FAIRification of DMRichR Pipeline: Advancing Epigenetic Research on Environmental and Evolutionary Model Organisms

Wassim Salam[1,2] https://orcid.org/0009-0001-0372-195X
Marcin W. Wojewodzic[2,3,*] https://orcid.org/0000-0003-2501-5201
Dagmar Frisch[4,*] https://orcid.org/0000-0001-9310-2230

[1] Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany
[2] Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo, Norway
[3] Department of Chemical Toxicology, Norwegian Institute of Public Health, Oslo, Norway
[4] Department of Evolutionary and Integrative Ecology, Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

*Corresponding authors:
dagmar.frisch@igb-berlin.de; MarcinWlodzimierz.Wojewodzic@fhi.no

Key words: Ecology, Epigenetics, *Daphnia pulex*, DMRichR, Methylation, Non-model organism, risk assessment

Abstract

Bioinformatics tools often prioritize humans or human-related model organisms, overlooking the requirements of environmentally relevant species, which limits their use in ecological research. This gap is particularly challenging when implementing existing software, as inadequate documentation can delay the innovative use of environmental models for modern risk assessment of chemicals that can cause aberration in methylation patterns. The establishment of fairness in ecological and evolutionary studies is already constrained by more limited resources in these fields of study, and an additional imbalance in tool availability further hinders comprehensive ecological research. To address these gaps, we adapted the DMRichR package, a tool for epigenetic analysis, for use with custom, non-model genomes. As an example we here use the crustacean *Daphnia*, a keystone grazer in aquatic ecosystems. This adaptation involved the modification of specific code, computing three new species-specific packages (BSgenome, TxDb, and org.db), and computing a CpG islands track using the makeCGI package. Additional adjustments to the DMRichR package were also necessary to ensure proper functionality. The developed workflow can now be applied not only to different *Daphnia* species that were previously unsupported, but also to any other species for which an annotated reference genome is available.

Introduction

Epigenetic research is a crucial field in ecological research for understanding how organisms adapt to environmental challenges (Thiebaut, Hemerly and Ferreira 2019; McGuigan, Hoffmann and Sgrò 2021; Lamka *et al.* 2022) as well as for application in ecotoxicological studies that involve non-model species (Vandegehuchte and Janssen 2014; Šrut 2021). However, the related bioinformatics tools are predominantly oriented towards humans or model organisms for humans, often neglecting the requirements for environmental model organisms. This disparity hinders their development and application in ecological research or modern risk assessment for chemicals. Additionally, a lack of clear documentation or the absence of necessary dependencies further complicates the implementation of existing software for these organisms.

Differential genome-wide methylation analysis involves the comparison of methylation patterns across different conditions to understand the impacts on gene regulation and consequently gene expression (Parle-Mcdermott and Harrison 2011; Li and Tollefsbol 2021). Currently, a multitude of tools exists that are suitable for Differentially Methylated Region (DMR) analysis, each of which makes use of a different differential methylation test. These include but are not limited to: Fisher's exact test, BSmooth, MethylKit, MethylSig, DSS, Metilene, RADMeth, Biseq. Each tool uses a unique approach, with none of them consistently outperforming others during benchmark testing (Piao *et al.* 2021).

One recent tool that stands out for its robust use of a Bayesian framework is DMRichR (v1.7.8) (Hansen, Langmead and Irizarry 2012; Korthauer *et al.* 2019; Laufer *et al.* 2021). It is a powerful resource in epigenetic research which contrasts with other DMR analysis tools by combining both the dmrseq (v1.15.1) and bsseq (v1.38.0) algorithms to identify differentially methylated regions (DMRs) with greater precision and accuracy. It improves analysis by incorporating prior knowledge and probabilistic models to better capture true methylation changes, thus reducing false positives. This method prioritizes methylation sites with higher coverage, using them to infer nearby sites with less data. To identify DMRs, this approach focuses on the comparison of groups of samples (e.g. treatments) rather than within-group levels, and allows the analysis of samples even at a lower sequencing depth of 1-5x (Laufer 2023). DMRs are identified in two steps: first, pooling and weighting data from high-coverage sites to detect differences, and then statistically testing these regions to identify significant changes across the genome (Korthauer *et al.* 2019; Laufer 2023). The

biggest notable drawback of DMRichR is that it is primarily available for typical model species and its application can be challenging for researchers with limited bioinformatics experience. As of its last update in October 2023, only 15 species are supported, which include *Homo sapiens* and *Mus musculus*, while excluding many environmentally important species or evolutionary models such as *Daphnia*.

One of such important additions to the suite of taxa to be used with DMRichR are aquatic organisms. *Daphnia* are keystone organisms in aquatic ecosystems, serving as a vital link in the food web between primary producers and higher trophic levels (Miner *et al.* 2012; Ogorelec *et al.* 2020). In the last decade, the planktonic crustacean *Daphnia* has emerged as an important model in ecological and evolutionary research, and is supported by institutions such as the National Institutes of Health (NIH) (Ebert 2005, 2022). Despite their extensive use in genetic, ecotoxicological, and ecological research (Colbourne *et al.* 2022), genomic resources are typically built for *D. pulex* and *D. magna*, but even for these, bioinformatic tools remain scarce. Providing the implementation of DMRichR for these taxa will facilitate access to the pipelines in the research community, and could potentially drive modern chemical risk assessment when epigenetic effects are of interest.

The use of custom genomes with the R package DMRichR can be achieved by making appropriate modifications in the code of annotationDatabases.R (Laufer 2023). This requires in a first step the computation of three new packages (Fig.1): BSgenome, TxDb and org.db (for description of functionality see below). Next, a CpG islands track must be computed, which can be done using the makeCGI package (Irizarry, Wu and Feinberg 2009; Wu *et al.* 2010). Lastly, a few modifications have to be made to packages used within DMRichR's code, namely within Dmrseq, Annotatr and ChIPseeker.

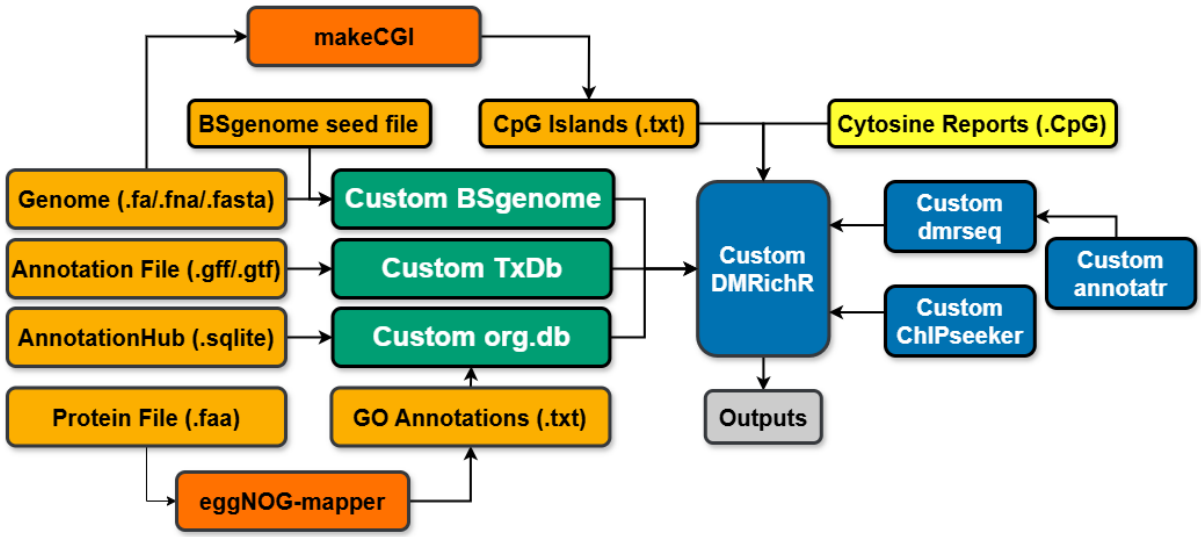

Figure 1. Steps required to adapt DMRichR to a new organism with an annotated genome. **Yellow:** User input files (methylation calls) resulting from a Whole-Genome Bisulfite Sequencing (WGBS) experiment; **Light-orange:** User input files with information on the reference genome; **Dark-orange:** Intermediary tools for producing input files; **Green:** Newly-computed species-specific packages; **Blue:** Modified versions of DMRichR and additional packages used within it; **Grey:** Results generated by DMRichR, which include Blocks, Differentially Methylated Regions (DMRs), Smoothed Individual Methylation Values, Heatmaps.

Implementation

An overview of the workflow is provided in Fig. 1. While here we executed the workflow specifically for *Daphnia pulex,* it could also be applied more generally for any other species with an annotated reference genome following the same steps. The R code (v4.3.2) (R Core Team 2023) associated with each package's computation will be published in Appendix 1. In the next sections we describe the generation or modification made to provide all packages and steps for DMRichR to function with the *Daphnia pulex* genome and the cytosine report files produced by Bismark (Krueger and Andrews 2011).

BSgenome

The BSgenome package enables users to efficiently manage and analyze whole genome

sequences. It provides tools for tasks such as extracting genomic sequences and conducting genome-wide analyses. Apart from DNA methylation analysis, this has additional value for use in other applications such as sequence alignment, variant calling, histone modification (ChiPseq) analysis, and motif discovery (Pagès 2024). The first step is to write a "seed file" following BSgenome guidelines (Pagès 2020), which contains package metadata and instructions on how the package should be compiled. The *D. pulex* ASM2113471v1 genome assembly (NCBI 2021) was used to produce the seed file BSgenome.Dpulex.NCBI.ASM2113471v1-seed (Appendix 2). Following recommended naming conventions, the package BSgenome.Dpulex.NCBI.ASM2113471v1 was computed (Appendix 3).

TxDb

Using either a GFF (General Feature Format) or GTF (Gene Transfer Format) file, a transcript annotation package can be computed. It contains transcript-related annotations (like exons, introns, UTRs), which in addition to being a central package for DMRichR, may be useful for applications such as transcriptomics, RNA-Seq data analysis, Chip-Seq annotation, and gene structure analysis. The *D. pulex* GTF file (NCBI 2022) was used, and in a two-step process, the package TxDb.Dpulex.NCBI.ASM2113471v1.knownGene was computed (Appendix 4).

org.db

This package contains mappings between a central identifier (e.g., Entrez gene IDs) and other identifiers (e.g. gene symbol, gene name, gene ontology, chromosome)(Morgan and Arora 2014). The AnnotationHub package already contains a *D. pulex* database in SQLite format, with ”GID” (Entrez ID) as a central key for its tables. A GO (Gene Ontology) annotation file was obtained by passing the *D. pulex* protein file (protein.faa.gz)(NCBI 2022) to eggNOG-mapper (Huerta-Cepas et al. 2019; Cantalapiedra et al. 2021) with default options. By combining these components we computed org.Duplex.eg.db, following org.db package naming conventions (Appendix 5).

CpG Islands

By using the makeCGI package (Irizarry, Wu and Feinberg 2009; Wu et al. 2010), CpG Islands of any available annotated genome can be de novo discovered to build a CpG islands track,

which is an integral component for the functionality of the DMRichR analysis. Computation of the BSgenome package is a prerequisite to this step (Fig. 1), as makeCGI loads the specified genome from the latter. The posterior probability is one of many parameters that the user can modify, which affects how the package decides what defines a CpG Island. From the CGI files of different organisms already available on the Hao Wu Lab website (Appendix 6), a posterior probability of 0.99 was chosen for all genomes except for that of the fruit fly *Drosophila melanogaster* (0.975). makeCGI was executed with BSgenome.Dpulex.NCBI.ASM2113471v1 (Appendix 7), with a chosen posterior probability similar to that of *D. melanogaster*, because the *Daphnia* genome has a higher resemblance to that of the fruit fly than to the other listed genomes. The genome of *Daphnia*, like that of the fruit fly, is characterized by a small number of methylated bases (Asselman et al. 2016; Kusari et al. 2017). A text file containing CpG Islands entries was therefore produced (Appendix 8). These parameters might need further tuning for other species.

Modifications of DMRichR and Additional Packages

Multiple snippets of code were adjusted within the DMRichR package. Below we provide a short overview of the changes, which can be viewed in detail in Appendix 9 :

- Added new genome “Dpulex” and integrated CGI annotations
- Integrated new BSgenome, TxDb and org.db packages in “annotationDatabases.R”
- Modifications were made to arguments such as “minInSpan”, “bpSpan”, “maxGapSmooth”, “maxGap”, “minNumRegion” and “blockSize”. This was done because *D. pulex* has not only a significantly smaller size but also low and sparse methylation. Lowering the threshold of default arguments allows for more methylation data to be captured.
- Added argument “cytosineReportFormat” (default NULL). Setting a value of “nf-core/methylseq” (Ewels et al. 2023) would enable DMRichR to process cytosine reports generated by nf-core/methylseq which are produced with a slightly different naming convention than when produced by Bismark.
- The packages dmrseq, annotatr and ChIPseeker were adjusted to allow the integration of the new *D. pulex* genome. The respective changes can be seen in Appendix 10, 11 and 12.

Case Study

Proper functionality of the modified R package DMRichR is demonstrated using sample data from an unpublished study involving Whole-Genome Bisulfite Sequencing data from *Daphnia pulicaria*, a member of the *Daphnia pulex* species complex (Dudycha and Tessier 1999). The provided input example presents a cytosine report generated from 10 samples (5 control vs. 5 experimental), and for demonstration purposes has been reduced to report methylation only on chromosome NC_060022.1. The code for this test run is provided in Section 4, which contains instructions on setup and installation. Detailed instructions about how to test the customized DMRichR package can be found in **Appendix 13**. The cytosine reports of each sample (Appendix 14), which are used as input for DMRichR, were produced by nf-core/methylseq (v2.4.0) (Ewels *et al.* 2023). The DMR plot shown in Fig. 2 below displays one of many DMRs obtained by this case study run, which successfully demonstrates the FAIRification of the DMRichR pipeline.

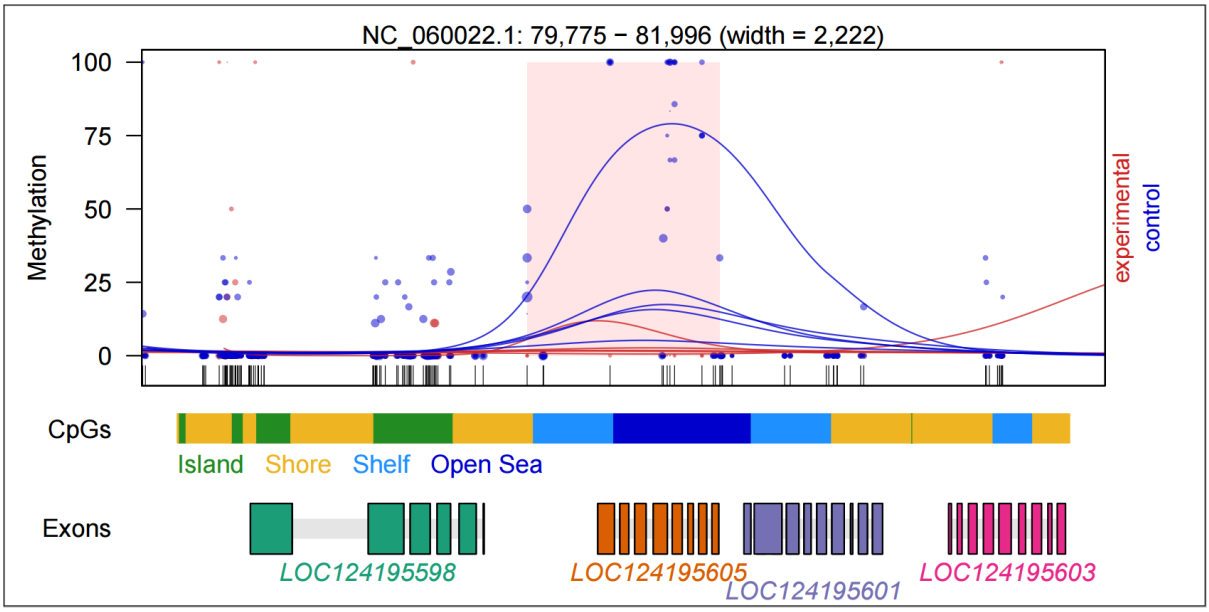

Figure 2. DMR plot displaying a DMR consisting of 13 CpGs with 23% hypomethylation in experimental samples compared to control samples. The methylation level of each CpG site in an individual sample is shown as a point, with its size directly proportional to its coverage. Smoothed methylation levels are represented by lines, color-coded as blue for control samples and red for experimental samples. A track of CpG and gene annotations are additionally displayed under the plot, retrieved from the computed CGI track and org.db package respectively.

Data and Code availability

All analyses were performed using R Statistical Software (v4.3.2) (R Core Team 2023) and Bioconductor (v3.18) (Huber *et al.* 2015). The BSgenome seed-file for *D. pulex*, the code used to compute the above-mentioned packages and the CpG islands list, as well as the sample data (cytosine reports) used in the DMRichR test-run will be made publicly available upon publication. For ease of use, the installations of the computed *D. pulex* packages (BSgenome, TxDb and org.db) were seamlessly integrated into the custom DMRichR package. However, they are as well readily available to use independently of DMRichR. This can prove useful for specific applications, some of which are mentioned in the respective package sections.

Conclusion

Integrating support for the *D. pulex* genome into the DMRichR package represents a significant advancement in the field of ecological and evolutionary genomics. It strengthens the capacity for high-resolution analysis of DNA methylation patterns in *D. pulex*, and enhances the possibility of using whole genome methylation in a modern risk assessment for chemicals. The incorporation of support for the *D. pulex* genome to DMRichR thus allows researchers to leverage this tool's robust functionalities to investigate epigenetic modifications and efficiently use sparse information across different treatments with greater precision. This adaptation not only facilitates deeper insights into the adaptive mechanisms and environmental responses of *D. pulex* but also creates possible use for risk assessment using epigenetics that is still underexplored in ecological studies.

The workflow described here sets a precedent for similar enhancements in other species. The process involves the careful annotation of the target species' genome, followed by integration into the DMRichR framework, thereby enabling the broader scientific community to extend these powerful analytical capabilities to a diverse array of organisms. With an annotated reference genome, increasingly available for many non-model species, the workflow we have described and tested here is particularly beneficial for researchers studying methylation patterns in ecologically and evolutionary significant species, as it bridges the gap between advanced bioinformatics tools and ecological research, fostering a more comprehensive understanding of epigenetic regulation in varied environmental contexts. Lastly, the packages produced in this work contribute not only to the

advancement of differential methylation analysis, but also to other applications improving FAIRness of these tools for environmental research.

Acknowledgments

For computing time we would like to thank the High-Performance Computing Service of ZEDAT, Freie Universität Berlin.

Funding

DF acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation – Project number 461099895) and from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 658714. MWW acknowledges funding from Marie Sklodowska Curie action FP7-PEOPLE-2013-IEF (EU, GB). DF and MWW acknowledge NERC Biomolecular Analysis Facility (Liverpool, GB, NBAF998). The bioinformatic analysis was partially funded by the German Federal Ministry of Education and Research (BMBF, Förderkennzeichen 033W034A).

Competing Interests

None declared.

Availability and Implementation: Code and data will be made available after peer review, upon publication in github and Zenodo.

Author Contributions

WS: Data Curation - lead, Investigation - lead, Writing - original draft; DF & MW: Conceptualisation - equal, Investigation - equal, Supervision - equal, Writing; review and editing - equal; Data Production - equal

268 **References**

269 Colbourne JK, Shaw JR, Sostare E et al. Toxicity by descent: A comparative approach for
270 chemical hazard assessment. Environ Adv 2022;9:100287.
271 Dudycha JL, Tessier AJ. Natural genetic variation of life span, reproduction and juvenile
272 growth in *Daphnia*. Evolution 1999;53:1744–56.
273 Ebert D. Introduction to Daphnia Biology. Ecology, Epidemiology, and Evolution of
274 Parasitism in Daphnia [Internet]. National Center for Biotechnology Information (US),
275 2005.
276 Ebert D. Daphnia as a versatile model system in ecology and evolution. EvoDevo 2022;13:16.
277 Ewels P, Hüther P, Sateesh P et al. nf-core/methylseq: [2.4.0] Gillespie Gaia. 2023, DOI:
278 10.5281/zenodo.8029942.
279 Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing
280 reads to differentially methylated regions. Genome Biol 2012;13:R83.
281 Huber W, Carey VJ, Gentleman R et al. Orchestrating high-throughput genomic analysis with
282 Bioconductor. Nat Methods 2015;12:115–21.
283 Irizarry RA, Wu H, Feinberg AP. A species-generalized probabilistic model-based definition of
284 CpG islands. Mamm Genome Off J Int Mamm Genome Soc 2009;20:674–80.
285 Korthauer K, Chakraborty S, Benjamini Y et al. Detection and accurate false discovery rate
286 control of differentially methylated regions from whole genome bisulfite sequencing.
287 Biostat Oxf Engl 2019;20:367–83.
288 Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq
289 applications. Bioinformatics 2011;27:1571–2.
290 Lamka GF, Harder AM, Sundaram M et al. Epigenetics in Ecology, Evolution, and
291 Conservation. Front Ecol Evol 2022;10.
292 Laufer B. Introduction to DMRichR. 2023.
293 https://www.benlaufer.com/DMRichR/articles/DMRichR.html
294 Laufer B, Hwang H, Jianu JM et al. Low-pass whole genome bisulfite sequencing of neonatal
295 dried blood spots identifies a role for RUNX1 in Down syndrome DNA methylation
296 profiles. Hum Mol Genet 2021;29:3465–76.
297 Li S, Tollefsbol TO. DNA methylation methods: Global DNA methylation and methylomic
298 analyses. Methods San Diego Calif 2021;187:28–43.
299 McGuigan K, Hoffmann AA, Sgrò CM. How is epigenetics predicted to contribute to climate

change adaptation? What evidence do we need? Philos Trans R Soc Lond B Biol Sci 2021;376:20200119.

Miner BE, De Meester L, Pfrender ME et al. Linking genes to communities and ecosystems: *Daphnia* as an ecogenomic model. Proc R Soc B Biol Sci 2012;279:1873–82.

Ogorelec Z, Wunsch C, Kunzmann A et al. Large daphniids are keystone species that link fish predation and phytoplankton in trophic cascades. Fundam Appl Limnol Arch Für Hydrobiol 2020;194, DOI: 10.1127/fal/2020/1344.

Pagès H. BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. Bioconductor 2024.

Parle-Mcdermott A, Harrison A. DNA Methylation: A Timeline of Methods and Applications. Front Genet 2011;2, DOI: 10.3389/fgene.2011.00074.

Piao Y, Xu W, Park KH et al. Comprehensive Evaluation of Differential Methylation Analysis Methods for Bisulfite Sequencing Data. Int J Environ Res Public Health 2021;18:7975.

R Core Team. R: A language and environment for statistical computing. 2023.

Šrut M. Ecotoxicological epigenetics in invertebrates: Emerging tool for the evaluation of present and past pollution burden. Chemosphere 2021;282:131026.

Thiebaut F, Hemerly AS, Ferreira PCG. A Role for Epigenetic Regulation in the Adaptation and Stress Responses of Non-model Plants. Front Plant Sci 2019;10, DOI: 10.3389/fpls.2019.00246.

Vandegehuchte MB, Janssen CR. Epigenetics in an ecotoxicological context. Mutat Res Genet Toxicol Environ Mutagen 2014;764–765:36–45.

Wu H, Caffo B, Jaffee HA et al. Redefining CpG islands using hidden Markov models. Biostatistics 2010;11:499–514.