coding sequences of *Apis mellifera*, *Daphnia pulex*, *Crassostrea gigas*, *Homo sapiens* and *P. virginalis* (Fig. **3.5B** to **3.5F**). The CpGo/e distributions of *C. gigas* and *H. sapiens* both with a genome-wide methylation level above 1 % (Wang et al., 2014; Lister et al., 2009) were more shifted towards low CpGo/e values (Fig. **3.5D** and **3.5E**) compared to the distributions of *A. mellifera*

and *D. pulex* (Fig. **3.5B** and **3.5C**) both with a genome-wide methylation level below 1 % (Asselman et al., 2016; Lyko et al., 2010). Particularly, the protein-coding sequences of *P. virginalis* showed a CpG depletion similar to *H. sapiens* (Fig. **3.5F**) indicating the presence of historical germline DNA methylation in *P. virginalis*.

Identification of a conserved and active DNA methylation system in P. virginalis

To identify the DNA methylation system, the assembled transcriptome of *P. virginalis* was aligned against the protein sequences of the water flea *Daphnia pulex*, which was the only known crustacean with an annotated transcriptome. This approach identified a complete DNA methylation system in *P. virginalis* consisting of single homologues for Dnmt1 and Dnmt3 DNA methyltransferase and the Tet DNA dioxygenase, respectively (Fig. **3.6A**). The comparison of virtually translated protein sequences to established honeybee and human homologues revealed proteins containing all the known protein domains in the correct order (Fig. **3.6A**). Interestingly, a long C-terminal sequence of the *P. virginalis* Dnmt3 distinguishes the Dnmt3 homologue from the established protein sequences (Fig. **3.6A**). It is possible that the C-terminus is an assembly artifact, but the sequence was assembled by two independent assembly approaches and different data sets. These results suggest that the identified proteins are a maintenance DNA methyltransferase (Dnmt1), de novo DNA methyltransferase (Dnmt3) and DNA hydroxymethylase (Tet).

To confirm the expression of the marbled crayfish DNA methylation system, mRNA of adult animals was isolated from various tissues (heart, hepatopancreas, abdominal muscle and claw muscle) and analyzed by qRT-PCR. Consistent with a function as maintenance methyltransferase, Dnmt1 was moderately expressed in all tissues (Fig. **3.6B**). In comparison to Dnmt1, the expression of Dnmt3 appeared more tissue-specific, whereas mRNA levels of Tet were the highest among all tissues (Fig. **3.6B**). Nonetheless, all three enzymes were expressed in the analyzed tissues.

In a previous study the presence of DNA methylation in the marbled crayfish genome (Günter Vogt et al., 2008) was analyzed by capillary electrophoresis with laser-induced detection of fluorescently labeled nucleotides. Consequently, mass spectrometry was performed to determine the DNA methylation level and DNA hydroxymethylation level quantitatively in three tissues (ovary, hepatopancreas and abdominale muscle) form an adult animal. The analysis of 5-methylcytosine revealed highly consistent methylation levels of 2.4 - 2.52 % (Fig. **3.6C**), which are comparable to the levels observed in mammalian tissues (Gama-Sosa et al., 1983). In contrast, the low but significant 5-hydroxymethylcytosine levels in *P. virginalis* adult tissues (5.4 -

9.3 ppm) were substantially lower than the levels described in the majority of the mammalian tissues (0.1%) (Globisch et al., 2010) and more than two orders of magnitude below the highest level detected in brain, here as control mouse brain (Fig. **3.6D**) (Kriaucionis & Heintz, 2009). In summary, the results demonstrate the presence of a conserved and active DNA methylation system in *P. virginalis*.

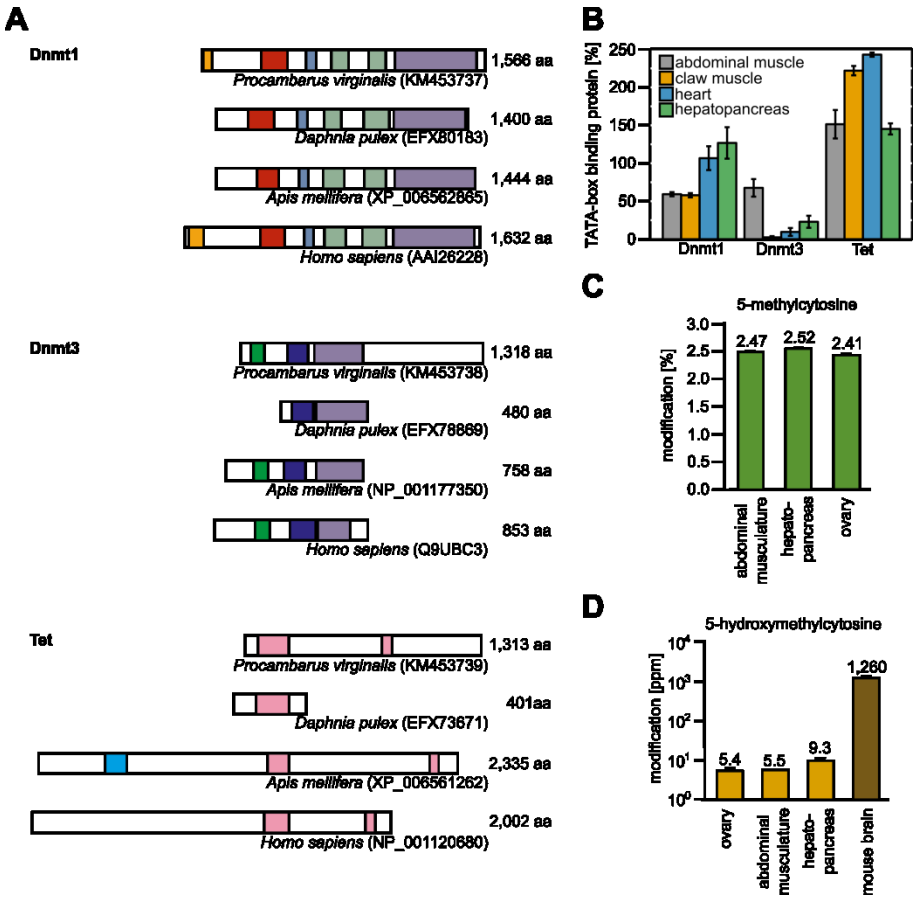


Figure 3.6 DNA methylation system in *P. virginalis*.

(**A**) Virtually translated protein sequences of Dnmt1, Dnmt3 and Tet are shown in comparison with three reference organisms: *Daphnia pulex*, *Apis mellifera* and *Homo sapiens*. Accession numbers are indicated in brackets and conserved domains are shown as coloured boxes. Dnmt1: DMAP1 binding domain (orange), replication foci domain (red), CXXC zinc finger domain (blue), bromo adjacent homology domain (green) and catalytic domain (purple). Dnmt3: PWWP domain (green), zinc finger domain (blue) and catalytic domain (purple). Tet: Rrn6 domain (blue) and catalytic domain (pink). (**B**) Expression of Dnmt1, Dnmt3 and Tet in various adult tissues, relative to the TBP (TATA-box binding protein) housekeeping gene. Represented are averaged values from measurement of three technical and two biological replicates. (**C** and **D**) Quantitative analysis of genomic 5-methylcytosine (**C**) and 5-hydroxymethylcytosine (**D**) levels of various tissues form an adult marbled crayfish by mass spectrometry (performed by Katharina Schmid). (**D**) Adult mouse brain DNA was included as reference for detection of 5-hydroxymethylcytosine.

3.2 The Methylome of *P. virginalis*

The ratio of CpG depletion in coding-sequences (Fig. **3.5F**) and mass-spectrometric analyses of the DNA (Fig. **3.6C**) confirmed the presence of DNA methylation in *P. virginalis*. To analyze the methylome of *P. virginalis* at single-base resolution, whole-genome bisulfite sequencing (WGBS) was performed. Specific examples of the subsequently described methylation patterns are shown in the Appendix.

3.2.1 DNA Methylation Characteristics

The sequencing of *P. virginalis* hepatopancreas sample HD2 resulted in 33 Gb sequence information. Roughly 76 % of the processed reads could be mapped to the draft *P. virginalis* genome (provided by Julian Gutekunst) covering 82 % of the genome and 64 % of all CpG dinucleotides with an average strand-specific base coverage of 8.4 x (per covered CpG). Since mitochondrial DNA is unmethylated (Hong et al., 2013; Liu et al., 2016), the assembled mtDNA sequence of *P. virginalis* (section 3.1.1) was used to determine the bisulfite conversion efficiency. This approach confirmed a high bisulfite conversion rate of 99.77 %, thus confirming the high quality of the dataset.

General characteristics of the P. virginalis methylome

As already indicated by the analysis of the CpG depletion in the *P. virginalis* transcriptome (Fig. **3.5F**), whole-genome bisulfite sequencing confirmed that the methylation in *P. virginalis* is targeted to CpG dinucleotides (Fig. **3.7A**). Moreover, the methylation level of CpG dinucleotides displayed a bimodal distribution (Fig. **3.7B**) as observed for other organisms with DNA methylation e.g. *Apis mellifera*, *Crassostrea gigas* or *H. sapiens* (Raddatz et al., 2013; Xiaotong Wang et al., 2014). In addition, the methylation observed in *P. virginalis* was symmetric (Fig. **3.7C**), consistent with the symmetry of CpG dinucleotides. These results show that the *P. virginalis* methylome shares the basic features of Dnmt1-Dnmt3-dependent animal methylomes (Zemach et al., 2010).

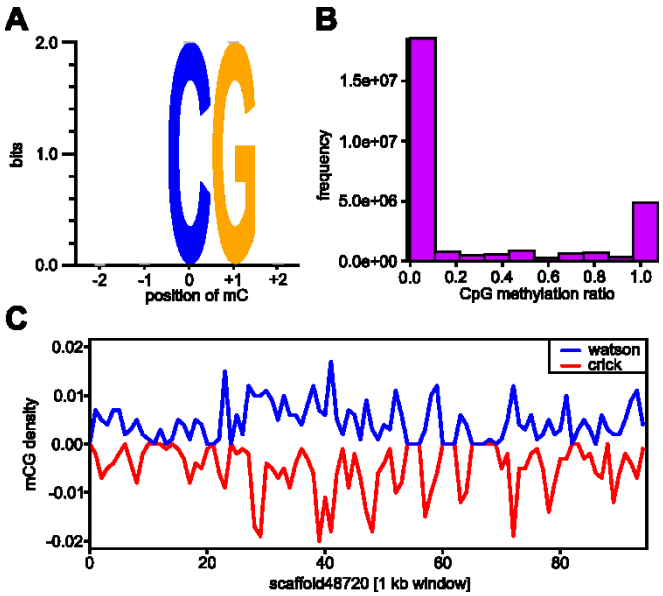


Figure 3.7 DNA methylation characteristics in *P. virginalis*.

(**A**) Nucleotide proportion of the two nucleotides downstream and upstream of methylated cytosines. (**B**) Distribution of the average methylation level for each CpG (methylation ratio). (**C**) Strand specific density of methylated CpGs across the scaffold 48720 (Watson strand: blue, Crick strand: red). The density was calculated by dividing the number of methylated CpGs (methylation ratio ≥ 0.8 and coverage ≥ 3) by the length using a 1 kb non-overlapping sliding window.

Gene bodies are targets of DNA methylation

Mammalian methylomes are characterized by an ubiquitous DNA methylation pattern, whereas some invertebrate methylomes show a mosaic methylation pattern while others are characterized by a sporadic methylation pattern (Breiling & Lyko, 2015; Schübeler, 2015). To characterize the methylation pattern in *P. virginalis* the 20 longest scaffold sequences were analyzed. Interestingly, 25 % of the analyzed scaffolds were ubiquitously methylated (e.g. Fig. **3.8A**), while 5 % were sporadically methylated (e.g. Fig. **3.8B**) and 70 % displayed a mosaic DNA methylation pattern (e.g. Fig. **3.8C**). The two latter patterns were not the result of low coverage. However, the majority of analyzed scaffolds showed a mosaic DNA methylation pattern implying that the DNA methylation is targeted to specific genomic regions.

As methylated gene bodies were observed while analyzing methylation patterns (Fig. **3.8C**), the average methylation of gene regions was calculated by averaging the methylation levels of individual CpGs. The average methylation of coding-exons (CDS), exons, introns and 3'UTRs were approximately twice as high or even higher than the genome background (Fig. **3.8D**). Exons had a lower methylation level (32 %) compared to introns with a maximum of 42 % methylation. This methylation pattern is different from the described

preference of DNA methylation for exons over introns in most organisms (Feng et al., 2010; Lister et al., 2009). However, these results confirm gene body methylation in *P. virginalis* which is considered to be a basal evolutionary feature of eukaryotic methylomes (Feng et al., 2010; Sarda et al., 2012; Zemach et al., 2010).

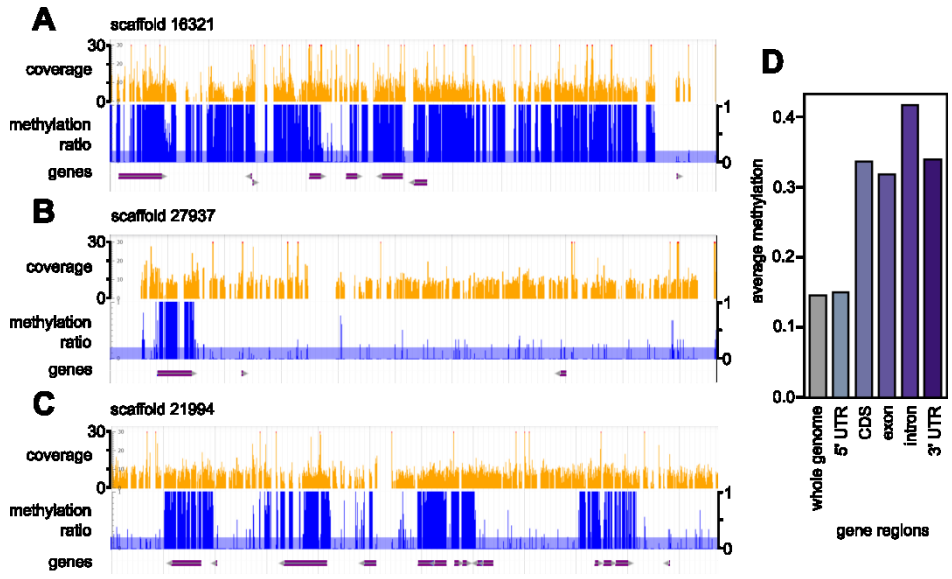


Figure 3.8 Methylation pattern and targets.

Ubiquitous (**A**), sporadic (**B**) and mosaic (**C**) DNA methylation are shown by methylation ratios of each CpG (blue vertical bars) along the scaffolds 16321 (773 kb), 27937 (665 kb) and 21994 (734 kb), respectively. Methylation ratios below 0.2 are marked as bisulfite conversion artefacts (transparent blue horizontal bar). The predicted gene features within the scaffolds are illustrated below each methylation panel (purple). Corresponding coverage (orange vertical bas, pink: coverage > 30) of the scaffolds is depicted above the methylation panel. (**D**) Methylation level of predicted genes divided into untranslated regions (5'UTR and 3'UTR), protein-coding sequences (CDS), exons and introns are shown together with the genome-wide methylation level (grey bar).

3.2.2 Gene Body Methylation

To investigate the methylation pattern of gene bodies in *P. virginalis*, the DNA methylation levels were analyzed across genes. The methylation within gene bodies was increased (53 %) relative to the upstream and downstream regions (Fig. **3.9A**) and dropped sharply to the background level (39 %) around the transcription start site (TSS) and transcription termination site (TTS). This methylation pattern is similar to patterns described for the majority of invertebrates and distinct from patterns in *Apis mellifera* and *Bombyx mori* which methylation levels showed a peak shortly after the TSS and a minor peak before the TTS (Zemach et al., 2010). Furthermore, the methylation patterns in *P. virginalis* are different from the patterns

described for mammals which show only decreased methylation levels at the TSS (Feng et al., 2010; Zemach et al., 2010).

Since some invertebrates showed a bimodal distribution of gene body methylation (C Falckenhayn et al., 2013; Suzuki et al., 2013; Xiaotong Wang et al., 2014), genes were binned based on their methylation level. Roughly 26 % of the genes were entirely unmethylated (< 0.1), whereas 41 % of the genes were highly methylated (> 0.7; Fig. **3.9B** and example Fig. **3.9C**). Thus, gene body methylation in *P. virginalis* was bimodally distributed indicating that DNA methylation is targeted to a subset of genes.

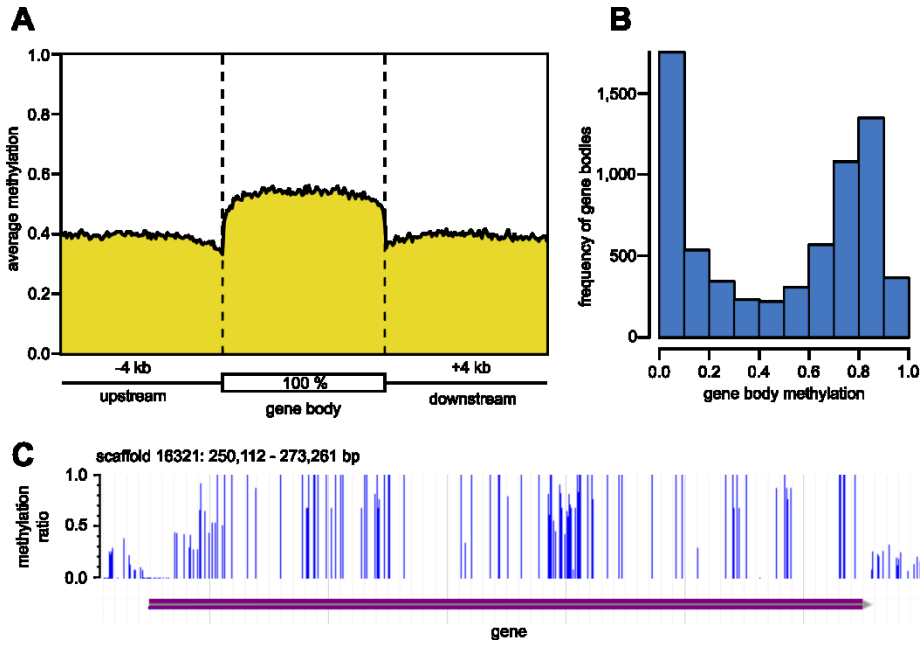


Figure 3.9 Gene body methylation.

(**A**) Distribution of average methylation level along the predicted gene bodies. Starting 4 kb upstream and ending 4 kb downstream of the transcription start site (TSS) and transcription termination site (TTS), respectively (indicated by vertical dashed lines). (**B**) Distribution of gene body methylation levels. (**C**) Example of gene body methylation in scaffold 16321. Methylation ratios of each CpG (blue vertical bars) and the predicted gene (horizontal purple bar below the methylation panel) are illustrated.

Gene body methylation is targeted to a nonrandom subset of genes

Since depletion of CpG dinucleotides in coding sequences is associated with accumulated deamination of methylated cytosines to thymines in the germline (section 3.1.3) (Shen et al., 1994; Yi & Goodisman, 2009), the ratio of CpG depletion (CpGo/e value) of gene bodies was calculated and divided into three groups. On average gene bodies with a low CpGo/e value (< 0.6) displayed a higher methylation level than genes with a high CpGo/e value

(≥ 1.2; Fig. **3.10A**) indicating an inverse correlation between CpGo/e value and methylation. This finding is coherent with previous observations in *Apis mellifera* and *Schistocerca gregaria* (C Falckenhayn et al., 2013; Lyko et al., 2010) and suggests that gene bodies with a low CpGo/e value are preferentially methylated in *P. virginalis*.

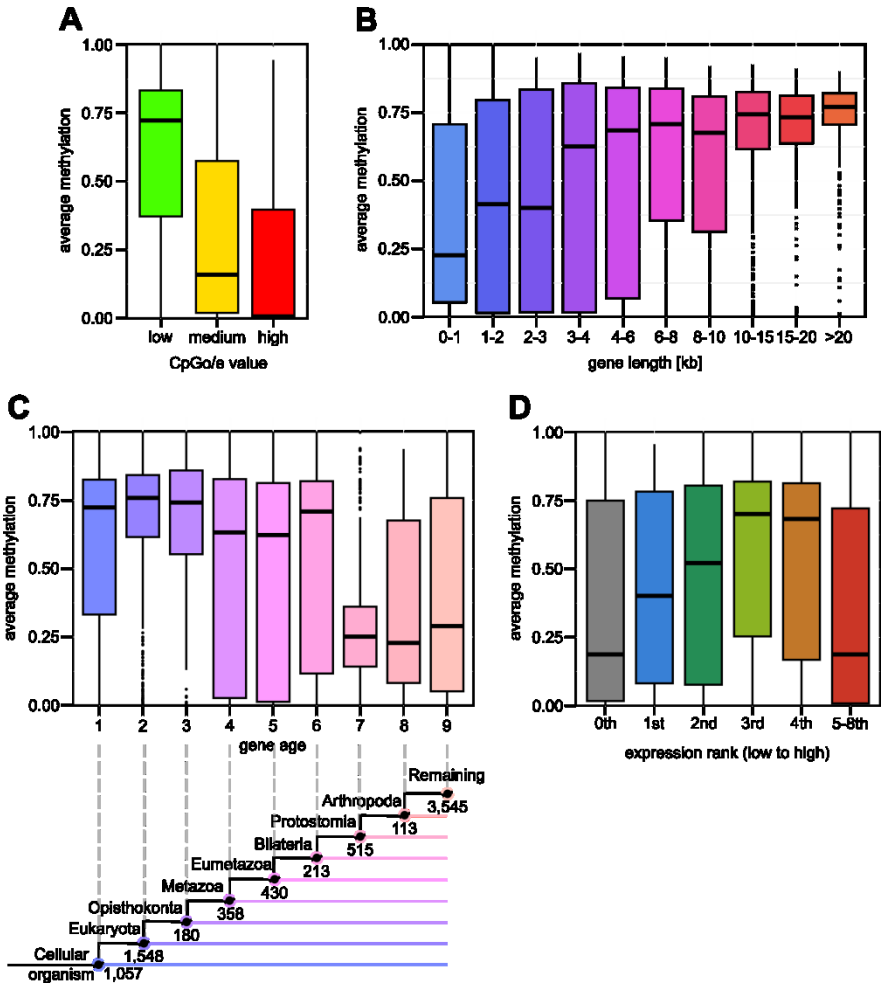


Figure 3.10 Feature of target genes.

Distribution of gene body methylation levels across genes classified in different CpGo/e groups (**A**), length intervals (**B**), age groups (**C**) and expression ranks (**D**). (**A**) Normalized CpG content [amount of observed CpGs to amount of expected CpGs (o/e)] was grouped into low (< 0.6), medium (≥ 0.6, < 1.2) and high (≥ 1.2). (**C**) All predicted *P. virginalis* genes were translated into protein sequences and mapped to different phylogenetic nodes with 1 representing the oldest and 9 the youngest groups. Phylostrata and the correponding number of mapped *P. virginalis* genes are indicated below the panel. (**D**) The 0^{th} rank represents all unexpressed genes (TPM = 0), while all expressed genes (TPM > 0) were distributed into 8 bins from least expressed (1^{st} rank) to most expressed ($5\text{-}8^{th}$) genes.

It has been reported that highly methylated genes in insects are shorter than genes with a low methylation level, whereas in other invertebrates and plants the opposite methylation

pattern was observed (Sarda et al., 2012; Takuno & Gaut, 2012; Xiaotong Wang et al., 2014). To test whether DNA methylation in *P. virginalis* correlates with gene length, the genes were grouped into different length categories. The gene body methylation level increased with longer gene length (Fig. **3.10B**). Genes with a length of more than 10 kb had the highest average methylation level, whereas genes shorter than 1 kb had the lowest level, indicating that long genes are preferentially methylated in *P. virignalis*.

It has been suggested that sequence conservation of highly methylated genes is a common feature in invertebrates (Sarda et al., 2012; Suzuki et al., 2007). Hence, the *P. virginalis* genes were classified into 9 phylostrata representing different evolutionary ages. Genes that originated after Bilateria (phylostratum 7-9), showed a lower methylation level (Fig. **3.10C**) than genes that originated before Metazoa (phylostratum 1-3), indicating that in *P. virginalis* young genes are less likely to be methylated than older genes.

To investigate the relationship of gene body methylation and gene expression in *P. virginalis*, RNA-Seq was performed with the same sample material as used for WGBS. The expression of genes was determined as TPM value (transcripts per kilobase million) and genes were binned into several expression ranks. Highly expressed (rank 5-8) and unexpressed (rank 0) genes displayed the lowest methylation level, whereas genes with a moderate expression were more highly methylated (Fig. **3.10D**) suggesting a parabolic relationship of gene body methylation to gene transcription. This result is consistent with observations in plants and other invertebrates (Zemach et al., 2010; Zilberman et al., 2007). Thus, these results suggest that gene body methylation in *P. virginalis* is targeted to a nonrandom subset of genes sharing several features.

3.2.3 Housekeeping Gene Methylation

As it was indicated that the DNA methylation targets a nonrandom subset of genes (section 3.2.2), additional analyses were performed to identify the targeted gene set.

Housekeeping genes are main targets of gene body methylation

Based on the observations described in section 3.2.2, the following criteria were defined to classify the genes into targeted and non-targeted genes for subsequent characterization. Genes with a low CpGo/e value (< 0.6), long gene sequence (≥ 10 kb), evolutionary conserved protein sequence (age node ≤ 3) and moderate expression rank (3^{rd} - 4^{th}) were categorized into the group of targeted genes, while genes not meeting one of those criteria into the non-targeted genes. Indeed, the average methylation level of genes meeting the defined criteria was notably

increased (73 % and median 76 %) compared to the average methylation level of the non-targeted group (32 % and median 20 %; Fig. **3.11A**) confirming that gene body methylation in *P. virginalis* targets genes with similar features.

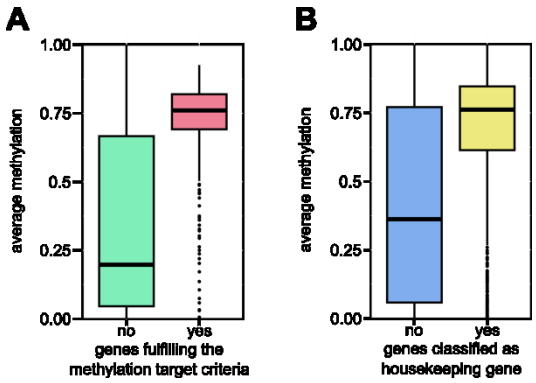


Figure 3.11 Housekeeping gene methylation.

Distribution of gene body methylation levels across genes fulfilling the methylation target criteria (**A**) and across genes classified as housekeeping genes (**B**). (**A**) Methylation target criteria are: low CpGo/e value (< 0.6), long gene sequence (≥ 10 kb), evolutionary conserved protein sequence (age node ≤ 3) and moderate expression rank (3^{rd} – 4^{th}). Genes matching all criteria are identified as methylation target (group "yes") and genes not matching one of these criteria as non-methylation target (group "no"). (**B**) Protein sequences of predicted genes were mapped to a list of human housekeeping genes (Eisenberg & Levanon, 2013) and genes with a significant hit (e-value < $1e^{-10}$) are identified as housekeeping genes (group "yes").

The observed characteristics of genes targeted by DNA methylation (Fig. **3.10**) are shared features of housekeeping genes and thus, suggest that housekeeping genes could be preferentially methylated. The *P. virginalis* genes were aligned to a published list of human housekeeping genes (Eisenberg & Levanon, 2013). Genes classified as housekeeping genes displayed an increased methylation level (mean 66 % and median 76 %) compared to the non-housekeeping genes (mean 42% and median 36 %; Fig. **3.11B**), which was similar for the comparison between target and non-target genes (Fig. **3.11A**). Consistently, the averaged CpGo/e value, gene length, gene age and expression rank of the housekeeping genes met the applied criteria for the methylation targets (CpGo/e 0.43, length 12,044 bp, gene age node 2.1, expression rank 3.1). In the group of target genes, 73.7 % of the genes were classified as housekeeping genes and only 1.7 % of genes in the other gene group. As such, housekeeping genes were 44 fold enriched in the group of genes targeted by DNA methylation. These results confirmed that gene body methylation targets housekeeping genes in *P. virginalis*.

Housekeeping gene methylation might fine-tune expression

As it was observed that gene body methylation shows a parabolic relationship with transcription (Fig. **3.10D**), this analysis was repeated for housekeeping genes. This showed that housekeeping genes with moderate expression were highly methylated, whereas highly and lowly expressed housekeeping genes displayed lower methylation levels (Fig. **3.12A**). Additionally, methylation levels along unexpressed housekeeping genes remained constant at the genome wide level of 9.2 %, while moderately expressed housekeeping genes showed the characteristic gene body methylation pattern with a methylation plateau at 71% in the gene body and a decreased methylation of 50 % upstream and downstream (Fig. **3.12B**). In summary, these results suggest that DNA methylation might fine-tune the expression of housekeeping genes in *P. virginalis*.

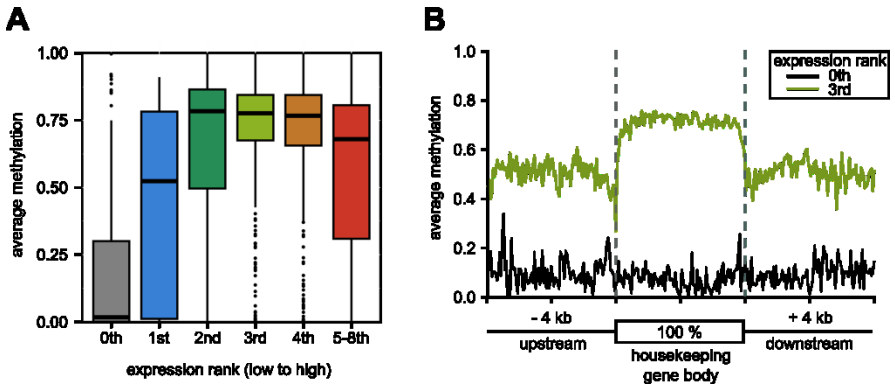


Figure 3.12 Housekeeping gene methylation might fine-tune expression.

(**A**) Distribution of gene body methylation levels across housekeeping genes grouped into different expression ranks. The 0^{th} rank represents all unexpressed genes (TPM = 0), while all expressed genes (TPM > 0) were distributed into 8 bins from least expressed (1^{st} rank) to most expressed ($5\text{-}8^{th}$) genes (**B**) Distribution of average methylation level along the gene bodies of unexpressed (expression rank 0^{th}: black line) and moderate expressed (3^{rd} expression rank: green line) housekeeping genes. Starting 4 kb upstream and ending 4 kb downstream of the transcription start site (TSS) and transcription termination site (TTS), respectively (indicated by vertical dashed lines).

3.2.4 Repeat Methylation

Since repetitive elements in invertebrates are reported to be unmethylated in several insect species, e.g. *Apis mellifera* or *Bombyx mori* (Feng et al., 2010; Zemach et al., 2010), but methylated in other species like the desert locust or the pacific oyster (C Falckenhayn et al., 2013; Feng et al., 2010; Xiaotong Wang et al., 2014), the methylation of repeats in *P. virginalis* was analyzed.

Hypomethylation of transposable elements and repeats

To test whether DNA methylation targets repetitive elements in *P. virginalis*, methylation levels were determined for repeat elements. The methylation within repetitive elements was reduced (methylation level of 21 %) relative to the immediate flanking regions (methylation level of 28 %) and increased with rising distance from the elements (methylation level of up to 32%; Fig. **3.13A**). A similar observation was reported for other invertebrates with unmethylated repeats (Feng et al., 2010; Zemach et al., 2010) suggesting that transposable elements and repeats are hypomethylated in *P. virginalis*. Nevertheless, some repeat elements were indeed methylated (Fig. **3.13B** and **3.13C**). Approximately 63 % of repetitive elements were unmethylated (< 0.1), while 17 % were highly methylated (> 0.7; Fig. **3.13B**) indicating that repeat methylation in *P. virginalis* might be targeted to a specific set of repeat elements.

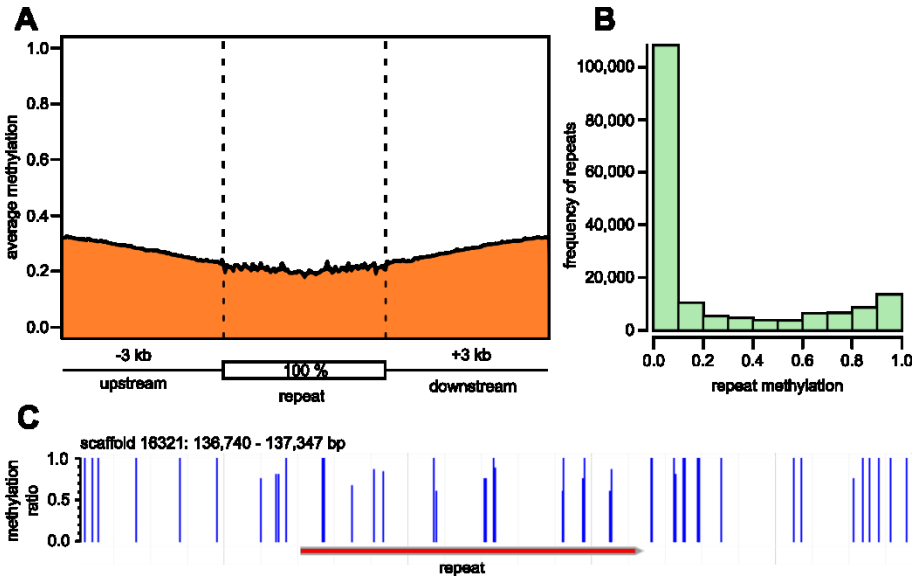


Figure 3.13 Repeat methylation.

(**A**) Distribution of average methylation level from 3 kb upstream and 3 kb downstream of the annotated repeats. Repeat start site and end site are indicated by vertical dashed lines. (**B**) Distribution of repeat methylation levels. (**C**) Example of repeat methylation in scaffold 16321. Methylation ratios of each CpG (blue vertical bars) and the annotated repeat (horizontal red bar below the methylation panel) are illustrated.

DNA transposons and old repeats are methylated

As young repeat elements in particular short interspersed elements (SINEs) were targets of DNA methylation in the pacific oyster *Crassostrea gigas* (Xiaotong Wang et al., 2014), the methylation of repeat classes and repeat divergence rates was analyzed in *P. virginalis*. DNA transposons showed the highest methylation level among all repeat classes (Fig. **3.14A**). The

methylation level of DNA transposons was twice as high (46 %) as the average repeat methylation level (Fig. **3.13A**) and close to the average gene body methylation level (Fig. **3.8D** and **3.9A**). This indicates that DNA methylation of repeats in *P. virginalis* is mainly targeted to DNA transposons. Especially old repetitive elements (divergence rate ≥ 21 %) had a higher methylation level than younger elements (Fig. **3.14B**) suggesting that some repeat elements might gain methylation over evolutionary time.

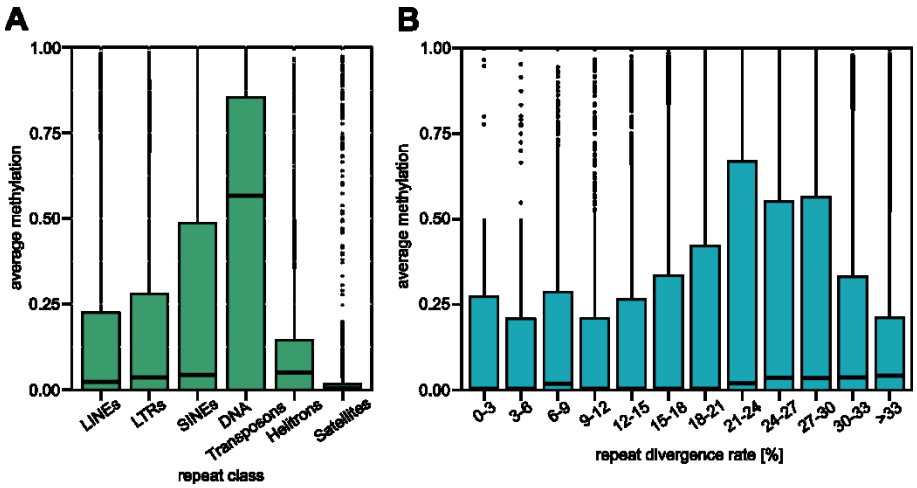


Figure 3.14 Features of target repeats.

Distribution of methylation levels across repeats classified into different repeat classes (**A**) and different repeat divergence rates (**B**). (**B**) The divergence rate of a repeat was determined by the sequence difference between the identified *P. virginalis* repeat and the sequence in the repeat library.

Methylation of repeats located within genes

Since transposable elements can be incorporated into genes as new exons (Sorek, 2007) and even contribute to entire genes (Feschotte & Pritham, 2007; Volff, 2006), the methylation of repeat elements and their location within the genome were analyzed. Indeed, repeats located within genes were higher methylated (average methylation 0.4), whereas repeats outside of genes were lower methylated (average methylation 0.2; Fig. **3.15**). Moreover, around 26 % of highly methylated repeats (average methylation ≥ 0.8) were incorporated into genes and only 10 % of the slightly methylated repeats (average methylation ≤ 0.2) were part of a gene. Thus, repetitive elements inside of genes are 2.7 x enriched in the group of repeats with a high methylation level compared to the group of low methylated repeats indicating that repetitive elements located within gene bodies are more likely to be methylated.

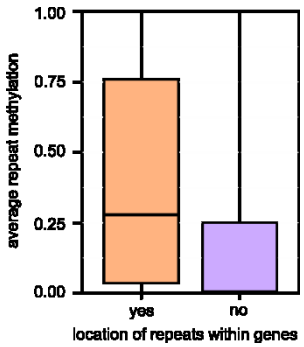


Figure 3.15 Repeat methylation as possible consequence of gene body methylation.
Distribution of repeats located within genes (orange) and outside of genes (purple).

3.3 Conservation of Gene Body Methylation

To further characterize DNA methylation in *P. virginalis*, additional whole-genome bisulfite sequencing (WGBS) of several individuals and distinct tissues was performed (for details see materials and methods table **3.2**). Since gene bodies are the main targets of DNA methylation in *P. virginalis*, the generated data were used for comparison of gene body methylation level between individuals, tissues and species.

3.3.1 Between Individuals

To investigate methylation differences in gene body methylation between individuals, the methylation patterns of the hepatopancreas were compared between two different individuals of our laboratory population. Notably, only 1.28 % (81 out of 6,333) of the compared genes displayed an absolute methylation difference higher than 0.2 between the individuals (Fig. **3.16**). Therefore, the inter-individual comparison of gene body methylation levels showed a high reproducibility of tissue-specific DNA methylation patterns (Fig. **3.16A**) for individuals reared under similar conditions.

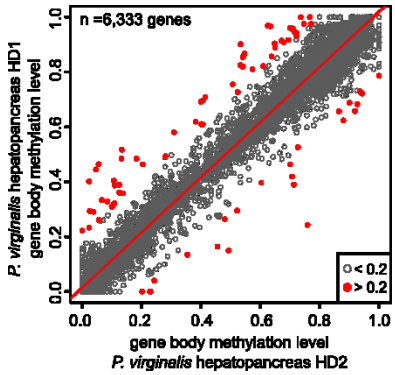


Figure 3.16 Reproducibility of tissue-specific gene body methylation patterns in *P. virginalis*.
Scatterplot of the average methylation level of a gene in both samples. The average metyhlation was calculated by the mean of at least 5 CpGs with a coverage of at least 3 in both samples. Depicted is the calculated regression line (red). Absolute methylation differences between the samples are colour coded: dark grey (≤ 0.2) and purple (> 0.2). Comparison: HD2 hepatopancreas vs. HD1 hepatopancreas.

3.3.2 Between Tissues

It has been reported that *C. intestinalis* sperm and muscle cells display an identical set of methylated and unmethylated genes (Suzuki et al., 2013). As this is the only known study investigating tissue variability of gene body methylation in an invertebrate animal (Suzuki et al., 2013), the gene body methylation levels were compared between tissues in *P. virginalis.* The percentage of genes with an absolute methylation difference greater than 0.2 was slightly higher between hepatopancreas and abdominal muscle (2.88 %; Fig. **3.17A**) than between hepatopancreas and gills (0.66 %; Fig. **3.17B**), which may be related to the particularly low sequencing coverage of the abdominal muscle sample (Table **5.2**). Overall, the comparison of different tissues from the same individual displayed a similar reproducibility of the methylation patterns in comparison to tissue-specific DNA methylation patterns (Fig. **3.16**). This suggests that gene body methylation in *P. virginalis* is tissue-invariant, which represents a major difference from the tissue-specificity of mammalian methylomes (Hon et al., 2013; Kundaje et al., 2015; Ziller et al., 2013).

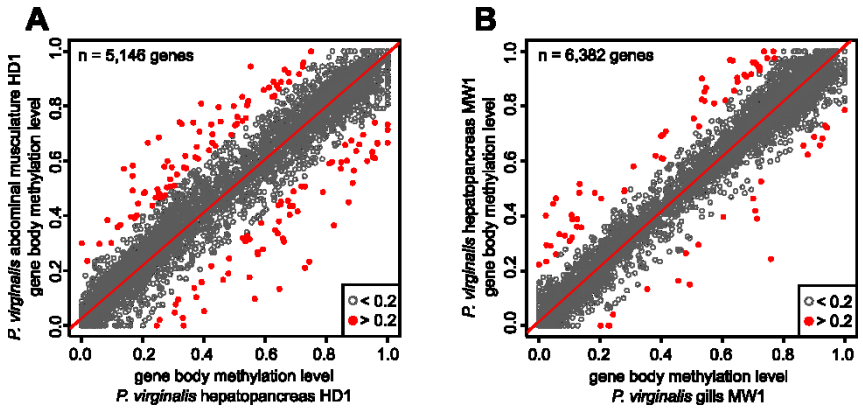


Figure 3.17 Reproducibility of inter-tissue gene body methylation patterns in *P. virginalis*.
Scatterplot of the average methylation level of a gene in both samples. The average metyhlation was calculated by the mean of at least 5 CpGs with a coverage of at least 3 in both samples. Depicted is the calculated regression line (red). Absolute methylation differences between the samples are colour coded: dark grey (≤ 0.2) and purple (> 0.2). (**A**) inter-individual: HD2 hepatopancreas vs. HD1 hepatopancreas. (**A**) HD1 hepatopancreas vs. HD1 abdominal musculature and (**B**) MW1 hepatopancreas vs. MW1 gills.

3.3.3 Between Species

Organisms with varying ploidy levels like the watermelon *Citrullus vulgaris* or the pond loach *Misgurnus anguillicaudatus* display differences in DNA methylation between the ploidy levels (Gardiner et al., 2015; Li et al., 2011; Zhang et al., 2016; Zhang et al., 2015; Zhou et al., 2016). Since *P. virginalis* is a triploid variant of the the diploid mother species *P. fallax* (Martin et al., 2016), genome-wide DNA methylation levels were compared between both species. To quantitatively determine the 5-methylcytosine level, mass-spectrometry was performed for abdominal musculature of three *P. virginalis* and three *P. fallax* adult animals. Remarkably, the detected global DNA methylation level was higher in *P. fallax* (2.92 %) than in *P. virginalis* (2.41 %; Fig. **3.18A**) suggesting that some genomic regions might be differentially methylated between both species.

As the detected 5-methylcytosine levels differ between *P. fallax* and *P. virginalis*, whole-genome bisulfite sequencing (WGBS) of *P. fallax* was performed. The gene body methylation patterns were compared between *P. fallax* and *P. virginalis*, because gene bodies are the main targets of DNA methylation in *P. virginalis*. A comparison of gene body methylation levels between the hepatopancreases of two *P. fallax* individuals revealed a small fraction (1.04 %) of genes with an absolute methylation divergence > 0.2 (Fig. **3.18B**). Therefore, a high inter-individual similarity was observed in P. fallax (Fig. **3.16**). Interestingly, when comparing the same tissue from *P. virginalis* and *P. fallax*, 3.79 % (240 out of 6,303) of the genes displayed an absolute methylation difference > 0.2 (Fig. **3.18C**). Finally, the overall methylation difference

between *P. fallax* and *P. virginalis* was higher (0.058) than between individuals from the same species (0.034 and 0.038, respectively; Fig. **3.18D**). However, the gene body methylation divergence between the species was lower than expected, based on the genome-wide methylation variation detected by mass-spectrometry (Fig. **3.18A**).

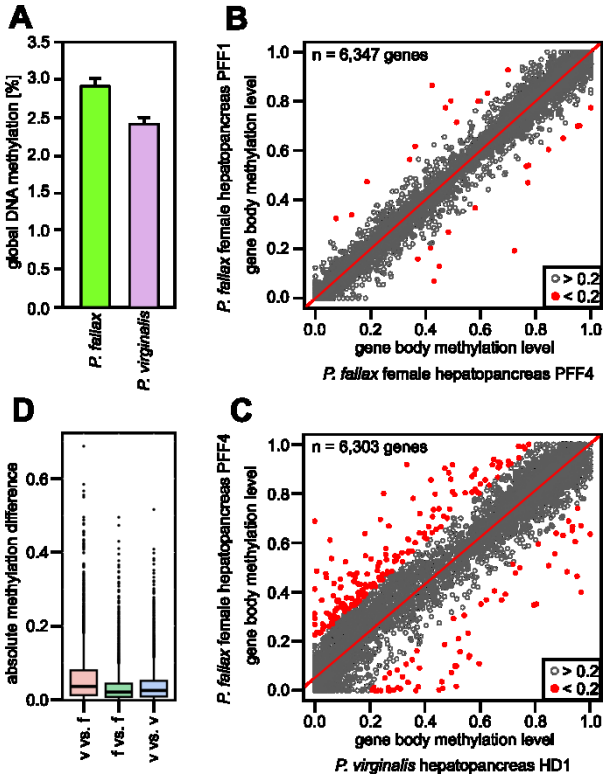


Figure 3.18 Comparison of DNA methylation between *P. fallax* and *P. virginalis*.
Comparison of global DNA methylation (**A**) and gene body methylation (**B** – **D**) between *P. fallax* and *P. virginalis*. (**A**) Quantitative analysis of genomic 5-methylcytosine levels of abdominal muscle form three adult *P. fallax* and *P. virginalis* individuals by mass spectrometry (performed by Katharina Schmid). (**B** and **C**) Scatterplot of the average methylation level of a gene in both samples. The average metyhlation was calculated by the mean of at least 5 CpGs with a coverage of at least 3 in both samples. Depicted is the calculated regression line (red). Absolute methylation differences between the samples are colour coded: dark grey (≤ 0.2) and purple (> 0.2). (**B**) *P. fallax* females inter-individual: PFF4 hepatopancreas vs. PFF1 hepatopancreas. (**C**) inter-species: *P. fallax* female (PFF4) hepatopancreas vs *P. virginalis* (HD1) hepatopancreas. (**D**) Boxplot of absolute methylation differences in gene bodies of two hepatopancreas samples. Pink: *P. virginalis* (v) vs. *P. fallax* (f). Green: *P. fallax* female 1 vs. *P. fallax* female 4. Blue: *P. virginalis* HD1 vs. *P. virginalis* HD2.

4 Discussion

In 2010 Zemach et al. and Feng et al. could show that the DNA methylation between vertebrates and invertebrates is different indicating that the DNA methylation in invertebrates may has a different role as in mammals. Since then several studies analyzed the methylomes of invertebrates, mainly insects (Cunningham et al., 2015; Falckenhayn et al., 2012; Glastad et al., 2011; Lyko et al., 2010; Xu Wang et al., 2013; Xiang et al., 2010). The only crustacean species among the non-insect invertebrates is *Daphnia pulex*, even though crustaceans comprise more than 40,000 species with a high phenotypic diversity and many crustaceans are keystone species with ecological and environmental relevance for their habitats (Colbourne et al., 2011; Günter Vogt, 2008). To broaden the knowledge about DNA methlyation in crustaceans, the methylome of the marbled crayfish was characterized. The findings presented in this study contribute additional information to the evolutionary conservation of gene body and repeat methylation in invertebrates and showed that housekeeping genes are the main targets of gene body methylation.

4.1 The Marbled Crayfish - an Independent Asexual Species

The common species concepts define speciation in a manner which is suitable for sexual reproducing organisms (Wheeler & Maier, 2000). For example, the biological species concept circumscribes new species based on its genetic isolation to other species in combination with the ability of sexual reproduction (Myr, 2000). Parthenogenetic organisms are per se genetically isolated from its species of origin (Martin et al., 2010) and reproduce asexually. Therefore, the biological species concept does not apply to these organisms (Myr, 2000). Consequently, the taxonomic treatment of parthenogenetic organisms is problematic and several contradictory suggestions have been made for their taxonomic treatment (Martin et al., 2010). Nevertheless, Martin et al. (2010) suggested to establish the parthenogenetic form of *Procambarus fallax*, the marbled crayfish, as a new species, if additional data confirm a regional wild population and/or a single origin. Chucholl and Pfeiffer (2010) described the first stable wild population of marbled crayfish in Germany confirming the first point of Martin et al. (2010). Since then several wild populations of marbled crayfish have been reported substantiating this criterion (Liptak et al., 2016; Lőkkös et al., 2016; Novitsky & Son, 2016). To consider the marbled crayfish as a new species and example of asexual speciation, it is indispensable to prove a single origin.

The complete mitochondrial genome sequence was assembled for the marbled crayfish (Fig. **3.1A**) and four individuals were analyzed, two from distinct laboratory populations (Heidelberg and Petshop) and two from different, stable wild populations (Moosweiher and

Madagascar) (Chucholl & Pfeiffer, 2010; J. P. G. Jones et al., 2009). The mitochondrial DNA sequences of the four marbled crayfish were identical (Fig. **3.1B**) indicating that they emerged from the same parthenogenetic lineage of *P. fallax*.

These results confirm a single origin of the marbled crayfish, which is especially important, since other species have populations of asexual lineages e.g. the freshwater snail *Potamopyrgus antipodarum* (Dybdahl & Lively, 1995). In contrast to the marbled crayfish, the asexual populations of the *P. antipodarum* evolved several times independently (Neiman, Jokela, & Lively, 2005). Similarly, the water flea *Daphnia* cyclically arrests in parthenogenesis and is capable to resume sexual reproduction under suitable conditions (Ebert, 2005). Taken together, asexual reproduction in the freshwater snail and *D. pulex* occurs naturally and is a kind of survival strategy. Consequently, the asexual lineages of these animals do not represent new species.

Since marbled crayfish and *P. fallax* are morphologically similar (Martin et al., 2010), it cannot be ruled out that wild populations of mixed sexual and asexual reproducing individuals of *P. fallax* have been failed to notice. If mixed *P. fallax* wild populations of different origins would exist, the marbled crayfish would not represent a new species similar to asexual lineages of *D. pulex* and the freshwater snail. However, the assembled mitochondrial DNA sequence can now be used to distinguish between multiple origins or single origin of asexual reproducing *P. fallax* descendants. Nevertheless, the results point towards a single origin of the marbled crayfish populations dating them back to the first population reported in 1995 (Günter Vogt et al., 2004). As such, the marbled crayfish meet the criteria for asexual speciation mentioned by Martin et al. (2010) and therefore, should be considered as the independent species *Procambarus virginalis* as suggested by Martin et al. (2010). These results were part of a publication describing the marbled crayfish as an independent species (Procambarus virginalis) (Günter Vogt et al., 2015).

4.2 The *P. virginalis* Transcriptome - Good Quality of the First Assembly

Flow cytometric analysis of *P. virginalis* haemocytes revealed a genome size larger than the human genome (Fig. **3.2**) indicating that the genome assembly will be challenging and time consuming. Thus, the less complex transcriptome was assembled using a normalized RNA-Seq library prepared from various tissues.

Standard assembly statistics only reflect genome biases and methodologies e.g. average sequence length or fractions of undetermined bases (gaps) and do not to represent the completeness of genes (Simão et al., 2015). Quality assessment of *de novo* assembled

genomes and transcriptomes is especially challenging, since no established reference assembly is available as blue print (Iwasaki et al., 2016). Nonetheless, comprehensive sequence analyses to other close related species are used to estimate the quality of a new assembly (Colbourne et al., 2011; Richards et al., 2008; Simão et al., 2015; Tenlen et al., 2016).

The majority of arthropodan orthologs were completely assembled in the *P. virginalis* transcriptome and among the organisms with a first assembly only the transcriptomes of *Bombyx mori* and *Daphnia pulex* displayed a higher fraction of complete arthropodan orthologs (Fig. **3.3A**). Additionally, since the *P. virginalis* transcriptome was assembled from sequenced reads of a normalized library prepared from four different tissues instead of all tissue-types and some transcripts are only transcribed under specific environmental conditions, the *P. virginalis* transcriptome most probably does not contain all *P. virginalis* transcripts. However, the analysis of the DNA methylation system in *P. virginalis* revealed complete protein sequences for Dnmt1, Dnmt3 and Tet, while the protein sequences of Dnmt3 and Tet were incomplete in *D. pulex* (Fig. **3.6A**). Additionally, assembled *P. virginalis* protein sequences were confirmed by mass-spectrometry (Fig. **3.3C**). Moreover, the majority of transcripts could be annotated (Fig. **3.4B**) and showed sequence similarity to other organisms (Fig. **3.4D**). Even though the *P. virginalis* transcriptome does not contain all *P. virginalis* transcripts, the majority of assembled sequences seem to be complete and unlikely to be assembly artifacts.

Furthermore, the classification of the *P. virginalis* transcript sequences into bilaterian, pancrustacean, curstacean, decapodan and astacoidean transcripts revealed that the highest fraction are bilaterian-specific proteins (Fig. **3.4D**). This result is consistent with observations in other organisms like *Drosophila melanogaster* or *Homo sapiens*, but different to *Daphnia pulex* (Colbourne et al., 2011; Richards et al., 2008). The genome of *D. pulex* encodes for a minimum set of 31,000 genes and only 26 % are bilaterian-specific, whereas over 36 % are without detectable homology to other species (Colbourne et al., 2011). Therefore, Colbourne et al. (2011) concluded that more than a third of genes in *D. pulex* are *Daphnia*-specific which might play important roles in its ecoresponse. *D. pulex* is only one crustacean lineage out of more than 40,000 known species (Colbourne et al., 2011) and to the time of its publication the only crustacean with a published genome. Consequently, Colbourne et al. (2011) did not compare the genes of *Daphnia pulex* to another crustacean species. Hence, it is almost impossible to exactly define the amount of *Daphnia*-specific genes without comparison to a close related non-*Daphnia* species. Therefore, it is extremely likely that a considerable amount of *Daphnia*-specific genes is actually crustacean- or branchiopoda-specific. Though, since the publication of the *Daphnia pulex* genome, several crustacean transcriptomes have been assembled like *Litopenaeus vannamei* or *Pontastacus leptodactylus* (C. Li et al., 2012; Manfrin et al., 2013).

Thus, the protein sequences of *D. pulex*, *L. vannamei* and *P. leptodactylus* were included in the classification of the *P. virginalis* transcripts (Fig. **3.4C**) reducing the amount of non-homologous sequences (Fig. **3.4D**). This fraction of unique proteins is more similar to the portion of lineage-specific genes reported for other genomes e.g. *Strongylocentrotus purpuratus* or *Tribolium castaneum* (Colbourne et al., 2011; Richards et al., 2008) and consequently, maximal one fifth of the *P. virginalis* transcriptome might be *Procambarus*-specific. However, it is more likely that comparison to species of the Cambaridae family will further decrease the fraction of non-homologous sequences in the *P. virginalis* transcriptome.

Taken together, the first draft assembly of the *P. virginalis* transcriptome has a good quality and thus, is suitable to support the genome assembly and to get a first impression about the evidences for the *P. virginalis* methylome. Nevertheless, the transcriptome can be further improved by sequencing a broader range of tissue types and incorporation of the genome information into the assembly process.

4.3 *P. virginalis* - a Remarkable Crustacean Methylome

Initial analysis of the *P. virginalis* transcriptome revealed solid evidences for a methylome: first a methylation-dependent CpG depletion in protein coding sequences similar to *H. sapiens* (Fig. **3.5**), second a conserved DNA methylation system (Fig. **3.6A**), and third comparable high levels of 5-methylcytosine and remarkably low levels of 5-hydroxymethylcytosine (Fig. **3.6C** and **3.6D**). Based on these primary observations, whole-genome bisulfite sequencing (WGBS) of *P. virginalis* hepatopancreas was performed for a characterization of its methylome. The first examination of the WGBS-data showed that the methylome of *P. virginalis* shares the key features of animal methylomes (Zemach & Zilberman, 2010): CpG-specific, bimodal and symmetric methylation (Fig. **3.7**). Furthermore, the majority of analyzed *P. virginalis* scaffolds displayed a mosaic methylation pattern (Fig. **3.8C**), which is typical for an invertebrate methylome, indicating that the DNA methylation is targeted to specific genomic regions. Animal methylomes on the one hand share methylation of gene bodies and on the other hand are distinct in the methylation of transposable elements (Feng et al., 2010; Zemach et al., 2010; Zemach & Zilberman, 2010). Hence, the methylation of gene bodies and repeat elements were in focus of the subsequent study.

4.3.1 Conserved Gene Body Methylation

Invertebrate methylomes show a bimodal distribution of gene body methylation indicating that a specific set of genes are methylated (Cunningham et al., 2015; Falckenhayn et al., 2012; Lyko et al., 2010; Suzuki et al., 2013; Suzuki et al., 2007; Xiaotong Wang et al., 2014; Xu Wang et al., 2013). While the majority of analyses describes a negative correlation of CpG-density and gene body methylation, other characteristics for targeted gene methylation are reported sporadically (Cassandra Falckenhayn et al., 2012; Lyko et al., 2010; Suzuki et al., 2013, 2007; Xu Wang et al., 2013; Xiang et al., 2010). For example, Suzuki et al. (2007 and 2013) studied the correlation between gen body methylation and expression, whereas Cunningham et al. (2015) performed an gene ontology enrichment analyses of methylated genes. Consequently, a summarizing, in depth analysis of the gene body methylation characteristics might help to understand which genes are targeted by methylation and which features of methylated genes might be conserved among invertebrates.

Gene bodies in *P. virginalis* showed the typical gene body plateau methylation pattern of invertebrates (Fig. **3.9A**) (Feng et al., 2010; Zemach et al., 2010). Notably, introns were higher methylated than exons which is rarely observed in other animals (Fig. **3.8C**) (Feng et al., 2010; Lister et al., 2009). However, primary methylation analysis of the first draft *Locusta migratoria* genome revealed an methylation preference of introns over exons, which was even more pronounced than in *P. virginalis* (Xianhui Wang et al., 2014). It is likely that the automatic annotation could not identify all exons within the intronic regions of the *P. virginalis* genome, because the program *ab initio* predict genes based on signal detection of e.g. splice donor sites (Picardi & Pesole, 2010). Hence, these unidentified exons may contribute to a higher methylation level of introns. Nonetheless, further characterization of gene body methylation revealed preferential targeting of a subset of genes (Fig. **3.9B**) with following features: high CpG-depletion, long gene body sequence, evolutionary conservation and moderate expression (Fig. **3.10**). These characteristics are shared by housekeeping genes (Eisenberg & Levanon, 2013) and indeed, housekeeping genes displayed an increased methylation level compared to the non-housekeeping genes (Fig. **3.11B**). Nevertheless, the housekeeping genes were classified by sequence similarity to a list of human housekeeping genes (Eisenberg & Levanon, 2013). Consequently, some housekeeping genes in *P. virginalis* might not be identified or wrongly classified. However, comparison of the gene body methylation between several species like *Nicrophorus vespilloides* and *Nasonia vitripennis* displayed a high overlap between the highly methylated genes (Cunningham et al., 2015; Sarda et al., 2012) suggesting that only a minor fraction of *P. virginalis* genes might be misclassified.

Former publications about invertebrate methylation occasionally observed similar features of the targeted genes sets (length, age, CpGo/e, expression) and some performed gene ontology analyses which revealed an enrichment in housekeeping functions like metabolic process (Lyko et al., 2010; Suzuki et al., 2013; Suzuki et al., 2007; Xiaotong Wang et al., 2014). Therefore, the methylated sets were sometimes described as genes with "housekeeping gene" features (Cunningham et al., 2015; Suzuki et al., 2013, 2007; Xiaotong Wang et al., 2014; Xu Wang et al., 2013). The here described results in *P. virginalis* show for the first time that housekeeping genes are indeed the main targets of gene body methylation which might be conserved among invertebrates.

4.3.2 Housekeeping Gene Methylation May Facilitates Environmental Adaptability

Gene body methylation is widely conserved in eukaryotes and its discovery is rather recent (Suzuki et al., 2013). In mammals the genomes are ubiquitously methylated (Breiling & Lyko, 2015; Schübeler, 2015) and methylation occurs within gene bodies of active and inactive genes (Schübeler, 2015). Moreover, gene expression is modulated by tissue-specific methylation of regulatory regions like enhancers or promoters (Hon et al., 2013; Kundaje et al., 2015; Ziller et al., 2013). In contrast, methylation in invertebrates is generally associated with gene expression (Sarda et al., 2012; Xiang et al., 2010; Zemach et al., 2010) and targets housekeeping genes in *P. virginalis* (section 4.3.1). Even though gene body methylation is evolutionary conserved, the molecular and functional level is poorly understood (P. A. Jones, 2012; Sarda et al., 2012; Schübeler, 2015; Singer et al., 2015; Suzuki et al., 2013).

Interestingly, the gene expression of all genes as well as the housekeeping gene expression in *P. virginalis* revealed the typical parabolic relationship to gene body methylation, with moderately expressed housekeeping genes displaying the highest methylation level and housekeeping genes expressed at both extremes the lowest (Fig. **3.10D** and **3.12**). Hence, the results suggest that DNA methylation of housekeeping genes fine-tunes their expression.

The parabolic relationship of housekeeping gene methylation and expression might be explained by a model for transcription-coupled DNA methylation as described by Zilberman et al. (2006) based on their observations in *Arabidopsis thaliana*. During transcription polymerases disrupt the chromatin structure and preinitiation complexes (PICs) can form initiating an aberrant transcription (Fig. **4.1A**) (Zilberman et al., 2007). This aberrant transcript is then processed by Dicer into short interfering RNA (siRNA) which leads to the methylation of the homologous DNA as it was observed in *A. thaliana* (Chan, Henderson, & Jacobsen, 2005; Zilberman et al., 2007). The PIC formation depends on the transcription rate; highly expressed genes are occupied by

closely spaced polymerases (Fig. **4.1B**) and chromatin structures at low expressed genes are rarely disrupted, both preventing PIC formation (Zilberman et al., 2007). However, if a similar mechanism as described by Zilberman et al. (2007) is involved in housekeeping gene methylation in *P. virginalis*, needs to be addressed in future studies.

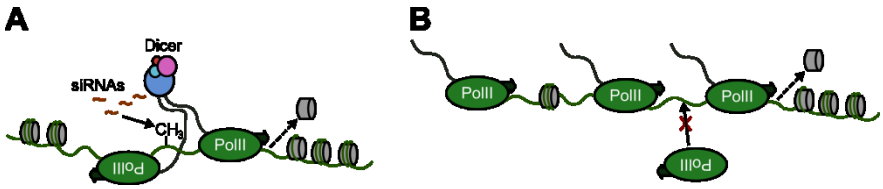


Figure 4.1 Model for transcription-coupled DNA methylation.
(**A**) Nucleosome disruption at a moderate expressed gene by a transiting polymerase, exposing a cryptic imitation site. Formation of a preinitiation complex (PIC) producing an aberrant transcript which is processed by Dicer into siRNAs causing methylation of the gene. (**B**) Nucleosome disruption at a highly expressed gene occupied by closely spaced polymerases which prevents PIC formation and consequenlty methylation of the gene. Figure is adopted from (Zilberman et al., 2007).

The impact of DNA methylation on housekeeping gene expression is further supported by the high reproducibility of tissue-specific gene body methylation patterns between individuals kept under the same conditions (Fig. **3.16**). In addition, gene bodies between different tissues showed identical methylation patterns (Fig. **3.17**) substantiating a possible role of DNA methylation in tissue-invariant expression of housekeeping genes. Interestingly, Suzuki et al. (2013) reported identical sets of methylated and non-methylated genes in different tissues of *Ciona intestinalis*, but in contrast did not associate their observation to a potential regulation of housekeeping gene expression by DNA methylation.

The results indicate a possible role of DNA methylation in fine-tuning of housekeeping expression in *P. virginalis* and is considerably different from its role in tissue-specific regulation of gene expression in mammals (Hon et al., 2013; Kundaje et al., 2015; Ziller et al., 2013). Thus, *P. virginalis* is an interesting model organism for environmental epigenetics. It is assumed that the methylome of an organism can be influenced by environmental signals to adapt to the new conditions (Feinberg, 2010; Lyko & Maleszka, 2011). Since the DNA methylation in mammals is tissue-specific and crucial in the regulation of gene expression during cell differentiation (Smith & Meissner, 2013), it seems that the methylation patterns are somewhat static and only minor changes can occur without altering the cellular identity (Fig. **4.2A**). In contrast, environmentally induced methylation changes in *P. virginalis* would probably lead to altered housekeeping gene expression remaining the cellular identity unaffected (Fig. **4.2B**). In summary, the here described results imply that housekeeping gene expression is regulated by DNA methylation and might be

a conserved feature of invertebrate methylomes. Hence, invertebrates and especially the clonal *P. virginalis* are meaningful model organisms to study the molecular basis by which DNA methylation connects the genome to the environment.

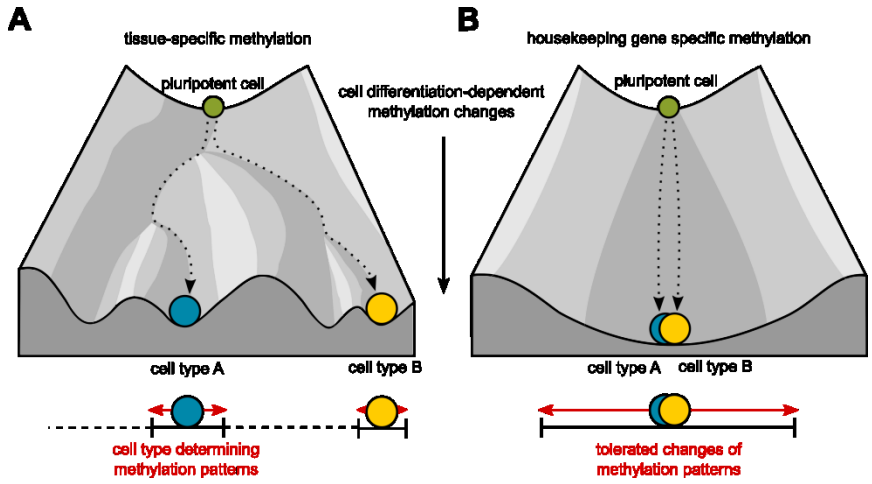


Figure 4.2 Schematic illustration for the range of tolerated methylation changes.
Methylation changes tolerated by species with (**A**) tissue-specific methylation patterns and (**B**) housekeeping gene specific methylation patterns. A pluripotent cell (green) can differentiate into cell type A (blue) or B (yellow) which is accompanied by methylation changes. (**A**) Tissue-specific methylation patterns, which differentiate the cell types from each other, display a narrowed range of tolerated methylation changes. (**B**) Housekeeping gene specific methylation patterns are highly similar between cell types and display a wider range of tolerated methylation changes. The figure is based on Waddigton's Classical Epigenetic Landscape proposed in 1957 (Goldberg et al., 2007).

4.3.3 Repeat Methylation Biased by Gene Body Methylation

While gene body methylation is a basal evolutionary feature of eukaryotic methylomes (Feng et al., 2010; Sarda et al., 2012; Zemach et al., 2010), the evolutionary conservation of repeat methylation is controversial. Zemach et al. (2010) concluded that repeat methylation was lost during early animal evolution and evolved independently in the vertebrate lineage to silence transposable elements (Zemach & Zilberman, 2010). Since then, several invertebrates with repeat methylation were reported, e.g. *Schistocerca gregaria* and *Crassostrea gigas* (Cassandra Falckenhayn et al., 2012; Xiaotong Wang et al., 2014), but also invertebrates with unmethylated repeats like *Nasonia vitripennis* were described (Xu Wang et al., 2013) implying that invertebrates use either cytosine-methylation mediated or cytosine-methylation independent mechanisms to silence transposable elements, like the Piwi-piRNA pathway (Aravin, Hannon, & Brennecke, 2007). Since the majority of analyzed invertebrates are insects, it is crucial to determine the methylation of repeat elements in other invertebrate families to further understand repeat methylation in invertebrates.

The majority of repetitive elements in *P. virginalis* displayed the typical pattern for hypomethylated repeat elements (Fig. **3.13A** and **3.13B**) as observed in other invertebrates (Feng et al., 2010; Zemach et al., 2010). However, a minor fraction was highly methylated (Fig. **3.13B** and **3.13C**) with DNA transposons being preferentially methylated (Fig. **3.14A**). As repeats can be incorporated into genes during evolution (Feschotte & Pritham, 2007; Sorek, 2007; Volff, 2006), repetitive elements located inside of gene bodies were higher methylated than repeat elements outside of genes (Fig. **3.15**), which is coherent with the priority methylation of older repeats (Fig. **3.14B**). In summary, repetitive elements in *P. virginalis* were hypomethylated and increased methylation levels of some repeat elements might be explained by their location within methylated gene bodies.

This is the first detailed analyses of repeat methylation in an invertebrate genome and confirms hypomethylation of repetitive elements in invertebrates, which suggest that methylation independent mechanisms may be utilized to silence transposable elements in *P. virginalis*. Moreover, the reported repeat methylation level in other invertebrates were either around the genome-wide methylation level or increased in a specific group of repeat elements (Cassandra Falckenhayn et al., 2012; Kao et al., 2016; Xianhui Wang et al., 2014; Xiaotong Wang et al., 2014). Together with the conservation of gene body methylation in these invertebrates, it might be possible that the published repeat methylation is biased by gene body methylation similar to the observation in *P. virginalis*. Moreover, Suzuki et al. observed that the methylation status of roughly six transposons in *C. intestinalis* was determined by its insertion site (Suzuki et al., 2007). This may also explain why Zemach et al. reported that transposable elements in *C. intestinalis* are hypomethylted, while Feng et al. observed a moderate methylation (Feng et al., 2010; Zemach et al., 2010). Furthermore, repetitive elements in *Nasonia vitripennis* are rarely methylated and the methylation of some elements is associated with activation rather than silencing, similar to gene body methylation (Xu Wang et al., 2013).

Concluding, gene body methylation might explain the inconsistency of published repeat methylation within invertebrates and may contribute to the discussion about the mechanisms of repeat silencing in invertebrates. However, a more detailed repetition of the published repeat methylation analyses will clarify the impact of gene body methylation on repeat methylation in invertebrates. Nevertheless, analysis of repeat methylation in additional species of other invertebrate families remain decisive to understand the evolution of repeat methylation.

4.4 Polyploidization - First Insights Into Methylation Changes

Some species vary in their ploidy level, especially plants like the watermelon *Citrullus vulgaris* which can be diploid, triploid and tetraploid (A. Li et al., 2011). Several studies were performed to analyze the DNA methylation changes which are associated with altered ploidy level (Gardiner et al., 2015; A. Li et al., 2011; Xiao et al., 2013; H.-Y. Zhang et al., 2016; J. Zhang et al., 2015; Zhou et al., 2016). However, the general methylation adaptations necessary for the polyploidization (generating a viable organism) are poorly understood. Since the *P. virginalis* reproduces parthenogenetically and is the triploid descendant of the diploid *P. fallax* (Martin et al., 2016), the *P. virginalis*-*P. fallax* pair might be a useful model for the understanding of the DNA methylation changes caused by polyploidization. Since DNA methylation contribute to the dosage compensation (Feil & Berger, 2007; Heard & Ditsteche, 2006; Martienssen & Colot, 2001), a first step was taken in this study by comparing gene body methylation patterns between *P. fallax* and *P. virginalis*.

The global 5-methylcytosine (5mC) level was increased in the diploid *P. fallax* compared to the triploid *P. virginalis* (Fig. **3.18A**). A decreased 5mC level in the triploid form was also observed in watermelon and *Salvia* (A. Li et al., 2011). However, in the triploid form of pear and *Poplar* the 5mC level was increased relative to the diploid form (A. Li et al., 2011). Notably, the observed global methylation differences were not reflected by gene body methylation (Fig. **3.18C**). Even though the gene body methylation divergence between *P. fallax* and *P. virginalis* was increased relative to the inter-individual comparison within the same species (Fig. **3.18D**), the observed variation was lower than expected and could not explain the difference in the global 5mC levels between *P. fallax* and *P. virginalis*. Thus, the gene body methylation patterns between both species are highly conserved.

The observed high reproducibility of gene body methylation patterns between *P. fallax* and *P. virginalis* suggests that the dosage compensation of the genes might be controlled by other epigenetic mechanisms like histone modifications (Feil & Berger, 2007; Heard & Ditsteche, 2006; Martienssen & Colot, 2001). However, in this study the methylation data were not correlated to differences in gene expression. Therefore, it cannot be ruled out that the overall but minor hypomethylation in *P. virginalis* relative to *P. fallax* is associated with differences in gene expression. For example, in humans hypomethylated exons of highly expressed genes were classified as potential enhancers involved in transcription elongation (Singer et al., 2015). As the methylation levels of *P. virginalis* and *P. fallax* were compared over the entire gene length, it might be possible that only specific regions of the genes display a high methylation difference which may correlate with altered gene expression.

Nevertheless, the minor difference in gene body methylation compared to the high global differences between both species might not be unexpected, since housekeeping genes are the main targets of methylation and other studies reported similar observations. For example, in the loach *Misgurnus anguillicaudatus,* global hypomethylation was observed with increasing ploidy level, but the genes tended to be rather hypermethylated than hypomethylated (Zhou et al., 2016). This supports that the main methylation changes from diploid *P. fallax* to triploid *P. virginalis* may not occur in gene bodies but in other genomic regions like promoters or repeats. Further, sub-genome-specific promoters were differentially methylated in polyploid wheat (Gardiner et al., 2015) and in polyploid rice more genes were differentially expressed than methylated, while the methylation of transposable elements was altered (H.-Y. Zhang et al., 2016; J. Zhang et al., 2015).

Finally, flow cytometric analysis of haemocytes from *P. fallax* and *P. virginalis* revealed an 1.4 x instead of 1.5 x increased genome content in the triploid *P. virginalis* compared to the diploid *P. fallax* (Günter Vogt et al., 2015). This indicates that some genetic information is either completely lost (no alleles) or partially lost (only two alleles instead of three). Thus, it might be possible that genome parts, which are critical in dosage compensation, are genetically regulated by the loss of the additional allele and not epigenetically regulated by DNA methylation. This may also explain the difference in the global methylation level, since mass-spectrometry detects methylation independent of the genomic context.

In summary, the global methylation differences between *P. fallax* and *P. virginalis* are not reflected by the gene body methylation differences and thus, the current results are not sufficient to explain the methylation changes during polyploidization and future studies need to address this problem in detail.

4.5 Conclusion and Outlook

Taken together, the findings presented in this doctoral thesis indicate that housekeeping genes are targeted by DNA methylation which might be evolutionary conserved among invertebrates. The sequence comparison of mitochondrial DNA between several marbled crayfish populations made an important contribution to the discussion about its taxonomic treatment and revealed that the marbled crayfish is a new asexual species termed *Procambarus virginalis*. The good quality of the assembled *P. virginalis* transcriptome enabled the characterization of the methylation-dependent CpG-depletion and the methylation machinery in *P. virginalis* confirming solid evidences for the existence of the *P. virginalis* methylome. The *P. virginalis* methylome showed characteristics typical for other invertebrate methylomes like gene body methylation.

Moreover, the observed influence of gene body methylation on repeat methylation in *P. virginalis* might explain the inconsistency of published repeat methylation within invertebrates and may be vital for the discussion about the evolutionary conservation of repeat methylation. Finally, the *P. virginalis* methylome was characterized by tissue-invariant housekeeping gene body methylation which might play a role in the fine-tuning of housekeeping gene expression. These features of the methylome enable *P. virginalis* to become an interesting model organism for environmental epigenetics. Furthermore, the gene body methylation patterns between the diploid *P. fallax* and the descendant triploid *P. virginalis* were highly conserved demonstrating that additional studies are necessary to identify the regions of polyploidization-dependent DNA methylation changes which explain the observed global methylation differences between both species.

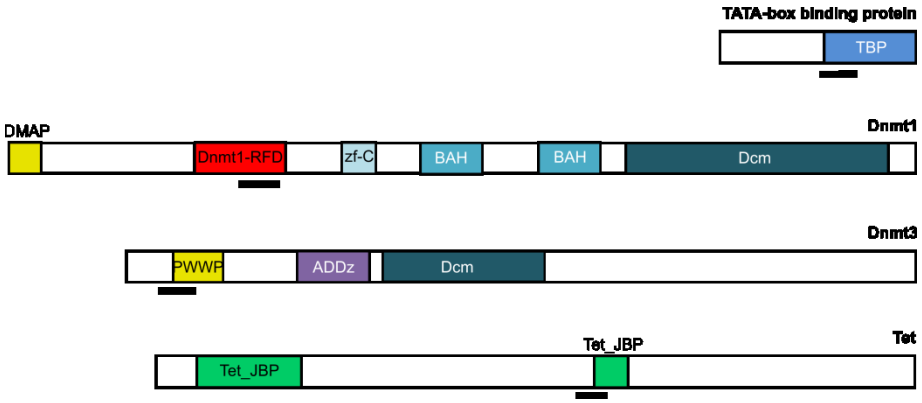
Currently common molecular biology techniques are limited for the application on *P. virginalis*. For example, specific antibodies need to be generated and established e.g. against Dnmt3. Dnmt3-antibodies could be used for the co-immunoprecipitation of possible interaction partners binding at the remarkably long C-terminal part of Dnmt3. Furthermore, *P. virginalis* cell cultures and a procedure to generate transgenic *P. virginalis* individuals are currently not available. Thus, knock-out/-down experiments to investigate molecular mechanisms like repeat silencing or transcription-coupled DNA methylation described by Zilberman et al. (2006) cannot be performed. Nevertheless, the new "Assay for Transposase-Accessible Chromatin using Sequencing" (ATAC-Seq) can be easily applied (Buenrostro et al., 2015) to perform epigenomic profiling of open chromatin analyzing chromatin accessibility, nucleosome positioning and factor occupancy by DNA-binding proteins (Buenrostro et al., 2013). Integration of the methylome and expression data with the ATAC-Seq data will provide new insights into the interplay between DNA methylation, gene expression and chromatin structure in *P. virginalis*. This will broaden the knowledge about the *P. virginalis* epigenome and may give an idea how the expression of tissue-specific genes and housekeeping genes is epigenetically regulated.

5 Appendix

5.1 Primer Location and Amplicon Sequences for Table 2.1

Location of primers which were used for qRT-PCR within the targeted proteins and the corresponding amplicon sequence are illustrated for Dnmt1, Dnmt3, Tet and TBP.

A



B

Amplicon			Targeted Enzyme		Primer
5'-3' Sequence	**Name**	**Length [bp]**	**Name**	**Domain**	**Name**
	TBP2	122	TATA-box BP	TBP	CasF_027 (fw); CasF_028 (rv)
CCACAGCTACAGAACATCGTTTCTACAGTCAACTTAAATTGTAAGCTCGACCTAAAGAAAATAGCTTTGCATGCTCGTAATGCC GAATATAATCCCAAACGTTTTGCAGCCGTCATCATGAG					
5'-3' Sequence	Dnmt1.2	150	Dnmt1	Dnmt1-RFD	CasF_007 (fw); CasF_008 (rv)
GGAGAAGGCACTGATTGGATTCTCTACTTCATATGCTGAATATATACTAATGGATCCAAGTGACACGTACGCTCCATTTGTTGA TGCTGTTAGAGAGAAGATTTACATTAGTAAAATAGTGATTGAGTTTCTGGTGAACAACGATGATGC					
5'-3' Sequence	Dnmt3.1	133	Dnmt3	PWWP	CasF_009 (fw); CasF_010 (rv)
GAATGGAACATCAGCACCTGCTAATTCTGTATCCAGTACTCACTATGGAAGACTTGTGTGGGCCAAGATTTCAGGTTCCAGATC CTGGCCAGCTGTCATTGTGAACCATGAAGATTGTGGAATGAGAGCACCG					
5'-3' Sequence	Tet3	100	Tet	Tet-JBP	CasF_025 (fw); CasF_026 (rv)
CCAGTAGAAGTGATCAACAGTGTAATAAACCCAGAGAACCAGAAACAGTAATCAAACAGAGTGACAATGTTGAGAATTTCCACG ATCCAGATATTGGAGG					

Figure 5.1 Location of primer and amplicon sequences used for qRT-PCR.

(**A**) Location of primer and amplicon sequences within the protein sequences are indicated by black horizontal bars. Depicted are the virtually translated protein sequences of TATAbox binding protein (TBP), Dnmt1, Dnmt3 and Tet. The conserved domains are shown as coloured boxes. TBP: TATAbox binding domain (blue). Dnmt1: DMAP1 binding domain (yellow), replication foci domain (red), CXXC zinc finger domain (azure) and catalytic domain (dark blue). Dnmt3: PWWP domain (yellow), zinc finger domain (purple) and catalytic domain (dark blue). Tet: catalytic domain (green). (**B**) Corresponding amplicon sequences of the used qRT-PCR primers. Primer sequences are shown in table **2.1**.

5.2 Not Used Data Sets

Complete list of data sets, which were not used for the analyses, are listed together with their corresponding sequencing approach and sample information.

Table 5.1 Overview of sequenced but not analyzed samples.

Sequencing overview of samples, which were not used for the analyses, are listed per animal ID, sequencing approach and tissue. RNA-Seq: whole transcriptome sequencing. Hepato: hepatopancreas. Antennal: antennal glands (green glands). abdM: abdominal muscle.

species	strain/ sex	animal ID	tissue	seqtype
P. virginalis	Heidelberg	HD1	abdM	RNA-Seq DKFZ
			hepato	RNA-Seq DKFZ
			antennal	RNA-Seq DKFZ
		HD2	abdM	RNA-Seq DKFZ
	Petshop	Pet4	abdM	RNA-Seq DKFZ
			hepato	RNA-Seq DKFZ
	Moosweiher	MW1	hepato	RNA-Seq DKFZ
P. fallax	female	PFF1	abdM	RNA-Seq DKFZ
		PFF3	abdM	RNA-Seq DKFZ
		PFF4	abdM	RNA-Seq DKFZ

5.3 Coverage of WGBS Data Sets

Complete list of WGBS data sets and their corresponding fraction of covered CpGs and strand-specific fold base coverage of those CpGs are listed per animal ID and tissue.

Table 5.2 Coverage of WGBS data sets.

Coverage overview of WGBS data sets listed per animal ID and tissue. Faction: portion of CpGs covered by at least one read. Per base: strand-specific fold coverage of each covered CpG.

Species	strain/ sex	animal ID	tissue	fraction	per base
P. virginalis	Heidelberg	HD1	abdM	42%	8.6 x
			hepato	62%	10.1 x
		HD2	hepato	64%	8.4 x
	Moosweiher	MW1	hepato	62%	7.8 x
			gills	67%	15.4 x

P. fallax	female	PFF1	hepato	58%	10.7 x
		PFF4	abdM	57%	10.6 x
			hepato	57%	10.2 x

5.4 Species Used for Phylostratigraphic Analyses

Complete list of organisms, which were used for the phylostratigraphic analyses, are listed together with their corresponding age node (level number), phylostrata (level name) and total amount of used protein sequences for the phylostrata.

Table 5.3 List of Species used in phylostratigraphic analysis.

Phylostrata and corresponding species used as representatives for phylostratigraphic analyses.

Level number	Level name	Species	# Protein sequences
1	Cellular organism	*uncultured bacterium*	7,480,913
		Escherichia coli	
		Stigmatella aurantiaca DW4/3-1	
		Streptomyces hygroscopicus ATCC 53653	
		Streptomyces sp. AA4	
		Burkholderia multivorans ATCC 17616	
		Streptomyces ghanaensis ATCC 14672	
		Streptomyces viridochromogenes DSM 40736	
		Acaryochloris marina MBIC11017	
		Agrobacterium tumefaciens str. C58	
		Amycolatopsis mediterranei U32	
		Bacillus cereus 03BB108	
		Bradyrhizobium japonicum USDA 110	
		Chitinophaga pinensis DSM 2588	
		Clostridium carboxidivorans P7	
2	Eukaryota	*Arabidopsis thaliana*	677,949
		Chlamydomonas reinhardtii	
		Dictyostelium discoideum AX4	
		Entamoeba histolytica HM-1:IMSS	
		Giardia lamblia ATCC 50803	
		Paramecium tetraurelia strain d4-2	
		Physcomitrella patens subsp. patens	
		Phytophthora infestans T30-4	

		Picea sitchensis	
		Plasmodium falciparum 3D7	
		Tetrahymena thermophila	
		Thalassiosira pseudonana CCMP1335	
		Toxoplasma gondii ME49	
		Trichomonas vaginalis G3	
		Trypanosoma brucei TREU927	
3	Opisthokonta	*Ashbya gossypii ATCC 10895*	257,365
		Aspergillus fumigatus Af293	
		Aspergillus nidulans FGSC A4	
		Candida dubliniensis CD36	
		Cryptococcus neoformans var. neoformans JEC21	
		Gibberella zeae PH-1	
		Magnaporthe oryzae 70-15	
		Monosiga brevicollis MX1	
		Neurospora crassa OR74A	
		Pichia pastoris GS115	
		Saccharomyces cerevisiae S288c	
		Scheffersomyces stipitis CBS 6054	
		Schizosaccharomyces pombe	
		Ustilago maydis 521	
4	Metazoa	*Amphimedon queenslandica*	13,802
5	Eumeta-zoa	*Hydra magnipapillata*	96,838
		Nematostella vectensis	
		Trichoplax adhaerens	
6	Bilateria	*Bos taurus*	2,048,839
		Canis lupus familiaris	
		Danio rerio	
		Gallus gallus	
		Homo sapiens	
		Mus musculus	
		Pan troglodytes	
		Rattus norvegicus	
		Saccoglossus kowalevskii	
		Strongylocentrotus purpuratus	
		Tetraodon nigroviridis	

		Xenopus laevis	
7	Protostomia	*Aplysia californica*	228,974
		Lottia gigantea	
		Crassostrea gigas	
		Capitella teleta	
		Helobdella robusta	
8	Arthropoda	*Steganacarus magnus*	44,218
		Loxosceles reclusa	
		Ixodes ricinus	
		Achipteria coleoptrata	
		Platynothrus peltifer	

5.5 GpCo/e Distributions for Figure 3.5

Calculated GpCo/e distributions are depicted for protein-coding sequences analyzed for historical germline methylation.

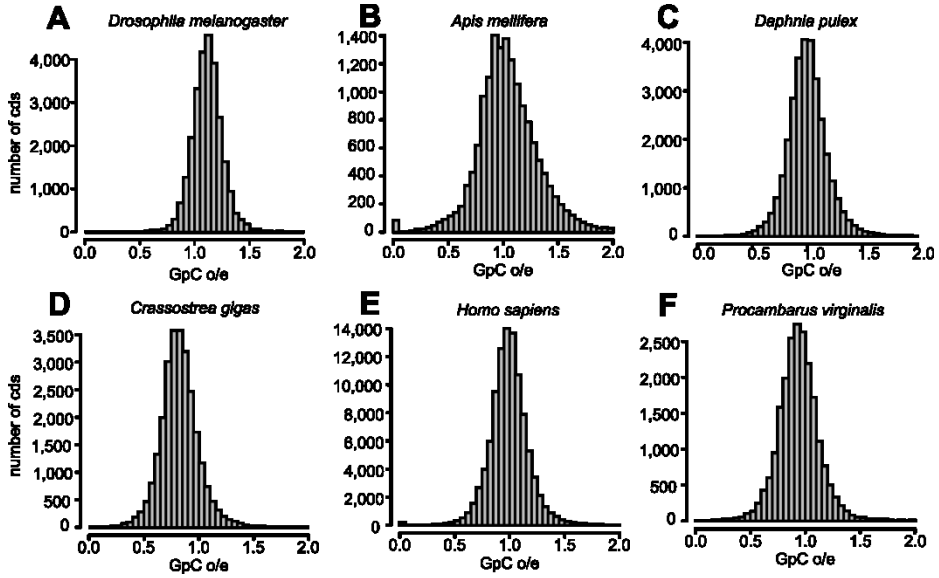


Figure 5.2 Control GpCo/e values.

Control plots for evolutionary CpG depletion in protein-coding sequences (cds) of various species (Fig. **3.6**). Distribution of normalized GpC [amount of observed GpCs to amount of expected GpCs (o/e)] content. Plots A to E are ordered from the lowest to the highest genome-wide methylation level: (**A**) *Drosophila melanogaster* (lacking DNA methylation) (Raddatz et al., 2013), (**B**) *Apis mellifera* (0.11 %) (Lyko et al., 2010), (**C**) *Daphnia pulex* (0.25 %) (Asselman et al., 2016), (**D**) *Crassostrea gigas* (1.96 %) (Xiaotong Wang et al., 2014) and (**E**) *Homo sapiens* (3.93 %) (Lister et al., 2009). **(F)** Distribution of GpCo/e values in *P. virginalis* protein-coding sequences.

5.6 Examples of Gene Body and Repeat Methylation

Apollo Example of Gene Body Methylation for Figure 3.9 and 3.10

Screenshots of Apollo Genome Browser displaying examples of methylated and unmethylated gene bodies and examples of gene body methylation in short vs. long genes, old vs. young genes and unexpressed vs. moderate expressed genes in P. virginalis hepatopancreas (sample HD2).

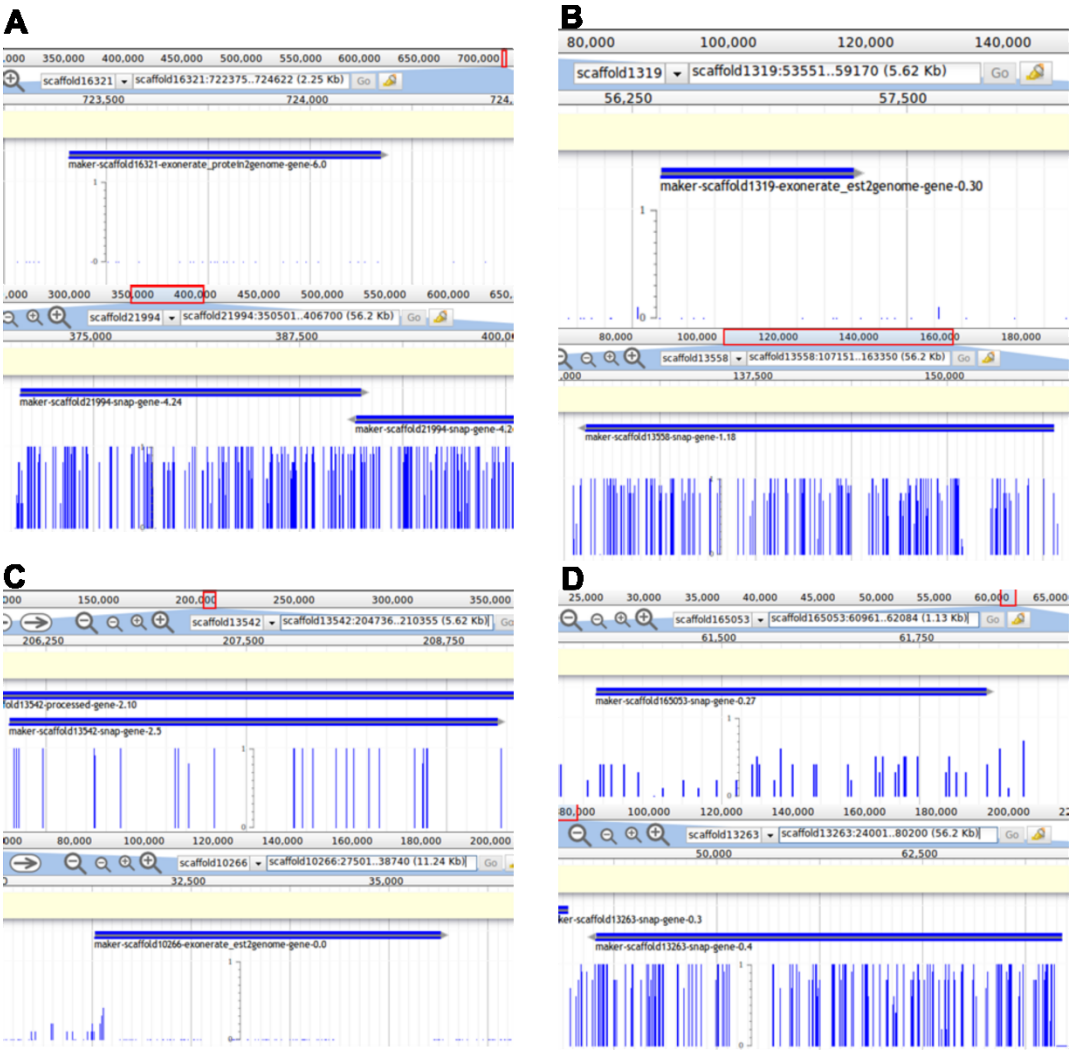


Figure 5.3 Examples of gene body methylation and feature of target genes.

Corresponding examples for Fig. **3.9** and **3.10**. Methylation ratios of each CpG (blue vertical bars) and the predicted gene (horizontal blue bar above the methylation panel) are illustrated. (**A**) Bimodal gene body methylation: complete unmethylated gene (top) and heavily methylated gene (bottom). (**B**) Length: short gene (< 1 kb, top) and long gene (> 20 kb, bottom). (**C**) Gene age: old gene (age group 1, top) and young gene (age group 9, bottom). (**D**) expression: unexpressed gene (rank 0th, top) and moderate expressed gene (rank 3rd, bottom).

Apollo Examples of Repeat Methylation for Figure 3.14

Screenshots of Apollo Genome Browser displaying examples of repeat methylation in two different repeat classes and with different divergence rate in P. virginalis hepatopancreas (sample HD2).

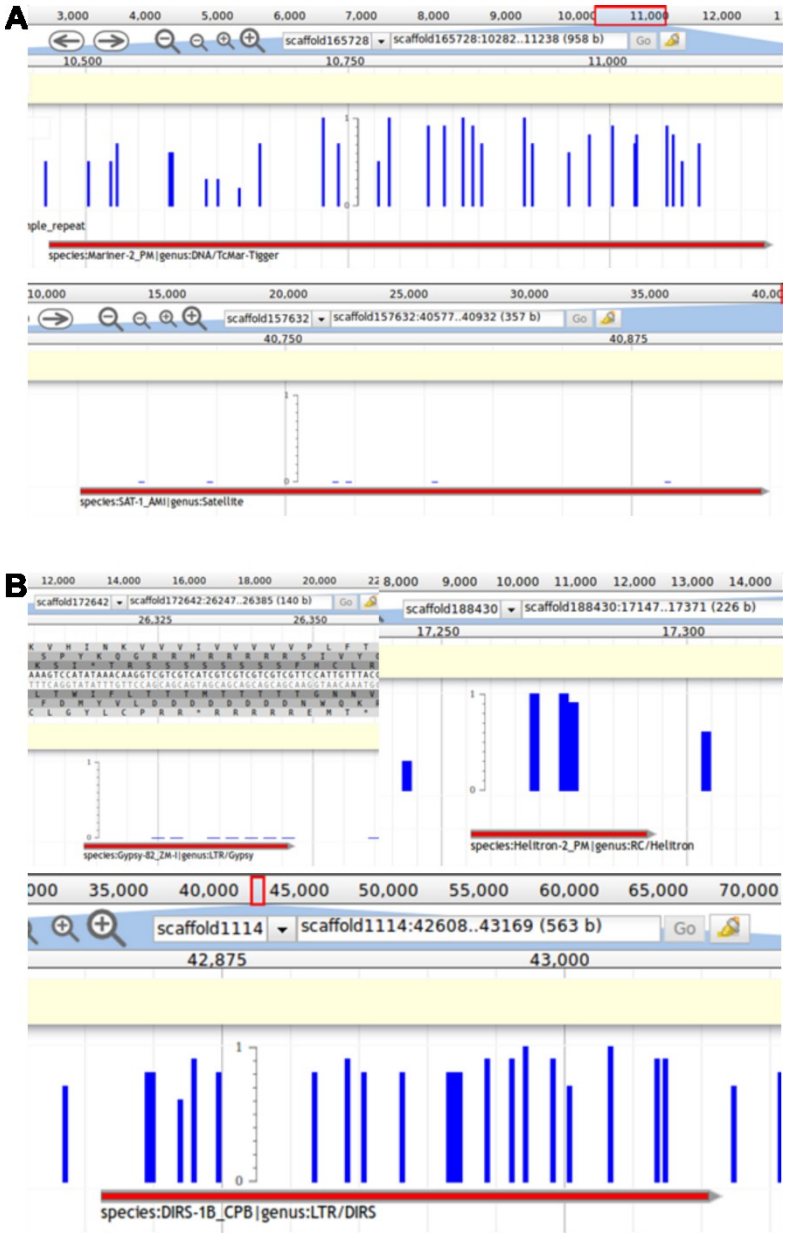


Figure 5.4 Examples of repeat methylation features.

Corresponding example for Fig. **3.14**. Methylation ratios of each CpG (blue vertical bars) and the annotated repeat (horizontal red bar below the methylation panel) are illustrated. (**A**) Repeat class: DNA-transposon (top) and satellite (bottom). (**B**) Repeat divergence rate: low diverged repeats (0 % and 2.8 %, top left and right) and high diverged repeat (36 %, bottom).

Apollo Example of Repeat Methylation for Figure 3.15

Screenshots of Apollo Genome Browser depicting examples of repeat methylation in repeats located inside vs. outside of genes in P. virginalis hepatopancreas (sample HD2).

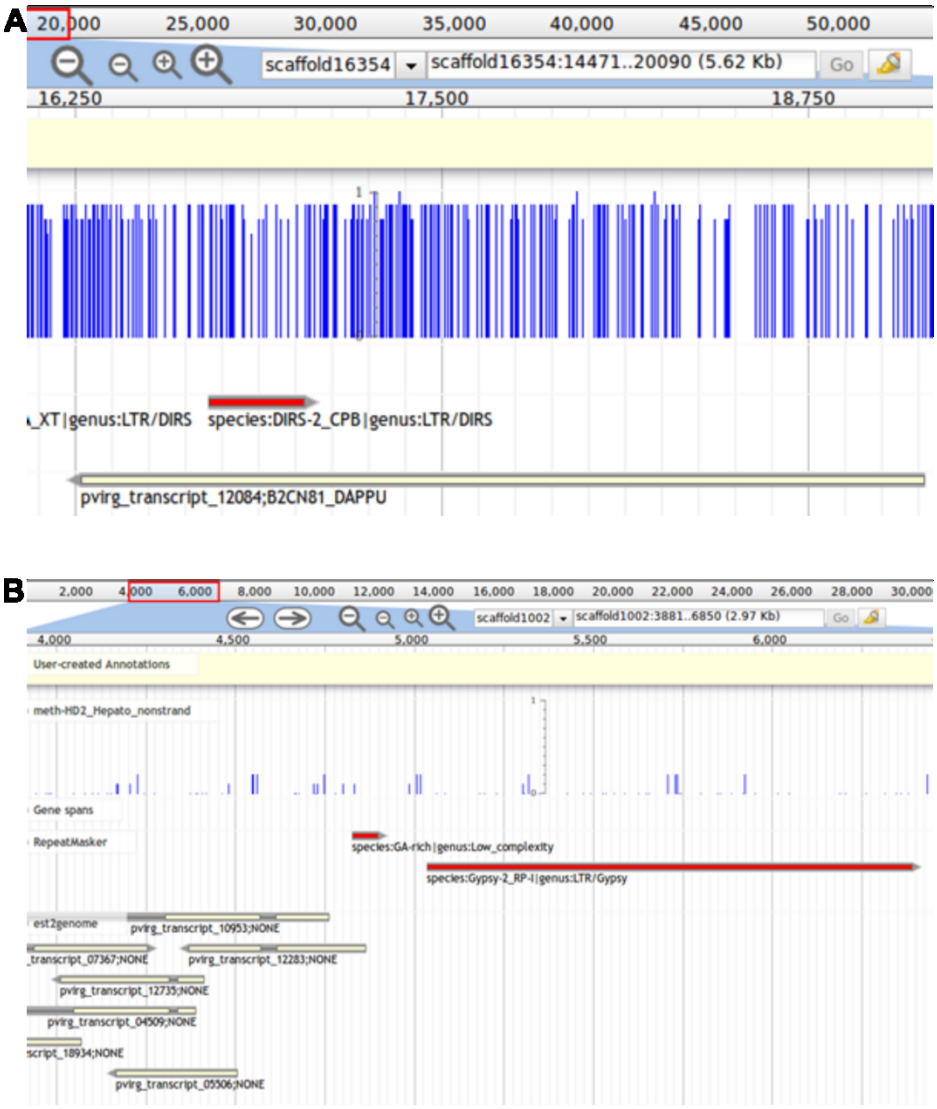


Figure 5.5 Examples of repeat methylation within genes and outside of genes.

Corresponding examples for Fig. **3.15**. Repeats with a high divergence rate (> 24 %) in scaffold 16354 (**A**) and 1002 (**B**), respectively. Methylation ratios of each CpG (blue vertical bars, top), the annotated repeat (horizontal red bar, middle) and the predicted genes (horizontal light yellow bar, bottom) are illustrated. (**A**) High diverged repeat (33 %) within a predicted gene. (**B**) High diverged repeats (24 % and 34 %, respectively) outside of predicted genes.

List of Publications

Theissinger K, Falckenhayn C, Blande D, Toljamo A, Gutekunst J, Makkonen J, Jussila J, Lyko F, Schrimpf A, Schulz R and Kokko H. (2016). *De Novo assembly and annotation of the freshwater crayfish Astacus astacus transcriptome.* Mar Genomics. 28:7-10.

Vogt G, Falckenhayn C, Schrimpf A, Schmid K, Hanna K, Panteleit J, Helm M, Schulz R and Lyko F. (2015). *The marbled crayfish as a paradigm for saltational speciation by autopolyploidy and parthenogenesis in animals.* Biol Open. 4(11):1583-94.

Falckenhayn C, Boerjan B, Raddatz G, Frohme M, Schoofs L and Lyko F. (2013). *Characterization of genome methylation patterns in the desert locust Schistocerca gregaria.* J Exp Biol. 216(8):1423-9.

Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C and Maleszka R. (2010). *The honey bee epigenomes: differential methylation of brain DNA in queens and workers.* PLoS Biol. 8(11):e1000506.

References

02.01 Gewässergüte Chemie. (2004). Berlin.

Aravin, A. a, Hannon, G. J., & Brennecke, J. (2007). The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race. *Science*, *318*(November), 761–764.

Asselman, J., De Coninck, D. I., Pfrender, M. E., & De Schamphelaere, K. A. (2016). Gene body methylation patterns in *Daphnia* are associated with gene family size. *Genome Biology and Evolution*, *8*(4), evw069. http://doi.org/10.1093/gbe/evw069

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, *21*, 381–395. http://doi.org/10.1038/cr.2011.22

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., … Zhang, J. (2015). UniProt: A hub for protein information. *Nucleic Acids Research*, *43*(D1), D204–D212. http://doi.org/10.1093/nar/gku989

Bernstein, B. E., Meissner, A., & Lander, E. S. (2007). The Mammalian Epigenome. *Cell*, *128*(4), 669–681. http://doi.org/10.1016/j.cell.2007.01.033

Bird, A. (2007). Perceptions of epigenetics. *Nature*, *447*(7143), 396–8. http://doi.org/10.1038/nature05913

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, *8*(7), 1499–1504. http://doi.org/10.1093/nar/8.7.1499

Blake, J. A., Christie, K. R., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., … Westerfeld, M. (2015). Gene ontology consortium: Going forward. *Nucleic Acids Research*, *43*(D1), D1049–D1056. http://doi.org/10.1093/nar/gku1179

Bohman, P., Edsman, L., Martin, P., & Scholtz, G. (2013). The first Marmorkrebs (Decapoda: Astacida: Cambaridae) in Scandinavia. *BioInvasions Records*, *2*(3), 227–232. http://doi.org/10.3391/bir.2013.2.3.09

Breiling, A., & Lyko, F. (2015). Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics {&} Chromatin*, *8*(1), 24. http://doi.org/10.1186/s13072-015-0016-6

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open

chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–8. http://doi.org/10.1038/nmeth.2688

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, *2015*(January), 21.29.1-21.29.9. http://doi.org/10.1002/0471142727.mb2129s109

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST plus: architecture and applications. *BMC Bioinformatics*, *10*(421), 1. http://doi.org/Artn 421\nDoi 10.1186/1471-2105-10-421

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., … Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196. http://doi.org/10.1101/gr.6743907

Chan, S. W. L., Henderson, I. R., & Jacobsen, S. E. (2005). Gardening the genome: DNA methylation in Arabidopsis thaliana. *Nature Reviews Genetics*, *6*(5), 351–360. http://doi.org/10.1038/nrg1664

Chucholl, C., & Pfeiffer, M. (2010). First evidence for an established Marmorkrebs (Decapoda, Astacida, Cambaridae) population in Southwestern Germany, in syntopic occurrence with Orconectes limosus (Rafinesque, 1817). *Aquatic Invasions*, *5*(4), 405–412. http://doi.org/10.3391/ai.2010.5.4.10

Clark, T. A., Murray, I. A., Morgan, R. D., Kislyuk, A. O., Spittle, K. E., Boitano, M., … Korlach, J. (2012). Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Research*, *40*(4), 1–12. http://doi.org/10.1093/nar/gkr1146

Clarke, J., Wu, H., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, *4*(April), 265–270. http://doi.org/10.1038/nnano.2009.12

Claros, M. G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., & Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. *Biology*, *1*(2), 439–59. http://doi.org/10.3390/biology1020439

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., … Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA

methylation patterning. *Nature*, *452*(7184), 215–219. http://doi.org/10.1038/nature06745

Colbourne, J. K., Pfrender, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., … Boore, J. L. (2011). The ecoresponsive genome of Daphnia pulex. *Science (New York, N.Y.)*, *331*(6017), 555–561. http://doi.org/10.1126/science.1197761

Colella, S., Shen, L., Baggerly, K. A., Issa, J. P., & Krahe, R. (2003). Sensitive and quanitative universal Pyrosequencing methylation analysis of CpG sites. *Biotechniques*, *35*(1), 146–150.

Crandall, K. A. (2010). Procambarus fallax. http://doi.org/http://dx.doi.org/10.2305/IUCN.UK.2010-3.RLTS.T153961A4569411.en

Cunningham, C. B., Ji, L., Wiberg, R. A. W., Shelton, J., McKinney, E. C., Parker, D. J., … Moore, A. J. (2015). The Genome and Methylome of a Beetle with Complex Social Behavior, Nicrophorus vespilloides (Coleoptera: Silphidae). *Genome Biology and Evolution*, *7*(12), 3383–96. http://doi.org/10.1093/gbe/evv194

Dimmer, E., Huntley, R., Barrell, D., Binns, D., Draghici, S., Camon, E., … Lovering, R. (2008). The Gene Ontology - Providing a Functional Role in Proteomic Studies. *Practical Proteomics*, *25*(March 2008), 2–11. http://doi.org/10.1002/pmic.200800002

Dümpelmann, C., & Bonacker, F. (2012). Erstnachweis des Marmorkrebses Procambarus fallax f . virginalis (Decapoda : Cambaridae) in Hessen. *Forum Flusskrebse*, *18*(Hlug 2010), 3–14.

Duncan, E. J., Gluckman, P. D., & Dearden, P. K. (2014). Epigenetics, plasticity, and evolution: How do we link epigenetic change to phenotype? *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *322*(4), 208–220. http://doi.org/10.1002/jez.b.22571

Dybdahl, M. F., & Lively, C. M. (1995). Diverse, endemic and polyphyletic clones in mixed populations of a freshwater snail (Potamopyrgus antipodarum). *Journal of Evolutionary Biology*, *8*(3), 385–398. http://doi.org/10.1046/j.1420-9101.1995.8030385.x

Ebert, D. (2005). Introduction to Daphnia Biology. In D. Ebert (Ed.), *Ecology, Epidemiology and Evolution of Parasitism in Daphnia* (pp. 5–18). Bethesda: National Library of Medicine (US), National Center for Biotechnology Information.

Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in*

Genetics, *29*(10), 569–574. http://doi.org/10.1016/j.tig.2013.05.010

Falckenhayn, C., Boerjan, B., Raddatz, G., Frohme, M., Schoofs, L., & Lyko, F. (2012). Characterization of genome methylation patterns in the desert locust Schistocerca gregaria. *The Journal of Experimental Biology*, *4*, 1423–1429. http://doi.org/10.1242/jeb.080754

Falckenhayn, C., Boerjan, B., Raddatz, G., Frohme, M., Schoofs, L., & Lyko, F. (2013). Characterization of genome methylation patterns in the desert locust Schistocerca gregaria. *J Exp Biol*, *216*(Pt 8), 1423–1429. http://doi.org/10.1242/jeb.080754

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., & Bock, C. (2015). Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Reports*, *10*(8), 1386–1397. http://doi.org/10.1016/j.celrep.2015.02.001

Feil, R., & Berger, F. (2007). Convergent evolution of genomic imprinting in plants and mammals. *Trends in Genetics*, *23*(4), 192–199. http://doi.org/10.1016/j.tig.2007.02.004

Feinberg, A. P. (2010). Epigenomics reveals a functional genome anatomy and a new approach to common disease. *Nature Biotechnology*, *28*(10), 1049–52. http://doi.org/10.1038/nbt1010-1049

Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., … Jacobsen, S. E. (2010). Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(19), 8689–8694. http://doi.org/10.1073/pnas.1002720107

Feschotte, C., & Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu Rev Genet*, *41*, 331–368. http://doi.org/10.1126/scisignal.2001449.Engineering

Fisher, A. G. (2002). Cellular identity and lineage choice. *Nat Rev Immunol*, *2*(12), 977–982. http://doi.org/10.1038/nri958

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., … Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, *7*(6), 461–5. http://doi.org/10.1038/nmeth.1459

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., … Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine

residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(5), 1827–31. http://doi.org/10.1073/pnas.89.5.1827

Fu, Y., Luo, G.-Z., Chen, K., Deng, X., Yu, M., Han, D., … He, C. (2015). N6 - Methyldeoxyadenosine Marks Active Transcription Start Sites in Chlamydomonas. *Cell*, *161*, 879–892. http://doi.org/10.1016/j.cell.2015.04.010

Gama-Sosa, M. A., Midgett, R. M., Slagel, V. A., Githens, S., Kuo, K. C., Gehrke, C. W., & Ehrlich, M. (1983). Tissue-specific differences in DNA methylation in various mammals. *BBA - Gene Structure and Expression*, *740*(2), 212–219. http://doi.org/10.1016/0167-4781(83)90079-9

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, *31*(13), 3784–3788. http://doi.org/10.1093/nar/gkg563

Glastad, K. M., Hunt, B. G., Yi, S. V., & Goodisman, M. A. D. (2011). DNA methylation in insects: On the brink of the epigenomic era. *Insect Molecular Biology*, *20*(5), 553–565. http://doi.org/10.1111/j.1365-2583.2011.01092.x

Globisch, D., Münzel, M., Müller, M., Michalakis, S., Wagner, M., Koch, S., … Carell, T. (2010). Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS ONE*, *5*(12), 1–9. http://doi.org/10.1371/journal.pone.0015367

Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: A Landscape Takes Shape. *Cell*, *128*(4), 635–638. http://doi.org/10.1016/j.cell.2007.02.006

Goll, M. G., & Bestor, T. H. (2005). Eukaryotic Cytosine Methyltransferases. *Annual Review of Biochemistry*, *74*(1), 481–514. http://doi.org/10.1146/annurev.biochem.74.010904.153721

Goll, M. G., Kirpekar, F., Maggert, K. a, Yoder, J. a, Hsieh, C.-L., Zhang, X., … Bestor, T. H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science*, *311*(5759), 395–8. http://doi.org/10.1126/science.1120976

Gravina, S., Dong, X., Yu, B., Vijg, J., Gravina, S., Ganapathi, S., … Andrews, S. (2016). Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome. *Genome Biology*, *17*(1), 150. http://doi.org/10.1186/s13059-016-1011-3

Greer, E. L., Blanco, M. A., Gu, L., Sendinc, E., Liu, J., Aristizabal-Corrales, D., … Shi, Y. (2015). DNA Methylation on N6-Adenine in C . elegans. *Cell*, *161*, 868–878.

http://doi.org/10.1016/j.cell.2015.04.005

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - A baiting and iterative mapping approach. *Nucleic Acids Research*, *41*(13), 1–9. http://doi.org/10.1093/nar/gkt371

Hayatsu, H., Wataya, Y., Kai, K., & Ida, S. (1970). Reaction of sodium bisulfite with uracil, cytsoine and their derivatives. *Biochemistry*, *9*(14), 2858–2865.

He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., … Xu, G.-L. (2011). Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science*, *333*(September), 1303–1308.

Heard, E., & Ditsteche, C. M. (2006). Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes & Development*, *20*, 1848–1867.

Heimer, K. (2010, August 18). Invasion of self-cloning crayfish alarms Madagascar. *Deutsche Presse-Agentur Wire Story*. Hamburg. Retrieved from http://www.earthtimes.org/articles/news/339974,alarms-madagascar-feature.html

Holdich, D., & Pöckl, M. (2007). Invasive crustaceans in European inland waters. In F. Gherardi (Ed.), *Freshwater bioinvaders: profiles, distribution, and threats* (pp. 29–75). The Netherlands: Springer.

Hon, G. C., Rajagopal, N., Shen, Y., McCleary, D. F., Yue, F., Dang, M. D., & Ren, B. (2013). Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature Genetics*, *45*(10), 1198–206. http://doi.org/10.1038/ng.2746

Hong, E. E., Okitsu, C. Y., Smith, A. D., & Hsieh, C.-L. (2013). Regionally specific and genome-wide analyses conclusively demonstrate the absence of CpG methylation in human mitochondrial DNA. *Molecular and Cellular Biology*, *33*(14), 2683–90. http://doi.org/10.1128/MCB.00220-13

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*(10), R115. http://doi.org/10.1186/gb-2013-14-10-r115

Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucelosides by paper chromatography. *Journal of Biological Chemistry*, *175*(1), 315–332.

Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., … Fan, G. (2016). Simultaneous profiling of

transcriptome and DNA methylome from a single cell. *Genome Biology*, *17*(1), 88. http://doi.org/10.1186/s13059-016-0950-z

Huang, X., & Madan, a. (1999). CAP 3: A DNA sequence assembly program. *Genome Research*, *9*(906), 868–877. http://doi.org/10.1101/gr.9.9.868

Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., … Boysen, G. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (New York, N.Y.)*, *333*(6047), 1300–3. http://doi.org/10.1126/science.1210597

Iwasaki, Y., Nishiki, I., Nakamura, Y., Yasuike, M., Kai, W., Nomura, K., … Ototake, M. (2016). Effective de novo assembly of fish genome using haploid larvae. *Gene*, *576*(2), 644–649. http://doi.org/10.1016/j.gene.2015.10.015

Jones, J. P. G., Rasamy, J. R., Harvey, A., Toon, A., Oidtmann, B., Randrianarison, M. H., … Ravoahangimalala, O. R. (2009). The perfect invader: A parthenogenic crayfish poses a new threat to Madagascar's freshwater biodiversity. *Biological Invasions*, *11*(6), 1475–1482. http://doi.org/10.1007/s10530-008-9334-y

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews. Genetics*, *13*(7), 484–92. http://doi.org/10.1038/nrg3230

Jurkowska, R. Z., Jurkowski, T. P., & Jeltsch, A. (2011). Structure and function of mammalian DNA methyltransferases. *Chembiochem : A European Journal of Chemical Biology*, *12*(2), 206–22. http://doi.org/10.1002/cbic.201000195

Jurkowski, T. P., Meusburger, M., Phalke, S., Helm, M., Nellen, W., Reuter, G., & Jeltsch, A. (2008). Human DNMT2 methylates tRNA Asp molecules using a DNA methyltransferase-like catalytic mechanism. *Rna*, *14*(1987), 1663–1670. http://doi.org/10.1261/rna.970408.transferase

Kanehisa, M. (1996). (KEGG) Toward pathway engineering: A new database of genetic and molecular pathways. *Science & Technology Japan*, *59*, 34–38. Retrieved from http://www.kanehisa.jp/docs/archive/stj.pdf

Kao, D., Lai, A. G., Stamataki, E., Rosic, S., Konstantinides, N., Jarvis, E., … Aboobaker, A. (2016). The genome of the crustacean *Parhyale hawaiensis* : a model for animal development, regeneration, immunity and lignocellulose digestion. *bioRxiv*, 65789. http://doi.org/10.1101/065789

Kawai, T., & Takahata, M. (Eds.). (2010). *Biology of Crayfish*. Sapporo, Japan: Hokkaido University Press.

Kriaucionis, S., & Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, *324*(5929), 929–930. http://doi.org/10.1126/science.1169786

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., … Ziegler, S. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. http://doi.org/10.1038/nature14248

Kuo, K. C., McCune, R. A., & Gehrke, C. W. (1980). Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA. *Nucleic Acids Research*, *8*(20), 4763–4778.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357–359. http://doi.org/10.1038/nmeth.1923

Laszlo, A. H., Derrington, I. M., Brinkerhoff, H., Langford, K. W., Nova, I. C., Samson, J. M., … Gundlach, J. H. (2013). Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(47), 18904–9. http://doi.org/10.1073/pnas.1310240110

Laura-Jayne Gardiner, Mark Quinton-Tulloch, Lisa Olohan, Jonathan Price, Neil Hall, A. H. (2015). A genome-wide survey of DNA methylation in hexaploid wheat. *Genome Biology*, *16*(December), 273. http://doi.org/10.1186/s13059-015-0838-3

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, *3*, 1–8. http://doi.org/10.1016/j.bdq.2015.02.001

Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews. Genetics*, *11*(3), 204–220. http://doi.org/10.1038/nrg2719

Li, A., Hu, B. Q., Xue, Z. Y., Chen, L., Wang, W. X., Song, W. Q., … Wang, C. G. (2011). DNA Methylation in Genomes of Several Annual Herbaceous and Woody Perennial Plants of Varying Ploidy as Detected by MSAP. *Plant Molecular Biology Reporter*, *29*(4), 784–793.

http://doi.org/10.1007/s11105-010-0280-3

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. http://doi.org/10.1186/1471-2105-12-323

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2009). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, *26*(4), 493–500. http://doi.org/10.1093/bioinformatics/btp692

Li, C., Weng, S., Chen, Y., Yu, X., Lü, L., Zhang, H., … Xu, X. (2012). Analysis of Litopenaeus vannamei Transcriptome Using the Next-Generation DNA Sequencing Technique. *PLoS ONE*, *7*(10). http://doi.org/10.1371/journal.pone.0047442

Li, E., Bestor, T. H., & Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, *69*(6), 915–926. http://doi.org/10.1016/0092-8674(92)90611-F

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. http://doi.org/10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. http://doi.org/10.1093/bioinformatics/btp352

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. http://doi.org/10.1093/bioinformatics/btl158

Liptak, B., Mrugala, A., Pekarik, L., Mutkovic, A., Grula, D., Petrusek, A., & Kouba, A. (2016). Expansion of the marbled crayfish in Slovakia: beginning of an invasion in the Danube catchment? *Journal of Limnology*, *75*(2), 18. http://doi.org/10.4081/jlimnol.2016.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., … Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–22. http://doi.org/10.1038/nature08514

Liu, B., Du, Q., Chen, L., Fu, G., Li, S., Fu, L., … Bin, C. (2016). CpG methylation patterns of human mitochondrial DNA. *Scientific Reports*, *6*(23421), 23421.

http://doi.org/10.1038/srep23421

Lőkkös, A., Müller, T., Kovács, K., Várkonyi, L., Specziár, A., & Martin, P. (2016). The alien, parthenogenetic marbled crayfish (Decapoda: Cambaridae) is entering Kis-Balaton (Hungary), one of Europe's most important wetland biotopes. *Knowledge and Management of Aquatic Ecosystems*, (417), 16. http://doi.org/10.1051/kmae/2016003

Luo, G.-Z., Blanco, M. A., Greer, E. L., He, C., & Shi, Y. (2015). DNA N6-methyladenine: a new epigenetic mark in eukaryotes? *Nature Reviews, Molecular Cell Biology*, *16*(12), 705–710. http://doi.org/10.1038/nrm4076

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., & Maleszka, R. (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. http://doi.org/10.1016/j.tig.2011.01.003

Manfrin, C., Tom, M., de Moro, G., Gerdol, M., Guarnaccia, C., Mosco, A., … Giulianini, P. G. (2013). Application of D-Crustacean Hyperglycemic Hormone Induces Peptidases Transcription and Suppresses Glycolysis-Related Transcripts in the Hepatopancreas of the Crayfish Pontastacus leptodactylus - Results of a Transcriptomic Study. *PLoS ONE*, *8*(6). http://doi.org/10.1371/journal.pone.0065176

Marchler-Bauer, A., & Bryant, S. H. (2004). CD-Search: Protein domain annotations on the fly. *Nucleic Acids Research*, *32*(WEB SERVER ISS.), 327–331. http://doi.org/10.1093/nar/gkh454

Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y., & Bryant, S. H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Research*, *30*(1), 281–3. http://doi.org/10.1093/nar/30.1.281

Martienssen, R. a., & Colot, V. (2001). DNA Methylation and Epigenetic Inheritance in Plants and Filamentous Fungi. *Science*, *293*(5532), 1070–1074. http://doi.org/10.1126/science.293.5532.1070

Martin, P., Dorn, N. J., Kawai, T., van der Heiden, C., & Scholtz, G. (2010). The enigmatic Marmorkrebs (marbled crayfish) is the parthenogenetic form of Procambarus fallax (Hagen,

1870). *Contributions to Zoology*, *79*(3), 107–118.

Martin, P., Kohlmann, K., & Scholtz, G. (2007). The parthenogenetic Marmorkrebs (marbled crayfish) produces genetically uniform offspring. *Naturwissenschaften*, *94*(10), 843–846. http://doi.org/10.1007/s00114-007-0260-0

Martin, P., Thonagel, S., & Scholtz, G. (2016). The parthenogenetic Marmorkrebs (Malacostraca: Decapoda: Cambaridae) is a triploid organism. *Journal of Zoological Systematics and Evolutionary Research*, *54*(1), 13–21. http://doi.org/10.1111/jzs.12114

Marzano, F. N., Scalici, M., Chiesa, S., Gherardi, F., Piccinini, A., & Gibertini, G. (2009). The first record of the marbled crayfish adds further threats to fresh waters in Italy. *Aquatic Invasions*, *4*(2), 401–404. http://doi.org/10.3391/ai.2009.4.2.19

Maxam, a M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(2), 560–4. http://doi.org/10.1073/pnas.74.2.560

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., & Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, *33*(18), 5868–5877. http://doi.org/10.1093/nar/gki901

Myr, E. (2000). The Biological Species Concept. In Q. D. Wheeler & R. Maier (Eds.), *Species Concepts and Phylogenetic Theory* (pp. 17–29). New York: Columbia University Press.

Neiman, M., Jokela, J., & Lively, C. M. (2005). Variation in asexual lineage age in Potamopyrgus antipodarum, a New Zealand snail. *Evolution; International Journal of Organic Evolution*, *59*(9), 1945–1952. http://doi.org/10.1111/j.0014-3820.2005.tb01064.x

Ning, L., Liu, G., Li, G., Hou, Y., Tong, Y., & He, J. (2014). Current Challenges in the Bioinformatics of Single Cell Genomics. *Frontiers in Oncology*, *4*(January), 7. http://doi.org/10.3389/fonc.2014.00007

Novitsky, R. A., & Son, M. O. (2016). The first records of Marmorkrebs [Procambarus fallax (Hagen , 1870) f . virginalis] (Crustacea , Decapoda , Cambaridae) in Ukraine. *Ecologica Montenegrina*, *5*(Churcholl 2014), 44–46.

Picardi, E., & Pesole, G. (2010). Computational Methods for Ab Initio and Comparative Gene Finding. In O. Carugo & F. Eisenhaber (Eds.), *Data Mining Techniques for the Life*

Sciences (Vol. 609, pp. 269–284). Humana Press. http://doi.org/10.1007/978-1-60327-241-4

R Core Development Team. (2013). R: A language and environment for statistical computing. *R Found Stat Comput*, *1*.

Raddatz, G., Guzzardo, P. M., Olova, N., Fantappié, M. R., Rampp, M., Schaefer, M., … Lyko, F. (2013). Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(21), 8627–31. http://doi.org/10.1073/pnas.1306723110

Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, a P., & Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(10), 5237–5242. http://doi.org/10.1073/pnas.97.10.5237

Ratel, D., Ravanat, J., Berger, F., & Wion, D. (2006). N6-methyladenine : the other methylated base of DNA. *BioEssays*, *28*(3), 309–315. http://doi.org/10.1002/bies.20342

Razin, A., & Cedar, H. (1977). Distribution of 5-methylcytosine in chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(7), 2725–2728. http://doi.org/10.1073/pnas.74.7.2725

Razin, A., & Sedate, J. (1977). Analysis of 5-Methylcytosine in DNA. *Analytical Biochemistry*, *77*, 370–377.

Richards, S., Gibbs, R. a, Weinstock, G. M., Brown, S. J., Denell, R., Beeman, R. W., & Gibbs, R. (2008). The genome of the model beetle and pest Tribolium castaneum. *Nature*, *452*(7190), 949–55. http://doi.org/10.1038/nature06784

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441–448. http://doi.org/10.1016/0022-2836(75)90213-2

Sarda, S., Zeng, J., Hunt, B. G., & Yi, S. V. (2012). The evolution of invertebrate gene body methylation. *Molecular Biology and Evolution*, *29*(8), 1907–1916. http://doi.org/10.1093/molbev/mss062

Schaefer, M., Pollex, T., Hanna, K., Tuorto, F., Meusburger, M., Helm, M., & Lyko, F. (2010). RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes*

and Development, *24*(15), 1590–1595. http://doi.org/10.1101/gad.586710

Scholtz, G., Braband, A., Tolley, L., Reimann, A., Mittmann, B., Lukhaup, C., … Vogt, G. (2003). Ecology: Parthenogenesis in an outsider crayfish. *Nature*, *421*(6925), 806. http://doi.org/10.1038/421806a

Schübeler, D. (2015). Function and information content of DNA methylation. *Nature*, *517*(7534), 321–326. http://doi.org/10.1038/nature14192

Shen, J. C., Rideout, W. M., & Jones, P. A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Research*, *22*(6), 972–6. http://doi.org/10.1093/nar/22.6.972

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. http://doi.org/10.1093/bioinformatics/btv351

Simon, J. C., Delmotte, F., Rispe, C., & Crease, T. (2003). Phylogenetic relationships between parthenogens and their sexual relatives: The possible routes to parthenogenesis in animals. *Biological Journal of the Linnean Society*, *79*(1), 151–163. http://doi.org/10.1046/j.1095-8312.2003.00175.x

Singer, M., Kosti, I., Pachter, L., & Mandel-Gutfreund, Y. (2015). A diverse epigenetic landscape at human exons with implication for expression. *Nucleic Acids Research*, *43*(7), 3498–3508. http://doi.org/10.1093/nar/gkv153

Smallwood, S. a, Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., … Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, *11*(8), 817–20. http://doi.org/10.1038/nmeth.3035

Smit, A., Hubley, R., & Green, P. (2013). RepeatMasker Open-4.0.

Smith, Z. D., & Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nature Reviews. Genetics*, *14*(3), 204–20. http://doi.org/10.1038/nrg3354

Soderlund, C., Nelson, W., Willer, M., & Gang, D. R. (2013). TCW: Transcriptome Computational Workbench. *PLoS ONE*, *8*(7), 1–10. http://doi.org/10.1371/journal.pone.0069401

Song, J., Rechkoblit, O., Bestor, T. H., & Patel, D. J. (2011). Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science (New York,*

N.Y.), *331*(6020), 1036–40. http://doi.org/10.1126/science.1195380

Sorek, R. (2007). The birth of new exons: mechanisms and evolutionary consequences. *RNA (New York, N.Y.)*, *13*(10), 1603–8. http://doi.org/10.1261/rna.682507

Suzuki, M. M., Kerr, A. R. W., De Sousa, D., & Bird, A. (2007). CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Research*, *17*(5), 625–631. http://doi.org/10.1101/gr.6163007

Suzuki, M. M., Yoshinari, A., Obara, M., Takuno, S., Shigenobu, S., Sasakura, Y., … Nakayama, A. (2013). Identical sets of methylated and nonmethylated genes in Ciona intestinalis sperm and muscle cells. *Epigenetics & Chromatin*, *6*(1), 38. http://doi.org/10.1186/1756-8935-6-38

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., … Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (New York, N.Y.)*, *324*(5929), 930–5. http://doi.org/10.1126/science.1170116

Takuno, S., & Gaut, B. S. (2012). Body-methylated genes in arabidopsis thaliana are functionally important and evolve slowly. *Molecular Biology and Evolution*, *29*(1), 219–227. http://doi.org/10.1093/molbev/msr188

Tatusov, R. L., Koonin, E. V, & Lipman, D. J. (2012). A RTICLES A Genomic Perspective on Protein Families. *Science*, *278*(1997), 631–637. http://doi.org/10.1126/science.278.5338.631

Tenlen, J. R., Smith, F. W., Wang, J. R., Kiera, A., Nishimura, E. O., Tintori, S. C., … Osborne, E. (2016). Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences*, *113*(36), E5364–E5364. http://doi.org/10.1073/pnas.1613046113

Tessarz, P., & Kouzarides, T. (2014). Histone core modifications regulating nucleosome structure and dynamics. *Nature Reviews, Molecular Cell Biology*, *15*(11), 703–708. http://doi.org/10.1038/nrm3890

Tomizawa, S., Kobayashi, H., Watanabe, T., Andrews, S., Hata, K., Kelsey, G., & Sasaki, H. (2011). Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. *Development (Cambridge, England)*, *138*(5), 811–20. http://doi.org/10.1242/dev.061416

Tost, J., Dunker, J., & Gut, I. G. (2003). Analysis and quantification of multiple methylation variable positions in CpG islands by Pyrosequencing. *Biotechniques*, *35*(1), 152–156.

Vanyushin, B. F., Tkacheva, S. G., & Belozersky, A. N. (1970). Rare Bases in Animal DNA. *Nature*, *225*(5236), 948–949. JOUR. Retrieved from http://dx.doi.org/10.1038/225948a0

Vogt, G. (2008). How to minimize formation and growth of tumours: Potential benefits of decapod crustaceans for cancer research. *International Journal of Cancer*, *123*(12), 2727–2734.

Vogt, G. (2008). The marbled crayfish: A new model organism for research on development, epigenetics and evolutionary biology. *Journal of Zoology*, *276*(1), 1–13. http://doi.org/10.1111/j.1469-7998.2008.00473.x

Vogt, G., Falckenhayn, C., Schrimpf, A., Schmid, K., Hanna, K., Panteleit, J., … Lyko, F. (2015). The marbled crayfish as a paradigm for saltational speciation by autopolyploidy and parthenogenesis in animals. *bioRxiv*, 25254. http://doi.org/10.1101/025254

Vogt, G., Huber, M., Thiemann, M., van den Boogaart, G., Schmitz, O. J., & Schubart, C. D. (2008). Production of different phenotypes from the same genotype in the same environment by developmental variation. *The Journal of Experimental Biology*, *211*(Pt 4), 510–23. http://doi.org/10.1242/jeb.008755

Vogt, G., Tolley, L., & Scholtz, G. (2004). Life stages and reproductive components of the marmorkrebs (marbled crayfish), the first parthenogenetic decapod Crustacean. *Journal of Morphology*, *261*(3), 286–311. http://doi.org/10.1002/jmor.10250

Volff, J. N. (2006). Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays*, *28*(9), 913–922. http://doi.org/10.1002/bies.20452

Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., Zhao, D., … Kang, L. (2014). The locust genome provides insight into swarm formation and long-distance flight. *Nature Communications*, *5*, 2957. http://doi.org/10.1038/ncomms3957

Wang, X., Li, Q., Lian, J., Li, L., Jin, L., Cai, H., … Zhang, G. (2014). Genome-wide and single-base resolution DNA methylomes of the Pacific oyster Crassostrea gigas provide insight into the evolution of invertebrate CpG methylation. *BMC Genomics*, *15*, 1119. http://doi.org/10.1186/1471-2164-15-1119

Wang, X., Wheeler, D., Avery, A., Rago, A., Choi, J. H., Colbourne, J. K., … Werren, J. H. (2013). Function and Evolution of DNA Methylation in Nasonia vitripennis. *PLoS Genetics*, *9*(10). http://doi.org/10.1371/journal.pgen.1003872

Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M., & Kriventseva, E. V. (2013). OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research*, *41*(D1), 358–365. http://doi.org/10.1093/nar/gks1116

Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L., & Schübeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, *37*(8), 853–62. http://doi.org/10.1038/ng1598

Werren, J. H., Richards, S., Desjardins, C. A., Niehuis, O., Gadau, J., Colbourne, J. K., & The Nasonia Genome Working Group. (2010). Functional and Evolutionary Insights from the Genomes of Three Parasitoid Nasonia Species. *Science*, *327*(January), 343–349.

Wheeler, Q. D., & Maier, R. (Eds.). (2000). *Species Concepts and Phylogenetic Theory*. New York: Columbia University Press.

World Weather Online. (2012a). Antananarivo Monthly Climate Average, Madagascar. Retrieved September 12, 2016, from http://www.worldweatheronline.com/v2/weather-averages.aspx?q=Antananarivo, Madagascar

World Weather Online. (2012b). Stockholm Monthly Climate Average, Sweden. Retrieved September 12, 2016, from http://www.worldweatheronline.com/stockholm-weather-averages/stockholms-lan/se.aspx

Wyatt, G. R. (1950). Recognition and Estimation of 5-Methylcytosine in Nucleic Acids. *Journal of Biochemistry*, *48*, 581–584.

Xi, Y., & Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, *10*(1), 232. http://doi.org/10.1186/1471-2105-10-232

Xiang, H., Zhu, J., Chen, Q., Dai, F., Li, X., Li, M., … Wang, J. (2010). Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nature Biotechnology*, *28*(5), 516-U181. http://doi.org/10.1038/nbt0710-756d

Xiao, J., Song, C., Liu, S., Tao, M., Hu, J., Wang, J., … Liu, Y. (2013). DNA Methylation Analysis of Allotetraploid Hybrids of Red Crucian Carp (Carassius auratus red var.) and Common

Carp (Cyprinus carpio L.). *PLoS ONE*, *8*(2). http://doi.org/10.1371/journal.pone.0056409

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., … Wang, J. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660–1666. http://doi.org/10.1093/bioinformatics/btu077

Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., … Chen, S. (2012). FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS ONE*, *7*(12), 1–6. http://doi.org/10.1371/journal.pone.0052249

Yang, Y., & Smith, S. a. (2013). Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*, *14*(1), 328. http://doi.org/10.1186/1471-2164-14-328

Yi, S. V, & Goodisman, M. A. D. (2009). methylation in animals Methylation in Animal Genomes, 551–556.

Zemach, A., McDaniel, I. E., Silva, P., & Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, *328*(5980), 916–919. http://doi.org/10.1126/science.1186366

Zemach, A., & Zilberman, D. (2010). Evolution of eukaryotic DNA methylation and the pursuit of safer sex. *Current Biology*, *20*(17), R780–R785. http://doi.org/10.1016/j.cub.2010.07.007

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. http://doi.org/10.1101/gr.074492.107

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., … Chen, D. (2015). N^6-methyladenine DNA modification in Drosophila. *Cell*, *161*(4), 893–906. http://doi.org/10.1016/j.cell.2015.04.018

Zhang, H.-Y., Zhao, H.-X., Wu, S.-H., Huang, F., Wu, K.-T., Zeng, X.-F., … Wu, X.-J. (2016). Global Methylation Patterns and Their Relationship with Gene Expression and Small RNA in Rice Lines with Different Ploidy. *Frontiers in Plant Science*, *7*(July), 1002. http://doi.org/10.3389/fpls.2016.01002

Zhang, J., Liu, Y., Xia, E.-H., Yao, Q.-Y., Liu, X.-D., & Gao, L.-Z. (2015). Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression. *Proceedings of the National Academy of Sciences*, 201515170. http://doi.org/10.1073/pnas.1515170112

Zhou, H., Ma, T.-Y., Zhang, R., Xu, Q.-Z., Shen, F., Qin, Y.-J., … Li, Y.-J. (2016). Analysis of Different Ploidy and Parent–Offspring Genomic DNA Methylation in the Loach Misgurnus anguillicaudatus. *International Journal of Molecular Sciences*, *17*(8), 1299. http://doi.org/10.3390/ijms17081299

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., & Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics*, *39*(1), 61–9. http://doi.org/10.1038/ng1929

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., … Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*, *500*(7463), 477–81. http://doi.org/10.1038/nature12433

Acknowledgment

My sincere thanks go to Prof. Dr. Frank Lyko, who did not only gave me the great chance to conduct my PhD thesis in his lab about this interesting topic but, also to bring forward my own ideas about my wish model organism and who head always confidence in my accomplishment.

A special thanks to the former and current co-workers of the Lyko lab, who gave me the feeling to be part of the big department family and thus, making the life in the exile easier for me. Furthermore, I want thank the "Coffee Breakers" for the cheerful times and the warm discussions about life.

I would like to acknowledge and extend my heartfelt gratitude to the following researchers and co-workers of the Lyko lab, who shared their personal experience with me. Günter, Julian and Fanny for helping advice in bioinformatics approaches and discussions of my results, especially Julian and Fanny for the hilarious time in our office. Katharina and Tanja for encouraging and helping me through my horrible time when nothing seemed to work, especially Katharina who did a lot of maintenance work for the marbled crayfish population in the lab.

I want also thank Prof. Dr. Ana Martin-Villalba for being my second referee and to all my members of my thesis advisory committee for their time and suggestions.

Above all, I am most grateful to my family, friends and my husband Julian, words alone cannot express what I owe them for their encouragement and whose patient love enabled me to complete this PhD thesis. Specifically my husband, who may went through an even tougher time because of me.

And a special thanks to my parents, who made all things possible.