Comprehensive analysis of methylation data in non-model plant species

Dissertation

„kumulativ“

zur Erlangung des Grades eines

Doktor der Naturwissenschaften

(Dr. rer.nat.)

des Fachbereichs Biologie der Philipps-Universität Marburg

vorgelegt von

Sultan Nilay Can

aus Ankara, Türkei

Marburg an der Lahn

Mai 2021

Die vorliegende Dissertation wurde von April 2018 bis Mai 2021 am Fachbereich Biologie, Pflanzenzellbiologie unter Leitung von Prof. Dr. Stefan A. Rensing und Dr. Noé Fernández-Pozo angefertigt.

Vom Fachbereich Biologie der Philipps-Universität Marburg (Hochschulkennziffer 1180) als Dissertation angenommen am

Erstgutachter: Prof. Dr. Stefan A. Rensing

Zweitgutachter: Prof. Dr. Lars Opgenoorth

Prof. Dr. Dominik Heider

Dr. Noé Fernández-Pozo

Tag der Disputation:

For my parents, Salime and Yusuf Can,
for giving me the greatest gift ever
— an education

&

For my hero, my brother Önay,
For his advice, his love, and his faith.
Since he always understood.

"The human takes places in the universe with its heart, not with its body."
—Yasar Kemal, *The Legend of Ararat*

I Publications and contributions

All research work achieved during my time as a Ph.D. student in this thesis is either published in peer-reviewed scientific journals or under review at the time of submitting this thesis. The list of publications and my contribution to each publication are listed below.

I.I Publications contributing to this thesis

A blind and independent benchmark study for detecting differentially methylated regions in plants (Paper-I)

Clemens Kreutz, Nilay S Can, Ralf Schulze Bruening, Rabea Meyberg, Zsuzsanna Mérai, Noe Fernandez-Pozo, Stefan A Rensing

Bioinformatics *(2020), Volume 36, Issue 11, Pages 3314–3321,*
https://doi.org/10.1093/bioinformatics/btaa191

My contribution: The literature review of DMR tools was done by me together with former bachelor Julia Ott in Rensing Lab. Simulated datasets with selected DMR callers were analyzed by me. *P. abies* datasets were assayed together with Ralf Schulze Bruening.

Thesis chapter 3.1.

The EpiDiverse plant Epigenome-Wide Association Studies (EWAS) pipeline (Paper-II)

Sultan Nilay Can, Adam Nunn, Dario Galanti, David Langenberger, Claude Becker, Katharina Volmer, Katrin Heer, Lars Opgenoorth, Noe Fernandez-Pozo, and Stefan A. Rensing

Epigenomes *(2021), Vol 5, Issue 2, Page 12, https://doi.org/10.3390/epigenomes5020012*

My contribution: The software development and implementation of the pipeline with Nextflow were done by me and Adam Nunn. Abstract of the Python script for missing data imputation with beta distribution was written by Dario Galanti, but major syntax modifications were done by me. Most of the current literature for EWAS studies and tools was screened by me. Implementation of methods with new input types (DMPs and DMRs with both methods), scripts for new graphical outputs such as histograms of p-values, sequence dot plots, Manhattan plots were done by me, Adam Nunn, and Dario Galanti.

Amendments to decrease the running time and disk usage were done by me and Adam Nunn. I prepared all of the figures in the manuscript.

Thesis chapter 3.2.

EpiDiverse Toolkit: a pipeline suite for the analysis of ecological plant epigenetics (Paper-III)

Adam Nunn, Sultan Nilay Can, Christian Otto, Mario Fasold, Bárbara Díez Rodríguez, Noé Fernández Pozo, Stefan A. Rensing, Peter F. Stadler and David Langenberger

NAR Genomics and Bioinformatics *(2021), submitted*

My contribution: I, Adam Nunn, Peter F. Stadler, Christian Otto, David Langenberger, Mario Fasold worked collaboratively to trace software bugs, discuss the structure and implementation of the pipelines with Nextflow. The benchmark for the DMR and the EWAS pipelines was done by me and developed in cooperation with Adam Nunn. *P. nigra* dataset was processed by Bárbara Díez Rodríguez, Adam Nunn, and me for testing the EWAS pipeline. Preparation of figures was done by Adam Nunn and me.

Thesis chapter 3.3

I.II Other scientific publications not contributing to this thesis

"Introduction to Ecological Plant Epigenetics", a free textbook on ecological plant epigenetics.

Adam Nunn, Adrián Contreras Garrido, Anupoma Niloya Troyee, Bárbara Díez Rodríguez, Bhumika Dubay, Cristian Peña, Daniela Ramos-Cruz, Dario Galanti, Iris Sammarco, María Estefanía López, Morgane van Antro, Nilay Can, Paloma Perez-Bello Gil, Panpan Zhang, Samar Fatma

GitBook*, not publicly available yet*

My contribution: I wrote the „Differential Methylation "chapter and revised the „Epigenetics in Evolution "chapter.

II Zusammenfassung

Eines der Ziele der Pflanzen-Epigenetik ist der Nachweis differentieller Methylierung, die nach bestimmten Behandlungen oder in variablen Umgebungen auftreten kann. Dies kann mit einer Einzelbasenauflösung mit Standardmethoden für die Ganzgenom-Bisulfit-Sequenzierung (WGBS) und die Bisulfit-Sequenzierung mit reduzierter Repräsentation (RRBS) erreicht werden. Ein weiteres wichtiges Ziel ist es, Sequenziermethoden in Kombination mit Bisulfit-Behandlung anzuwenden, um Genetik und Epigenetik mit phänotypischen Merkmalen in Verbindung zu bringen. In den letzten 19 Jahren ist dies durch sogenannte genomweite Assoziationsstudien (GWAS) und epigenomweite Assoziationsstudien (EWAS) möglich geworden, wobei Letztere darauf abzielen, die potenziellen Biomarker zwischen phänotypischen Merkmalen und epigenetischer Variation aufzudecken. In der Praxis sind derartige Studien auf Softwarepakete oder "Bioinformatik-Pipelines" angewiesen, die die erforderlichen Rechenprozesse routinemäßig und zuverlässig durchführen. Diese Arbeit beschreibt mehrere solcher Pipelines, die im Rahmen von EpiDiverse, einem Innovative Training Network (ITN) (https://epidiverse.eu/, Zugriff am 1.2.2021), entwickelt wurden, das umfassende Untersuchungen zu Pipelines für WGBS, differenziell methylierte Regionen (DMR), EWAS und Einzelnukleotid-Polymorphismus (SNP)-Analysen ermöglicht.

Hier stelle ich die Benchmark-Untersuchungmit DMR-Tools, die EWAS-Pipeline und Bioinformatik-Pipelines vor, die im EpiDiverse-Toolkit implementiert sind.

Zunächst habe ich gemeinsam mit den Co-Autoren durch die Analyse von DMR-Tools mit simulierten Datensätzen mit sieben verschiedenen Tools (metilene, methylKit, MOABS, DMRcate, Defiant, BSmooth, MethylSig) und vier Pflanzenarten (*Aethionema arabicum*, *Arabidopsis thaliana*, *Picea abies* und *Physcomitrium patens*) gezeigt, dass metilene eine überlegene Performanz in Bezug auf die Gesamtgenauigkeit und den Recall aufweist. Aus diesem Grund haben wir beschlossen, dieses Tool als Standard-DMR-Caller in der EpiDiverse-DMR-Pipeline einzusetzen.

Anschließend führte ich erweiterte Funktionen der EWAS-Pipeline über das GEM R-Paket hinaus ein, wie z.B. grafische Ausgaben, neuartige Imputation fehlender Daten, Kompatibilität mit neuen Eingabetypen usw. Dann deckte ich den Effekt fehlender Daten mit dem Datensatz von *Picea abies* (Fichte) auf und konnte zeigen, dass die Pipeline eine logische Imputation von fehlenden Daten aufweist. Des Weiteren ergab sich eine signifikante Überlappung zwischen den Analyseergebnissen der Pipeline und der *Quercus lobata* (Tal-Eiche).

Durch umfangreichen Benchmark mit verschiedenen Tools wurde eine Gruppe von Pipelines veröffentlicht, wobei sich das EpiDiverse-Toolkit für die Arbeit mit WGBS-Datensätzen eignet (https://github.com/EpiDiverse, Zugriff am 1.2.2021).

III Abstract

One of the goals of plant epigenetics is detecting differential methylation that may occur following specific treatments or in variable environments. This can be achieved with a single-base resolution with standard methods for whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS). Another important goal is to exploit sequencing methods in combination with bisulfite treatment to associate genetics and epigenetics with phenotypic traits. In the past 19 years, this has become possible using so-called genome-wide association studies (GWAS) and epigenome-wide association studies (EWAS), the latter of which aims to reveal the potential biomarkers between phenotypic traits and epigenetic variation. In practice, such studies rely on software packages or "bioinformatics pipelines" which make the requisite computational processes routine and reliable. This thesis describes several such pipelines, developed within the framework of EpiDiverse, an Innovative Training Network (ITN) (https://epidiverse.eu/, accessed on 1 May 2021) carrying out comprehensive studies on pipelines for WGBS, differentially methylated region (DMR), EWAS, and single nucleotide polymorphism (SNP) analyses.

Here I introduce the benchmark study with DMR tools, the EWAS pipeline, and bioinformatics pipelines implemented within the EpiDiverse toolkit.

At first, by analyzing DMR tools with simulated datasets with seven different tools (metilene, methylKit, MOABS, DMRcate, Defiant, BSmooth, MethylSig) and four plant species (*Aethionema arabicum*, *Arabidopsis thaliana*, *Picea abies*, and *Physcomitrium patens*), together with the coauthors, we showed that metilene has a superior performance in terms of overall precision and recall. Therefore, we set it as a default DMR caller in the EpiDiverse DMR pipeline.

Afterward, I introduced extended features of the EWAS pipeline beyond the GEM R package e.g., graphical outputs, novel missing data imputation, compatibility with new input types, etc. Then I revealed the effect of missing data with the *Picea abies* (Norway spruce) data and showed the pipeline presents logical missing data imputation. Furthermore, I obtained a significant overlap between the pipeline and *Quercus lobata* (valley oak) analysis results.

By extensive benchmark with various tools, a group of pipelines became publicly available, whereby the EpiDiverse toolkit suits for people working with WGBS datasets (https://github.com/EpiDiverse, accessed on 1 May 2021).

IV Thesis structure

Chapter 1 ("General introduction") covers a general presentation to my research topics and provides an introduction to the reader to comprehend the basics of epigenetics, bioinformatics, and fundamentals of my research topics.

Chapter 2 ("Questions and objectives") shows the main focus and punchlines of this work.

In **chapter 3** ("Publications"), all publications listed in chapter 1 are summarized as in **chapter 3.1** ("A blind and independent benchmark study for detecting differentially methylated regions in plants") the benchmark of seven DMR callers (metilene [1], methylKit [2], MOABS [3], DMRcate [4], Defiant [5], BSmooth [6], MethylSig [7]) were performed in terms of precision and recall. Metilene overperformed others and used as a default tool in the EpiDiverse DMR pipeline. **Chapter 3.2** ("The EpiDiverse plant Epigenome-Wide Association Studies (EWAS) pipeline"), the EpiDiverse EWAS pipeline was evaluated with two datasets of non-model plant species, namely valley oak (*Q. lobata*) with 58 samples [8] and Norway spruce (*P. abies*) with 28 samples (derived from [9] and unpublished data) to test its reliability and the performance. Finally, **chapter 3.3** ("EpiDiverse Toolkit: a pipeline suite for the analysis of ecological plant epigenetics"), covers all EpiDiverse bioinformatics pipelines to perform an extensive WGBS data analysis.

Chapter 4 ("Concluding remarks") describes significant arguments with the studies outlined in Chapters 2,3, and 4.

Chapter 5 ("Outlook") covers an observation where my main research objectives are treated separately and together. The strengths and weaknesses of the thesis are highlighted and possible amendment ideas for future actions are listed.

Table of Contents

Abbreviations

A. thaliana	*Arabidopsis thaliana*
Ae. arabicum	*Aethionema arabicum*
BS-seq	Bisulfite sequencing
DMPs	Differentially methylated positions
DMRs	Differentially methylated regions
DNMT	DNA methyltransferase
epiRILs	Epigenetic recombinant lines
EWAS	Epigenome-wide association studies
F. vesca	*Fragaria vesca*
FDA	Food and drug administration
FLC	Flowering locus
FNs	False negatives
FPs	False positives
GOA	Gene Ontology Analysis
GWAS	Genome-wide association studies
ITN	Innovative Training Network
LMM	Linear mixed models
MTD	Methyl-transferase domain
P. patens	*Physcomitrium patens*
P. nigra	*Populus nigra*
SNP	Single nucleotide polymorphism
T. arvense	*Thlaspi arvense*
TNs	True negatives
TPs	True positives
TEs	Transposable elements
WGBS	Whole-genome bisulfite sequencing

Chapter 1

General introduction

1 General introduction

1.1 Why study plants?

Plants dominantly produce atmospheric oxygen (O_2) through photosynthesis [10] to sustain all aerobic life on earth. Nearly everything we consume for eating comes directly or indirectly from plants and up to 50% of food and drug administration (FDA) approved drugs during the last 40 years are from plants [11].

We owe a lot of our understanding of genetics and molecular biology to plants. For example, Gregor Mendel who is acknowledged as a founder of modern genetics theory, worked to shed a light on the laws of inheritance with breeding experiments with massive amount pea plants (*Pisum sativum*) [12] and his work remained unpopular and unrecognized till the beginning of 20th century. The term "mutation" was first introduced by Hugo de Vries who relied on Mendel's works and this discovery established the fundamentals of the mutation theory of evolution [13]. If we were to continue with the contribution of plant-based research to genetics and molecular biology, it will not be surprising to mention Barbara McClintock who discovered transposable elements (TEs) with maize [14] and received the 1983 Nobel Prize in Physiology or Medicine. Other remarkable discoveries can be listed in the epigenetics area as an exception to the Mendelian rules.

Epigenetic modifications/mechanisms are an on/off mechanism of genes affecting the appearance and the development of an organism. The realization of this concept is mainly owned by Waddington, who published a study showing the inheritance of a characteristic acquired in a population in reaction to an environmental stimulus in 1956 [15]. Waddington coined the term "epigenetics", and it was used by David Nanney to distinguish between cellular control system types [16]. The outstanding discovery of a naturally occurring epigenetic mutation was revealed in 1999 with toadflax (*Linaria vulgaris*) and scientists keep uncovering the significance of this field since then.

1.2 Working with a model and non-model plant species

Model organisms are species easy to maintain and breed in laboratory conditions to help scientists to understand biological processes. The advantage of a model organism is that it can breed in large numbers with very short generation time therefore several generations can be followed at once. Also, model organisms can be used to generate detailed genetic maps and produce a large number of offspring. The vascular plant *A. thaliana* genome was the first plant genome to be sequenced [17] and presents a widely accepted model organism in plant and crop science. The successful consolidation of *A. thaliana* into different research fields contributed to our understanding of key concepts in biology. Moreover, another species moss *Physcomitrium patens* (*P. patens*), the first bryophyte genome to be sequenced [18] gives ideas about the divergence of the embryophytes (land plants). This process has begun 450 million years ago (Ma) and early diverging lineages as bryophytes, such as hornworts, mosses, and liverworts helped to rebuild the evolutionary structure and for us to understand the conquest of land by the plant [18].

However, the new technologies for genome sequencing helped to get partial (transcriptome) or complete (whole genome sequencing) DNA sequences for non-model plant species too. Therefore, non-model plants became very attractive to study adaptation to extreme environmental conditions and metabolites that can be used in the food or medical industry. The study of non-model plant genomics may reveal the biochemical pathways and genetic factors which contributed to flowering, biotic or abiotic stress tolerance, and pest resistance. Revealing the mechanisms involved during the survival of plants under extreme conditions may give clues about plant response mechanisms.

1.3 Epigenetics: mechanisms, concepts, and effects on plants

Several different molecular elements seem to be involved in the determination of how gene sequences behave and act, whether in classical genetics or in a way that highly deviates from this area. The epigenetic regulatory system that deviates from the typical Mendelian sense is not necessarily novel, besides the set of molecular mechanisms cannot be reduced to a simple universal 'code'. Epigenetics involves genetic control with DNA or histone modifications (acetylation, phosphorylation, ubiquitylation, sumoylation) which might change gene activity without altering an individual's DNA sequence and accordingly the accessibility of genetic information [19] Figure 1. In particular, DNA methylation is a covalent modification in which a methyl group (CH_3) is added to the fifth carbon of the cytosine. Furthermore, DNA methylation in particular is the most frequently studied mechanism in plants since it can easily be accessed. DNA methylation can silence genes and/or repetitive elements through a process that changes the chromatin structure therefore its accessibility [20]. DNA methylation allows these silenced states to be inherited via cellular divisions.

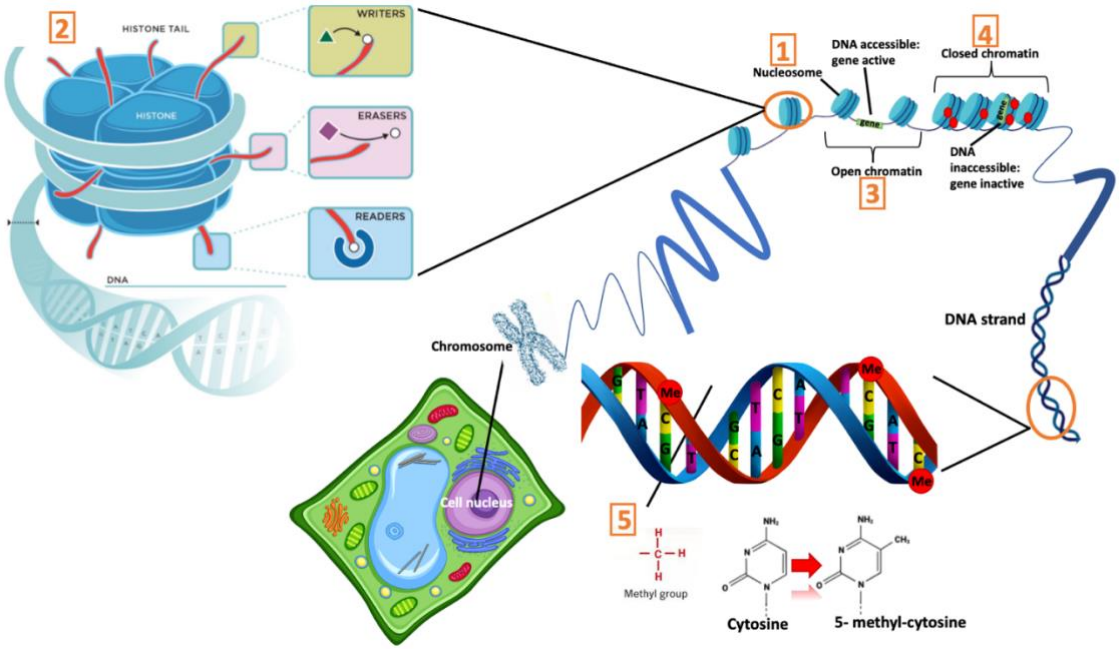


Figure 1: **Epigenetic modifications.** A nucleosome is a basic structural unit of DNA (1) consists of DNA wound around eight proteins called histones (2) and looks like thread wrapped around a spool. Histone modifications such as acetylation, methylation, and phosphorylation are added to histones by enzymes called "writers", removed by enzymes called "erasers" (2). Other proteins called "readers" do not change the histone structure but identify a specific modification pattern by binding them or including additional proteins to regulate gene activity (2). When chromatin is unmethylated and accessible (euchromatin), then it is active for transcription (3). When chromatin is methylated and in closed form (heterochromatin), it is inactive for transcription (4). A methyl group (Me) is an organic compound that consists of one carbon atom bonded to three hydrogen atoms (CH_3). When methyl group attaches to the Cytosine, DNA methylation occurs, and this modification may change the gene activity (5). (Image in (2) was taken from www.zenithepigenetics.com, copyright of Richard E. Ballermann).

DNA methylation variation can affect phenotypes [21] by regulating gene expression and can be shaped by environmental variation, arise stochastically or it can be driven by genetic variants traits [22]. Importantly, DNA methylation as one of the most powerful factors may cause heritable epigenetic information, and this means a lack of resetting of epigenetic marks between generations. This type of inheritance is more prominent in plants compared to animals. Epialleles are specific DNA methylation patterns of a genetic locus among many examples of the aforementioned transgenerational inheritance [23]. For instance, in *Linaria vulgaris* (toadflax), naturally occurring mutant alleles were linked together with the revertant phenotype and showed high DNA methylation at the Lcyc locus but low gene expression [24]. In addition to this example, Cnr mutants in tomato showed colorless, non-ripening fruits caused by the epiallele at the LeSPL–CNR locus and mutant phenotypes showed increased DNA methylation at the promoter region in another study [25]. Many other epialleles had been defined in the model organism *A. thaliana* and they all seem to be in TEs or other repetitive sequences [26]. The association between DNA methylation and the repression of TEs possibly affects short- or long-term adaptation to changing environmental conditions [27-30]. Li et al., 2018 showed that 2,311 TE loci seem to be a target of positive selection and contributed to the adaptation of *A. thaliana* [31]. Another example of adaptation with the abiotic stress response and vernalization is also found to be affected by histone modifications. For example, vernalization in *A. thaliana* is regulated by flowering locus (FLC) when cold triggers H3K27me3 and H3K9me deposition in the FLC chromatin [32].

DNA methylation occurs in different nucleotide contexts. In animals, mostly Cs in CG contexts are methylated but there are rare exceptions with this [33]. On the contrary, DNA methylation in plants can be symmetrical (CG and CHG contexts) or asymmetrical (CHH, where H represents A, T, or C) [34-37]. Methylation in each sequence context is maintained through cell divisions by different enzymatic machinery [38] Figure 2.

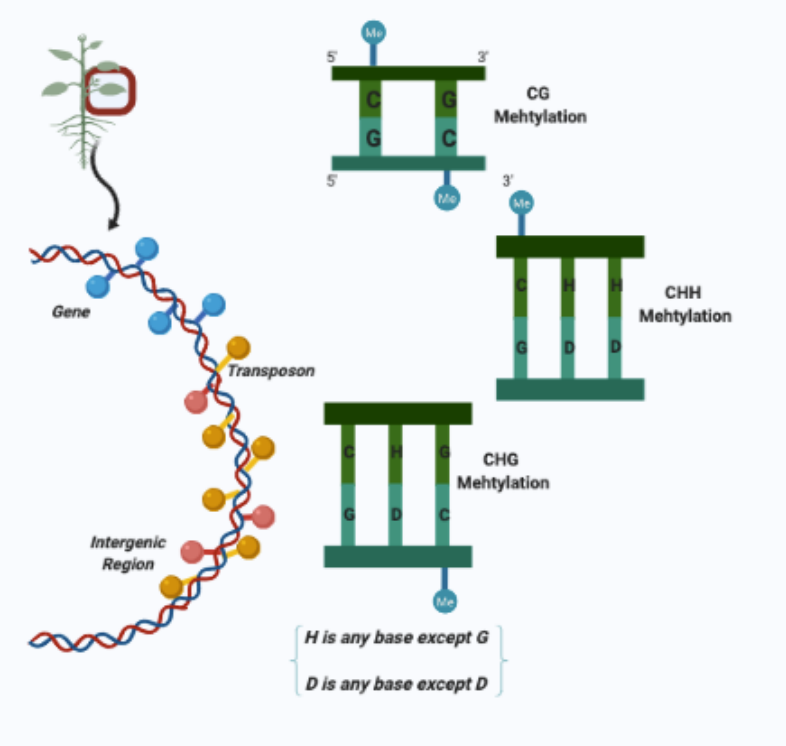


Figure 2: **CG, CHG, and CHH context DNA methylations.** DNA methylation can be in symmetrical (CG and CHG) or asymmetrical (CHH) contexts in plants whereby it can establish or enhance the epigenetic silencing of genes, intergenic regions, and transposons. Regulation of these contexts is maintained by four DNA methyltransferase (DNMT) families that share a conserved methyl-transferase domain (MTD). Created with BioRender.com.

Whole-genome bisulfite sequencing (WGBS) [39] and reduced representation of bisulfite sequencing (RRBS) [40] incorporate bisulfite conversion with high-throughput sequencing and determine DNA methylation at a single base resolution. Bisulfite conversion is a process where genomic DNA is denatured, and sodium bisulfite is applied to cause deamination of unmethylated cytosines into uracils while keeping methylated cytosines unchanged. One of the crucial goals in epigenetics is to detect differentially methylated positions (DMPs) and regions (DMRs) arising from different treatments or environments [41] which are established between different individuals. DMPs can be called with various statistical approaches. DMRs are named genomic regions where multiple adjacent positions show differential methylation [42]. Many types of DMR callers can be preferred depending on the experimental design. Some of them are listed on a timeline basis in Figure 3.

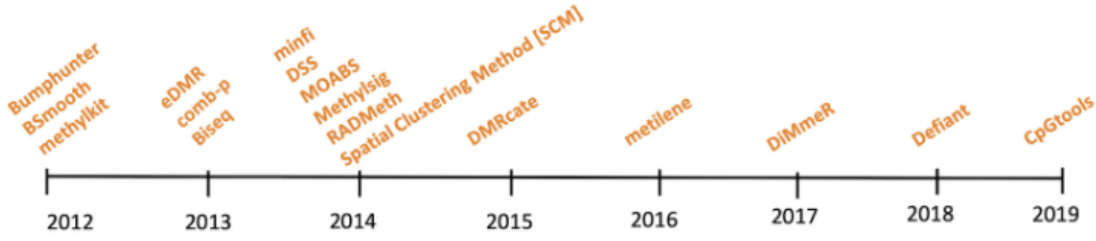


Figure 3: **Timeline and strategies of DMR callers.** Some widely used statistical approaches to call differential methylation were listed on the timeline for DMP and DMR calling.

1.4 Association studies to examine genetic and epigenetic modifications

Scientists developed association studies to understand the background of complex traits to associate genetic and epigenetic variants with phenotypic traits. They tested genetic variation across the genome between individuals to show potential genotype-genotype association by GWAS which have been widely used with complex human studies [43, 44]. GWAS has become a powerful tool also for plants to study complex and important traits in agriculture [45]. It has been used in many crop species such as maize (*Zea mays*), wheat (*Triticum aestivum*), rice (*Oryza sativa*), soybean (*Glycine max*), sorghum (*Sorghum bicolor*), barley (*Hordeum vulgare*), cotton (*Gossypium hirsutum*), and the model species thale cress (*Arabidopsis thaliana*) [46-48]. GWAS was also used with genetic engineering studies e.g., transgenic drought tolerance in maize was developed after the discovery of ZmVPP1 [49] which encouraged researchers to prefer genome-editing studies [50]. Furthermore, GWAS has been used to disclose genomic regions related to physiological, agronomic, and fitness traits such as plant height, stress tolerance, flowering time, kernel number, and grain yield [46, 47, 51]. However, studies showed that many human diseases including cancer showed epigenetic connection [46, 47, 52], as a result EWAS has emerged as a counterpart of GWAS [53]. EWAS is an effective type of analysis to uncover the association between epigenetics and biological traits [42]. Epigenetic cannot be thought of independently of genetics and transcriptomics. Observing the effect of epigenetics on phenotypes can help us to understand the basis of that phenotype with other association studies such as eQTL, meQTL, and TWAS. Expression quantitative trait loci (eQTL) are genetic variants that have an impact on gene expression levels. Methylation quantitative trait loci (meQTL) are genetic variants that affect DNA methylation patterns. Transcriptome-wide association study (TWAS) integrates GWAS and gene expression for finding gene and trait relationships. So, a combination of indirect relations of all these with EWAS can increase the power of meta-analysis (Figure 4).

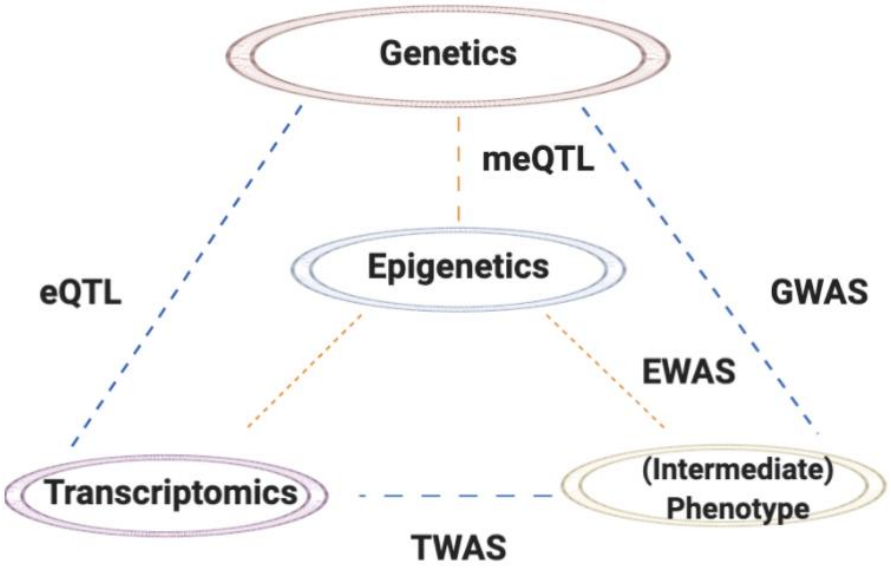


Figure 4: **Relationship between omics and association studies between epigenetics.** The epigenetic connection has to be made with sequence variants in genetics and changes in transcriptomics. Metabolomics and proteomics may also be helpful to reveal the impact of epigenetics on intermediate phenotype. For example, quantitative trait loci (eQTL) could be used to reveal the genetic variants have an effect on gene expression levels on the genome. One can prefer methylation quantitative trait loci (meQTL) to reveal an association between genetic and epigenetic variants. Transcriptome-wide association studies (TWAS) is for studying the gene expression levels with phenotypic traits. An epigenome-wide association studies (EWAS) is a research of association between epigenetic marks and a particular phenotype in different individuals. Genome-wide association studies (GWAS) is a study to reveal a set of genetic variants associated with a trait. The assembly and interaction of these studies can yield a great power to understand meta-analysis. Figure modified from Cazaly, E et al., 2019 and created with BioRender.com.

It has been shown that transgenerational epigenetic marks could be transmitted to offspring through mitosis with vegetative propagation, and meiosis with sexual reproduction [54]. As already mentioned, *Linaria vulgaris* mutants can be a good example of transgenerational epigenetic inheritance [24]. Germline in plants is inherited from somatic cells and therefore can support the heritability of epigenetic marks. Given their sessile lifestyle, plants have a higher sensitivity to detect environmental changes, which may cause epigenetic changes in cell lines that generate a germline [27]. Some studies showed that stress-related DNA methylation in Arabidopsis can be transgenerational [55, 56]. So, heritable epialleles may influence plant phenotypic traits, evolution, and fitness. Conclusively, we can assume that epigenetic variation might have a bigger impact on clonal than sexual plants. Clonal plants, by bypassing meiosis, circumvent the huge epigenetic resetting that occurs during this stage. Thus, it has been proposed that the inheritance of epigenetic marks across clonal generations is higher than sexual generations, and that epigenetic variation might shape the evolutionary trajectory of clonal plants even more than sexual plants [57].

Epigenetic changes are dynamic and bring causality or consequence problems between phenotype and epigenetic mechanisms which is a major challenge of EWAS [58]. Epigenetic recombinant lines (epiRILs) usage may overcome this issue where individuals have the same genetic background [59] and the association between epigenetics and phenotypic traits can be ensured. Other common issues shared by GWAS and EWAS can be listed as missing data and/or huge raw datasets [60]. Missing data hitch can be solved by a robust estimation/imputation with the beta distribution that explains the characteristic of DNA

methylation distribution in the genome [61]. A huge dataset problem can be solved by splitting data per chromosome/scaffold that may also help to reduce running time and disk space usage during analyses.

Unfortunately, there are limited studies with plants using EWAS (for example, a PubMed search for "ewas plant" returned 14 on 26 April 2021, while "ewas human" returned 403 hits). To set an example EWAS was used to associate climatic and spatial variables with DNA methylation in *Quercus lobata* (valley oak) [8] and was successfully applied to *Elaeis guineensis* (oil palm) to identify epigenetic changes that cause the different phenotype [62]. Another highlighted study with vegetatively propagated *Pinus pinea* (stone pine) trees showed that a high degree of DNA methylation is associated with the different levels of phenotypic plasticity [63]. Various EWAS tools were implemented to reveal epigenetics, genetics, and phenotypic trait associations however many of them cannot do missing data imputation such as EWAS: epigenome-wide association study software v2.0 [64] or hard-coded with human array-like GLINT [65]. But for example, an R package called GEM [66] is compatible with all species and not hardcoded for the specific ones and allows to include a genetic variation with missing data imputation. As a result, EWAS is a promising field to release further information, in addition to that, the amount of research is novel and narrow in scope. These criteria had awakened the interest to investigate these methods in this dissertation.

1.5 Bioinformatics pipelines

Pipelines in bioinformatics describe a set of processing steps to transfer raw data into something interpretable. In other words, high-throughput bioinformatics pipeline frameworks convert raw data into the desired format and enable the analysis of meta and sequence data. A workflow engine is a software that could be used to execute a pipeline or a workflow. Scripting languages such as Bash (https://www.gnu.org/software/bash/, accessed on 1 May 2021), Python (https://www.python.org/, accessed on 1 May 2021), Perl (https://www.perl.org/, accessed on 1 May 2021), and R (https://www.r-project.org/, accessed on 1 May 2021) can be preferred to implement workflows and pipelines. Nextflow (https://www.nextflow.io/, accessed on 1 May 2021) is one of the most famous workflow engines that facilitates or provides a modern, scalable, portable, and reproducible environment to develop pipelines in a user-friendly way [67].

1.6 EpiDiverse consortium and research project number five (RP05)

EpiDiverse (https://epidiverse.eu/en, accessed on 1 May 2021) is a Marie Skłodowska-Curie ITN that aims at the study of epigenetic variation in wild, non-model plant species such as *Fragaria vesca* (wild strawberry), *Populus nigra* (black poplar), and *Thlaspi arvense* (field pennycress). The network brings research groups together from ecology, molecular (epi)genetics, and bioinformatics to explore the genomic basis, molecular mechanisms, and ecological significance of epigenetic variation in natural plant populations. My research project (RP05) focuses on state-of-the-art bioinformatics and epigenetics methods to develop a set of pipelines and tools for epigenomics data analysis in non-model plants. The pipelines implemented under this project aim to understand the genetic and epigenetic changes happening in plants from different natural populations. This project also aimed to

answer how climatic/phenotypic variables associate with DNA methylation. Moreover, collaborative research projects 6,7, and 8 study how the inheritance process is carried through generations via sexual and asexual reproduction and how they differ in annual (*T. arvense*) and perennial species (*P. nigra* and *F. vesca*).

Chapter 2

Questions and objectives

2 Questions and objectives

The main focus of this work was to develop a set of pipelines for non-model plant species for the EpiDiverse consortium to analyze WGBS data. The first step was a benchmark study with DMR tools and implementation of the EpiDiverse EWAS pipeline followed by pipelines in the EpiDiverse toolkit.

The DMR benchmark tries to suggest an optimal statistical method to call DMRs for chosen species from several groups of plants such as bryophytes, gymnosperms, and angiosperms with different levels of replication and sequencing depth.

The EpiDiverse EWAS pipeline aims to be compatible with WGBS data for all species and allows users to analyze multiple input types with methylation calls including differential methylation. A novel method with DMRs input can yield interesting results for people to focus on specific regions associated with phenotypic traits. The EpiDiverse EWAS pipeline serves a novel, trustworthy missing data imputation, comprehensive graphical outputs to observe results better, and splits data per chromosome to decrease running time and space usage.

The EpiDiverse toolkit aimed and succeeded to be convenient, user-friendly, and includes an open-access set of pipelines not only for plant ecologists but also for other scientists to work with WGBS data. Therefore, the implemented pipelines enable users to perform multiple analyses such as mapping the bisulfite sequencing (BS-seq) data with the WGBS pipeline, calling differential methylation by the DMR pipeline. In other respects, performing EWAS with the EWAS pipeline, calling variants, and discriminating between genetic and epigenetic variation with a novel algorithm using the SNP pipeline are also possible.

Chapter 3

Publications

3 Publications

3.1 A blind and independent benchmark study for detecting differentially methylated regions in plants (Paper-I)

BS-seq is a widely used and state-of-the-art technique for mapping methylation and one of the aims of plant epigenetics is to reveal differential methylation. Differential methylation may take place due to environmental changes or specific treatments. This study aimed to compare the performances of seven algorithms (metilene [1], methylKit [2], MOABS [3], DMRcate [4], Defiant [5], BSmooth [6], MethylSig [7]) for calling DMRs with four plant species namely *Ae. Arabicum* (angiosperm), *A. thaliana* (angiosperm), *P. abies* (gymnosperm), and *P. patens* (bryophytes) using simulated datasets based on experimental BS-seq data. An independent design was set to reduce the potential bias. It was possible to generate a decision tree to select the optimal algorithm based on the data structure. Metilene showed outstanding performance and therefore we suggest the usage of metilene as a default approach.

3.1.1 Paper

Following is the electronic publication.

Bioinformatics, 36(11), 2020, 3314–3321
doi: 10.1093/bioinformatics/btaa191
Advance Access Publication Date: 17 March 2020
Original Paper

Genome analysis

A blind and independent benchmark study for detecting differentially methylated regions in plants

Clemens Kreutz[1,2,*], Nilay S. Can[3], Ralf Schulze Bruening[3], Rabea Meyberg[3], Zsuzsanna Mérai[4], Noe Fernandez-Pozo[3] and Stefan A. Rensing[3,5]

[1]Faculty of Medicine and Medical Center, Institute of Medical Biometry and Statistics, University of Freiburg, 79104 Freiburg, Germany, [2]Centre for Integrative Biological Signalling Studies (CIBSS), University of Freiburg, 79104 Freiburg, Germany, [3]Plant Cell Biology, Faculty of Biology, University of Marburg, 35043 Marburg, Germany, [4]Gregor Mendel Institute of Molecular Plant Biology, Austrian Academy of Sciences, Vienna BioCenter (VBC), 1030 Vienna, Austria and [5]Centre for Biological Signaling Studies (BIOSS), University of Freiburg, 79104 Freiburg, Germany

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on August 20, 2019; revised on January 31, 2020; editorial decision on March 12, 2020; accepted on March 13, 2020

Abstract

Motivation: Bisulfite sequencing (BS-seq) is a state-of-the-art technique for investigating methylation of the DNA to gain insights into the epigenetic regulation. Several algorithms have been published for identification of differentially methylated regions (DMRs). However, the performances of the individual methods remain unclear and it is difficult to optimally select an algorithm in application settings.

Results: We analyzed BS-seq data from four plants covering three taxonomic groups. We first characterized the data using multiple summary statistics describing methylation levels, coverage, noise, as well as frequencies, magnitudes and lengths of methylated regions. Then, simulated datasets with most similar characteristics to real experimental data were created. Seven different algorithms (*metilene*, *methylKit*, *MOABS*, *DMRcate*, *Defiant*, *BSmooth*, *MethylSig*) for DMR identification were applied and their performances were assessed. A blind and independent study design was chosen to reduce bias and to derive practical method selection guidelines. Overall, *metilene* had superior performance in most settings. Data attributes, such as coverage and spread of the DMR lengths, were found to be useful for selecting the best method for DMR detection. A decision tree to select the optimal approach based on these data attributes is provided. The presented procedure might serve as a general strategy for deriving algorithm selection rules tailored to demands in specific application settings.

Availability and implementation: Scripts that were used for the analyses and that can be used for prediction of the optimal algorithm are provided at https://github.com/kreutz-lab/DMR-DecisionTree. Simulated and experimental data are available at https://doi.org/10.6084/m9.figshare.11619045.

Contact: ckreutz@imbi.uni-freiburg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A large number of methods and algorithms have been established during recent years for analyzing high-throughput data and often several approaches exist for the same task. Because the performance of such competing algorithms is usually context-specific (Shippy *et al.*, 2006; Su *et al.*, 2014; Webb-Robertson *et al.*, 2015), i.e. depends on characteristics of the investigated biological system as well as on the amount and quality of the data, clear and general rules for selecting the most suitable algorithm in practice are missing. Therefore, benchmark studies for deriving such guidelines are essential. In this study, we compared multiple tools to detect differentially methylated regions (DMRs) in plant bisulfite sequencing data (BS-seq, also known as Bisulfite-Seq or Methyl-seq).

DNA methylation can have a remarkable effect on gene expression and cellular activity and analyzing it via deep sequencing is an important approach to study its impact on transcription. BS-seq is the most common method to study DNA methylation to nucleotide-level. BS-seq can be combined with long-read sequencing for real-time bisulfite sequencing (Yang and Scott, 2017). Other methods, such as ChIP-seq exists to study histone modifications (Chen *et al.*, 2018), and novel methods, such as TET-assisted pyridine borane sequencing, could be the future for investigation of

DNA methylation at nucleotide level since it is a cheaper and more reliable alternative to BS-seq (Liu et al., 2019). In BS-seq, to identify the DNA methylation status, DNA samples are treated with sodium bisulfite before sequencing. This way, methylated cytosines (including 5-Methylcytosine and 5-Hydroxymethylcytosine) remain unchanged, whereas unmethylated cytosines are converted to uracils.

For the detection of DMRs, many different approaches are available (Robinson *et al.*, 2014), and in our application setting, it was unclear how to optimally choose from these competing methods. Existing studies provide a heterogeneous and inconsistent picture, with limited information for the usage on plant BS-seq data where, in contrast to e.g. mammals, DNA methylation can be observed not only in *CG* context but also in *CHG* and *CHH* contexts (H = A, C or T) (Sahu *et al.*, 2013). In addition, DMR algorithms tested in plants, are often limited to *Arabidopsis thaliana*, but DNA methyltransferases, genome size, methylation patterns and transposable elements differ between plant species (Bewick and Schmitz, 2017; Lang *et al.*, 2008). Thus, studying multiple species is important.

Based on recent DMR benchmark studies it is not evident which algorithms are most suitable for DMR detection in plants. For example, Jühling *et al.* (2016) introduced *metilene* and compared it with three other DMR algorithms: *MOABS* (Sun *et al.*, 2014), *BSmooth* (Hansen *et al.*, 2012) and *BiSeq* (Hebestreit *et al.*, 2013) using experimental datasets from human and simulated datasets based on human Chromosome 10. *Metilene* showed superior specificity, sensitivity and accuracy and particularly outperformed the others when the methylation background was variant and the methylation differences were subtle. A similar benchmark is done with another tool, *Defiant* (Condon *et al.*, 2018) comparing it with six other DMR tools: *BSmooth*, *methylKit* (Akalin *et al.*, 2012), *MethylSig* (Park *et al.*, 2014), *metilene*, *MOABS* and *RADMeth* (Dolzhenko and Smith, 2014), using the simulated data from Jühling *et al.* (2016) and experimental BS-seq data from rat. *Defiant* showed comparable performance with metilene for simulated data and identified more DMRs in the experimental dataset; however, several low coverage DMRs were only detected by metilene. In Gong and Purdom (2018), *MethCP* was presented in comparison with *BSmooth*, *HMM-Fisher*, *DSS*, *methylKit* and *metilene* in an analysis of simulated datasets. Here, *MethCP* and *metilene* showed the highest rate of true positives among all the tools tested. In Catoni *et al.* (2018), experimental data from human cell lines, rice and *A.thaliana* were analyzed to assess the performance of *DMRcaller* in comparison with *methylKit* and *MethylSig*, resulting in better performance of *DMRcaller*. In all of the studies aforementioned, superior performance of the newly presented approach was concluded, presenting only partially consistent outcomes as compared with previous studies.

The heterogeneity of the outcomes might originate from the fact that different types of data with distinct characteristics were used for assessment. Most DMR analyses are based on human or mammals and on the *CG* methylation context, consequently ignoring plant methylation contexts *CHH* and *CHG*. Only two of the previous benchmark studies (Catoni *et al.*, 2018; Gong and Purdom, 2018) partly focused on DMRs in plants. However, the designs of both studies were not based on typical data and they do not assess the tradeoff between true positives and false positives. In Catoni *et al.* (2018), wild-type was compared with knockouts of methyltransferases. These knockouts are known to prevent methylation and, therefore, only sensitivity could be evaluated by treating all DMR predictions exclusively as true positives. In Gong and Purdom (2018), random permutations are applied to generate uninformative data and then only specificity was assessed by treating all predictions as false positives.

Different plant groups differ in their DNA methyltransferase complements and context-specific DNA methylation (Bewick *et al.*, 2017). Here, we include data of the prime plant model, *A.thaliana*, and of a relative from the same order (Brassicales), *Aethionema arabicum*, representing low coverage data (Table 1). We also include data of *Picea abies*, in which methylation frequencies are different (Heer *et al.*, 2018) and of a moss, *Physcomitrella patens*, which again shows differences, namely different properties of gene body methylation (Lang *et al.*, 2018).

In this study, we, therefore, intended to analyze realistically simulated data including multiple plant species from several plant groups, i.e. a bryophyte (*patens*), a gymnosperm (*P.abies*) and two angiosperms (*Ae.arabicum* and *A.thaliana*). All of these datasets are again different from those of mammals, in both terms, methylation context (including CHG and CHH contexts) and distribution of methyl groups. As described in the following, we took experimental data from these plants as templates and calibrated a simulation framework for generating data with attributes most similar to our experimental datasets. We evaluated the performance of *BSmooth*,

Table 1. Overview about the experimental BS-seq datasets

	Abbreviation	Description	DNA context	Analyzed region	Analyzed range (bp)	Analyzed data points	Coverage
Data context	Aetar-C-CG	*Ae.arabicum*, Cyprus ecotype, seeded at 20°C	CG	Scaffold 65	2 720 010	29 356	3.11
	Aetar-C-CHG	*Ae.arabicum*, Cyprus ecotype, seeded at 20°C	CHG	Scaffold 65	2 725 648	48 748	2.29
	Aetar-C-CHH	*Ae.arabicum*, Cyprus ecotype, seeded at 20°C	CHH	Scaffold 65	2 728 141	214 994	2.11
	Aetar-c-CG	*Ae.arabicum*, Cyprus ecotype, seeded at 25°C	CG	Scaffold 65	2 720 259	27 989	3.26
	Aetar-T-CHG	*Ae.arabicum*, Turkey ecotype, seeded at 20°C	CHG	Scaffold 65	2 730 515	59 415	2.60
	Aetar-T-CHH	*Ae.arabicum*, Turkey ecotype, seeded at 20°C	CHH	Scaffold 65	2 731 240	266 096	2.36
	Arath-CG	*A.thaliana*, root tissue	CG	Chr 2	19 696 777	903 767	34.72
	Arath-CHG	*A.thaliana*, root tissue	CHG	Chr 2	19 696 771	977 493	33.96
	Arath-CHH	*A.thaliana*, root tissue	CHH	Chr 2	3 952 731	1 000 000	31.36
	Picab-M-GG	*P.abies*, 500 m above sea level	CG	50 scaffolds	2 822 478	33 101	41.38
	Picab-M-CHG	*P.abies*, 500 m above sea level	CHG	26 scaffolds	1 554 819	61 918	21.00
	Picab-M-CHH	*P.abies*, 500 m above sea level	CHH	3 scaffolds	243 324	61 918	14.53
	Picab-P-GG	*P.abies*, 1200 m above sea level	CG	50 scaffolds	2 822 266	32 967	30.86
	Picab-P-CHG	*P.abies*, 1200 m above sea level	CHG	26 scaffolds	1 586 240	61 918	17.01
	Picab-P-CHH	*P.abies*, 1200 m above sea level	CHH	3 scaffolds	246 862	61 918	11.90
	Phypa-G-CG	*patens*, Gransden ecotype	CG	Chr 27	5 294 665	182 243	11.32
	Phypa-G-CHG	*patens*, Gransden ecotype	CHG	Chr 27	5 295 410	200 348	10.74
	Phypa-G-CHH	*patens*, Gransden ecotype	CHH	Chr 27	5 298 384	1 347 668	8.02
	Phypa-R-CG	*patens*, Reute ecotype	CG	Chr 27	5 294 552	178 799	12.12
	Phypa-R-CHG	*patens*, Reute ecotype	CHG	Chr 27	5 295 374	197 297	12.25
	Phypa-R-CHH	*patens*, Reute ecotype	CHH	Chr 27	5 298 334	1 330 264	8.67

Note: Representative genomic regions were analyzed to characterize attributes of the data and for optimizing the simulation parameters. The provided coverage is the average number of reads of the analyzed genomic positions. For each data context, 1 000 000 data points were simulated.

Defiant, *DMRcate* (Peters *et al.*, 2015), *metilene*, *methylKit*, *MethylSig* and *MOABS*. Other tools, such as BiSeq (Hebestreit *et al.*, 2013), *BEAT* (Akman *et al.*, 2014), *DSS* (Feng *et al.*, 2014), *RnBeads* (Assenov *et al.*, 2014), *M3D* (Mayo *et al.*, 2015), were not included because it has been claimed that they cannot handle methylation in non-*CG* context (Catoni *et al.*, 2018).

Recently, neutral benchmark studies that do not focus on introducing a new method and guarantee that the authors have no preferences have been demanded (Boulesteix *et al.*, 2017). Moreover, knowledge about the underlying truth provides information for the definition of configuration parameters that is not available in real experimental setting and thus might lead to overoptimistic assessment. To avoid such biases, we chose a blinded study designs, i.e. data simulations and assessments were strictly separated from the DMR analyses. Based on this blinded study design, we derived a decision tree that can serve as algorithm selection rule in new applications.

2 Materials and methods

BS-seq data of four plant species from different taxonomic groups were used to create simulated realistic datasets to evaluate the performance of seven DMR prediction tools. Conditions and characteristics of the methylation data from these experimental samples are summarized in Table 1.

2.1 Methylation data

For *Ae.arabicum* (Brassicaceae), representing one of the angiosperm species from this study, dried seed samples of two different ecotypes (Turkey and Cyprus) were used to create simulated datasets (for sample preparation see Supplementary information). *Arabidopsis thaliana*, another Brassicaceae, was selected to be analyzed in this study as the most well-studied plant model. To create the simulated datasets for *A.thaliana*, methylation call files from bisulfite-treated root samples from Seymour *et al.* (2014) were used. Data from *P.abies* were selected to represent gymnosperms. Methylation calls from Heer *et al.* (2018) of trees situated at about 1200 meters above sea level (MASL) and clones from those trees, planted 24 years ago at 520 MASL were used to create the simulated data. *patens* was selected as a model plant to represent bryophytes. Simulated datasets were generated using methylation calls derived from adult gametophores of two different ecotypes, Gransden (Lang *et al.*, 2018) and Reute (Meyberg *et al.*, 2019). Twenty-one BS-seq datasets of four plant species in three methylation contexts (CG, CHG and CHH) were selected for benchmarking. These datasets differ in their sequencing depth, from very low (<5×) in *Ae.arabicum*, low (10×) in *patens*, medium (10–20×) in *P.abies* and high (30×) in *A.thaliana* (Table 1). The raw data from *A.thaliana*, *P.abies* and *patens* are available in the Sequence Read Archive with accession numbers: PRJEB6701, PRJEB26494, SRR4454535 and SRR9901085. The scaffold 65 data used for *Ae.arabicum* is available at https://doi.org/10.6084/m9.figshare.11619045.

2.2 DMR algorithm configuration

BSmooth v1.18, *Defiant* v1.1.3, *DMRcate* v1.18, *methylKit* v1.8.1, *MethylSig* v0.5.2, *metilene* v0.2.7 and *MOABS* v1.3.2 were used to predict DMR for all simulated datasets. Tools were used with default parameters and when possible, *P*-value thresholds were set to 0.05, the minimum methylation difference was set to 10%, the window size was set to 200 bp for tools using windows to calculate DMRs, and the minimum coverage was set to 3 for all tools and datasets, with the following exceptions. *Metilene* has no option to filter by minimum coverage, so minimum coverage was 0. Defiant was not able to produce results for *Ae.arabicum* datasets, with very low coverage, so the minimum coverage was set to 0 in these cases and 3 was used for all other cases.

A modified version of *metilene* v0.2.7 was provided by the developers to be able to process *CHH* and *CHG* contexts. This version included a parameter (G) for the chunk size to prevent software crashes due to the high frequency of *CHH* and *CHG* patterns. The chunk size was set to 200 for all analyses using *metilene* for the contexts *CHH* and *CHG*. The chunk size parameter was included in *metilene* v0.2.8.

2.3 Simulating data

Simulated data were generated and utilized to assess the performances of seven DMR identification methods. For data simulation, we utilized the *WGBSSuite*, a software package that has been developed for simulated whole genome bisulfite sequencing data (Rackham *et al.*, 2015).

The *WGBSSuite* simulates the data by the following procedure. In a first step, a discrete-state Markov model is used to subdivide genomic regions into *CG islands* and deserts. The 2×2 transition matrix T has two free parameters T_{12} for the transition probabilities from desert to islands and T_{21} from *CG islands* to deserts. The other two entries of the transition matrix T_{11} and T_{22} are given by normalization, i.e. $T_{11} = 1 - T_{12}$ and $T_{22} = 1 - T_{21}$. Each simulation is initialized in the desert state.

In the second step, the methylation sites are drawn from exponential distributions. The distance of subsequent sites simulated according to $\Delta\text{pos} = \max\{1, \text{round}(\Delta x)\}$ with $\text{Prob}(\Delta x) = (1/R)e^{-\Delta x/R}$ for $\Delta x > 0$. In the classical setting, these sites correspond to consecutive *CG* locations. For our setup in plants, the sites may also correspond to *CHG* or *CHH*, depending on the analyzed context. Different densities for *CG islands* and deserts are defined by two rates $R = R_{\text{island}}$ or $R = R_{\text{desert}}$. The larger these rates, the closer are the simulated positions of the methylation sites.

As a third step, the methylation status is simulated. Here, a non-homogeneous HMM is used and the transitions between states are modulated by the distances of the methylation sites. The simulation suite describes methylation via four states, an unmethylated state S_1, a transition state from unmethylated to methylated S_2, a methylated state S_3 and a transition state S_4 from methylated to unmethylated. There is a 4×4 transition matrix Π that does only allow transitions in plausible order $S_1 \to S_2 \to S_3 \to S_4$ and, therefore, has four off-diagonal elements $\Pi_{12}, \Pi_{23}, \Pi_{34}, \Pi_{41}$. In our analyses, the transition rates from the intermediate states S_2 and S_4 were fixed to the default value $\Pi_{23} = \Pi_{41} = 0.02$, the two additional off-diagonal elements Π_{12} and Π_{34} controlling transitions, i.e. the lengths of the methylated and unmethylated regions, were estimated from the experimental data as discussed in the following section. This four-state HMM describes the methylation changes between consecutive genomic sites. The dependency on the distances of consecutive sites is included by a decay of the probabilities for remaining in a state, i.e. the decrease in diagonal elements is multiplied with a factor e^{-d} with decay parameter d.

All intervals that are assigned to the methylated state feature a joint simulation parameter P_{meth} for the probability of observing a methylated read by sequencing. Unmethylated regions have an analogous parameter P_{un}. Unmethylated regions have probability P_{un} close to zero, methylated regions have P_{meth} close to one. The probabilities for the intermediate states S_2 and S_4 are given by the arithmetic mean $0.5P_{\text{un}} + 0.5P_{\text{meth}}$ in the *WGBSSuite*. Moreover, there are two so-called 'error' parameters E_{un} and E_{meth} that control fluctuating probabilities for the measurements. Again, the error parameters for the intermediate states S_2 and S_4 are calculated by the arithmetic mean. Similarly, there are two parameters for the mean number of reads in methylated and unmethylated regions termed M_{meth} and M_{un}. The mean number of reads for the intermediate states is again given by the average of both parameters.

For simulating differential methylation, a fraction F of the intervals with a specific methylation state are drawn and then assumed as differentially methylated. Then, the direction of differential methylation is drawn according to a so-called 'balance' parameter (Rackham *et al.*, 2015). The balance parameter was set to 0.5 which corresponds to equal probabilities for up- and downregulation. All simulation parameters are summarized in Supplementary Tables S1 and S2. The magnitudes of the differential methylation levels, i.e. the difference of the probabilities P_{meth} or P_{un} are used as effect size that is evaluated over the whole range for assessing the power of the approaches.

2.4 Optimizing simulation parameters

To simulate the data realistically, the simulation parameters of the *WGBSSuite* were calibrated to our experimental datasets. For quantifying similarity of simulated and experimental data, we characterized datasets by calculating several attributes. At this point, it is important to characterize methylated as well as nonmethylated regions independently. To be able to discriminate regions with low methylation levels from high levels, a custom smoothing approach was applied as described in the Supplementary information.

For regions with high methylation levels that are termed presumably methylated (PM) in the following, and for regions that are presumably unmethylated (PU), 16 attributes were calculated:

1. Mean and SD of the number of reads (⇒ 2 attributes for PM, 2 attributes for PU).
2. Mean and SD of the methylation levels (⇒ 2 attributes for PM, 2 attributes for PU).
3. Mean and SD of the $\log_{10}$ distances of read positions (⇒ 2 attributes).
4. Mean and SD of the $\log_{10}$ lengths of PM regions according to our smoothing approach (⇒ 2 attributes).
5. Mean and SD of the $\log_{10}$ lengths of PU regions (⇒ 2 attributes).
6. Smoothing threshold for the number of reads (⇒ 1 attribute).
7. Smoothing threshold for the methylation level (⇒ 1 attribute).

Comparing the attributes $a_1^{\text{sim}},\ldots,a_{16}^{\text{sim}}$ for the simulated data with the attributes $a_1^{\text{obs}},\ldots,a_{16}^{\text{obs}}$ observed for real BS-seq data provided 16 residuals

$$\text{res}_i = \frac{a_i^{\text{sim}} - a_i^{\text{obs}}}{w_i},\ i = 1,\ldots,16 \tag{1}$$

which were used to estimate 13 simulation parameters that have to be defined in the *WGBSSuite*. The weights $w_1,\ldots,w_{16}$ are required to ensure, that each residual enters with similar magnitude to the objective function to prevent that one attribute dominates. For calculating these weights, we uniformly drew 100 parameter vectors $\rightarrow\theta$ and calculated 100 residual vectors and calculated deviations as indicators for the typical spread of each attribute. The weights w_i were then defined as the inverse of these SDs to obtain residuals res_i with a typical size of about 1.

To prevent ill-conditioning, the residuals in (1) were augmented with additional regularizing residuals

$$\text{res}_i^{\text{prior}} = \frac{\theta_i - \theta_i^{\text{prior}}}{\text{SD}(\theta_i^{\text{prior}})},\ i = 1,\ldots,13 \tag{2}$$

which can be interpreted as weak Gaussian priors for each of the 13 simulation parameters resulting in a total of 29 residuals. The target values of these penalties for θ_i are given by the average $\theta_i^{\text{prior}} = 0.5\,U_i + 0.5\,L_i$ between upper U_i and lower bounds L_i and the priors' SDs $\text{SD}(\theta_i^{\text{prior}}) = U_i - L_i$ was chosen equal to the parameter ranges.

2.4.1 Least-squares optimization

$$\hat{\theta} = \arg\min_{\theta} V(\theta) \text{ with } V(\rightarrow\theta) = \sum_{i=1}^{29} \text{res}_i^2 \tag{3}$$

using Matlab's trust region nonlinear least-squares optimization algorithm lsqnonlin (Coleman and Li, 1996) was then performed as described in the Supplementary. Supplementary Table S1 summarizes the simulation parameters as well as the assumed upper and lower bounds. The *WGBSSuite* was originally implemented in R (Rackham *et al.*, 2015). Since evaluating the objective function in Matlab and calculating the residuals in R was too time-consuming for optimization with respect to run-time, we re-implemented the core of the *WGBSSuite* in Matlab by translating the code line by line. Optimization of the simulation parameters was performed in Matlab because optimization of the nonsmooth, stochastic objective function (3) is a numerically challenging task and it is known that optimization algorithms implemented in Matlab have strong performance (Raue *et al.*, 2013). The original implementation in the *WGBSSuite* simulates methylation differences with constant magnitude. We added the functionality of changing the level of methylation differences continuously. The change of methylation probabilities was randomly drawn according to a uniform distribution $\pm U(0,\Delta_{\max})$ in the range between 0 and $\Delta_{\max} = 0.5$.

2.5 Assessing DMR predictions

The performance of DMR predictions was assessed in terms of precision and recall. For this evaluation, the number of correctly/incorrectly classified DNA positions was counted using the simulation data generated after calibrating the simulation parameters to individual datasets. For such simulated data, the underlying true differential methylation is known which enables assessment in terms of precision and recall. *Recall* is a synonym for *sensitivity* and is equivalent to the true-positive rate

$$\text{recall} = TP/P \tag{4}$$

that is given by the ratio of the number TP of sites which were correctly classified as part of DMRs, relative to the total number of positives $P = TP + FN$, i.e. relative to the number of sites in all predicted DMRs.

Precision is a synonym for the *positive predictive value* and is defined as the fraction of predictions that are correct

$$\text{precision} = \frac{TP}{TP + FP}\quad , \tag{5}$$

i.e. the true positives relative to the sum of true positives and false positives FP. In our setup, the false positives FP coincide with the number of sites which are in predicted DMRs but are in fact not within true DMRs. The F1-score is defined as the harmonic mean

$$\frac{1}{\text{F1}} = \frac{1}{2}\left(\frac{1}{\text{precision}} + \frac{1}{\text{recall}}\right) \tag{6}$$

of precision and recall which is equivalent to

$$\text{F1} = \frac{2TP}{2TP + FP + FN} \tag{7}$$

because $P = TP + FN$ where FN denotes false negatives, i.e. differentially methylated genomic positions that are not within predicted DMRs.

Because strong differences are easier to detect, the performance in general depends on the underlying magnitude of the differential methylation. Since this magnitude is unknown, we applied a power-calculation approach with two groups of samples, each with three replicates and evaluated the dependency of the F1-score on the true underlying magnitude of differential methylation. For this purpose, the simulated DMRs were sorted according to the true differential methylation level Δ and $\text{F1}(\Delta)$ was calculated for all DMRs which have at least a differential methylation level Δ.

In this analysis a favorable approach is indicated by a superior curve

$$\text{F1}_{\text{algorithm A}}(\Delta) > \text{F1}_{\text{algorithm B}}(\Delta) \Rightarrow \text{A better than B} \tag{8}$$

because such an algorithm enables a more reliable identification of DMRs with differential levels up to Δ.

3 Results

3.1 Study design

The major goal of this study is to compare selected DMR callers in a realistic performance comparison using simulated datasets based on real data. Therefore, a blinded study design was chosen, i.e. the

DMR tools were applied to simulated BS-seq data without knowledge about the data. Blinded study designs are very common in other fields of research, e.g. for clinical studies. Recently, they were demanded for benchmark studies in the field of computation biology (Peters et al., 2018; Weber *et al.*, 2019). The advantages of blinded benchmark studies are discussed in Boulesteix *et al.* (2017) and Kreutz (2016). Experimental data taken from different plant species served as templates for simulating realistic data. Characterization, calibration of the simulation parameters and generation of realistic simulated datasets was performed. For each experimental dataset (as summarized in Table 1), a corresponding simulation dataset with 1e6 genomic positions and three replicates in two groups was generated by Group I (in Freiburg). Subsequently, the simulated datasets were analyzed with selected DMR callers in a blind way by Group II (in Marburg). No information about the frequency or magnitudes of methylation or differential methylation was provided. In such a setting, each parameter of the individual algorithms has to be chosen as the suggested default or is adapted based on plausibility arguments or by manual inspection of preliminary outcomes like in real application settings. This strategy enables a realistic and almost unbiased scoring, because tuning of configuration parameters for improving the outcomes is prohibited by the study design. After analysis of the simulated data, the predicted DMRs for the evaluated tools were sent back to Group I to be compared with the true DMRs. Then, the outcomes were assessed by calculating F1-scores and by ranking the methods. The experimental data templates as well as the simulated datasets are publicly available at https://github.com/kreutz-lab/DMR-DecisionTree.

3.2 Inclusion and exclusion of DMR approaches

After generating first simulated datasets, the blinded analyses were performed in Marburg between August 2017 and September 2018. For our study, we selected the DMR approaches that were available at that time and had been applied according to our knowledge in the context of plants. A special characteristic of plants is that DNA methylation occurs at *CHH* motifs. Since this motif is not symmetric, methylation is asymmetric on both DNA strands in contrast to methylation at *CG* and *CHG* residues. We, therefore, excluded BiSeq (Hebestreit *et al.*, 2013), *BEAT* (Akman *et al.*, 2014), *DSS* (Feng *et al.*, 2014), *RnBeads* (Assenov *et al.*, 2014), *M3D* (Mayo *et al.*, 2015) because it has been claimed that they cannot handle methylation in non-*CG* context (Catoni *et al.*, 2018).

After *defiant* was published (Condon *et al.*, 2018) in February 2018, we decided to include *defiant* because of the performance benefits observed in Condon *et al.* (2018) and since the approach relies on reasonable statistical foundations. After information about the underlying truth and about the performances of the approaches had been shared, we could not include additional approaches since this would have violated our blinded study design.

3.3 F1-scores

We first analyzed the dependency of the chosen F1-scores on the true underlying methylation level as described in Section 2.5. Figure 1A shows that on average *metilene* and *DMRcate* are superior over the whole range. The horizontal axis corresponds to the parameter Δ which controls the probability difference of observing a methylated read in the *WGBSSuite*. $\Delta = 0.5$, as an example, corresponds to a probability difference equal to 0.5 between the two groups of samples. The individual power curves for each algorithm and each dataset are shown as Supplementary Figure S6. Figure 1B shows a boxplot of the F1-scores for the 21 data contexts after averaging over the whole Δ range. The boxes indicate 25% and 75% percentiles, the lines denote the whole range of the F1-scores except outliers. This indicates that even for the best-performing methods, there are few datasets with inferior performance.

3.4 Multivariate analysis

The performance in terms of the F1-score for a simulated dataset depends on two major effects, the chosen DMR algorithm as well as on the simulation data context. The simulation dataset reflects attributes of the datasets which are in our setting determined from the experimental datasets used as templates for realistic simulation. To analyze and disentangle both effects, a linear model

$$\mathrm{F1}_{ad} = I + A_a + D_d + \varepsilon_{ad},\ a = 2,\ldots,8;\ d = 2,\ldots,21 \qquad (9)$$

for the F1-score observed for algorithm a and data context d was used to estimate the impact A_a of the individual algorithms as well as the impact D_d of the data attributes. As intercept I, the best-performing approach (*metilene*) and the Phypa-R-CG data context was chosen. The outcome of the multivariate analysis is provided as Supplementary Table S3. The estimate $\hat{I} = 0.75$ represents a reference F1-score. In this parameterization, the other estimated effects denote changes added to this intercept. ε_{ad} represents unexplained Gaussian noise. The regression model (9) can be used to assess the significance of performance differences between DMR approaches while accounting/adjusting for performance differences originating from data characteristics.

Figure 2 shows the estimated effects and 95% confidence intervals. The intercept $\hat{I} = 0.75$ is highlighted by gray shading. Negative effects indicate loss of performance on average compared to the reference analysis represented by intercept. The performance loss for *Defiant* (-0.12 with $P = 0.026$) and *BSmooth* (-0.16 with $P = 0.004$) is significant. For *methylKit* (-0.27 and $P = 3.9\mathrm{e}-6$), *MethylSig* (-0.35 with $P \leq 1\mathrm{e}-8$) and *MOABS* (-0.22 with

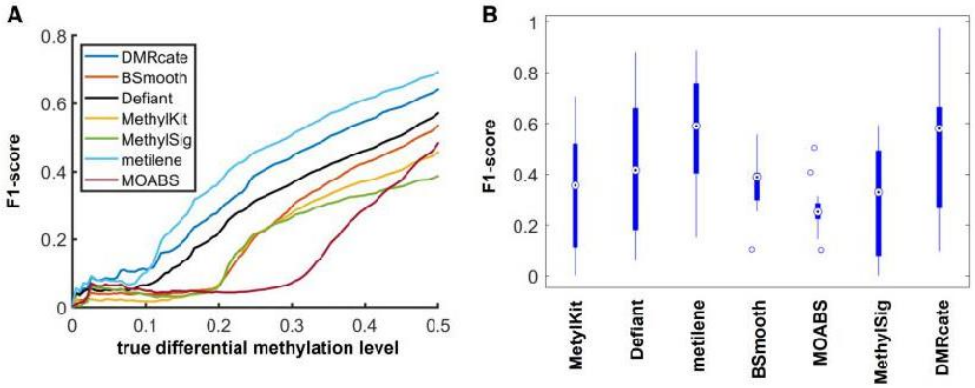


Fig. 1. (**A**) The increase in the F1-scores with the true differential methylation level. For this depiction, the outcomes of all data contexts were analyzed jointly. *Metilene* (light blue) and *DMRcate* (blue) exhibit superior performances over the whole range. (**B**) The distribution of the F1-scores averaged over all true differential levels Δ for the 21 individual datasets as boxplot. Again, *DMRcate* and *metilene* outperform the other approaches

$P = 1e-4$) the estimated performance losses in terms of the F1-score are in the range -0.22 to -0.35 and are very significant. The estimates for the different data contexts indicate the impact on the F1-score originating from the experimental data template. Note that in general increasing the sample size enhances significance of real performance differences. Therefore, an increased number of simulated and analyzed benchmark datasets could lead to further significant effects that might be too small to be significant for our chosen sample size. Despite remarkable effects, there seems to be no general tendency with respect to DNA context. By trend, datasets from *Ae.arabicum* yielded a performance decline compared to other plants, probably based on the much lower coverage of these datasets (Table 1). The comprehensive results of this statistical analysis are available as Supplementary Table S4.

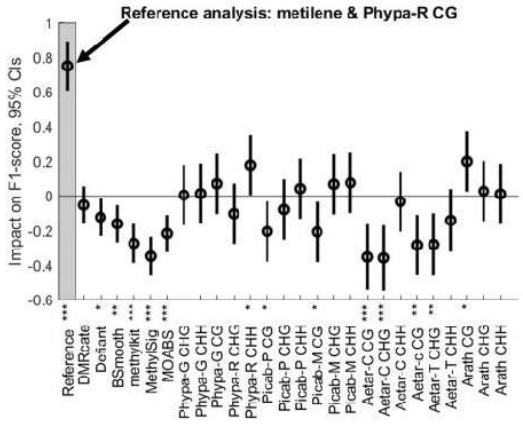


Fig. 2. Estimated impact on the F1-score of the individual algorithms and of the data contexts. *Metilene* as superior method has been chosen as a reference point (gray shading). For Phypa-R-CG context, an F1-score equals to $\hat{I} = 0.75$ is obtained. The other black dots indicate the change of the F1-score if the algorithm is and/or the background is switched. This means that negative values (below the black horizontal line) decreases the performance. Error bars indicate 95% confidence intervals, i.e. all bars that do not cross the black line are significant according to a 95% significance level. Effects that are significantly different from zero are indicated by $*P<0.05$, $**P<0.01$ and $***P<0.001$. Hence, all algorithms except *DMRcate* are significantly worse on average over all data contexts

Metilene that had on average the best performance is also faster than most other tools on average. Only *Defiant* is clearly faster and *methylKit* exhibits a minor benefit in terms of computation times on a linux computer with Intel(R) Xeon(R) CPU E5-2609-0 2.40 GHz and 62.9 GB RAM. Details about runtimes and memory usage are provided as Supplementary information.

3.5 Ranking of the algorithms and selection guidelines

As a next step for deriving general rules for selecting a DMR approach, we calculated ranks over the methods' F1-scores for each data context. Figure 3 shows ranks of the algorithms as boxplot in panel (A) and as heatmap in panel (B). In our notation, a rank equals to one corresponds to the best performance, the least-performing approach has rank zero. *metilene* shows superior performance, but there are four outliers with inferior performance. The second best-performing approach is *DMRcate*.

The black rectangles in Figure 3B indicate four data contexts where *metilene* is not among the two best-performing approaches, all of them from *Ae.arabicum*. *BSmooth* performs well for these four datasets and is, therefore, a well-suited alternative for these data contexts, although it has inferior performance on average. Therefore, one strategy for properly choosing high-performing DMR methods would be using *metilene* by default and switching to *BSmooth* for such special cases.

There is one setting, Aetar-C-CHH (first row in Fig. 3B) where this strategy has only medium performance (ranks for *metilene* and *BSmooth* are 0.28 and 0.57). *DMRcate* performs optimal for this dataset and can serve also as substitute for a second data (Aetar-T-CHH) set where *metilene* performs bad (rank = 0.33). Therefore, our suggested strategy is using *metilene* as default and switching to *BSmooth* or *DMRcate*. Choosing among these three algorithms always yields optimal performance with a superior rank (see Supplementary Fig. S5).

For generalizing this outcome, we identified data attributes as predictors for the performance and for guiding algorithm selection. We investigated the same data attributes that were used to characterize similarity between simulated and experimental data to derive a decision tree for optimal algorithm selection. Figure 4 shows the resulting data attributes which indicate optimal selection between two algorithms (Fig. 4A; *metilene* and *BSmooth*) or between three algorithms (Fig. 4B; *metilene*, *BSmooth* and *DMRcate*) which is our recommendation. At https://github.com/kreutz-lab/DMR-DecisionTree, we provide a Matlab implementation of the calculation of the data attributes and for evaluating this decision tree. This implementation can be

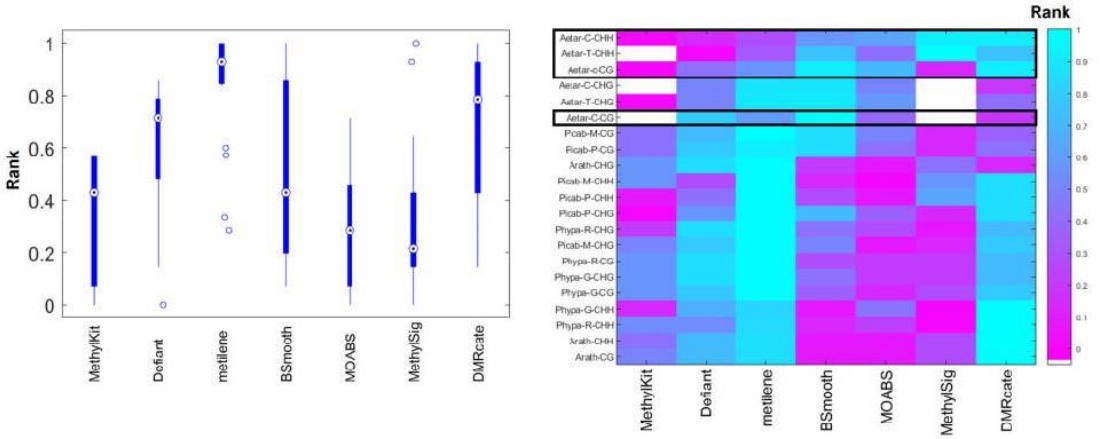


Fig. 3. (**A**) The ranking of the methods (best = 1, worst = 0) over each data context as boxplots. For this depiction, ranks were calculated via $R_{\text{method}} = \text{mean}_i(\text{rank}_{\text{methods}}(F1(\Delta_i)))$ i.e. ranks were calculated for all methods along the evaluated methylation differences Δ_i. *Metilene* shows superior performance with a median rank equals to 0.93. The depiction as heatmap shown in **B** indicates that there are data context (highlighted by the black boxes) where *metilene* has inferior performance. This observation demands for an algorithm selection guideline based on attributes of the analyzed dataset. There are seven white colors indicating datasets where *MethylSig* or *methylKit* could not complete because of too low coverage. (Color version of this figure is available at *Bioinformatics* online.)

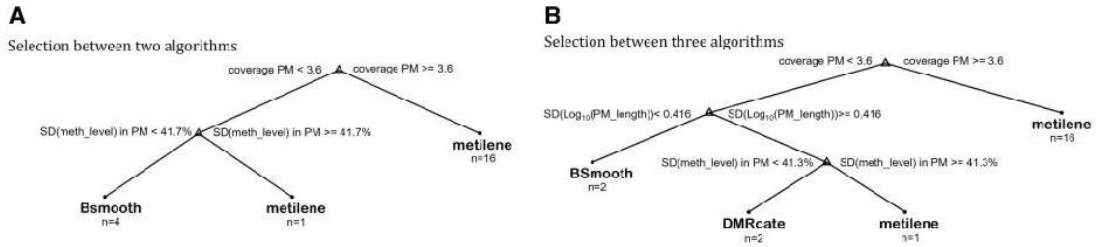


Fig. 4. Two decision trees based on data attributes indicating optimal selection between two and three approaches. (**A**) The selection guideline derived to choose between two algorithms. *Metilene* as the overall best algorithm is the default. For settings where *metilene* has bad performance, *BSmooth* serves as replacement. The decision tree indicates that in the case where the average coverage in PM regions is smaller than 3.6 and the measured methylation levels have large variability, *BSmooth* should be selected according to our analyses. (**B**) The optimal selection guideline which requires a data-based selection between three methods, namely between *metilene* by *BSmooth* and *DMRcate*. Again, metilene can be considered as default. *BSmooth* is selected for low coverage in PM (<3.6) and small SD of the length of PMs (<0.416). *DMRcate* is selected for low coverage in PM (<3.6), large SD of the length of PMs (>0.416) and small SD of the measured methylation levels (<41.3%). If methylation levels within PMs are consistently increased, this SD is small. If methylation levels within PMs are consistently increased over all read positions in the region, this SD is small. Conversely, it is large if only a subset of the read positions in the region indicate methylation

applied to any new BS-seq dataset to predict the best-performing DMR approach based on our study.

4 Discussion and summary

After decades of rapid progress in the development of high-throughput experimental techniques, and establishment of computational methods for analyzing high-throughput data in parallel, a large number of computational algorithms and software packages for statistical analysis of omics-data are nowadays available. Most of the few available benchmark studies failed to derive generally superior algorithms because context-dependency is a common feature for benchmarking of computational methods. Nevertheless, context-specific application limits and the relationships between data characteristics and performance advantages are widely unknown. Therefore, it is typically difficult or even impossible to select optimally performing computational methods in applications for analysis of high-throughput data. Usage of suboptimal or inappropriate approaches, in turn, can lead to insignificant, misleading or even erroneous results and conclusions. Therefore, guidelines for the selection of computational methods that reliably and efficiently work in a given context are an important requirement for reliable research and for the transfer of computational methods to experimental practice.

In this study, we present a benchmark study comprising seven prominent methods for identification of DMRs were applied for the plant methylation contexts: *CG*, *CHG* and *CHH*. We tested DMR algorithms for 21 experimental datasets from plants belonging to three taxonomic groups. Calibration of the simulation parameters individually to each experimental dataset enabled simulation of realistic benchmark data. Previous benchmark studies provide only heterogeneous and fragmentary information for plants. Catoni *et al.* (2018) assessed the performance of MethCP, their own new approach, with *MethlySig* and *methylKit* for an *A.thaliana* experiment where the differential methylation between wild-type and methyltransferase knockout plants was calculated. Methyltransferase knockouts block methylation almost entirely and, therefore, cause uniform and unphysiological levels of underlying methylation differences. Moreover, they observed inconsistent outcomes for *CG*, *CHG* and *CHH*, that diminished the utility of that study for our purposes. In fact, most studies in the literature comparing the performance of DMR approaches were performed by jointly presenting a new approach. Such studies, however, are easily biased because they are not independent (Boulesteix *et al.*, 2013, 2018) and they are at least partly performed to demonstrate benefits of the new method they are presenting.

In contrast to previous studies, we tried to design a benchmark study which has minimal bias and is tailored to DMR data from plants. In our study, the DMR approaches were applied to simulated data but without any knowledge about the true differential methylation like in real application settings. This blinded study design provides an assessment in terms of precision and recall which is comparable to application settings where configurations parameters have to be selected based on default suggestions and manual inspection of preliminary outcomes. One shortcoming of the blinded study design is that it is not possible to extend the study by additional methods because this prohibits blinded configurations and evaluations of the methods. Therefore, we could not add other DMR approaches after the first assessments were conducted.

Our benchmark study showed clear advantages of some methods although no approach outperformed all other methods for all datasets. Overall, *metilene* exhibited superior performance and we suggest usage of *metilene* as default approach. However, we also found few data settings where *metilene* has inferior performance. This is in agreement with previous studies (Condon *et al.*, 2018; Gong and Purdom, 2018; Jühling *et al.*, 2016). We found that for our datasets, where *metilene* showed lower performance, either *DMRcate* or *BSmooth* were superior and could, therefore, serve as better options. In agreement with Condon *et al.* (2018), *Defiant* also performed well in our setting but the performance of *metilene* was superior.

We could find data attributes like coverage or lengths of PM regions that guide to the optimal algorithm choice for all our datasets. We provide publicly available code that can be used to calculate these attributes. We derived a decision tree based on these attributes indicating the optimal choice. The optimal rule derived for our scope use *metilene* as default and guides the usage of *BSmooth* and *DMRcate* as substitutes. *BSmooth* should be selected in case of small coverage and small variation of the lengths of DMRs. *DMRcate* is the method of choice for small coverage, large variation of the length of DMRs and small SD of the measured methylation levels.

Analogous guidelines are common in some traditional statistical fields, e.g. for selecting statistical tests. As an example, for two-group comparison, the *t*-test has superior power for normally distributed errors, but it is suggested to use the Wilcoxon rank-sum test as substitute in case of unknown noise distribution or if the data contains outliers. Such rules are an important requirement for reliability of research in all bioinformatics fields for the transfer of new computational approaches to basic research as well as clinical and industrial applications.

Our outcomes are in line with the common tendency in the literature that for comprehensive computational analysis of high-throughput data there is usually not a single outperforming approach (Shippy *et al.*, 2006; Su *et al.*, 2014; Webb-Robertson *et al.*, 2015). Instead, the performances depend on attributes of the analyzed data. These insights demand for algorithm selection guidelines based on attributes of the data.

A relevant remaining question is a common issue for all benchmark studies, namely to which extent the outcomes translate into other applications and biological contexts. Since BS-seq is a generic experimental technique that is applicable independently of the biological background, the question reduces to generalization and

validity for new datasets with distinct attributes. The only solution to this universal issue is emphasizing the importance of comprehensive benchmark studies, establishment of standards for the designs of benchmark studies as well as a broad set of representative benchmark problems that are routinely utilized for development of new algorithms. The benchmark data which was generated within this study contributes to these efforts.

In terms of the data attributes that we used for characterizing our experimental data and for optimizing the similarity of the simulated datasets, we observed that the differences between the three DNA contexts (*CG*, *CHG* and *CHH*) are larger than the differences between the different plants (see Supplemental Information, Fig. 4). This observation in combination with the fact that we investigated plants from three different taxonomic groups indicates that the data used in our study samples a broad range of BS-seq data characteristics, thus we expect rather general outcomes. This speculation, however, can only be proven if larger sets of different organisms are evaluated which is beyond the scope of this study.

Our study pinpoints four important general aspects for performing meaningful benchmark studies. (i) We chose a blinded design to diminish biased assessment. (ii) The study is independent since none of the authors were involved in the development of any DMR methods tested. (iii) A multivariate statistical model has been applied for assessing significance and for decomposing the impact of algorithm selection from other effects. Moreover, (iv) a data-based decision rule has been derived to select between competing computational methods which is applicable to any new dataset.

Acknowledgements

We also acknowledge funding by the European Research Area Network for Coordinating Action in Plant Sciences (ERA-*CAPS*) for the 'SeedAdapt' consortium project (https://www.seedadapt.eu), and the *Austrian Science Fund FWFI3979*.

Funding

This work was supported by the German Ministry of Education and Research [EA:Sys FKZ031L0080] and by the German Research Foundation (DFG) under Germany's Excellence Strategy [CIBSS-EXC-2189-2100249960-390939984]. Parts were funded by the European Training Network *EpiDiverse* (https://epidiverse.eu) that received funding from the EU Horizon 2020 program under Marie Skłodowska-Curie [agreement number 764965]. We also acknowledge funding by the European Research Area Network for Coordinating Action in Plant Sciences (ERA-CAPS) for the 'SeedAdapt' consortium project (https://www.seedadapt.eu), and the Austrian Science Fund FWFI3979.

Conflict of Interest: none declared.

References

Akalin,A. *et al.* (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.*, **13**, R87.

Akman,K. *et al.* (2014) Genome-wide quantitative analysis of DNA methylation from bisulfite sequencing data. *Bioinformatics*, **30**, 1933–1934.

Assenov,Y. *et al.* (2014) Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods*, **11**, 1138–1140.

Bewick,A.J. *et al.* (2017) Chromomethylases and gene body DNA methylation in plants. *Genome Biol.*, **18**, 65.

Bewick,A.J. and Schmitz,R.J. (2017) Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.*, **36**, 103–110.

Boulesteix,A.L. *et al.* (2013) A plea for neutral comparison studies in computational sciences. *PLoS One*, **8**, e61562.

Boulesteix,A.L. *et al.* (2017) Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. *BMC Med. Res. Methodol.*, **17**, 138.

Boulesteix,A.-L. *et al.* (2018) On the necessity and design of studies comparing statistical methods. *Biom. J.*, **60**, 216–218.

Catoni,M. *et al.* (2018) DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucleic Acids Res.*, **46**, e114.

Chen,X. *et al.* (2018) ChIP-seq: a powerful tool for studying protein–DNA interactions in plants. *Mol. Biol.*, **27**, 171–180.

Coleman,T. and Li,Y. (1996) An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optimiz.*, **6**, 418–445.

Condon,D.E. *et al.* (2018) Defiant: (DMRs: easy, fast, identification and ANnoTation) identifies differentially methylated regions from iron-deficient rat hippocampus. *BMC Bioinformatics*, **19**, 31.

Dolzhenko,E. and Smith,A.D. (2014) Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*, **15**, 215.

Feng,H. *et al.* (2014) A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.*, **42**, e69.

Gong,B. and Purdom,E. (2018) MethCP: differentially methylated region detection with change point models. *bioRxiv*, doi:http://dx.doi.org/10.1101/265116.

Hansen,K.D. *et al.* (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.*, **13**, R83.

Hebestreit,K. *et al.* (2013) Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, **29**, 1647–1653.

Heer,K. *et al.* (2018) Detection of somatic epigenetic variation in Norway spruce via targeted bisulfite sequencing. *Ecol. Evol.*, **8**, 9672–9682.

Jühling,F. *et al.* (2016) Metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res.*, **26**, 256–262.

Kreutz,C. (2016) New concepts for evaluating the performance of computational methods. *IFAC-PapersOnLine*, **49**, 63–70.

Lang,D. *et al.* (2008) Exploring plant biodiversity: the *Physcomitrella* genome and beyond. *Trends Plant Sci.*, **13**, 542–549.

Lang,D. *et al.* (2018) The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.*, **93**, 515–533.

Liu,Y. *et al.* (2019) Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.*, **37**, 424–429.

Mayo,T.R. *et al.* (2015) M3D: a kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics*, **31**, 809–816.

Meyberg,R. *et al.* (2019) Characterization of evolutionarily conserved key players affecting eukaryotic flagellar motility and fertility using a moss model. *bioRxiv*, doi:https://doi.org/10.1101/728691.

Park,Y. *et al.* (2014) MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, **30**, 2414–2422.

Peters,B. *et al.* (2018) Putting benchmarks in their rightful place: the heart of computational biology. *PLoS Comput. Biol.*, **14**, e1006494.

Peters,T.J. *et al.* (2015) *De novo* identification of differentially methylated regions in the human genome. *Epigenet. Chromatin*, **8**, 6.

Rackham,O.J.L. *et al.* (2015) WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools. *Bioinformatics*, **31**, 2371–2373.

Raue,A. *et al.* (2013) Lessons learned from quantitative dynamical modeling in systems biology. *PLoS One*, **8**, e74335.

Robinson,M.D. *et al.* (2014) Statistical methods for detecting differentially methylated loci and regions. *Front. Genet.*, **5**, 324.

Sahu,P.P. *et al.* (2013) Epigenetic mechanisms of plant stress responses and adaptation. *Plant Cell Rep.*, **32**, 1151–1159.

Seymour,D.K. *et al.* (2014) Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. *PLoS Genet.*, **10**, e1004785.

Shippy,R. *et al.* (2006) Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nat. Biotechnol.*, **24**, 1123–1131.

Su,Z. *et al.* (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat. Biotechnol.*, **32**, 903.

Sun,D. *et al.* (2014) MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.*, **15**, R38.

Webb-Robertson,B.-J.M. *et al.* (2015) Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J. Proteome Res.*, **14**, 1993–2001.

Yang,Y. and Scott,S.A. (2017) DNA methylation profiling using long-read single molecule real-time bisulfite sequencing (SMRT-BS). *Methods Mol. Biol.*, **1654**, 125–134.

Weber,L.M. *et al.* (2019) Essential guidelines for computational method benchmarking. *Genome Biol.*, **20**, 125.

3.1.2 Further applicability of this work

Amongst all DMR callers, metilene was chosen as superior and further investigated in the development of the EpiDiverse DMR pipeline (https://github.com/EpiDiverse/dmr, accessed on 1 May 2021). The pipeline was applied and used for the EWAS benchmark published by Can et al., 2021 [68].

3.2 The EpiDiverse plant Epigenome-Wide Association Studies (EWAS) pipeline (Paper-II)

Association studies emerged from the necessity of elucidating the relationship between complex traits, genetics, and epigenetics. The pioneering study of this field GWAS led its counterpart so-called EWAS in the epigenetics area. The scarcity of data and limitations with current EWAS tools and plant species brought the need for a comprehensive and compatible tool. Therefore, the EpiDiverse EWAS pipeline was developed to allow multiple inputs, a non-hard coded system for all species, multiple graph options with outputs for a user to observe results better, and unique missing data imputation to study potential epigenetic markers associated with genetics and/or phenotypic traits as publicly available (https://github.com/EpiDiverse/ewas, accessed on 1 May 2021).

3.2.1 Paper

Following is the electronic publication

epigenomes

MDPI

Article

The EpiDiverse Plant Epigenome-Wide Association Studies (EWAS) Pipeline

Sultan Nilay Can [1], Adam Nunn [2,3], Dario Galanti [4], David Langenberger [2], Claude Becker [5,6], Katharina Volmer [7], Katrin Heer [8,9], Lars Opgenoorth [10], Noe Fernandez-Pozo [1] and Stefan A. Rensing [1,10,11,*]

1 Plant Cell Biology, Department of Biology, University of Marburg, 35043 Marburg, Germany; nilaycan@biologie.uni-marburg.de (S.N.C.); noe.fernandezpozo@biologie.uni-marburg.de (N.F.-P.)
2 ecSeq Bioinformatics GmbH, 04103 Leipzig, Germany; adam.nunn@ecseq.com (A.N.); david.langenberger@ecseq.com (D.L.)
3 Bioinformatics Group, Department of Computer Science, University of Leipzig, 04107 Leipzig, Germany
4 Plant Evolutionary Ecology, Institute of Evolution and Ecology, University of Tübingen, Auf der Morgenstelle 5, 72076 Tübingen, Germany; dario.galanti@uni-tuebingen.de
5 Gregor Mendel Institute of Molecular Plant Biology (GMI), Austrian Academy of Sciences, Vienna BioCenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria; claude.becker@gmi.oeaw.ac.at
6 Genetics, Faculty of Biology, Ludwig-Maximilians-University München, 82152 Martinsried, Germany
7 Department of Forest Genetic Resources, Nordwestdeutsche Forstliche Versuchsanstalt (NW-FVA), 37079 Göttingen, Germany; Katharina.Volmer@nfp.niedersachsen.de
8 Conservation Biology, Department of Biology, University of Marburg, 35043 Marburg, Germany; katrin.heer@uni-marburg.de
9 Plant Ecology and Geobotany, Department of Biology, University of Marburg, 35043 Marburg, Germany
10 Centre for Biological Signaling Studies (BIOSS), University of Freiburg, 79104 Freiburg, Germany; Lars.Opgenoorth@Staff.Uni-Marburg.DE
11 SYNMIKRO Center for Synthetic Microbiology, University of Marburg, 35043 Marburg, Germany
* Correspondence: stefan.rensing@biologie.uni-marburg.de

Abstract: Bisulfite sequencing is a widely used technique for determining DNA methylation and its relationship with epigenetics, genetics, and environmental parameters. Various techniques were implemented for epigenome-wide association studies (EWAS) to reveal meaningful associations; however, there are only very few plant studies available to date. Here, we developed the EpiDiverse EWAS pipeline and tested it using two plant datasets, from *P. abies* (Norway spruce) and *Q. lobata* (valley oak). Hence, we present an EWAS implementation tested for non-model plant species and describe its use.

Keywords: EWAS; GWAS; plant epigenetics; DNA methylation; non-model species; pipeline

Citation: Can, S.N.; Nunn, A.; Galanti, D.; Langenberger, D.; Becker, C.; Volmer, K.; Heer, K.; Opgenoorth, L.; Fernandez-Pozo, N.; Rensing, S.A. The EpiDiverse Plant Epigenome-Wide Association Studies (EWAS) Pipeline. *Epigenomes* **2021**, *5*, 12. https://doi.org/10.3390/epigenomes5020012

Academic Editors: Cao Xuan Hieu and Vu Thi Ha Giang

Received: 15 March 2021
Accepted: 20 April 2021
Published: 4 May 2021

1. Introduction

Epigenetics describes DNA or chromatin modifications that might change transcriptional activity without altering the DNA sequence and might be propagated somatically or through the germline. Epigenetic modifications such as DNA methylation and histone modifications (acetylation, phosphorylation, ubiquitylation, sumoylation) may affect the chromatin structure and, thereby, the access to genetic information [1]. Of these epigenetic modifications, methylation currently is the most intensively studied in plants as it can be easily assessed. DNA methylation is an epigenetic modification consisting of the addition of a methyl group (CH_3) to the fifth carbon of the cytosine (C). Epigenetic mechanisms can alter phenotypic traits [2]. It was shown that DNA methylation may play a crucial role in gene expression regulation, e.g., of plant defense response under various environmental stresses [3]. DNA methylation may lead to heritable epigenetic information and transgenerational epigenetics describes the lack of resetting mechanisms of epigenetic states between generations. Epialleles are responsible for this heritable phenotypic variation and plants seem to have this type of inheritance in contrast to mammals [4]. There are very few known

examples of natural epialleles, suggesting that epiallelic variation is very rare in nature, compared to allelic variation [5]. One of the first discovered natural plant phenotypes not based on a change in the DNA sequence was *Linaria vulgaris* (toadflax) [6]. This study revealed that mutant alleles showed high DNA methylation but low gene expression and a clear coincidence between the revertant phenotype and the degree of DNA methylation at the Lcyc locus [6]. Another example of epimutation alleles was first described by Barbara McClintock in maize by focusing on the effect of suppressor–mutator (Spm) transposons on gene expression [7]. Moreover, Cnr mutants in tomato showed colorless, non-ripening fruits and this was found to be caused by a mutant allele at the LeSPL–CNR locus [8]. The mutant phenotype was found to be associated with increased DNA methylation at the promoter region; the upstream promoter of LeSPL–CNR coincides with a TE that is heavily methylated in both wildtype and Cnr mutant. Finally, many epimutable alleles have been defined in Arabidopsis, and they all seem to involve TEs or other repetitive sequences [5]. DNA methylation also leads to transposon mobility, potentially affecting both short- and long-term adaptation to environmental conditions [9–11]. An example of epigenetic adaptation to temperature is vernalization observed in *Arabidopsis thaliana* ecotypes, which is regulated by flowering locus (FLC), where cold stress triggers H3K27me3 and H3K9me deposition in the FLC chromatin [12].

DNA methylation may occur in different nucleotide contexts. In animals, only Cs in CG contexts are methylated, whereas in plants [13], DNA methylation can be symmetrical (CG and CHG contexts) or asymmetrical (CHH, where H represents A, T, or C) [14–17]. The different contexts have different maintenance mechanisms [18]. Whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS) are widely used methods to determine DNA methylation at a single-base resolution [19,20]. One goal in plant epigenetics is to detect positions or regions that are differentially methylated due to treatment or different environments, and multiple samples are required since differentially methylated positions (DMPs) are called using statistical approaches [21]. Differentially methylated regions (DMRs) are genomic regions where multiple adjacent positions reveal differential methylation [22].

Interest in understanding the genetic architecture of complex traits led to association studies to relate genetics and epigenetics with phenotypic traits. Testing genetic variation across genomes of individuals to reveal genotype–phenotype associations is made possible by genome-wide association studies (GWAS), which have frequently been used for human disease studies [23,24] and enabled the detection of many genetic variants significantly associated with complex human diseases. Results obtained from GWAS have been clinically reliable and help to develop new treatments for multiple diseases from diabetes to schizophrenia [25,26]. In plants, GWAS is a powerful tool for understanding complex traits, useful to discover the genetics related to important traits in agriculture and to accelerate breeding programs [27]. It has been applied to many crop species, e.g., maize (*Zea mays*), wheat (*Triticum aestivum*), rice (*Oryza sativa*), soybean (*Glycine max),* sorghum (*Sorghum bicolor*), barley (*Hordeum vulgare*), cotton (*Gossypium hirsutum*), and the model species Arabidopsis [28–31]. Moreover, GWAS is used to reveal genomic regions related to physiological, agronomic, and fitness traits such as plant height, stress tolerance, flowering time, kernel number, and grain yield [28–30,32], and identified genes connected with geographical deviation and adaptation in rice domestication [33]. Additionally, GWAS have also been used with genetic engineering, e.g., transgenic drought tolerance in maize was developed after the discovery of ZmVPP1 [34], and this led to an increment of studies using genome editing on target genes [35]. However, many diseases and disorders in humans including cancer show an epigenetic association [28–30]. Due to that, epigenome-wide association studies (EWAS) as a counterpart of GWAS have also been used in human studies [36]. EWAS is a powerful method to reveal epigenetic variation associated with biological traits [22,37]. Transgenerational epigenetic marks can be transmitted to descendants through mitosis (in case of vegetative propagation) or meiosis (sexual reproduction) [38]. The methylation variation of the same gene between different

plants is called epialleles and can lead to different phenotypes that are heritable. Mutants of *Linaria vulgaris* are an example of transgenerational epigenetic inheritance [6]. Mechanisms involved in transgenerational inheritance of epigenetic marks are not fully understood but data showed that DNA methylation easily passes through generations and many studies focus on this mark [39]. Histone modification can also affect gene function and phenotype; however, it has been largely ignored in EWAS due to technological limitations and sample availability. Germline cells in plants are inherited from somatic cells and therefore can contribute to the heritability of epigenetic marks. Plants can sense environmental changes during their vegetative growth, and it may lead to epigenetic changes in cell lines that generate a germline [9]. Studies showed that stress-induced transgenerational reactions depend on DNA methylation in Arabidopsis [40,41]. Therefore, heritable epialleles may affect plant evolution, phenotypic traits, and fitness. Since many of the plants go through asexual propagation, meiotic epigenetic resetting does not occur, and information is carried to the next generation more effectively than in sexual reproduction [42]. Epigenetic changes are dynamic, making it difficult to discriminate significant relationship between phenotype and epigenetic mechanisms—a major challenge of EWAS [43] and common issues both for GWAS and EWAS are dealing with missing and big data [44]. Thus far, there has been very scarce use of EWAS for plants (for example, a PubMed search for "ewas plant" returned seven hits 19 February 2021, while "ewas human" returned 131). Published examples include DNA methylation variation in *Quercus lobata* (valley oak) associated with climatic gradients [45], and EWAS has been successfully applied to identify the epigenetic change that causes the metastable somaclonal variant in *E. guineensis* (oil palm) [46]. Another study with stone pine (*Pinus pinea)* showed that there was a remarkable level of phenotypic plasticity. Vegetatively propagated *P. pinea* trees showed a high degree of DNA methylation under different environmental conditions [47]. Several EWAS tools have been developed, yet most of them are hardcoded for human studies such as *GLINT* [48] or not able to deal with missing data such as EWAS: epigenome-wide association study software v2.0 [49]. However, there is one tool not hard coded that also accounts for genetic data, is compatible with all species, and allows missing data imputation, namely, the *GEM R package* [50], hence chosen for this study.

Here, we present the EpiDiverse EWAS pipeline, developed in the context of the EpiDiverse ITN network (https://epidiverse.eu/, accessed on 1 March 2021), which studies the effects of epigenetics in natural variation, stress responses, and acclimations of plants [51]. We aimed at realizing parts of the research agenda of EpiDiverse outlined in Richards et al. (2017) [51]. In EpiDiverse, a set of bioinformatics pipelines was developed to facilitate epigenetic analyses based on DNA methylation, especially for non-model plants. These pipelines are modular and scalable and can easily connect their inputs and outputs (Figure 1), providing a suite of useful tools for whole-genome bisulfite sequencing (WGBS) methylation calling, single nucleotide polymorphism detection (SNP), differentially methylated position, and region (DMP and DMR) detection and EWAS (https://github.com/EpiDiverse, accessed on 1 March 2021). The software included in the WGBS and DMR pipelines was selected from the best performing tools in benchmarking studies [52,53]. Here, we describe and test the performance of the EpiDiverse EWAS pipeline using four different input types from two non-model plant datasets and test the effect of missing data.

2. Results and Discussion

2.1. EpiDiverse EWAS Pipeline Workflow

The EpiDiverse EWAS pipeline is based on functions implemented in the GEM R package [50] and extends them by multiple features that allow the use of methylation calls and differential methylation data, with optional analysis of methylation quantitative trait loci (methQTLs) for diploid organisms from variant call data in which methQTL is an epigenetic marker that coincides with a quantitative trait locus (QTL). Additionally, missing data filtering with the GEM R package was modified, and estimation is conducted with beta distribution because there is evidence that the existing method biases the calculation

of FDR values. This issue is made apparent when the methylation data are subdivided, e.g., by chromosome/scaffold: since the global methylation values are calculated from the remaining positions for the sample, the *p*-values themselves vary wildly for the same positions, depending on how many other positions are present during the analysis.

Additional graphs are generated (sequence dot plots, Manhattan plots) to help the user to understand results better and observe more visual outputs. The EpiDiverse EWAS pipeline performs epigenome-wide association studies, employing three models implemented in the GEM R package. We preferred GEM over other EWAS tools because, for example, *GLINT* [48] is hard coded for use in Illumina human methylation arrays, and *EWAS: epigenome-wide association study software v2.0* [49] is not able to estimate missing data.

The EWAS pipeline is part of the EpiDiverse toolkit, which provides tools for mapping WGBS data and methylation calling (WGBS pipeline), calculation of differential methylation (DMR pipeline), and estimation of genetic variants from bisulfite sequencing (SNP pipeline) (Figure 1). Bisulfite sequencing raw data can be processed by the WGBS pipeline to produce methylation calls, variant files in the SNP pipeline, and DMPs and DMRs in the DMR pipeline. The EWAS pipeline can use as input any combination of results from the other EpiDiverse pipelines (Figure 1), or users can provide their input files.

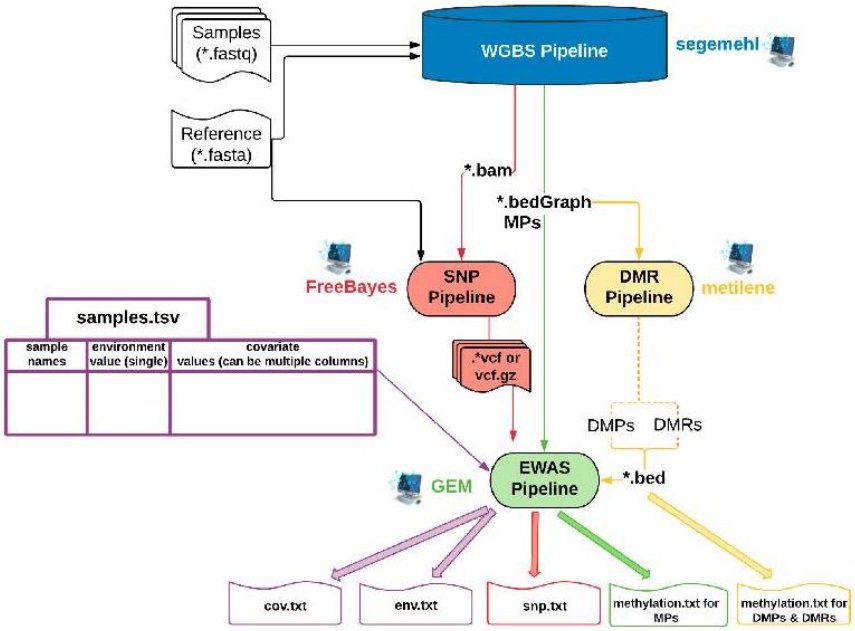


Figure 1. The EpiDiverse EWAS pipeline workflow and its interaction with the WGBS, SNP, and DMR EpiDiverse pipelines. Utilized packages or software were specified next to pipeline names. The EpiDiverse epigenome-wide association studies (EWAS) pipeline requires a tab-separated sample.tsv file (shown with purple frame) to specify climatic data and covariate(s) for group determination (can be sampling site, geographical location, or treatment group) and methylation data. As methylation input types, it accepts methylation calls (green arrow) and differentially methylated positions/differentially methylated regions (DMPs/DMRs) (yellow arrow), which can be provided by the whole-genome bisulfite sequencing (WGBS) and the DMR pipelines, respectively. The EWAS pipeline allows running three different models to find epigenetic markers associated with the environment (E), genetic variation (G), or the combination of both (GxE). G and GxE models need single nucleotide polymorphism (SNP) information (red arrow), which can be directly calculated by the SNP pipeline using bisulfite sequencing data, or, as for all other inputs, it can be provided by users. See Figures S1–S4 for more detail. * indicates multiple files with the same extension in a specified directory.

The pipeline was built using Nextflow [54], a workflow tool for running tasks across multicompute infrastructures in a portable manner. It comes with docker containers to facilitate the installation process. The dependencies for the pipeline can be managed by Conda (https://docs.conda.io/en/latest/miniconda.html, accessed on 1 March 2021), Singularity [55], and/or Docker [56].

2.1.1. Input Types for the EWAS Pipeline

Input can be derived from other EpiDiverse pipelines (WGBS, SNP, DMR) or user-provided and is combined with a user-provided, tab-separated sample sheet file to submit EWAS analysis (Figure 1). This sample sheet file has sample identifiers in the first column, environment values in the second column, and single/multiple covariates after the environmental values. This file is required to use sample names as a key and derive covariates (Figure S4). To account for genetic interaction, the EWAS pipeline needs an SNP genotype matrix encoded by 1, 2, 3 for major homozygote (AA), heterozygote (AB), and minor homozygote (BB) variants in vcf, bcf, or zipped vcf.gz format. The SNP pipeline can be used to extract genetic variation files in vcf.gz format to derive this input.

Cytosine methylation calls in all contexts (CG, CHG, and CHH) in bedGraph format as separated files (Figure S1) are used as methylated position (MP) input. Since each file represents a methylome per sample, these single bedGraph files are united with the bedtools unionbedg function [57] to generate a single methylation file with all samples as columns (Figure S2). In some cases, some positions are not covered by all sample methylomes, generating missing data (shown as NA), which may arise due to regions of low coverage sequencing. The pipeline can also be fed with differential methylation data in DMP/DMR bed format such as provided by the EpiDiverse DMR pipeline (Figure S3). The union of DMPs is simply intersected with the MPs with bedtools intersect to obtain individual methylation values per sample [57]. DMR input can be processed in two different ways—either (i) the MPs included in the region are analyzed or (ii) average methylation is calculated for all positions in that region for all samples, and regions are provided as identifiers (Figure 2). If no specific input parameter is indicated, the pipeline will automatically start the run with suitable models using the provided inputs (Table 1).

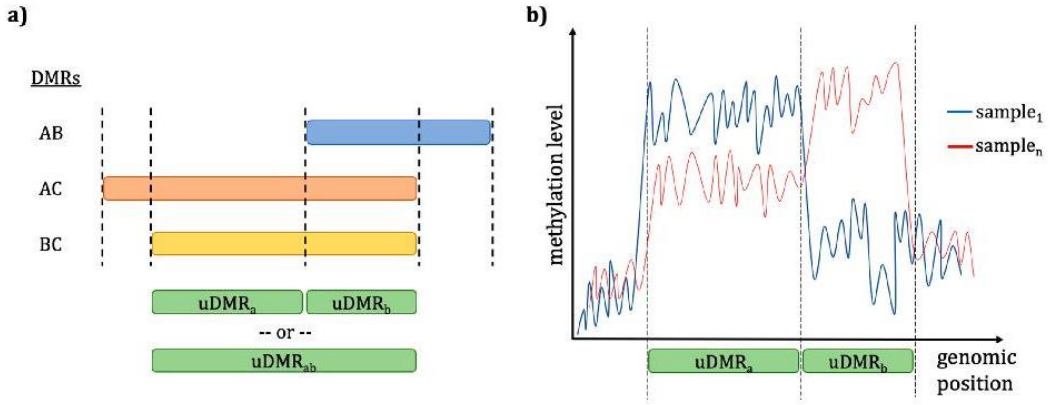


Figure 2. Average-over-region method with the DMRs input type. Overview to show (**a**) when differentially methylated regions (DMRs) arise from multiple pairwise comparisons between groups (e.g., AB, AC, BC) they are intersected to form distinct, nonoverlapping union DMRs (uDMR) according to a minimum fraction of supporting comparisons provided by the user (e.g., X = 0.5 in this sample). These uDMRs can be merged or taken as independent for further analysis. The resulting uDMRs are intersected with the methylated positions in (**b**) to derive average methylation levels in each sample for each region, which can then be carried forward as unique identifiers for EWAS. When only a single set of DMRs are provided to the pipeline, the regions are simply taken as is for the averaging process. This averaging process is repeated for all samples.

Table 1. Required inputs and file formats for the EpiDiverse EWAS pipeline. All possible input types and requirements for different models to run the epigenome-wide association studies (EWAS) pipeline. Tab-separated "sample sheet" file and methylated positions (MPs) input are required for all runs, differentially methylated positions (DMPs), and differentially methylated regions (DMRs) are needed if users would like to run the pipeline with these inputs. Genetic variants are necessary for G and GxE models.

Input	Description	File(s) Formats	Required for Which Runs?	Required for Which Model?
sample sheet	Sample list, which includes sample names as key variables, single environment/phenotype data, and covariate(s).	txt	Required for all runs	Required for all models
MPs	Context-specific methylation calls per sample.	bedGraph	Required for all runs	Required for all models
DMPs	Context-specific differentially methylated positions.	bed	Required to run the pipeline with DMPs	Allowed for all models
DMRs	Context-specific differentially methylated regions.	bed	Required to run the pipeline with DMRs	Allowed for all models
Genetic variants	Genetic markers either in single or multisample formats.	vcf or vcf.gz	Required to run the G and GxE models	Required for G and GxE models

The advantage of using MPs alone is that no prior assumption about pairwise comparisons or sample grouping has to be applied and that the full data are used. If there is good reason to believe that DMRs capture the hypervariable regions where DNA methylation differences are occurring, then it is an advantage to include them. This reduces the number of multiple tests of MPs/DMPs and can be based on meaningful a priori knowledge. On the other hand, when using DMPs and/or DMRs, the data size is reduced, resulting in lower running time for the EWAS pipeline itself. To decide which samples should be compared in a pairwise fashion to create DMPs or DMPs, assumptions need to be made that might bias the results and might not capture all the relevant information.

2.1.2. Available Models

The models E, G, and GxE of the GEM tool suite are available in the EpiDiverse EWAS pipeline (Figures 1 and 3). The Emodel is performed for detecting methylation markers associated with environmental parameters using linear regression *lm* ($M \sim E + cvrt$), with an *i*th vector from the methylation matrix (M), a jth vector from the environment matrix (E), and covariate(s) (cvrt) matrix, which is required for the grouping of samples, is used as a matrix-based iterative correlation (Figure S4) [50]. The output of Emodel is a list of potential epigenetic markers significantly correlated with a specific environmental factor.

The Gmodel is used for detecting methylation markers associated with genotype data using a linear regression *lm (M~G + cvrt)*, with an ith vector of methylation matrix (M), a jth vector of SNP genotype matrix (G), and covariate matrix. Gmodel creates a methQTL genome-wide map by performing a matrix-based iterative correlation between SNP and methylation matrices. The output of the Gmodel is a list of marker–SNP pairs, in which the SNP is coupled with the C of interest.

The GxE is used to reveal an association between genetic variation and environmental factors that may affect heritable or nonheritable DNA methylation levels, using again a linear regression *lm (M~G x E + cvrt)*, in which environmental values are combined with the covariate file. The output of the GxE is a list of marker–SNP–env triplets in which the environment parameter is a factor divided into genotype groups (AA, AB, and BB) to explain the significant C of interest. This model's output hypothesizes that the relationship between methylation and environment can be better understood by a division of genotype groups. As discussed in the "Removal of genetic variants that might be interpreted as significant epigenetic marks" Section the genetic variance may influence epigenetic variance via loss or gain of a methylation site and this can be addressed while intersecting different model outputs.

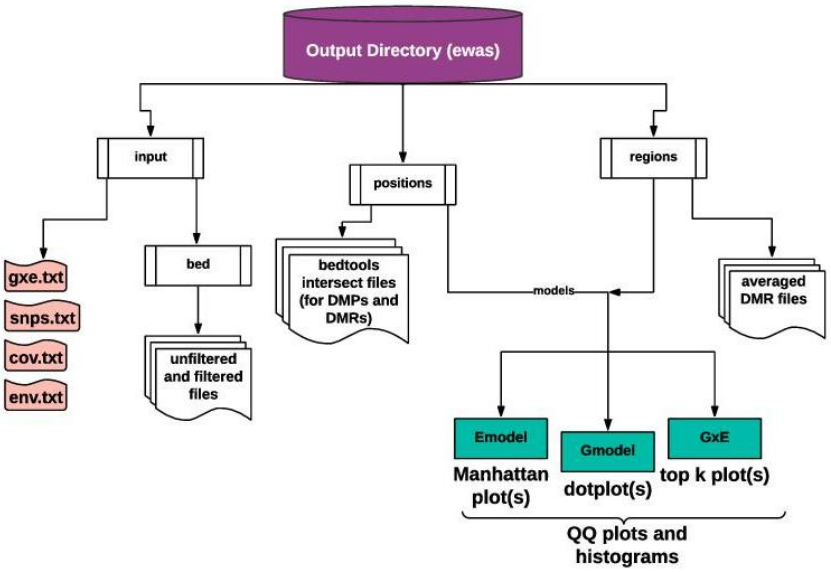


Figure 3. EpiDiverse EWAS pipeline output directory structure. EpiDiverse EWAS pipeline generates input directory as default and positions directory with methylation calls, DMPs, and regions directory DMRs. Input directory covers gxe.txt, snps.txt, cov.txt, and env.txt files and bed directory with merged bedGraph files as unfiltered and filtered and missing data estimated. Both positions and regions directories have three subdirectories for outputs and graphs with Emodel, Gmodel, and GxE names. Q–Q plots and histograms are produced with all models.

2.1.3. NA Filtering and Imputation with Methylation and SNP Datasets

Regardless of which markers are used with the EpiDiverse EWAS pipeline, a common problem has to be solved related to missing data. The pipeline unites methylation calls and DMP and DMR outputs from independently sequenced samples with the bedtools unionbedg function to derive the main methylation matrix. This unifying process leads to some positions having missing data (NAs). This problem may occur due to stochastic coverage variability or due to reference bias when a deletion affects a portion of the samples. Both cases can result in some loci being covered in some samples and not in others, causing incomplete datasets where NAs have to be excluded or estimated.

GEM replaces missing methylation values by calculating the global methylation for the rest of the sample and does not discard any positions with a high amount of missing data. As outlined above, the generic GEM method biases the calculation of FDR values (when the methylation data is divided by chromosome/scaffold, while the global methylation is calculated from the existing positions per sample, *p*-values vary for the same position; see "NA filtering and imputation with methylation and SNP datasets" for more details). Simply removing all positions where even one sample has missing data can be an alternative solution but can also reduce the total size of the dataset significantly in large cohorts. Instead, we implemented a strategy similar to that implemented in metilene [39], in which missing methylation values are imputed based on a beta distribution using the values of the remaining samples at the same position [58]. The reasoning is that if we are unable to provide real data to the model, then the next best thing is to provide estimated data that have the minimum possible impact on the model. Any significant associations that arise on markers with estimated data should therefore be driven by the samples for which we do have real data. Missing data imputation is also carried out for the SNP data using BEAGLE [40], based on a proportion of samples with missing data according to a user-defined threshold.

2.1.4. Text and Graphical Outputs

Every model generates an unfiltered, filtered, and NA-imputed methylation file for each context (Figure 3). The Emodel output lists the model statistics for each epigenetic marker (rows) in the format "ID | beta | stats | pvalue | FDR," where ID is for chromosome/scaffold names, beta is a beta coefficient in a linear model, stats is the t-statistics for the marker of interest, pvalue is the probabilistic score of an individual marker, and FDR are false discovery rate corrected *p*-values (*q*-values; Figure S5). The output of the Gmodel is a list of marker–SNP pairs, in which the SNP is the appropriate couple to explain the marker of interest. The only different column from the Emodel output is additional "snp" column next to the ID column; "ID | snp | beta | stats | pvalue | FDR" (Figure S6). The output of the GxE is a list of marker–SNP–env triplets where the environment is a factor divided into genotype groups (AA, AB, and BB) to explain the significant marker of interest, and output is otherwise the same as for Gmodel.

The EWAS pipeline provides multiple output plots such as *p*-value Q–Q plots and histograms for all models (Figures S7 and S8), Manhattan plots for Emodel (Figure S9), sequence dot plots for Gmodel (Figure S10), and genotype interaction plots for the GxE (Figure S11). Each visualization is implemented using the ggplot2 package in R (https://github.com/tidyverse/ggplot2, accessed on 1 September 2019).

2.2. Evaluation of the EpiDiverse EWAS Pipeline

Two published datasets of non-model plant species, namely, valley oak (*Quercus lobata*) with 58 samples [45] and Norway spruce (*Picea abies*) with 28 samples (derived from [59] and unpublished data), were used to test the reproducibility of the results and the performance of the EpiDiverse EWAS pipeline.

Q. lobata is a long-lived California endemic tree species with a ~730 Mbp genome [60]. Gugger et al. (2016) used RRBS to analyze whether climate is associated with variation in DNA methylation levels in 58 naturally occurring trees collected across climate gradients [45].

P. abies (Norway Spruce) is also a long-lived (conifer) tree species with a 20 Gbps draft genome [61]. Heer et al. (2018) analyzed eight *P. abies* trees in a targeted bisulfite sequencing approach, employing the SeqCapEpi Kit (NimbleGen). They sampled four trees (ortets) located in Bavaria, Germany, at ~1200 m above sea level (a.s.l.) and four clones that originated from those trees (ramets) planted at ~500 m a.s.l. [59]. In the present study, these data were extended with additional clones to test missing data management, replicate the results of the previous study, and determine the effects of input types. Those additional clones originated either from Germany or Sweden and were planted between 1970 and 1973 at several locations in Germany.

2.2.1. Analysis of *Q. lobata* Dataset

In the original publication, it was suggested that single-methylation variants (SMVs), which are MPs are involved in response to the local environment and the acclimation to a climate in a long-lived tree species, valley oak [45]. The authors found 43 significant SMVs associated using several climatic variables. In total 38, 1, and 1 SMVs in CG, CHG, and CHH context were found to be associated with maximum temperature (tmax). A single CG–SMV associated with the minimum temperature (tmin) and single CHG and CHH SMVs associated with growing season growing degree days above 5 °C (GSDD5) were found to be significant. CG–SMVs showed stronger associations with climatic variables than other types of SMVs and SNPs.

We used this dataset to test whether these findings could be replicated using the EpiDiverse EWAS pipeline (Table 1). When running Emodel, a total of 33 out of 38 tmax related CG–SMVs are shared, and the EWAS pipeline found 47 SMVs in total for this context (Table 2). Likewise, the single tmin-related CG–SMV and the tmax-related one (CHG, CHH), are shared. Results are also similar for GSDD5 in CG context and CWD in CG and CHH context. Hence, there is good agreement between the two analyses with the

EWAS pipeline results containing the majority of the published results, while detecting a few more significant positions, in particular in the CHH context for tmin.

Table 2. Comparison of EpiDiverse EWAS Emodel output for valley oak with the published data.

CG	tmax [1]	tmin [2]	GSDD5 [3]	CWD [4]
Gugger et al., 2016	38	1	0	0
EpiDiverse EWAS pipeline	47	2	0	0
shared amount	33	1	not applicable	not applicable
Shared % based on Gugger et al., 2016	86.42%	100%	100%	100%
CHG				
Gugger et al., 2016	1	0	1	0
EpiDiverse EWAS pipeline	1	0	0	1
shared amount	1	Not applicable	0	0
Shared % based on Gugger et al., 2016	100%	100%	0%	0%
CHH				
Gugger et al., 2016	1	0	1	0
EpiDiverse EWAS pipeline	3	16	0	0
shared amount	1	not applicable	0	not applicable
Shared % based on Gugger et al., 2016	100%	0%	0%	100%

[1] tmax: maximum temperature, [2] tmin: minimum temperature, [3] CWD: (an integrated measure of water availability or stress considering rainfall, evapotranspiration, and basin hydrology), and [4] GSDD5 (growing season growing degree days above 5 °C).

Open reading frames (ORFs) harboring significant MPs related to spatial and climatic variables were blasted against the NCBI nonredundant protein database and the closest hits were analyzed (Table S1). Significant MPs uniquely found by the EpiDiverse EWAS pipeline seem to be connected to relevant studies in the literature both for climatic and spatial variables (cf. Supplementary Section Blastx analysis with the *Q. lobata* dataset).

Some of the differences found between the two methods might be explained by differences in the estimation of missing data. Loci with more than 10% missing data were discarded from the previous analysis [45], and missing data were estimated by the EpiDiverse EWAS pipeline [58].

The previous study [45] used a multivariate method called RDA with a kinship matrix from methylation data. RDA is a forced classification method analogous to linear regression for cases that have multiple dependent variables (e.g., SMVs) and multiple independent variables (e.g., climate and spatial variables). RDA may not always be feasible with few variables, and this is especially true when there is a large proportion of unconstrained variation, i.e., the variation in the response matrix that is nonredundant with the variation in the explanatory matrix. Another thing that one cannot always be sure about which data to use to obtain the kinship matrix. In summary, the two methods perform similarly, with the EpiDiverse EWAS pipeline detecting a few more significantly correlated positions.

2.2.2. Analysis of *P. abies* Dataset

Heer et al. (2018) hypothesized that the methylation percentage between clones from the two environments at a global level was similar and proposed that the methylation patterns remained largely stable during the life history of the trees [59].

We tested whether the sampling locations differ in terms of climatic variables. Due to a violation of normality using the Shapiro–Wilk test with 0.05 *p*-value, we carried out a nonparametric Wilcoxon test to compare locations between Goeppingen, Harsefeld, Neuhaus for clones apart from ramets, Bavarian forest national park for ortets, and Übersee for ramets in Germany. Precipitation (prcp) is significantly different between all sites (Figure S12 upper, left), whereas maximum (tmax) (Figure S12 upper, right), and minimum temperatures (lower) (Figure S12c) were found to be different only between some.

Independently from the EpiDiverse EWAS pipeline, coalescence analysis between the SNP (genetic) and the methylation data (epigenetic) was performed to determine whether the samples are congruent and also to narrow down pairwise comparisons for DMP and DMR calling. Coalescence analysis with CG context showed a highly dissimilar pattern between the SNP and averaged methylation data per sample for the clones apart from ramets (Figure 4, please see Figures S13–S18 for other contexts and non-averaged methylation dendrograms). Since the similarity trees are based on genetic and epigenetic variance were dissimilar, three clustering approaches were employed to call DMPs/DMRs based on (i) trees' locations, (ii) SNP clustering, and (iii) methylation call clustering (Figures S13–S18). In the previous study [59] ramet vs. ortet analysis was performed and revealed potentially interesting results that same ID ortets and ramets clustered together. Hence, this comparison was repeated here as one of the pairwise comparisons. An independent run outside of the EpiDiverse EWAS pipeline was conducted with an unsupervised method, kWIP [62], which found ramet and ortet pairs clustering (Figure S19). This outcome also confirms the same clustering with the previous study [59].

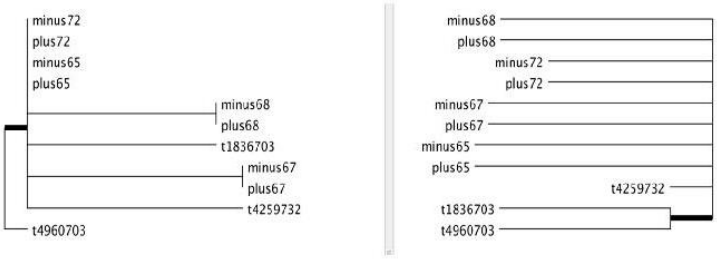


Figure 4. Coalescence analysis between the SNP and methylation data for the CG context. SNP (left) and averaged methylation call values (right) cluster comparison for the CG context. This comparison yielded a 72% topological score indicating a relatively high fraction of clades/branches present in both trees (cf. 3. Methods for details). The thick branches represent deviating topologies. Minus refers to ramets and plus to ortets. Cf. Figures S13–S18 for additional analyses.

DMP/DMR Analysis Considerations Using Different Callers

No significant DMPs were found with $q < 0.1$ filterings by the EpiDiverse DMR pipeline using the default DMR caller metilene [63]. Hence, DMPs between "ramet vs. ortet" were compared with the output of other DMR callers, methylkit [64] and defiant (implemented in the EpiDiverse toolkit) [65], which yielded results with $q < 0.1$ (See Figure S20 and Supplemental Section "DMP/DMR analysis using different callers" for details). If one desires to use DMPs and/or DMRs as input instead of methylated positions (MPs), the alternative solution can be pairwise clustering to reveal differential positions and/or regions and a user has to define which groups to compare. It should be kept in mind that an unsupervised clustering may not always yield proper and distinct groups to achieve DMPs and/or DMRs.

Filtering Missing Data after Uniting Individual Methylomes

Bedtools unionbed function for unifying process creates some missing positions due to, e.g., varying coverage, resulting in some markers being covered in some samples and not in others, cf. Section 2.1.3 for more details (Table S2a,b).

Therefore, we filtered methylated positions' data so that only those positions present across all samples remained. This led to only 7%, 7%, and 5.5% of data remaining in CG, CHG, and CHH contexts, respectively (Table S2c,d,e). To quantify the effect of missing data, we performed an iterative filtering analysis of 0.1 increments with filter_NA parameter using covariates based on the geographic location of trees, methylation, and SNP data (Figures S21–S32). Covariates only with the location of trees and combinations with it

yielded the highest number of intersections between results. SNP and methylation-based covariates showed no significant outputs.

The Intersection of Positions with All Inputs and Models for the CG Context

In order not to bias the data via NA correction, a zero-tolerance missing data threshold was used in all subsequent *P. abies* analyses.

It was shown that gene body CG methylation is relatively stable across seasons [66]. Hence, for this test study, the EWAS run with all models was performed in the CG context only. Significant positions in all model outputs were intersected for location-based clustering using precipitation environment and CG context data with the UpsetR R package [67] for all input types (Figure 5). We selected precipitation for this study because it showed significant differences between all sample locations.

In summary, G and GxE models with DMPs as input is the combination that yields the highest number of significant positions. The maximum number of groups that share a position is seven (check the vertical line on the far right in Figure 5). Moreover, 20, 182, 9713, 37,026, and 77,405 terms are, respectively, shared by five, four, three, two, and single groups. It makes sense to test several inputs and models for higher sensitivity.

Depending on the input type and model, the output in terms of significant positions varies considerably.

Removal of Genetic Variants That Might Be Interpreted as Significant Epigenetic Marks

To determine potentially problematic overlap between genetic and epigenetic variation we intersected all models. Only one position was found to be shared between Emodel MP, G & GxE models DMP input, "MA_160146:1616-1617", i.e., position 1616/17 on *P. abies* contig MA_160146 (Figure S33). A total of 16 SNPs were found to be correlated with this CG, 16 for the GxE and one for the Gmodel, 17 SNPs are in common between G and GxE models, and only five of them are on CG bases. Additionally, those 17 shared SNPs were intersected with the output of the EWAS runs and it was observed that the G and GxE DMRs and averaged outputs have an intersection with these 17 SNPs. In conclusion, the optimal intervention as post hoc analysis should be excluding the intersected CG identifiers from Emodel, in other words, shared positions between Emodel and Gmodel/GxE should be removed from the Emodel output if the aim is to discover epigenetic associations that arise purely due to environmental factors (the same is valid for Gmodel when obtaining solely genetic-related identifiers). Please check the "Removal of genetic variants that might be interpreted as significant epigenetic marks" chapter in Supplementary for more details.

Emodel Output Gene Ontology (GO) Analysis

To determine environmentally associated positions, we conducted an Emodel analysis, integrating all available climatic data and all contexts. Current climatic datasets through the CHELSA database (Climatologies at high resolution for the earth's land surface areas) [68] are maximum, minimum temperature, and precipitation. Since on the level of individual positions, there is scarce overlap between all input types (MPs, DMPs, and DMRs), we decided to analyze the results on the level of Gene Ontology (GO) level (Figures S34–S36) to check concordance in different model outputs and found enriched GO terms for significant cytosines. Such analyses can be useful to reduce noise and to determine the conserved overlap of datasets [69].

To determine whether GO terms are concordant between different input types we concentrated on the usually most meaningful biological process (BP) ontology and found that 47% of terms overlap with the previous study in which only ortets vs. ramets were analyzed [59] (Figure S34). The highest number of significant terms (26) in the Emodel analysis resulted from MP input, CHH context, and tmin climatic data (Figure S37).

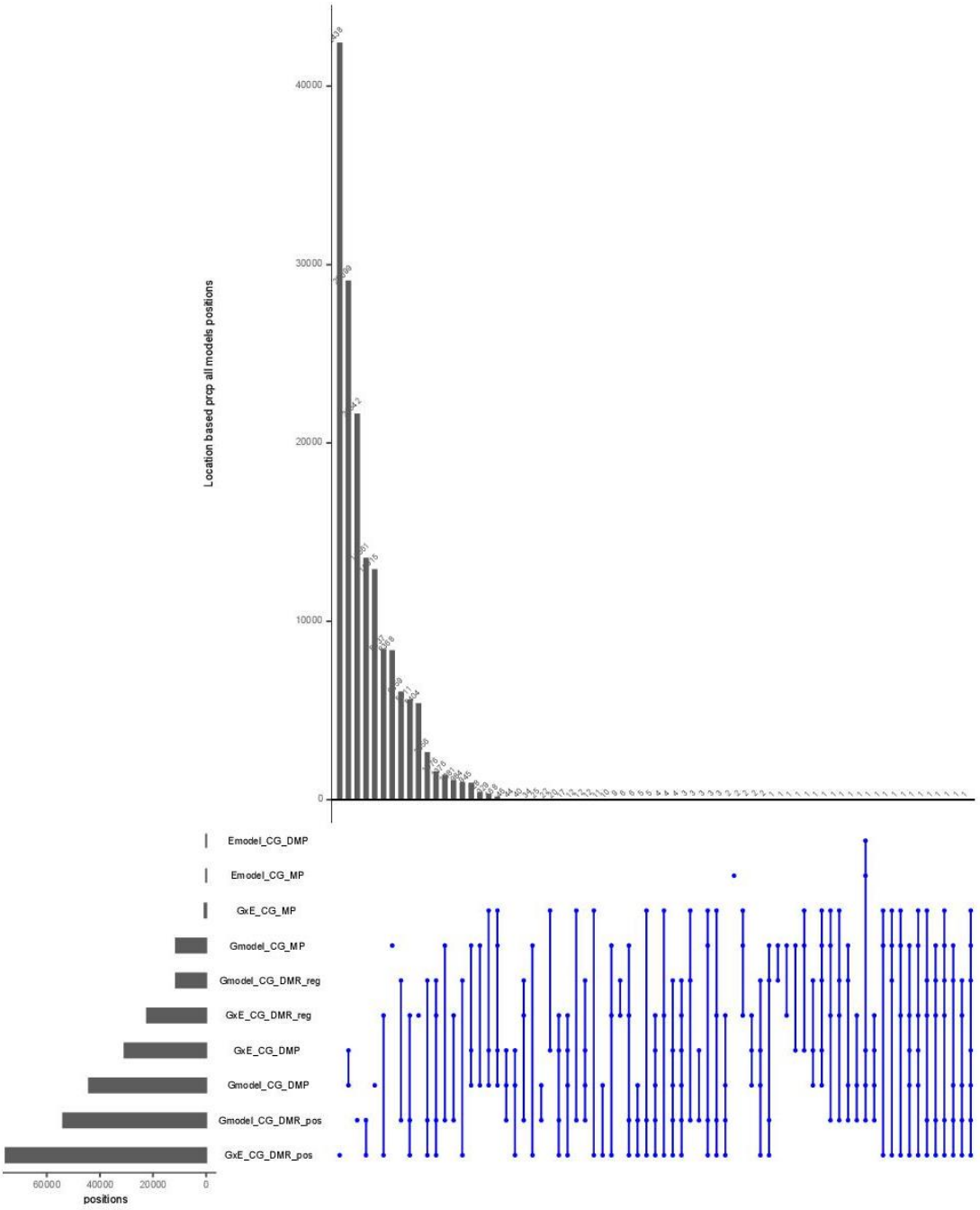


Figure 5. Upset plot of significant positions for all models with CG context using locations of trees as covariates with precipitation environmental data. Vertical lines refer to shared terms between classes on the left side. A maximum shared number of terms is between G and GxE models with DMP input type. Overall, 39% of the terms are shared and 61% are unique to single inputs. The highest number of unique elements are found for GxE DMR input with 42,438 terms, and the lowest is with two terms for the Emodel CG MP input.

CG Context G and GxE GO Analysis

Since the G and GxE models are computationally intensive, we ran them only on precipitation data, which was found to be significantly different between all locations (Figure S12), and in CG context, since it was shown that CG methylation is relatively stable seasonally [66,70]. G and GxE GO outputs (Figures S38–S41) were intersected to distinguish the effect of climatic data vs. SNP data. The filtered output results ($q < 0.05$) for CG context with Gmodel show that all four input types yield significant terms (Table S3). In total, 69% of BP terms are shared and only 31% are unique to single inputs (Figure S38).

Enriched BP terms that are found with G and GxE models include, e.g., "monoterpene," "metabolic organic biopolymer," and "phenylpropanoid." The term "response to temperature stimulus" was found with Gmodel DMP and GxE MP inputs. Other prominent terms that might be related to precipitation were "water transport" (all inputs of GxE, and DMP input with Gmodel) and "water homeostasis" (Gmodel with DMR average method).

The *P. abies* dataset analysis showed that most of the GO terms between different models are shared and only a few are unique to a given model. The term "phenylpropanoid metabolic process" is shared between all outputs except for Gmodel MPs and DMPs. Phenolic extracts in *P. abies* were reported to exhibit antifungal [71], antibacterial [72], and antioxidative activity [73]. Additionally, phenolic compound-related terms were found significantly in higher numbers with needles of damaged trees [74]. It was shown by other studies that the colonization of trees by various bark beetle species was related to the released number of monoterpenes [75]. Terms found with "water transfer" seem to exist in studies for drought resistance [76,77].

In summary, Gmodel with averaged DMRs as input is the combination that yields the highest number of terms. However, it should be noted that the GxE also found a (low) number of unique terms and that DMRs and MPs as inputs also yielded unique terms. Hence, for higher sensitivity or a consensus approach, it makes sense to test several input types and models (Figure 6), in particular since results also vary per context, as shown for the Emodel study (Figure S37).

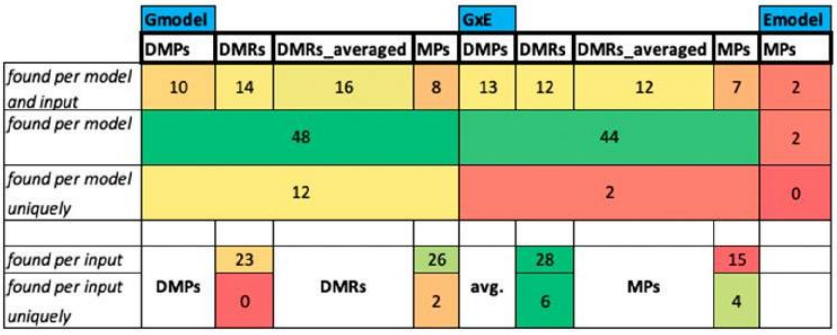


Figure 6. Subset of BP GO terms related to "water", "root", "shoot", and "defense" per input type under three models (G, GxE, and E), in GC context for precipitation. Several BP GO terms matching "water", "root", "shoot", or "defense" are shown per model and input type. Cells are colored from green = high to red = low.

An assumption is needed to decide which groups to compare to derive DMPs and DMRs, but no assumption is required with MP input type. Covariates are used for grouping samples, and the user may prefer multiple of them. Computation time may take longer while using the MP type input, but it should be noted that using methylated positions are suggested to be used as default in case a user does not have differential methylation values, and it would be advantageous to use the whole (unbiased) methylation data. Additionally, a user should consider the time that will be spent to obtain the differential methylation before executing EWAS. Therefore, we recommend MP input for a start and in absence of

concrete ideas (such as ramet vs. ortet in the *P. abies* dataset) for pairwise groupings. If those are present, averaged DMRs might be preferable based on the number of terms that can be derived from them (Figure 6).

2.3. Conclusions

The EpiDiverse EWAS pipeline allows the analysis of MPs and differential methylation data. It presents logical missing data imputation with beta distribution and produces multiple graphs with each model in the GEM package to help the user to observe results better. We reanalyzed data published previously [45] and found a significant amount of overlap in terms of significant MPs related to spatial and climatic variables.

In terms of the *Q. lobata* dataset [45], we found that nearly all significant C's could be reproduced, although the underlying statistical methods differ. Missing data estimation as implemented in the EpiDiverse EWAS pipeline suggests that beta distribution is a robust and accurate choice for approximation, as inferred from the significant amount of overlap. Most importantly, nearly all of the unique C's only found by the EWAS pipeline seem to have meaningful associations with spatial and climate variables in the literature.

We used the *P. abies* dataset to determine the overlap between different GEM models and input types. We found that the choice of model and input depends on the user's research question. G and GxE models detected more significant GO terms, compared to the Emodel terms in GC context (Figure 6, and averaged DMRs are superior to the other input types in terms of how many terms can be detected. As a hierarchical controlled vocabulary, gene ontology helps to group meaningful biological functions that might be missed in individual gene descriptions. Different genes related to the same biological function may have GO terms in common. Finding most of the GO terms overlapping between different analyses shows a large part of the findings of these analyses are shared on the level of the ontological vocabulary and its underlying functionality, e.g., the biological process enacted. Yet, Emodel found the highest number of terms in the CHH context (Figure S36). Most of the detected GO terms overlapped between different models, inputs, and contexts, suggesting robust results regardless of the model and, to some extent, input type. However, most models and input type combinations yield a certain fraction of uniquely found terms, suggesting that a consensus approach (using several models and input types and using their overlap) might make sense (Figure 6).

In terms of input filtering, we found that *p*-value-filtered metilene DMPs do not lead to severely different results from using *q*-value-filtered methylkit or defiant DMPs. Unsupervised clustering using kWIP to derive groups for DMP and DMR analysis was found to be a potential replacement for a priori grouping.

In summary, we present the EpiDiverse EWAS pipeline as a versatile tool to perform plant EWAS analyses, either using the output of the other EpiDiverse pipelines or custom data.

3. Materials and Methods

3.1. The EpiDiverse EWAS Pipeline

The EpiDiverse EWAS pipeline is available on GitHub (https://github.com/EpiDiverse/ewas/tree/master, accessed on 1 March 2021). The pipeline was set up using Nextflow 20.07.1 revisions [54] and is based on the GEM R package [50].

3.2. Analysis of Q. lobata Data

As described in the original publication, mature leaves from *Q. lobata* were sampled at each of 58 locations spread along the foothills of the Coastal and Sierra Nevada ranges [45]. Positions with more than 10% missing data, less than 10X coverage, and a 10% range of variation (the difference between the maximum and minimum methylation per position) were filtered out. The authors considered four different climate variables and integrated amount of water availability or stress, considering evapotranspiration, basin hydrology, and rainfall, mean minimum temperature of the coldest month (tmin), mean maximum

temperature of the warmest month (tmax), and growing season growing degree days above 5°C (GSDD5). A multivariate method called redundancy analysis (RDA) was employed to test the variation explained by SMVs and SNPs, and positions were filtered with multiple testing ($q < 0.1$).

To mirror these analyses, the EWAS pipeline run was performed with 10× coverage, q value < 0.1, with a maximum of 10% of missing data, and different standard deviation values per position (0.028, 0.0176, and 0.0197 for CG, CHG, and CHH, respectively) to replicate the results in the previously published study [45]. These parameters produced the same amount of data produced in Gugger et al. (2016) with negligible differences in terms of FDR calculation (Figure S41).

3.3. Analysis of P. abies Data

The *P. abies* dataset was chosen to perform a comprehensive test for measuring the performance and parameters of the EWAS pipeline. In total, 28 samples were used, composed of four original trees or ortets (ID = 65, 67, 68, 72), four clones that originated from those ortets or ramets (ID = 65, 67, 68, 72), and 20 clone trees originated from three trees (ID = 4259732, 4960703, 186370), two located in Germany and one in Sweden. Clones from these trees were planted by the Northwest German Forest Research Institute as part of their project "fit for clim" (https://www.fitforclim.de/, accessed on 1 March 2021) with varying climatic conditions in Germany, namely, Neuhaus (with unique numbers or Mitte/middle, oben/up, unten/low extensions based on the position in the tree they were sampled from), Göppingen (G extension), and Harsefeld (H extension) (Figure S42).

The EpiDiverse WGBS pipeline with the segemehl standalone tool [66] was used for methylation calling with the options –noDedup, –SE, and –unique, using the high confidence gene set as reference [52]. Overall, 15 to 85 million reads per sample were left after trimming, yielding a coverage of 8- to 26-fold (Table S4). Although the duplicate ratio was quite high due to the linear relationship of sequencing depth and duplicate level, this is probably not due to PCR bias (see Supplement, Ratio of PCR Duplicates for *P. abies* Datasets, Figure S43).

The EpiDiverse DMR pipeline with the metilene software was used to call DMRs with default settings and –sig 0.1 and –diff 20 for DMPs. The EpiDiverse SNP pipeline was used with the –variants parameter to derive SNP variants per sample as separated .vcf files. All files were compressed with bgzip, indexed, and finally merged and filtered to keep variants that have been successfully genotyped in 100% of individuals, a minimum quality score of 30, and a minor allele count of three. This final file was used with the –SNPs parameter for the EWAS pipeline run. The EpiDiverse EWAS pipeline takes individual variant files to merge and filter them.

The EpiDiverse EWAS pipeline run was conducted separately for each methylation context while disabling the other two, e.g., –noCHH –noCHG parameters used for a CG run, –distance parameter was set to 2000, and –coverage to 5. G and GxE models were run in 10 separate runs, and all positions with noninfinite t statistics were discarded. Separate outputs were merged, and FDR calculation was carried out.

Hierarchical clustering (HC) with the Euclidean distance method with ctc package in R was performed on genetic variability (SNPs) and on epigenetic variability (methylated positions). The idea was that if there is no difference between the two resulting topologies, either of them might be used to derive groups. Coalescence analysis was performed to check for commonalities/differences of the topologies, and methylation calls and SNP HC graphs were compared via the compare2trees standalone tool [67]. The SNP tree is composed of 11 samples, including the four ramets, the four ortets (ramets and ortets are assumed to be genetically identical), and the three new clones. The methylation data comprises 28 samples: 4 from the ramets, 4 from the ortets, and 20 from the new clones. Two versions of the methylation tree were used, with all samples separate, and using averaged methylation per position for the same tree ID located in different locations. Branch thickness in the result of the compare2trees software is used to show the topological

score, which is the percentage fraction of clades/branches that are present in the actual tree that is also present in the estimated tree; thicker branches refer to a lower score (Figure 4 and Figures S13–S18). The kWIP tool was used to cluster methylation data for four ramets and ortets.

GO bias analyses were performed with the GOSTAT pipeline while intersecting the results from the EWAS pipeline with GO term annotations [68].

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/epigenomes5020012/s1, Figure S1: WGBS output directory and a bedGraph formatted dataset, Figure S2: Methylome input and the final file after bedtools unionbedg process, Figure S3: DMPs or DMRs output directory and a bed formatted dataset, Figure S4: Required inputs to run EpiDiverse EWAS pipeline with different models, Figure S5: The output of the Emodel, Figure S6: The output of the Gmodel, Figure S7: Q–Q plots, Figure S8: Histogram plots, Figure S9: Manhattan plots, Figure S10: Sequence dot plots, Figure S11: Top significant k-plots, Figure S12: Nonparametric Wilcoxon test to compare climatic datasets for locations of trees, Figure S13: Coalescence analysis with SNP and averaged methylation data for the CG context, Figure S14: Coalescence analysis with SNP and not averaged methylation data for the CG context, Figure S15: Coalescence analysis with SNP and averaged methylation data for the CHG context, Figure S16: Coalescence analysis with SNP and not averaged methylation data for the CHG context, Figure S17: Coalescence analysis with SNP and averaged methylation data for the CHH context, Figure S18: Coalescence analysis with SNP and not averaged methylation data for the CHH context, Figure S19: fastq raw files HC with k32 performed by kWIP software, Figure S20: Intersection of significant C's with *p*- and *q*-values on gene level for methylkit, metilene, and defiant DMR callers., Figure S21: Intersection of outputs with different filter_NA values for MPs input using all covariates, Figure S22: Intersection of outputs with different filter_NA values for DMPs input using all covariates, Figure S23: Intersection of outputs with different filter_NA values for DMRs input using all covariates, Figure S24: Intersection of outputs with different filter_NA values for MPs input using only location-methylation-based covariates, Figure S25: Intersection of outputs with different filter_NA values for DMPs input using only location-methylation-based covariates, Figure S26: Intersection of outputs with different filter_NA values for MPs input using only location-SNP-based covariates, Figure S27: Intersection of outputs with different filter_NA values for DMPs input using only location-SNP-based covariates, Figure S28: Intersection of outputs with different filter_NA values for DMRs input using only location-SNP-based covariates, Figure S29: Intersection of outputs with different filter_NA values for MPs input using only SNP-methylation-based covariates, Figure S30: Intersection of outputs with different filter_NA values for MPs input using only location-based covariates, Figure S31: Intersection of outputs with different filter_NA values for DMPs input using only location-based covariates, Figure S32: Intersection of outputs with different filter_NA values for DMRs input using only location-based covariates, Figure S33: Intersection of shared SNPs and significant common markers between G and GxE models, Figure S34: Intersection of significant BP GO terms between location-based Emodel output and a previous study [5] with UpsetR package, Figure S35: Intersection of significant MF GO terms between location-based Emodel output and a previous study with UpsetR package, Figure S36: Intersection of significant CC GO terms between location-based Emodel output and a previous study with UpsetR package, Figure S37: Highlighted GO terms based on Emodel, Figure S38: Intersection of significant BP GO terms between all models, only CG context and precipitation data for location-based clustering, and a previous study with UpsetR package, Figure S39: Intersection of significant MF GO terms between all models, only CG context and precipitation data for location-based clustering, and a previous study with UpsetR package, Figure S40: Intersection of significant CC GO terms between all models, only CG context and precipitation data for location-based clustering, and a previous study with UpsetR package, Figure S41: Gugger et al. (2016) methylation and climatic data processing and analysis by the EpiDiverse EWAS pipeline, Figure S42: Location of *P. abies* trees (a), additional clone information (b), and grouping of trees (c), Figure S43: PCR duplicate analysis, Table S1: *Q. lobata* blastx analysis, Table S2: Missing data estimation of EpiDiverse EWAS pipeline (a) and GEM R package (b) [9]. Missing data statistics for P. abies dataset with CG (c), CHG (d), and CHH (e) contexts, Table S3: EWAS output and GO statistics, Table S4: Statistics of additional *P. abies* samples.

Author Contributions: Conceptualization and design: S.N.C., S.A.R., and N.F.-P.; methodology: S.N.C., A.N., S.A.R., and N.F.-P.; software: S.N.C., A.N., and D.G.; investigation: S.N.C., S.A.R., and

N.F.-P.; writing—original draft preparation: S.N.C., S.A.R., and N.F.-P.; supervision: S.A.R., N.F.-P., K.H., L.O. and D.L.; data generation and curation: K.V., K.H., L.O., and C.B.; revising of the Article for intellectual content: S.N.C., S.A.R., and N.F.-P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported and funded by the European Training Network EpiDiverse (https://epidiverse.eu, accessed on 1 March 2021), which received funding from the EU Horizon 2020 program under Marie Skłodowska-Curie grant agreement No 764965. K.H. is grateful for funded by the DFG (HE 7345/2-1) for exome capture and sequencing.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The pipeline is available at https://github.com/EpiDiverse/ewas/tree/master, accessed on 1 March 2021. BAM Files containing the mapped reads for ramet and ortets are available at European Nucleotide Archive (ENA, www.ebi.ac.uk/ena/, accessed on 1 September 2019) under the project PRJEB26494, raw read fastq accessions under ERR2591764:ERR2591771. Mapped reads for the other 20 clone trees are under the project PRJNA703787, raw read fastq accessions are under the SRA run accessions SRR13760855: SRR13760874.

Acknowledgments: We would like to thank Pan Hong, developer of the GEM package, for her valuable and constructive suggestions during the development of the pipeline and Alwin Janssen for help with the *P. abies* sampling. We also would like to thank Iris Sammarco, Bárbara Díez Rodríguez, and Marc W. Schmid for discussions, as well as Paul F. Gugger and Sorel Fitz-Gibbon for sharing the concatenated and reversed reference scaffolds for analysis of *Q. lobata* datasets.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fuchs, J.; Demidov, D.; Houben, A.; Schubert, I. Chromosomal histone modification patterns—From conservation to diversity. *Trends Plant Sci.* **2006**, *11*, 199–208. [CrossRef] [PubMed]
2. Weinhold, B. Epigenetics: The Science of Change. *Environ. Heal Perspect.* **2006**, *114*, A160–A167. [CrossRef]
3. Sudan, J.; Raina, M.; Singh, R. Plant epigenetic mechanisms: Role in abiotic stress and their generational heritability. *Biotech* **2018**, *8*, 172. [CrossRef] [PubMed]
4. Quadrana, L.; Colot, V. Plant Transgenerational Epigenetics. *Annu. Rev. Genet.* **2016**, *50*, 467–491. [CrossRef]
5. Weigel, D.; Colot, V. Epialleles in plant evolution. *Genome Biol.* **2012**, *13*, 249. [CrossRef]
6. Cubas, P.; Vincent, C.; Coen, E. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **1999**, *401*, 157–161. [CrossRef]
7. McClintock, B. Genetic Control of Differentiation, Brookhaven Symposia in Biology. In *The Control of Gene Action in Maize*; Royal Society: London, UK, 1965; Volume 18, pp. 162–184.
8. Manning, K.; Tör, M.; Poole, M.; Hong, Y.; Thompson, A.J.; King, G.J.; Giovannoni, J.J.; Seymour, G.B. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat. Genet.* **2006**, *38*, 948–952. [CrossRef]
9. Mirouze, M.; Paszkowski, J. Epigenetic contribution to stress adaptation in plants. *Curr. Opin. Plant Biol.* **2011**, *14*, 267–274. [CrossRef]
10. McCue, A.D.; Nuthikattu, S.; Reeder, S.H.; Slotkin, R.K. Gene Expression and Stress Response Mediated by the Epigenetic Regulation of a Transposable Element Small RNA. *PLoS Genet.* **2012**, *8*, e1002474. [CrossRef] [PubMed]
11. Chinnusamy, V.; Zhu, J.K. Epigenetic regulation of stress responses in plants. *Curr. Opin. Plant Biol.* **2009**, *12*, 133–139. [CrossRef]
12. Kim, D.H.; Sung, S. Accelerated vernalization response by an altered PHD-finger protein in Arabidopsis. *Plant Signal. Behav.* **2017**, *12*, e1308619. [CrossRef]
13. Yi, S.V. Insights into Epigenome Evolution from Animal and Plant Methylomes. *Genome Biol. Evol.* **2017**, *9*, 3189–3201. [CrossRef]
14. Colot, V.; Rossignol, J.L. Eukaryotic DNA methylation as an evolutionary device. *Bioessays* **1999**, *21*, 402–411. [CrossRef]
15. Feng, S.; Jacobsen, S.E.; Reik, W. Epigenetic Reprogramming in Plant and Animal Development. *Science* **2010**, *330*, 622–627. [CrossRef] [PubMed]
16. Gent, J.I.; Dong, Y.; Jiang, J.; Dawe, R.K. Strong epigenetic similarity between maize centromeric and pericentromeric regions at the level of small RNAs, DNA methylation and H3 chromatin modifications. *Nucleic Acids Res.* **2011**, *40*, 1550–1560. [CrossRef]
17. Lister, R.; O'Malley, R.C.; Tonti-Filippini, J.; Gregory, B.D.; Berry, C.C.; Millar, A.H.; Ecker, J.R. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **2008**, *133*, 523–536. [CrossRef] [PubMed]
18. Law, J.A.; Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **2010**, *11*, 204–220. [CrossRef]

19. Gu, H.; Smith, Z.D.; Bock, C.; Boyle, P.; Gnirke, A.; Meissner, A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc.* **2011**, *6*, 468–481. [CrossRef]
20. Kishore, K.; Pelizzola, M. Identification of Differentially Methylated Regions in the Genome of Arabidopsis thaliana. *Methods Mol. Biol.* **2018**, *1675*, 61–69.
21. Tsai, P.C.; Bell, J.T. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *Int. J. Epidemiol.* **2015**, *44*, 1429–1441.
22. Rakyan, V.K.; Down, T.A.; Balding, D.J.; Beck, S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **2011**, *12*, 529–541. [CrossRef]
23. Tam, V.; Patel, N.; Turcotte, M.; Bossé, Y.; Paré, G.; Meyre, D. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **2019**, *20*, 467–484. [CrossRef] [PubMed]
24. Bush, W.S.; Moore, J.H. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.* **2012**, *8*, e1002822.
25. Visscher, P.M.; Brown, M.A.; McCarthy, M.I.; Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **2012**, *90*, 7–24. [CrossRef] [PubMed]
26. Visscher, P.M.; Wray, N.R.; Zhang, Q.; Sklar, P.; McCarthy, M.I.; Brown, M.A.; Yang, J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **2017**, *101*, 5–22. [CrossRef]
27. Cortes, L.T.; Zhang, Z.; Yu, J. Status and prospects of genome-wide association studies in plants. *Plant Genome* **2021**, *14*, e20077. [CrossRef]
28. Ersoz, E.S.; Yu, J.; Buckler, E. Applications of Linkage Disequilibrium and Association Mapping in Crop Plants. In *Genomics-Assisted Crop Improvement*; Varshney, R.K., Tuberosa, R., Eds.; Springer: Dordrecht, The Netherlands, 2007; pp. 97–119.
29. Liu, H.-J.; Yan, J. Crop genome-wide association study: A harvest of biological relevance. *Plant J.* **2019**, *97*, 8–18. [CrossRef]
30. Sukumaran, S.; Yu, J. Association Mapping of Genetic Resources: Achievements and Future Perspectives. *Genom. Plant Genet. Resour.* **2013**, 207–235. [CrossRef]
31. Varshney, R.K.; Ribaut, J.-M.; Buckler, E.S.; Tuberosa, R.; Rafalski, J.A.; Langridge, P. Can genomics boost productivity of orphan crops? *Nat. Biotechnol.* **2012**, *30*, 1172–1176. [CrossRef]
32. Gupta, P.K.; Kulwal, P.L.; Jaiswal, V. Association mapping in plants in the post-GWAS genomics era. *Adv. Genet.* **2019**, *104*, 75–154. [CrossRef]
33. Chen, E.; Huang, X.; Tian, Z.; Wing, R.A.; Han, B. The genomics of Oryza species provides insights into rice domestication and heterosis. *Annu. Rev. Plant Biol.* **2019**, *70*, 639–665. [CrossRef]
34. Wang, S.-B.; Feng, J.-Y.; Ren, W.-L.; Huang, B.; Zhou, L.; Wen, Y.-J.; Zhang, J.; Dunwell, J.M.; Xu, S.; Zhang, Y.-M. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* **2016**, *6*, 19444. [CrossRef]
35. Zhang, Y.; Massel, K.; Godwin, I.D.; Gao, C. Applications and potential of genome editing in crop improvement. *Genome Biol.* **2018**, *19*, 210. [CrossRef]
36. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367. [CrossRef] [PubMed]
37. Lappalainen, T.; Greally, J.M. Associating cellular epigenetic models with human phenotypes. *Nat. Rev. Genet.* **2017**, *18*, 441–451. [CrossRef]
38. Heard, E.; Martienssen, R.A. Transgenerational epigenetic inheritance: Myths and mechanisms. *Cell* **2014**, *157*, 95–109. [CrossRef] [PubMed]
39. Kalisz, S.; Purugganan, M.D. Epialleles via DNA methylation: Consequences for plant evolution. *Trends Ecol. Evol.* **2004**, *19*, 309–314. [CrossRef] [PubMed]
40. Boyko, A.; Blevins, T.; Yao, Y.; Golubov, A.; Bilichak, A.; Ilnytskyy, Y.; Hollander, J.; Meins, F., Jr.; Kovalchuk, I. Transgenerational adaptation of Arabidopsis to stress requires DNA methylation and the function of Dicer-like proteins. *PLoS ONE* **2010**, *5*, e9514. [CrossRef]
41. Lang-Mladek, C.; Popova, O.; Kiok, K.; Berlinger, M.; Rakic, B.; Aufsatz, W.; Jonak, C.; Hauser, M.-T.; Luschnig, C. Transgenerational Inheritance and Resetting of Stress-Induced Loss of Epigenetic Gene Silencing in Arabidopsis. *Mol. Plant* **2010**, *3*, 594–602. [CrossRef] [PubMed]
42. Latzel, V.; Gonzalez, A.P.R.; Rosenthal, J. Epigenetic Memory as a Basis for Intelligent Behavior in Clonal Plants. *Front. Plant Sci.* **2016**, *7*, 1354. [CrossRef]
43. Paul, D.S.; Beck, S. Advances in epigenome-wide association studies for common diseases. *Trends Mol. Med.* **2014**, *20*, 541–543. [CrossRef] [PubMed]
44. Verma, M. Genome-wide association studies and epigenome-wide association studies go together in cancer control. *Future Oncol.* **2016**, *12*, 1645–1664. [CrossRef] [PubMed]
45. Gugger, P.F.; Fitz-Gibbon, S.; PellEgrini, M.; Sork, V.L. Species-wide patterns of DNA methylation variation in Quercus lobata and their association with climate gradients. *Mol. Ecol.* **2016**, *25*, 1665–1680. [CrossRef]
46. Ong-Abdullah, M.; Ordway, J.M.; Jiang, N.; Ooi, S.-E.; Kok, S.-Y.; Sarpan, N.; Azimi, N.; Hashim, A.T.; Ishak, Z.; Rosli, S.K.; et al. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* **2015**, *525*, 533–537. [CrossRef]

47. Sáez-Laguna, E.; Guevara, M.-Á.; Díaz, L.-M.; Sánchez-Gómez, D.; Collada, C.; Aranda, I.; Cervera, M.-T. Epigenetic Variability in the Genetically Uniform Forest Tree Species Pinus pinea L. *PLoS ONE* **2014**, *9*, e103145. [CrossRef]
48. Rahmani, E.; Yedidim, R.; Shenhav, L.; Schweiger, R.; Weissbrod, O.; Zaitlen, N.; Halperin, E. GLINT: A user-friendly toolset for the analysis of high-throughput DNA-methylation array data. *Bioinformatics* **2017**, *33*, 1870–1872. [CrossRef] [PubMed]
49. Xu, J.; Zhao, L.; Liu, D.; Hu, S.; Song, X.; Li, J.; Lv, H.; Duan, L.; Zhang, M.; Jiang, Q.; et al. EWAS: Epigenome-wide association study software 2.0. *Bioinformatics* **2018**, *34*, 2657–2658. [CrossRef]
50. Pan, H.; Holbrook, J.D.; Karnani, N.; Kwoh, C.K. Gene, Environment and Methylation (GEM): A tool suite to efficiently navigate large scale epigenome wide association studies and integrate genotype and interaction between genotype and environment. *BMC Bioinform.* **2016**, *17*, 1–8. [CrossRef] [PubMed]
51. Richards, C.L.; Alonso, C.; Becker, C.; Bossdorf, O.; Bucher, E.; Colomé-Tatché, M.; Durka, W.; Engelhardt, J.; Gaspar, B.; Gogol-Döring, A.; et al. Ecological plant epigenetics: Evidence from model and non-model species, and the way forward. *Ecol. Lett.* **2017**, *20*, 1576–1590. [CrossRef]
52. Nunn, A.; Otto, C.; Stadler, P.F.; Langenberger, D. Comprehensive benchmarking of software for mapping whole genome bisulfite data: From read alignment to DNA methylation analysis. *Briefings Bioinform.* **2021**, *10*. [CrossRef]
53. Kreutz, C.; Can, N.S.; Bruening, R.S.; Meyberg, R.; Mérai, Z.; Fernandez-Pozo, N.; Rensing, S.A. A blind and independent benchmark study for detecting differentially methylated regions in plants. *Bioinformatics* **2020**, *36*, 3314–3321. [CrossRef]
54. Di Tommaso, P.; Chatzou, M.; Floden, E.W.; Barja, P.P.; Palumbo, E.; Notredame, C. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **2017**, *35*, 316–319. [CrossRef] [PubMed]
55. Kurtzer, G.M.; Sochat, V.; Bauer, M.W. Singularity: Scientific containers for mobility of compute. *PLoS ONE* **2017**, *12*, e0177459. [CrossRef]
56. Merkel, D. Docker: Lightweight Linux containers for consistent development and deployment. *Linux. J.* **2014**, *2014*, 1075–3583.
57. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef] [PubMed]
58. Raineri, E.; Dabad, M.; Heath, S. A Note on Exact Differences between Beta Distributions in Genomic (Methylation) Studies. *PLoS ONE* **2014**, *9*, e97349. [CrossRef]
59. Heer, K.; Ullrich, K.K.; Hiss, M.; Liepelt, S.; Brüning, R.S.; Zhou, J.; Opgenoorth, L.; Rensing, S.A. Detection of somatic epigenetic variation in Norway spruce via targeted bisulfite sequencing. *Ecol. Evol.* **2018**, *8*, 9672–9682. [CrossRef] [PubMed]
60. Sork, V.L.; Fitz-Gibbon, S.T.; Puiu, D.; Crepeau, M.; Gugger, P.F.; Sherman, R.; Stevens, K.; Langley, C.H.; Pellegrini, M.; Salzberg, S.L. First Draft Assembly and Annotation of the Genome of a California Endemic OakQuercus lobataNée (Fagaceae). *G3 Genes Genomes Genet.* **2016**, *6*, 3485–3495. [CrossRef]
61. Nystedt, B.; Street, N.R.; Wetterbom, A.; Zuccolo, A.; Lin, Y.-C.; Scofield, D.G.; Vezzi, F.; Delhomme, N.; Giacomello, S.; Alexeyenko, A.; et al. The Norway spruce genome sequence and conifer genome evolution. *Nature* **2013**, *497*, 579–584. [CrossRef]
62. Murray, K.D.; Webers, C.; Ong, C.S.; Borevitz, J.; Warthmann, N. kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLoS Comput. Biol.* **2017**, *13*, e1005727. [CrossRef]
63. Jühling, F.; Kretzmer, H.; Bernhart, S.H.; Otto, C.; Stadler, P.F.; Hoffmann, S.D. Metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res.* **2016**, *26*, 256–262. [CrossRef]
64. Akalin, A.; Kormaksson, M.; Li, S.; E Garrett-Bakelman, F.; E Figueroa, M.; Melnick, A.; E Mason, C. methylKit: A comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **2012**, *13*, R87. [CrossRef] [PubMed]
65. Condon, D.E.; Tran, P.V.; Lien, Y.-C.; Schug, J.; Georgieff, M.K.; Simmons, R.A.; Won, K.-J. Defiant: (DMRs: Easy, fast, identification and ANnoTation) identifies differentially Methylated regions from iron-deficient rat hippocampus. *BMC Bioinform.* **2018**, *19*, 1–12. [CrossRef]
66. Ito, T.; Nishio, H.; Tarutani, Y.; Emura, N.; Honjo, M.N.; Toyoda, A.; Fujiyama, A.; Kakutani, T.; Kudoh, H. Seasonal Stability and Dynamics of DNA Methylation in Plants in a Natural Environment. *Genes* **2019**, *10*, 544. [CrossRef]
67. Conway, J.R.; Lex, A.; Gehlenborg, N. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **2017**, *33*, 2938–2940. [CrossRef] [PubMed]
68. Karger, D.N.; Conrad, O.; Böhner, J.; Kawohl, T.; Kreft, H.; Soria-Auza, R.W.; Zimmermann, N.E.; Linder, H.P.; Kessler, M. Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **2017**, *4*, 170122. [CrossRef] [PubMed]
69. Wilhelmsson, P.K.I.; Chandler, J.O.; Fernandez-Pozo, N.; Graeber, K.; Ullrich, K.K.; Arshad, W.; Khan, S.; Hofberger, J.A.; Buchta, K.; Edger, P.P.; et al. Usability of reference-free transcriptome assemblies for detection of differential expression: A case study on Aethionema arabicum dimorphic seeds. *BMC Genom.* **2019**, *20*, 1–19. [CrossRef]
70. Mathieu, O.; Reinders, J.; Čaikovski, M.; Smathajitt, C.; Paszkowski, J. Transgenerational Stability of the Arabidopsis Epigenome Is Coordinated by CG Methylation. *Cell* **2007**, *130*, 851–862. [CrossRef]
71. Belt, T.; Altgen, M.; Mäkelä, M.; Hänninen, T.; Rautkari, L. Cellular level chemical changes in Scots pine heartwood during incipient brown rot decay. *Sci. Rep.* **2019**, *9*, 1–10. [CrossRef] [PubMed]
72. Välimaa, A.-L.; Honkalampi-Hämäläinen, U.; Pietarinen, S.; Willför, S.; Holmbom, B.; Von Wright, A. Antimicrobial and cytotoxic knotwood extracts and related pure compounds and their effects on food-associated microorganisms. *Int. J. Food Microbiol.* **2007**, *115*, 235–243. [CrossRef] [PubMed]
73. Kähkönen, M.P.; Hopia, A.I.; Vuorela, H.J.; Rauha, J.P.; Pihlaja, K.; Kujala, T.S.; Heinonen, M. Antioxidant activity of plant extracts containing phenolic compounds. *J. Agric. Food Chem.* **1999**, *47*, 3954–3962. [CrossRef] [PubMed]

74. Ganthaler, A.; Stöggl, W.; Mayr, S.; Kranner, I.; Schüler, S.; Wischnitzki, E.; Sehr, E.M.; Fluch, S.; Trujillo-Moya, C. Association genetics of phenolic needle compounds in Norway spruce with variable susceptibility to needle bladder rust. *Plant Mol. Biol.* **2017**, *94*, 229–251. [CrossRef] [PubMed]
75. Zhao, T.; Krokene, P.; Björklund, N.; Långström, B.; Solheim, H.; Christiansen, E.; Borg-Karlson, A.-K. The influence of Ceratocystis polonica inoculation and methyl jasmonate application on terpene chemistry of Norway spruce, Picea abies. *Phytochemistry* **2010**, *71*, 1332–1341. [CrossRef] [PubMed]
76. Kohler, M.; Kunz1, J.; Herrmann, J.; Hartmann, P.; Jansone, L.; Puhlmann, H.; Wilpert, K.V.; Bauhus, J. The Potential of Liming to Improve Drought Tolerance of Norway Spruce [*Picea abies* (L.) Karst.]. *Front. Plant Sci.* **2019**, *10*, 382. [CrossRef]
77. Kivimäenpää, M.; Sutinen, S.; Karlsson, P.E.; Selldén, G. Cell Structural Changes in the Needles of Norway Spruce Exposed to Long-term Ozone and Drought. *Ann. Bot.* **2003**, *92*, 779–793. [CrossRef]

3.2.2 Further applicability of this work

The pipeline is available under https://github.com/EpiDiverse/ewas (accessed on 1 May 2021) including test datasets for making it possible for anyone to access and analyze their data. It is used by other projects namely RP06, RP07, and RP08 in the EpiDiverse consortium as in collaboration with my project. Those projects aim to study the natural DNA methylation variation and to elucidate its association with climatic and/or phenotypic variation using non-model species like *F. vesca*, *T. arvense*, and *P. nigra*, respectively.

3.3 EpiDiverse Toolkit: a pipeline suite for the analysis of ecological plant epigenetics (Paper III)

The broadening view of next-generation sequencing studies with plant epigenetics brings its computational challenges too. Existing tools with model species may not help users to utilize these methods in non-model species. Therefore, we developed a toolkit for plant ecologists to work with WGBS data. The toolkit serves bioinformatics pipelines for mapping, calling of methylation values and differential methylation between chosen groups, performs EWAS and a novel application to call variants.

3.3.1 Paper

Following is the electronic publication.

Published online DD MM YYYY *NAR Genomics and Bioinformatics, YYYY, Vol. xx, No. xx* **1–5**
doi:10.1093/nar/gkn000

EpiDiverse Toolkit: a pipeline suite for the analysis of ecological plant epigenetics

Adam Nunn [1,2], Sultan Nilay Can [3], Christian Otto [1], Mario Fasold [1], Bárbara Díez Rodríguez [3], Noe Fernandez-Pozo [3], Stefan A. Rensing [3], Peter F. Stadler [2] and David Langenberger [1,*]

[1]ecSeq Bioinformatics GmbH, Leipzig, 04103, Germany, [2]Institute for Computer Science, University of Leipzig, Leipzig, 04107, Germany and [3]Deparment of Biology, University of Marburg, Marburg, 35043, Germany

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT

The expanding scope and scale of next generation sequencing experiments in ecological plant epigenetics brings new challenges for computational analysis. Existing tools built for model data may not address the needs of users looking to apply these techniques to non-model species, particularly on a population or community level. Here we present a toolkit suitable for plant ecologists working with whole genome bisulfite sequencing; it includes pipelines for mapping, the calling of methylation values and differential methylation between groups, epigenome-wide association studies, and a novel implementation for both variant calling and discriminating between genetic and epigenetic variation.

INTRODUCTION

Model organisms such as *Arabidopsis thaliana* have helped lay the foundation for our understanding of plant epigenetics(1, 2, 3). Now, the increasingly competitive costs of next generation sequencing (NGS) have opened the door for plant ecologists to apply these lessons on the population and community level, to gain more specific insight into non-model species(4). The EpiDiverse Toolkit addresses the challenges of expanding scope and scale for existing computational techniques, with a suite of pipelines for the analysis of DNA methylation from bisulfite sequencing (bs-seq; methylC-seq) data.

During NGS library preparation, sodium bisulfite treatment facilitates the conversion of unmethylated cytosine to uracil while leaving 5-methylcytosine (5mC) positions intact(5). This necessitates specialised or adapted tools to carry out conventional downstream procedures such as mapping(6) and variant calling(7). For non-model plant species this is further confounded by poor quality reference genomes, with additional difficulties due in part to a high tolerance for polyploidy and high rates of heterozygosity(8). Finally, DNA methylation can occur in additional sequence contexts (CHG, CHH) which in contrast to CG are not prevalent in mammalian data(9).

The tools presented herein (Figure 1) are implemented with Nextflow(10), building on best-practice concepts outlined by nf-core(11). They are intended to be efficient, intuitive for novice users, optimisable for laptop, HPC cluster or the cloud, and scalable from small lab studies to field trials with large populations. Dependencies are as simple as installing Nextflow alongside one of either Bioconda(12), Docker, or Singularity on a POSIX compatible system, facilitating a high level of reproducibility. The platform will be maintained and expanded upon as new tools are developed in the future.

MATERIALS AND METHODS

Test Data

A subset of 23 independent, whole genome bisulfite sequencing libraries (150 bp long paired-end reads) of the deciduous tree species *Populus nigra* were selected from the repository hosted by the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB44879 (https://www.ebi.ac.uk/ena/browser/view/PRJEB44879). The libraries were sequenced under the broader initiative of the EpiDiverse consortium according to the procedures outlined by Díez Rodríguez *et al.* (manuscript in prep.). This subset represents two clone populations (Supplementary Table S1) derived from cuttings originating from field sites in Germany and Lithuania and cultivated together under common garden conditions. Measurements of leaf flavonol content from the parent generation were derived from observations taken in the field by Díez Rodríguez *et al.* (manuscript in prep.). The reference genome was obtained from the repository hosted by the ENA at EMBL-EBI under accession number PRJEB44889 (https://www.ebi.ac.uk/ena/browser/view/PRJEB44889).

*To whom correspondence should be addressed. Tel: +49 341 425 891 99; Fax: +49 341 33187-962; Email: david.langenberger@ecseq.com

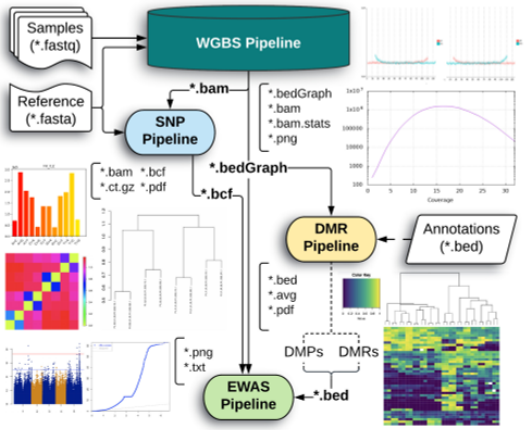


Figure 1. Overview of the EpiDiverse Toolkit. The WGBS data forms the foundation of the analysis, and each downstream pipeline is built to work either in cooperation with one another or, optionally, with independently-generated input data. All pipelines output runtime metadata, tracing and further visualisation in addition to what is shown here. The full output is described for each pipeline in the documentation on Github.

Whole Genome Bisulfite Sequencing (WGBS)

Mapping of bisulfite sequencing data can be carried out either in 'high-throughput mode', with a low memory footprint and a runtime suitable for rapid analysis of population data, or 'high-sensitivity mode', with a demonstrable improvement in precision-recall and downstream methylation analysis for non-model plant species(13). Multiple samples can be processed in parallel, and quality control (QC) is performed with a combination of published tools and in-house scripting. Methylation values based on coverage are called with MethylDackel (https://github.com/dpryan79/MethylDackel), which also provides QC for M-bias analysis and overlapping paired-end reads.

Variant Calling and Sample Clustering (SNP)

As single nucleotide polymorphisms (SNPs) in a cytosine-to-thymine context are obscured in bisulfite data(7), neither variant calling nor sample methylation clustering can be resolved using conventional methods. A simple post-processing procedure for *in silico* manipulation of both base qualities and base nucleotides in bisulfite contexts, following alignment, has been shown to facilitate conventional SNP calling on WGBS data(14). This heuristic method has been implemented herein and enables a) downstream analysis with tools that are already well-established for DNA-seq such as Freebayes(15), and b) sample clustering with kWIP(16) which uses k-mer diversity to estimate a distance matrix.

Differential Methylation (DMR)

Pairwise comparisons of methylation profiles between groups of samples can be made with metilene(17), to derive either regions (DMRs) or positions (DMPs) while correcting for multiple comparisons. Due to the non-parametric statistical test, each methylation context (CG, CHG, CHH) can be analysed independently without *a priori* assumptions about the distribution of methylation values. A recent benchmark demonstrated a higher sensitivity for finding DMRs with metilene in comparison to other tools(18).

Epigenome-wide Association Studies (EWAS)

For a given population of samples, the output derived from previous aspects of the toolkit can be combined and processed using the EWAS pipeline(19) for analysis using the GEM suite(20), in order to study the association between epigenetics, genetics, and environmental metadata through the identification of quantitative trait loci (QTL). These QTLs can be discovered either by taking the full set of methylated positions, in any methylation context, or by first subsetting according to DMPs/DMRs, or even by taking the DMPs/DMRs themselves in place of methylated positions for use as genomic markers. The confounding genetic component can be resolved in each case by providing the SNPs derived in the first place from the same bisulfite data, without the need for conventional whole genome sequencing data alongside.

RESULTS

The 23 independent WGBS libraries were first mapped in 'high-throughput mode' with the EpiDiverse WGBS pipeline, resulting in mapping rates ranging from 78.38% to 80.44% under default parameter settings (Supplementary Table S2). The global methylation level in all contexts is reported in Supplementary Figure S1, alongside a principal component analysis demonstrating the unsupervised grouping of all samples based on the variation in shared methylated sites.

Following alignment, variant calling was performed with the EpiDiverse SNP pipeline to identify SNPs from bisulfite-treated data based on sequence masking and base quality manipulation(14). The total number of variants in each sample are summarised in Supplementary Table S3. Alternatively, the pipeline can attempt to mask short variants and normalise the genetic diversity between samples. As studies on population epigenetics tend to centre around species with low genetic diversity (cf. hierarchical clustering tree on genetic information in Supplementary Figure S2a), a hierarchical clustering based on sequence k-mer diversity(16) after masking short variants can instead give an indication of grouping based on DNA methylation patterns (Supplementary Figure S2b). Such an analysis can facilitate the identification of discrete groups prior to calling differentially methylated positions / regions, without limiting the analysis to only those methylated positions that are shared across all samples by a minimum

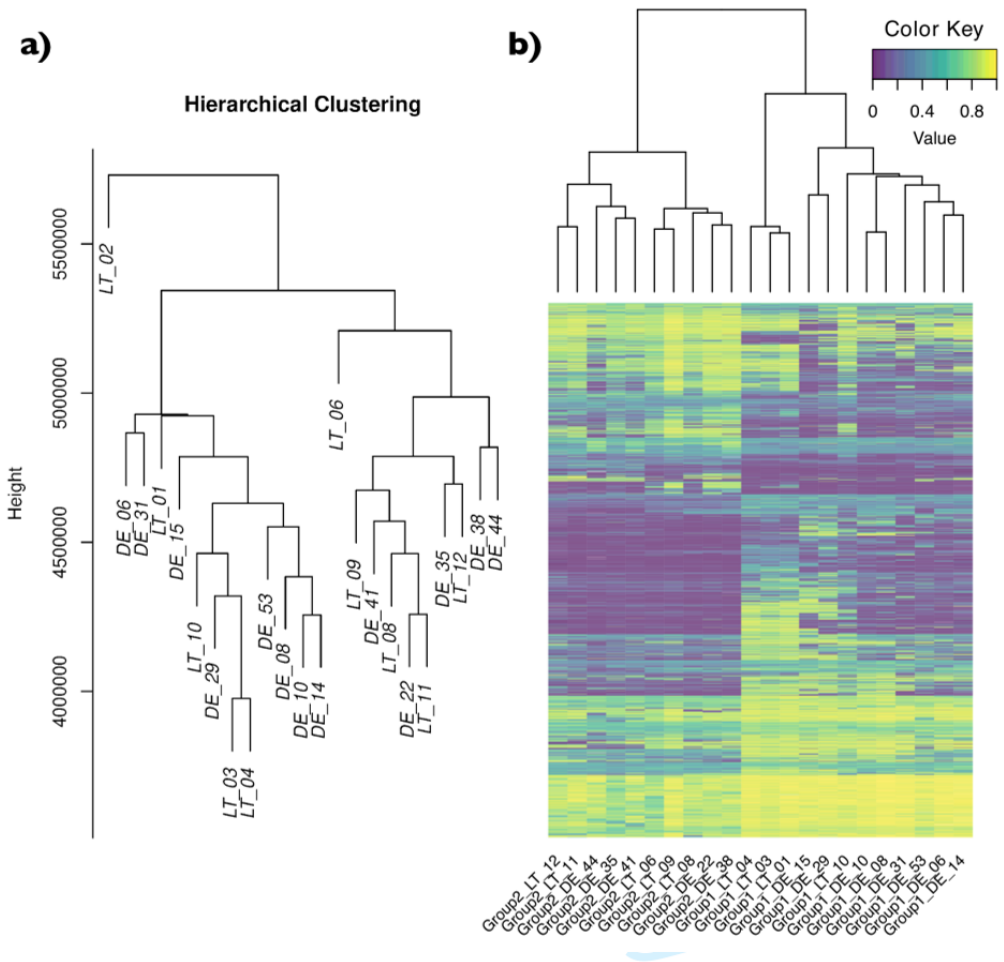


Figure 2. **(a)** Hierarchical clustering of methylated sites (all contexts) derived from the cohort of *P. nigra* samples from populations in Germany and Lithuania, and **(b)** the resulting heatmap of significant DMRs ($q < 0.05$) obtained after cutting the hierarchical tree at 5.25×10^6 to form two discrete groups (leaving LT_02 as outlier). Either plot can be obtained using the EpiDiverse toolkit.

threshold on sequencing depth. Otherwise, the distance matrix can instead be estimated from the methylation values in the conventional approach following per-sample methylation calling.

Appropriate groupings are dependant on the specific experimental design of each study. Once identified they can be subsequently evaluated for differential methylation with the EpiDiverse DMR pipeline, which analyses either all possible pairwise comparisons of groups or each group in relation to a designated control group. Here, methylated sites (all contexts) obtained from the cohort of German and Lithuanian populations of *P. nigra* were subject to hierarchical clustering and the resulting tree cut at approximately 5.25×10^6 to form two discrete groups and one outlier (Figure 2a). The total number of significant DMRs ($q < 0.05$) resulting from the pairwise comparison of these groups are given in Supplementary Table S4, and the corresponding heatmap showing the differential methylation level across the range of selected samples is shown in Figure 2b. Interestingly, the heatmap in some instances shows greater congruency with the clustering based on kWIP in Supplementary Figure S2b (e.g. LT_10, DE_41, DE_44, and a distinct clade with LT_01, LT_03, LT_04), indicating the potential utility as an alternative approach.

Finally, the cumulation of results from the WGBS and DMR pipelines were combined into a small analysis with EpiDiverse EWAS, based on the methylated sites

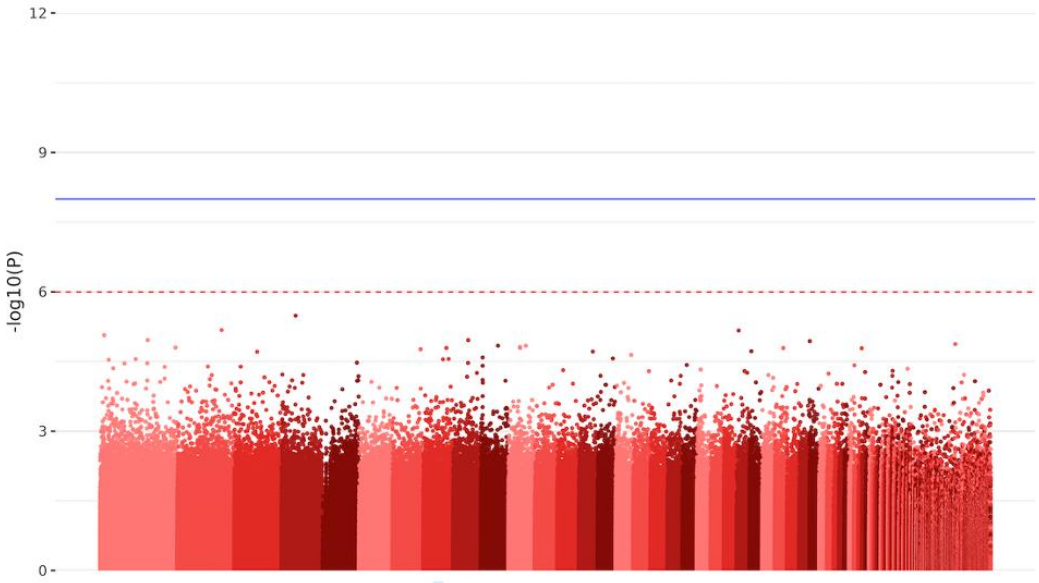


Figure 3. Manhattan plot demonstrating the total number of tested positions during EWAS, from the cohort of *P. nigra* samples obtained from populations in Germany and Lithuania. At this level, no positions were found to be significant ($p<1\times10^{-8}$) or even suggestive ($p<1\times10^{-6}$) based on common thresholds selected to account for the burden of multiple testing. The plot is obtained automatically from the EWAS pipeline output (E-model).

in CG context and subset according to the significant DMRs discovered in the same context, using leaf flavonol content measured in the parent generation as a phenotypic trait. In the case of *P. nigra* the resulting manhattan plot (E-model) in Figure 3 reveals initially no significant QTLs below the common significance threshold of $p<1\times10^{-8}$, or even below the suggestive significance threshold of $p<1\times10^{-6}$, based on the global analysis of all methylated sites. The same analysis when conducted however at the region-level revealed a total of 92 significant QTLs ($q<0.25$) which could be taken forward for further investigation (Supplementary Table S5). A brief inspection of these regions intersected with functional annotations in the *P. nigra* genome returned some features potentially relevant to flavonol content, including genes with homology to *ascorbate-specific transmembrane electron transporter 1*, *caspase family protein*, and *mechanosensitive ion channel protein 3* alongside also *methyltransferases PMT2/PMT24*. Furthermore, the incorporation of SNP data into the G-model aspect of the EWAS pipeline can help to resolve any underlying genetic component which may be driving such associations with epigenetic markers.

A typical drawback of any (epi)genome-wide association study is the high burden of multiple testing, necessitating the use of a controlling procedure which can often be excessively conservative due to the high number of negative tests, thus obscuring many genuine biological findings which may be present within the dataset. The common significance threshold of $p<1\times10^{-8}$ is based on a Bonferroni adjustment limited to a maximum of 1 million tests, regardless of the true number of tests. It is often argued with genetic data that a lack of true sample independence owing to linkage disequilibrium between SNPs can facilitate the use of this more heuristic variant of the Bonferroni adjustment, but statistically speaking this may be less than ideal. A more robust solution would be to reduce the total number of tests in the first place based on *a priori* knowledge. The EWAS pipeline therefore provides a mechanism to subset data based on any such regions provided by the end-user, for example here with DMRs obtained from the DMR pipeline, with the aim to reduce the majority of negative tests while still capturing the majority of positive tests. Though true positives may still be missed, depending largely on the selection criteria of such regions, often more can be gained relative to the global analysis of all methylated positions.

CONCLUSION

The EpiDiverse Toolkit provides a suite of software pipelines for the analysis of ecological plant epigenetics, which adheres to the principles of "FAIR" (i.e. Findable, Accessible, Interoperable and Re-usable). The toolkit combines common procedures, such as mapping and methylation calling, with novel implementations for short variant calling and combining all results within a robust variation of EWAS, with each aspect benchmarked specifically for non-model plant species. This provides

a consistent, repeatable framework which not only streamlines computational analyses within-species, but also facilitates more general comparisons between different organisms which may have evolved very different mechanisms involving DNA methylation.

DATA AVAILABILITY

All pipelines are open-source and publicly available through the https://github.com/EpiDiverse domain. The data used for analysis was generated by the European Training Network "EpiDiverse" to be published in the European Nucleotide Archive, and is otherwise available upon reasonable request to the authors.

FUNDING

The European Training Network "EpiDiverse" received funding from the EU Horizon 2020 program under Marie Skłodowska-Curie grant agreement No 764965.

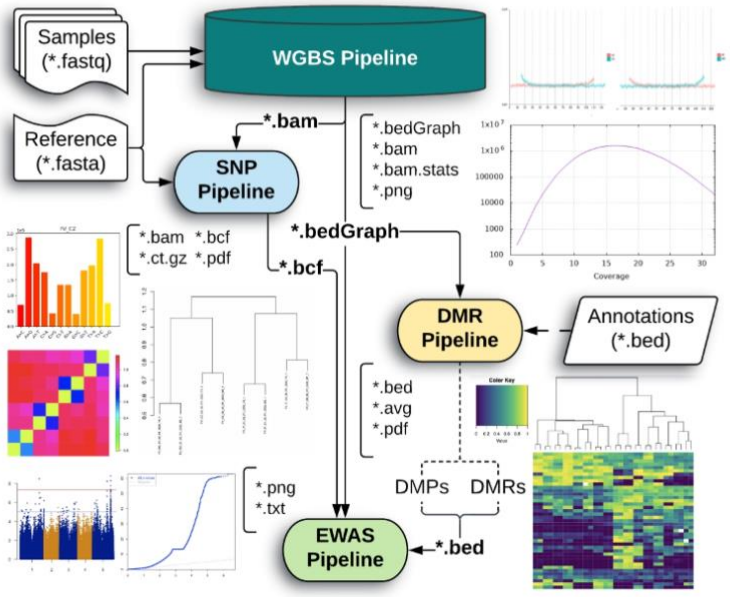
ACKNOWLEDGEMENTS

We would like to thank all the members of the EpiDiverse Consortium for their active and invaluable support in discussing, developing and testing these tools.

Conflict of interest statement. None declared.

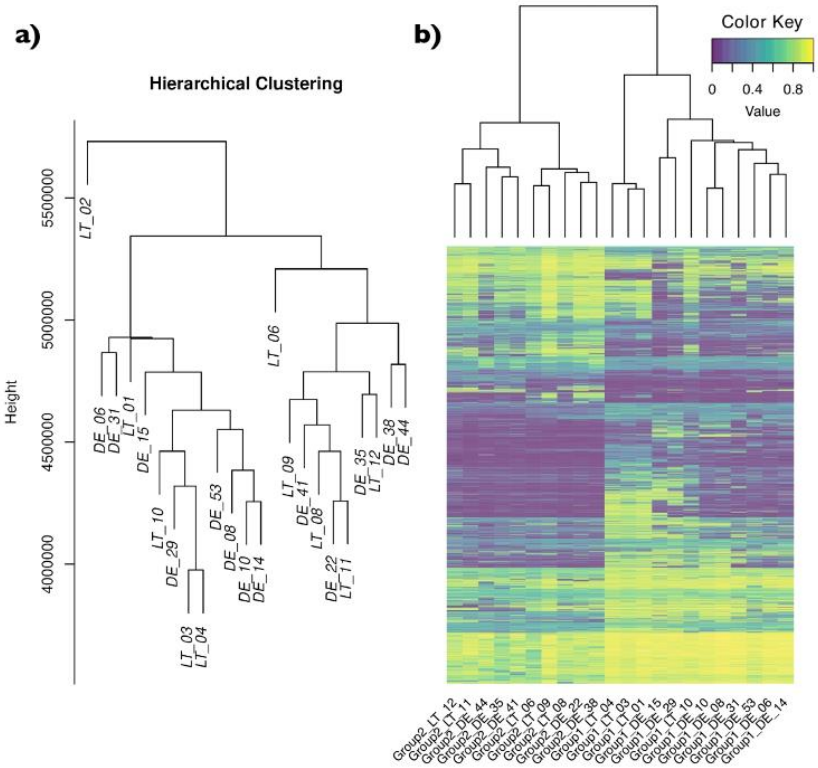
REFERENCES

1. Bossdorf, O., Arcuri, D., Richards, C. L., & Pigliucci, M. (2010) Experimental alteration of DNA methylation affects the phenotypic plasticity of ecologically relevant traits in Arabidopsis thaliana. *Evolutionary Ecology*, **24**(3), 541-553.
2. Boyko, A., & Kovalchuk, I. (2010) Transgenerational response to stress in Arabidopsis thaliana. *Plant signaling & behavior*, **5**(8), 995-998.
3. 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.
4. Richards, C. L., Alonso, C., Becker, C., Bossdorf, O., Bucher, E., Colomé-Tatché, M., ... & Grosse, I. (2017) Ecological plant epigenetics: Evidence from model and non-model species, and the way forward. *Ecology letters*, **20**(12), 1576-1590.
5. 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*, **89**(5), 1827-1831.
6. Tran, H., Porter, J., Sun, M. A., Xie, H., & Zhang, L. (2014). Objective and comprehensive evaluation of bisulfite short read mapping tools. *Advances in bioinformatics*, **2014**.
7. Liu, Y., Siegmund, K. D., Laird, P. W., & Berman, B. P. (2012) Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome biology*, **13**(7), R61.
8. Schatz, M. C., Witkowski, J., & McCombie, W. R. (2012) Current challenges in de novo plant genome sequencing and assembly. *Genome biology*, **13**(4), 243.
9. Feng, S., Cokus, S. J., Zhang, X., Chen, P. Y., Bostick, M., Goll, M. G., ... & Ukomadu, C. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, **107**(19), 8689-8694.
10. di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017) Nextflow enables reproducible computational workflows. *Nature Biotechnology*, **35**(4), 316-319.
11. Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., ... & Nahnsen, A. (2020) The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, **38**(3), 276-278.
12. Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., ... & Köster, J. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, **15**(7), 475-476.
13. Nunn, A., Otto, C., Stadler, P. F., & Langenberger, D. (2021) Erratum to: Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA methylation analysis. *Briefings in Bioinformatics*, **bbab183**.
14. Nunn, A., Otto, C., Stadler, P. F., & Langenberger, D. (2021) Manipulating base quality scores enables variant calling from bisulfite sequencing alignments using conventional Bayesian approaches. *bioRxiv preprint bioRxiv:2021.01.11.425926*
15. Garrison, E., & Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.
16. Murray, K. D., Webers, C., Ong, C. S., Borevitz, J., & Warthmann, N. (2017) kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLoS computational biology*, **13**(9), e1005727.
17. Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., & Hoffmann, S. (2016) metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome research*, **26**(2), 256-262.
18. Kreutz, C., Can, N. S., Bruening, R. S., Meyberg, R., Mérai, Z., Fernandez-Pozo, N., & Rensing, S. A. (2020) A blind and independent benchmark study for detecting differentially methylated regions in plants. *Manuscript submitted for publication.*
19. Can, S. N., Nunn, A., Galanti, D., Langenberger, D., Becker, C., Volmer, K., Heer, K., Opgenoorth, L., Fernandez-Pozo, N. & Rensing, S.A. (2021) The EpiDiverse Plant Epigenome-Wide Association Studies (EWAS) Pipeline. *Epigenomes*, **5**(2), p.12.
20. Pan, H., Holbrook, J. D., Karnani, N., & Kwoh, C. K. (2016) Gene, Environment and Methylation (GEM): a tool suite to efficiently navigate large scale epigenome wide association studies and integrate genotype and interaction between genotype and environment. *BMC bioinformatics*, **17**(1), 299.



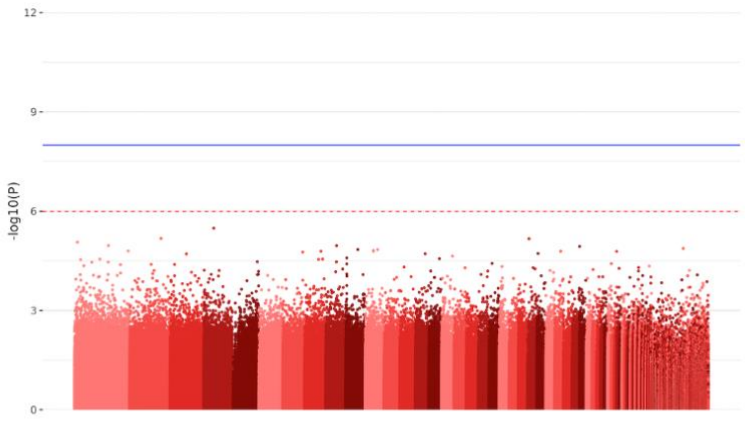
Overview of the EpiDiverse Toolkit. The WGBS data forms the foundation of the analysis, and each downstream pipeline is built to work either in cooperation with one another or, optionally, with independently-generated input data. All pipelines output runtime metadata, tracing and further visualisation in addition to what is shown here. The full output is described for each pipeline in the documentation on Github.

760x626mm (600 x 600 DPI)



(a) Hierarchical clustering of methylated sites (all contexts) derived from the cohort of *P. nigra* samples from populations in Germany and Lithuania, and **(b)** the resulting heatmap of significant DMRs (q<0.05) obtained after cutting the hierarchical tree at 5.25×10^6 to form two discrete groups (leaving LT_02 as outlier). Either plot can be obtained using the EpiDiverse toolkit.

776x723mm (600 x 600 DPI)



Manhattan plot demonstrating the total number of tested positions during EWAS, from the cohort of *P. nigra* samples obtained from populations in Germany and Lithuania. At this level, no positions were found to be significant ($p<1x10^{-8}$) or even suggestive ($p<1x10^{-6}$) based on common thresholds selected to account for the burden of multiple testing. The plot is obtained automatically from the EWAS pipeline output (E-model).

388x214mm (72 x 72 DPI)

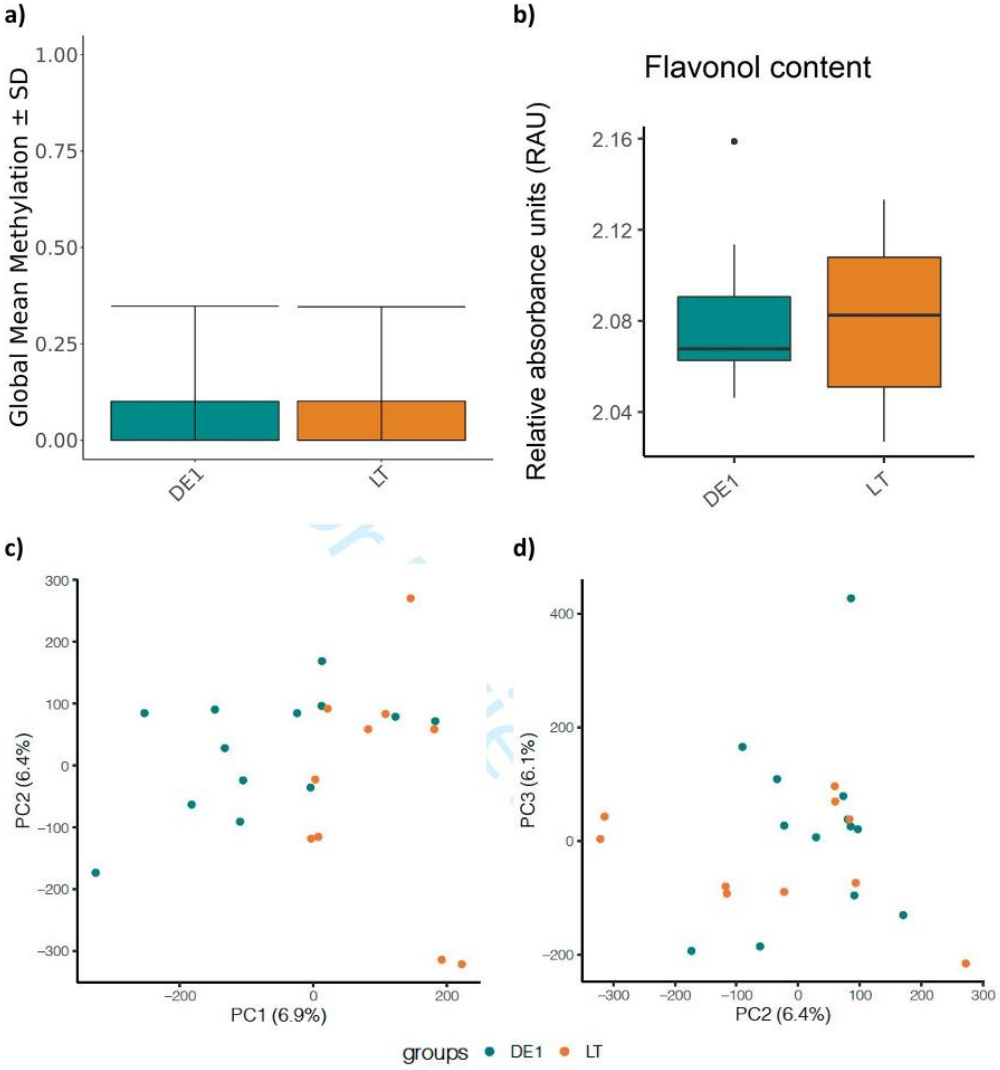


Figure S1. Descriptive statistics of the German (DE1) and Lithuanian (LT) populations of *Populus nigra* samples demonstrating **a)** global methylation levels in all contexts following cultivation under common garden conditions, **b)** leaf flavonol content as measured from samples collected in the field, and Principal Component Analysis (PCA) based on the per-site methylation levels in all contexts between **c)** component 1 and 2, and **d)** component 2 and 3.

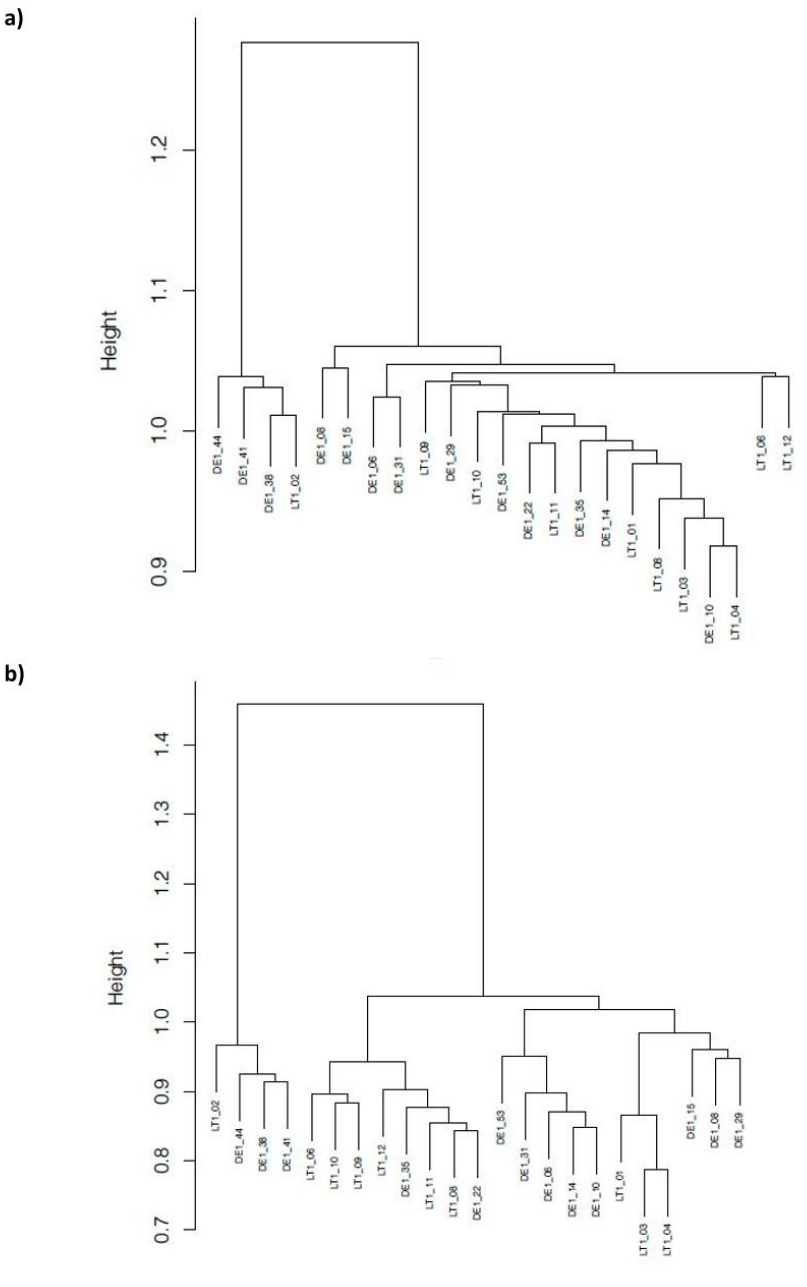


Figure S2. Hierarchical clustering of mappable FASTQ reads by k-mer diversity, using kWIP, following either **a)** bisulfite masking, or **b)** masking short variants to normalise genetic diversity.

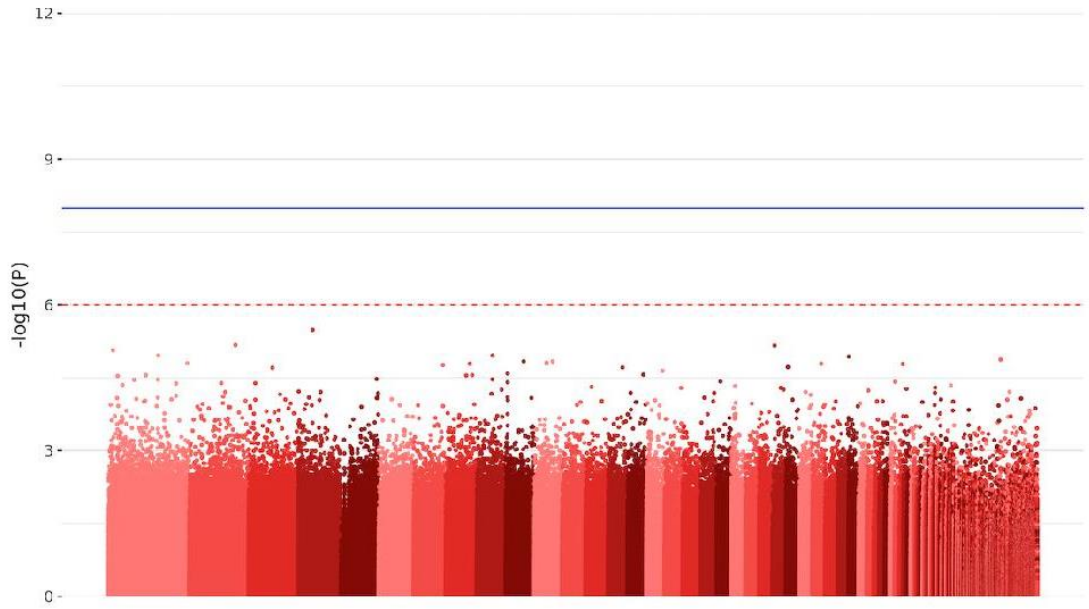


Figure 3. Manhattan plot demonstrating the total number of tested positions during EWAS, from the cohort of *P. nigra* samples obtained from populations in Germany and Lithuania. At this level, no positions were found to be significant ($p<1\times10^{-8}$) or even suggestive ($p<1\times10^{-6}$) based on common thresholds selected to account for the burden of multiple testing. The plot is obtained automatically from the EWAS pipeline output (E-model).

with functional annotations in the *P. nigra* genome returned some features potentially relevant to flavonol content, including genes with homology to *ascorbate-specific transmembrane electron transporter 1*, *caspase family protein*, and *mechanosensitive ion channel protein 3* alongside also *methyltransferases PMT2/PMT24*. Furthermore, the incorporation of SNP data into the G-model aspect of the EWAS pipeline can help to resolve any underlying genetic component which may be driving such associations with epigenetic markers.

A typical drawback of any (epi)genome-wide association study is the high burden of multiple testing, necessitating the use of a controlling procedure which can often be excessively conservative due to the high number of negative tests, thus obscuring many genuine biological findings which may be present within the dataset. The common significance threshold of $p<1\times10^{-8}$ is based on a Bonferroni adjustment limited to a maximum of 1 million tests, regardless of the true number of tests. It is often argued with genetic data that a lack of true sample independence owing to linkage disequilibrium between SNPs can facilitate the use of this more heuristic variant of the Bonferroni adjustment, but statistically speaking this may be less than ideal. A more robust solution would be to reduce the total number of tests in the first place based on *a priori* knowledge. The EWAS pipeline therefore provides a mechanism to subset data based on any such regions provided by the end-user, for example here with DMRs obtained from the DMR pipeline, with the aim to reduce the majority of negative tests while still capturing the majority of positive tests. Though true positives may still be missed, depending largely on the selection criteria of such regions, often more can be gained relative to the global analysis of all methylated positions.

CONCLUSION

The EpiDiverse Toolkit provides a suite of software pipelines for the analysis of ecological plant epigenetics, which adheres to the principles of "FAIR" (i.e. Findable, Accessible, Interoperable and Re-usable). The toolkit combines common procedures, such as mapping and methylation calling, with novel implementations for short variant calling and combining all results within a robust variation of EWAS, with each aspect benchmarked specifically for non-model plant species. This provides a consistent, repeatable framework which not only streamlines computational analyses within-species, but also facilitates more general comparisons between different organisms which may have evolved very different mechanisms involving DNA methylation. The pipelines are open-source and publicly available through the https://github.com/EpiDiverse domain.

FUNDING

The European Training Network "EpiDiverse" received funding from the EU Horizon 2020 program under Marie Skłodowska-Curie grant agreement No 764965.

ACKNOWLEDGEMENTS

We would like to thank all the members of the EpiDiverse Consortium for their active and invaluable support in discussing, developing and testing these tools.

Conflict of interest statement. None declared.

REFERENCES

1. Bossdorf, O., Arcuri, D., Richards, C. L., & Pigliucci, M. (2010) Experimental alteration of DNA methylation affects the phenotypic plasticity of ecologically relevant traits in Arabidopsis thaliana. *Evolutionary Ecology*, **24**(3), 541-553.
2. Boyko, A., & Kovalchuk, I. (2010) Transgenerational response to stress in Arabidopsis thaliana. *Plant signaling & behavior*, **5**(8), 995-998.
3. 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.
4. Richards, C. L., Alonso, C., Becker, C., Bossdorf, O., Bucher, E., Colomé-Tatché, M., ... & Grosse, I. (2017) Ecological plant epigenetics: Evidence from model and non-model species, and the way forward. *Ecology letters*, **20**(12), 1576-1590.
5. 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*, **89**(5), 1827-1831.
6. Tran, H., Porter, J., Sun, M. A., Xie, H., & Zhang, L. (2014). Objective and comprehensive evaluation of bisulfite short read mapping tools. *Advances in bioinformatics*, **2014**.
7. Liu, Y., Siegmund, K. D., Laird, P. W., & Berman, B. P. (2012) Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome biology*, **13**(7), R61.
8. Schatz, M. C., Witkowski, J., & McCombie, W. R. (2012) Current challenges in de novo plant genome sequencing and assembly. *Genome biology*, **13**(4), 243.
9. Feng, S., Cokus, S. J., Zhang, X., Chen, P. Y., Bostick, M., Goll, M. G., ... & Ukomadu, C. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, **107**(19), 8689-8694.
10. di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017) Nextflow enables reproducible computational workflows. *Nature Biotechnology*, **35**(4), 316-319.
11. Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., ... & Nahnsen, A. (2020) The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, **38**(3), 276-278.
12. Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., ... & Köster, J. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, **15**(7), 475-476.
13. Nunn, A., Otto, C., Stadler, P. F., & Langenberger, D. (2020) Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA methylation analysis. *bioRxiv preprint bioRxiv:2020.08.28.271585*
14. Nunn, A., Otto, C., Stadler, P. F., & Langenberger, D. (2021) Manipulating base quality scores enables variant calling from bisulfite sequencing alignments using conventional Bayesian approaches. *bioRxiv preprint bioRxiv:2021.01.11.425926*
15. Garrison, E., & Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.
16. Murray, K. D., Webers, C., Ong, C. S., Borevitz, J., & Warthmann, N. (2017) kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLoS computational biology*, **13**(9), e1005727.
17. Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., & Hoffmann, S. (2016) metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome research*, **26**(2), 256-262.
18. Kreutz, C., Can, N. S., Bruening, R. S., Meyberg, R., Mérai, Z., Fernandez-Pozo, N., & Rensing, S. A. (2020) A blind and independent benchmark study for detecting differentially methylated regions in plants. *Manuscript submitted for publication.*
19. Pan, H., Holbrook, J. D., Karnani, N., & Kwoh, C. K. (2016) Gene, Environment and Methylation (GEM): a tool suite to efficiently navigate large scale epigenome wide association studies and integrate genotype and interaction between genotype and environment. *BMC bioinformatics*, **17**(1), 299.

3.3.2 Further applicability of this work

The EpiDiverse toolkit is available under https://github.com/EpiDiverse (accessed on 1 May 2021) including test datasets for making it possible for anyone to access and analyze their data. It covers WGBS, DMR, SNP, and EWAS pipelines to perform comprehensive analyses with all species. Pipelines available with the toolkit have been used by all projects under the EpiDiverse consortium.

Chapter 4

Concluding remarks

4 Concluding remarks

In summary, the DMR benchmark shares unbiased analysis with a minimal bias. This study is independent since none of the authors participated in the development of any DMR methods tested. The data decision rule to select between computational methods is also applicable to any other dataset. Our study showed advantages of some statistical approaches, none of the tools outperformed others for all four datasets but metilene showed outstanding performance in terms of recall and precision.

In other respects, the work with EWAS presented in this thesis not only shared a versatile, extensive, and user-friendly pipeline but also shared open-source scripts implemented within it. During the last two years, the EpiDiverse EWAS pipeline was tested in some studies and presented reliable and logical results ([68] and unpublished study). In terms of *Q. lobata* analysis in Can et al., 2021 study [68], we have seen that nearly all significant Cs were reproduced although statistical methods were different between the EpiDiverse pipeline and Gugger et al., 2016 study [8]. Missing data estimation with beta distribution yielded robust and accurate approximation as highlighted from the significant amount of overlap. Some unique Cs found only by the EWAS pipeline were seen to be associated with meaningful literature with climatic and spatial variables. Considering *P. abies* dataset (derived from [9] and unpublished data), we deduced that choice of model and input depends on the user's research objective. To sum up, we shared the EpiDiverse EWAS pipeline to perform EWAS analysis either using the output from other EpiDiverse pipelines or custom data in accurate format.

Finally, we built a set of tools for plant ecologists with whole-genome bisulfite data to accomplish an extensive WGBS [69], DMR [70], SNP [71], and EWAS [68] analyses and succeeded with publicly available toolkit under https://github.com/EpiDiverse (accessed on 1 May 2021). This set of pipelines enables to conduct of a broad range of analyses with WGBS data and its sufficient to have proper input formats to run all pipelines separately or together.

Chapter 5

Outlook

5 Outlook

Regarding the DMR benchmark study, researchers can generate their simulated datasets with provided scripts using additional statistical approaches to extend this benchmark study. It could be really interesting to see the performance of newly added DMR tools with other species that were not used in this study.

Data sets, scripts, and conclusions in this thesis were used and will be used for further research. Ascending the need for a versatile EWAS tool for all species makes the EpiDiverse pipeline valuable and efficient. Linear mixed models (LMM) could be considered to implement in a future release for prospective users who want to encounter relatedness in other words kinship matrix. Kinship matrix can be estimated either from genetic or epigenetic data between individuals to control the false-positive ratio. However, LMM is not suitable to distinguish genetically driven epigenetic variation from environmentally induced epigenetic variation. Also, gene ontology analysis (GOA) for a specific species could be possible to perform in future versions of the EpiDiverse EWAS pipeline. I hope to see its contribution to the studies of researchers who work with EWAS. Furthermore, hopefully, the pipeline and its results will be useful for EpiDiverse projects namely RP06, RP07, and RP08 in collaboration with mine.

Concerning the EpiDiverse toolkit, researchers can do mapping, calling DMPs/DMRs with variants, and perform EWAS with their datasets. We give the opportunity of using publicly available scripts provided by each pipeline under the EpiDiverse toolkit to perform specific runs besides provided options. Next releases may contain pipelines for revealing the association between expression and DNA methylation data. It could also be possible to adapt existing tools to compatible with EpiGBS [72] and RRBS [39] data formats in the future.

6 References

1. Juhling, F., et al., *metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data.* Genome Res, 2016. **26**(2): p. 256-62.
2. Akalin, A., et al., *methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles.* Genome Biol, 2012. **13**(10): p. R87.
3. Sun, D., et al., *MOABS: model based analysis of bisulfite sequencing data.* Genome Biol, 2014. **15**(2): p. R38.
4. Peters, T.J., et al., *De novo identification of differentially methylated regions in the human genome.* Epigenetics Chromatin, 2015. **8**: p. 6.
5. Condon, D.E., et al., *Defiant: (DMRs: easy, fast, identification and ANnoTation) identifies differentially Methylated regions from iron-deficient rat hippocampus.* BMC Bioinformatics, 2018. **19**(1): p. 31.
6. Hansen, K.D., B. Langmead, and R.A. Irizarry, *BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions.* Genome Biol, 2012. **13**(10): p. R83.
7. Park, Y., et al., *MethylSig: a whole genome DNA methylation analysis pipeline.* Bioinformatics, 2014. **30**(17): p. 2414-22.
8. Gugger, P.F., et al., *Species-wide patterns of DNA methylation variation in Quercus lobata and their association with climate gradients.* Mol Ecol, 2016. **25**(8): p. 1665-80.
9. Heer, K., et al., *Detection of somatic epigenetic variation in Norway spruce via targeted bisulfite sequencing.* Ecol Evol, 2018. **8**(19): p. 9672-9682.
10. Dismukes, G.C., et al., *The origin of atmospheric oxygen on Earth: the innovation of oxygenic photosynthesis.* Proc Natl Acad Sci U S A, 2001. **98**(5): p. 2170-5.
11. Veeresham, C., *Natural products derived from plants as a source of drugs.* J Adv Pharm Technol Res, 2012. **3**(4): p. 200-1.
12. Schacherer, J., *Beyond the simplicity of Mendelian inheritance.* C R Biol, 2016. **339**(7-8): p. 284-8.
13. Stamhuis, I.H., O.G. Meijer, and E.J. Zevenhuizen, *Hugo de Vries on heredity, 1889-1903. Statistics, Mendelian laws, pangenes, mutations.* Isis, 1999. **90**(2): p. 238-67.
14. McClintock, B., *The control of gene action in maize.* Genetic Control of Differentiation, Brookhaven Symposia in Biology, Royal Society, London, 1965. **18**: p. 162–184.
15. H., W.C., *Genetic Assimilation of the Bithorax Phenotype.* Society for the Study of Evolution, 1956. **10**(1): p. 1-13.
16. Nanney, D.L., *EPIGENETIC CONTROL SYSTEMS.* Proc Natl Acad Sci U S A, 1958. **44**(7): p. 712-7.
17. Weigel, D. and R. Mott, *The 1001 genomes project for Arabidopsis thaliana.* Genome Biol, 2009. **10**(5): p. 107.
18. Rensing, S.A., et al., *The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants.* Science, 2008. **319**(5859): p. 64-9.
19. Miller, J.L. and P.A. Grant, *The role of DNA methylation and histone modifications in transcriptional regulation in humans.* Subcell Biochem, 2013. **61**: p. 289-317.
20. Fuchs, J., et al., *Chromosomal histone modification patterns--from conservation to diversity.* Trends Plant Sci, 2006. **11**(4): p. 199-208.
21. Weinhold, B., *Epigenetics: the science of change.* Environ Health Perspect, 2006. **114**(3): p. A160-7.
22. Sudan, J., M. Raina, and R. Singh, *Plant epigenetic mechanisms: role in abiotic stress and their generational heritability.* 3 Biotech, 2018. **8**(3): p. 172.
23. Quadrana, L. and V. Colot, *Plant Transgenerational Epigenetics.* Annu Rev Genet, 2016. **50**: p. 467-491.
24. Cubas, P., C. Vincent, and E. Coen, *An epigenetic mutation responsible for natural variation in floral symmetry.* Nature, 1999. **401**(6749): p. 157-61.
25. Manning, K., et al., *A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening.* Nat Genet, 2006. **38**(8): p. 948-52.
26. Weigel, D. and V. Colot, *Epialleles in plant evolution.* Genome Biol, 2012. **13**(10): p. 249.

27. Mirouze, M. and J. Paszkowski, *Epigenetic contribution to stress adaptation in plants.* Curr Opin Plant Biol, 2011. **14**(3): p. 267-74.
28. McCue, A.D., et al., *Gene expression and stress response mediated by the epigenetic regulation of a transposable element small RNA.* PLoS Genet, 2012. **8**(2): p. e1002474.
29. Chinnusamy, V. and J.K. Zhu, *Epigenetic regulation of stress responses in plants.* Curr Opin Plant Biol, 2009. **12**(2): p. 133-9.
30. Jansz, N., *DNA methylation dynamics at transposable elements in mammals.* Essays Biochem, 2019. **63**(6): p. 677-689.
31. Li, Z.W., et al., *Transposable Elements Contribute to the Adaptation of Arabidopsis thaliana.* Genome Biol Evol, 2018. **10**(8): p. 2140-2150.
32. Kim, D.H. and S. Sung, *Accelerated vernalization response by an altered PHD-finger protein in Arabidopsis.* Plant Signal Behav, 2017. **12**(5): p. e1308619.
33. Dyachenko, O.V., et al., *Human non-CG methylation: are human stem cells plant-like?* Epigenetics, 2010. **5**(7): p. 569-72.
34. Colot, V. and J.L. Rossignol, *Eukaryotic DNA methylation as an evolutionary device.* Bioessays, 1999. **21**(5): p. 402-11.
35. Feng, S., S.E. Jacobsen, and W. Reik, *Epigenetic reprogramming in plant and animal development.* Science, 2010. **330**(6004): p. 622-7.
36. Gent, J.I., et al., *Strong epigenetic similarity between maize centromeric and pericentromeric regions at the level of small RNAs, DNA methylation and H3 chromatin modifications.* Nucleic Acids Res, 2012. **40**(4): p. 1550-60.
37. Lister, R., et al., *Highly integrated single-base resolution maps of the epigenome in Arabidopsis.* Cell, 2008. **133**(3): p. 523-36.
38. Law, J.A. and S.E. Jacobsen, *Establishing, maintaining and modifying DNA methylation patterns in plants and animals.* Nat Rev Genet, 2010. **11**(3): p. 204-20.
39. Gu, H., et al., *Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling.* Nat Protoc, 2011. **6**(4): p. 468-81.
40. Kishore, K. and M. Pelizzola, *Identification of Differentially Methylated Regions in the Genome of Arabidopsis thaliana.* Methods Mol Biol, 2018. **1675**: p. 61-69.
41. Tsai, P.C. and J.T. Bell, *Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation.* Int J Epidemiol, 2015. **44**(4): p. 1429-1441.
42. Rakyan, V.K., et al., *Epigenome-wide association studies for common human diseases.* Nat Rev Genet, 2011. **12**(8): p. 529-41.
43. Tam, V., et al., *Benefits and limitations of genome-wide association studies.* Nat Rev Genet, 2019. **20**(8): p. 467-484.
44. Bush, W.S. and J.H. Moore, *Chapter 11: Genome-wide association studies.* PLoS Comput Biol, 2012. **8**(12): p. e1002822.
45. Tibbs Cortes, L., Z. Zhang, and J. Yu, *Status and prospects of genome-wide association studies in plants.* Plant Genome, 2021. **14**(1): p. e20077.
46. Ersoz E.S., Y.J., Buckler E.S., *Applications of Linkage Disequilibrium and Association Mapping in Crop Plants. In: Varshney R.K., Tuberosa R. (eds) Genomics-Assisted Crop Improvement*, in *Genomics-Assisted Crop Improvement*, D. Springer, Editor. 2007, Springer, Dordrecht. p. pp 97-119.
47. Sukumaran S., Y.J., *Association Mapping of Genetic Resources: Achievements and Future Perspectives*, in *Genomics of Plant Genetic Resources*. 2013, Springer, Dordrecht. p. 207-235.
48. Varshney, R.K., et al., *Can genomics boost productivity of orphan crops?* Nat Biotechnol, 2012. **30**(12): p. 1172-6.
49. Wang, S.B., et al., *Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology.* Sci Rep, 2016. **6**: p. 19444.
50. Zhang, Y., et al., *Applications and potential of genome editing in crop improvement.* Genome Biol, 2018. **19**(1): p. 210.
51. Gupta, P.K., P.L. Kulwal, and V. Jaiswal, *Association mapping in plants in the post-GWAS genomics era.* Adv Genet, 2019. **104**: p. 75-154.
52. Liu, H.J.a.Y., J., *Crop genome-wide association study: a harvest of biological relevance.* Plant J, 2019. **97**: p. 8-18.

53. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
54. Heard, E. and R.A. Martienssen, *Transgenerational epigenetic inheritance: myths and mechanisms.* Cell, 2014. **157**(1): p. 95-109.
55. Boyko, A., et al., *Transgenerational adaptation of Arabidopsis to stress requires DNA methylation and the function of Dicer-like proteins.* PLoS One, 2010. **5**(3): p. e9514.
56. Lang-Mladek, C., et al., *Transgenerational inheritance and resetting of stress-induced loss of epigenetic gene silencing in Arabidopsis.* Mol Plant, 2010. **3**(3): p. 594-602.
57. Latzel, V., A.P. Rendina González, and J. Rosenthal, *Epigenetic Memory as a Basis for Intelligent Behavior in Clonal Plants.* Front Plant Sci, 2016. **7**: p. 1354.
58. Paul, D.S. and S. Beck, *Advances in epigenome-wide association studies for common diseases.* Trends Mol Med, 2014. **20**(10): p. 541-3.
59. Hu, Y., et al., *Prediction of Plant Height in Arabidopsis thaliana Using DNA Methylation Data.* Genetics, 2015. **201**(2): p. 779-93.
60. Verma, M., *Genome-wide association studies and epigenome-wide association studies go together in cancer control.* Future Oncol, 2016. **12**(13): p. 1645-64.
61. Houseman, E.A., et al., *Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions.* BMC Bioinformatics, 2008. **9**: p. 365.
62. Ong-Abdullah, M., et al., *Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm.* Nature, 2015. **525**(7570): p. 533-7.
63. Sáez-Laguna, E., et al., *Epigenetic variability in the genetically uniform forest tree species Pinus pinea L.* PLoS One, 2014. **9**(8): p. e103145.
64. Xu, J., et al., *EWAS: epigenome-wide association study software 2.0.* Bioinformatics, 2018. **34**(15): p. 2657-2658.
65. Rahmani, E., et al., *GLINT: a user-friendly toolset for the analysis of high-throughput DNA-methylation array data.* Bioinformatics, 2017. **33**(12): p. 1870-1872.
66. Pan, H., et al., *Gene, Environment and Methylation (GEM): a tool suite to efficiently navigate large scale epigenome wide association studies and integrate genotype and interaction between genotype and environment.* BMC Bioinformatics, 2016. **17**: p. 299.
67. Di Tommaso, P., et al., *Nextflow enables reproducible computational workflows.* Nat Biotechnol, 2017. **35**(4): p. 316-319.
68. Can, S.N., et al., *The EpiDiverse Plant Epigenome-Wide Association Studies (EWAS) Pipeline.* Epigenomes, 2021. **5**: p. 12.
69. Nunn, A., et al., *Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA methylation analysis.* bioRxiv, 2020. **10.1101**(2020.08.28.271585).
70. *Kreutz, C.; Can, N.S.; Bruening, R.S.; Meyberg, R.; Mérai, Z.; Fernandez-Pozo, N.; Rensing, S.A. A blind and independent benchmark study for detecting differentially methylated regions in plants. Bioinformatics 2020,36, 3314–3321.*
71. Nunn, A., et al., *Manipulating base quality scores enables variant calling from bisulfite sequencing alignments using conventional Bayesian approaches.* BioRxiv, 2020. **10.1101**(2021.01.11.425926): p. 17.
72. van Gurp, T.P., et al., *epiGBS: reference-free reduced representation bisulfite sequencing.* Nat Methods, 2016. **13**(4): p. 322-4.
73. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

7 Supporting information

7.1 A blind and independent benchmark study for detecting differentially methylated regions in plants

Supporting material can be found at: https://academic.oup.com/bioinformatics/article-abstract/36/11/3314/5809142?redirectedFrom=fulltext#supplementary-data (accessed on 1 May 2021).

7.2 The EpiDiverse Plant Epigenome-Wide Association Studies (EWAS) Pipeline

Supporting material can be found at**:**

https://www.mdpi.com/article/10 .3390/epigenomes5020012/s1 (accessed on 1 May 2021).

7.2.1 Additional analyses that were not shown in the main text

This part of analysis was not included in Kreutz et al., 2020 study [70] but related with it. A list was generated with strict parameters to measure DMR callers' accuracy and named "Manual DMRs". We assumed that this list is a sum of true positives (TPs) considering significant methylated cytosines. Script for manual DMRs aims to collect all TPs with specific parameters which take two distinct datasets. Additional parameters apart from datasets are mean methylation difference between samples per position, read coverage, and distance between DMRs for the same methylation context. Parameters can be explained like:

- Input1 (Dataset1 i.e.: ecotype1, context: CHH)
- Input1 (Dataset2 i.e.: ecotype2, context: CHH)
- Minimum coverage → This parameter is to remove positions that have lower coverage than this value.
- Minimum mean methylation difference → This parameter is to remove positions that have lower methylation difference than this value.
- Distance → This parameter is the maximum distance between two positions to be named as a DMR (base pairs (bps)).

Coverage scores for *Ae. arabicum* are (1.25, 1.29), (0.78, 1.73), and (1.45, 1.53) for CHH, CHG, and CpG contexts for C25-T25 (Cyprus 25 °C and Turkey 25 °C) ecotypes. They are like (0.68, 0.95), (0.90, 1.28) and (0.78, 1.13) for CHH, CHG and CpG contexts for C20-T20 ecotypes (Cyprus 20 °C and Turkey 20 °C). Therefore, minimum coverage was set to three for all species due to critical low values of *Ae. arabicum*. The window size was set to 200 bps, and the mean methylation difference between samples to 60% to collect enough DMRs for all species.

The script starts to detect a distance between a methylated position and a consecutive position to it. If this distance does not exceed 200bps, then it takes its consecutive position as a start point and checks the distance between its closest next position. This process goes until exceeding the 200bps limit and this region is called a "block". Blocks that have both hyper and hypomethylation were filtered and printed "WARNING! Opposite DMRs" onto the output. As the last step, blocks that have more than two elements were called a "region". After all, these blocks were named as "true DMRs" and merged into a single file to get the final "Manual DMR" list (Figure S1).

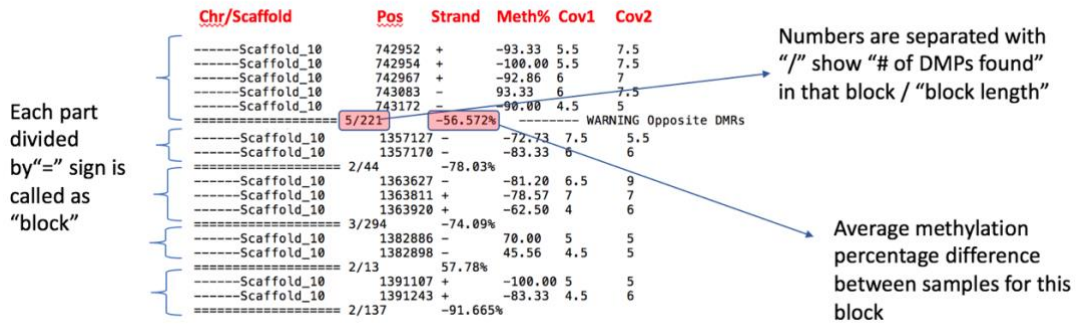


Figure S1: **Example output of a manual DMR script for *Ae.arabicum* dataset.** First column shows the chr/scaffold (chromosome/scaffold) information. The second column shows the positions of DMPs. The third column keeps the strand information. The methylation difference between the two datasets is kept in the fourth column. Coverage information is in the fifth and sixth columns per dataset.

A confusion matrix had to be set to statistically test the significance of the difference between the tool's DMRs and the manual DMRs.

$$\text{Sensitivity} = \frac{TP}{(TP+FN)}$$

$$\text{Specificity} = \frac{TN}{(TN+FP)}$$

$$\text{Recall} = \frac{TP}{(TP+FP)}$$

$$\text{Precision} = \frac{TP}{(TP+FN)}$$

$$F_1\text{score} = 2* \frac{(Precision * Recall)}{(Precision+Recall)}$$

- **Sensitivity** (also called as **true positive rate**) calculates the number of actual positives that are correctly identified.
- **Specificity** (also called as **true negative rate**) calculates the number of actual negatives that are correctly identified.
- **Recall** is referred to the true positive rate or sensitivity.
- **Precision** refers to how close estimates from varied samples are to each other.
- **F_1score** is the harmonic mean of precision and recall an also called as traditional F-measure or balanced F-score.

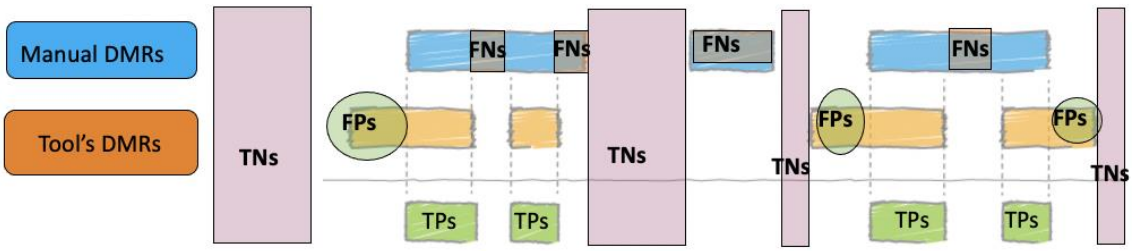


Figure S2: **Confusion matrix calculation for manual DMRs (green) and the tool's DMRs (orange).** True Positives (TPs) are calculated as the intersection of "tool's list" AND "manual list". True Negatives (TNs) are calculated as the intersection of "complement of tool's list" AND "complement of the manual list. False Negatives (FNs) are calculated as the intersection of "complement of tool's list" AND "manual list". If the manual list has an intersection with a complement of tool's list, then it is a wrong negative sign. False Positives (FPs) are calculated as the intersection of "complement of manual DMRs" AND "tool's DMRs.

The analysis starts with comparing the manual DMRs and the tool's DMRs. "Bedtools complement" is used to find the complement regions of these two lists. Complement of the manual DMRs and complement of tool's DMRs named as a "complement_manual_dmrs" and "complement_tool_dmrs" respectively. After that, "bedtools intersect" [73] was used to find overlapping positions between the four files; manual list, complement_of_manual list, tool's list, complement_of_tool's list. As the last step, the intersection of confusion matrix values (TP's, TN's, FP's, and FN's) was considered per position. The reason behind this last intersection step was to decrease the number of positions that were counted more than once.

Confusion matrix can also be explained like this:

```
1. bedtools complement –i tools_dmrs –g all_Cytosines_for_this_context #Complement of
   a tool in interest
2. bedtools complement –i manual_dmrs –g all_Cytosines_for_this_context #Complement of
   manual DMRs
3. bedtools intersect -a tool_dmrs.bed -b manual_dmrs.bed | bedtool intersect -a - -b
   all_methylated_Cytosines_for_this_context.txt | wc –l #TPs
4. bedtools intersect -a tool_dmrs.bed -b complement_manual_dmrs.bed | Bedtools
   intersect -a - -b all_methylated_Cytosines_for_this_context.txt | wc –l #FPs
5. bedtools intersect -a complement_tool_dmrs.bed -b manual_dmrs.bed | Bedtools
   intersect -a - -b all_methylated_Cytosines_for_this_context.txt | wc –l #FNs
6. bedtools intersect -a complement_tool_dmrs.bed -b complement_manual_dmrs.bed |
   Bedtools intersect -a - -b all_methylated_Cytosines_for_this_context.txt | wc -l
   #TNs
```

Five main criteria used in benchmarking but F1 scores are constituted the final decision. Criteria are:

- Real running time in seconds with simulated datasets. Please check Kreutz et al., 2020 for more information [70].
- Peak RAM in KBs with real datasets.
- Specificity scores with real datasets.

- Sensitivity scores with real datasets.
- F_1 scores with both real & simulated datasets

Confusion matrix values (TP, FP, TN, and FN), recall ($\frac{TP}{(TP+FP)}$), precision ($\frac{TP}{(TP+FN)}$), specificity ($\frac{TN}{(TN+FP)}$) and sensitivity ($\frac{TP}{(TP+FN)}$) were calculated for all species per methylation context. Ranks were given with the Excel RANK function as inversely correlated with the scores. Higher recall, precision, specificity, sensitivity, and F1 scores for both real ($F1_{real}$) and simulated($F1_{sim}$) values got the highest rank as 1. The final benchmarking criteria were based on the $\frac{(2*(F1sim)+F1real)}{3}$ formula. Figures from S3 to S6 outlines the rank statistics per species and S7 shows overall statistics for all species together.

tool	average rank (sim & real F1)	average sensitivity	average rank of sensitivity	average specificity	average rank of specificity
Metilene	1,33	0,2223	3,67	0,9022	3,33
DMRCate	2,33	0,2126	2,67	0,9257	5,33
Defiant	2,33	0,3585	2,67	0,9958	3,33
MethylKit	4,67	0,3122	3,33	0,9892	5,00
MOABS	5,00	0,3671	3,33	0,9996	2,33
MethylScore	5,33	0,0758	5,33	0,9975	1,67
MethylSig	7,00	0,0000	7,00	0,0000	7,00
Bsmooth	7,00	0,0000	7,00	0,0000	7,00

Figure s3: **$F1_{simulated}$ and $F1_{real}$ scores ranking for *P.patens* datasets.** Averaged ranks for simulated and real F1 scores with sensitivity and specificity. Metilene seems to be the best and is followed by DMRcate.

tool	average rank (sim & real F1)	average sensitivity	average rank of sensitivity	average specificity	average rank of specificity
Defiant	1,67	0,2855	2,00	0,9218	4,00
Metilene	3,67	0,2617	2,33	0,9232	3,00
MethylScore	3,67	0,2260	3,67	0,9729	1,67
MethylKit	4,33	0,2716	3,00	0,9159	4,33
Bsmooth	4,33	0,0000	6,00	0,0000	6,00
MOABS	4,33	0,0477	4,00	0,9626	2,00
DMRCate	6,00	0,0000	6,00	0,0000	6,00
MethylSig	7,00	0,0000	6,00	0,0000	6,00

Figure s4: **$F1_{simulated}$ and $F1_{real}$ scores ranking for *P.abies* datasets.** Averaged ranks for simulated, real F1 scores with sensitivity and specificity. Defiant seems to be the best and is followed by metilene.

tool	average rank (sim & real F1)	average sensitivity	average rank of sensitivity	average specificity	average rank of specificity
Metilene	2,00	0,1188	2,67	0,9988	4,17
Defiant	3,20	0,0596	4,17	0,9999	1,67
MOABS	3,60	0,2995	3,50	0,9937	2,50
MethylSig	3,80	0,1643	3,83	0,9662	6,17
MethylScore	4,00	0,1078	3,17	0,9996	3,83
MethylKit	4,20	0,1350	2,50	0,9992	2,83
Bsmooth	4,40	0,0101	6,00	0,9999	6,17
DMRCate	5,80	0,0000	6,33	0,0000	7,17

Figure s5: ***$F1_{simulated}$ and $F1_{real}$ scores ranking for Ae. arabicum datasets.*** Averaged ranks for simulated and real F1 scores with sensitivity and specificity. Metilene seems to be the best and is followed by defiant.

tool	average rank (sim & real F1)	average sensitivity	average rank of sensitivity	average specificity	average rank of specificity
Metilene	1,18	0,0217	2,33	0,9982	4,33
DMRCate	1,42	0,0894	2,67	0,6195	7,00
Defiant	1,84	0,0235	2,00	0,9950	5,67
MethylKit	3,15	0,0034	4,33	0,9993	3,00
MethylScore	3,15	0,0016	4,67	1,0000	2,33
MethylSig	4,13	0,0000	6,00	1,0000	1,00
MOABS	4,76	0,0019	5,00	1,0000	3,00
Bsmooth	5,44	0,0000	7,00	0,0000	8,00

Figure S6: ***$F1_{simulated}$ and $F1_{real}$ scores ranking for A. thaliana datasets.*** *Averaged ranks for simulated and real F1 scores with sensitivity and specificity. Metilene seems to be the best and is followed by DMRcate.*

tool	average rank (sim & real F1)	how many NAs?	average sensitivity	average rank of sensitivity	average specificity	average rank of specificity
Metilene	2,5333	0	0,1111	2,69	0,8995	4,46
Defiant	3,0667	0	0,1156	3,15	0,9209	3,69
Defiant&Metilene	3,0667	0	0,0443	4,46	0,9228	2,31
DMRCate	4,6667	2	0,2494	4,08	0,6622	7,08
MethylScore	4,6667	0	0,0677	4,23	0,9223	3,23
MethylKit	5,2000	0	0,1351	3,38	0,9200	4,08
Bsmooth	5,6667	3	0,0013	6,58	0,1250	7,00
MOABS	5,6667	0	0,2234	3,92	0,9200	2,77
MethylSig	6,0667	2	0,0758	5,38	0,6767	5,38

Defiant&Metilene
is the overlap of both
(F1sim*2 + F1real) / 3

NA, not applicable
if tool could not be run on a dataset

Figure S7: **Final benchmarking table for all species.** Averaged ranks calculated as $\frac{(2*(F1_{sim})+F1_{real})}{3}$ for all species. The "how many NAs" column (the 2nd column) gives details about the tool's ability to run depending on the total number of replicates, coverage, and missing data, please check Kreutz et al., 2020 for more information [70]. Average sensitivity and average specificity columns show averaged ranks. Metilene showed superior performance and followed by defiant.

7.3 Documentation, installation, and interpretation of EpiDiverse DMR and EWAS pipelines

7.3.1 Wiki Documentation of the EpiDiverse Toolkit

The EpiDiverse/template pipeline is part of the EpiDiverse Toolkit (https://app.gitbook.com/@epidiverse/s/project/epidiverse-pipelines/best-practice-pipelines, accessed on 1 May 2021) and a best-practice suite of tools intended for the study of Ecological Plant Epigenetics (https://app.gitbook.com/@epidiverse/s/project/, accessed on 1 May 2021). Links to general guidelines and pipeline-specific documentation can be found below:

7.3.2 Installation and pipeline configuration

To start using the EpiDiverse analysis pipelines, follow the steps below:

7.3.2.1 Install Nextflow

Nextflow runs on most POSIX systems (Linux, Mac OSX etc). It can be installed by running the following commands:

```
1. java -version # Make sure that Java v8+ is installed
2. curl -fsSL get.nextflow.io | bash # Install Nextflow v19.09+
3. mv netxflow ~/bin # Add Nextflow binary to your $PATH
4. #sudo nv nextflow /usr/local/bin # OR system-wide installation
```

See nextflow.io (accessed on 1 May 2021) for further instructions on how to install and configure Nextflow itself.

7.3.2.2 Install the pipeline

7.3.2.2.1 Automatic

The pipelines themselves need no installation - Nextflow will automatically fetch them from GitHub if e.g., epidiverse/wgbs is specified as the pipeline name.

7.3.2.2.2 Offline

The above method requires an internet connection so that Nextflow can download the pipeline files. If you're running on a system that has no internet connection, you'll need to download and transfer the pipeline files manually:

```
1. curl -L https://github.com/epidiverse/wgbs/archive/1.0.0.zip -o epidiverse # Download
   the latest release of the pipeline e.g., (see
   https://github.com/epidiverse/wgbs/releases (accessed on 1 May 2021))
2. unzip epidiverse-wgbs-v1.0.0.zip
3. cd /path/to/my/data
4. nextflow run /path/to/pipelines/epidiverse-wgbs-v1.0.0 [PARAMETERS]
```

7.3.2.2.3 Development

If you would like to make changes to the pipeline, it's recommended to fork on GitHub and then clone the files. Once cloned you can run the pipeline directly as above.

7.3.2.3 Pipeline configuration

By default, the pipelines run with the -profile standard configuration profile. This uses several sensible defaults for process requirements and is suitable for running on a simple (if powerful!) basic server. You can see this configuration in conf/base.config from the base directory of each pipeline repository.

Be warned of two important points about the default configuration:

1. The default profile uses the local executor

    - All jobs are run in the login session. If you're using a simple server, this may be fine. If you're using a compute cluster, this is bad as all jobs will run on the head node.
    - See the Nextflow docs (https://www.nextflow.io/docs/latest/executor.html, accessed on 1 May 2021) for information about running with other hardware backends. Most job scheduler systems are natively supported.

2. Nextflow will expect all software to be installed and available on the `$PATH`

7.3.2.3.1 Configuration profiles

Nextflow can be configured to run on a wide range of different computational infrastructures. In addition to pipeline-specific parameters, you will need to define system-specific options.

For more information, please see the Nextflow documentation:
(https://www.nextflow.io/docs/latest/, accessed on 1 May 2021).

Whilst most parameters can be specified on the command line, it is usually sensible to create a configuration file for your environment. A template for such a config can be found in assets/custom.config from the base directory of each pipeline repository. If you are the only person to be running this pipeline, you can create your config file as `~/.nextflow.config` and it will be applied every time you run Nextflow. Alternatively, save the file anywhere and reference it when running the pipeline with config /path/to/config.

If you think other people are using the pipeline who would benefit from your configuration (e.g., other common cluster setups), please let us know. We can add a new preset configuration profile which can be used by specifying `-profile <name>` when running the pipeline.

The pipelines already come with several such config profiles - see the installation appendices and usage documentation for more information.

7.3.2.3.2 Software dependencies: bioconda

If you're unable to use either Docker or Singularity but you have conda installed, you can use the bioconda environment that comes with the pipeline. Using the predefined -profile conda configuration when running the pipeline will take care of this automatically.

If you prefer to build your environment, running this command will create a new conda environment with all of the required software installed:

```
1. conda env create -f environment.yml
2. conda clean -a #recommended, not essential
3. conda activate wgbs #Name depends on the version
```

The env/environment.yml file can be found from the base directory of the pipeline repository. Note that you may need to download this file from the GitHub project page if Nextflow is automatically fetching the pipeline files. Ensure that the bioconda environment file version matches the pipeline version that you run.

7.3.2.3.3 Software dependencies: Docker and Singularity

With either Docker (https://docs.docker.com/engine/install/, accessed on 1 May 2021) or Singularity (https://sylabs.io/guides/3.0/user-guide/, accessed on 1 May 2021) installed, you can use the predefined -profile docker or profile singularity configurations when running the pipeline to take care of the software.

If you prefer to use your container, running the pipeline with the option with-singularity <container> or -with-docker <container> and pointing towards a specific image will allow it to be automatically fetched and used.

If running offline with Singularity, you'll need to download and transfer the Singularity image first:

```
o singularity pull --name epidiverse-[PIPELINE]-[VERSION].simg docker://epidiverse # Once
  transferred, use -with-singularity but specify the path to the image file:
o nextflow run /path/to/epidiverse/[PIPELINE] -with-singularity /path/to/epidiverse
```

7.3.2.4 Running on EpiDiverse infrastructure

To run the pipeline on the EpiDiverse (https://epidiverse.eu/en, accessed on 1 May 2021) servers (epi or diverse), use the command line flag `-profile epi` or `-profile diverse` respectively. This tells Nextflow to submit jobs using the SLURM job executor and use a pre-built conda environment for software dependencies.

There are also three shortcuts available for EpiDiverse species which can be used in place of
--reference in pipelines that require a reference genome.

```
o --thlaspi
o --fragaria
o --populus
```

7.3.3 The EpiDiverse EWAS pipeline documentation

EpiDiverse/ewas is a bioinformatics analysis pipeline for performing epigenome-wide association studies (EWAS) from methylated positions and/or regions, with optional analysis of methQTLs for diploid organisms from variant call data.

The workflow processes a population of sample bed files, usually derived from the EpiDiverse-WGBS pipeline (https://github.com/EpiDiverse/wgbs, accessed on 1 May 2021), and formats them with bedtools (https://github.com/arq5x/bedtools2, accessed on 1 May 2021) for analysis with the R package GEM (https://github.com/fastGEM/GEM, accessed on 1 May 2021). Output in the form of DMPs or DMRs from the EpiDiverse-DMR (https://github.com/EpiDiverse/dmr, accessed on 1 May 2021) pipeline can also be given to pre-filter the number of positions and reduce multiple comparisons. In addition, the union set of DMRs can themselves be used as independent markers within EWAS in place of individual positions. Sample variants, usually derived from the EpiDiverse-SNP (https://github.com/EpiDiverse/snp, accessed on 1 May 2021) pipeline, can additionally be provided to test the association of methylation-SNP pairs as methQTLs. The pipeline provides visualization in the form of Manhattan plots for Emodel, sequence dot plots for Gmodel, genotype interaction plots for the GxE model, and p-value QQ-plots + histograms for all models, each with ggplot2 (https://github.com/tidyverse/ggplot2, accessed on 1 May 2021).

The pipeline is built using Nextflow (https://www.nextflow.io/, accessed on 1 May 2021) a workflow tool to run tasks across multiple compute infrastructures in a very portable manner. It comes with docker containers making installation trivial and results highly reproducible.

7.3.3.1 Quick Start

1. Install Nextflow (https://www.nextflow.io/, accessed on 1 May 2021)
2. Install one of docker (https://docs.docker.com/engine/install/, accessed on 1 May 2021), singularity (https://sylabs.io/guides/3.0/user-guide/, accessed on 1 May 2021) or conda (https://docs.conda.io/en/latest/miniconda.html, accessed on 1 May 2021)
3. Download the pipeline and test it on a minimal dataset with a single command

```
1. NXF_VER=20.07.1 nextflow run epidiverse/ewas -profile test,
   <docker|singularity|conda>
```

4. Start running your analysis!

```
1. NXF_VER=20.07.1 nextflow run epidiverse/ewas -profile <docker|singularity|conda> \
   --input /path/to/wgbs/directory --samples /path/to/samples.tsv
```

See the usage documentation (https://github.com/EpiDiverse/ewas/blob/master/docs/usage.md, accessed on 1 May 2021) for all of the available options when running the pipeline.

7.3.3.2 Workflow

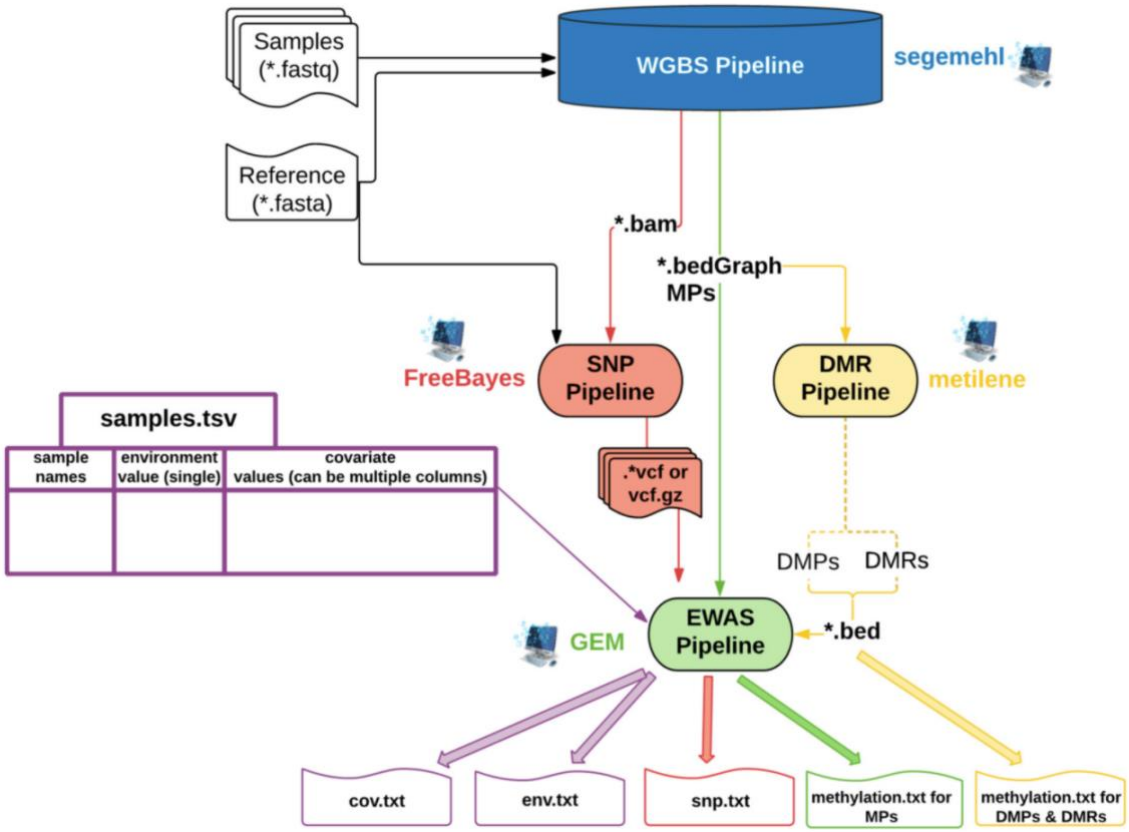


Figure S8: **Workflow of EpiDiverse EWAS pipeline.** The EpiDiverse EWAS pipeline workflow and its interaction with the WGBS, SNP, and DMR EpiDiverse pipelines. Utilized packages or software were specified next to pipeline names. The EpiDiverse epigenome-wide association studies (EWAS) pipeline requires a tab-separated sample.tsv file (shown with purple frame) to specify climatic data and covariate(s) for group determination (can be sampling site, geographical location, or treatment group) and methylation data. As methylation input types, it accepts methylation calls (green arrow) and differentially methylated positions/differentially methylated regions (DMPs/DMRs) (yellow arrow), which can be provided by the whole-genome bisulfite sequencing (WGBS) and the DMR pipelines, respectively. The EWAS pipeline allows running three different models to find epigenetic markers associated with the environment (E), genetic variation (G), or the combination of both (GxE). G and GxE models need single nucleotide polymorphism (SNP) information (red arrow), which can be directly calculated by the SNP pipeline using BS-seq data, or, as for all other inputs, it can be provided by users. See Figures S1–S4 for more detail. * indicates multiple files with the same extension in a specified directory (original paper Figure 1, [68]).

7.3.3.3 Running the EWAS pipeline

This chapter describes the parameter options used by the EpiDiverse EWAS pipeline. The main command for running the pipeline is as follows: `nextflow run epidiverse/ewas [OPTIONS]`

Note that the pipeline will create files in your working directory:

```
○   work/           # Directory containing the nextflow working files
○   ewas/           # Finished results (configurable, see below)
○   .nextflow.log   # Log file from Nextflow
○   .nextflow/      # Nextflow cache and history information
```

Inputs and outputs

--input <ARG> [REQUIRED] Specify the input path for the directory containing outputs from the WGBS pipeline. The pipeline searches for bedGraph files in'`*/bedGraph/{sample_name}.{context}.bedGraph`' format, where sample names must correspond to the sample sheet and context can be either "CpG", "CHG", or "CHH".

--samples <ARG> [REQUIRED] Specify the path to the sample sheet file containing information regarding sample names and corresponding environment and covariate values. The file must contain at least three tab-separated columns: 1) sample names, 2) environment value, 3) covariate values, with further columns optional for additional covariates.

Example samples.tsv file:

#ID	env	cov1	cov2	.....	cov4
sample1	34	1	1		
sample2	42	1	1		
sample3	21	1	2		
sample4	56	1	2		
sample5	76	2	3		
sample6	65	2	3		
sample7	11	2	4		
sample8	22	2	4		
.....					

There can be multiple covariate columns, but the environmental factor can only have one.

--DMPs <ARG> Specify the path to the DMR pipeline output directory to run EWAS analyses in addition to methylated positions filtered by significant DMPs. The pipeline searches for bed files in '`*/{context}/metilene/*/*.bed`' format where context can be either "CpG", "CHG", or "CHH".

--DMRs <ARG> Specify the path to the DMR pipeline output directory to run EWAS analyses in addition to methylated positions filtered by significant DMRs. In addition, the

pipeline will call the union of all significant regions and attempt to run EWAS with whole regions as markers. The pipeline searches for bed files in '*/{context}/metilene/*/*.bed' format where context can be either "CpG", "CHG", or "CHH".

--**SNPs** <**ARG**> ONLY SUITABLE FOR DIPLOID ORGANISMS. Specify the path to the SNP pipeline output directory to enable EWAS analyses Gmodel and GxEmodel which attempt to create a genome-wide methQTL map. The pipeline searches for VCF files in '*/vcf/{sample_name}.{extension}' where sample names must correspond to the sample sheet and the extension can be any standard vcf extension readable by 'bcftools' and defined with –extension parameter. Alternatively, the path to a single multi-sample VCF file can be provided.

--**extension** <**ARG**> Specify the extension to use when searching for VCF files e.g., *.vcf *.bcf or *.vcf.gz [default: *.vcf.gz]

--**output** <**ARG**> A string that will be used as the name for the output results directory, which will be generated in the working directory. [default: ewas]

Model Decision

--**Emodel** Run analysis with "E model". Disables other models unless they are also specified. If no individual model is specified, then all that is possible with the provided inputs will run in parallel [default: off]

--**Gmodel** Run analysis with "G model". Disables other models unless they are also specified. If no individual model is specified, then all that is possible with the provided inputs will run in parallel [default: off]

--**GxE** Run analysis with "GxE model". Disables other models unless they are also specified. If no individual model is specified, then all that is possible with the provided inputs will run in parallel [default: off]

--**noCpG** Disables EWAS analysis in CpG context. Note: at least one methylation context is required for analysis. [default: off]

--**noCHG** Disables EWAS analysis in CHG context. Note: at least one methylation context is required for analysis. [default: off]

--**noCHH** Disables EWAS analysis in CHH context. Note: at least one methylation context is required for analysis. [default: off]

Input Filtering

--**coverage** <**ARG**> Specify the minimum coverage threshold to filter individual methylated positions from the –input directory before running analyses [default: 0]

--**input_FDR** <**ARG**> Specify the minimum FDR significance threshold to include DMPs and/or DMRs from the respective –DMPs and –DMRs directories [default: 0.05]

--proportion <ARG> Minimum proportion of samples that must share a DMP and/or DMR for it to be considered in the analysis [default: 0.2]

--merge When running EWAS using the union set of DMRs as markers, specify to merge adjacent sub-regions into larger regions before methylation averaging and subsequent analysis [default: off]

SNP Filtering

NB: PROVIDING VARIANTS ONLY SUITABLE FOR DIPLOID ORGANISMS.

--max_missing <ARG> Variants that were successfully genotyped in a given proportion of individuals. It can take values from 0 to 1, where 1 means no missing data allowed [default: 0.5]

--mac <ARG> Minor allele count [default: 3]

--minQ <ARG> Minimum quality score [default: 30]

Output Filtering

--Emodel_pv <ARG> Set the p-value to run the "E model". Note: this filter is hardcoded as "1" for Q-Q plot generation and the user-given value is applied to Manhattan plots [default: 0.0001]

--Gmodel_pv <ARG> Set the p-value to run the "G model". [default: 0.0001]

--GxE_pv <ARG> Set the p-value to run the "GxE model". [default: 0.0001]

--output_FDR <ARG> Specify the maximum FDR threshold for filtering EWAS post-analysis [default: 0.05]

Visualization

--kplots <ARG> Specify the number of plots to generate for the top k significant results in the "GxE model" [default: 10]

--distance <ARG> Specify the distance threshold to define cis and trans methQTLs in the dot plot generated for the "G model" output [default: 5000]

7.3.3.4 Understanding the results of the EWAS pipeline

This chapter describes the output produced by the EWAS pipeline.

Pipeline overview

The pipeline is built using Nextflow (https://www.nextflow.io/, accessed on 1 May 2021) and processes data using the following steps:

- Pre-processing -Sample filtering and parsing the sample sheet
- bedtools unionbedg -Combining all samples into single files
- bedtools intersect -Intersecting methylated positions based on DMPs and/or DMRs
- Averaging regions -Calculating average methylation values for given DMRs
- Processing variants -Filtering and merging input variant call file(s)
- E model -Testing the association between methylated positions on the given environmental trait
- G model -Generating methQTL maps for methylated positions and corresponding SNPs
- GxE model -Testing the interaction between methQTLs and the associated environmental trait
- Pipeline Info -Reports from nextflow about the pipeline run

7.3.3.5 Output Directory Structure

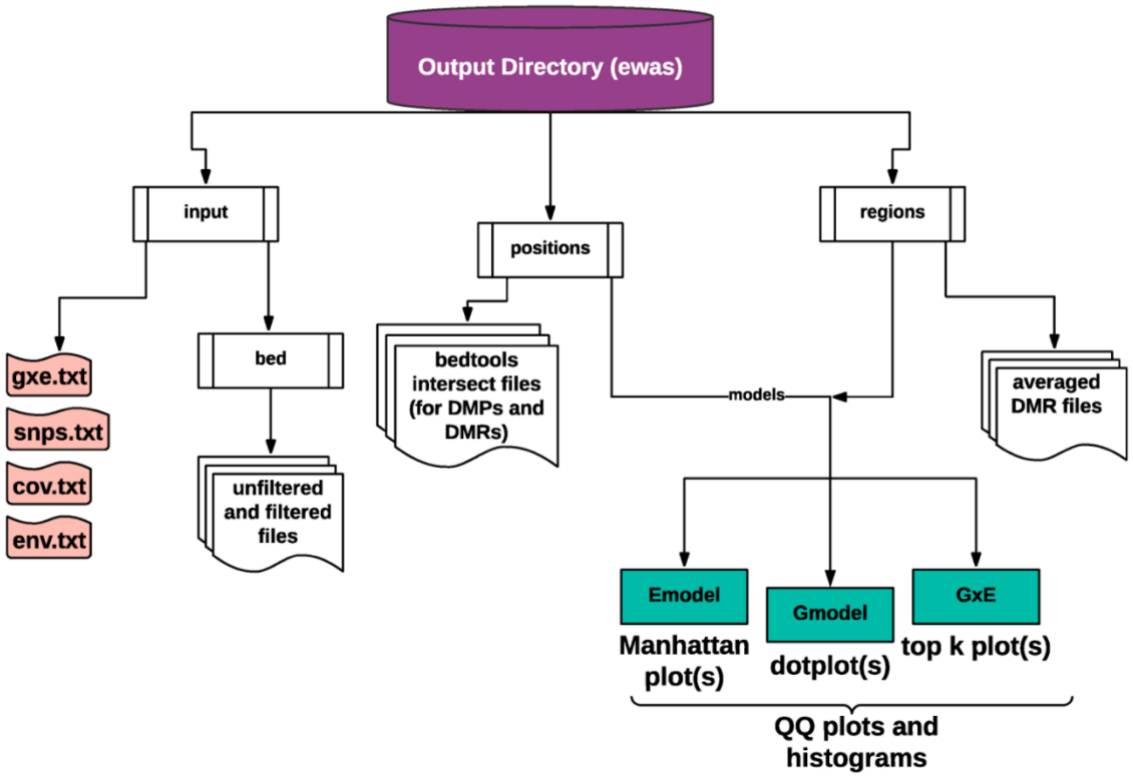


Figure S9: **EpiDiverse EWAS pipeline output directory structure.** EpiDiverse EWAS pipeline generates input directory as default and positions directory with methylation calls, DMPs, and regions directory DMRs. Input directory covers gxe.txt, snps.txt, cov.txt, and env.txt files and bed directory with merged bedGraph files as unfiltered and filtered and missing data estimated. Both positions and regions directories have three subdirectories for outputs and graphs with Emodel, Gmodel, and GxE names. Q–Q plots and histograms are produced with all models (original paper Figure 3, [68]).

Pre-processing

The pipeline requires individual bedGraphs (in each specified methylation context) and a sample sheet with corresponding sample names, environmental trait values, and covariate values to run. The sample sheet is processed into the format required to run GEM, and individual sample bedGraphs are filtered according to coverage. If DMP or DMR comparisons are given, then these will be filtered for a user-specified significance threshold on positions/regions before downstream analysis.

Output directory: ewas/input/

- cov.txt
- env.txt
- gxe.txt

NB: Only saved if the GxE model is enabled during the pipeline run.

Bedtools unionbedg

Following sample pre-processing, the entire collection of each input type is merged into single files per each methylation context.

Output directory: ewas/input/bed

{CpG,CHG,CHH}.bedGraph.bed

chrom	start	end	g1_vs_g2	g1_vs_g3	g2_vs_g3
scaffold_53	166683	166807	NA	NA	0.037
scaffold_53	227390	227644	NA	NA	0.006
scaffold_53	309090	309149	NA	0.000	NA
scaffold_53	309149	309180	0.017	0.000	NA
scaffold_53	309180	309262	0.017	NA	NA
scaffold_53	309535	309715	NA	0.000	0.000
...					

Figure S10: **The output of bedtools unionbedg process for methylation calls.** From *. bedGraph files each position denotes the methylation value for each input sample, and all files denote the presence/absence of a given position/region for individual samples/comparisons by the use of "NA".

{CpG,CHG,CHH}.DMPs.bed

NB: Only saved if DMPs are given during the pipeline run.

{CpG,CHG,CHH}.DMRs.bed

NB: Only saved if DMRs are given during the pipeline run.

chrom	start	end	g1_vs_g2	g1_vs_g3	g2_vs_g3
scaffold_53	166683	166807	NA	NA	0.037
scaffold_53	227390	227644	NA	NA	0.006
scaffold_53	309090	309149	NA	0.000	NA
scaffold_53	309149	309180	0.017	0.000	NA
scaffold_53	309180	309262	0.017	NA	NA
scaffold_53	309535	309715	NA	0.000	0.000
...					

Figure S11: **The output of bedtools unionbedg process for DMRs.** From *. bed files each position denotes the REGION for each input sample, and all files denote the presence/absence of a given position/region for individual samples/comparisons by the use of "NA".

Bedtools intersect

If DMPs or DMRs are given as input types, then the respective union files from bedtools unionbedg are intersected with the entire dataset of individual methylated positions to filter them by positions that are contained within significant DMPs and/or DMRs. This reduces the total number of multiple comparisons when running EWAS and improves sensitivity. All files resulting from this process contain methylated positions denoting the respective methylation value for each input sample.

Output directory: ewas/positions/

{CpG,CHG,CHH}.bedGraph.bed

{CpG,CHG,CHH}.DMPs.bed

NB: Only saved if DMPs are given during the pipeline run.

{CpG,CHG,CHH}.DMRs.bed

NB: Only saved if DMRs are given during the pipeline run.

chrom	start	end	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8	sample9
scaffold_53	390	391	0.50	1.00	NA	NA	0.33	1.00	1.00	0.50	0.00
scaffold_53	392	393	NA	1.00	NA	NA	NA	0.00	NA	NA	NA
scaffold_53	581	582	0.66	0.75	1.00	1.00	0.66	1.00	0.66	0.75	1.00
scaffold_53	583	584	0.87	1.00	1.00	1.00	0.75	0.83	0.77	0.71	1.00
scaffold_53	671	672	0.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
scaffold_53	673	674	0.87	0.93	0.66	0.50	0.85	0.83	0.90	1.00	1.00
...											

Figure S12: **The output of bedtools intersect process.** Respective union files from bedtools unionbedg for DMPs/DMRs are intersected with the entire dataset of individual methylated positions

Averaging regions

If DMRs are given as an input type, then the respective union files from bedtools unionbedg are overlayed with the individual methylated positions to calculate the average methylation value for each sample in each region. These independent regions are then given to the EWAS analysis in addition to the standard test on individual positions. If the –merge option is specified, then regions that are immediately adjacent to each other are merged before calculating the average methylation.

Output directory: ewas/regions/

{CpG,CHG,CHH}.region.bed or {CpG,CHG,CHH}.merged.bed

Union bed files containing the average methylation over each region for each input sample.

Processing variants

If SNPs are given as an input type, then a combination of bcftools and vcftools is given to filter positions and extract a genotype matrix encoded as 1,2,3 for major allele homozygote (AA), heterozygote (AB), and minor allele homozygote (BB) for all SNPs across all samples. The resulting matrix is passed on to the "G model" and "GxE model" where appropriate.

Output directory: ewas/input/

- snps.txt
- Genotype matrix required by GEM.

E model

The E model tests the association between the methylation value on given positions/regions and the environmental trait value specified for each sample in the sample sheet.

Output directory: ewas/{positions,regions}/Emodel

*.log

The log of the stderr from GEM Emodel

*.txt

The full results from GEM Emodel output

*.filtered_*_FDR.txt

The results from GEM Emodel filtered by FDR threshold

cpg	beta	stats	pvalue	FDR
MA_1063600_3380	0.000201348214721003	19.1252478505865	1.9366697826534e-16	1.76740852332208e-09
MA_130823_2338	0.00374034501820006	14.8030798020365	7.06691481350611e-14	3.22463994297191e-07
MA_124616_3396	0.00229308051378857	14.2178616353006	1.74734130014078e-13	5.31542330152315e-07
MA_659042_4763	0.00349331275816753	13.2810848935699	7.92517963527306e-13	1.80813349824787e-06
MA_101037_18934	0.00650347618011456	13.0467326159581	1.17172366089198e-12	2.13863447840995e-06
MA_45879_4444	-0.00204511827706292	-11.3429585756145	2.37823546131811e-11	3.61730366774381e-05
...				

Figure S13: **The output of the Emodel.** This file has "cpg|beta|stats|pvalue|FDR" columns where cpg is for significant chr/scaffold names, beta is a beta coefficient in a linear model, stats is the t-statistics for the C in interest, pvalue is the probabilistic score of a C and FDR is corrected p-values, in other words, q-values.

*.jpg

Q-Q plots on the full results

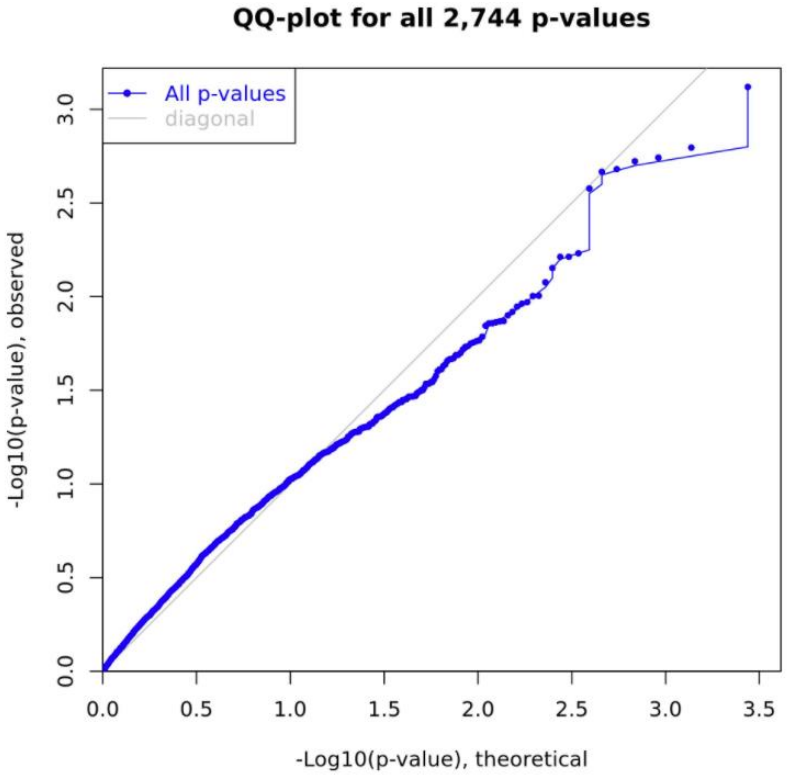


Figure S14: **QQ plot.** QQ plots are generated with all models and give a theoretical vs observed distribution of all p-values. The x-axis is an indicator of normal distribution and ranges between [-4,4]. A total number of p-values can be seen in the header (original paper Figure S7, [68]).

*.png

Manhattan plots on the filtered results

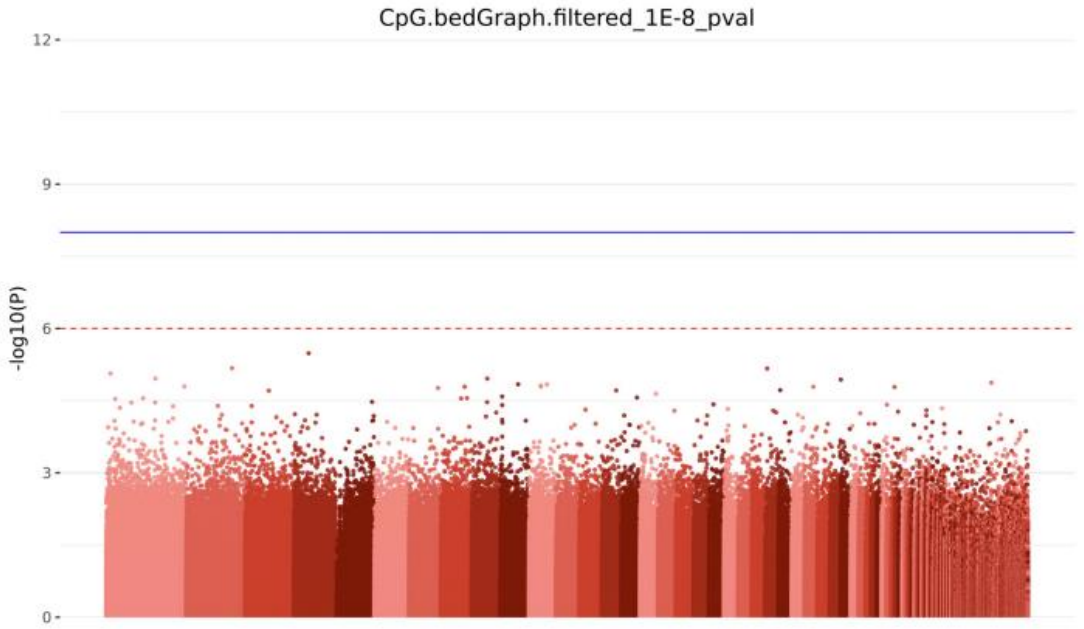


Figure S15: **Manhattan plot.** Example of Manhattan plot output for CpG context and Emodel from the EWAS pipeline, representing all positions below p-value 1e-8. The dashed red line is a suggestive threshold (10-6 by default) and the blue line shows the epigenome-wide significance threshold (suggestive threshold / 100) to narrow down highly significant biomarkers above the lines (original paper Figure S9, [68]).

G model

The G model generates genome-wide methQTL maps to test for an association between the methylation value on given positions/regions and any SNPs which co-occur in each sample. As the matrix of methylated positions vs SNPs is orders of magnitude larger than methylated positions alone, this analysis is divided among individual scaffolds and combined at the end for FDR calculation.

Output directory: ewas/{positions,regions}/Gmodel

*.txt

The full results from GEM Gmodel output

*.filtered_*_FDR.txt

cpg	snp	beta	stats	pvalue	FDR
MA_1063600_3380	SNP962	0.17808859	42.28204	1.482360e-111	1.482360e-106
MA_130823_2338	SNP700	-0.21534752	-18.43573	1.761554e-47	8.807769e-43
MA_124616_3396	SNP578	-0.15171656	-16.70323	9.169281e-42	3.056427e-37
MA_659042_4763	SNP690	0.10567235	13.47239	5.237893e-31	1.309473e-26
MA_101037_18934	SNP589	0.07781375	13.07099	1.112935e-29	2.225870e-25
MA_45879_4444	SNP703	0.13979006	12.55871	5.390763e-28	8.984606e-24
...					

Figure S16: **The output of the Gmodel.** It is a list of C-SNP pairs, where the SNP is the appropriate couple to explain CG in interest. The only different column from the Emodel output is the additional "snp" column next to the ID column. The output of the GxE model is the same.

*.png

Sequence dot plots indicating the relative positions of SNPs and methylated positions in significant methQTLs

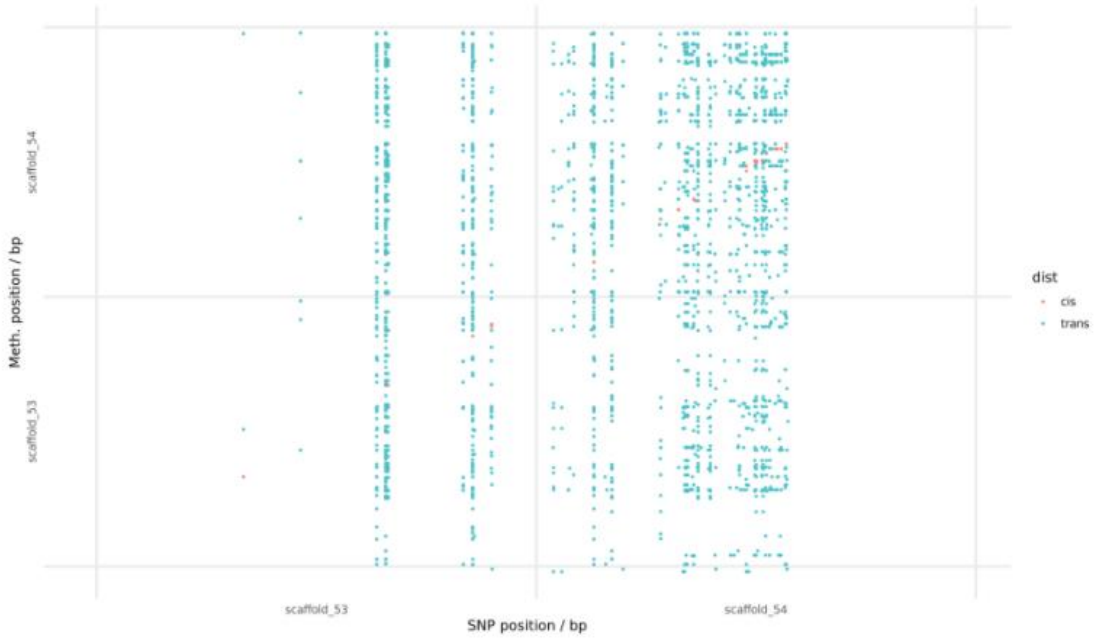


Figure S17: **Sequence dot plot.** They are generated with the G model using Plotly with relative positions of SNPs and methylated positions in significant methQTLs. Cis and trans SNP-Cytosine pairs are marked as red and blue respectively. Scaffold/chr names are written on axes (original paper Figure S10, [68]).

GxE model

The GxE model tests for the interaction between methQTLs and the environmental trait. As the matrix of methylated positions vs SNPs is orders of magnitude larger than methylated positions alone, this analysis is divided among individual scaffolds and combined at the end for FDR calculation.

Output directory: ewas/{positions,regions}/GxE

*.txt

The full results from GEM GxE model output

*.filtered_*_FDR.txt

The results from the GEM GxE model filtered by FDR threshold

cpg	snp	beta	stats	pvalue	FDR
MA_1063600_3380	SNP962	0.17808859	42.28204	1.482360e-111	1.482360e-106
MA_130823_2338	SNP700	-0.21534752	-18.43573	1.761554e-47	8.807769e-43
MA_124616_3396	SNP578	-0.15171656	-16.70323	9.169281e-42	3.056427e-37
MA_659042_4763	SNP690	0.10567235	13.47239	5.237893e-31	1.309473e-26
MA_101037_18934	SNP589	0.07781375	13.07099	1.112935e-29	2.225870e-25
MA_45879_4444	SNP703	0.13979006	12.55871	5.390763e-28	8.984606e-24
...					

Figure S18: **The output of the GxE model.** It is a list of C-SNP pairs, where the SNP is the appropriate couple to explain CG in interest. The only different column from the Emodel output is the additional "snp" column next to the ID column. The output of the G model is the same.

/.png

Plots for the top K most significant interactions and the associations with the environmental trait for major allele homozygote (AA), heterozygote (AB), and minor allele homozygote (BB) for all SNPs across all samples.

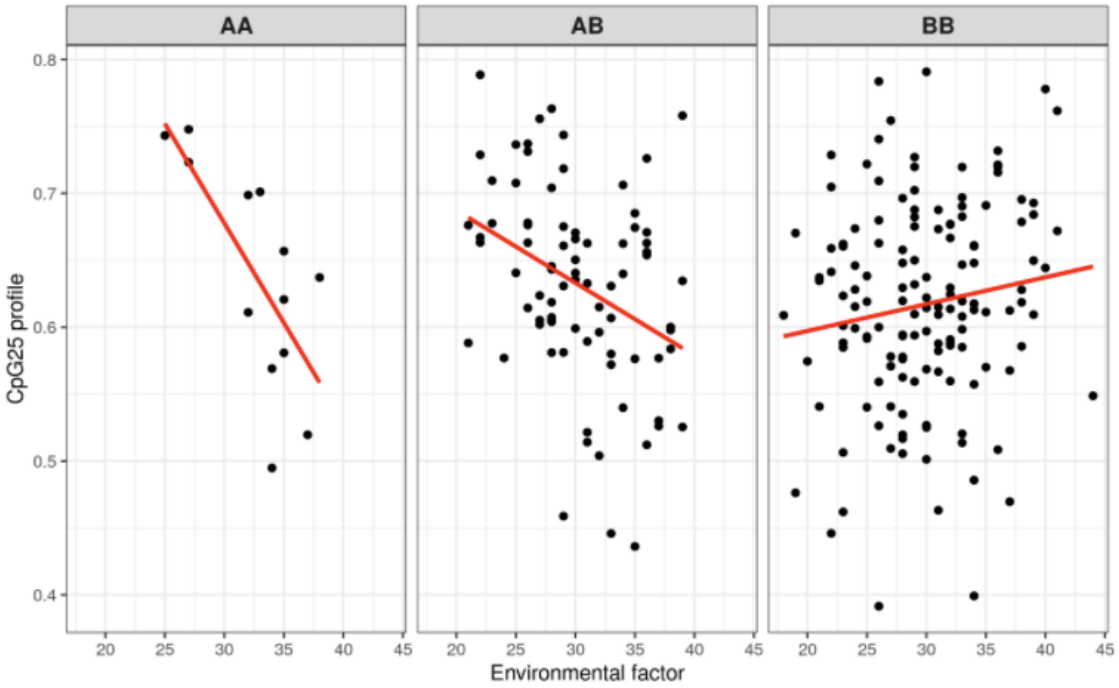


Figure S19: **Top k-plots.** Top k-plots are generated with the phenotypic trait for major allele homozygote (AA), heterozygote (AB), and minor allele homozygote (BB) for all SNPs across all samples are generated with the GxE model. Significant Cytosine name is indicated at the left, not necessarily three alleles have to be produced with each C, the red line shows the slope of the linear relationship between individuals, dots (samples) and environmental factor (climatic data) is shown on the x-axis, methylation beta values are seen on the y-axis. Ref: Pan et al., 2016 (original paper Figure S11, [68]).

Pipeline Info

Nextflow has several built-in reporting tools that give information about the pipeline run.

Output directory: template/

dag.svg: DAG graph giving a diagrammatic view of the pipeline run.

NB: If Graphviz (http://www.graphviz.org/, accessed on 1 May 2021) was not installed when running the pipeline, this file will be in DOT format (https://graphviz.org/doc/info/lang.html, accessed on 1 May 2021) instead of SVG.

- report.html: Nextflow report describing parameters, computational resource usage, and task bash commands used.
- timeline.html: A waterfall timeline plot showing the running times of the workflow tasks.
- trace.txt: A text file with machine-readable statistics about every task executed in the pipeline.

7.3.3.6 Credits

These scripts were originally written for use by the EpiDiverse European Training Network, by Nilay Can ([@nilaycan] (https://github.com/nilaycan, accessed on 1 May 2021)) and Adam Nunn ([@bio15anu] (https://github.com/bio15anu, accessed on 1 May 2021)).

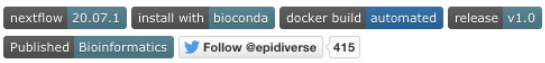
This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 764965.

7.3.3.7 Citation

If you use epidiverse/ewas for your analysis, please cite it using the following DOI: https://doi.org/10.3390/epigenomes5020012, accessed on 1 May 2021.

7.3.4 The EpiDiverse DMR pipeline documentation

EpiDiverse-DMR Pipeline



EpiDiverse/dmr is a bioinformatics analysis pipeline for calling differentially methylated positions (DMPs) or regions (DMRs) from non-model plant species.

The workflow processes raw methylation data from bedGraphs resulting from the EpiDiverse/wgbs pipeline (https://github.com/EpiDiverse/wgbs, accessed on 1 May 2021), which are then grouped for analysis with bedtools unionbedg (https://github.com/arq5x/bedtools2, accessed on 1 May 2021). Each pairwise comparison between groups is performed with metilene (https://www.bioinf.uni-leipzig.de/Software/metilene/, accessed on 1 May 2021), and downstream visualization is carried out with R-packages ggplot2 (https://github.com/tidyverse/ggplot2, accessed on 1 May 2021) and gplots (https://github.com/cran/gplots, accessed on 1 May 2021) to produce distribution plots and heatmaps.

See the output documentation (https://github.com/EpiDiverse/ewas/blob/master/docs/output.md, accessed on 1 May 2021) for more details of the results.

The pipeline is built using Nextflow (https://www.nextflow.io/, accessed on 1 May 2021), a workflow tool to run tasks across multiple compute infrastructures in a very portable manner. It comes with docker containers making installation trivial and results highly reproducible.

7.3.4.1 Quick Start

```
1. #Install Nextflow (https://www.nextflow.io/, accessed on 1 May 2021)
2. #Install one of docker (https://docs.docker.com/engine/install/, accessed on 1 May
   2021), singularity (https://sylabs.io/guides/3.0/user-guide/, accessed on 1 May
   2021) or conda (https://docs.conda.io/en/latest/miniconda.html, accessed on 1 May
   2021)
3. NXF_VER=20.07.1 nextflow run epidiverse/dmr -profile
   test,<docker|singularity|conda> # Download the pipeline and test it on a minimal
   dataset with a single command
4. NXF_VER=20.07.1 nextflow run epidiverse/dmr -profile <docker|singularity|conda> \
   --input /path/to/wgbs/bam --samples /path/to/samples.tsv # Start running your
   analysis!
```

See the usage documentation (https://github.com/EpiDiverse/dmr/blob/master/docs/usage.md, accessed on 1 May 2021) for all of the available options when running the pipeline.

7.3.4.2 Workflow

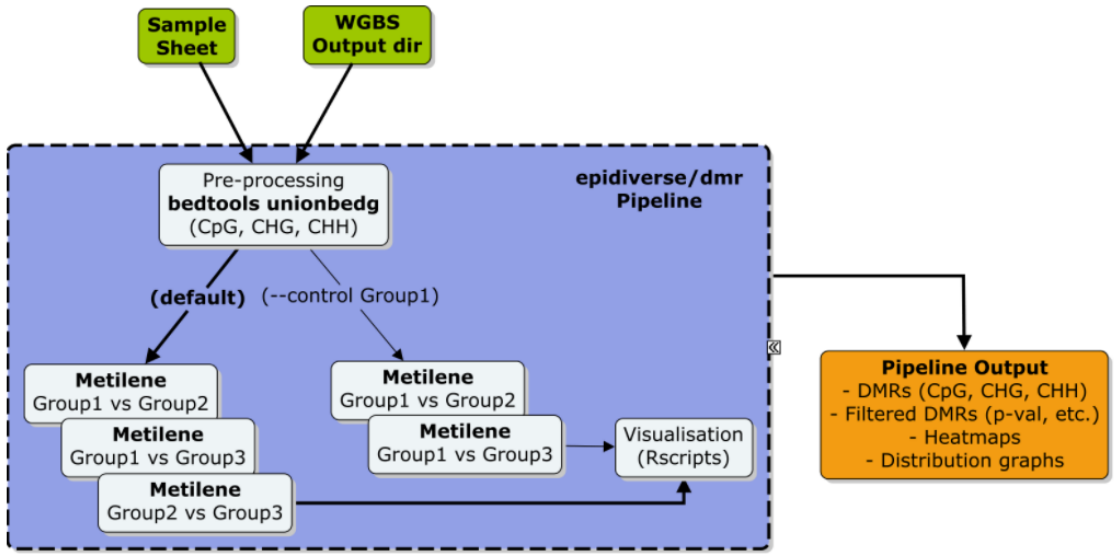


Figure S20: **Workflow of EpiDiverse DMR pipeline.** Required inputs are tab-separated sample sheets with sample name identifiers and methylation calls from the WGBS pipeline output directory. Pre-processing starts with bedtools unionbedg then metilene tries to reveal differential methylation besides some graphical outputs with R scripts.

7.3.4.3 Running the DMR pipeline

The main command for running the pipeline is as follows:

```
nextflow run epidiverse/dmr [OPTIONS]
```

Note that the pipeline will create files in your working directory:

```
work/           # Directory containing the nextflow working files
dmrs/           # Finished results (configurable, see below)
.nextflow.log   # Log file from Nextflow
.nextflow/      # Nextflow cache and history information
```

Inputs and Outputs

--input <ARG> [REQUIRED] Specify the path to the directory containing each sample output from the EpiDiverse/wgbs pipeline (https://github.com/EpiDiverse/wgbs, accessed on 1 May 2021), to be taken forward for analysis. All the sub-directories must correspond to sample names in the file provided to the --samples parameter, and contain within each one a bedGraph directory with files in'*.bedGraph' format. From the wgbs pipeline, only the bedGraphs are necessary for this pipeline to run.

For more information, please refer to the EpiDiverse/wgbs documentation (https://github.com/EpiDiverse/wgbs/blob/master/docs/usage.md, accessed on 1 May 2021) to view the relevant directory structure.

--samples <ARG> [REQUIRED] Specify the path to the "samples.tsv" file, containing information regarding sample names and corresponding groupings/replicates to determine how samples in the input directory should be analyzed. The file must contain three *tab-delimited* columns: 1) sample names, corresponding to the sub-directories in the --input directory. 2) group names, for grouping samples, replicates together. 3) replicate names to provide easy-to-read alternatives for complicated sample names.

Example "samples.tsv" file:

sampleA_1 groupA rep1
sampleA_2 groupA rep2
sampleB_1 groupB rep1

NB: Samples present in the `--input` directory will be ignored if not specified in the "samples.tsv" file.

--output <ARG> A string that will be used as the name for the output results directory, which will be generated in the working directory. This directory will contain subdirectories for each set of reads analyzed during the pipeline. [default: dmrs]

--control <ARG> Specify a string that corresponds to a *group name* in the provided "samples.tsv" file, and the pipeline will run DMR comparisons for each group relative to this group. Otherwise, the pipeline will run all possible pairwise comparisons between groups if no control group is specified. [default: off]

--dmp Specify that DMPs should be analyzed instead of DMRs. [default: off]

DMR/DMP Calling

All options in this section are relevant to DMR calling, while only some are also applicable to DMP calling.

--cov <ARG> Specify the minimum coverage threshold to filter methylated positions *before* running the analyses. [default: 5]

--sig <ARG> Specify the maximum q-value threshold for filtering DMP/DMRs post-analysis. [default: 0.05]

--diff <ARG> Specify the minimum differential methylation level (percent) for filtering DMP/DMRs post-analysis. [default: 10]

--CpN <ARG> Minimum number of Cs a DMR needs to contain to be reported. Not relevant to DMPs. [default: 10]

--gap <ARG> Minimum distance (bp) between Cs that are not to be considered as part of the same DMR. Not relevant to DMPs. [default: 146]

--resample <ARG> Minimum proportion of group samples that must be present in a given position to resample missing data [default: 0.8]

--bonferroni Specify Bonferroni method for multiple comparison testing, otherwise, Benjamini-Hochberg FDR will be used by default. [default: off]

--segSize <ARG> Give a hard cutoff for pre-segmenting regions before DMR identification. Higher values improve runtimes in CHG and CHH context but limit the capacity to identify DMRs that overlap the cutoff location. Can be turned off with 0. [default: 1000]

--segContext <ARG> Give a comma-delimited string of methylation contexts where you wish to apply the heuristic --segSize parameter. [default: CHH]

Additional Parameters

--noCpG Do not call DMP/DMRs in the context. Note: at least one methylation context is required for analysis. [default: off]

--noCHG Do not call DMP/DMRs in the CHG context. Note: at least one methylation context is required for analysis. [default: off]

--noCHH Do not call DMP/DMRs in the CHH context. Note: at least one methylation context is required for analysis. [default: off]

--debug Specify to prevent Nextflow from clearing the work dir cache following a successful pipeline completion. [default: off]

--version When called with nextflow run epidiverse/dmr --version this will display the pipeline version and quit.

--help When called with nextflow run epidiverse/dmr --help this will display the parameter options and quit.

7.3.4.4 Understanding the results of the DMR pipeline

This chapter describes the output produced by the DMR pipeline.

Pipeline overview

The pipeline is built using Nextflow (https://www.nextflow.io/, accessed on 1 May 2021 and processes data using the following steps:

1. bedtools unionbedg - compiling pairwise comparisons between groups
2. metilene - calling DMPs/DMRs
3. Visualization - distributions and heatmaps
4. Pipeline Info - reports from nextflow about the pipeline run

7.3.4.5 Output Directory Structure

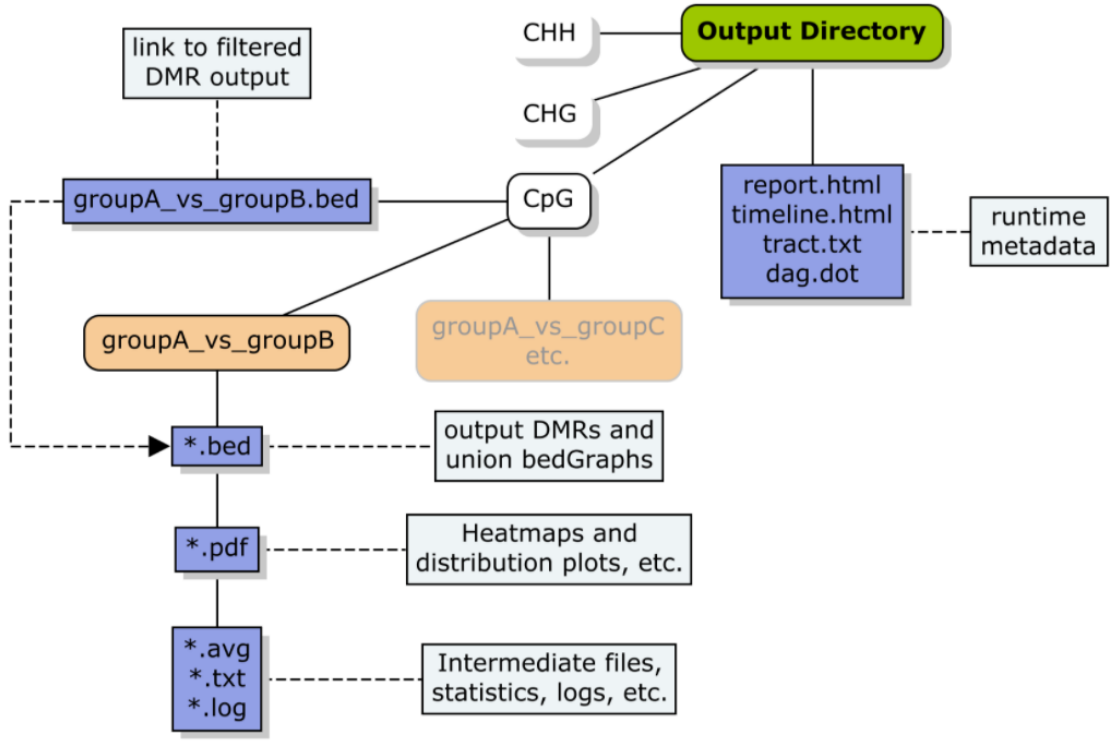


Figure S21: **EpiDiverse DMR pipeline output directory structure.** Output is separated by contexts and shares report.html, timeline.html, tract.txt, and dag.dot files in the output directory. Each context subdirectory has filtered and unfiltered results as .bed format. Heatmaps, distribution plots, etc. are produced as a .pdf besides some intermediate data, statistics, logs, etc.

bedtools unionbedg

The first step of the pipeline is to identify which samples belong to the different groups according to the "samples.tsv" file and produce input files for *each pairwise comparison* between the specified groups by combining the appropriate samples using bedtools unionbedg (https://github.com/arq5x/bedtools2, accessed on 1 May 2021).

Output directory: dmrs/{CpG,CHG,CHH}./input/

e.g., groupA_vs_groupB.bed

- There will be one file according to this naming convention for *each pairwise comparison* according to the number of groups that have been identified in the "samples.tsv" file.
- The number of columns corresponds to the total number of samples in the groups, giving the relevant methylation information (proportion) on each row.

chr	pos	groupA_rep1	groupA_rep2	groupB_rep1
Chr1	109	1.00	0.00	0.57
Chr1	110	1.00	0.00	0.00
Chr1	115	1.00	0.00	1.00
Chr1	116	0.95	0.85	1.00
Chr1	161	0.72	0.77	0.88
Chr1	162	0.80	0.90	1.00
Chr1	310	0.00	0.33	0.57
Chr1	311	0.00	0.75	0.78
...				

Figure S22: **Example groupA_vs_groupB.bed file.** This file has chr, pos, and methylation information per sample.

metilene

DMP/DMR calling is carried out using metilene (https://www.bioinf.uni-leipzig.de/Software/metilene/, accessed on 1 May 2021), using the input generated from bedtools unionbedg in the previous step.

Output directory: dmrs/{CpG,CHG,CHH}./metilene

For each type of output here there will be one file following the naming convention for *each pairwise comparison* according to the number of groups that have been identified in the "samples.tsv" file. The FIRST group in the name is considered the "control" or "comparison" group and e.g., methylation difference is given *relative* to this group.

- e.g., groupA_vs_groupB/groupA_vs_groupB.log

Logfile for each pairwise comparison performed by metilene

- e.g., groupA_vs_groupB/groupA_vs_groupB.bed

The raw, unfiltered output from metilene. This contains all identified DMP/DMRs even if they fall outside the given q-value significance threshold (default 0.05).

Check the metilene documentation (https://www.bioinf.uni-leipzig.de/Software/metilene/Manual/#10_output, accessed on 1 May 2021) to understand this format.

- e.g., groupA_vs_groupB/groupA_vs_groupB.0.05.bed

#chr	start	end	CpN	meth. diff.	significance	length
Chr1	433046	433090	10	−0.791500	0.027401	44
Chr1	656192	656381	10	−0.798500	0.039691	189
Chr1	670553	670943	15	−0.615333	0.034995	390
Chr1	1092268	1092338	12	−0.705000	0.009411	70
Chr1	1344363	1344554	15	−0.559333	0.0022926	191
Chr1	1581203	1581283	12	−0.684167	0.040138	80
Chr1	1593081	1593265	17	−0.660294	0.0047519	184
Chr1	1884168	1884365	13	−0.620385	0.013316	197
...						

Figure S23: **Example groupA_vs_groupB.0.05.bed file.** The is file contains the streamlined output from metilene which is filtered according to the parameters set for the pipeline run. This output is used for all downstream analysis and visualization.

7.3.4.6 Visualization

The pipeline will generate heatmaps and density distribution plots for the DMP/DMRs generated for each pairwise comparison using R packages gplots gplots (https://github.com/cran/gplots, accessed on 1 May 2021) and ggplot2 (https://github.com/tidyverse/ggplot2, accessed on 1 May 2021)

NB: - these will only be generated if DMRs are found within the significance threshold, otherwise, Nextflow will report that these processes have 'failed' harmlessly.

Output directory: dmrs/{CpG,CHG,CHH}./visual

For each type of output here there will be one file following the naming convention for *each pairwise comparison* according to the number of groups that have been identified in the "samples.tsv" file. The FIRST group in the name is considered the "control" or "comparison" group and e.g., methylation difference is given *relative* to this group.

e.g., groupA_vs_groupB/groupA_vs_groupB.0.05.avg

e.g., groupA_vs_groupB/groupA_vs_groupB.0.05.txt

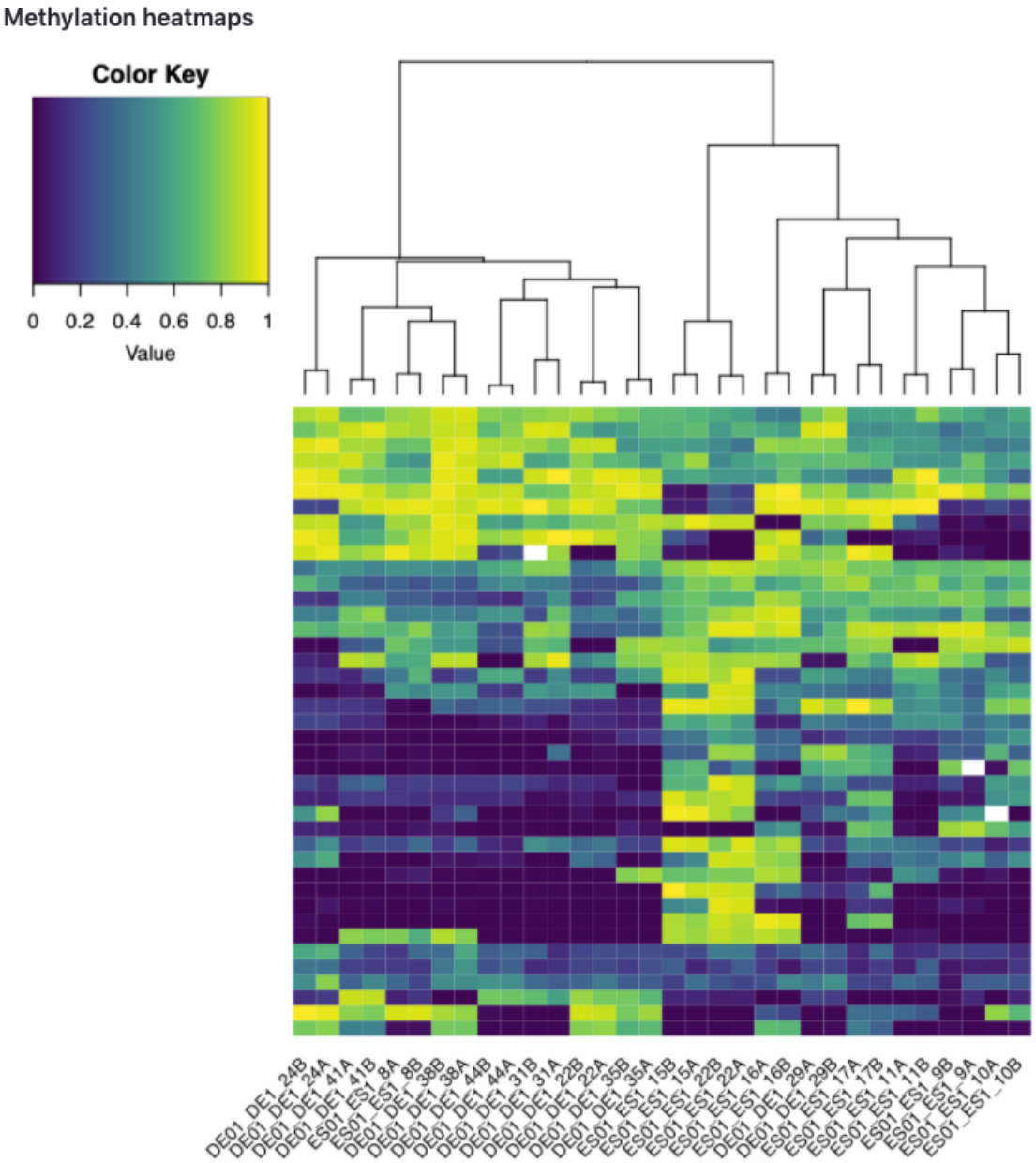


Figure S24: **Methylation heatmaps.** Raw input used to generate heatmaps, consisting of the average methylation level for each region for each sample used to derive the given pairwise comparison. A heatmap showing the relative methylation differences between samples.

e.g., groupA_vs_groupB/groupA_vs_groupB.0.05_Heatmap.pdf

7.3.4.7 Distributions

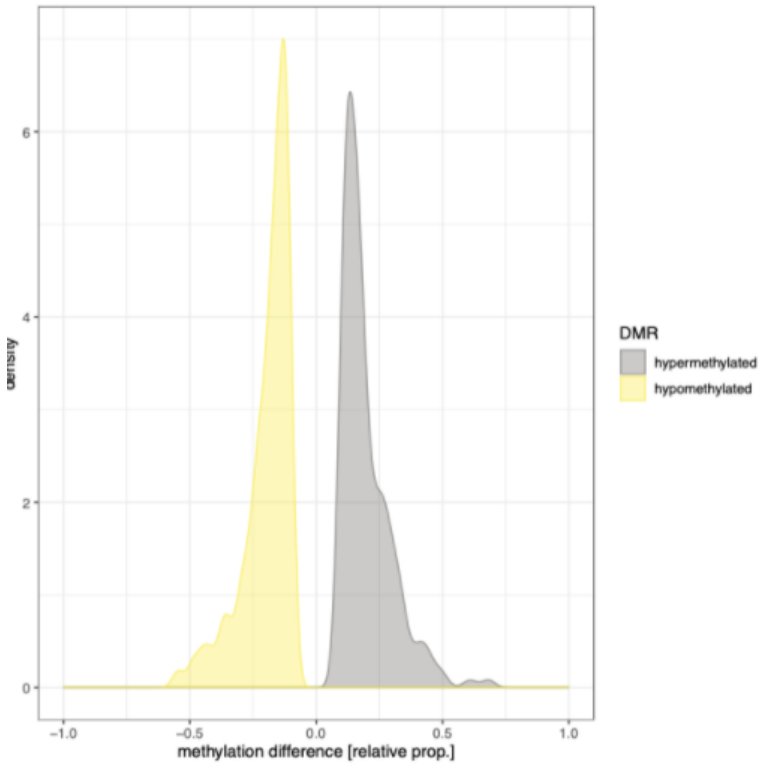


Figure S25: **Methylation difference.** A plot showing the distribution of methylation differences among hyper- and hypo-methylated regions.

- e.g., groupA_vs_groupB/groupA_vs_groupB.0.05_DensDiff.pdf

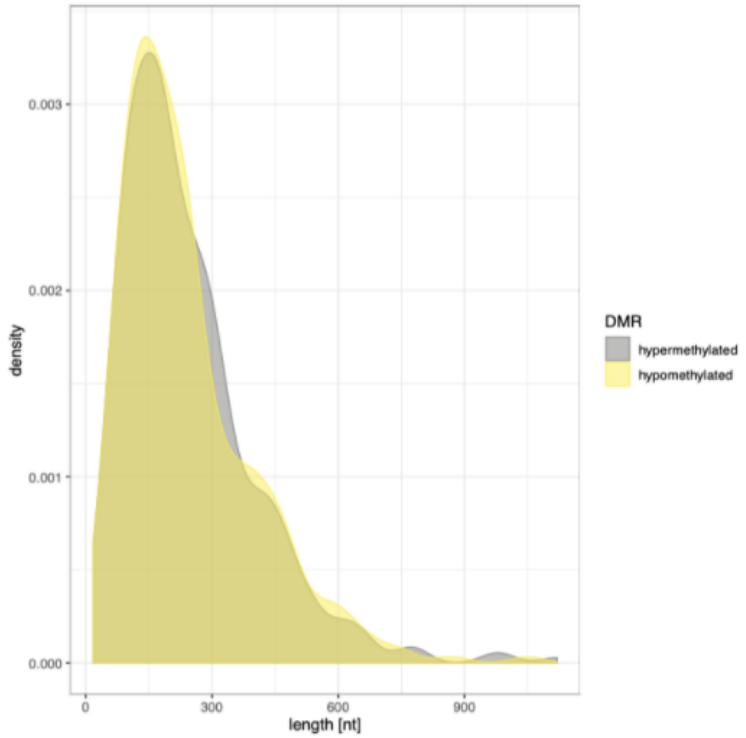


Figure S26: **DMR lengths.** A plot showing the distribution of methylation differences among hyper- and hypo-methylated regions.

- e.g., groupA_vs_groupB/groupA_vs_groupB.0.05_DensLen.pdf

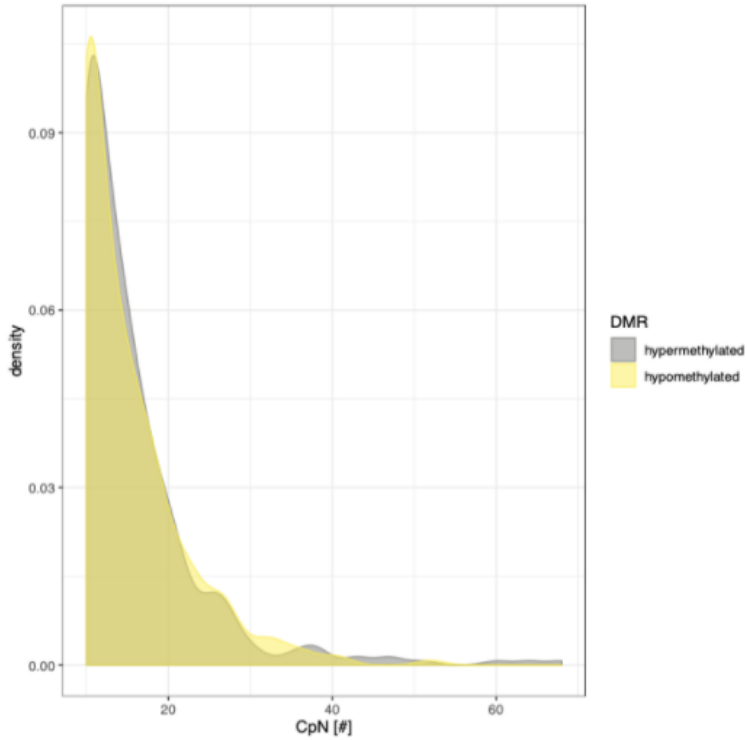


Figure S27: **CpN content.** A plot showing the distribution of CpN contents among hyper- and hypo-methylated regions.

- e.g., groupA_vs_groupB/groupA_vs_groupB.0.05_DensCpN.pdf

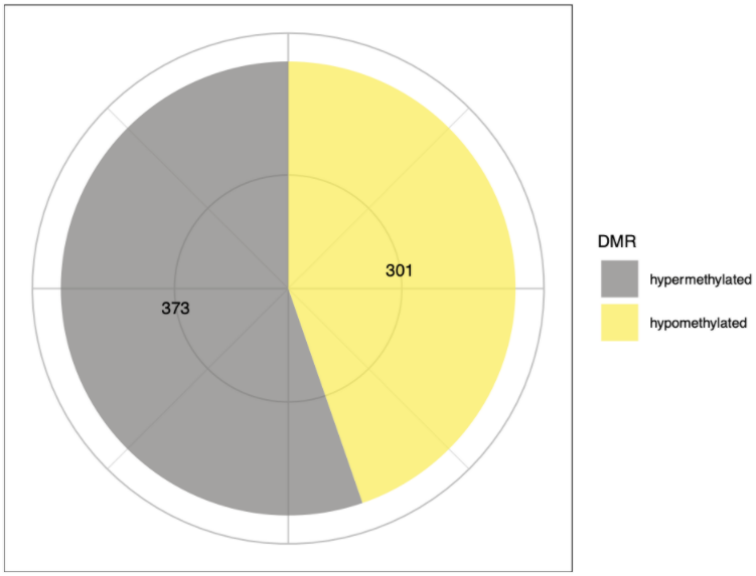


Figure S28: **Hyper- vs hypo-methylation counts.** A simple pie chart showing the number of hyper- and hypo-methylated regions.

- e.g., groupA_vs_groupB/groupA_vs_groupB.0.05_Piechart.pdf

7.3.4.8 Credits

These scripts were originally written for use by the EpiDiverse European Training Network, by Adam Nunn ([@bio15anu] (https://github.com/bio15anu, accessed on 1 May 2021)) and Nilay Can ([@nilaycan] (https://github.com/nilaycan, accessed on 1 May 2021)).

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 764965.

7.3.4.9 Citation

If you use epidiverse/dmr for your analysis, please cite it using the following DOI: https://doi.org/10.1093/bioinformatics/btaa191, accessed on 1 May 2021.

7.3.5 Additional Parameters for all pipelines

--debug Specify to prevent Nextflow from clearing the work dir cache following a successful pipeline completion. [default: off]

--version When called with nextflow run epidiverse/name_of_the_pipeline --version this will display the pipeline version and quit.

--help When called with nextflow run epidiverse/name_of_the_pipeline --help this will display the parameter options and quit.

7.3.6 Software Dependencies

There are different ways to provide the required software dependencies for the pipeline. The recommended method is to use the Conda, Docker, or Singularity profiles as provided by the pipeline.

-profile Use this parameter to choose a preset configuration profile. See chapter "Installation and pipeline configuration" for more information about profiles.
Profiles available with the pipeline are:

- **standard**
  - The default profile used if -profile is not specified.
  - Uses sensible resource allocation for, runs using the local executor (native system calls), and expects all software to be installed and available on the $PATH.
  - This profile is mainly designed to be used as a starting point for other configurations and is inherited by most of the other profiles below.
- **conda**
  - Builds a conda environment from the environment.yml file provided by the pipeline
  - Requires conda to be installed on your system.
- **docker**
  - Launches a docker image pulled from epidiverse/ewas
  - Requires docker to be installed on your system.
- **singularity**
  - Launches a singularity image pulled from epidiverse/ewas
  - Requires singularity to be installed on your system.
- **epi|diverse**
  - Designed to be used on the EpiDiverse clusters epi or diverse
  - Launches jobs using the SLURM executor.
  - Uses pre-built conda environments to provide all software requirements.
- **custom**
  - No configuration at all. Useful if you want to build your config from scratch and want to avoid loading in the default base config for process resource allocation.

If you wish to provide your package containers it is possible to do so by setting the standard or custom profile and then providing your custom package with the command line flags below. These are not required with the other profiles.

-with-conda <ARG> Flag to enable conda. You can provide either a pre-built environment or a *.yml file.
-with-docker <ARG> Flag to enable docker. The image will automatically be pulled from Dockerhub.
-with-singularity <ARG> Flag to enable the use of singularity. The image will automatically be pulled from the internet. If running offline, follow the option with the full path to the image file.

7.3.7 Other command-line parameters

-work-dir <ARG> Specify the path to a custom work directory for the pipeline to run with (e.g., on a scratch directory)

-params-file <ARG> Provide a file with specified parameters to avoid typing them out on the command line. This is useful for carrying out repeated analyses. A template params file assets/params.txt has been made available in the pipeline repository.

-config <ARG> Provide a custom config file for adapting the pipeline to run on your computing infrastructure. A template config file assets/custom.config has been made available in the pipeline repository. This file can be used as a boilerplate for building your custom config.

-resume [<ARG>] Specify this when restarting a pipeline. Nextflow will use cached results from any pipeline steps where the inputs are the same, continuing from where it got to previously. Give a specific pipeline name as an argument to resume it, otherwise, Nextflow will resume the most recent. NOTE: This will not work if the specified run finished successfully, and the cache was automatically cleared. (See: --debug)

-name <ARG> Name for the pipeline run. If not specified, Nextflow will automatically generate a random mnemonic.

7.3.8 Troubleshooting

7.3.8.1 Singularity issues

Sometimes Singularity runs into problems when pulling multiple images at the same time for a pipeline run. In these instances, it is sometimes better to pull the images manually into the directory that the pipeline will be run from, using for instance the following code:

```
singularity pull --name epidiverse-[PIPELINE]-[ VERSION].simg \
docker://epidiverse/[PIPELINE]:[VERSION]
```

Check the Nextflow documentation (https://www.nextflow.io/docs/latest/, accessed on 1 May 2021) for more information about configuring Singularity.

7.3.8.2 Extra resources and getting help

If you still have an issue with running the pipeline, then feel free to contact us at info@epidiverse.eu or by opening an issue in the respective pipeline repository on GitHub asking for help.

If you have problems that are directly related to Nextflow and not our pipelines, then check out the Nextflow Gitter channel (https://gitter.im/nextflow-io/nextflow, accessed on 1 May 2021) or the google group:

(https://groups.google.com/forum/?pli=1&authuser=0#!forum/nextflow, accessed on 1 May 2021).

8 Acknowledgments

Beginning of all, I would like to express my gratitude to all involved in the preparation of this long bioinformatics essay.

First and foremost, I want to thank my research supervisors Prof. Rensing and Dr. Fernández-Pozo that accepted me as a Ph.D. student. Without your assistance and support, this thesis would have never been accomplished. My personal development as a scientist went far beyond anything I could imagine, I'm grateful for this.

Immediately after, I would like to address my respectful thanks towards Lars Opgenoorth, Katrin Heer, Christian Otto, Detlef Weigel, Claude Becker, and Pan Hong the people who were involved in the validation survey and constructive critics for this research project.

A sincere thanks to Önay Can, Iris Sammarco, and Esen Erkilic for their diligent proofreading of this thesis.

I am also thankful and feel fortunate for knowing Adam Nunn, your worthful help and our nerd discussions were priceless. My soulmate Iris Sammarco, Morgane van Antro, Dario Galanti, and Bárbara Díez Rodríguez for their help, discussions, friendship, and joyful moments we had. Whole EpiDiverse consortium for being a great family. David Langenberger for being the hidden hero of my project.

I am deeply grateful to Rabea Meyberg, Leonie Verhage, Lars Bröker, Julien Reinhard, and Marco Göttig for making me quickly adapt to Marburg in a short time. Thanks to all AG Rensing group members for making no day in the lab without joy.

I prost here many friends, Güzin Erdem, Nehir Günce Dasci, Onur Akyürek, Ecem Uzun, Can Yildirim, Sema Yilmaz, Tugce Nur Bozkurt, Derya Bagci, Volkan Dogan, and Ostplatz WG in Leipzig for keeping me partially in touch with the real world.

I would like to acknowledge the support, encouragement, patience, attention, and great love of my parents, Salime and Yusuf Can. Furthermore, I owe so much to my brother Önay Can, thanks for walking beside me and being my best friend. I would like to continue with thanking to my sister-in-law Günes, niece Maya Frida, my nephews Toprak Martin, Uzay Carl for their unconditional love. My aunt, world's most peaceful soul Yasemin Erkilic, my uncle Mahmut Erkilic, my cousins Zahide Erkilic, and Esen Erkilic for unwavering support and guidance in Germany.

Like J.R.R. Tolkien says:

"The Road goes ever on and on //
Down from the door where it began //
Now far ahead the Road has gone //
And I must follow, if I can".

9 Curriculum Vitae

Personal information

Name	Sultan Nilay Can
Born	17th Sep. 1989
in	Ankara, Turkey

Scientific education

2018 – 2021	Ph.D. Thesis title: "Comprehensive analysis of methylation data in non-model plant species", Philipps-Universität Marburg, Germany
2013 - 2017	M.Sc. Bioinformatics, Thesis title: "Mapping and analysis of human disease network map (*diseasome*) on mouse genotype & phenotype network", Middle East Technical University, Ankara, Turkey
2008 - 2013	Bachelor of Science in Statistics, Middle Technical University, Ankara, Turkey

Skills

Languages	Turkish (native)
	English (fluent)
	German (certificate from Philipps-Universität Marburg Sprachenzentrum, A.2.2

Publications

2021

Can, S.N.; Nunn, A.; Galanti, D.; Langenberger, D.; Becker, C.; Volmer, K.; Heer, K.; Opgenoorth, L.; Fernandez-Pozo, N.; Rensing, S.A. The EpiDiverse Plant Epigenome-Wide Association Studies (EWAS) Pipeline. Epigenomes 2021, 5, 12. https://doi.org/10.3390/epigenomes5020012

Nunn, Adam; **Can, Sultan Nilay**; Otto, Christian; Fasold, Mario; Díez Rodríguez, Bárbara; Fernandez-Pozo, Noe; Rensing, Stefan; Stadler, Peter F.; Langenberger, David, NARGAB-2021-086, EpiDiverse Toolkit: a pipeline suite for the analysis of ecological plant epigenetics. Submitted.

2020

Clemens Kreutz, **Nilay S Can**, Ralf Schulze Bruening, Rabea Meyberg, Zsuzsanna Mérai, Noe Fernandez-Pozo, Stefan A Rensing, A blind and independent benchmark study for detecting differentially methylated regions in plants, Bioinformatics, Volume 36, Issue 11, June 2020, Pages 3314–3321, https://doi.org/10.1093/bioinformatics/btaa191

Conferences and talks

2020

3rd EpiDiverse summer school and the annual meeting (virtual), UMR, **talk** "The EpiDiverse EWAS pipeline and the hands-on session"

2018

Black Forest Workshop 2018, the 1st Black Forest Flagellated Plant Workshop (2018), **poster**, "Differentially methylated regions (DMR) tools benchmark with plant species"

Other

The EpiDiverse DMR pipeline development (https://github.com/EpiDiverse/dmr)

The EpiDiverse EWAS pipeline development (https://github.com/EpiDiverse/ewas)

10 Declarations

Die hier vorgelegte Arbeit “Comprehensive analysis of methylation data in non-model plant species”, wurde ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt. Alle Daten, die direkt oder indirekt aus anderen Quellen übernommenen wurden, sind unter Angabe der Quellen gekennzeichnet. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Dies ist mein erster Promotionsversuch.

Marburg, 2021

_____

(Sultan Nilay Can)