Received: 19 December 2023 | Revised: 16 August 2024 | Accepted: 2 September 2024

DOI: 10.1111/1755-0998.14021

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Benchmarking sample pooling for epigenomics of natural populations

Ryan J. Daniels[1] | Britta S. Meyer[2] | Marco Giulio[1] | Silvia G. Signorini[1,3] | Nicoletta Riccardi[4] | Camilla Della Torre[3] | Alexandra A.-T. Weber[1]

[1]Department of Aquatic Ecology, Swiss Federal Institute of Aquatic Science and Technology (Eawag), Dübendorf, Switzerland

[2]Department of Biology, Research Unit for Evolutionary Immunogenomics, University of Hamburg, Hamburg, Germany

[3]Department of Biosciences, University of Milan, Milan, Italy

[4]CNR—Water Research Institute, Verbania, Italy

Correspondence
Ryan J. Daniels and Alexandra A.-T. Weber, Department of Aquatic Ecology, Swiss Federal Institute of Aquatic Science and Technology (Eawag), Dübendorf, Switzerland.
Email: jryan.daniels@gmail.com and alexandra.weber@eawag.ch

Funding information
Eidgenössische Anstalt für Wasserversorgung Abwasserreinigung und Gewässerschutz

Handling Editor: Maren Wellenreuther

Abstract

DNA methylation (DNAm) is a mechanism for rapid acclimation to environmental conditions. In natural systems, small effect sizes relative to noise necessitates large sampling efforts to detect differences. Large numbers of individually sequenced libraries are costly. Pooling DNA prior to library preparation may be an efficient way to reduce costs and increase sample size, yet there are to date no recommendations in ecological epigenetics research. We test whether pooled and individual libraries yield comparable DNAm signals in a natural system exposed to different pollution levels by generating whole-epigenome data from two invasive molluscs (*Corbicula fluminea*, *Dreissena polymorpha*) collected from polluted and unpolluted localities (Italy). DNA of the same individuals were used for pooled and individual epigenomic libraries and sequenced with equivalent resources per individual. We found that pooling effectively captures similar genome-wide and global methylation signals as individual libraries, highlighting that pooled libraries are representative of the global population signal. However, pooled libraries yielded orders of magnitude more data than individual libraries, which was a consequence of higher coverage. We would therefore recommend aiming for a high initial coverage of individual libraries (15×) in future studies. Consequently, we detected many more differentially methylated regions (DMRs) with the pooled libraries and a significantly lower statistical power for regions from individual libraries. Computationally pooled data from the individual libraries produced fewer DMRs and the overlap with wet-lab pooled DMRs was relatively low. We discuss possible causes for discrepancies, list benefits and drawbacks of pooling, and provide recommendations for future epigenomic studies.

KEYWORDS
Asian clam, DNA methylation, effect size, Mollusc, pollution, power, zebra mussel

1 | INTRODUCTION

Epigenetics includes the study of the stable but reversible molecular modifications that do not alter the DNA sequence itself (Fallet et al., 2020; Gallego-Fabrega et al., 2015; Paro et al., 2021; Venney et al., 2023). These modifications can be established and removed in response to stimuli (Paro et al., 2021), making them far more variable and potentially noisy compared to genetic variation (Tsai & Bell, 2015). As epigenetic interactions can regulate gene expression (Fallet et al., 2020), this mechanism has received considerable attention (Brander et al., 2017; Marin et al., 2019; Mounger et al., 2021). DNA methylation (DNAm) is the most frequently studied epigenetic modification, particularly in ecological epigenetic research focusing on comparisons between populations, this is in part due to the similarity of the wet-lab and bioinformatic procedures to routine population genomics (Fallet et al., 2020; Lamka et al., 2022). For instance, of the available methods, whole epigenome sequencing (hereafter WepiGS) for example whole-genome bisulfite sequencing (WGBS) and whole-genome enzymatic-conversion sequencing (EMseq) offer the highest resolution available as changes can be tracked as base-pairs across the entire genome (Fallet et al., 2020; Paro et al., 2021; Venney et al., 2023; Ziller et al., 2014).

Epigenetic biomarkers have emerged as promising tools for investigating the drivers of responses to environmental stressors such as temperature and pollution (see Fallet et al., 2020; Jeremias et al., 2020; Venney et al., 2023), in much the same way that genetics has been (e.g. Weber et al., 2013). Notwithstanding, so far the effects of xenobiotics on DNAm were assessed only on a limited number of species and mostly under controlled laboratory conditions (see Ardura et al., 2018; Harney et al., 2022). There is a clear need to increase representation in research to include more ecologically relevant species through the investigation of natural populations subjected to pollution (Šrut, 2021).

Recent reviews of ecological epigenetic research have highlighted the absence of established 'best practices' (Laine et al., 2022), and gaps in taxonomic and geographic sampling, and the lack of adequate replication across a broad range of research topics but particularly in population-level studies (Lamka et al., 2022). Methylation effect sizes in ecological settings tend to be small, so large numbers of samples (e.g. >100 individuals per population or condition) are required to detect differences (Lea et al., 2017). Increasing sample size is not always possible in the case of rare or endangered species, and in most cases the maximum sample size is limited by budget. Indeed, the preparation of individual libraries and sequencing have a strong impact on research costs. In WepiGS studies, data are typically obtained at the individual level as this is the current best-practice, however researchers may be interested in population-wide signals in which case individual variation within a population is not a primary interest. Indeed, ecologists are often most interested the diversity of ecologically important phenotypes and their interaction with environmental conditions across broad regions (Laine et al., 2022; e.g. Han et al., 2020; Tolley et al., 2019). While the cost of sequencing has strongly decreased since its advent (Jobling et al., 2014) and is still decreasing, wet laboratory costs including individual library preparation remain a major obstacle for large sample sizes in many ecological epigenetic research projects. Hence, optimizing these steps is crucial to obtaining data with the highest statistical power in a cost-effective manner.

One means to decrease costs associated with library preparation would be to pool the DNA from individual samples from the same population or condition prior to library preparation. The pooled libraries would thus represent the average signal of the individuals contained therein, with the advantage to prepare a single library. Pooling of DNA samples is commonly used in population genomics, where accurate population allele frequencies can be obtained from a large number of pooled samples (Konczal et al., 2013; Ozerov et al., 2013). Furthermore, pooling has also been used in transcriptomic studies, as it has been shown that pooling RNA samples and reducing coverage are effective ways to optimize costs while maintaining sufficient power in differential expression analyses (Assefa et al., 2020). However, so far, few studies compared the effects of sample pooling using DNAm data and discussion is hampered by the lack of comparative studies (Laine et al., 2022). One of the available studies showed consistent results between individually run samples and pooled samples, with correlation coefficients for CpG array data >.98 (Gallego-Fabrega et al., 2015). Two further studies focusing on mass-spectrometry data from individual and pooled DNA produced strong evidence that pooled DNA samples provide reliable estimates of group DNA methylation averages and showed that the agreement holds up with a range of individuals in a pool (Docherty et al., 2009, 2010). To date, however, the comparison between individual and pooled samples has not been done with sequencing data (whether targeting a reduced fraction of the genome such as bsRADseq, or using whole-genome data (e.g. WGBS or EMseq)). While pooling samples has a strong potential to increase power and reduce costs, there are important considerations related to methylation data which have led to recommending against sample pooling (see Laine et al., 2022; Lea et al., 2017; Ziller et al., 2014). First, methylation data are more variable than genomic data by virtue of their inducibility and reversibility (Tsai & Bell, 2015), with several studies reporting changes in the timescale of days to weeks in response to stressors (see review by Venney et al., 2023). Second, methylation patterns may be tissue specific (Laine et al., 2022; Lee et al., 2017), thus individual samples are not only temporal and spatial snapshots but also somatically heterogeneous. More distantly related cell-types manifest notably divergent methylation patterns, underscoring a significant limitation in methylation analyses (Blake et al., 2020; Ziller et al., 2014). Biases may be introduced if inter-individual (or inter-tissue) variation cannot be accounted for (Teschendorff et al., 2017). Finally, a particular concern has been that pooling masks variation prevents inclusion of covariates (Tsai & Bell, 2015; Ziller et al., 2014) and ultimately requires more biological replicates to account for the hidden variation (Futschik & Schlötterer, 2010). Most importantly, when samples are pooled, there is no way to return to the individual data, so any covariates of interest in the data that were not expected or previously identified in the original pooling design will be masked.

Up to now, the benefits and drawbacks of sample pooling in whole-genome DNAm studies have not been formally compared, and there are currently no clear recommendations on the pertinence of pooling DNA for epigenomics of natural populations. To address this gap, we investigated empirically the effects of sample pooling in DNAm by using whole-epigenome data (WepiGS) from two invasive freshwater bivalves from polluted and unpolluted sites. We set out to use the same resource investment in sequencing for individual and pooled libraries. The aims were: (1) to test whether global DNAm signals from pooled and individual libraries are equivalent, (2) to compare the overlap between differentially methylated regions between polluted and unpolluted sites arising from individual and pooled datasets and (3) to incorporate our observations into a list of benefits and drawbacks of sample pooling, and describe a set of recommendations on the pertinence of sample pooling for future ecological epigenetic projects.

2 | METHODS AND MATERIALS

2.1 | Sampling

Adult individuals of the Asian clam (*Corbicula fluminea*, O. F. Müller 1774) and the zebra mussel (*Dreissena polymorpha*, Pallas 1771) were collected by SCUBA diving at either polluted or unpolluted sites in Lake Maggiore, Italy, and frozen at −20°C upon arrival in the laboratory (Table 1). Sampling permits were not necessary as both species are invasive. Sampling sites were chosen based on the multi-year monitoring of legacy persistent organic pollutants (POPs). The monitoring has been running in Lake Maggiore since 1996 (https://www.cipais.org/web/) (see Table S1 for further details). The Baveno site (polluted locality) is located within the Pallanza Basin, which receives water inputs from the Toce River which is affected by industrial contamination of DDx and Hg. The site is often exceeding the probable effect concentration thresholds for sediments (Guzzella et al., 2018; Marziali et al., 2021) and the concentration of legacy POPs measured *D. polymorpha*, in the freshwater mussel *Unio elongatulus* and eggs of *Podiceps cristatus* (see Table S1) was higher with respect to other sites of the lake (CIPAIS, 2021, 2022; Parolini et al., 2013; Riva et al., 2010). Conversely, Magadino and Cannobio sites (non-polluted) are located in the North side of the lake and are less affected by contamination of legacy pollutants both in sediments and in *D. polymorpha* (CIPAIS, 1999, 2022). No measurements are available for *Corbicula fluminea* at these localities.

2.2 | DNA extraction

We performed DNA extractions for 40 individuals (summary of the experimental design in Figure 1), consisting of 10 samples for each species and for each site. Extractions were performed using foot tissue with the DNeasyBlood and Tissue DNA extraction kit (Qiagen Cat.no. 69504), following the manufacturer's

TABLE 1 Summary of sampling design.

Species	No. of specimens	Locality	Coordinates (decimal degrees)	Sampling date	Depth (m)	Environment (treatment group)	Acronym
Corbicula fluminea	10	Baveno	45.915011 N, 8.503474 E	09.05.18	4	Polluted	CP
	10	Magadino	46.153440 N, 8.852953 E	13.08.20	4	Non-polluted	CNP
Dreissena polymorpha	10	Baveno	45.915011 N, 8.503474 E	22.05.20	6	Polluted	DP
	10	Cannobio	46.092616 N, 8.691536 E	22.05.20	6	Non-polluted	DNP

Note: All localities are in Lake Maggiore, Italy.

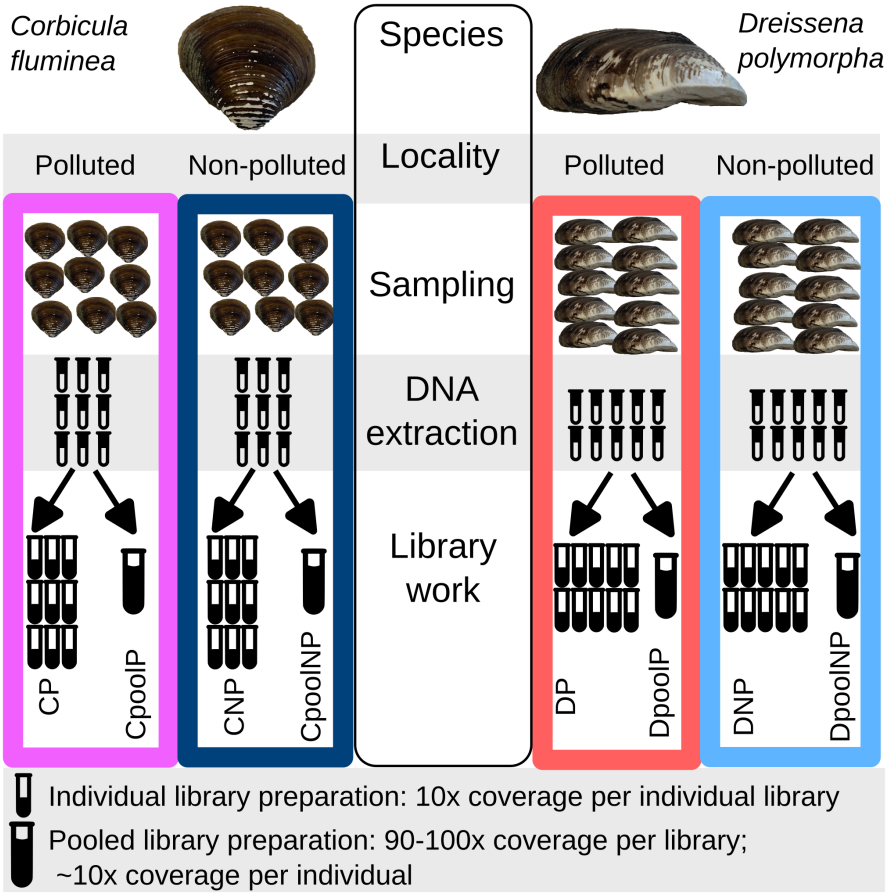

FIGURE 1 Experimental design of the study. Individuals from two species (*Corbicula fluminea* and *Dreissena polymorpha*) were collected at polluted and non-polluted localities in Lake Maggiore, Italy. See Table 1. Individual DNA extractions were performed. The same DNA extractions were used to construct 38 individual and four pooled libraries. Individual and pooled libraries were sequenced at an equivalent per-individual sequencing coverage (i.e. ~10× per individual).

recommendations (Elution in 110 μL elution buffer). DNA integrity was examined using agarose gel electrophoresis and DNA concentration was measured using Qubit 2.0 (Invitrogen). Two *Corbicula* extractions failed (1 CP, 1 CNP), leaving 38 DNA extractions for individual library preparation.

2.3 | Pooling design, enzymatic conversion, library preparation and sequencing

To ensure individual and pooled libraries were comparable, the pooled libraries were prepared using the same individually extracted DNA (Table S2; Figure 1). The concentration of each individual DNA extraction was determined using Qubit HS dsDNA assay. The same amount of DNA (300 ng) was then sheared in a total volume of 60 μL using a Qsonica sonicator (Q800R2 instrument) using different shearing times depending on the level of DNA integrity previously assessed using agarose gel electrophoresis: (1) samples with high-molecular weight DNA were sheared 2′45″ min; (2) samples with semi-degraded DNA were sheared 9′00″–11′30″ min; (3) samples with highly degraded DNA were not sheared. For the samples with highly degraded DNA, control DNA was sheared individually (9 min) and then added to the sample DNA.

For the individual libraries, 25 μL of sheared DNA was used as input volume for library preparation. For the pooled libraries, 2.5 μL of sheared DNA of each individual was pooled to have a starting volume of 25 μL for library preparation (for the *Corbicula* libraries, 2.5 μL of H_20 was added to reach a total volume of 25 μL). The following four pooled libraries were generated; (1) *Dreissena* pool polluted (DpoolP) representing the 10 *D. polymorpha* individuals from the polluted locality; (2) *Dreissena* pool non-polluted (DpoolNP) representing the 10 *D. polymorpha* individuals from the non-polluted locality; (3) *Corbicula* pool polluted (CpoolP) representing the 9 *C. fluminea* individuals from the polluted locality; (4) *Corbicula* pool non-polluted (CpoolNP) representing the 9 *C. fluminea* individuals from the non-polluted locality. We prepared a total of 42 libraries consisting of 38 individual libraries and 4 pooled libraries. We used an enzymatic technique (EMseq) to convert unmethylated cytosines in thymidine as it minimizes DNA damage. We used the NEB Next Enzymatic Methyl-seq Kit (New England Biolabs Cat.no. E7120S). Control DNA (CpG methylated pUC19 and unmethylated lambda) used to estimate conversion rates was added to each DNA extraction before shearing as per the manufacturer's instructions (New England Biolabs). Library preparation was done following the manufacturer's instructions except that we used half volumes of all reagents.

As we wanted to compare pooled and individual libraries on a resource cost basis, we aimed to obtain the same mean sequencing coverage per sample from both the individual libraries and the pooled libraries. Another way to optimize costs further would be to reduce the sequencing effort for pooled libraries, but we do not specifically test this in this study.

We thus combined individual libraries in equimolar concentrations and the pooled libraries in a molar concentration x-fold higher than the individual libraries (i.e. 10× for *D. polymorpha* pools and 9× for *C. fluminea* pools) as the individual and pooled libraries of a particular species were sequenced on the same lane. Specifically, the concentration of each library was determined by Qubit, using the HS dsDNA assay and the average library size was determined by using a TapeStation 4150 Instrument (Agilent Technologies). Molarities were then calculated by the following formula:

$$\text{Conc.}\,[\text{nM}] = \frac{\text{Conc.}\,[\text{ng}/\mu\text{L}] \times 106}{660[\text{g}/\text{mol}] \times \text{average library bp}}.$$

For the *Dreissena* sequencing pool with 20 individual libraries and 2 pooled libraries, 100 fmol were used per individual library and 1000 fmol per pooled library, which resulted in a total amount of 4000 fmol. For the *Corbicula* sequencing pool with 18 individual libraries and 2 pooled libraries, 75 fmol were used per individual library and 675 fmol per pooled library, which resulted in a total amount of 2700 fmol. These two sequencing pools were sequenced on two lanes of a S4 flowcell on an Illumina Novaseq 6000 sequencer (150 bp paired-end) at the Functional Genomics Center, Zürich.

2.4 | Quality control and mapping

In total, 18 *C. fluminea* and 20 *D. polymorpha* individuals were sequenced at an average of 74 (±9.3) million reads (Table S2). The four pooled libraries were sequenced at an average of 620 (±66) million reads. The reads were quality-assessed using FastQC v.0.11.9 (Andrews, 2019) and MultiQC v.1.9 (Ewels et al., 2016). Adapters were identified and removed using Trim Galore! v.0.6.6 (Krueger, 2020) with default settings. To correct for bias of methylation percentage at the read ends, reads were trimmed off 10 bases on both the 3′ and 5′ ends (as recommended; https://felixkrueger.github.io/Bismark/bismark/library_types/). Default settings were retained for all other trimming steps, including the removal of low-quality bases (`--quality 20`) and dropping reads shorter than 20 bases (`--length 20`). Enzyme conversion efficiency was assessed using the two control DNA. The high quality reads having passed QC were then aligned to the respective publicly available reference genomes; *C. fluminea* (Zhang et al., 2021) and *D. polymorpha* (McCartney et al., 2022).

Alignment, de-duplication and methylation extraction were performed with Bismark v.0.24.2 (Krueger & Andrews, 2011). Briefly, we first converted reference genomes computationally for alignment and then indexed using Bowtie2 v.2.4.4 (Langmead & Salzberg, 2012) with default settings (command `bismark_genome_preparation`). Alignment was run with directionality specified using the default alignment score (`--score_min L,0,-1.2`). As part of the QC for the trimmed reads, we compared the number of read-pairs, the level of read duplication and the alignment efficiency between sites within species. The raw Fastq data from each library was split into six files of equal size for parallel alignment. The files were concatenated with Bismark (–multiple) for deduplication. Methylation extraction was performed with default settings, including the –exclude_overlap flag, which only considers data from one of the two strands available in case of overlap between forward and reverse reads. Tests were performed using base R functions including the Shapiro–Wilk test (`shapiro.test`) for univariate normality (Shapiro & Wilk, 1965), the Bartlett test (`bartlett.test`) for homogeneity of variance (Bartlett, 1937) and the ANOVA performed using the `lm` and `summary.aov` functions.

2.5 | Single nucleotide polymorphism detection

Any C to T single nucleotide polymorphisms (SNPs) in our dataset would be incorrectly interpreted as an unmethylated cytosine by Bismark (Krueger & Andrews, 2011). We therefore removed potential variants in CpG sites using the BS-SNPer tool (Gao et al., 2015). Variants were identified for each sequenced library independently using the default settings and a minimum coverage of 10× to correspond to filters applied in MethylKit downstream.

2.6 | Coverage filtering and computational pooling

We processed the aligned reads for CpG sites (dinucleotide sequence of 5′–CG–3′ within a DNA molecule) with the `MethylKit` R package, v.1.24.0 (Akalin et al., 2012) available through Bioconductor (Huber et al., 2015). For each species we excluded unplaced contigs. We decided to retain bases with at least 10 reads (i.e. minimum coverage of 10×).

We note that the mean coverage per individual library was 10× and filtering for 10 reads will cause notable data loss but we opted for this value as it provides a resolution of 10% for changes in methylation.

We further excluded over-represented sites, which may reflect sequencing bias, by removing the sites in the 99.9th percentile of coverage. Regions of 1000 bp size were formed as non-overlapping blocks using the tile function in `MethylKit` with default options (sliding windows of 1000 bp and regions of 1000 bp). In another mollusc, *Crassostrea virginica*, CpG methylation islands have a median length of 1024 bp (minimum, 500 bp; Venkataraman et al., 2020). Our tiles may thus capture many islands in part or in whole.

We further performed computational pooling of individual library data to compare with the wet-lab pooled library data. Computational pooling is a *post-hoc* process that sums up the coverage within each site using the individual library data and creates one library per site or population. We used the individual libraries after filtering (described above) as input data and pooled using the `pool` function in

MethylKit. For the individual libraries, we only included sites that were present in at least 75% of the libraries.

2.7 | Evaluation of concordance between pooled and individual libraries

2.7.1 | Genome-wide global CpG methylation levels

To test for an agreement between the pooled and individual libraries, we fit an overall correlation of the CpG methylation estimates for all samples in a pairwise fashion using Pearson's correlations with the getCorrelation function in MethylKit. We estimated the correlations per chromosome and discussed the coefficient averaged across chromosomes.

To describe the relationship between the signal in the pooled and individual libraries and between the polluted and non-polluted sites, we performed clustering based on the % methylation estimates using a principal component analysis (PCA). To estimate an error on the PCA coordinates, we performed a jackknife over linkage groups, estimating the standard error (Busing et al., 1999; as in Montinaro et al., 2015). To confirm that jackknife iterations were reporting a similar clustering signal, we tested for a correlation between the PC loading matrix across jackknife iterations using the Tucker's coefficient (Lorenzo-Seva & ten Berge, 2006; Peres-Neto & Jackson, 2001).

To prevent any conflicts in the directional of components between jackknife iterations, we used a Procrustes transformation to align each iteration of the PCA with the PCA of the full dataset (Peres-Neto & Jackson, 2001). The transformation coefficients were examined to ensure that no matrix needed excessive transformation to align as this would indicate a big difference in the signal. Where PCs were strongly correlated across jackknife iterations, we proceeded to estimate the error.

2.7.2 | Differential methylation in response to pollution

We tested whether there was overlap between regions showing differential methylation (DMRs) between polluted and non-polluted sites from the individual, pooled and computationally pooled libraries. Differential methylation for the individual libraries was estimated using a logistic regression (Cramer & Howitt, 2004). This regression cannot be conducted with one sample per site (i.e. pooled libraries), so differential methylation was estimated for the pooled and computationally pooled libraries using the Fisher's exact test (Fisher, 1934) (Table S3). *p*-values were corrected for multiple testing under a sliding linear model method (Wang et al., 2011) and we report the *q*-values. Regions were considered to have significant differential methylation (i.e. DMR) with $q < .01$ and a mean methylation difference of at least 10%. To understand the direction of hyper/hypo-methylation, all tests were performed with the following orders for sites: 'Pollution site' versus 'Non-Pollution site'. The number of regions in common between tests were visualized with ggupset (Ahlmann-Eltze, 2020) package in R.

2.7.3 | Methylation profiles by genetic context

To examine if the library preparation schemes recovered qualitatively different DMRs (i.e. with a different genetic context), we profiled the distribution of the DMRs detected according to available annotation features. As the annotation file for *Corbicula fluminea* was not available for this work, we present only the result of *Dreissena polymorpha*. We used the available gene annotations from the UCSC website (as of 2024-03-25, GCF_020536995.1_UMN_Dpol_1.0_genomic). We categorized DMRs as intergenic, exon or intron, with further subcategories where relevant.

2.8 | Estimates of recovered power

To gauge the available power in our dataset, we estimated the recovered power per region in the contrast of polluted and non-polluted sites. The power of a test is defined as the probability that it correctly rejects the null hypothesis when the alternative hypothesis is true. First, we define the mean difference in methylation (hereafter MDM) as the difference in the mean methylation estimates between the populations. We then identified regions which we considered to be non-responsive to pollution as regions with MDM of ~0. This measure takes into account both variance within and between sites and allows for some level of artificial variance due to errors. With the individual libraries power estimates were based on a t-test. Effect sizes were estimated as done by Mansell et al. (2019). We estimated Cohen's *d*, which is the expected difference in means divided by the standard deviation across all samples (Cohen, 1988). The MDM at each locus was based on that calculated in the estimation of DMRs, with $\alpha = .01$ and the observed sample sizes per site per species. We consider only loci with 100% overlap across all samples. The power values were calculated using the pwr.t.test function in the R package pwr (Champely, 2018). For the pooled and computationally pooled libraries, adjustments were needed to replicate the Fisher's exact test. With binomial count data, the variance is a function of the mean (Everitt & Hothorn, 2010), and this allows us to estimate the standard deviation as the square root of the variance function using only the proportions.

$$\text{Variance} = \frac{p_1}{(1-p_1)} n_1 + \frac{p_2}{(1-p_2)} n_2.$$

The effect size was estimated using the ES.h function which uses an arcsine transformation. The power was estimated using the pwr.2p2n.test function in the R package pwr. The pwr.2p2n.test test considers a two-proportion test with unequal sample sizes (i.e. coverage in this context) under the null hypothesis that there is no difference in the site means. The region-specific coverage value was used in the calculation.

2.9 | Estimates of the necessary sampling effort for significant detection

We estimated the distribution of the necessary sampling effort to detect statistically significant differences between polluted and non-polluted sites at each region. Sampling effort estimates were made with the `pwr.t.test` function for the individual libraries and with the `pwr.2p.test` function for the pooled and computationally pooled data. We set the power threshold to 80% (`power = 0.8, n = NULL`) in all cases and we assumed equal sampling effort. In this estimate the sampling effort for individual libraries is measured as the number of biological replicates (each providing a methylation estimate as a continuous number). Sampling effort for the pooled and computationally pooled data are measured as the interaction of the coverage and the number of biological replicates (count data as either methylated or unmethylated read).

2.10 | Laboratory costs estimation

We summarized the costs per sample in a hypothetical scenario where 12 populations from two sites have been sampled (Table 3). We estimated the cost of creating a 'bulk pooled', 'nested pooled' and 'individual libraries' with 8, 4 and 1 individual(s) per library, respectively. The costs were based on quotes as of 2023 in Swiss Francs including local taxes. These costs exclude all procedures which are equivalent between the pooled and the individual libraries (such as sample collection, DNA extraction, DNA quality control and sequencing of libraries, assuming equivalent sequencing depth per individual). We aimed to obtain equivalent sequencing depth per individual in individual and pooled libraries to have the same resource investment in both sites. Reducing sequencing effort of the pooled libraries may be a way to further decrease costs, however we do not specifically test this in this study.

3 | RESULTS

3.1 | Quality control and mapping

All reads were of high quality with an average per base Phred score >32. Filtering by conversion rate efficiencies resulted in the removal of four *C. fluminea* samples with less than 98.5% conversion efficiency and a further two *D. polymorpha* due to possible over-conversion and poor recovery of the control sequences. All four pooled libraries had adequate conversion rates. For the remaining samples the conversion levels of the unmethylated lambda control in the CpG context were 99.34 ± 0.20%, while maintaining methylation levels of 96–98.3% on the pUC19 control (Table S4). For the individual libraries, the average number of reads after filtering and end-trimming was 75 ± 9.3 million (μ ± SD) for *C. fluminea* and 70 ± 9.7 million (μ ± SD) for *D. polymorpha* (Table S2). The pooled libraries had reads slightly under 10× the value of a single individual library; 669 ± 10 million (μ ± SD) for *C. fluminea* and 636 ± 30 (μ ± SD) million for *D. polymorpha*. The statistical comparison of the read QC measures between pollution sites within species showed that all groups had a normally distributed number of duplicated reads and proportion of aligned reads (p-value >.05, Shapiro–Wilk test).

3.2 | Variant detection

Possible SNPs at C/G sites were removed to prevent misinterpretation as C to T enzymatic conversion. We detected a total of 15,781 unique variants across all libraries for *C. fluminea* and 18,216 for *D. polymorpha* (Table S5). The vast majority (95%–96%) of these variants were unique to single individuals (15,108 and 19,190 for *C. fluminea* and *D. polymorpha*, respectively). All variants were removed from downstream analyses.

TABLE 2 Summary of data loss throughout QC.

				Before filter		After filter		Proportion loss	
Species	Treatment	Pool	Libraries	Mean	SD	Mean	SD	Mean	SD
C. fluminea	NP	N	7	31,876,245	1,263,490	664,004	267,223	0.98	0.79
	P	N	7	30,608,103	2,281,176	532,142	290,244	0.98	0.87
	NP	Y	1	46,177,902	–	28,113,914	–	0.39	–
	P	Y	1	45,844,340	–	27,100,500	–	0.41	–
D. polymorpha	NP	N	8	51,086,375	1,789,694	568,528	285,859	0.99	0.84
	P	N	10	47,273,860	3,660,842	334,821	174,217	0.99	0.95
	NP	Y	1	76,126,510	–	48,759,866	–	0.36	–
	P	Y	1	75,759,688	–	47,910,153	–	0.37	–

Note: Presented are the mean (±SD) number of CpG sites before and after applying a filter for 10× minimum coverage, as well as the amount of data lost as a proportion.

3.3 | Larger data loss in individual libraries compared with pooled libraries for equivalent sequencing effort

We applied a conservative minimum coverage filter of 10× for all work. While the per-individual sequencing effort was the same between individual and pooled libraries, the individual libraries yielded several orders of magnitude fewer sites than the pooled libraries for both species (Table 2). A steep data reduction for individual libraries is expected as the average coverage was 10× per individual sequence and we applied a filter for 10× coverage.

Specifically, 97%–99% of the individual library data was filtered out at this step, while 36%–41% of the pooled library data was filtered out. This resulted in approximately 75,000 to 1,200,000 sites per individual library, while we obtained approximately 27–48 million sites per pooled library (Table 2). Coverage values are variable among individual libraries, but there were no large deviations from the mean to warrant exclusion in all but one site (Table S2).

We recovered 14,461 and 2410 regions common to 75% of the individuals for *C. fluminea and D. polymorpha*, respectively, representing a decrease of over 80% relative to the average number of regions available after filtering and tiling. In contrast, the number of regions for the pooled library data declined by 52% (868,103 and 1,180,209 for *C. fluminea* and *D. polymorpha*, respectively). The number of regions retrieved from the computationally pooled datasets were a marginal 7% loss (*C. fluminea*) and 10% gain (*D. polymorpha*) relative to the pre-united individual libraries.

The observed data loss is due to two major steps: (1) initial minimum coverage filtering per individual and (2) union step to find the regions common among individuals. This result is certainly a consequence of our sequencing strategy (i.e. equivalent sequencing effort per individual in pooled and individual libraries); however, this was done on purpose to compare the data of pooled and individual libraries obtained with the same resource investment in sequencing.

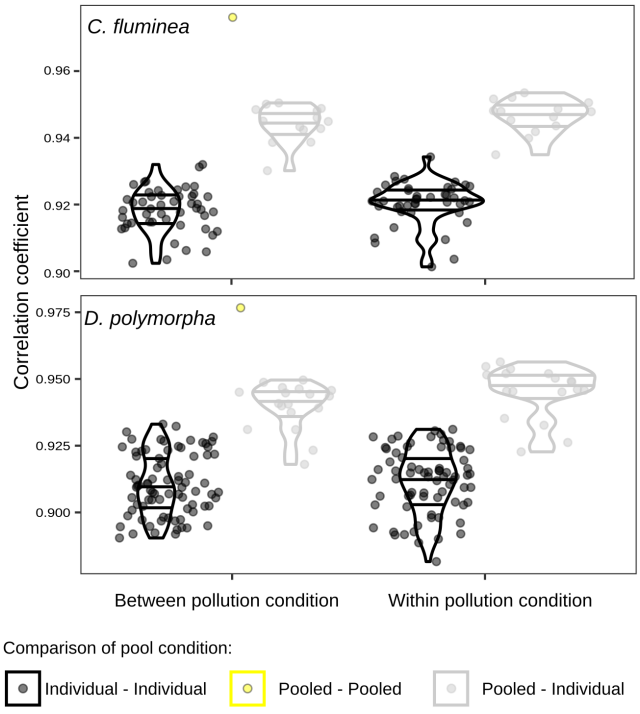

FIGURE 2 Scatter plots of the correlation coefficients for percent methylation between pollution sites. Pearson correlation coefficients are based on the per-region % methylation for each pair of libraries when using all samples.

3.4 | Evaluation of agreement between pooled and individual libraries

3.4.1 | Global DNA methylation and correlations of genome-wide methylation levels

C. fluminea had slightly lower global methylation levels compared to *D. polymorpha* (~15.48% vs. 19.67%; Table S2). These values are at the upper end of those reported for other molluscs (5%–15%; see Fallet et al., 2020). Differences in methylation were negligible between the polluted and non-polluted sites, and between the pooled and individual libraries for both species (*C. fluminea*, individual libraries, polluted 15.40 ± 0.74% vs. non-polluted 15.44 ± 0.57%, pooled libraries 15.5% vs. 15.6%; *D. polymorpha*, individual libraries polluted 19.79 ± 2.15% vs. non-polluted 19.58 ± 0.57%, pooled libraries 19.6% vs. 19.7%).

We examined the correlation of methylation percentage values between individual and pooled libraries to test for congruence between the two datasets that are expected to be equivalent. Correlations were slightly stronger for *C. fluminea* (.90–.99) compared to *D. polymorpha* (.88–.98; Figure 2), but overall similar trends were detected. The percent methylation values were positively correlated between individual and pooled libraries with the pairwise correlation coefficients not going below .92, irrespective of the pollution site or species. This is expected if pooled libraries demonstrate the same signal as the individual libraries and where data have not been centered (see Xu et al., 2015 for the importance of centring).

Unexpectedly, however, individual libraries correlated only slightly better with pooled libraries of the same site compared to correlations across sites (Figure 2). For both species, pooled libraries correlated best with each other (~.988). These results highlight that pollution has a weaker influence on correlation than the pooling method. For both species, the individual libraries from the pollution site had the lowest correlation coefficients for within site correlations (across both species: pollution .88–.93 vs. non-pollution .90–.93). This suggests that there may be a pollution-related response in methylation estimates influencing variation. Overall, we found that genome-wide methylation levels of individual and pooled libraries were well correlated following our expectations.

3.4.2 | Agreement among PCA jackknife iterations

To understand the similarity among samples, we tested for clustering using a principal component analysis on the percent methylation values. We measured heterogeneity of the signal across the genome with a standard error based on a delete-one jackknife. This measured changes to the PCA coordinates when removing a linkage group with each iteration of the PCA. For both species and nearly all PCA iterations, the Tucker's coefficient was greater than .90 indicating an overall agreement in signal between the global PCA and each iteration (see Figures S1). For both species, three of the jackknife iterations gave notably lower Tucker's coefficients and greater Procrustes D values than the remaining iterations, indicating a disproportionate influence from the respective linkage group (*C. fluminea* LG05, LG08, LG12; *D. polymorpha* NC_068364.1, NC_068365.1, NC_068370.1). The result does argue that some linkage groups may have notable divergences from the majority of the genome.

3.4.3 | PCA of genome-wide methylation levels

In the *C. fluminea* PCA (all libraries, 94,912 regions; Figure 3a), the primary eigenvector captured ~10% of the variance and the first 11 vectors accounted for the top 90% of the variance. There was no clear difference between pollution level and no clear differences in the variation within groups, with two outlying samples in the non-polluted group.

In the *D. polymorpha* PCA (all libraries, 190,879 regions; Figure 3c), the first eigenvector captured a significant part of the variance (27%), largely describing variation within the polluted population. The remaining vectors captured similar, but small, proportions of the variance (~8%–5%). Again we see no clear support for a directional or consistent difference in central tendency of either pollution groups (Figure 3d).

For both species the pooled libraries plotted central to all the individual libraries. This was confirmed by estimating the mean position of individual libraries on PC1 and PC2. With one exception, the mean positions were nearly equivalent to their pooled counterparts (For *D. polymorpha* NP: individual libraries −2.33, −0.56 vs. pooled libraries −1.62, 0.86; P: individual libraries 2.42, 0.55 vs. pooled libraries 5.50, 0.37. For *C. fluminea* NP: For individual libraries −3.99, −8.34 vs. pooled libraries −0.18, 7.68). Only the polluted site for *C. fluminea* had the mean position notably away from the observed pooled libraries (individual libraries 4.35, 9.99 vs. pooled libraries −2.36, −3.85).

3.4.4 | Comparison of differential methylation in individual versus pooled libraries

We used the pollution condition to assess if differential methylation estimates were similar between individual and pooled libraries. For both species, there were orders of magnitude more DMR in either wet-lab or computationally pooled libraries compared with individual libraries (Figure 4). This is not surprising given the

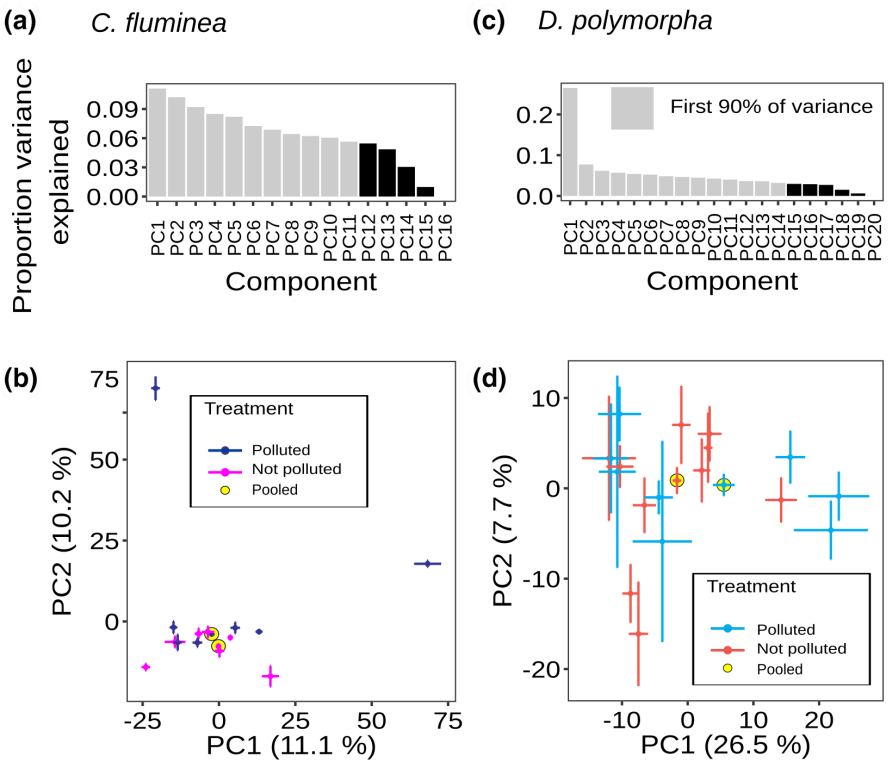

FIGURE 3 Principal component analysis of the genome-wide percent methylation. Panels (a) and (c) show the variance explained by each component for *C. fluminea* and *D. polymorpha*, respectively. Highlighted bars show the components that make up the top 90% of the variance. Panels (b) and (d) show the first two components with standard error bars based on the delete-one jackknife in *C. fluminea* and *D. polymorpha*, respectively. The percent variation explained by each axis is indicated on the axis label.

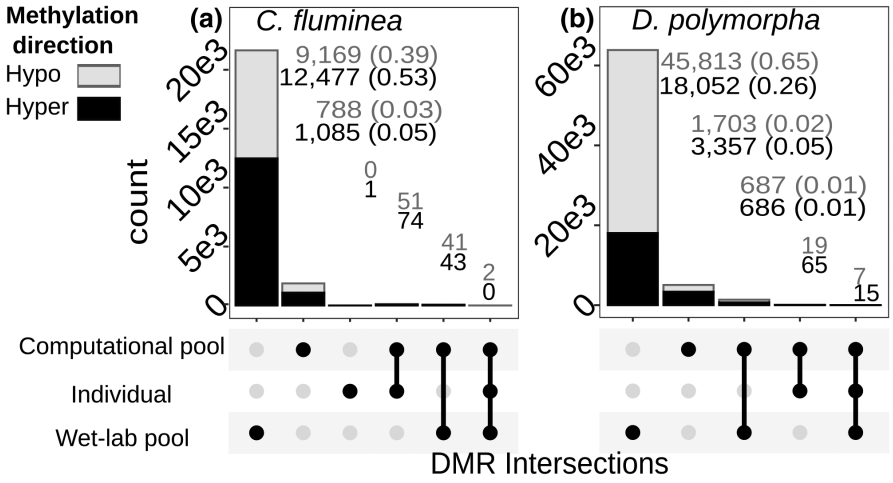

FIGURE 4 Intersection of the identified differentially methylated regions (DMRs) from contrast of the polluted and non-polluted localities for the individual, pooled and computationally pooled libraries for (a) C. fluminea and (b) D. polymorpha. The indivdiual, pooled and computationally pooled libraries are each a 'set' of DMRs as show by the rows at the bottom. The 'intersections' (columns) are the DMRs shared between sets. The dot-plot in the bottom shows how DMRs from each set are distributed among intersections. The central barplot shows the number of DMRs within a particular intersection. Numbers in the plot indicate the DMR counts which are either hyper- or hypo-methylated for each bar (proportion of total DMRs in brackets when >0.01). Font colours correspond to the legend.

large difference in input data. Both the pooled and computationally pooled library of *C. fluminea* produced more hypermethylated regions than hypomethylated regions (12,517 vs. 9201 for pooled and 1191 vs. 879 for computationally pooled). In contrast individual libraries produced less than 130 DMRs in total with more hypermethylated regions.

Hypomethylated regions outnumbered hypermethylated regions for *D. polymorpha* with the wet-lab pooled data (46,450 vs. 18,581, hypo- and hypermethylated, respectively). This was not the case with computationally pooled libraries (4066 vs. 2244, hyper- and hypo-methylated, respectively). Individual libraries similarly produced more hypermethylated regions (80 vs. 25) but again less than 130 regions in total (Figure 4).

Despite the large number of regions detected by the tests with the pooled and computationally pooled data, there was a relatively low overlap in the identified regions. The incongruency was most pronounced for *C. fluminea* where the overlapping DMRs amounted to <5% of the identified DMRs from the computationally pooled data. For *D. polymorpha*, there was a substantially greater proportion of shared DMRs between the computationally pooled and wet-lab pooled data, with 29% of the hypermethylated and 17% of the hypomethylated DMRs shared.

When put in the context of regions overlapping across datasets, we see that for both species the number of DMRs detected is directly proportional to the number of input regions (see Figure 4; Figure S2). Most regions in common between library datasets were also identified as DMRs for those datasets. For both species, only a single region shared between individual libraries and either wet-lab or computationally pooled data was not also identified as a DMRs in the respective datasets (*C. fluminea* one of 128 regions). For overlapping regions between wet-lab and computationally pooled data, more than half of the regions were identified as DMRs in both pools (~50% for *C. fluminea* and 71%–77% for *D. polymorpha*) and these proportions were consistent for hypomethylation and hypermethylation (see Figure 4; Figure S2).

There were no distinct shifts in the proportion of hypermethylated regions relative to hypomethylated regions between datasets for *C. fluminea* (ratio of hyper:hypo-methylated; wet-lab pooled 1.36, computationally pooled 1.35, individual libraries 1.41). In sharp contrast, the ratios changed notably for *D. polymorpha* (wet-lab pooled 0.40, computationally pooled 1.81, individual libraries 3.2). This likely reflects the general high methylation across the genome of invertebrates being better represented in the wet-lab pool of *D. polymorhpa* by virtue of the greater number of regions recovered. Overall this indicates that when regions are recovered across different datasets, the majority do have consistent outcomes.

3.5 | Methylation profiles by genetic context

We see that DMRs are over-represented in genic regions (exon + intron) while intergenic regions made up 18%–35% of the detected DMRs (Figure 5). Introns were the largest constituent of the genic regions (44%–52%) with exons making up 22%–27% of DMRs. The DMRs detected in the two library pooling schemes show overall consistency in the distribution among contexts for both hypermethylation and hypomethylation. The similarity in profile was also seen when considering sub-divisions in genic context where approximately equal proportions of regions overlapped with start codons and stop codons. The individually sequenced libraries produced too few DMRs to reliably profile their distribution across

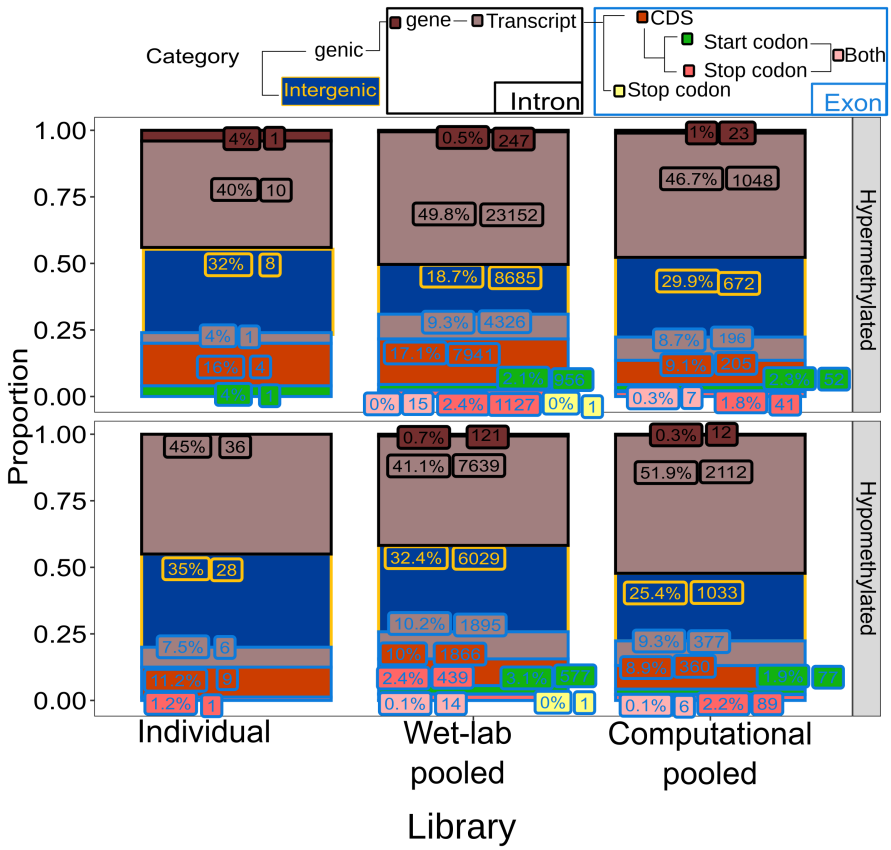

FIGURE 5 Distribution of regions used in the analyses, split by the three library preparation schemes and grouped according genic context. Results are shown for *Dreissena polymorpha* for the number of differentially methylated regions detected with hypermethylation (top panel) and hypomethylation (bottom panel) split by association with different genic contexts. Library preparation schemes are individual libraries (left), wet-lab pooled libraries (centre) and computationally pooled libraries (right). The number of regions and proportion of total regions for each sub-category are indicated on the plot. Colours correspond to the barplot and figure legend.

regions but do suggest an overall agreement with the computationally pooled data.

3.6 | Estimates of the required sampling effort and the recovered power

To assess the available power in our data, we estimated the proportion of regions responsive to pollution (Table S6) the level of power recovered with our current sampling effort (Table S7) and the sampling effort needed to achieve power at a level of 80% at each region (Table S8). We used region-specific estimates of effect size and mean difference in methylation (MDM) between sites (i.e. polluted vs. non-polluted).

A low proportion of regions (<25%) were deemed to be non-responsive to pollution (i.e. with MDM = 0) across all contrasts. The wet-lab pooled and computationally pooled libraries for both species produced more regions non-responsive to pollution than the individual libraries (Table S6). Proportionally, wet-lab pooled libraries produced >10× more than individual libraries and computationally pooled produced 4–5× more than wet-lab pooled.

For the regions responsive to pollution (i.e. with |MDM| >0), the sampling effort estimates were very large. For the individual libraries, estimates were predominantly more than 100 individuals per site (Figure 6a,d). Sample effort below 100 was only achieved for ~26% (669 regions)–35% (4803 regions) of the total regions considered, and no regions had sufficient power with our sampling effort.

Similarly, in the wet-lab and computationally pooled data, estimates of required sampling effort were >1000× and almost exclusively >100× coverage for both species (Figure 6b,c,e,f). For a set of 10 individuals this would be the equivalent of ~100× coverage each to achieve adequate coverage for less than 50% of the variance distribution. We recovered adequate power at only ~4%–15% of regions in the pooled and computationally pooled libraries, highlighting that the majority of our regions were under-powered. Computationally pooled libraries produced greater proportions of regions with coverage estimates below 100× compared to pooled libraries (Table S8).

The bias towards under-powered regions can be understood by the distribution of MDM values. Across all comparisons, regions with the lowest MDM had the lowest power and largest sampling effort, which is to be expected as larger differences require fewer samples. Nearly the entire MDM distribution was predominantly under-powered and had very small estimates of MDM. For example, with wet-lab and computationally pooled data, the lowest 50% of the MDM distribution (~1% difference in methylation) had no regions with sampling effort below 100× coverage. Such low MDM would be below the detectable resolution with 10× coverage used in our study (Figure 6a,d; Table S8). For the wet-lab and computationally pooled data, it was only regions in the top 10% of MDM distribution which would be detectable with a 10× coverage as used in this study (>10% difference in methylation)

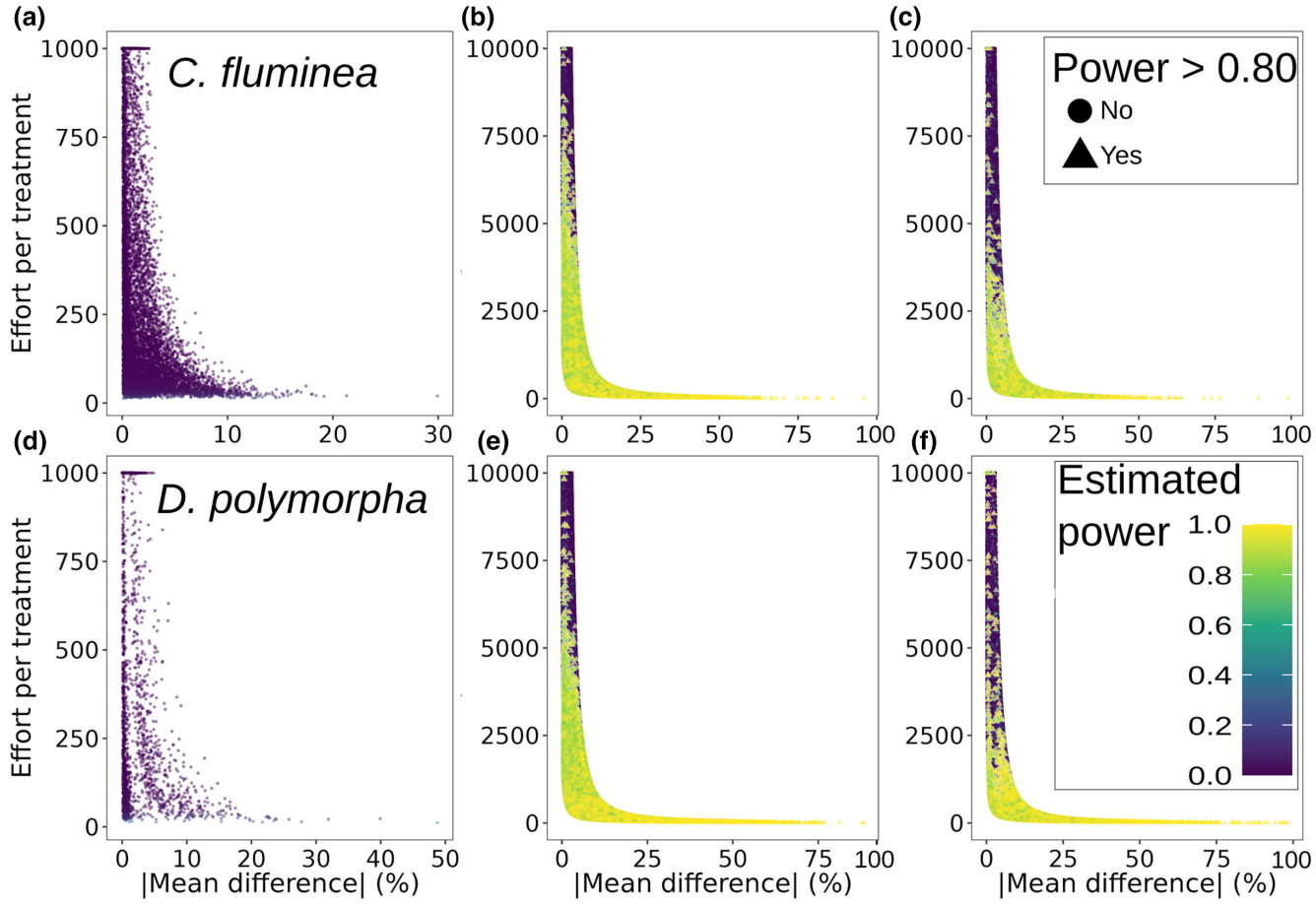

FIGURE 6 Estimates of the required per-region sampling effort and the achieved power. Panels (a)–(c) show the estimates for *C. fluminea*, and panels (d)–(f) show the estimates for *D. polymorpha* with separate plots for individual (panels a & d), wet-lab pooled (panels b & e) and computationally pooled libraries (panels c & f). Dots are individual loci and colours indicate the estimated power achieved. Note that in panels (a) and (d) regions with sampling effort estimates >1000 were capped to 1000 and in panels (b), (c), (e) and (f), regions with sampling effort estimates >10,000 were capped to 10,000. Sampling effort is measured as the number of biological replicates per site for individual libraries in panels (a) and (d), and sequencing coverage per pooled site for pooled libraries in panels (b), (c), (e) and (f). Regions with power >0.8 were plotted above other regions to aid visibility.

and only 10% of these regions had sampling effort estimates below 100× coverage (1%–3% of the total regions) (Table S8). The computationally pooled data of *D. polymorpha* was a notable exception here as the upper 25% of the regions reached the detection limit of 10× coverage making up as much as 8% of the total regions. Individual libraries could produce a greater proportion of regions with achievable sampling effort for lower MDM bins (Table S8). For example, the bottom 95% of the MDM distribution (MDM <9%) had ~28% of regions with achievable sampling effort estimates. However, we note that these differences would still be below the detection limit for each individual sample.

With regards to differences between the species, we see that *D. polymorpha* had a greater proportion of regions with achievable coverage, this being double the proportions for *C. fluminea* in the pooled data and computationally pooled data. For the individual libraries this proportion was only slightly lower than that of *C. fluminea* (28% vs. 33%). Similarly, a greater proportion of regions achieved sufficient power with our sampling for *D. polymorpha*.

Finally we also note the greater MDM for *D. polymorpha*, as much as 49% for individual libraries (vs. 30% *C. fluminea*). For the pooled and computationally pooled data, the two species had equal ranges for MDM, ~88%–100%. These differences may be attributed to sample sizes between the species, the library pooling and the possible differences in natural variation.

4 | DISCUSSION

4.1 | Individual and pooled libraries provide similar genome-wide methylation estimates

Here we examined if cost-effective pooled whole-epigenome libraries provide equivalent biological results to individually sequenced libraries. We found that pooled libraries produced a congruent epigenetic signature with individual libraries at the genome-wide level as seen with pairwise correlations, the PCA and the predominant

TABLE 3 Cost comparison of pooled and individual libraries.

	Individual	Nested pooled	Gross pooled
Pooling details			
Populations	12	12	12
Libraries	192	48	24
Samples per library	1	4	8
Total samples	192	192	192
Material/Protocol total			
NEBNext® Enzymatic Methyl-seq Kit[a]	9374	2604	1302
General consumables[b]	192	48	24
Library quantification[c]	192	48	24
Library quality control[c]	960	240	120
Total	10,718	2940	1470
	55.82	15.31	7.66
Cost ratio (Individual:Pooled)		3.6	7.3

Note: Estimated costs per specimen in a scenario of sampling 12 populations for a total of 192 biological replicates and no technical replicates. Costs are provided in Swiss Francs (CHF) with local 2023 prices. Costs are likely to vary among countries based on local factors. Prices include local taxes. Where costs are equal between pooled and individual libraries, we omitted such costs (i.e. individual DNA extraction, quality control, shearing, sequencing depth).

[a]1×24 reaction kit for pooled and 2×96 reaction kit for individual libraries.

[b]Pipette tips, general reagents, gloves, tubes etc.

[c]Tapestation D1000 screen tape, reagents and consumables.

congruence among DMRs but we note that the signal at specific regions were not necessarily congruent largely due to insufficient power and detection resolution. Our estimates of global methylation were also stable between pooled and individual libraries. These results are in line with previous research supporting a global or genome-wide correspondence of DNA methylation levels from pooled and individual libraries using different ways of measuring DNA methylation (Docherty et al., 2009, 2010; Gallego-Fabrega et al., 2015). In our data, both pooled and individual libraries showed a negligible difference in methylation between the polluted and non-polluted groups and for both species but in all cases DMRs were detected.

4.2 | Pooled libraries provide more data than individual libraries for an equivalent sequencing effort while reducing wet-lab costs

We found that there is at least a 7-fold decrease in the cost per sample for 'bulk pooled' libraries compared with individual libraries (Table 3). The reduction depends on the number of biological replicates in each pool and this allows a great degree of flexibility to balance sample sizes, coverage and cost. The pooled libraries also produced orders of magnitude more final regions (between 20 and 50× more than individual libraries) and lost notably less data throughout the workflow. There are a number of steps in the workflow at which data are lost. From our results we have pointed out a few noteworthy stages: lower alignment efficiencies compared to conventional WGS sequence alignment (30%–70% read pairs discarded), filtering for specific subset of sites from the sequence data, data discarded from the overlap between paired reads, deduplication and other post-alignment steps/biases (~80% reduction relative to the expected coverage), the removal of suspected SNPs, coverage filtering (~36%–41% loss for pooled libraries, over 98% loss for individual libraries), finding loci in common across samples (52% loss for pooled libraries, 81%–95% loss for individual libraries), filtering out sites/regions that are not responsive to pollution (1%–21% regions lost) and finally, the loss of sites/regions which have insufficient power to detect differences at the chosen MDM resolution (75%–95% of regions). This result is certainly in part a consequence of our sequencing strategy (i.e. equivalent sequencing effort per individual in pooled and individual libraries), however this was done deliberately to compare the data of pooled and individual libraries obtained with the same resource investment in sequencing. This result highlights the need for a particularly high coverage of individual WepiGS libraries (e.g. >15× initial coverage) to obtain a sufficient number of sites for final comparisons.

When estimating the required sampling efforts to reach sufficient power per region, pooled libraries produced as much as 90× more regions with an achievable sampling effort for regions with detectable MDM (103 vs. 9358 regions for *D. polymorpha*). Achievable is defined here as regions with estimates of <100 samples per site (individual libraries) or coverage estimates below 100× (pooled libraries). The achieved power was greater for the pooled libraries while the individual libraries for both species had no regions that reached the threshold of 0.8 power. All of this strongly supports that pooled libraries produce sufficient data more reliably compared with individual libraries given similar per-sample coverage of 10× on average and with a subsequent filter of at least 10 reads per methylation call and considering the resolution offered by the used coverage and filtering combination.

Low power is a concern even in well controlled studies, for example of clinical ADHD (van Dongen et al., 2019). Theoretically, reasonable power (80%, $p < .05$) can be achieved with small sample sizes (100 cases-control pairs) and even with small effect sizes (as low as 4.5%) for array-based work (Tsai & Bell, 2015), and empirically, 95% power can be achieved with arrays ($p < 10E\text{-}6$) with >43 pooled samples (Gallego-Fabrega et al., 2015). However, the work by Dongen et al. detected very few differentially methylated loci (<20 out of >400K; 0.005% of available loci) despite being able to directly or indirectly link loci to previous GWAS results, vouching for the accuracy of the signals detected. Their meta-analysis found no loci overlap between previous studies despite using the same genotyping array (see also Kaplow et al. (2015) where overlap between studies was less than 53%, or 8 of 15 loci). In WepiGS, specific loci and the distribution of sequencing effort across the genome cannot be guaranteed so we expect far lower power at any particular site.

In practice, our pooled libraries detected several thousand DMRs while the individual libraries detected <130 DMRs out of millions of initially available sites. These values are similar to that reported elsewhere (e.g. pooled data 598; Venkataraman et al., 2020). These differences will make a meaningful impact on the return of investment and the possible scope of downstream interpretation in research. We also employed computational pooling which pools the data of the individual libraries *post-hoc*. Here we found it to successfully mitigate much of the data loss that individual libraries suffered from during the QC process. Computationally pooled libraries produced smaller volumes of input data and DMRs relative to the wet-lab pooled.

4.3 | Pooled and computationally pooled libraries provided different DMRs in our dataset

Beyond the global signal, our results showed that genome-wide congruence does not necessarily imply corresponding DMR signals for pooled and individual libraries. We found reasonable agreement in DMRs between individual, wet-lab pooled and computationally pooled libraries for the regions which did overlap across datasets (50%–77%) but a substantial amount of shared regions were not detected across methods which is contrasting with the correlations and PCA.

These differences may arise from several factors: (1) the individuals compared between the pooled and computationally pooled libraries were not exactly the same as four *C. fluminea* and two *D. polymorpha* individuals were excluded from downstream analyses as they failed conversion rate quality control; (2) lack of normalization of individual data before computational pooling; (3) stochasticity in the library preparation and sequencing processes (e.g. differential PCR during library preparation; cryptic biases in sequencing among the specimens of the pooled libraries); and (4) differences in achieved power. Based on these results, we discuss below the benefits and drawbacks of sample pooling, as well as possible improvements and ways forward.

4.4 | Benefits and drawbacks of sample pooling and recommendations

There are important opportunities offered by WepiGS for ecological and evolutionary studies, and authors have made clear that it is crucial to optimally use resources and consider trade-offs before initiating a project (Laine et al., 2022). We emphasize that there is not a single optimal solution for all projects and that the decision to sequence pooled or individual libraries depends on the scientific question of a particular project and should be planned at very early stages. Here we put our work in context and provide a summary of the key benefits and drawbacks of pooling libraries for WepiGS, as well as their implications (summary in Table 4).

Starting with the benefits, we showed that pooled libraries can be up to seven-fold more cost effective than individual libraries, when comparing wet lab costs. These costs are likely to be a limiting factor into the future, given that sequencing costs are constantly decreasing. There is flexibility in cost adjustment when the pooling scheme varies (e.g. deciding how many pools to prepare) but there is presently limited research on the trade-offs of different degrees of pooling. Another important benefit of pooled libraries is that the number of individuals per pool can be increased; typically the number of individuals per investigated population is between 10 and 20, however it has been shown that a larger number of individuals is required to achieve sufficient power in natural populations (Lea et al., 2017; Tsai & Bell, 2015). For instance, when a predictor variable explains 15% of the difference between populations, 125 individuals per population are needed to reach 50% power (Lea et al., 2017). A third advantage was that a larger proportion of the sequencing data from pooled libraries can be used, resulting in many-fold increase in retrieved loci in our dataset. This was not only due to the higher coverage of pooled libraries, but also because there was a single union step using the pooled datasets (i.e. finding the loci in common among libraries). This union step typically leads to a large loss of data when using individual libraries. Thus, individual average coverage (10× in our study) should be substantially higher than the filtering threshold (10 reads in our study) if many biological replicates will be united at many loci. Ziller et al. (2014) have argued instead for increased sensitivity with more biological replicates and coverage rather than coverage alone, but this would increase the costs substantially. Finally, if a high sequencing coverage is not necessary (e.g. in our case 100× per pooled library), researchers can decide to lower the sequencing effort per pooled library, possibly decreasing even more the project costs.

Using pooled libraries has several drawbacks, though, the most important one being that there is no possibility of going back to the individual data. Hence, researchers should be extremely careful when thinking about the pooling design and make sure that every covariate that may impact the signal in the data has been taken into account (e.g. sampling locality, sampling time in the year, sex, age, tissue, experimental condition, etc.). If these covariate can be clearly identified and separated in sub-pools, then pooling the DNA of samples may be a good option to increase power and decrease costs. In contrast, if covariates cannot be identified or if the variability in the data are not known (e.g. first epigenomic assessment), we would recommend against pooling.

Individual libraries provide more flexibility and higher resolution as groups and comparisons can be done a posteriori (e.g. testing the impact of different covariates in a pilot study) and data can be reused for future projects (e.g. adding individuals from different populations or time points, or different comparisons can be made). Other drawbacks of using pooled libraries arise from the data analysis side. For instance, individual samples that failed cannot be excluded (e.g. low conversion rate; low amount of sequencing data), and thus equal conversion rates and sequencing depth for all individuals in a pool is assumed.

Furthermore, we observed that a large amount of computational resources was required to process the pooled datasets (e.g. alignment, methylation calling) and that the currently widely used bioinformatic tools have limited functionality/customizability (e.g. MethylKit running in R) and otherwise require users to develop workarounds for

TABLE 4 Benefits and drawbacks of DNA pooling before library preparation.

Topics	Implications	Libraries	
		Individual	Pooled
Costs	Higher wet lab costs for individual libraries. Cost savings can be adjusted according to the pooling scheme (see Table 3 for details)	−	+
Power	Increased number of individuals included in a pool improves accuracy of population-level metrics (e.g. Response to treatments, differences between environmental conditions) and increases power to detect differences	−	+
Power/costs	20–30-fold more data when pooled libraries are sequenced at an equivalent sequencing effort to the individual libraries (see Table 2 for details). Additional cost savings are possible if sequencing effort of pooled libraries is reduced	−	+
Flexibility	Individual information (covariates) cannot be used with pooling. Nested pooling (pooling by condition, e.g. sampling locality, sex, age, tissue, experimental condition) is needed to measure variability in the data. Data reuse for subsequent projects is challenging	+	−
Data analyses	Differences in individual conversion rates or individual sequencing depth are not taken into account when samples are pooled. Possible biased representation of some samples in the pool cannot be accounted for	+	−
	Greater computational resources needed resulting from greater data volumes of pooled libraries. Many tools for methylation analyses are not adapted to handle large datasets from pooled libraries	+	−
	Fewer, less flexible statistical tests are available for pooled datasets	+	−

many tasks. Moving data between tools to achieve piecemeal analyses is often time-consuming and discouraged (Laine et al., 2022). Researchers could decide to filter out invariant sites early in the data analysis, and/or decrease the sequencing depth of the pools (i.e. less than 10× per sample), however the minimum sequencing depth to obtain meaningful population methylation rates is not known. We note that this issue arises when study organisms have large genomes (in our case 1.6 and 1.8 Gb), but it may be less of a problem for organisms with smaller genome sizes (e.g. less than 1 Gb). Finally, we also noticed that there are fewer and less flexible statistical tests available for data analysis (e.g. Logistic regression cannot be used with two samples). A workaround to this problem may be to create subpools per condition (i.e. nested pooling) to still be able to use the logistic regression analysis. We hope that new tools that can handle large pooled epigenomic datasets will be developed in the future.

To conclude, individual libraries provide greater flexibility and control therefore they are the best option for a first epigenomic dataset where covariate variation is unknown, or when samples are rare or limited. However, there are situations in which pooling DNA before library preparation would be the best option for population-level signals (Futschik & Schlötterer, 2010; Kaplow et al., 2015), to increase power and decrease costs. For instance, when the number of individuals per population is not limited (e.g. abundant species), in well-studied systems where epigenomic variation is already characterized and researchers want to increase power in follow-up studies. In these systems, either a clear separation of covariates is possible or organisms are small and whole organisms are used for DNA extraction, making sure that all covariates are captured in a single DNA extraction (Harney et al., 2022). Finally, pooling would be particularly well-suited in systems with small genome sizes to facilitate downstream analyses.

4.5 | Possible improvements and ways forward

We have shown that pooled libraries provide estimates of genome-wide global methylation levels that are comparable to individual

libraries. However, signals of differential methylation at specific regions were not necessarily congruent, most likely as a result of large differences in the number of loci retrieved and the recovered power, which was a consequence of our sequencing effort strategy. Another possible cause was that six individuals were excluded from the analyses due to low conversion rates and low sequencing data, resulting in true differences between the pooled and computationally pooled datasets. In addition, stochasticity in the library preparation and sequencing processes may have led to further discrepancies between these datasets.

Genetic variation is a source of confounding effects with methylation variation (Venney et al., 2023). For example, in *Ostrea lurida* (oyster) as much as 27% of methylation can be explained by inter-individual genomic variation (Silliman et al., 2023). While we removed likely SNPs prior to the DMR analysis, some detected DMRs may arise from C/T polymorphisms not detected or we may have false positives when the reference is divergent form the study populations. However, these polymorphisms would impact equally pooled and individual datasets.

These epigenomic datasets were the first ones for the two species of interest, *C. fluminea* and *D. polymorpha*. Therefore, global DNA methylation levels were previously unknown, as well as the level of covariate variation. Based on these results, we would recommend to perform a pilot study using individual libraries to assess these metrics and make an informed decision on whether or not to pool a large number of individuals in subsequent studies before considering pooling. Furthermore, we acknowledge that we did not perform simulations in this study, because we wanted to focus on empirical data to explore commonalities and differences between individual and pooled libraries produced in the lab. Thus, we aimed to obtain a very practical result close to a real experiment. In future studies, it would be interesting to simulate the minimum coverage of a pool required to obtain reliable population-level DNA methylation rates, as a way of facilitating downstream analyses and further decreasing project costs. To conclude, our study brings important insights on the relevance of pooling DNA of individuals before library preparation in epigenomic studies of natural populations, and we believe that it will help researchers in making informed decisions for future epigenomic projects.

AUTHOR CONTRIBUTIONS

RJD: Analyses and writing; BSM: Supervision and writing; SS and MG: Laboratory work and writing; CDT: Design and sample contribution; NR: Design and sample contribution; SS: Design and sample contribution; AATW: Design, supervision, writing and funding.

ACKNOWLEDGEMENTS

The project was funded by the Swiss Federal Institute of Aquatic Science and Technology (Eawag). We also thank Niklaus Zemp for bioinformatic support. Sequencing was performed at the Functional Genomics Center Zürich (FGCZ), Switzerland. Data analyses were performed at the Euler High—Performance Computer Centre in Zürich, Switzerland (https://sis.id.ethz.ch/services/hpc/). Data produced and analysed in this paper were generated in collaboration with the Genetic Diversity Centre (GDC), ETH Zürich. Finally, we would like to thank V. Laine and two anonymous reviewers who provided comments to improve previous versions of this manuscript. RJD was supported by the National Research Foundation of South Africa. SS received funds from Short Term Scientific Missions provided by the COST Action "Conservation of freshwater mussels: a pan-European approach" (CONFREMU) CA1823. Open access funding provided by ETH-Bereich Forschungsanstalten.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interests.

DATA AVAILABILITY STATEMENT

Raw sequence reads are deposited in the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) (BioProject: PRJNA1019700). Custom code used in this manuscript is available on a public Gitlab archive (https://gitlab.com/RJDaniels/public-code-methylseq-pooling-2023.git).

BENEFITS GENERATED

Benefits from this research accrue from the sharing of our data and results on public databases as described above.

ORCID

Ryan J. Daniels https://orcid.org/0000-0003-4985-3541
Britta S. Meyer https://orcid.org/0000-0002-2549-1825
Alexandra A.-T. Weber https://orcid.org/0000-0002-7980-388X

REFERENCES

Ahlmann-Eltze, C. (2020). *ggupset: Combination matrix axis for 'ggplot2' to create 'UpSet' plots*. https://CRAN.R-project.org/package=ggupset

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., & Mason, C. E. (2012). methylKit: A comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, *13*, R87.

Andrews, S. (2019). *FastQC: A quality control tool for high throughput sequence data*. https://github.com/s-andrews/FastQC/releases/tag/v0.11.9

Ardura, A., Clusa, L., Zaiko, A., Garcia-Vazquez, E., & Miralles, L. (2018). Stress related epigenetic changes may explain opportunistic success in biological invasions in antipode mussels. *Scientific Reports*, *8*(1), 10793.

Assefa, A. T., Vandesompele, J., & Thas, O. (2020). On the utility of RNA sample pooling to optimize cost and statistical power in RNA sequencing experiments. *BMC Genomics*, *21*, 384.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, *160*, 268–282.

Blake, L. E., Roux, J., Hernando-Herraez, I., Banovich, N. E., Perez, R. G., Hsiao, C. J., Eres, I., Cuevas, C., Marques-Bonet, T., & Gilad, Y. (2020). A comparison of gene expression and DNA methylation patterns across tissues and species. *Genome Research*, *30*, 250–262.

Brander, S. M., Biales, A. D., & Connon, R. E. (2017). The role of epigenomics in aquatic toxicology. *Environmental Toxicology and Chemistry*, *36*, 2565–2573.

Busing, F. M. T. A., Meijer, E., & van der Leeden, R. (1999). Delete-m jackknife for unequal m. *Statistics and Computing*, *9*, 3–8.

Champely, S. (2018). *pwr: Basic functions for power analysis*. https://CRAN.R-project.org/package=pwr

CIPAIS. (1999). *Ricerche sulla distribuzione e gli effetti del DDT nell'ecosistema Lago Maggiore*. Rapporto finale sui risultati delle indagini. Ed. Commissione internazionale per la protezione acque italo-svizzere: 85.

CIPAIS. (2021). *Indagine sulle sostanze pericolose nell'ecosistema del Lago Maggiore. Programma triennale 2019–2021*. Campagna 2020. Ed. Commissione Internazionale per la protezione delle acque italo-svizzere: 115.

CIPAIS. (2022). Ricerche sulla distribuzione e gli effetti del DDT nell'ecosistema Lago Maggiore. Programma 2019–2021. Rapporto finale sui risultati delle indagini. Ed. Commissione internazionale per la protezione acque italo-svizzere: 193.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cramer, D., & Howitt, D. (2004). *The Sage dictionary of statistics* (1st ed.). SAGE.

Docherty, S. J., Davis, O. S. P., Haworth, C. M. A., Plomin, R., & Mill, J. (2009). Bisulfite-based epityping on pooled genomic DNA provides an accurate estimate of average group DNA methylation. *Epigenetics & Chromatin*, *2*, 255–258.

Docherty, S. J., Davis, O. S. P., Haworth, C. M. A., Plomin, R., & Mill, J. (2010). DNA methylation profiling using bisulfite-based epityping of pooled genomic DNA. *Methods*, *52*, 255–258.

Everitt, B., & Hothorn, T. (2010). *A handbook of statistical analyses using R* (2nd ed.). CRC Press, Taylor & Francis.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*, 3047–3048.

Fallet, M., Luquet, E., David, P., & Cosseau, C. (2020). Epigenetic inheritance and intergenerational effects in mollusks. *Gene*, *729*, 144166.

Fisher, R. A. (1934). Statistical methods for research workers. In F. A. E. Crew & D. W. Cutler (Eds.), *Biological monographs and manuals* (Vol. 5, pp. 1–317). Edingburgh, UK.

Futschik, A., & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, *186*, 207–218.

Gallego-Fabrega, C., Carrera, C., Muiño, E., Montaner, J., Krupinski, J., & Fernandez-Cadenas, I. (2015). DNA methylation levels are highly correlated between pooled samples and averaged values when analysed using the Infinium HumanMethylation 450 BeadChip array. *Clinical Epigenetics*, *7*, 78.

Gao, S., Zou, D., Mao, L., Liu, H., Song, P., Chen, Y., Zhao, S., Gao, C., Li, X., Gao, Z., Fang, X., Yang, H., Ørntoft, T. F., Sørensen, K. D., & Bolund, L. (2015). BS-SNPer: SNP calling in bisulfite-seq data. *Bioinformatics*, *31*, 4006–4008.

Guzzella, L. M., Novati, S., Casatta, N., Roscioli, C., Valsecchi, L., Binelli, A., Parolini, M., Solcà, N., Bettinetti, R., Manca, M., Mazzoni, M., Piscia, R., Volta, P., Marchetto, A., Lami, A., & Marziali, L. (2018). Spatial and temporal trends of target organic and inorganic micropollutants in Lake Maggiore and Lake Lugano (Italian-Swiss water bodies): Contamination in sediments and biota. *Hydrobiologia*, *824*, 271–290.

Han, F., Jamsandekar, M., Pettersson, M. E., Su, L., Fuentes-Pardo, A., Davis, B., Bekkevold, D., Berg, F., Casini, M., Dahle, G., Farrell, E. D., Folkvord, A., & Andersso, L. (2020). Ecological adaptation in Atlantic herring is associated with large shifts in allele frequencies at hundreds of loci. *eLife*, *9*, e61076.

Harney, E., Paterson, S., Collin, H., Chan, B. H. K., Bennett, D., & Plaistow, S. J. (2022). Pollution induces epigenetic effects that are stably transmitted across multiple generations. *Evolution Letters*, *6*, 118–135.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*, 115–121.

Jeremias, G., Gonçalves, F. J. M., Pereira, J. L., & Asselman, J. (2020). Prospects for incorporation of epigenetic biomarkers in human health and environmental risk assessment of chemicals. *Biological Reviews*, *95*, 822–846.

Jobling, M., Hollox, E., Hurles, M., Kivisild, T., & Tyler-Smith, C. (2014). *Human evolutionary genetics* (2nd ed.). Garland Science, Taylor & Francis Group.

Kaplow, I. M., MacIsaac, J. L., Mah, S. M., McEwen, L. M., Kobor, M. S., & Fraser, H. B. (2015). A pooling-based approach to mapping genetic variants associated with DNA methylation. *Genome Research*, *25*, 907–917.

Konczal, M., Koteja, P., Stuglik, M. T., Radwan, J., & Babik, W. (2013). Accuracy of allele frequency estimation using pooled RNA-Seq. *Molecular Ecology Resources*, *14*, 381–392.

Krueger, F. (2020). *Trim Galore*. https://github.com/FelixKrueger/TrimGalore/releases/tag/0.6.6

Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, *27*, 1571–1572.

Laine, V. N., Sepers, B., Lindner, M., Gawehns, F., Ruuskanen, S., & van Oers, K. (2022). An ecologist's guide for studying DNA methylation variation in wild vertebrates. *Molecular Ecology Resources*, *23*, 1488–1508.

Lamka, G. F., Harder, A. M., Sundaram, M., Schwartz, T. S., Christie, M. R., DeWoody, J. A., & Willoughby, J. R. (2022). Epigenetics in ecology, evolution, and conservation. *Frontiers in Ecology and Evolution*, *10*. https://doi.org/10.3389/fevo.2022.871791

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, *9*, 357–359.

Lea, A. J., Vilgalys, T. P., Durst, P. A. P., & Tung, J. (2017). Maximizing ecological and evolutionary insight in bisulfite sequencing data sets. *Nature Ecology & Evolution*, *1*, 1074–1083.

Lee, I., Rasoul, B. A., Holub, A. S., Lejeune, A., Enke, R. A., & Timp, W. (2017). Whole genome DNA methylation sequencing of the chicken retina, cornea and brain. *Scientific Data*, *4*, 170148.

Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). TuckerCongruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*, 57–64.

Mansell, G., Gorrie-Stone, T. J., Bao, Y., Kumari, M., Schalkwyk, L. S., Mill, J., & Hannon, E. (2019). Guidance for DNA methylation studies: Statistical insights from the Illumina EPIC array. *BMC Genomics*, *20*, 366.

Marin, P., Genitoni, J., Barloy, D., Maury, S., Gibert, P., Ghalambor, C. K., & Vieira, C. (2019). Biological invasion: The influence of the hidden side of the (epi)genome. *Functional Ecology*, *34*, 385–400.

Marziali, L., Guzzella, L., Salerno, F., Marchetto, A., Valsecchi, L., Tasselli, S., Roscioli, C., & Schiavon, A. (2021). Twenty-year sediment contamination trends in some tributaries of Lake Maggiore (northern Italy): Relation with anthropogenic factors. *Environmental Science and Pollution Research International*, *28*, 38193–38208.

McCartney, M. A., Auch, B., Kono, T., Mallez, S., Zhang, Y., Obille, A., Becker, A., Abrahante, J. E., Garbe, J., Badalamenti, J. P., Herman, A., Mangelson, H., Liachko, I., Sullivan, S., Sone, E. D., Koren, S., Silverstein, K. A. T., Beckman, K. B., & Gohl, D. M. (2022). The genome of the zebra mussel, *Dreissena polymorpha*: A resource for comparative genomics, invasion genetics, and biocontrol. *G3: Genes, Genomes, Genetics*, *12*, jkab423.

Montinaro, F., Busby, G. J. J., Pascali, V. L., Myers, S., Hellenthal, G., & Capelli, C. (2015). Unravelling the hidden ancestry of American admixed populations. *Nature Communications*, *6*, 6596.

Mounger, J., Ainouche, M. L., Bossdorf, O., Cavé-Radet, A., Li, B., Parepa, M., Salmon, A., Yang, J., & Richards, C. L. (2021). Epigenetics and

the success of invasive plants. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *376*, 20200117.

Ozerov, M., Vasemägi, A., Wennevik, V., Niemelä, E., Prusov, S., & Kent, M. (2013). Cost-effective genome-wide estimation of allele frequencies from pooled DNA in Atlantic salmon (*Salmo salar* L.). *BMC Genomics*, *14*, 12.

Paro, R., Grossniklaus, U., Santoro, R., & Wutz, A. (2021). *Introduction to epigenetic*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-6867-3

Parolini, M., Pedriali, A., & Binelli, A. (2013). Chemical and biomarker responses for site-specific quality assessment of the Lake Maggiore (northern Italy). *Environmental Science and Pollution Research International*, *20*, 5545–5557.

Peres-Neto, P. R., & Jackson, D. A. (2001). How well do multivariate data sets match? The advantages of a procrustean superimposition approach over the Mantel test. *Oecologia*, *129*, 169–178.

Riva, C., Parolini, M., Binelli, A., & Provini, A. (2010). The case of pollution of Lake Maggiore: A twelve-year study with the bioindicator mussel *Dreissena polymorpha*. *Water, Air, and Soil Pollution*, *210*, 75–86.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611.

Silliman, K., Spencer, L. H., White, S. J., & Roberts, S. B. (2023). Epigenetic and genetic population structure is coupled in a marine invertebrate. *Genome Biology and Evolution*, *15*, 1–18.

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

Teschendorff, A. E., Breeze, C. E., Zheng, S. C., & Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-wide association studies. *BMC Bioinformatics*, *18*, 105.

Tolley, K. A., Hopkins, K. P., & da Silva, J. M. (2019). Genetic structure associated with habitat diversification supports the independent evolution of ecomorphs *Bradypodion pumilum*. *African Journal of Herpetology*, *68*, 77–89.

Tsai, P.-C., & Bell, J. T. (2015). Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *International Journal of Epidemiology*, *44*, 1429–1441.

van Dongen, J., Zilhão, N. R., Sugden, K., Hannon, E. J., Mill, J., Caspi, A., Agnew-Blais, J., Arseneault, L., Corcoran, D. L., Moffitt, T. E., Poulton, R., Franke, B., Boomsma, D. I., Heijmans, B. T., 't Hoen, P. A. C., van Meurs, J., Isaacs, A., Jansen, R., Franke, L., … Heijmans, B. T. (2019). Epigenome-wide association study of attention-deficit/hyperactivity disorder symptoms in adults. *Biological Psychiatry*, *86*, 599–607.

Venkataraman, Y. R., Downey-Wall, A. M., Ries, J., Westfield, I., White, S. J., Roberts, S. B., & Lotterhos, K. E. (2020). General DNA Methylation Patterns and Environmentally-Induced Differential Methylation in the Eastern Oyster (*Crassostrea virginica*). *Frontiers in Marine Science*, *7*. https://doi.org/10.3389/fmars.2020.00225

Venney, C. J., Anastasiadi, D., Wellenreuther, M., & Bernatchez, L. (2023). The evolutionary complexities of DNA methylation in animals: From plasticity to genetic evolution. *Genome Biology and Evolution*, *15*, evad216.

Wang, H.-Q., Tuominen, L. K., & Tsai, C.-J. (2011). SLIM: A sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics*, *27*, 225–231.

Weber, A. A.-T., Dupont, S., & Chenuil, A. (2013). Thermotolerance and regeneration in the brittle star species complex *Ophioderma longicauda*: A preliminary study comparing lineages and Mediterranean basins. *Comptes Rendus Biologies*, *336*, 572–581.

Xu, Z., Niu, L., Li, L., & Taylor, J. A. (2015). ENmix: A novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Research*, *44*, e20.

Zhang, T., Yin, J., Tang, S., Li, D., Gu, X., Zhang, S., Suo, W., Liu, X., Liu, Y., Jiang, Q., Zhao, M., Yin, Y., & Pan, J. (2021). Dissecting the chromosome-level genome of the Asian clam (*Corbicula fluminea*). *Scientific Reports*, *11*, 15021.

Ziller, M. J., Hansen, K. D., Meissner, A., & Aryee, M. J. (2014). Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nature Methods*, *12*, 230–232.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Daniels, R. J., Meyer, B. S., Giulio, M., Signorini, S. G., Riccardi, N., Della Torre, C., & Weber, A.-T. (2024). Benchmarking sample pooling for epigenomics of natural populations. *Molecular Ecology Resources*, *00*, e14021. https://doi.org/10.1111/1755-0998.14021