Supplementary Material and Methods to Chapter 4

Genome and Transcriptome Sequencing

We sequenced whole body transcriptomes and the genomes of *Ctenocephalides felis*, *Liposcelis bostrychophila*, and *Panorpa germanica* using HiSeq2000 and HiSeq4000 sequencing technology. Genomic DNA and mRNA were extracted from the samples summarized in table 1. DNA and mRNA extraction, DNA shearing, DNA size-selection, library preparation, DNA sequencing, and read demultiplexing were done by BGI-Shenzhen following Illumina standard protocols for Illumina sequencers and proprietary in-house tools at BGI.

We used a k-mer approach to estimate the approximate size of the three genomes by analyzing the 17-mer frequency distribution in the unprocessed raw reads of all libraries with the software jellyfish version 2.2.10 (Marcais & Kingsford 2011). The specific setting for running jellyfish were: jellyfish count -t 16 -C -m 17 -s 4G -o 17mer_out --min-qual-char=?. The estimated genome sizes are given in table 2. The genome sizes served as baseline to calculate the read coverage of the various libraries.

Transcriptome Assembly

Raw RNAseq reads were screened for adapters and trimmed using Cutadapt (version 1.18; Martin 2011) as implemented in Trim Galore (version 0.5.0; https://github.com/FelixKrueger/TrimGalore). Reads were trimmed requiring a minimum Phred score of 25 and retaining only reads that have a minimum length of 75 bp after trimming. The trimmed reads were *de-novo* assembled using Trinity (version 2.8.3; Grabherr et al. 2011) under default settings. The default settings include an *in-silico* normalization of the reads to remove highly redundant sequences before the assembly process.

Genome Assembly

Raw genomic reads were screened for adapters and low quality bases and trimmed using trimmomatic version 0.33 (Bolger et al. 2014) using the following parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:25 MINLEN:50. We performed an internal assemblathon comparing the results of two different assemblers – Platanus 1.2.4 (Kajitani et al. 2014) and Allpaths-LG version 52488 (Gnerre et al. 2011). How sequencing libraries are combined in the three steps of the assembly with (contig-assembly, scaffolding, and gap-closing) has an impact on the obtained result. We thus deliberately tested 21 different set-ups for Platanus in order to identify the best combination (see table 3). Since no 8 kbp mate-pair libraries were sequenced for *C. felis*, only settings 15-20 (see table 3) were used for the assemblies of its genome.

For Allpaths-LG, we combined all four libraries in a single assembly. To this end, we defined the 250 bp PE library as *fragment* library, the 800 bp PE library as *jumping* library with an *inward* read orientation, and the 3 kbp and 8 kbp mate-pair libraries as *jumping* libraries with an *outward* read orientation. The estimated standard deviation of the fragments sizes (for the *fragment* library) and insert sizes (for the *jumping* library) were set to ca. 10 % of the average number of bases in the fragments (20 bp for the *fragment* library) and the average number of bases in the inserts (80 bp, 300 bp, and 800 bp for the 800 bp, 3 kbp, and 8 kbp *jumping* libraries, respectively).

We used QUAST 3.2 (Gurevich et al. 2013) to characterize all 22 assemblies, adding contigs to the comparison (-s option). Furthermore, we mapped 5 million randomly selected read pairs from the 800 bp PE library on the assemblies, using the short read mapper provided within the CLC Genomics Workbench 12.0 (QIAGEN). The fraction of mapped reads and completely mapped read-pairs was used in the evaluation of assemblies (described below).

Based on the estimated genome size, the metrics calculated by QUAST, and the read mapping statistics we calculated a score to select the best assembly based on the following properties:

1) The deviation of the total assembled genome size to the estimated genome size.

A reasonable good assembly contains few but long continuous sequences. Therefore, it is preferable to have large proportions of the assembled bases in long contigs. We calculate the score $s_{total\ length}$ as follows:

$s_{total\ length}$ = 1 - (|”estimated genome size” - “Total number of bp in contigs >= 0 bp”| +
|”estimated genome size” - “Total number of bp in contigs >= 1000 bp”| +
|”estimated genome size” - “Total number of bp in contigs >= 5000 bp”| +
|”estimated genome size” - “Total number of bp in contigs >= 10000 bp”| +
|”estimated genome size” - “Total number of bp in contigs >= 25000 bp”| +
|”estimated genome size” - “Total number of bp in contigs >= 50000 bp” |) / (6 * ”estimated genome size”)

If the assembled genome size is in proximity to the estimated genome size, the result is a score between 1/6 (all bases are assembled in contigs < 1000 bp) and 1 (all bases are assembled in a single contig that has the length of the estimated genome size). If the assembled genome size is multiple times larger than the estimated genomes size, the resulting score will be negative.

2) The scaffold N50, corrected by the number of Ns in the scaffolds.

s_{N50} = log(“scaffold N50” - ((“scaffold N50” / 100.000) * “scaffold # N’s per 100.000 bp”))

3) The fraction of identified BUSCO genes.

s_{BUSCO} = (*complete BUSCO genes* + *fragmented BUSCO genes*) / *total number of BUSCO genes*

where the *total number of BUSCO genes* =2675.

4) The fraction of mapped reads and completely mapped read pairs.

$s_{mapping}$ = *fraction of mapped reads * fraction of mapped pairs*

The final score to judge the assemblies is then obtained by multiplying all four property scores:

$s_{final} = s_{total\ length} * s_{N50} * s_{BUSCO} * s_{mapping}$

We chose the assembly with the highest s_{final} score to be the final genome assembly to work with.

Platanus was not able to assemble reasonable genomes for *Panorpa germanica* and *Ctenocephalides felis*, respectively: No BUSCO gene could be identified in each of the 21 assemblies for *C. felis* and only 1 BUSCO gene was found in each of the 21 assemblies for *P. germanica*. We thus selected the genome assembled by Allpaths-LG as the best assembly for these two species (see table 4). Although, Platanus successfully assembled genomes for *Liposcelis bostrychophila* with for all 21 settings, Allpaths-LG also produced a superior genome assembly, as judged by the number of successfully identified BUSCO genes, the number of assembled base pairs and the number of contigs they are assembled in, and the N50 value. Consequently, the s_{final} score identified the Allpaths-LG assembly as the best genome assembly for *L. bostrychophila.* The best genome assembly for *S. ovinae* was generated by Platanus, using setting 3 (see table 3).

Annotation of Protein-coding Genes and Repeats

Protein-coding genes for *L. bostrychophila*, *S. ovinae*, *C. felis*, and were *de novo* annotated within the repeat-masked genome assemblies using MAKER MPI (version 2.31.10; Cantarel et al. 2008) following the iterative approach outlined by D. Card (https://gist.github.com/darencard/bb1001ac1532dd4225b030cf0cd61ce2).

In the initial round of gene annotation, genes were predicted based on the assembled transcriptomes of each species and additional external evidence consisting of high quality proteomes of 12 insect species as well as the complete Swiss Prot database (see table 6). Following the initial gene prediction, we conducted three iterative rounds using the *ab initio* gene predictors AUGUSTUS (version 3.3.2; Stanke et al. 2008) and SNAP (version 2.42.1; Korf 2004), which were trained using the gene models from the previous round. SNAP was trained using only gene models with an annotation edit distance score of 0.25 or better and a length of at least 50 amino acids. AUGUSTUS was trained using the BUSCO pipeline (version 3.1.0; Simão et al. 2015), the Edopterygota set of conserved genes, and the initial HMM models of *D. melanogaster* (options: -sp fly -l endopterygota_odb9). We extracted all gene models identified in the previous round that were supported by RNAseq evidence, including 1,000 bp of upstream and downstream flanking regions, as input for BUSCO.

We annotated transposable elements and other repeats in the genome assemblies of all 46 species included in this project following the pipeline exactly as described by Petersen et al. 2019. We used updated software versions for RepeatModeler Open-1.0.10 (Smit and Hubley 2008-2015), RepeatMasker 4.0.7 (Smit et al. 2013-2015), and RepBase version 20140131 (Jurka et al. 2005).

References

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M. 2008. MAKER : An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18:188–196.

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. PNAS 108:1513–1518.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075.

Jurka J, Kapitonov V, V, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005;110:462-467. doi: 10.1159/000084979

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, et al. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24:1384–1395.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10.

Marcais, G., Kingsford C. (2011): A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27, 764–770.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212.

Smit, AFA, Hubley, R. *RepeatModeler Open-1.0*. 2008-2015 <http://www.repeatmasker.org>.

Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*. 2013-2015 <http://www.repeatmasker.org>.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:637–644.

Petersen, M., Armisén, D., Gibbs, R.A. *et al.* Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evol Biol* **19,** 11 (2019). https://doi.org/10.1186/s12862-018-1324-9

Tables

Table 1: Samples used for gDNA and mRNA extraction and for NGS library preparation.

Species	Library	Origin	Sex	Number of samples	Tissue	Provider
C. felis	250 bp	Lab strain maintained at Dryden laboratory Kansas State University, Manhattan, KS, USA	female	72	whole body	Brian Wiegmann
C. felis	800 bp	Lab strain maintained at Dryden laboratory Kansas State University, Manhattan, KS, USA	male	180	whole body	Brian Wiegmann
C. felis	3 kbp	Lab strain maintained at Dryden laboratory Kansas State University, Manhattan, KS, USA	male and females´	360 males, 624 females	whole body	Brian Wiegmann
C. felis	100 bp (cDNA)	Lab strain maintained at Dryden laboratory Kansas State University, Manhattan, KS, USA	male	eight	whole body	Brian Wiegmann
C. felis	100 bp (cDNA)	Lab strain maintained at Dryden laboratory Kansas State University, Manhattan, KS, USA	female	eight	whole body	Brian Wiegmann
L. bostrychophila	250 bp, 800 bp, 3 kbp, 8 kbp	Lab strain maintained at Jodrell Laboratory Royal Botanic Gardens, Surrey, UK	female	ca. 2,500	whole body	Paul W. C. Green
L. bostrychophila	100 bp (cDNA)	Lab strain maintained at Jodrell Laboratory Royal Botanic Gardens, Surrey, UK	female	ca. 500	whole body	Paul W. C. Green

P. germanica	250 bp (also used for 800-bp library)	Germany, Rhineland-Palatinate, Albersweiler (June 8, 2014)	male	three	whole body	Oliver Niehuis
P. germanica	800 bp	Germany, Rhineland-Palatinate, Albersweiler (June 8, 2014; August 17, 2013)	male	13	whole body (3 samples), thorax (10 samples)	Oliver Niehuis
P. germanica	3 kbp (also used for 800-bp library	Germany, Rhineland Palatinate, Albersweiler (August 17, 2013)	male	ten	thorax	Oliver Niehuis
P. germanica	8 kbp	Germany, Rhineland Palatinate, Albersweiler (May 10, 2015; August 17, 2013)	male, female	six males, nine female	whole body (males), thorax (females)	Oliver Niehuis
P. germanica	100 bp (cDNA)	Germany, Rhineland Palatinate, Albersweiler (August 17, 2013)	male	two	whole body	Oliver Niehuis
P. germanica	100 bp (cDNA)	Germany, Rhineland Palatinate, Albersweiler (August 17, 2013)	female	two	whole body	Oliver Niehuis

Table 2: Number of reads and estimated coverage of the sequenced libraries of *Ctenocephalides felis*, *Liposcelis bostrychophila*, and *Panorpa germanica.*

Species	**Estimated genome size**	**Number of sequenced gDNA reads (estimated coverage)**				**Number of sequenced cDNA reads**	
		250[a]	**800**[b]	**3000**[b]	**8000**[b]	**males**[b]	**females**[b]
C. felis	~ 450 Mbp[c]	192.5 M[e] (~ 64x)	111.3 M[e] (~ 25x)	101.3 M[e] (~ 23x)	NA	104.7 M[f]	128.2 M[f]
L. bostrychophila	~ 258–515 Mbp[d]	194.2 M[e] (113–43x)	152.2 M[e] (59–30x)	66.8 M[e] (26–13x)	60.4 M[e] (23–12x)	NA	89.5 M[c]
P. germanica	~ 420 Mbp[c]	353.3 M[e] (~ 126x)	275.0 M[e] (~ 65x)	122.1 M[e] (~ 29 x)	173.1 M[e] (~ 41 x)	87.4 M[e]	90.4 M[e]

[a] 150-bp read length
[b] 100-bp read length
[c] +/- ~ 50 Mbp (no distinct peak in *k*-mer histogram, likely because of high degree of heterozygosity)
[d] k-mer histogram shows one prominent peak (515 Mbp) and an additional very weak second peak (258 Mbp). The first one is interoreted as representing sex chromosomes, the second one is interpreted as representing autosomes.
[d] HiSeq 2000
[e] HiSeq 4000

Table 3: Library combinations used for each of the three assembly steps of Platanus v. 1.2.4. 1 = 250 bp paired-end, 2 = 800 bp paired-end, 3 = 2 kbp mate-pair, 4 = 8 kbp mate-pair.

Setting	contig-assembly	scaffolding	gap-closing
0	1,2	1,2,3	1,2,3,4
1	1,2	1,2,3,4	1,2
2	1,2	1,2,3,4	1,2,3
3	1,2	1,2,3,4	1,2,3,4
4	1,2	1,2,3,4	2,3,4
5	1,2	2,3,4	1,2
6	1,2	2,3,4	1,2,3
7	1,2	2,3,4	1,2,3,4
8	1,2,3	1,2,3,4	1,2
9	1,2,3	1,2,3,4	1,2,3
10	1,2,3	1,2,3,4	1,2,3,4
11	1,2,3	2,3,4	1,2
12	1,2,3	2,3,4	1,2,3
13	1,2,3,4	2,3,4	1,2,3,4
14	1,2,3,4	2,3,4	1,2
15	1,2	2,3	1,2
16	1,2	2,3	1,2,3
17	1,2	1,2,3	2,3
18	1,2	1,2,3	1,2,3
19	1,2,3	2,3	1,2,3
20	1,2,3	1,2,3	1,2,3

Table 4: Assembler and BUSCO statistic of the best genome assembly for each species.

Species	Assembler	BUSCO, n=2675			
		Complete	Duplicated	Fragmented	Missing
C. felis	Allpaths-LG v. 52488	1,584 (59%)	123 (4.5%)	493 (18%)	598 (22%)
L. bostrychophila	Allpaths-LG v. 52488	2,156 (80%)	337 (12%)	367 (13%)	152 (5.6%)
P. germanica	Allpaths-LG v. 52488	1,293 (48%)	83 (3.1%)	591 (22%)	791 (29%)
S. ovinae	Platanus v1.2.4, setting 3	1,387 (51%)	45 (1.8%)	538 (19%)	757 (28%)

Table 5: Assembly statistics of the best genome assembly for each species calculated with QUAST and CLC Genomics Workbench. [1]scaffolds >= 0 bp, [2]scaffolds >= 500 bp, [3]no reads were mapped because only Allpaths-LG successfully produced an assembly.

	C. felis	***L. bostrychophila***	***P. germanica***	***S. ovinae***
estimated genome size [Mbp]	~ 450	~ 258	~ 420	~ 63
assembled genomes size[1] [bp]	477,127,871	395,601,244	304,488,381	43,580,081
assembled genomes size[2] [bp]	477,127,871	395,601,244	304,488,381	41,800,826
number of scaffolds[1]	32,101	32,152	67,529	13,321
number of scaffolds[2]	32,101	32,152	67,529	1,640
#N's per 100 kbp[2]	14,309.85	17,916.48	11,082.55	4,362.00
longest scaffold	856,730	1,802,688	345,957	1,138,786
N50[2]	56,317	166,666	7,763	354,095
L50[2]	2,022	595	9,337	34
mapped reads [%]	-[3]	83.01	-[3]	93.53
mapped read pairs [%]	-[3]	60.33	-[3]	62.16
GC (%)	29.56	33.79	31.05	32.87

Table 6: Proteomes used as external evidence for the annotation of protein-coding genes of *Panorpa germanica*, *Ctenocephalides felis*, and *Liposcelis bostrychophila*

Species	**Abbreviation**	**Order**	**Data source**	**Version**
Acyrthosiphon pisum	Apis	Hemiptera	AphidBase	NCBI 2.1
Tribolium castaneum	Tcas	Coleoptera	RefSeq	NCBI 102
Heliconius erato demophoon	Hera	Lepidoptera	LepBase	v1.0
Anoplophora glabripennis	Agla	Coleoptera	RefSeq	NCBI 101
Papilio polytes	Ppol	Lepidoptera	LepBase	v1.0
Apis mellifera	Amel	Hymenoptera	RefSeq	NCBI 104
Nasonia vitripennis	Nvit	Hymenoptera		OGS2
Drosophila melanogaster	Dmel	Diptera	RefSeq	NCBI Release 6
Aedes aegypti	Aaeg	Diptera	RefSeq	NCBI 101
Ctenocephalides felis	Cfel	Siphonaptera	RefSeq	NCBI 100
Zoothermopsis nevadensis	Znev	Isoptera	RefSeq	NCBI 100
Pediculus humanus	Phum	Psocodea	ensembl	PhumU2.2
Database				
Swiss prot		complete	uniprot.org	10.12.2018

Supplementary Figure Captions

Fig. 4.S1:Comparison of DNA methylation levels between holometabolous and hemimetabolous insects. Holometabolous insects display significantly higher average levels of A) global (Wilcoxon rank sum test, p-value= 0.00004454) and B) gene body (Wilcoxon rank sum test, p-value=0.00003332) methylation levels compared to hemimetabolous insects.

Fig. 4.S2: The gene body methylation patterns of all 46 species included in our analyses. Each point in the plot is a comparison of the average levels of DNA methylation between the exons and the introns of a single gene. Increasing density of data points on each plot is indicated by an enlargement of single data points and color alteration according to the legend in each figure.

Fig. 4.S3: Comparison of the levels of A) global and B) gene body DNA methylation among the four patterns of gene body methylation. EM insects display lower levels of both global and gene body DNA methylation compared to both BM and MM insects (Kruskal-Wallis *H* test, p-value global =0.000001056, p-value gene bodies=0.0000002894).

Fig. 4.S4: Comparison of repeat methylation levels between holometabolous and hemimetabolous insects. Hemimetabola show significantly higher levels of repeat methylation compared to Holometabola (Wilcoxon rank sum test p-value=0.0004609).

Fig. 4.S5: A) Levels of repeat methylation across insects. In many hemimetabolous species the avergae methylation levels of repeat sequences are similar or even higher than the genome average, whereas in Holometabola repeat sequences are universally hypomethylated. B) Gene bodies and repeat sequences show positive DNA methylation enrichment in Hemimetabola, whereas repeat sequences show negative DNA methylation enrichment in Holometabola.

Fig. 4.S6: Comparison of methylation levels between intragenic and intergenic repeat sequences across insects. Intragenic repeat sequences are consistently more highly methylated in all insect species with appreciable methylation levels.

Fig. 4.S7: Comparison of DNA methylation enrichment at intergenic and intragenic repeat sequences between hemimetabolous and holomeabolous insects. Both intragenic and intergenic repeat sequences showed positive and significantly higher DNA methylation enrichment in hemimetabolous insects compared to Holometabola (Wilcoxon rank sum test, p-value intragenic= 0.0103, p-value intergenic=0.04466).

Fig. 4.S8: A) Comparison of DNA methylation enrichment at intergenic and intragenic repeat sequences among insects with different gene body methylation patterns. BM insects showed positive and significantly higher DNA methylation enrichment at both intragenic and intergenic repeat sequences compared with insects with an EM or MM pattern, whereas in EM and MM insects intragenic repeats were hypomethylated (Kruskal-Wallis *H* test, p-value intergenic=0.01333, p-value intragenic=0.0003172). B) Scatterplot correlating gene body methylation levels with repeat methylation levels in EM (dark blue triangles), MM (light green squares), and BM (dark green points) insects. We identified a strong positive correlation between gene body methylation levels and repeat methylation levels (Spearman correlation coefficient=0.845, p-value<0.005).

Fig. 4.S9: A) Barplot comparing methylation levels between exons that contain repeat sequences and exons that are free of repeat sequences across all 46 insect species included in this study. B) The same comparison but between repeat-free and repeat-containing introns. C) Boxplots comparing methylation levels among repeat-free and repeat-containing exons and introns in EM species. Repeat-containing exons have marginally lower levels of DNA methylation than repeat-free introns (Wilcoxon signed-rank test, p-value=0.049). D) Boxplot comparison of the per species number of CG dinucleotides among repeat-free and repeat-containing exons and introns of EM species.

Fig. 4.S10: A) Comparison between the average methylation ratios of repeat-free and repeat-associated introns in each gene group of BM and MM species (Wilcoxon signed rank test BM: High gene body methylation group, p-value=0.2031; Low gene body methylation group, p-value=0.01953. MM: High gene body methylation group, p-

value=0.9453; Low gene body methylation group, p-value=0.007813). B) The same comparison as in A, but for exons (Wilcoxon signed rank test BM: High gene body methylation group, p-value=0.2031; Low gene body methylation group, p-value=0.4961. MM: High gene body methylation group, p-value=0.4609; Low gene body methylation group, p-value=0.07813). C) Comparison of the proportion of genic CG content of four different groups of introns between BM and MM insect species (Wilcoxon rank sum test: Highly methylated repeat-containing introns, p-value=0.0006355; Lowly methylated repeat-containing introns, p-value=0.01077) D) The same comparison as in C, but for exons.

Fig. 4.S11: Scatterplot correlating gene body methylation levels of a species with the proportion of introns that contain repeats for the same species. We identified a strong negative correlation between gene body methylation levels and the proportion of introns that contain repeats (Spearman correlation coefficient=-0.6104172 , p-value=0.000006657).

Fig. 4.S12: Reconstructed states of gene body methylation levels mapped onto internal edges and nodes of the insect tree using a color gradient. Warmer colors represent lower levels of DNA methylation.

Fig. 4.S13: Ancestral state reconstruction of insect gene body methylation patterns. BM: orange color; MM: Blue color; UM: Green color; EM: Pink color. Maximum likelihood methods for reconstructing ancestral states: MPPA (top left), JOINT (top right), MAP (bottom). For a more detailed view of the reconstruction, please view supplementary file X.

Supplementary Table Captions

Table 4.S1: Contains taxonomic information on all the species analyzed for chapter 4. Additionally, download links for methylomes that were mapped, for the reference genome assemblies, and for gene annotations are provided.

Table 4.S2: Contains average methylation ratios of all annotated genomic elements for all species analyzed. It also contains average methylation enrichment values for all species analyzed.

Table 4.S3: Contains the number of DNMT1 and DNMT3 copies identified for all species analyzed. Also, contains conversion rates only for species that a lambda phage spike-in was used as control.

Table 4.S4: Contains information on the mode of acquisition and the parts used for DNA extraction for the species used for whole genome bisulfite sequencing.