universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Epigenetic responses to cadmium pollution in *Lumbricus terrestris*

verfasst von / submitted by

Fabian Oberauer, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2024 / Vienna 2024

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 220
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Systems Biology
Betreut von / Supervisor:	Assoz. Prof. Dipl. -Ing. Dr. Ovidiu Paun

Acknowledgements

I would like to thank my supervisor Ovidiu Paun for the opportunity, talks and help along the full project. Also I would like to thank Maja Šrut for making the project possible, her inputs and help during the EM-seq. Thanks go to the whole Plant Ecological Genomics group for providing such an mentally stimulating and fun environment. I need to add special thanks to Florian for sharing part of his script and Katie who was the victim of most of my questions. My deepest thanks go to the University of Vienna, Veterinary University of Vienna, University of Innsbruck, the team behind the Life Science Computer Cluster. Additionally, I would like to thank PhD student Gerhard P. Aigner and master student Verena Pittl for their help with earthworm maintenance and sampling, and Vianey Paola Barrera Enriquez and Leanne Haggerty from the Darwin Tree of Life for providing pre-release access to the structural annotation of the *Lumbricus terrestris* genome.
Lastly I would like to thank my partner Nathi, for everything.

Funding

This work was funded by the Austrian Science Fund (FWF) projects M2262, P338350 and P35275.

Summary

The Anthropocene has led to extensive effects on ecosystems. In addition to climate change, soil pollution with heavy metals is a global phenomenon with major, not well understood effects on ecosystems. To better understand the long term effects of sub-lethal pollutant concentrations we use Cadmium (Cd), a highly toxic heavy metal, as an experimental pollutant for the ecologically important annelid *Lumbricus terrestris*, the common earthworm. We focus our analysis on epigenetic changes in DNA methylation in response to Cd exposure. We generated whole-genome libraries using the recently developed Enzymatic Methyl-sequencing (EM-seq), a methylation detecting method with less biases as bisulfite-treated sequencing (BS-seq). A bioinformatics analysis incorporating the methylation caller bismark and the R-package dss detected DNA methylation changes on a single base and region level. Differentially methylated regions (DMRs) have been intersected with gene models to explore candidate functional alterations around in the gene bodies or promoters. The analyses identified over 20,000 DMRs between Cd treated genomes and controls, which overlapped with over 10,000 gene regions. A subsequent GO term enrichment resulted in general metabolic terms, but a high Cd pollution condition (i.e., 25 mg/kg soil) resulted in enrichments for “stress response to metal ions” and “methylation” which have been absent for a low Cd treatment (i.e. 10 mg/kg soil). In this study we are able to advance our knowledge of the genomic distribution of DNA methylation in annelids and of the molecular effects heavy metals induce in *L. terrestris*, a key component of many soil communities.

Zusammenfassung

Das Anthropozän führte zu weitreichenden Effekten auf Ökosysteme wie den Klimawandel und die globale Bodenverschmutzung mittels Schwermetallen, welche gravierende, aber nicht komplett untersuchte Effekte verursachen. Um unser Verständnis über Langzeitfolgen von Schwermetallen im ökologisch wertvollen Annelid Lumbricus terrestris zu verbessern, benutzten wir Cadmium, ein giftiges Schwermetall, als experimentellen Schadstoff in nicht tödlichen Konzentrationen. Wir fokussierten uns auf resultierende epigenetische DNA Methylations Änderungen. Wir generierten whole-genome BS-libraries mittels Enzymatic Methyl-Sequenzierung (EM-seq), eine kürzlich entwickelte Technik, die im Vergleich zur herkömmlichen Bisulfit-behandelten Sequenzierung (BS-seq) weniger Bias aufweist. Die bioinformatische Analyse integrierte den Methylierungs-Caller bismark und das R-Paket dss, um DNA-Methylierungsveränderungen in Einzelbasen-Auflösung und Regionen im gesamten Genom zu erkennen. Differenziell methylierte Regionen (DMRs) wurden dann mit Genmodels verglichen, um Gene mit möglichen DNA-Methylierungsveränderungen in Genkörpern oder Promotoren zu identifizieren. Die Analyse fand 20,000 DMRs zwischen Kontrollen und den Cd behandelten Genomen, welche mit den Regionen von 10,000 Genen überlappen. Anschließende Analysen der Anreicherung von Begriffen aus der Gene Ontology (GO) zeigen allgemeine Stoffwechselbegriffe, während bei der hohen Konzentration von Cadmium (25mg/kg Erde) eine Anreicherung für "Stressreaktion auf Metallionen" und "Methylierung" zeigt, welche in niedrigeren Konzentrationen (10mg/kg Erde) fehlen. Diese Studie trägt dazu bei, unser Verständnis der DNA-Methylierung bei Anneliden und der Auswirkungen der Schwermetallverschmutzung in Böden zu erweitern.

1. Introduction

1.1. Cadmium pollution

Apart from the well documented climate change, pollution of ecosystems is another effect of the Anthropocene. Soil pollution with heavy metal presents a global challenge, exerting profound impacts on biota (Rodríguez Eugenio et al., 2018). Cadmium (Cd) stands as a highly toxic heavy metal and a common pollutant with widespread incidence (Patra et al., 2011), with human polluting sources including mining, agricultural activities (Irfan et al., 2021; Zou et al., 2021), industry and wastewater (Irfan et al., 2021). It is commonly found in urban areas (Audusseau et al., 2020), but also in rural and agricultural areas like soil-rice ecosystems (Zou et al., 2021). Even at low non-toxic concentrations of Cd have major effects on animal development (Audusseau et al., 2020; Rasnaca et al., 2022) and can have long-lasting effects in the environment (Šrut et al., 2017). Most physiological and molecular effects of Cd intoxication are not well understood, particularly in invertebrates. Cd causes a range of cellular and molecular changes in the exposed organisms leading to stress response mechanisms like expression of metallothioneins (MTs) (Mustonen et al., 2014). It has been shown to induce enzyme inhibition, generation of reactive oxygen species (ROS) and perturbation of apoptosis or cell proliferation (Fuller-Espie et al., 2011; Sandbichler and Höckner, 2016; Waisberg et al., 2003). The initial inhibition of DNA methyltransferases (DMNTs) by Cd and their subsequent overexpression may lead to global hypermethylation seen in many animals (Cribiu et al., 2018).

1.2. Annelids

The Phylum Annelida, or segmented worms, includes a highly diverse and ecologically important group of metazoans which can be found in marine, freshwater and terrestrial settings. They show a set of very diverse characteristics and feeding strategies. Over 20,000 species are currently described, but projections estimate over 30,000 species (Magalhães et al., 2021). Annelids are an integral, key element of their ecosystems: they act as ecological engineers by breaking down organic matter and turning over sediments (Capa and Hutchings, 2021). Annelids are often candidates for biomarkers and bioindicators (Abubakr, 2018; Capa and Hutchings, 2021; Pelosi and Römbke, 2018). They often show high tolerance to heavy metals and other contaminants by maintaining detoxification systems and accumulating excess in their body. They are the most abundant taxon in benthic communities (Dean, 2008) and soil macrofauna of temperate regions (Blouin et al., 2013). In stark contrast to their importance, overall basic research is lacking, their evolutionary history and relationships between extant taxa are still not fully understood, and model organisms are only slowly emerging (Seaver, 2016).

With about 9,000 known species the Clitellatas are their most taxonomically diverse class. This contains the earthworms and leeches with the former affecting the ecosystem in various, still not fully revealed, ways (Blouin et al., 2013). Like other annelids, some earthworms can accumulate heavy metals within their body to varying degrees (Audusseau et al., 2020), which can also lead to trophic transfer of heavy metals (Richardson et al., 2020). Worms of different ecological niches show interactions and change plant communities (Asshoff et al., 2010). Invasive species give remarkable opportunities to show these effects. They can alter the soil carbon dynamics in forests (Jennings and Watmough, 2016), change microbial and soil properties (Price-Christenson et al., 2020) and can destroy up to 30% of plant seeds (Cassin and Kotanen, 2016).

1.3. *Lumbricus terrestris*

The common earthworm, *Lumbricus terrestris,* is an hermaphroditic anecic organism, building permanent burrows to bring in leaf litter to consume and deposit their cast mainly between 0 to 100 cm underground, therefore contributing to soil mixing in these regions (Zicsi et al., 2011). Like other annelids they have some capacity of metal detoxification and accumulation (Audusseau et al., 2020) and they only have limited capacity to move above soil, with even some homing capacity increasing the spatial influence (Nuutinen and Butt, 2005). Thought to be native to Western Europe, *L. terrestris* is found as an invasive species in temperate to mild boreal climates in Australia, New Zealand, North America (where it also has an additional economic value as fish bait in Canada), and even within Europe - for example in isolated parts of Romania (Keller et al., 2020). This makes them, among others, good study organisms for the dynamics of invasive species. In general genetic diversity of populations within and between North America and Europe are comparable, hinting at multiple introductions, with estimates of genetic distance ranging from 3 - 5% (Gailing et al., 2012; Keller et al., 2020). However, a population differentiation as high as 15% was reported between Canadian and German populations (Velavan et al., 2009).

In 2023 a reference genome of *L. terrestris* was released by the Darwin Tree of Life Project (The Darwin Tree of Life Project Consortium et al., 2022) and we had the opportunity to use a pre-release version of the genome structural annotation.

1.4. Epigenetic signals

Epigenetics is a broad term first described in 1942 (Waddington, 1942) as causal interactions between genes and their products. In the early stages the term inspired two different focuses, one on genotype-phenotype relations, and another on long term gene regulation and cell inheritance. The remenands of these different views are still seen in the

slightly different definitions in different fields (Deans and Maggert, 2015). With the emergence of high throughput sequencing technologies it was possible to extend epigenetic analysis to non-model organisms, taking advantage of more complex ecological and evolutionary settings (Herrel et al., 2020).

Our general definition of epigenetics encompasses any modifications not relying on DNA sequences affecting gene expression, which include any alterations in DNA methylation, histone modification and miRNA (Moosavi and Motevalizadeh Ardekani, 2016). In this study we focus on DNA methylation changes as epigenetic signals without testing for expression changes, as the DNA sequence on its own can't explain the entire biological variation and its complexity (Maher, 2008). DNA methylation is understood as the modification of cytosine by addition of a methyl group (Lister and Ecker, 2009) in particular sequence contexts (in particular CpG context in animal genomes). It can be measured on a global genome level (e.g., Xu et al., 2000) and single-base level (e.g., Meissner, 2005). The regulatory effects of DNA methylation are complex (Jones, 2012). The prevailing notion that methylation occurring at transcription starting sites (TSS) serves to halt transcription relies on the structural characteristics of the respective genomic regions. Specifically, CpG islands exhibit the most pronounced effects in this regard, while TSS with lower GC content yield more varied outcomes. Investigations into the effects of methylation within gene bodies outside of TSS remain limited, but it is recognized that methylation at such regions can influence gene regulation and alternative splicing patterns (Planques et al., 2021; Yearim et al., 2015). The DNA methylation systems represent a conserved ancestral trait, shared between annelids and vertebrate (Planques et al., 2021)

1.5. DNA methylation and methods of detection: BS-seq and EM-seq

DNA methylation, or more precisely 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), is one of the multiple epigenetic marks. DNA methylation has a wide range of often not well understood biological effects and is in general associated with gene regulation and development (Smith and Meissner, 2013), but it can be also connected to adaptation to environmental stimuli on individual level (Artemov et al., 2017) or on population level, without relying on genetic variation.

DNA methylation can be highly position dependent, as methylation may not extend to the next cytosine, even if it is six base pairs apart (Weaver et al., 2004). Treatment with the harmless, inexpensive Sodium-Bisulfite with additional PCR amplification and sequencing (BS-seq) ensures high sensitivity on DNA methylation detection at a single-base level (Lister et al., 2008). Unmethylated cytosine forms with the bisulfate in acidic conditions an intermediate stage (5,6-dihydrocytosine-6-sulfonate) which is deaminated and by elimination forms a uracil residue in alkaline pH (Tanaka and Okamoto, 2007). 5mC and 5hmC are mostly protected from this

chemical reaction. In the resulting PCR and/or sequencing, the uracil is replaced by and read as thymine. By comparing treated reads with genomic data, DNA methylation states can be inferred (Krueger and Andrews, 2011; Trucchi et al., 2016).

Reduced representation bisulfite sequencing (Meissner, 2005) with its random approach and cost effective design is an effective tool based on the BS-seq approach. The method provides a genome-wide survey of DNA methylation, but only covers fractions instead of the full genome as in a whole-genome approach, and its application suffers from imprecise targeting of candidate regions (Paun et al., 2019).

In general mapping of BS-seq reads to reference genomes poses several problems. DNA strands are not symmetrically methylated, hence each strand needs to be considered separately. Methylation states may differ between cells, and the bisulfite-treated data must be aligned to an unchanged genome to estimate methylation states (Krueger et al., 2012). Mismatches of C to T indicate non-methylated cytosines in the data, but mismatches can also be attributed to genetic polymorphism. The issue of polymorphisms can be prevented by incorporating genomic reads into the method (e.g., Trucchi et al., 2016).

Treatment with bisulfite can also lead to biases due to DNA degradation, as the cytosine intermediate can also generate an abasic site after depyrimidination, which is known to cause strand scission during heating or alkaline conditions (Tanaka and Okamoto, 2007). This gives rise to biases, in particular as GC-rich regions are more prone to degradation leading to underrepresentation. In addition, methylated C are less likely to degrade linking reading depth with methylation state and therefore over-representing methylated regions (Olova et al., 2018).

Instead of a harsh treatment with bisulfite, heat and alkaline conditions, an alternative approach was developed using enzymatic reactions (enzymatic methyl sequencing, EM-seq). The method involves APOBEC3A, an enzyme which deaminates cytosines independent of methylation state, and TET2 and TG-BGT which in a preceding step convert methylated Cs into a product, that blocks deamination by APOBEC3A. This method is shown to decrease DNA degradation, therefore reducing bias compared to bisulfite treatment while still delivering the same signals (Vaisvila et al., 2021)

1.6. Aims and hypotheses

The understanding of DNA methylation regulation in annelids is currently limited. In this work, we used EM-seq to generate whole-genome estimates of DNA methylation at the single-base level in response of *L. terrestris* to different levels of environmental pollution with Cd. Based on the previous studies we formulate the following hypotheses:

- DNA methylation patterns change in response to Cd pollution in soil to a degree corresponding to pollution intensity.

- Differentially methylated regions (DMR) in response to Cd most often overlap with gene bodies rather than promoter regions.
- Genes with DMRs are enriched for cellular stress response and particularly heavy metal stress response and detoxification.

2. Methods

2.1. Experimental design

The experimental setup is described in detail in Aigner et al., 2022. Shortly, earthworms and soil were purchased from the company Proinsects (Germany) and were acclimated to laboratory conditions for four weeks. Earthworms were exposed to control and Cd-spiked soil (10 and 25 mg Cd per kg soil) for a period of 12 weeks in containers containing 150g wet soil/worm. Pictures from the experiment can be seen in Figure 1. For each experimental condition 25 earthworms were used and the exposures were run in triplicates. Earthworms were kept at 15 °C, 8/16 h light/dark cycle, in soil with 70% moisture content and fed weekly with horse manure. After 12 weeks, three earthworms from each replicate were sampled (nine earthworms per treatment). 1.5 cm tissue from the most distal part was sampled, flash frozen in liquid nitrogen, stored at -80°C and subsequently used for DNA extraction.

Figure 1: Pictures of experimental design. Left: Container; Right: *L. terrestris*.
Pictures provided by Maja Šrut.

2.2. Enzymatic methyl sequencing

Genomic DNA of 30 samples was extracted using GenElute Mammalian DNA miniprep kit (Sigma-Aldrich, USA) according to manufacturer's instructions. DNA was purified using the NucleoSpin gDNA Clean-up kit (Macherey-Nagel) and quantified using Nanodrop and PicoGreen protocols (Quant-iT™ PicoGreen™ dsDNA Assay Kit, Thermo Fisher Scientific). The DNA extractions were processed using the NEBNext® Enzymatic Methyl-seq Kit, obtained from New England Biolabs Inc (Vaisvila et al., 2021), by following the provided Instruction Manual (v. 7.0_4/23). For each sample 200 ng of extracted DNA was sheared at 4°C with 5 cycles of 15s on and 45 s off to a target average length of 400 bp using a Bioruptor® Pico sonication device (Diagenode, Denville, NJ, USA). For denaturation the recommended option of Formamide was

used. To increase the yield per library, the PCR amplification was set to six instead of four Cycles. Library quality control was done using Qubit® 3.0 Fluorometric Quantification (Life Technologies, Carlsbad, CA, USA) and a Bioanalyzer from Agilent. Libraries were sequenced on NovaSeq S4 300 cycles, as half lane (XP Workflow) by Vienna BioCenter Sequencing Facility.

2.3. Bioinformatic Analysis

All workflows and R, python and bash scripts and additional graphs are provided at: https://github.com/oberauerV/Epigenetics_LTer. Overview of the softwares used is presented in Table 1. All analyses were performed using the Life Science Computer Cluster (LiSC) of the University of Vienna.

Quality control was done using fastqc (Andrews, 2010), multiqc (Ewels et al., 2016) and phyloFlash (Gruber-Vodicka et al., 2020). Due to the missing of two barcodes and comparing the read counts of the sequenced samples, four samples were excluded from further analyses as two samples showed 500 and 2,500 reads respectively and and two samples, with the barcode matching the missing ones, showed much higher than expected read counts (~ 150 M, mean = 73 M).

Table 1. Softwares used for the bioinformatic analyses

Software	Version
Angsd	v.0941
bedtools	v.2.30.0
bismark	v.0.24.1
bsseq	v.1.36.0
busco	v.5.4.7
dss	v.2.48.0
fastqc	v.0.12.1
ggplot2	v.3.4.4
multiqc	v.1.18
OmicsBox	v.3.1.11
phyloFlash	v.3.4.2
R	v.4.3.2
Revigo	v.1.8.1
RStudio Server	2023.12.1 Build 402
seqkit	v.2.6.1
TrimGalore	v.0.6.10

For read quality trimming TrimGalore (Krueger, 2019) was used. Mapping and methylation calling was done using bismark (Krueger and Andrews, 2011) in two steps, first with unclipped reads, then with the clipped unmapped reads remaining from the first step. Clipping removed the first and last ten nucleotides from each read of each pair. The bam files were then combined together with bismark during deduplication. As reference genome, the Darwin Tree of Life (The Darwin Tree of Life Project Consortium et al., 2022) *L. terrestris* assembly (GenBank assembly accession: GCA_949752735.1) was used. Details and specifics are provided in the Workflow.

To call differentially methylated regions (DMR) we used the R (R Core Team, 2022) and the dss package (Feng et al., 2014) on a RStudio Server (RStudio Team, 2020) provided by the Life Science Computer Cluster (LiSC).

Pairwise differential methylation between the control and Cd-exposed samples (10 and 25 mg Cd/kg soil, respectively) was performed by using the two-group differentially-methylated-loci test function from dss. To estimate differentially methylated loci dss takes read depth, within-condition and between-condition variability. DMRs were called using the callDMR function. The resulting outputs were intersected using bedtools (Quinlan and Hall, 2010) to obtain three datasets: DMRs from low Cd condition, DMRs from high Cd condition and their intersection. Each of these datasets was further intersected with an structural annotation gff file for the genome, obtained from Darwin Tree of Life (pre-release version). In addition, a putative "promoter" gff file containing the regions 2,000 bp upstream of the genes was produced and used.

Functional annotations for the set of genes of the reference genome were assigned in OmicsBox (BioBam) using the Metazoa database. GO term enrichment analyses of the different sets of genes was done using the FIsher's exact test in OmicsBox, against the background of the GO terms in the reference genome.

The genetic structure of the accessions used in the experiment was checked with a PCA. For this purpose we masked bases in the bismark output bam files using a python script. The raw output bam files use a custom tag showing each position of the read and the methylation context if available (. for any base except C, small or large letter depending on methylation context of C). All Cs recognized by bismark were masked to N. An Angsd (Korneliussen et al., 2014) pipeline was used to generate the PCA from the C-masked bam alignment files. The data was filtered to allow a maximum 50% of accessions to be missing per site and include polymorphic sites with a minimum minor allele frequency 0.05. The PCA was plotted using ggplot2 (Wickham, 2016)

To contrast the genetic and epigenetic structure of accessions, an epigenetic PCA was also produced, based on methylation estimates of the BSmooth function from the bsseq

package (Hansen et al., 2012). The PCA was done using base R and plotting was done using ggplot2.

Region-by-region DNA methylation plots were made by using the plotMany function of bsseq, a BSmoothed object of the whole dataset and the DMR regions identified by dss.

3. Results

3.1. General

The trimmed dataset consisted of 26 samples from three different conditions. The control and low Cadmium (10mg Cd/kg soil) condition each are represented by nine samples and the high Cadmium (25 mg Cd/kg soil) by eight. The mapping rates of the samples can be found in Table 2. Overall the trimmed unclipped samples show an average mapping rate of 34.2% and methylation

Table 2. Mapping rates, CpG methylation level and the number of cytosines in the mapped reads per accession.

Sample	1st pass mapping - unclipped reads			2nd pass mapping - clipped reads		
	% Mapping rate	% CpG Methylated	# of C	% Mapping rate	% CpG Methylated	# of C
c12-1	30.6	13.1	1,575,560,227	4.5	9.1	137,800,419
c12-2	31.1	13.1	994,156,827	4.6	9.6	86,152,662
c12-3	28.1	13.9	899,793,084	4.5	10.6	88,322,236
c12-5	33.7	13.9	1,625,094,921	4.7	9.5	127,421,148
c12-6	33.2	12.8	1,403,215,684	4.6	9.1	110,042,575
c12-7	29.5	14.0	789,873,183	4.4	10.2	70,897,312
c12-9	32.8	12.9	1,382,699,735	5.1	9.8	122,193,863
c12-10	38.8	14.6	2,300,850,795	6.0	10.6	187,849,881
c12-11	38.7	11.0	1,808,388,613	6.1	7.7	150,909,955
10cd12-1	32.4	13.3	1,371,581,505	4.7	9.2	116,300,579
10cd12-2	34.0	12.8	1,749,034,507	5.0	9.2	146,419,662
10cd12-3	32.2	12.9	1,520,644,751	5.3	9.5	144,663,971
10cd12-4	39.7	13.4	2,248,564,340	5.4	9.1	155,807,985
10cd12-5	33.1	13.3	1,287,654,277	4.9	9.6	108,915,318
10cd12-6	33.4	12.2	1,395,681,253	5.1	9.0	122,211,096
10cd12-7	31.9	13.0	1,114,014,475	4.8	9.5	97,734,069
10cd12-9	38.3	14.4	1,682,874,042	5.9	10.2	136,463,697
10cd12-10	35.9	13.9	1,543,782,969	5.5	10.5	130,028,240
25cd12-1	38.6	12.5	1,940,933,216	5.4	8.6	142,783,736
25cd12-2	35.1	13.6	1,817,120,424	5.1	10.1	145,965,423
25cd12-3	35.0	12.5	1,317,684,462	5.1	8.9	107,419,134
25cd12-5	33.5	12.7	1,030,401,209	5.0	9.3	88,133,325
25cd12-6	30.7	12.9	1,323,852,970	4.7	9.8	120,940,426
25cd12-7	37.5	14.0	1,807,785,579	6.2	10.5	160,226,502
25cd12-9	32.8	12.8	1,199,402,978	5.1	9.4	106,184,059
25cd12-11	38.4	11.9	1,897,632,874	6.3	8.5	163,870,357

of 13.1%. Unmapped reads from the first mapping were clipped to remove possible artifacts on read-ends and were mapped with an average rate of 5.2% mapping and 9.5% methylation. All estimates of mapping only include uniquely mapping read-pairs, multi mapping reads are ignored by bismark and in the further analyses.

To get an overview of the clustering patterns in the data, PCAs on the genetic/genomic and the epigenomic datasets were performed (Figure 2). The epigenetic dataset consists of the methylation estimates by dss and bseq and the genomic dataset uses the bam alignment files where each cytosine position (marked by a custom tag in bismark) is masked.

Both the genetic and epigenetic PCA did not reveal a clear clustering of samples by treatment based on the genome-wide information treatment group. One outlier in the epigenetic dataset (10cd12-2) was observed, removing the sample led to an almost exact replica. Comparing the covariance matrices, patterns changed which didn't affect the PCA.

3.2. Differentially methylated regions

After methylation calling in bismark, methylation and coverage data were used as input for the dss package in R. The average coverage of all C in the dataset is 2.6x. The control condition was compared to each of the two experimental conditions. More DMRs were observed between the high Cd treatment and controls, compared to low Cd treatment. Table 3 shows a summary of the results.

Table 3. Summary statistics from dss. Length and # of CpG are describing the DMRs. Differentially methylated loci (DML) are given for the entire genome (i.e., also outside DMRs).

Comparison	DMRs			length		#CpG		all DMLs
	#	down	up	mean	SD	mean	SD	#
10 vs. C	20,324	10,544	9,780	227.24	262.11	24.93	30.15	197,567
25 vs. C	23,263	14,413	8,850	238.35	275.31	25.86	32.00	247,808
Intersection	6,913	4,346	2,390	328.77	355.42	37.37	42.42	56,901

Example plots showing the estimated methylation level at DMRs are presented in Figure 3. In the first two graphs of Figure 3 a distinct methylation level can be observed between controls (red) and both Cd treatments. The third and fourth graphs show trend changes specific for the high Cd condition, whereas the fifth a change specific for the low Cd treatment.

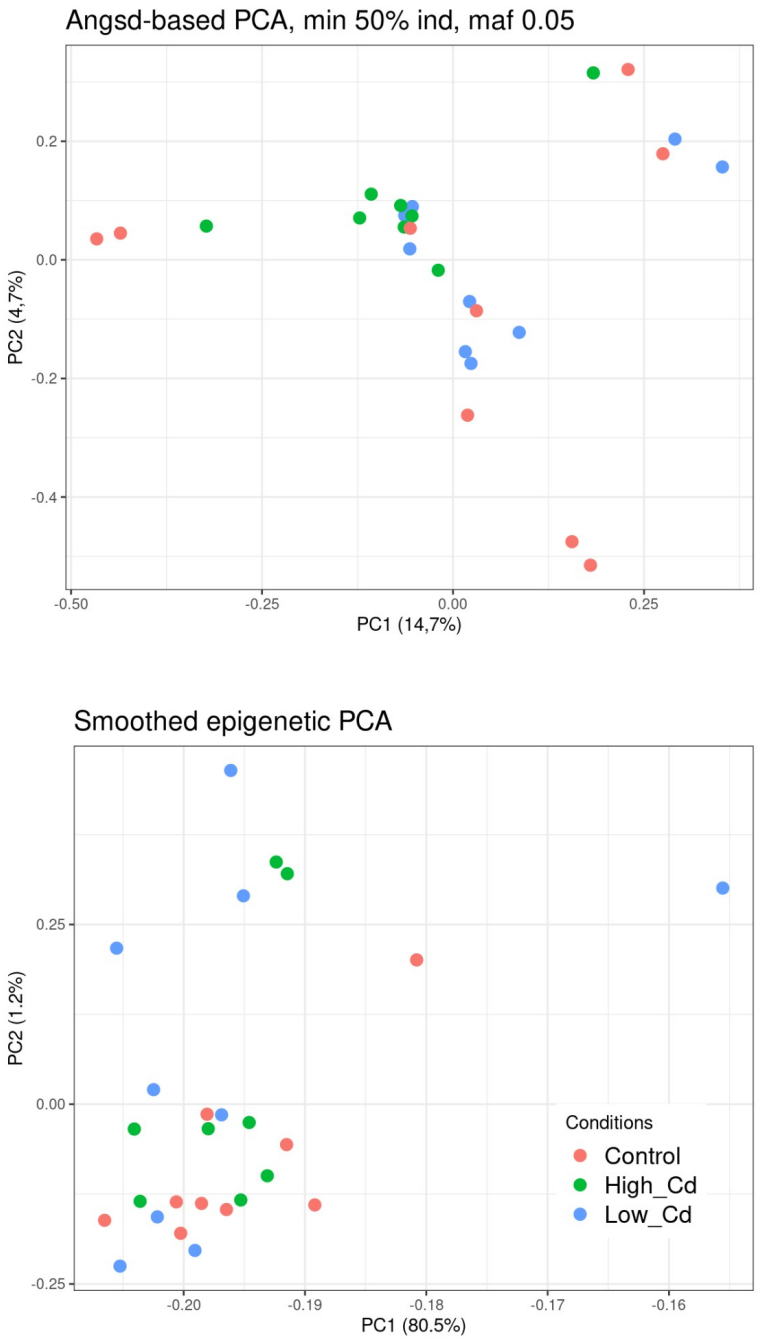

Figure 2: PCA of genomic and epigenetic data. genomic PCA, based on genotype likelihoods of the C masked bam output of bismark. epigenetic PCA, based on methylation estimates of C positions given by “dss” and “bsseq”. Each coloured dot represents an individual with their different conditions. red, control; blue, low Cd; green, high Cd.

Table 4. Number of genes with DMRs within gene bodies and/or promoters. For genes or promoters only unique entries are counted.

	# Genes	# Promoters	# Genes or promoters
C against 10	9,204	1,164	10,363
C against 25	10,675	1,386	12,059
intersection	3,113	409	3,113

3.3. DMR overlap with gene annotation

To estimate if genes could be affected by the DMRs we intersected them with all gene models in the genome annotation from the Darwin Tree of Life for the *L. terrestris* reference genome. The annotation contained 33,449 genes, and close to a third of them appeared to contain a DMR. Additionally we were interested in DMRs in promoter regions, but as the annotation file doesn´t contain promoters we generated a “promoter” annotation by taking the regions 2,000 bp upstream of each gene. Table 4 shows the number of genes affected. For GO term analysis a combined list of the genes that were affected by DMRs in gene bodies or promoters was used.

3.4. GO term enrichment

In light of the amount of genes a GO term enrichment was performed using the “Gene bodies or Promoters" gene lists. After a statistical analysis the dataset with the combined gene sets resulted in 18 significant GO Terms after FDR correction of the level of significance for multiple testing. The comparison of low Cd conditions against the controls led to no significant enrichments after FDR, whereas the comparison including the high Cd condition identified eight GO terms. REVIGO was used to generate semantic clustered scatter plots and hierarchical treeplots (Figure 4) of the 18 significant terms. To capture a more comprehensive picture of possible effects we also used non-FDR corrected GO terms, which lead to inclusion of the “response to metal ions” hierarchical label in the high Cd GO term list (Figure 5). This label is lost in the combined dataset and not found in the low Cd.

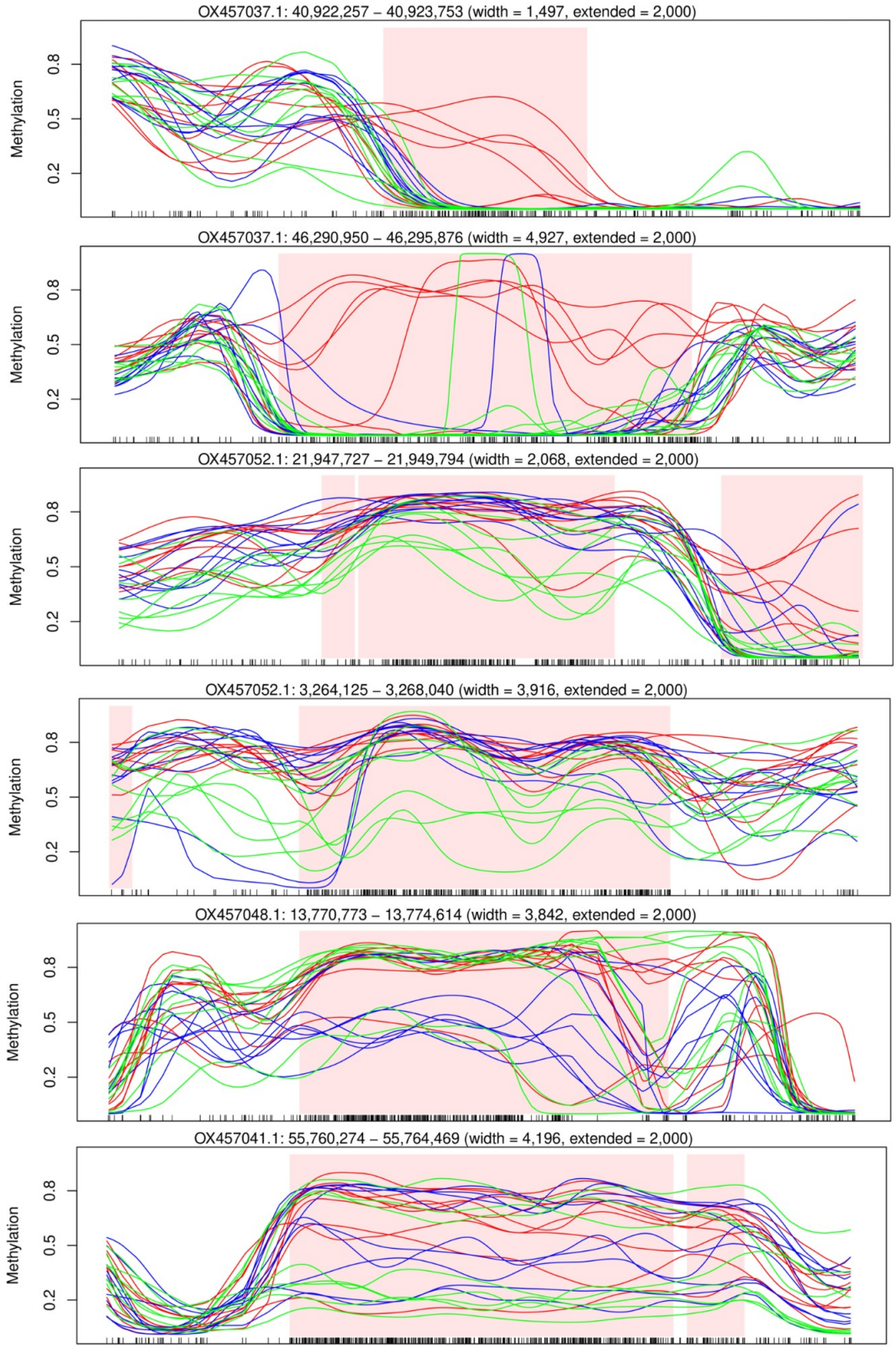

Figure 3: Examples of differentially methylated regions (DMRs). X-axis, genomic position; Y-axis, methylation level. The titles give chromosome, position and length of region. Coloured background indicates the DMR. Coloured lines give replicates of different conditions with red, control; blue, low Cd; green, high Cd.

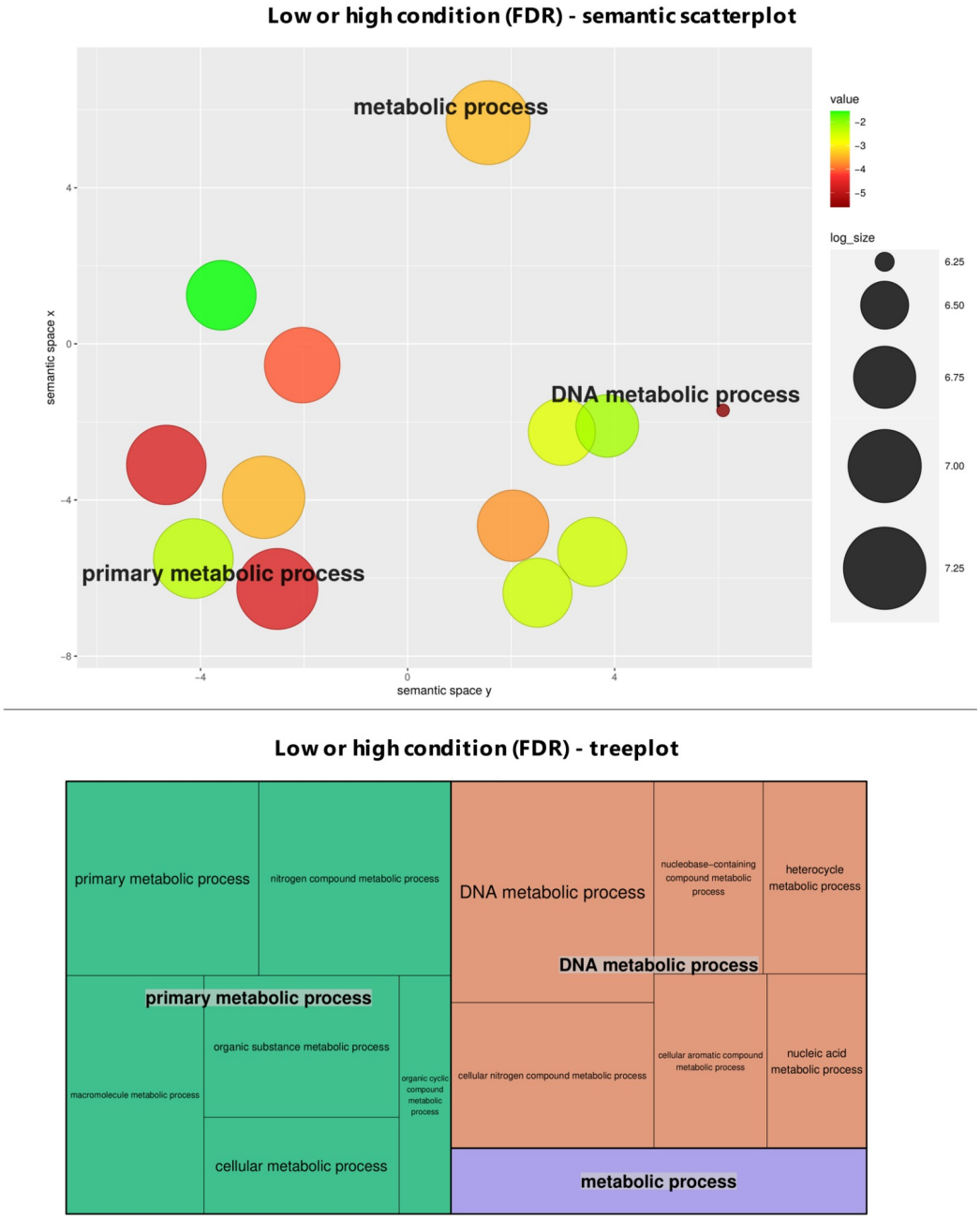

Figure 4: GO-Term Enriched Plots of combined dataset (FDR corrected); A: Semantic scatterplot and hierarchical treeplot of significant “Biological Process”-Terms

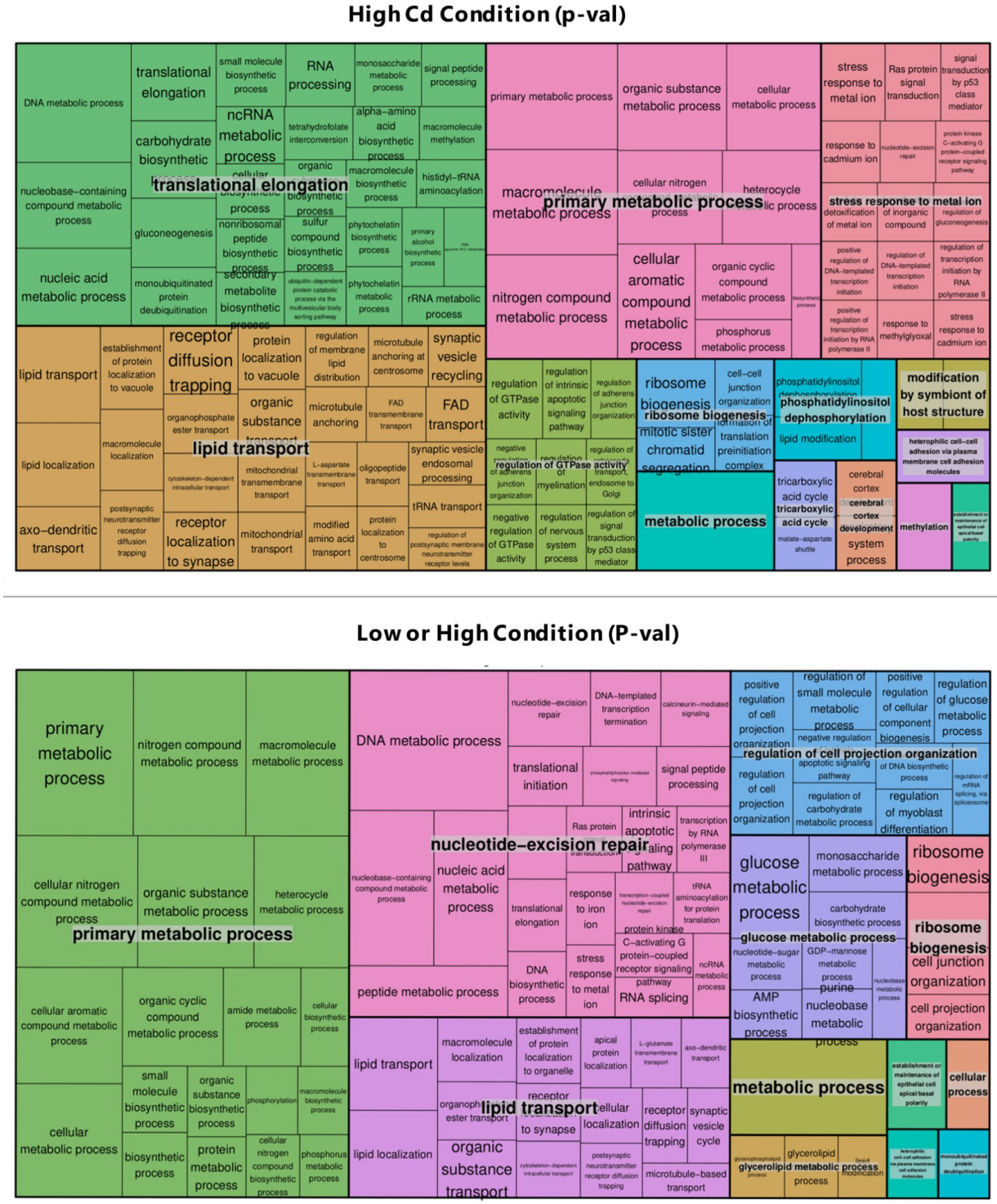

Figure 5: GO-Term Enriched tree plots (p-val); Treeplot of significant “Biological Process”-Terms of high Cd condition and the combined dataset.

4. Discussion

With this study whole genome DNA methylation in earthworm species *L. terrestris* was surveyed for the first time. We used the opportunity of the recently available genome and pre-release annotation file to evaluate the cytosine methylation and its changes at the nucleotide level in response to heavy metal Cd treatment. With an average coverage of 2.6x and the inability to unambiguously map BS-/EM-seq treated reads in repetitive regions this dataset may miss key features of epigenetic genome regulation (Marin et al., 2020; Putiri and Robertson, 2011). However, previous research indicates annelids share similar methylation patterns as vertebrates (Planques et al., 2021) and both, gene body methylation and methylations near TSS, can have major effects (Jones, 2012; Smith and Meissner, 2013; Yu et al., 2013). If we are able to reliably call methylation levels in these regions of interest we can be confident of the indicated responses. A potential reason for the low average coverage may be associated successful invasive history of *L. terrestris*. The reference genome used in this study is assembled from a specimen from the British Isles, but the samples used in our experiments originated from Canada where *L. terrestris* is considered invasive. European and North American populations maintain a high level of diversity and only show weak inbreeding patterns (Gailing et al., 2012; Keller et al., 2020; Velavan et al., 2009) and adaptation to novel environments may increase divergence (Marin et al., 2020). It is reasonable to expect that a recent increase in genetic diversity will affect the least regions constrained by selection like genes or regulatory elements. Therefore we expect that mapping rates are higher in functionally important regions. Indeed, our data contains regions with high coverage which are averaged out by regions with low coverage regions with single reads, likely due the divergence and stringent mapping settings needed for methylation calling. A major advantage of the “dss”-package is the inclusion of read depth filters for differentially methylation calling. In addition the smoothing process leads to improved methylation calls over lower coverage regions. Combining the assumptions that functional regions are more likely to be conserved, the use of multiple biological replicates per treatment, and the use of strict filtering for methylation calls based on read depth increase our confidence that the called DMRs, especially those around genes, are a true biological signal, and not an technical artifact or a random signal.

4.1. Methylation changes in response to environmentally relevant Cd pollution in the soil

Previous studies employing the same model organism and similar experimental setup showed that the low Cd concentrations (10 and 25 mg/kg) lead to hypermethylation in the *L. terrestris* genome on a global level (Aigner et al., 2022; Šrut et al., 2017). This study was

conducted in order to obtain a high-resolution picture of genome-wide methylation changes at the nucleotide level in the *L. terrestris* genome. The epigenetic PCA showed no clustering between the treatments, indicating that methylation was not consistently affected on a global level, excluding repetitive regions, by Cd treatment. However, each pairwise comparison showed DMRs with the high Cd condition affecting slightly more than the low Cd (23,263 vs 20,324), and an overlap of 6,913 DMRs identified in both. A deviation of the expected results are the direction of the methylation changes in the DMRs, as due to the previous observations of global methylation increases in response to Cd treatment (see above). We hence expected an increase in methylation levels in response to the treatments, but altogether more DMRs showed a decrease in methylation (20,611 vs 16,240 in at least one treatment) compared to control. Our results therefore suggest that hypermethylation as a response to Cd exposure in *L. terrestris* is more restricted to repetitive or less conserved regions. Hypermethylation in response to Cd is due to a compensatory overexpression of DMNT as result of an initial DMNT inhibition by Cd (Cribiu et al., 2018). DMRs with higher methylation in the Cd treatments as in control may indicate regions with no or less regulation via DNA methyltransferases while lower methylation levels may indicate regulated regions..

4.2. DMRs overlap with genes or gene regulatory elements

DNA methylation is commonly considered a “cis”-acting regulator, changing expression and/or splicing of gene products by changing the epigenetic signal at TSS and in gene bodies (Holliday and Pugh, 1975; Moarii et al., 2015). 19,309 genes or gene-promoter regions show an overlap with a DMR. Only 2,141 (11,09%) have an overlap in gene-promoter and DMR. This hints at a larger maintenance and regulation by gene body methylation, similar to vertebrates (Planques et al., 2021).

4.3. Genes affected by DMRs are linked to Cd as pollutant

In order to explore if the genes affected by DMR are involved in Cd stress responses, GO term enrichment was performed. We focused on biological processes, as a response to heavy metal would be part of this group. The genes overlapping with DMRs from the low Cd condition show no significant results after FDR, but for the high Cd condition and the combined genelist, significant terms were found. A subsequent summarization with Revigo (Supek et al., 2011) showed only general metabolic clusters and hierarchical labels (Figure 4). To gather a better trend of the response we tested for significant GO terms without FDR correction. The low Cd and combined gene list showed additional, more precise processes, but the high Cd condition showed methylation and response to heavy metals and Cd in semantic clusters and hierarchical terms (Figure 5). The methylation term could be a heritable epigenetic feature in

response to the DNMT inhibition resulting in a possible explanation of the long-lasting hypermethylation effect of Cd, but its occurrence only in the high Cd condition may indicate a regulatory response to Cd or heavy metal stress.

Hypermethylation is verified in the low Cd condition but effects on single-base level seem diffuse and dependent on the intensity of Cd exposure. Long term hypermethylation may also influence evolution as methylated C are more likely to transition to T leading to a higher mutation rate in methylated regions (Rideout et al., 1990). An epigenetic regulation in genes linked to Cd response is only evident in the high Cd condition. Combining the epigenetic dataset with expression data may clear up if low Cd leads to other transcriptional responses.

Analyzing the methylation plots, it is possible to see that despite some level of individual variation, DNA methylation changes can act on specific regions with high affinity and reproducibility between biological replicates. This dataset gives valuable insights that human pollutants in low, non-lethal concentrations can affect populations and that these effects may have further reaching implications for evolution. Annelids are an upcoming model system to study DNA methylation changes, specifically in regard to changing environmental conditions (e.g. pollution). Due to their abundance and importance in ecosystems they are often candidates for biomarkers and bioindicators. *L. terrestris* has a history as a successful invasive species and can therefore give insights into conservation biology and other fields. Particularly in regard to ecotoxicological epigenomics, earthworms could be good candidates to further study and develop predictive epigenetic biomarkers (Šrut, 2022).

The results of this study give insight into the not well understood DNA methylation mechanisms in annelids, but can also further the understanding of epigenetics and evolution overall. It is also a good base for cross-species comparisons and further research in this important ecological engineer found all over the world.

Data availability

All used scripts are available at: https://github.com/oberauerV/Epigenetics_LTer. The raw data will become publicly available from the short read archive (SRA) of GenBank upon publication of the study in a peer-reviewed journal.

5. References

Abubakr, A., 2018. Macro-Invertebrates (Annelida; Oligochaeta) As Bio-Indicator of Water Quality under Temperate Climatic Conditions. Int. J. Pure Appl. Biosci. 6, 726–737. https://doi.org/10.18782/2320-7051.5903

Aigner, G.P., Pittl, V., Fiechtner, B., Egger, B., Šrut, M., Höckner, M., 2022. Common mechanisms cannot explain time- and dose-dependent DNA methylation changes in earthworms exposed to cadmium. Sci. Total Environ. 812, 151468. https://doi.org/10.1016/j.scitotenv.2021.151468

Andrews, S., 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data.

Artemov, A.V., Mugue, N.S., Rastorguev, S.M., Zhenilo, S., Mazur, A.M., Tsygankova, S.V., Boulygina, E.S., Kaplun, D., Nedoluzhko, A.V., Medvedeva, Y.A., Prokhortchouk, E.B., 2017. Genome-Wide DNA Methylation Profiling Reveals Epigenetic Adaptation of Stickleback to Marine and Freshwater Conditions. Mol. Biol. Evol. 34, 2203–2213. https://doi.org/10.1093/molbev/msx156

Asshoff, R., Scheu, S., Eisenhauer, N., 2010. Different earthworm ecological groups interactively impact seedling establishment. Eur. J. Soil Biol. 46, 330–334. https://doi.org/10.1016/j.ejsobi.2010.06.005

Audusseau, H., Vandenbulcke, F., Dume, C., Deschins, V., Pauwels, M., Gigon, A., Bagard, M., Dupont, L., 2020. Impacts of metallic trace elements on an earthworm community in an urban wasteland: Emphasis on the bioaccumulation and genetic characteristics in Lumbricus castaneus. Sci. Total Environ. 718, 137259. https://doi.org/10.1016/j.scitotenv.2020.137259

Blouin, M., Hodson, M.E., Delgado, E.A., Baker, G., Brussaard, L., Butt, K.R., Dai, J., Dendooven, L., Peres, G., Tondoh, J.E., Cluzeau, D., Brun, J. -J., 2013. A review of earthworm impact on soil function and ecosystem services. Eur. J. Soil Sci. 64, 161–182. https://doi.org/10.1111/ejss.12025

Capa, M., Hutchings, P., 2021. Annelid Diversity: Historical Overview and Future Perspectives. Diversity 13, 129. https://doi.org/10.3390/d13030129

Cassin, C.M., Kotanen, P.M., 2016. Invasive earthworms as seed predators of temperate forest plants. Biol. Invasions 18, 1567–1580. https://doi.org/10.1007/s10530-016-1101-x

Cribiu, P., Chaumot, A., Geffard, O., Ravanat, J.-L., Bastide, T., Delorme, N., Quéau, H., Caillat, S., Devaux, A., Bony, S., 2018. Natural variability and modulation by environmental stressors of global genomic cytosine methylation levels in a freshwater crustacean, Gammarus fossarum. Aquat. Toxicol. 205, 11–18. https://doi.org/10.1016/j.aquatox.2018.09.015

Dean, H.K., 2008. The use of polychaetes (Annelida) as indicator species of marine pollution: a review. Rev Biol Trop 56.

Deans, C., Maggert, K.A., 2015. What Do You Mean, “Epigenetic”? Genetics 199, 887–896. https://doi.org/10.1534/genetics.114.173492

Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Feng, H., Conneely, K.N., Wu, H., 2014. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 42, e69–e69. https://doi.org/10.1093/nar/gku154

Fuller-Espie, S.L., Bearoff, F.M., Minutillo, M.A., 2011. Exposure of coelomocytes from the earthworm Eisenia hortensis to Cu, Cd, and dimethylbenz[a]anthracene: An in vitro study examining reactive oxygen species production and immune response inhibition. Pedobiologia 54, S31–S36. https://doi.org/10.1016/j.pedobi.2011.09.009

Gailing, O., Hickey, E., Lilleskov, E., Szlavecz, K., Richter, K., Potthoff, M., 2012. Genetic comparisons between North American and European populations of Lumbricus terrestris L. Biochem. Syst. Ecol. 45, 23–30. https://doi.org/10.1016/j.bse.2012.07.018

Gruber-Vodicka, H.R., Seah, B.K.B., Pruesse, E., 2020. phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes. mSystems 5, e00920-20. https://doi.org/10.1128/mSystems.00920-20

Hansen, K.D., Langmead, B., Irizarry, R.A., 2012. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 13, R83. https://doi.org/10.1186/gb-2012-13-10-r83

Herrel, A., Joly, D., Danchin, E., 2020. Epigenetics in ecology and evolution. Funct. Ecol. 34, 381–384. https://doi.org/10.1111/1365-2435.13494

Holliday, R., Pugh, J.E., 1975. DNA Modification Mechanisms and Gene Activity During Development: Developmental clocks may depend on the enzymic modification of specific bases in repeated DNA sequences. Science 187, 226–232. https://doi.org/10.1126/science.187.4173.226

Irfan, M., Liu, X., Hussain, K., Mushtaq, S., Cabrera, J., Zhang, P., 2021. The global research trend on cadmium in freshwater: a bibliometric review. Environ. Sci. Pollut. Res. 30, 71585–71598. https://doi.org/10.1007/s11356-021-13894-7

Jennings, B.W., Watmough, S.A., 2016. The Impact of Invasive Earthworms on Soil Respiration and Soil Carbon Within Temperate Hardwood Forests. Ecosystems 19, 942–954. https://doi.org/10.1007/s10021-016-9977-y

Jones, P.A., 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond.

Nat. Rev. Genet. 13, 484–492. https://doi.org/10.1038/nrg3230

Keller, E.L., Connolly, S.T., Görres, J.H., Schall, Jos.J., 2020. Genetic diversity of an invasive earthworm, Lumbricus terrestris, at a long-term trading crossroad, the Champlain Valley of Vermont, USA. Biol. Invasions 22, 1723–1735. https://doi.org/10.1007/s10530-020-02215-7

Korneliussen, T.S., Albrechtsen, A., Nielsen, R., 2014. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15, 356. https://doi.org/10.1186/s12859-014-0356-4

Krueger, F., 2019. Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. Babraham Bioinforma. Babraham Inst. Camb. U. K.

Krueger, F., Andrews, S.R., 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27, 1571–1572. https://doi.org/10.1093/bioinformatics/btr167

Krueger, F., Kreck, B., Franke, A., Andrews, S.R., 2012. DNA methylome analysis using short bisulfite sequencing data. Nat. Methods 9, 145–151. https://doi.org/10.1038/nmeth.1828

Lister, R., Ecker, J.R., 2009. Finding the fifth base: Genome-wide sequencing of cytosine methylation. Genome Res. 19, 959–966. https://doi.org/10.1101/gr.083451.108

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., Ecker, J.R., 2008. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. Cell 133, 523–536. https://doi.org/10.1016/j.cell.2008.03.029

Magalhães, W.F., Hutchings, P., Oceguera-Figueroa, A., Martin, P., Schmelz, R.M., Wetzel, M.J., Wiklund, H., Maciolek, N.J., Kawauchi, G.Y., Williams, J.D., 2021. Segmented worms (Phylum Annelida): a celebration of twenty years of progress through Zootaxa and call for action on the taxonomic work that remains . Zootaxa 4979, 190–211. https://doi.org/10.11646/zootaxa.4979.1.18

Maher, B., 2008. Personal genomes: The case of the missing heritability. Nature 456, 18–21. https://doi.org/10.1038/456018a

Marin, P., Genitoni, J., Barloy, D., Maury, S., Gibert, P., Ghalambor, C.K., Vieira, C., 2020. Biological invasion: The influence of the hidden side of the (epi)genome. Funct. Ecol. 34, 385–400. https://doi.org/10.1111/1365-2435.13317

Meissner, A., 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 33, 5868–5877. https://doi.org/10.1093/nar/gki901

Moarii, M., Boeva, V., Vert, J.-P., Reyal, F., 2015. Changes in correlation between promoter

methylation and gene expression in cancer. BMC Genomics 16, 873. https://doi.org/10.1186/s12864-015-1994-2

Moosavi, A., Motevalizadeh Ardekani, A., 2016. Role of Epigenetics in Biology and Human Diseases. Iran. Biomed. J. 246–258. https://doi.org/10.22045/ibj.2016.01

Mustonen, M., Haimi, J., Väisänen, A., Knott, K.E., 2014. Metallothionein gene expression differs in earthworm populations with different exposure history. Ecotoxicology 23, 1732–1743. https://doi.org/10.1007/s10646-014-1338-z

Nuutinen, V., Butt, K., 2005. Homing ability widens the sphere of influence of the earthworm L. Soil Biol. Biochem. 37, 805–807. https://doi.org/10.1016/j.soilbio.2004.10.008

Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R.V., Branco, M.R., Reik, W., 2018. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. 19, 33. https://doi.org/10.1186/s13059-018-1408-2

Patra, R.C., Rautray, A.K., Swarup, D., 2011. Oxidative Stress in Lead and Cadmium Toxicity and Its Amelioration. Vet. Med. Int. 2011, 1–9. https://doi.org/10.4061/2011/457327

Paun, O., Verhoeven, K.J.F., Richards, C.L., 2019. Opportunities and limitations of reduced representation bisulfite sequencing in plant ecological epigenomics. New Phytol. 221, 738–742. https://doi.org/10.1111/nph.15388

Pelosi, C., Römbke, J., 2018. Enchytraeids as bioindicators of land use and management. Appl. Soil Ecol. 123, 775–779. https://doi.org/10.1016/j.apsoil.2017.05.014

Planques, A., Kerner, P., Ferry, L., Grunau, C., Gazave, E., Vervoort, M., 2021. DNA methylation atlas and machinery in the developing and regenerating annelid Platynereis dumerilii. BMC Biol. 19, 148. https://doi.org/10.1186/s12915-021-01074-5

Price-Christenson, G.J., Johnston, M.R., Herrick, B.M., Yannarell, A.C., 2020. Influence of invasive earthworms (Amynthas spp.) on Wisconsin forest soil microbial communities and soil chemistry. Soil Biol. Biochem. 149, 107955. https://doi.org/10.1016/j.soilbio.2020.107955

Putiri, E.L., Robertson, K.D., 2011. Epigenetic mechanisms and genome stability. Clin. Epigenetics 2, 299–314. https://doi.org/10.1007/s13148-010-0017-z

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma. Oxf. Engl. 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. Vienna Austria.

Rasnaca, I., Kille, P., Newbold, L.K., Spurgeon, D.J., 2022. Impacts of Life-Time Exposure of Arsenic, Cadmium and Fluoranthene on the Earthworms' L. rubellus Global DNA

Methylation as Detected by msAFLP. Genes 13, 770. https://doi.org/10.3390/genes13050770

Richardson, J.B., Görres, J.H., Sizmur, T., 2020. Synthesis of earthworm trace metal uptake and bioaccumulation data: Role of soil concentration, earthworm ecophysiology, and experimental design. Environ. Pollut. 262, 114126. https://doi.org/10.1016/j.envpol.2020.114126

Rideout, W.M., Coetzee, G.A., Olumi, A.F., Jones, P.A., 1990. 5-Methylcytosine as an Endogenous Mutagen in the Human LDL Receptor and p53 Genes. Science 249, 1288–1290. https://doi.org/10.1126/science.1697983

Rodríguez Eugenio, N., McLaughlin, M.J., Pennock, D.J., 2018. Soil pollution: a hidden reality. Food and Agriculture Organization of the United Nations, Rome.

RStudio Team, 2020. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA.

Sandbichler, A., Höckner, M., 2016. Cadmium Protection Strategies—A Hidden Trade-Off? Int. J. Mol. Sci. 17, 139. https://doi.org/10.3390/ijms17010139

Seaver, E.C., 2016. Annelid models I: Capitella teleta. Curr. Opin. Genet. Dev., Developmental mechanisms, patterning and evolution 39, 35–41. https://doi.org/10.1016/j.gde.2016.05.025

Smith, Z.D., Meissner, A., 2013. DNA methylation: roles in mammalian development. Nat. Rev. Genet. 14, 204–220. https://doi.org/10.1038/nrg3354

Šrut, M., 2022. Environmental Epigenetics in Soil Ecosystems: Earthworms as Model Organisms. Toxics 10, 406. https://doi.org/10.3390/toxics10070406

Šrut, M., Drechsel, V., Höckner, M., 2017. Low levels of Cd induce persisting epigenetic modifications and acclimation mechanisms in the earthworm Lumbricus terrestris. PLOS ONE 12, e0176047. https://doi.org/10.1371/journal.pone.0176047

Supek, F., Bošnjak, M., Škunca, N., Šmuc, T., 2011. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLoS ONE 6, e21800. https://doi.org/10.1371/journal.pone.0021800

Tanaka, K., Okamoto, A., 2007. Degradation of DNA by bisulfite treatment. Bioorg. Med. Chem. Lett. 17, 1912–1915. https://doi.org/10.1016/j.bmcl.2007.01.040

The Darwin Tree of Life Project Consortium, Blaxter, M., Mieszkowska, N., Di Palma, F., Holland, P., Durbin, R., Richards, T., Berriman, M., Kersey, P., Hollingsworth, P., Wilson, W., Twyford, A., Gaya, E., Lawniczak, M., Lewis, O., Broad, G., Howe, K., Hart, M., Flicek, P., Barnes, I., 2022. Sequence locally, think globally: The Darwin Tree of Life Project. Proc. Natl. Acad. Sci. 119, e2115642118. https://doi.org/10.1073/pnas.2115642118

Trucchi, E., Mazzarella, A.B., Gilfillan, G.D., Lorenzo, M.T., Schönswetter, P., Paun, O., 2016.

Bs RAD seq: screening DNA methylation in natural populations of non-model species. Mol. Ecol. 25, 1697–1713. https://doi.org/10.1111/mec.13550

Vaisvila, R., Ponnaluri, V.K.C., Sun, Z., Langhorst, B.W., Saleh, L., Guan, S., Dai, N., Campbell, M.A., Sexton, B.S., Marks, K., Samaranayake, M., Samuelson, J.C., Church, H.E., Tamanaha, E., Corrêa, I.R., Pradhan, S., Dimalanta, E.T., Evans, T.C., Williams, L., Davis, T.B., 2021. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome Res. 31, 1280–1289. https://doi.org/10.1101/gr.266551.120

Velavan, T.P., Weller, S., Schulenburg, H., Michiels, N.K., 2009. High genetic diversity and heterogeneous parasite load in the earthworm Lumbricus terrestris on a German meadow. Soil Biol. Biochem. 41, 1591–1595. https://doi.org/10.1016/j.soilbio.2009.03.026

Waddington, C.H., 1942. The epigenotype. Endeavour 18–20.

Waisberg, M., Joseph, P., Hale, B., Beyersmann, D., 2003. Molecular and cellular mechanisms of cadmium carcinogenesis. Toxicology 192, 95–117. https://doi.org/10.1016/S0300-483X(03)00305-6

Weaver, I.C.G., Cervoni, N., Champagne, F.A., D'Alessio, A.C., Sharma, S., Seckl, J.R., Dymov, S., Szyf, M., Meaney, M.J., 2004. Epigenetic programming by maternal behavior. Nat. Neurosci. 7, 847–854. https://doi.org/10.1038/nn1276

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis.

Xu, M., Li, X., Korban, S.S., 2000. AFLP-Based detection of DNA methylation. Plant Mol Biol Rep. 18: 361-368.

Yearim, A., Gelfman, S., Shayevitch, R., Melcer, S., Glaich, O., Mallm, J.-P., Nissim-Rafinia, M., Cohen, A.-H.S., Rippe, K., Meshorer, E., Ast, G., 2015. HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. Cell Rep. 10, 1122–1134. https://doi.org/10.1016/j.celrep.2015.01.038

Yu, D.-H., Ware, C., Waterland, R.A., Zhang, J., Chen, M.-H., Gadkari, M., Kunde-Ramamoorthy, G., Nosavanh, L.M., Shen, L., 2013. Developmentally Programmed 3′ CpG Island Methylation Confers Tissue- and Cell-Type-Specific Transcriptional Activation. Mol. Cell. Biol. 33, 1845–1858. https://doi.org/10.1128/MCB.01124-12

Zicsi, A., Szlavecz, K., Csuzdi, C., 2011. Leaf litter acceptance and cast deposition by peregrine and endemic European lumbricids (Oligochaeta: Lumbricidae). Pedobiologia 54, S145–S152. https://doi.org/10.1016/j.pedobi.2011.09.004

Zou, M., Zhou, S., Zhou, Y., Jia, Z., Guo, T., Wang, J., 2021. Cadmium pollution of soil-rice ecosystems in rice cultivation dominated regions in China: A review. Environ. Pollut. 280, 116965. https://doi.org/10.1016/j.envpol.2021.116965

Appendix

Workflow

```
### Workflow
### Assumptions:
### All samples are paired-end reads
### Text file with list of samples (and conditions if needed)
### Quality assessment can be done after each step
### it is advised to change filenames to something meaningful, especially for the read data

## Adapter removal and Quality trimming using trim-galore v0.6.10

while read line; do
        sample=$(echo $line|awk '{print $2}')
        trim_galore --cores 20 --paired -o noclip --no_report_file sample/"$sample".R1.fq.gz \
                sample/"$sample".R2.fq.gz
done <samples

## genome preparation using Bismark v.0.24.1

bismark_genome_preparation genome/ #directory containing reference genome

## First alignment round using bismark
## optional: generate text file for containing the samples for each condition

while read line; do
        sample=$(echo $line|awk '{print $2}')
        condition=$(echo $line|awk '{print $3}')
        bismark -o bismark_"$condition" --multicore 22 -un --score_min L,0,-0.6 \
                --genome genome/ -1 noclip/"$sample".1.fq.gz -2 noclip/"$sample".2.fq.gz
done <samples

# move unmapped reads to separate directory
mv bismark_condition/*unmapped* unmapped/

# in each condition directory generate M-Bias plots to identify ignore statements in methylation
        extractor manually
export PERL5LIB=/lisc/user/oberauer/perl5/lib/perl5:$PERL5LIB #perl module GD::Graph needed
bismark_methylation_extractor --parallel 28 --mbias_only *.bam

## Clip unmapped reads using TrimGalore

while read line; do
        sample=$(echo $line|awk '{print $2}')
        trim_galore --cores 30 --paired --clip_R1 10 --clip_R2 10 --three_prime_clip_R1 10 \
                --three_prime_clip_R2 10 -o clip --no_report_file unmapped/"$sample".1*fq.gz \
                unmapped/"$sample".2*fq.gz
```

```
done <samples

## Aligment of trimmed reads

while read line; do
        sample=$(echo $line|awk '{print $2}')
        bismark -o bismark_clipped --multicore 22 --score_min L,0,-0.6 --genome genome/ \
                -1 clip/"$sample".1_clip.fq.gz -2 clip/"$sample".2_clip.fq.gz
done <samples

## Combine aligments and deduplication

while read line; do
        sample=$(echo $line|awk '{print $2}')
        dir=$(echo $line|awk '{print $3}')
        deduplicate_bismark --multiple --output_dir 2_dedup/ bismark_"$dir"/"$sample"_pe.bam \
                bismark_clipped/"$sample"_clip_pe.bam
done <samples

## Methylation extraction (.cov.gz file for dss input)
# when including ignore statements, manually check M-Bias
# CpG_context and Non_CpG_context are not needed in our analysis, removed for resources
while read line; do
        sample=$(echo $line|awk '{print $2}')
        bismark_methylation_extractor --parallel 28 --bedGraph --comprehensive --merge_non_CpG\
                --ignore 5 --ignore_3prime 3 --ignore_r2 5 --ignore_3prime_r2 3 \
                dedup/"$sample"_pe.multiple.deduplicated.bam
        rm CpG_context_"$sample"_pe.multiple.deduplicated.txt \
                Non_CpG_context_"$sample"_pe.multiple.deduplicated.txt  \
                "$sample"_pe.multiple.deduplicated.bedGraph.gz \
                "$sample"_pe.multiple.deduplicated.M-bias.txt \
                "$sample"_pe.multiple.deduplicated_splitting_report.txt
done <samples

##generate dss input filenames

gunzip *.cov.gz
while read line; do
        sample=$(echo $line|awk '{print $2}')
        awk '{print $1, $2, $5 + $6, $5}' "$sample".cov > "$sample".tsv
done <samples
```