Beyond the CpG: an integrative approach to decoding DNA methylation in immunometabolic health

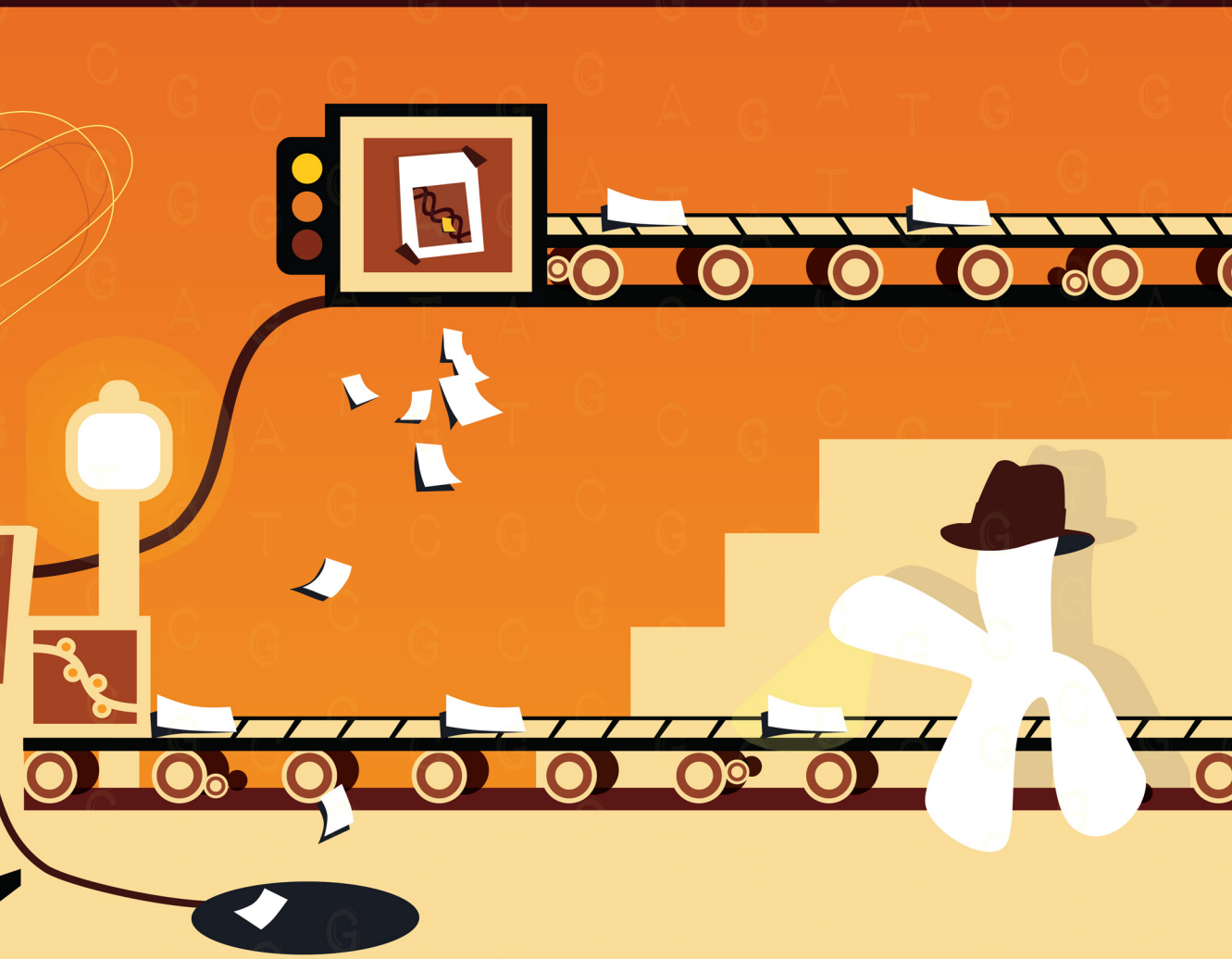
Sinke, L.J.

CHAPTER FIVE

DNAmArray

DNAmArray: streamlined workflow for the quality control, normalisation, and analysis of Illumina methylation array data

Lucy Sinke[1], Maarten van Iterson[1], Davy Cats[1], Tom Kuipers[1], and Bastiaan T. Heijmans[1]



[1] Leiden University Medical Centre, Leiden, The Netherlands

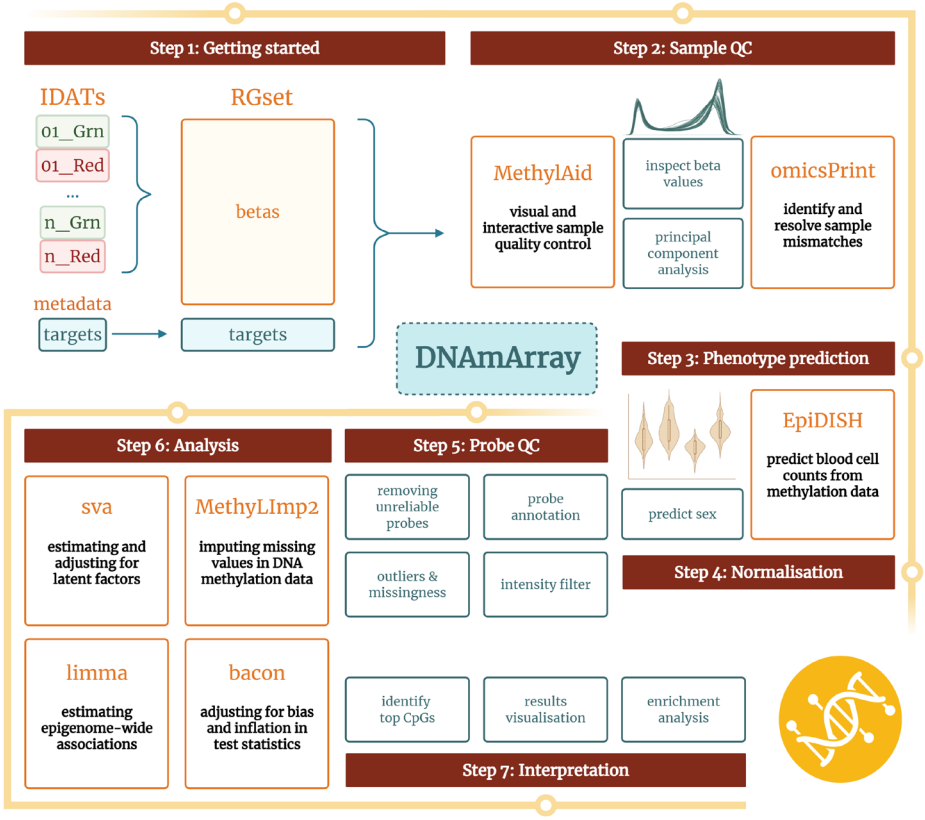
Published on Zenodo (2025)
DOI: 10.5281/zenodo.3332709

Abstract

Available packages for preprocessing, quality control, and analysis of DNA methylation data offer powerful and flexible functionality. However, these tools often exist as isolated elements, requiring users to navigate a complex landscape with limited guidance on how to implement them in a coherent workflow. A streamlined, well-documented pipeline that combines up-to-date processes and packages has potential to improve the accessibility of epigenome-wide analyses, facilitating good quality reporting that can keep up with data generation efforts. With the recent release of the Illumina Infinium Methylation EPIC v2.0 array in mind, we present *DNAmArray*, a well-documented workflow that details necessary fixes and enables the continued use of established methods in epigenomic studies.

DNAmArray is a modular R package and complementary workflow for the preprocessing, quality control, and analysis of DNA methylation array data, tailored for epigenome-wide association studies. Drawing on nearly a decade of experience analysing large-scale genomic data, most notably within the BIOS consortium, DNAmArray integrates widely accepted best practices and practical helper functions to streamline analyses in an up-to-date and user-friendly manner. The workflow takes raw IDAT files as input and transforms them via a series of clearly defined steps. Rather than prescribing a single best approach, DNAmArray transparently presents options, enabling users to make informed choices based on their specific data and research goals. Both the package and workflow are extensively validated on both Illumina 450K and EPIC arrays, with specific notes and adjustments provided for the EPIC v2.0 platform.

Graphical Abstract



Highlights

- *DNAmArray* is a modular workflow for accessible and user-friendly DNA methylation data processing and analysis.
- It is equipped with complementary documentation, which walks through an epigenome-wide analysis on publicly available example data.
- Both the pipeline and package are up-to-date and validated on Illumina 450k, EPIC, and EPIC v2.0 arrays.

Keywords: *DNA methylation; EWAS; data pipeline; reproducibility; Illumina*

Background

DNA methylation (DNAm) is a well-characterized epigenetic modification that plays a central role in regulating gene expression, cellular differentiation, and genomic stability[1–3]. Epigenome-wide association studies (EWAS) leverage DNAm data to identify CpG sites whose methylation is associated with phenotypes of interest, including immunometabolic traits, ageing, and environmental exposures[4–6]. Such studies have potential to offer mechanistic insights into the pathways underlying human health and disease.

To support these efforts, DNAm is often profiled using microarray-based technologies such as the Illumina Infinium HumanMethylation450 BeadChip (450K) and its successors, the Infinium MethylationEPIC (850K) and the recently released MethylationEPIC v2.0 array[7]. These platforms enable cost-effective, high-throughput measurement of methylation at hundreds of thousands of CpGs across the genome. However, despite technological advances, EWAS inherently share many statistical and technical challenges with genome-wide association studies (GWAS), including the need to correct for multiple testing, batch effects, and at times population structure. EWAS also present unique sources of confounding, including cell-type heterogeneity, age, and lifestyle influences[8]. To this end, several R packages have been developed to support various aspects of DNAm analysis, including *methylAid*[9], *omicsPrint*[10], and *bacon*[11], each providing specific tools for quality control (QC), preprocessing, and analysis. While powerful, these packages require thoughtful integration and manual customization to form a complete analysis pipeline compatible with the latest technology[12].

Here, we present *DNAmArray*, an R package and accompanying workflow designed to streamline the analysis of Illumina DNAm array data, including support for the EPIC v2.0 platform. *DNAmArray* integrates widely accepted methods from across the EWAS landscape with practical, in-house functions. It is informed by nearly a decade of work with large-scale datasets, including analysis of almost 4,000 samples from six Dutch biobanks as part of the Biobank-based Integrative Omics Study (BIOS) consortium. Rather than prescribing a rigid pipeline, *DNAmArray* emphasizes transparency, flexibility, and reproducibility by presenting multiple valid options in a modular framework. To aid adoption and implementation, we offer extensive documentation of the analysis workflow for an example EPIC dataset[13].

Results

Step 1 | Getting started

The example dataset used to illustrate the *DNAmArray* workflow is available from the NCBI Gene Expression Omnibus (GSE116339)[13]. It includes DNAm data profiled using the Illumina Infinium MethylationEPIC BeadChip array in 679 whole blood samples. The samples were collected from individuals enrolled in the Michigan PBB Registry, a longitudinal cohort established following an agricultural accident in the 1970s. This accident resulted in widespread exposure to polybrominated biphenyl (PBB) and, because PBB is both lipophilic and biologically stable, those exposed still have detectable levels in their blood many decades later. All samples were collected between 2004 and 2015, approximately 31 to 42 years after the initial exposure event. The resulting DNAm data has been previously analysed in an EWAS of PBB exposure and is well-suited for demonstrating preprocessing, normalisation, and analysis steps within the *DNAmArray* package.

Step 2 | Sample-level quality control

To ensure high data quality for downstream analyses, rigorous sample-level QC is conducted, detecting and excluding samples that exhibit signs of sample mixups or processing errors[14]. The input for this step is an RGset constructed from raw IDAT files of red and green channel intensities combined with sample metadata.

DNAmArray visualises sample quality using plots from the Bioconductor package, *methylAid*[9], utilising the control probes present on Illumina arrays. Each type of control probe is designed to evaluate a distinct stage of the DNAm protocol from bisulphite conversion through extension to hybridization. In addition, signals from negative control probes, which represent randomly permuted sequences and therefore should not hybridize to the DNA template, assess if a user-defined proportion of probes from each sample are visually distinct from background noise[15]. On the basis of these plots, outliers are detected and *DNAmArray* provides updated thresholds for not only the EPIC but also the EPICv2 array. All *MethylAid* plots accept custom thresholds and can be coloured by variables in the sample metadata (**Fig. 1a-d**). Furthermore, *DNAmArray* details how to explore and flag samples by inspecting their β-value distributions and outlines principal component approaches to identify sample clusters or outliers.

As an important part of sample-level QC, *DNAmArray* implements *omicsPrint*[10]. This package detects data linkage errors by comparing recorded sample relationships to those predicted from probes that contain common single nucleotide polymorphisms (SNPs). By examining genetic similarity, samples are inspected for duplicates, underlying family relationships, or sample mix-ups, which can then be resolved or corrected for.

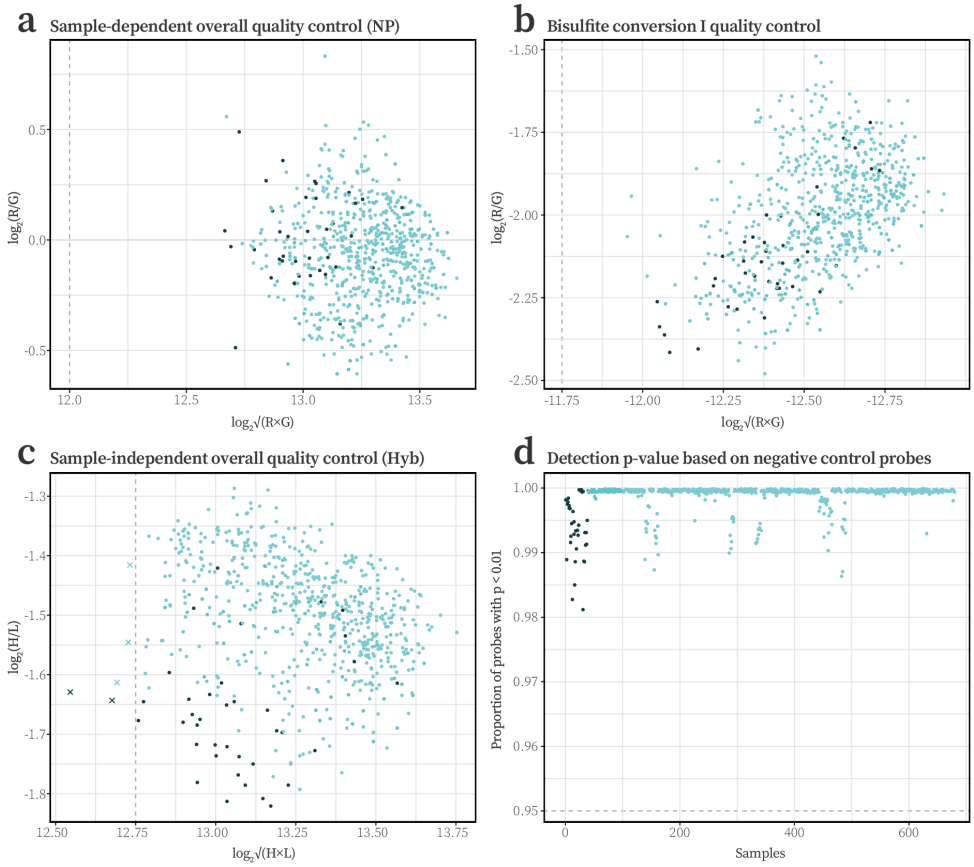


Figure 1 | *MethylAid* plots provided within *DNAmArray*. a) Sample-dependent overall QC based on non-polymorphic (NP) control probes, ensuring that both A and T (red) and C and G (green) nucleotides are detected in sufficient intensities. **b)** Bisulphite-conversion (BC) plot visualises if the conversion reaction was successful for all samples. Converted (C) probes are extended and measured in the defined colour channel (C1 and C2: green; C3 and C4: red). **c)** Sample-independent overall QC measured using hybridization (Hyb) controls. These synthetic targets complement the array perfectly and are present in the hybridization buffer at three concentration levels. This plot ensures distinction between high (H) and low (L) concentrations in the green channel. In the example data, poor hybridization is evident for five samples, which are removed from downstream analysis. **d)** Detection *p*-value plot displaying the proportion of probes (y-axis) that are distinguishable from the background noise (default $p < 0.01$). Background signal is determined using control probes of randomly permuted sequences and thereby not designed to hybridize to the DNA template. The user can specify both the proportion and *p*-value thresholds for this plot.

Step 3 | Predicting phenotypes

The *DNAmArray* workflow implements phenotype prediction, offering tools to validate sample identity, impute missing biological data, and adjust for variable factors. These predictions are particularly important considering the sensitivity of EWAS to biological confounding[22]. DNAm-based predictors of immune cell proportions are increasingly able to capture a larger amount of cellular diversity, and *DNAmArray* outlines how to estimate such subsets and add them to metadata and models (**Fig. 2a**)[23].

Furthermore, *DNAmArray* outlines steps from the *wateRmelon* package for sex prediction, offering a user-friendly way to confirm and resolve sample identity and check for less common karyotypes (**Fig. 2b**)[17]. To ensure optimal predictions, the workflow ensures this step is performed prior to probe masking. Many commonly masked probes, such as those on sex chromosomes, are essential components for prediction algorithms, and implementation on incomplete data is likely to reduce their accuracy. By comparing predicted and measured traits, *DNAmArray* increases confidence in data integrity, enriches metadata, and assists in downstream analyses.

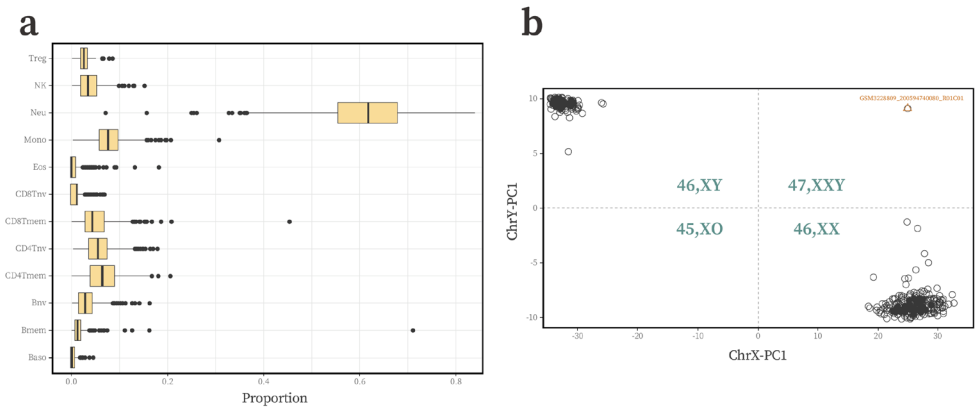


Figure 2 | Phenotype prediction in the *DNAmArray* workflow. a) Twelve immune cell type proportions are predicted from DNAm data using *EpiDISH*. **b)** DNAm at CpGs located on sex chromosomes is used to predict sex, including XXY and XO karyotypes.

Step 4 | Normalisation

Whilst β-values derived from the Illumina arrays are relatively robust due to their ratio-based formulation, this intrinsic stability does not preclude the presence of unwanted technical variation[16]. For studies aiming to detect subtle, biologically meaningful changes, refined correction of such technical variation is essential, as unaddressed technical noise can obscure methylation patterns. This can compromise both statistical power and reproducibility.

DNAmArray outlines implementation of multiple normalisation approaches, including *dasen* and functional normalization (*FunNorm*), using the *wateRmelon* Bioconductor package[17]. *FunNorm* is an approach designed specifically for Illumina methylation arrays that extends quantile normalization through the use of control probes[18]. Unlike traditional quantile normalization, which enforces identical distributions across samples, *FunNorm* retains biologically meaningful differences by estimating and removing variation associated with latent technical factors of the control probe signal. This control-informed approach offers a principled way to improve comparability across samples while minimizing the risk of overcorrection, particularly in biologically informative regions of the methylation spectrum (i.e., intermediate β-values around 0.5). By preserving these subtleties, *FunNorm* enhances the sensitivity of downstream differential methylation analysis[19].

Step 5 | Probe-level quality control

Data-driven probe-level QC is the final data processing step outlined by *DNAmArray*. It provides functions aimed at excluding unreliable data points, including those with zero intensity or based on fewer than three beads. Probes with more than 5% missing data across samples are removed to maximize data completeness and reduce bias. Additionally, probes previously established as unreliable due to cross-hybridization, ambiguous genomic mapping, or the presence of SNPs at critical binding sites are excluded[20]. Users are provided with curated probe masking objects for the 450K, EPIC, and EPIC v2.0 arrays, which flag CpGs located in ENCODE Blacklist regions[21]. To further safeguard against technical artifacts, *DNAmArray* includes options for detecting and handling spurious methylation values using outlier-based filtering approaches, which can be applied to both β and M values.

Following probe QC, the cleaned β-matrix, filtered probe annotations, and corresponding sample metadata are combined into a *SummarizedExperiment* object. This structure maintains internal consistency across assay data and phenotypic annotations and is compatible with existing Bioconductor packages and workflows.

Step 6 | EWAS pipeline

The *DNAmArray* package provides a comprehensive and customisable pipeline for performing EWAS, aimed at improving the accessibility of DNAm data analysis. To facilitate informed model specification, we outline steps to visualise relationships between potential covariates and methylation data, estimate latent factors, and incorporate random effects (**Fig. 3a**). Latent factors address sources of variation not captured by measured variables, and *DNAmArray* implements surrogate variable analyses (SVA) to estimate hidden batch effects in a data-driven approach. Proper specification of full and null models is described, alongside best practices for imputing DNAm using

methyLImp2[25]. By adjusting for both predicted confounding factors, such as estimated cell types, and unmeasured confounders including surrogate variables (SVs), the robustness and interpretation of resulting EWAS findings can be improved.

At its core, the EWAS component relies on the widely used *limma* package to perform linear modeling at all tested CpGs[24]. Steps are outlined to properly specify appropriate models including both fixed and random effects. Yet, even with proper model specification the test statistics from epigenome- and transcriptome-wide studies are susceptible to bias and inflation[11]. Therefore, *DNAmArray* implements *bacon*, a Bayesian method that controls these effects using the empirical null distribution. Correct implementation of this package is outlined, alongside complementary visualisations to assess performance (**Fig. 3b**).

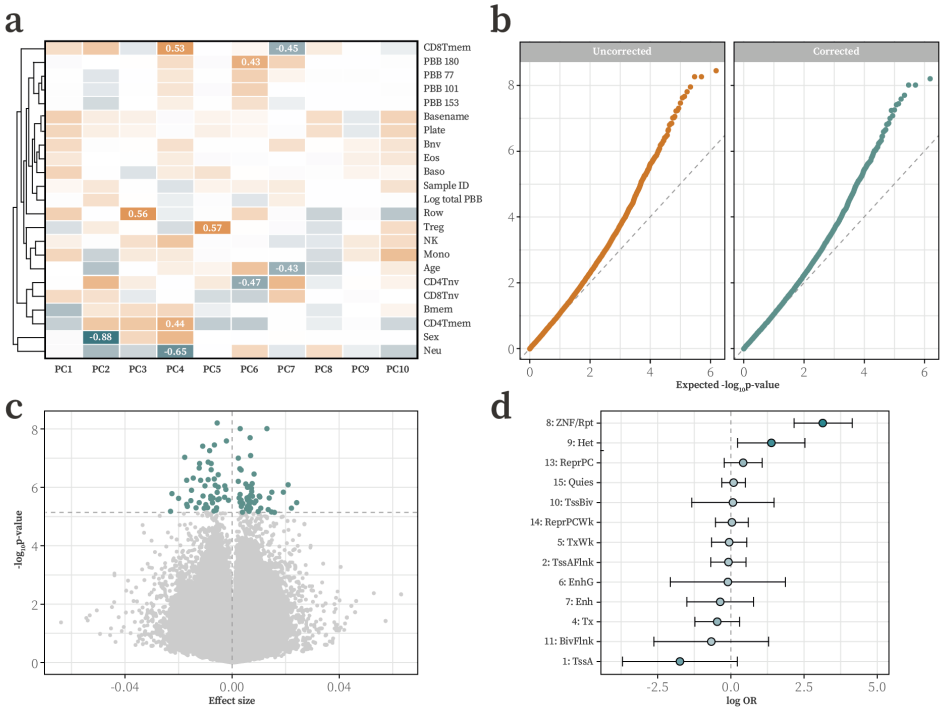


Figure 3 | Analysis steps within the *DNAmArray* workflow. a) Methods to properly specify models including heatmaps and principal component analysis (PCA) are included. **b)** Application of *bacon* allows test statistics to be adjusted for bias and inflation. **c)** EWAS results are visualized using volcano plots. **d)** Enrichment analyses, including for chromatin state, are described and visualised.

Step 7 | Interpretation of results

There is a pressing need to shift from hypothesis generation to biological interpretation in EWAS. Many CpGs have now been robustly associated with human health but remain incompletely characterised. Therefore, *DNAmArray* dedicates its final step to functional interpretation of results. Following standard steps to adjust for multiple testing, identify significant CpGs, and visualise results (**Fig. 3c**), the workflow guides users through probe annotation and chromatin state enrichment analyses (**Fig. 3d**). This serves to exemplify methods and helpful resources that be used to more fully characterise EWAS signals[26]. In this manner, *DNAmArray* paves the way for a user-driven improvement in EWAS interpretation in both novel and existing data.

Discussion

The development of *DNAmArray* addresses a gap in the transparency, reproducibility, and accessibility of EWAS analyses, particularly in the context of rapidly evolving array technologies. As the field transitions from the Illumina 450K to its successors, researchers face ongoing challenges in maintaining robust, reproducible, and up-to-date analysis workflows[27]. *DNAmArray* responds to these challenges by providing a modular, well-documented R package and workflow, integrating best practices from the epigenomics community with practical, experience-driven enhancements.

By validating steps on both 450K and EPIC arrays and providing clear guidance for EPIC v2.0, the workflow ensures continuity for researchers transitioning to new technologies. The emphasis on rigorous QC at both the sample and probe level improves the quality of input data for EWAS, and tools such as *methylAid*[9] and *omicsPrint*[10] alongside curated probe masking resources enable users to identify and address technical artifacts and data linkage errors that could otherwise compromise downstream analyses[20]. The inclusion of flexible, user-adjustable thresholds and visualisation options additionally allows QC procedures to be tailored to a range of datasets. Beyond CpG associations, the workflow highlights the importance of functional interpretation for translating EWAS findings into insights with biological context. By outlining clear steps for such analyses and describing extensions, *DNAmArray* helps bridge the gap between statistical results and biological understanding.

While *DNAmArray* represents a significant advancement, certain limitations remain. The workflow's focus on array-based technologies means that it is not suited to perform analysis on data outside of targeted CpG sites, and the accuracy of phenotype prediction and cell-type deconvolution depends on the quality and relevance of available reference data. As new array platforms and sequencing-based approaches emerge, ongoing maintenance and community engagement will be essential to ensure that *DNAmArray* remains current and widely applicable. Additionally, data pipelines can only process and adjust for measured variables. Other packages, such as *Omixer*[28], allow researchers to proactively reduce the impact of technical variation on the quality of their data and biological signals of interest, and study designs should rely on a combination of proactive measures alongside reproducible preprocessing and analysis to ensure robust results.

In summary, *DNAmArray* provides a timely, flexible, and user-friendly solution for the preprocessing, quality control, and analysis of Illumina DNAm array data. By integrating best practices with practical enhancements, and by supporting the latest array technologies, DNAmArray enables reproducible and interpretable EWAS, facilitating future discoveries in epigenetic epidemiology.

Author contributions

B.T.H. and M.v.I. conceived and initially designed the pipeline. L.S. and B.T.H. updated it to its current iteration, developed the complementary workflow, and tested the pipeline on example datasets. D.C. and T.K. provided bioinformatics support. T.K. additionally contributed scripts to implement *MethylAid* in the workflow.

Funding

Development of the workflow was funded by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007). The work of L.S. was supported by the Joint Programming Initiative 'a Healthy Diet for a Healthy Life' (JPI-HDHL) DIMENSION project [ZonMW project number: 529051021].

Declarations of interest

The authors declare no competing interests.

Data & code availability

DNAmArray is available from GitHub at `molepi/DNAmArray` and all versions are archived on Zenodo[29].

All software used is open source and freely available Unless stated otherwise, all calculations were performed using R version 4.2.2.

Acknowledgements

We are grateful to Jenny van Dongen for continued testing and improvement of the workflow. In addition, we thank Paul Hop, Jazmin Taubert, Yunfeng Liu, Manhoor Sulaiman, Thomas Jonkman, Helena Rasche, Elmar W. Tobi, Roderick Slieker, Wouter den Hollander, Rene Luijk, and Koen F. Dekkers for contributing to the testing and development of the package and associated workflow.

References

1. Hannon, E. *et al.* Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits. *Am J Hum Genet* **103** (5): 654-665 (2018).
2. Suelves, M. *et al.* DNA methylation dynamics in cellular commitment and differentiation. *Brief Funct Genomics* **15** (6): 443-453 (2016).
3. Héberlé, É. *and* Bardet, A. F. Sensitivity of transcription factors to DNA methylation. *Essays Biochem* **63** (6): 727-741 (2019).
4. Wielscher, M. *et al.* DNA methylation signature of chronic low-grade inflammation and its role in cardio-respiratory diseases. *Nat Commun* **13** (1): 2408 (2022).
5. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541** (7635): 81-86 (2017).
6. Dekkers, K.F. *et al.* Blood lipids influence DNA methylation in circulating cells. *Genome Biol* **17** (1): 138 (2016)
7. Noguera-Castells, A. *et al.* Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. *Epigenetics* **18** (1): 2185742 (2023).
8. Van Rooij, J. *et al.* Evaluation of commonly used analysis strategies for epigenome- A nd transcriptome-wide association studies through replication of large-scale population studies. *Genome Biol* **20** (1): 235 (2019).
9. Van Iterson, M. *et al.* MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics* **30** (23): 3435-3437 (2014).
10. Van Iterson, M. *et al.* omicsPrint: detection of data linkage errors in multiple omics studies. *Bioinformatics* **34** (12): 2142-2143 (2018).
11. van Iterson, M. *et al.* Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol* **18** (1): 19 (2017).
12. Ori, A. P. S. *et al.* Significant variation in the performance of DNA methylation predictors across data preprocessing and normalization strategies. *Genome Biol* **23** (1): 225 (2022).
13. Curtis, S. W. *et al.* Exposure to polybrominated biphenyl (PBB) associates with genome-wide DNA methylation differences in peripheral blood. *Epigenetics* **14** (1): 52-66 (2019).
14. Bhat, B. *and* Jones, G. T. Data Analysis of DNA Methylation Epigenome-Wide Association Studies (EWAS): A Guide to the Principles of Best Practice. *Methods Mol Biol* **2458**: 23–45 (2022).
15. Heiss, J. A. *and* Just, A. C. Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses. *Clin Epigenetics* **11** (1): 15 (2019).
16. Xu, Z. *and* Taylor, J. A. Reliability of DNA methylation measures using Illumina methylation BeadChip. *Epigenetics* **16** (5): 495-502 (2020).
17. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**: 293 (2013).
18. Fortin, J. P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol* **15** (12): 503 (2014).
19. Heiss, J. A. *and* Brenner, H. Between-array normalization for 450K data. *Front Genet* **6:** 92 (2015).
20. Zhou, W., Laird, P. W. *and* Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* **45** (4): e22 (2017).
21. Amemiya, H. M., Kundaje, A. *and* Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9** (1): 9354 (2019).
22. Teschendorff, A. E. *and* Zheng, S. C. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics* **9** (5): 757-768 (2017).
23. Zheng, S. C. *et al.* EpiDISH web server: Epigenetic dissection of intra-sample-heterogeneity with online GUI. *Bioinformatics* **36** (6): 1950-1951 (2020).
24. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43** (7): e47 (2015).
25. Plaksienko, A. *et al.* methyLImp2: faster missing value estimation for DNA methylation data. *Bioinformatics* **40** (1): btae001 (2024).
26. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38** (4): 576-589 (2010).
27. Solomon, O. *et al.* Comparison of DNA methylation measured by Illumina 450K and EPIC BeadChips in blood of newborns and 14-year-old children. *Epigenetics* **13** (6): 655-664 (2018).
28. Sinke, L., Cats, D. *and* Heijmans, B. T. Omixer: multivariate and reproducible sample randomization to proactively counter batch effects in omics studies. *Bioinformatics* **37** (18): 3051–3052 (2021).
29. Sinke, L. *et al.* DNAmArray: Streamlined workflow for the quality control, normalization, and analysis of Illumina methylation array data. *Zenodo* (2025).