

New *Phytologist* Supporting Information

Article title: DNA methylation in the wild: epigenetic transgenerational inheritance can mediate adaptation in clones of wild strawberry (*Fragaria vesca*)

Authors: Iris Sammarco, Bárbara Díez Rodríguez, Dario Galanti, Adam Nunn, Claude Becker, Oliver Bossdorf, Zuzana Münzbergová, Vít Latzel

Article acceptance date: 20 November 2023

The following Supporting Information is available for this article:

Fig. S1 Experimental design of the study.

Fig. S2 Principal component analysis for genetic variants, gene expression and DNA methylation coloured by mean temperature and precipitation.

Fig. S3 Boxplots showing methylation distribution across different genomic features for field and garden conditions.

Fig. S4 Total numbers of differentially methylated regions identified in different genomic contexts.

Fig. S5 Methylation level of differentially methylated regions in the garden, overlapping with genes and transposable elements.

Fig. S6 Distribution of the strongest differentially methylated region predictor in the garden.

Fig. S7 Number of *cis*-, *trans*-, climate-predicted and unexplained differentially methylated regions overlapping promoters, gene bodies or transposable elements.

Fig. S8 Dot plots for genome-wide association analyses for climate-predicted differentially methylated regions overlapping promoters.

Fig. S9 Circos plots showing correlations identified through mixOmics analyses between CG, CHG and CHH climate-associated DMRs and gene expression.

Methods S1 WGBS library preparation and sequencing.

Methods S2 Methylation and DMR calling.

Methods S3 DMR variance decomposition analysis.

Methods S4 Genome-wide association analysis.

Methods S5 RNA-sequencing.

See separate files:

Table S1 Origin sites and common garden characteristics of *Fragaria vesca* populations used in this study.

Table S2 Mapping statistics for WGBS analysis, SNP calling and RNA-seq.

Table S3 Correlation between number of *cis*-, *trans*-, climate-, unexplained-DMRs and number of genes and TEs.

Table S4 Enrichment ratio analysis of *trans*-predicted DMRs in CHG and CHH contexts, and unexplained DMRs in the CHH context, across different TE superfamilies.

Table S5 Gene Ontology enrichment analysis of predicted DMRs in *cis*, *trans*, climate-predicted and unexplained DMRs.

Table S6 List of climate-predicted DMRs found in the garden condition with a statistically significant correlation between promoter methylation and expression of the adjacent gene.

Table S7 List of correlations identified through mixOmics analyses between climate-predicted DMRs and gene expression.

Table S8 List of overlapping differentially expressed genes and environmentally linked DMRs that had significant correlations with gene expression.

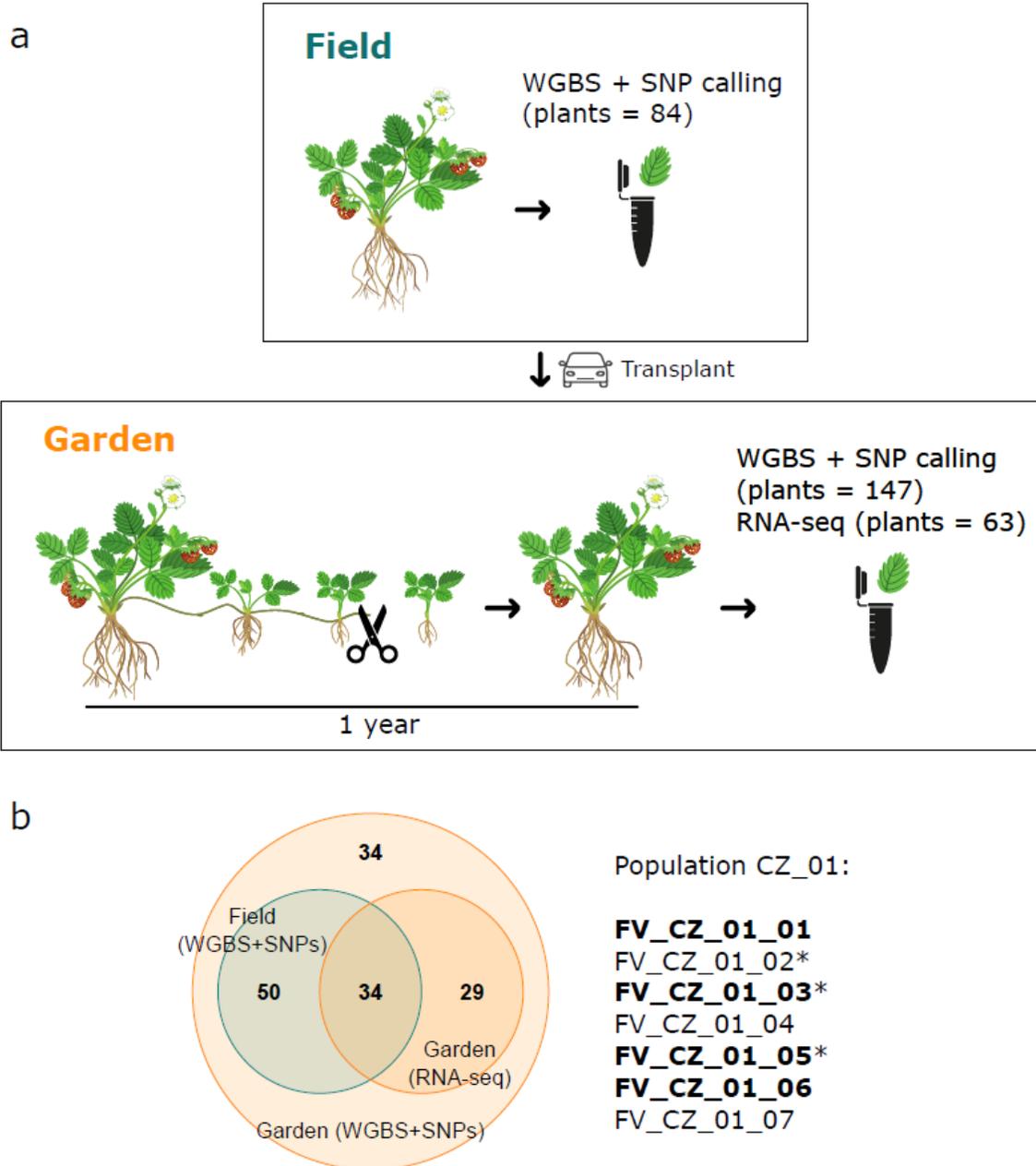


Figure S1: Experimental design of the study. (a) We performed whole genome bisulfite sequencing (WGBS) and inferred single nucleotide polymorphisms (SNPs) from these data from leaf tissue collected from 84 plants belonging to 21 populations of *Fragaria vesca* grown in their wild conditions (field, top) (see Fig. 1a for the sampling locations). We transplanted 147 plants

(7 plants per population) in a common garden (bottom), where we clonally propagated them for one year. For all the plants, we performed WGBS and inferred SNPs from leaf tissue of adult clonal offspring of at least the third generation, and RNA-sequencing for 63 individuals (3 plants per population). **(b)** The Venn diagram shows the overlap between the number of field (blue) and garden (orange) samples used for WGBS and SNP calling, and the garden samples used for RNA-sequencing. The provided sample list (right) represents population CZ_01 (used as an example). Bold entries denote samples with both field and garden WGBS and SNP data, non-bold entries represent samples with only garden WGBS and SNP data, and entries marked with asterisks (*) indicate samples used for RNA-sequencing.

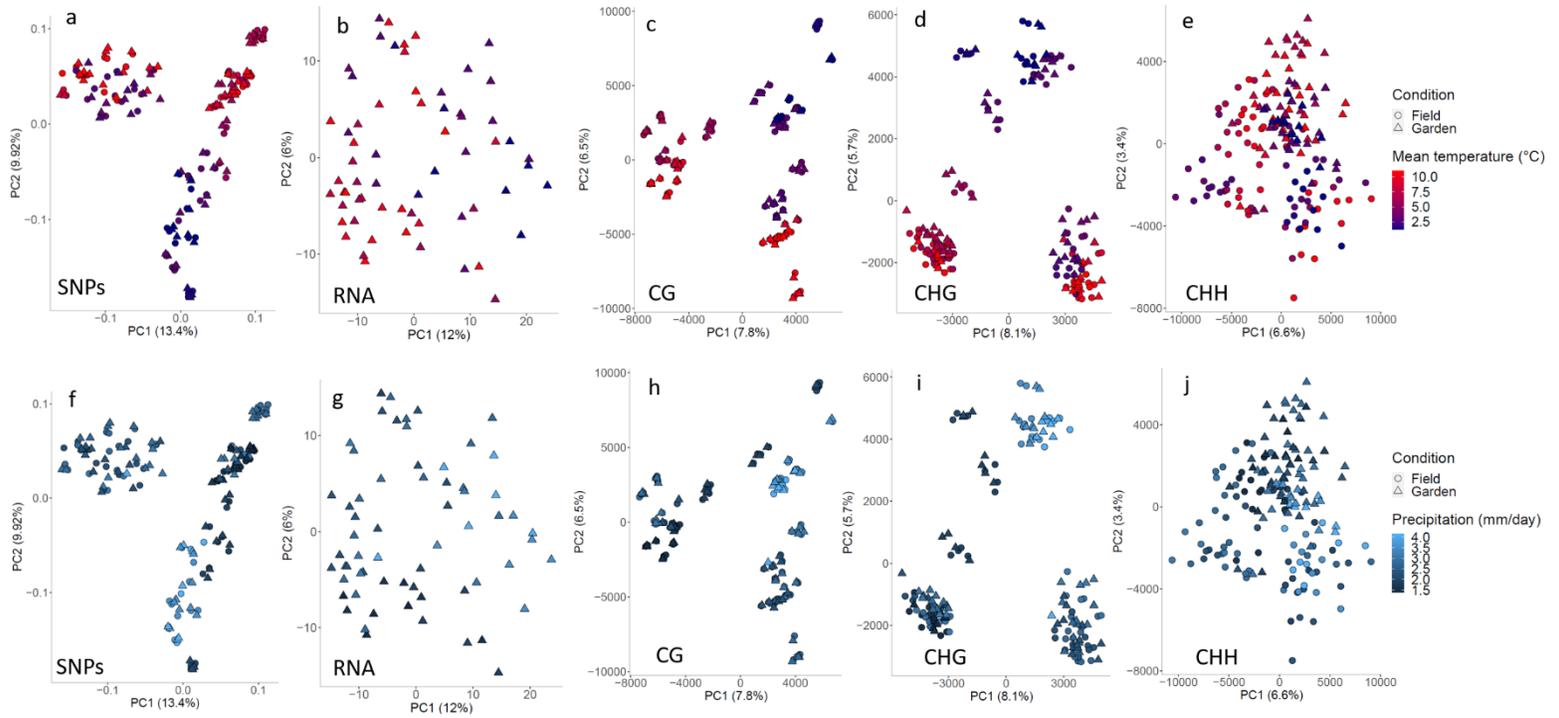


Figure S2: Principal component analysis (PCA) for genetic variants (SNPs), gene expression (RNA) and DNA methylation colored by mean temperature (a-e) and precipitation (f-j). SNPs (a, f), gene expression (b, g), and CG (c, h), CHG (d, i) and CHH (e, j) methylated positions. For SNPs and methylated positions, field plants: n = 84, garden plants: n = 84. For gene expression, garden plants: n = 63.

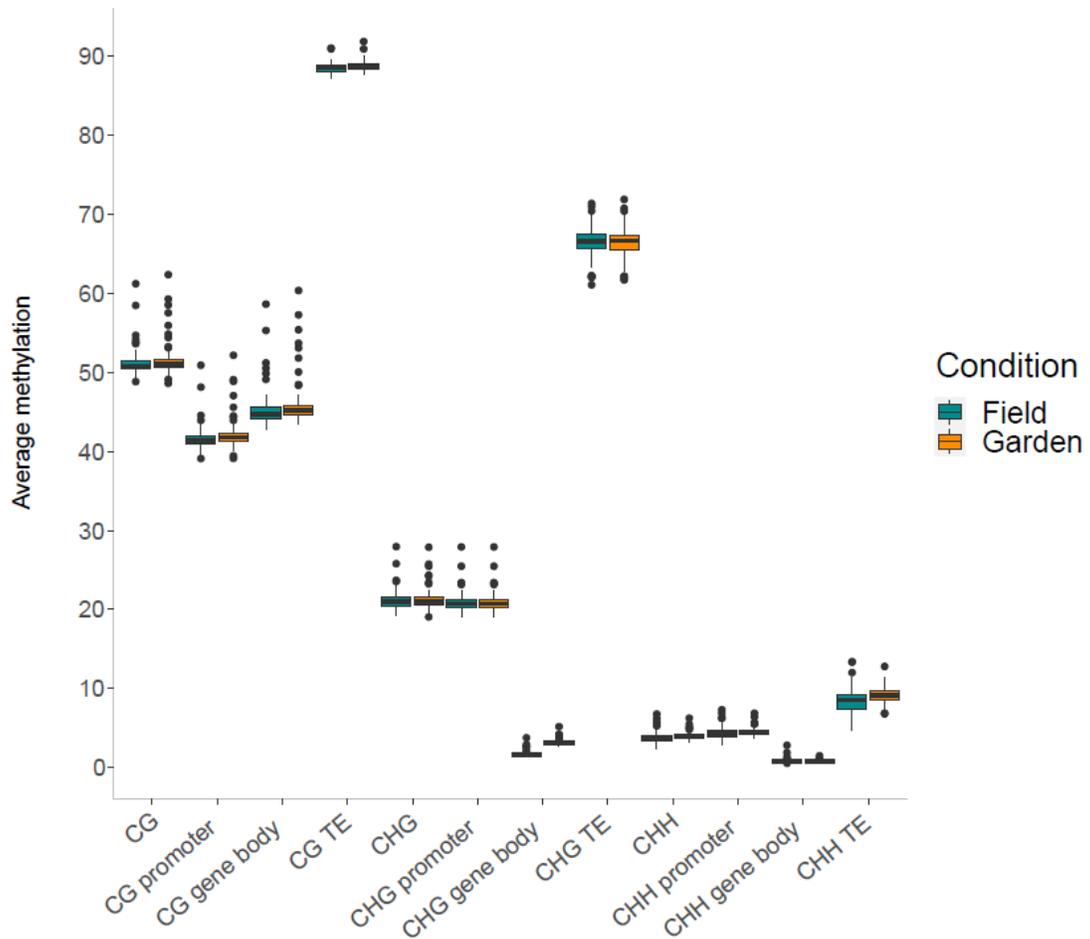


Figure S3: Boxplots showing methylation distribution across different genomic features for field and garden conditions. CG, CHG and CHH: genome-wide methylation levels in the three sequence contexts. TE: transposable elements.

Within each boxplot, the central line represents the median, while the box spans the interquartile range (IQR) from the first quartile (Q1) to the third quartile (Q3). Whiskers extend from the box to indicate variability outside the IQR. Any points beyond the whiskers' range are depicted as individual dots and considered outliers.

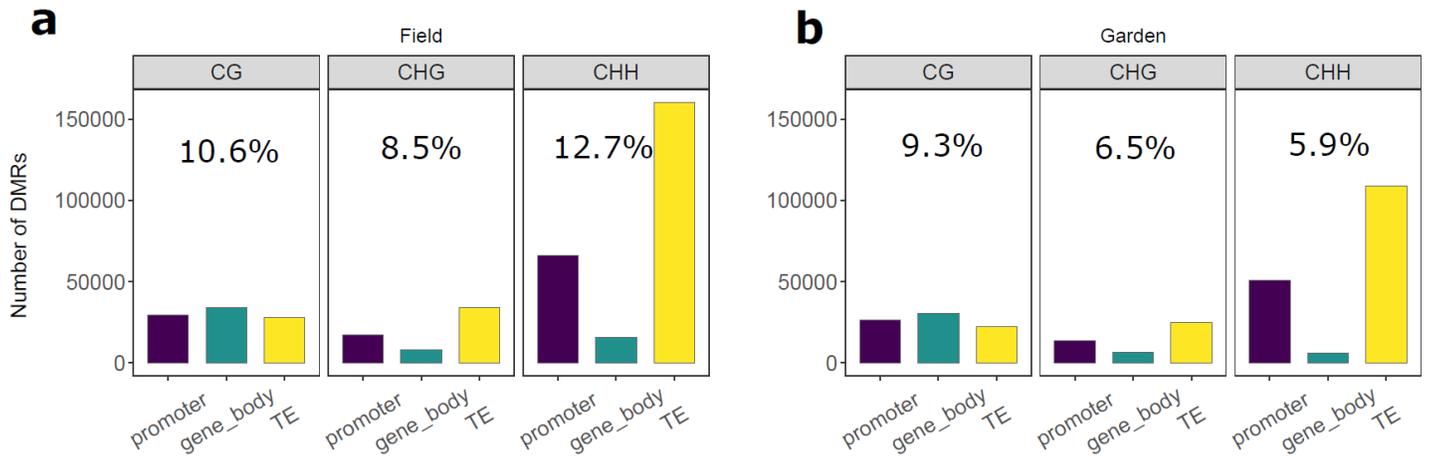


Figure S4: Total numbers of differentially methylated regions (DMRs) identified in different genomic contexts. (a) Plants in the field (n = 84). (b) Plants in the garden (n = 84). Note that a single DMR can overlap with both the promoter and gene body of the same gene. The number in each panel shows the frequencies of DMRs in the different sequence contexts. TE: transposable elements.

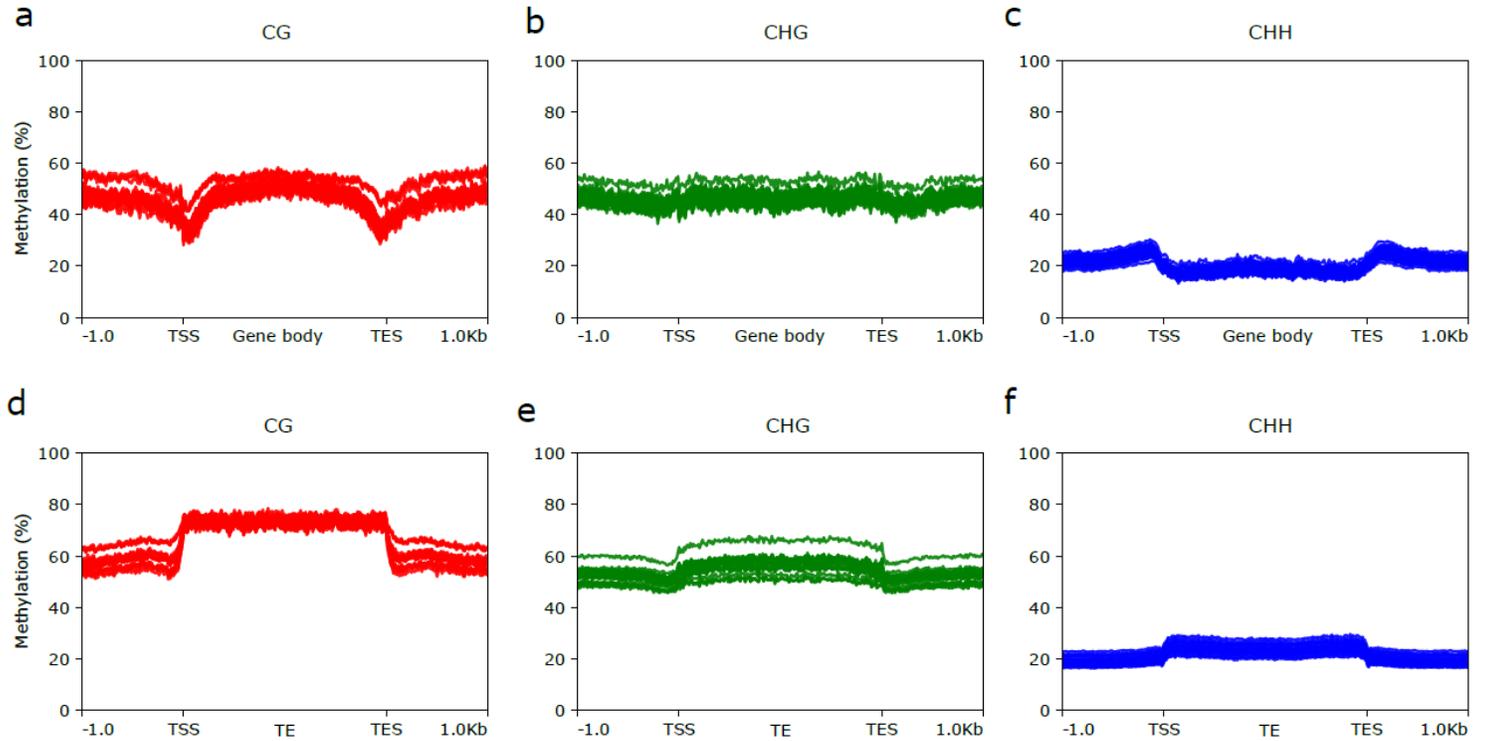


Figure S5: Methylation level of differentially methylated regions (DMRs) in the garden, overlapping with genes and transposable elements (TEs). (a) CG-DMRs, (b) CHG-DMRs and (c) CHH-DMRs overlapping with genes and their 1 kb-long upstream and downstream sequences. (d) CG-DMRs, (e) CHG-DMRs and (f) CHH-DMRs overlapping with TEs and their 1 kb-long upstream and downstream sequences. TSS: transcription start site; TES: transcription end site. Number of plants: 84.

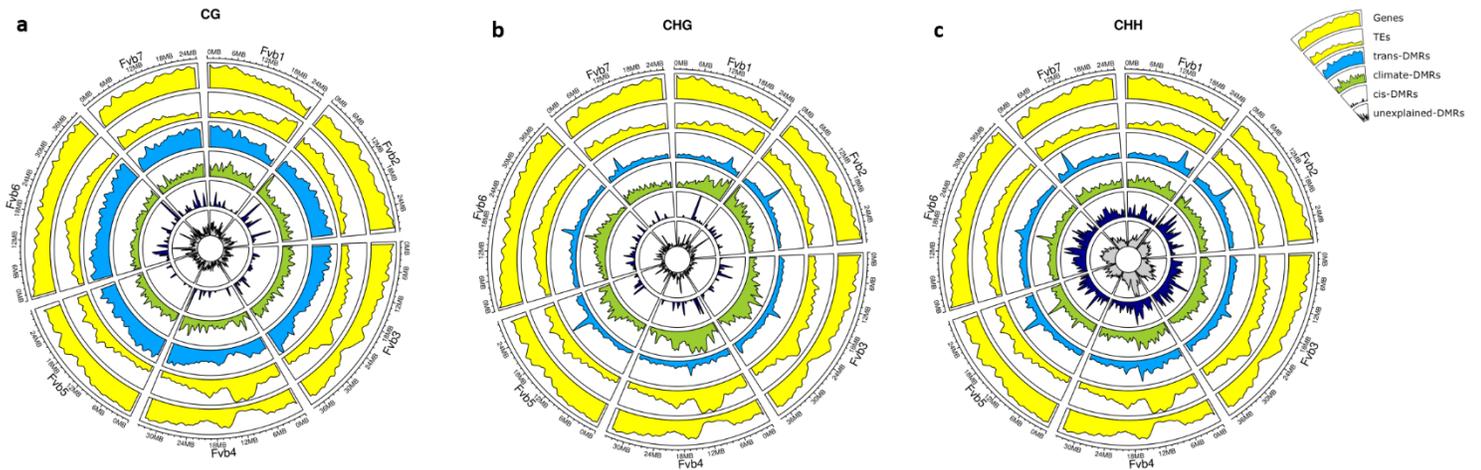


Figure S6: Distribution of the strongest differentially methylated region (DMR) predictor in the garden. Circos plots showing the density of garden-DMRs, gene and transposable element (TE) annotations for all the chromosomes (Fvb) in the CG **(a)**, CHG **(b)** and CHH **(c)** contexts. From outer to inner circles: gene and TE annotations (yellow), predicted DMRs in *trans* (light blue), climate-predicted DMRs (green), predicted DMRs in *cis* (dark blue) and unexplained DMRs (grey). DMRs where all the three predictors failed to explain >10% of the variance are classified as “unexplained”.

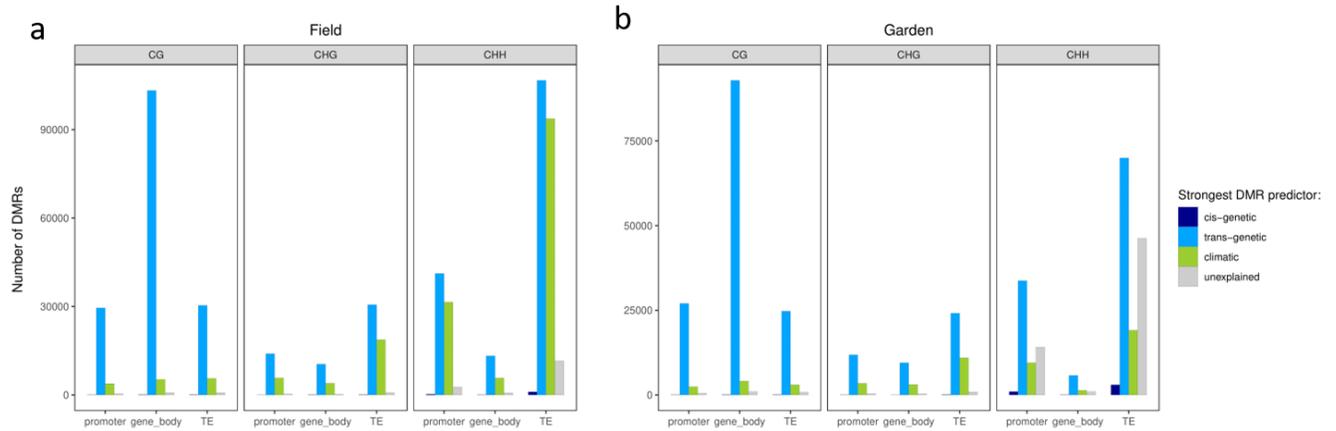


Figure S7: Number of *cis*-, *trans*-, climate-predicted and unexplained differentially methylated regions (DMRs) overlapping promoters, gene bodies or transposable elements (TEs). The number of DMRs shown in the plot exceeds the total number of DMRs identified in this study because a single DMR can overlap with both the promoter and gene body of the same gene. **(a)** Field plants: n = 84. **(b)** Garden plants: n = 84.

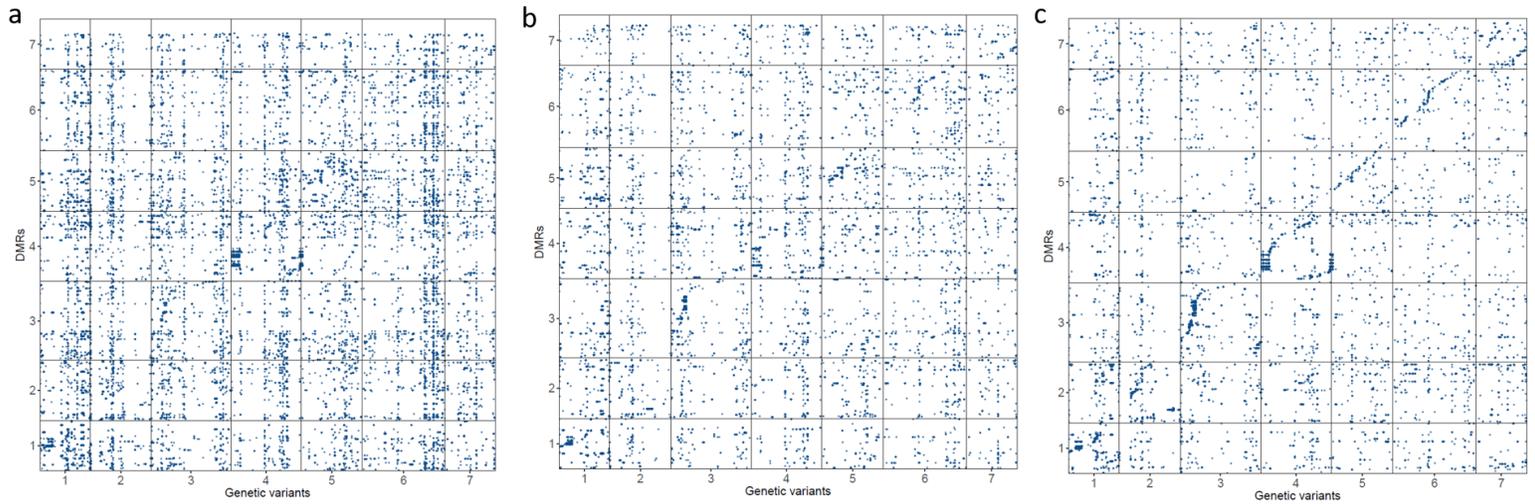


Figure S8: Dot plots for genome-wide association (GWA) analyses for climate-predicted differentially methylated regions (DMRs) overlapping promoters. (a) CG-, (b) CHG- and (c) CHH-DMRs. Each dot represents a significant association between a SNP and a climate-predicted DMR. Only associations that remain significant after Bonferroni correction are plotted. These DMRs were then classified as *indirectly* associated with the environment. The analysis was performed using all the available garden plants ($n = 147$).

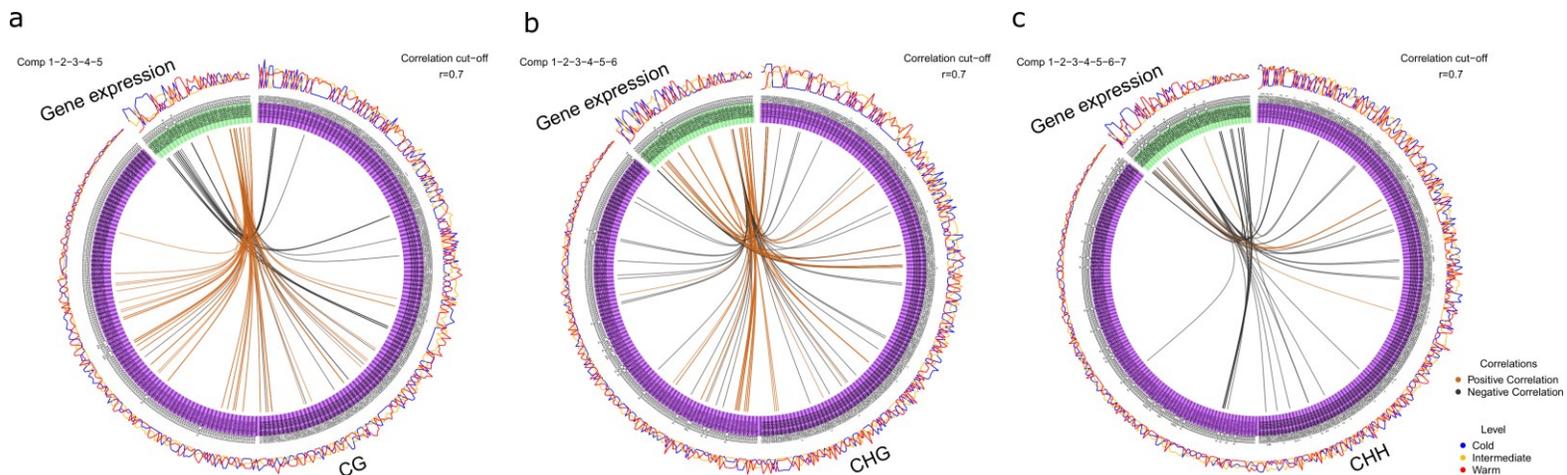


Figure S9: Circos plots showing correlations identified through mixOmics analyses between CG (a), CHG (b) and CHH (c) climate-associated DMRs and gene expression. The links within each Circos plot represent the positive (orange) and negative (black) correlations, with a correlation cutoff of $r = 0.7$. The blue, yellow and red lines outside the Circos plots correspond to the methylation and gene expression levels of the predictive variables under cold, intermediate and warm climatic conditions, respectively. These conditions were determined based on the mean temperature and precipitation from 2011-2018. The populations corresponding to these climatic conditions are as follows: FV_IT_05, FV_IT_07, FV_NO_03, FV_NO_04, FV_NO_05, FV_NO_06, FV_NO_07 (cold); FV_CZ_04, FV_CZ_05, FV_CZ_06, FV_CZ_07, FV_IT_02, FV_IT_06, FV_NO_01, FV_NO_02 (intermediate); FV_CZ_01, FV_CZ_02, FV_CZ_03, FV_IT_01, FV_IT_03, FV_IT_04 (warm). The number of components (comp) retained for each methylation context were 5 for CG, 6 for CHG, and 7 for CHH, after fine-tuning the parameters of our analysis. For variable selection, we kept 100 variables for both CG and CHG methylation, between 30 to 90 variables for CHH methylation, and 10 variables for gene expression across all analyses.

Methods S1

WGBS library preparation and sequencing

We extracted genomic DNA from individual plants using the Qiagen DNeasy Plant Mini Kit, following the manufacturer's instructions with minor modifications. To improve DNA quality and yield from *F. vesca*, a known recalcitrant species, we used an increased amount of buffer AP1 (600 μ l) together with 100 μ l of EDTA (0.5 M, pH=8) and PVPP (polyvinylpolypyrrolidone), and an increased amount of buffer P3 (260 μ l).

We then prepared libraries for WGBS using the NEBNext Ultra II DNA Library Prep Kit and EZ-96 DNA Methylation-Gold MagPrep (ZYMO). Briefly, we sonicated 200-300 ng of genomic DNA to a mean fragment size of \sim 350 bp using the Covaris instrument. We then performed end repair and 3' adenylation of sonicated DNA fragments, ligated the NEBNext adaptors, performed size selection with AMPure XP Beads (Beckman Coulter, Brea, CA), we treated the DNA with bisulfite, we performed PCR enrichment and index ligation using Kapa HiFi Hot Start Uracil+ Ready Mix (Agilent) (14 cycles). Finally, we sequenced paired-end reads on HiSeq X Ten (Illumina, San Diego, CA), using a sequencing coverage per sample of 30x.

Methods S2

Methylation and DMR calling

We used EpiDiverse WGBS pipeline to perform quality control (FastQC v0.11.9), base quality and adaptor trimming (cutadapt v2.10), bisulfite-aware mapping (erne-bs5 v2.1.1) and non-conversion rate calculation, duplicates detection (Picard MarkDuplicates v2.23.3), alignment statistics and methylation calling (MethylDackel v0.4.0) ('MethylDackel'; 'Picard toolkit', 2018; Andrews, 2010; Martin, 2011; Prezza *et al.*, 2016). In the mapping step, we used the most recent version of the genome of *F. vesca* v4.0.a2 (Edger *et al.*, 2018; Li *et al.*, 2019). On average, the sequencing produced 97,142,389 reads per sample (see Table S2 for detailed information), of which 96% mapped successfully to the genome after retaining only uniquely-mapping reads. We calculated the bisulfite non-conversion rate using the chloroplast genome, which is naturally unmethylated (Fojtová *et al.*, 2001), and we found an average bisulfite non-conversion rate among samples of 0.38% (see Table S2). We obtained individual bedGraph files of methylated positions for each sample and sequence context.

For PCA and RDA analyses, we combined the individual bedGraph files from both field and garden conditions in a multisample bedGraph file using custom scripts and bedtools (Quinlan & Hall, 2010). In order to compare the field with the garden conditions, we used only the samples for which we had WGBS data for both conditions ($n = 84$ per condition). We retained all the cytosines having coverage ≥ 5 in at least 80% of the samples (total methylated positions: 1,644,729, 2,574,494 and 12,335,916, respectively for CG, CHG and CHH). We then performed PCA on methylated positions that were transformed using the Hellinger transformation (Borcard *et al.*, 2018). This transformation involves dividing each value in the data matrix by the

sum of its respective row (*i.e.* samples) and subsequently taking the square root of the quotient. This transformation assigns lower weight to variables with low counts or a significant number of zeros.

We performed PCAs using custom scripts with the R function `prcomp` in the `stats` package (v3.5.1) (Sigg & Buhmann, 2008), and colored the plots using either country of origin, mean temperature or precipitation averaged over 7 years before the sampling year (2011-2018), as these were the only recent years available on the C3S Climate Data Store (CDS) website (<https://cds.climate.copernicus.eu/cdsapp#!/home>) (Cornes *et al.*, 2018).

To test for significance of the differences among countries, growth conditions, climatic conditions of origin of the plants and their joint effects, we performed RDA with the `RDA` function in the `vegan` package (v2.6.4) (Oksanen Jari *et al.*, 2020). For DNA methylation, we performed six separate RDA analyses, using in all of them Hellinger-transformed methylated positions as dependent variable and in 1) country of origin as a predictor and growth conditions (field, garden) and climate (mean temperature and precipitation averaged across 2011-2018) as covariates to account for the effect of growth condition and climate on the plants' methylomes; 2) climate of origin of the plants as a predictor and country and growth conditions as covariates; 3) growth conditions as a predictor and country and climate as covariates; 4) country, 5) climate or 6) growth conditions as a predictor and no covariates to calculate the variation in DNA methylation that is explained by each predictor individually, without controlling for the effects of other variables. In the Venn diagrams, we used the values derived from the partial RDA analyses to determine the area of the non-overlapping circles. To calculate the area of the overlapping circles, we subtracted the values obtained from the partial RDA

analyses from those obtained from the RDA analyses with the same predictors and without covariates. We tested the statistical significance of the RDA analyses using a permutation test with 499 permutations.

For DMR calling, we called DMRs separately for all the pairwise comparisons between the populations from the field, and the populations from the garden (*i.e.* we never compared a field population with a garden population). We used as input individual bedGraph files filtered for cytosine coverage ≥ 5 . Separately for field and garden conditions, we then combined the output bed files in a multisample bed file using custom scripts and bedtools (v2.27.1) (Quinlan & Hall, 2010), and we merged the overlapping DMRs obtained from all the pairwise comparisons with bedtools. We used only the samples for which we had WGBS data for both conditions ($n = 84$ per condition). We obtained 82,546 CG-DMRs, 49,459 CHG-DMRs and 211,363 CHH-DMRs for field, and 71,856 CG-DMRs, 37,795 CHG-DMRs and 138,807 CHH-DMRs for garden.

Methods S3

DMR variance decomposition analysis

To assess the amount of methylation variance explained by *cis*-genetic variants, *trans*-genetic variants and climatic variation, we performed a DMR variance decomposition analysis using the `marker_h2()` function from the R package `heritability` (v1.3) (Kruijer *et al.*, 2014). For both field and garden conditions, we ran three mixed models for each individual DMR, and we classified each DMR according to what the strongest predictor of its variance was. If no predictor explained >10% of the variance, the DMR was classified as unexplained (Galanti *et al.*, 2022). Each mixed model included one random factor matrix that captured one of the three predictors, as in (Galanti *et al.*, 2022):

```
cis_h2 = marker_h2(my_data$DMR_meth, my_data$sample, covariates = NULL, cis_ibs,
max.iter = 100)
```

```
trans_h2 = marker_h2(my_data$DMR_meth, my_data$sample, covariates = NULL, trans_ibs,
max.iter = 100)
```

```
env_h2 = marker_h2(my_data$DMR_meth, my_data$Pop, covariates = NULL, clim_norm,
max.iter = 100),
```

where “DMR_meth” represents a vector of DMR-methylation values, “sample” refers to sample names, “Pop” to population names, and `cis_ibs`, `trans_ibs` and `clim_norm` contain matrices of *cis*-variants, *trans*-variants and climatic variation, respectively.

For *cis*-variants, we used an Isolation-By-State (IBS) matrix generated with PLINK (v1.90b6.12) (<http://pngu.mgh.harvard.edu/purcell/plink/>) (Purcell *et al.*, 2007) using genetic variants within 50kb from the DMR middle point. For *trans*-variants, we used an IBS matrix obtained from

genetic variants filtered for Minor Allele Frequency (MAF) ≥ 0.01 and pruned for Linkage Disequilibrium (LD) with an LD threshold (r^2) of 0.8 for SNP pairs in a sliding window of 50 SNPs, sliding by 5. For climatic variation, we calculated a Euclidean distance matrix between climatic data from all the field sites, which we reversed and normalized to obtain a similarity matrix in a 0 to 1 range. We used the same matrix for both field and garden conditions. The climatic data included mean, maximum and minimum temperature, and precipitation, all averaged across 2011-2018, as these were the only recent years available on the website. We sourced climatic data at a resolution of $0.1 \times 0.1^\circ$ (v20.0e) from the European gridded dataset E-OBS, available through the C3S Climate Data Store (CDS) website (<https://cds.climate.copernicus.eu/cdsapp#!/home>) (Cornes *et al.*, 2018). For methylation variants, we use the merged DMRs obtained from all the pairwise comparisons with bedtools, separately for field and garden (see above). As above, we used only the samples for which we had WGBS data for both conditions ($n = 84$ per condition). We extracted average methylation of the resulting DMRs from all the samples with the function *regionCounts* from the R package methylKit (v1.16.1) (Akalin *et al.*, 2012), using a minimum cytosine coverage of 5.

Methods S4

Genome-wide association (GWA) analysis

We ran GWA analysis as described in (Galanti *et al.*, 2022), using the R package rrBLUP (4.6.1) (Endelman, 2011). For genetic variants, we imputed the missing genotype calls with BEAGLE 5.2 (Browning *et al.*, 2018) and filtered for MAF > 0.04. After filtering, we were able to retain 83,095 SNPs. We corrected for population structure using an IBS matrix obtained from variants filtered for MAF \geq 0.01 and pruned for Linkage Disequilibrium (LD) with an LD threshold (r^2) of 0.8 for SNP pairs in a sliding window of 50 SNPs, sliding by 5. We used individual average DMR-promoter methylation for each sequence context as phenotype, calculated with the *regionCounts* methylKit function (v1.16.1) (Akalin *et al.*, 2012), using a minimum cytosine coverage of 5. To determine the significance threshold, we applied the Bonferroni correction method.

Methods S5

RNA-sequencing

We extracted mRNA using the Nucleospin RNA Plus kit (Macherey Nagel), following the manufacturer's instructions with minor modifications. To improve RNA quality and yield from *F. vesca*, we used an increased amount of lysis buffer (500 µl) together with 100 µl of EDTA (0.5 M, pH=8) and PVPP (polyvinylpolypyrrolidone). The cDNA library and sequencing (PE150, 6 Gb per sample of raw data) were performed by Novogene Co., Ltd, Cambridge, using an Illumina NovaSeq 6000 platform. On average, we obtained 22,252,025 raw reads. We trimmed adaptors with cutadapt (v1.16) and assessed sequencing quality with MultiQC (v1.10.1) (Ewels *et al.*, 2016). We aligned the reads to the *Fragaria vesca* genome (v4.0.a2) using STAR (Spliced Transcripts Alignment to a Reference) (v2.7.1a) (Dobin *et al.*, 2013), assembled them into transcripts and quantified using StringTie (v2.1.5) (Kovaka *et al.*, 2019).

Table S1: Origin sites and common garden characteristics of *Fragaria vesca* populations used in this study. Country, population ID, population size, geographic coordinates, and climatic variables (mean, maximum, minimum temperature and precipitation) averaged over 7 years before the sampling year (2011-2018). For the common garden, the climatic variables refer to the year of cultivation of the plants in such conditions (2018-2019). For this condition, we do not report precipitation as these plants were watered regularly.

Table S2: Mapping statistics for whole genome bisulfite sequencing (WGBS) analysis, single nucleotide polymorphisms (SNP) calling and RNA-sequencing. WGBS: number of uniquely mapped reads, percentage of duplicate sequences, coverage and bisulfite non-conversion rate per sample. SNP calling: base count of Phred value > 20 (Q20) or 30 (Q30). RNA-seq: total number of raw reads and raw data, percentage of bases with Phred value > 20 or 30, total amount of mapped data and percentage of duplicate sequences.

Table S3: Correlation between number of *cis*-, *trans*-, climate-, unexplained-DMRs and number of genes and TEs. The correlation was calculated considering the DMR, gene and TE counts assigned to 1 kb genomic bins. Values represent the Pearson correlation coefficient (r). All the values are significant at $P < 0.05$. Field plants: $n = 84$, garden plants: $n = 84$.

Table S4: Enrichment ratio analysis of *trans*-predicted DMRs in CHG and CHH contexts, and unexplained DMRs in the CHH context, across different TE superfamilies. In all cases, the statistical tests assessing the random distribution of DMRs among TE superfamilies yielded

significant results ($P < 0.001$, Fisher's exact test). We employed an enrichment ratio threshold of 1. The TE superfamilies are named as per the TE annotation file: DTA for hat, DTC for CACTA, DTH for PIF/Harbinger, DTM for Mutator, DTT for Tc1/Mariner and LTR for long terminal repeat.

Table S5: Gene Ontology (GO) enrichment analysis of predicted DMRs in *cis*, *trans*, climate-predicted and unexplained DMRs. Only GO terms with an adjusted P -value < 0.05 are shown. BP: biological process, CC: cellular component.

Table S6: List of climate-predicted DMRs found in the garden condition with a statistically significant correlation between promoter methylation and expression of the adjacent gene. Only correlations with P -value < 0.05 are shown.

Table S7: List of correlations identified through mixOmics analyses between climate-predicted DMRs and gene expression. Only correlations with an absolute value of $r \geq 0.7$ are shown. This stringent cut-off ensures the inclusion of only highly correlated DMRs.

Table S8: List of overlapping differentially expressed genes (DEGs) and environmentally linked DMRs that had significant correlations with gene expression. Only DEGs with an adjusted P value < 0.05 and an absolute value of fold change (FC) ≥ 1.5 were included in the analysis.

References

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. 2012.

MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology* **13**: 1–9.

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

Borcard D, Gillet F, Legendre P. 2018. Spatial Analysis of Ecological Data. : 299–367.

Browning BL, Zhou Y, Browning SR. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. *American journal of human genetics* **103**: 338–348.

Cornes RC, van der Schrier G, van den Besselaar EJM, Jones PD. 2018. An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets. *Journal of Geophysical Research: Atmospheres* **123**: 9391–9409.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**: 15–21.

Edger PP, VanBuren R, Colle M, Poorten TJ, Wai CM, Niederhuth CE, Alger EI, Ou S, Acharya CB, Wang J, et al. 2018. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity.

GigaScience **7**.

Endelman JB. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* **4**: 250–255.

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

Fojtová M, Kovařík A, Matyášek R. 2001. Cytosine methylation of plastid genome in higher

plants. Fact or artefact? *Plant science : an international journal of experimental plant biology* **160**: 585–593.

Galanti D, Ramos-Cruzid D, Nunnid A, Rodríguez-Aré Valoid I, Scheepensid JF, Beckerid C, Bossdorfid O. 2022. Genetic and environmental drivers of large-scale epigenetic variation in *Thlaspi arvense* (NM Springer, Ed.). *PLOS Genetics* **18**: e1010452.

Kovaka S, Zimin A V., Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology* **20**: 1–13.

Kruijer W, Boer MP, Malosetti M, Flood PJ, Engel B, Kooke R, Keurentjes JJB, Van Eeuwijk FA. 2014. Marker-based estimation of heritability in immortal populations. *Genetics* **199**: 379–398.

Li Y, Pi M, Gao Q, Liu Z, Kang C. 2019. Updated annotation of the wild strawberry *Fragaria vesca* V4 genome. *Horticulture Research* **6**.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

MethylDackel. <https://github.com/dpryan79/MethylDackel/>

Oksanen Jari, Blanchet F. Guillaume, Friendly Michael, Kindt Roeland, Legendre Pierre, McGlinn Dan, Minchin Peter R., O'Hara R. B., Simpson Gavin L., Solymos Peter, et al. 2020.

CRAN - Package vegan.

Picard toolkit. 2018. <https://github.com/dpryan79/MethylDackel>

Prezza N, Vezzi F, Käller M, Policriti A. 2016. Fast, accurate, and lightweight analysis of BS-treated reads with ERNE 2. *BMC Bioinformatics* **17**: 69.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and

population-based linkage analyses. *American journal of human genetics* **81**: 559–575.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**: 841–842.

Sigg CD, Buhmann JM. 2008. Expectation-maximization for sparse and non-negative PCA. *Proceedings of the 25th International Conference on Machine Learning*: 960–967.