Bioinformatics, 2024, **40(7)**, btae423
https://doi.org/10.1093/bioinformatics/btae423
Advance Access Publication Date: 4 July 2024
Applications Note

OXFORD

Genome analysis

mLiftOver: harmonizing data across Infinium DNA methylation platforms

Brian H. Chen[1] **and Wanding Zhou** [2,3,*]

[1]California Pacific Medical Center Research Institute, Sutter Health, San Francisco, CA 94143, United States
[2]Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, United States
[3]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States

*Corresponding author. Department of Pathology and Laboratory Medicine, University of Pennsylvania, 3501 Civic Center Blvd., Philadelphia, PA 19104, United States. E-mail: wanding.zhou@pennmedicine.upenn.edu (W.Z.)

Associate Editor: Can Alkan

Abstract

Motivation: Infinium DNA methylation BeadChips are widely used for genome-wide DNA methylation profiling at the population scale. Recent updates to probe content and naming conventions in the EPIC version 2 (EPICv2) arrays have complicated integrating new data with previous Infinium array platforms, such as the MethylationEPIC (EPIC) and the HumanMethylation450 (HM450) BeadChip.

Results: We present mLiftOver, a user-friendly tool that harmonizes probe ID, methylation level, and signal intensity data across different Infinium platforms. It manages probe replicates, missing data imputation, and platform-specific bias for accurate data conversion. We validated the tool by applying HM450-based cancer classifiers to EPICv2 cancer data, achieving high accuracy. Additionally, we successfully integrated EPICv2 healthy tissue data with legacy HM450 data for tissue identity analysis and produced consistent copy number profiles in cancer cells.

Availability and implementation: mLiftOver is implemented R and available in the Bioconductor package SeSAMe (version 1.21.13+): https://bioconductor.org/packages/release/bioc/html/sesame.html. Analysis of EPIC and EPICv2 platform-specific bias and high-confidence mapping is available at https://github.com/zhou-lab/InfiniumAnnotationV1/raw/main/Anno/EPICv2/EPICv2ToEPIC_conversion.tsv.gz. The source code is available at https://github.com/zwdzwd/sesame/blob/devel/R/mLiftOver.R under the MIT license.

1 Introduction

The Infinium DNA methylation BeadChips (Illumina Inc., San Diego, CA, United States) (Bibikova *et al.* 2006) are widely used assay technologies for population-scale DNA methylation profiling, including meQTL studies (Min *et al.* 2021, Hawe *et al.* 2022), epigenetic risk scoring (Aref-Eshghi *et al.* 2020, Thompson *et al.* 2022), and epigenome-wide association studies (EWAS) (Li *et al.* 2019, Battram *et al.* 2022). Extensively employed in consortia projects, such as The Cancer Genome Atlas (TCGA), over 80 000 HumanMethylation450 (HM450) samples (Maden *et al.* 2021) and a comparable number of EPIC array methylation profiles have accumulated in the Gene Expression Omnibus (GEO). Compared to sequencing-based methods, Infinium arrays offer cost-effectiveness, high quantitative resolution (Zhou *et al.* 2019b), ease of use, and the ability to accommodate a wide range of DNA inputs (Lee *et al.* 2024). Their high throughput capabilities have accelerated clinical applications in areas such as cancer diagnosis (Capper *et al.* 2018), liquid biopsies (Li *et al.* 2022), and forensic science (Mannens *et al.* 2022). More recently, this technology has supported the creation of an extensive pan-mammalian DNA methylome atlas (Arneson *et al.* 2022, Ding *et al.* 2023, Haghani *et al.* 2023).

The arrays' probe naming system (i.e. "cg" number), beginning with the Infinium HumanMethylation27 BeadChip (HM27), has been a cornerstone for cross-referencing probes with unique CpG sites within the genome. Each cg number corresponds to a unique 122-mer sequence centered on the target cytosine-guanine dinucleotide (CpG site), with array probes designed against these sequences. Originally, the Infinium arrays featured a one-to-one design—one probe set per 122-mer sequence—enabling a mapping to the human genome and facilitating cross-referencing 122-mer IDs, or cg numbers, with genomic CpG locations. While both the 122-mer and probe sequences are susceptible to nonunique mapping, this referencing method is common in EWAS literature (Xiong *et al.* 2020, Min *et al.* 2021, Battram *et al.* 2022, Hawe *et al.* 2022) and has provided a convenient albeit imperfect system (e.g. from HM27, HM450 to EPIC, cg number-based probes can be directly compared) for indexing probe sequences or CpG sites within a genome assembly.

The main limitation of the original cg number system arises from its non-specificity—a single cg number could correspond to multiple probe designs targeting the same 122-mer sequence. Additionally, this framework did not allow the inclusion of multiple replicate probes (Bibikova *et al.* 2009), which would enhance the robustness of measurements. With the advent of newer Infinium array generations, like the EPIC version 2 (EPICv2) (Kaur *et al.* 2023, Noguera-Castells *et al.* 2023) and other non-human arrays (Arneson *et al.* 2022, Zhou *et al.* 2022), a more precise naming system was introduced. This new system retains the cg number as a prefix but

Received: 19 March 2024; **Revised:** 7 June 2024; **Editorial Decision:** 20 June 2024; **Accepted:** 3 July 2024

adds additional information to distinguish between probes, accounting for Infinium chemistry, strand orientation, and replicate indices (Zhou *et al.* 2022). However, while methodologically sound, introducing additional probe details can impede the integration of newly generated methylation data with legacy datasets using the antiquated probe naming system.

Moreover, the static probe content selection in Infinium technology reflects the evolving understanding of methylation biology (Zhou *et al.* 2017). Each array generation—HM27, HM450, EPIC, and EPICv2—has refined probe content to represent better emerging biological insights, like gene body methylation (Yang *et al.* 2014) and cis-regulatory element methylation (Neiman *et al.* 2017). However, integrating legacy data generated on older platforms may introduce missing probes, which remains technically challenging, especially for applications like computing epigenetic clocks (Horvath 2013) and cancer classification models (Capper *et al.* 2018), which require specific CpGs in a model. Although data imputation strategies can help fill missing values within samples, many methods, such as matrix factorization (Mazumder *et al.* 2010), cannot accommodate the complete missingness of a specific probe in the query dataset. How to continue leveraging the legacy data and predictive models on the ever-evolving Infinium platforms has become a pressing technical need.

To respond to this need, we introduce methylation LiftOver (mLiftOver), a tool designed to harmonize Infinium data efficiently across platforms, including the EPICv2 array. mLiftOver, handles probe ID conversion, replicate probe measurement resolution, and missing data imputation (Fig. 1A). It is compatible with the R/Bioconductor ecosystem and enables data conversion with varying stringency levels. We demonstrate its utility by applying it to public EPICv2 datasets, showcasing its high performance and utility in bridging different Infinium platforms.

2 Materials and methods

mLiftOver, developed in R, is a feature in the SeSAMe package (Zhou *et al.* 2018) and leverages the *ExperimentHub* (Pasolli *et al.* 2017) and the *sesameData* packages to organize empirical data for its operation (Fig. 1A). This tool can convert inputs of generic data types: Probe IDs as a string list, DNA methylation levels (beta values) as numerical matrices, and signal intensities as "SeSAMe::SigDF" objects. mLiftOver can also translate data to and from new and previous Infinium platforms. The tool generically identifies replicate probes as those sharing the same cg number prefix but differing in other design aspects, such as strand specification and Infinium chemistry (Fig. 1B). When integrating data between platforms with and without these suffixes, mLiftOver offers two data aggregation strategies: averaging beta values across replicates or selecting the replicate with the most significant signal detection, informed by detection *P*-values. The latter method can exclude probes with potential design issues, as indicated by the mask column within the "SigDF" object. When converting platforms without replicates to platforms with replicates, the same readings will be assigned to different replicates. *mLiftOver* is compatible with all existing Infinium platforms, including HM27, HM450, EPIC, EPICv2, and the Methylation Screening Array (MSA) (Goldberg *et al.* 2024). It also facilitates the conversion of raw signals stored as "SigDF" class objects, enabling integrated analyses such as copy number variation studies. Beyond signal conversion based on probe IDs, *mLiftOver* can incorporate empirical benchmarks from analyses where two platforms have profiled identical cell lines to filter platform-specific biases, thus enhancing data translation fidelity. We should note that mLiftOver does not address batch effects, so care should still be taken when designing and executing each experiment.

mLiftOver integrates publicly available datasets to facilitate the back-conversion of EPICv2 data to its antecedent platforms, EPIC and HM450. This reverse conversion process involves three steps: (i) translating probe IDs, (ii) filtering platform-specific biases, and (iii) imputing missing data by mapping the sample using an empirical nearest neighbor approach to samples within our comprehensive DNA methylome repository. By aligning with the closest matching tissue type, mLiftOver fills in gaps without relying on methylation levels from other samples in the dataset, thereby enabling single-sample dataset operations. We have conducted extensive analyses on 10 631 EPIC and 10 726 HM450 samples to establish a robust imputation baseline when either EPIC or HM450 is the target platform (Supplementary Table S1). This baseline collection of datasets spans 20 and 19 tissue types for HM450 and EPIC datasets, respectively, with blood as a focal tissue due to its prevalence in EWAS studies (Supplementary Fig. S1A). Additionally, we calculated the variance of beta values for each CpG site within the target tissue type to gauge the confidence of imputation for probes completely absent from the original array. Supplementary Figure S1A shows the standard deviation distribution by tissue type and assay platforms. These variance metrics are critical as they can serve as filters to eliminate methylation influences stemming from unaccounted variables, such as age. The imputation reference data is housed within the sesameData package, accessible via the "sesameDataGet" function. When mLiftOver detects missing data, it substitutes these gaps with the median methylation value for the respective tissue type. This tissue type is either deduced algorithmically or specified by the user, ensuring the replaced values align with the most probable biological context.

3 Results

To show the performance of mLiftOver, we benchmarked the accuracy of converted probe-level methylation readings using the EPIC and EPICv2 data profiling the same cell lines (GM12878, K562, and LNCaP) (Kaur *et al.* 2023). We first compared native EPIC data and converted data from EPICv2, then native EPICv2 data and harmonized data from EPIC, all profiling the same cell line (GM12878 or HCT116) (Fig. 1C). Conversions in both directions highly correlate with the native measurements from the target platform (Spearman $\rho = 0.988$) (Fig. 1C, first panel; Supplementary Fig. S1C). EPIC to EPICv2 conversion yields more probes due to the replicate probes with the same cg number prefix in EPICv2. Next, compared to native EPIC data, both replicate probe aggregation methods yielded similarly high measurement accuracy on 3481 probes with design replicates in EPICv2, with the methylation level averaging method slightly surpassing the detection *P*-value method (Fig. 1C, second panel; Supplementary Fig. S1D). For EPICv2 to EPIC conversion, we further considered data imputation. The imputed values alone also highly correlated with the native EPIC data

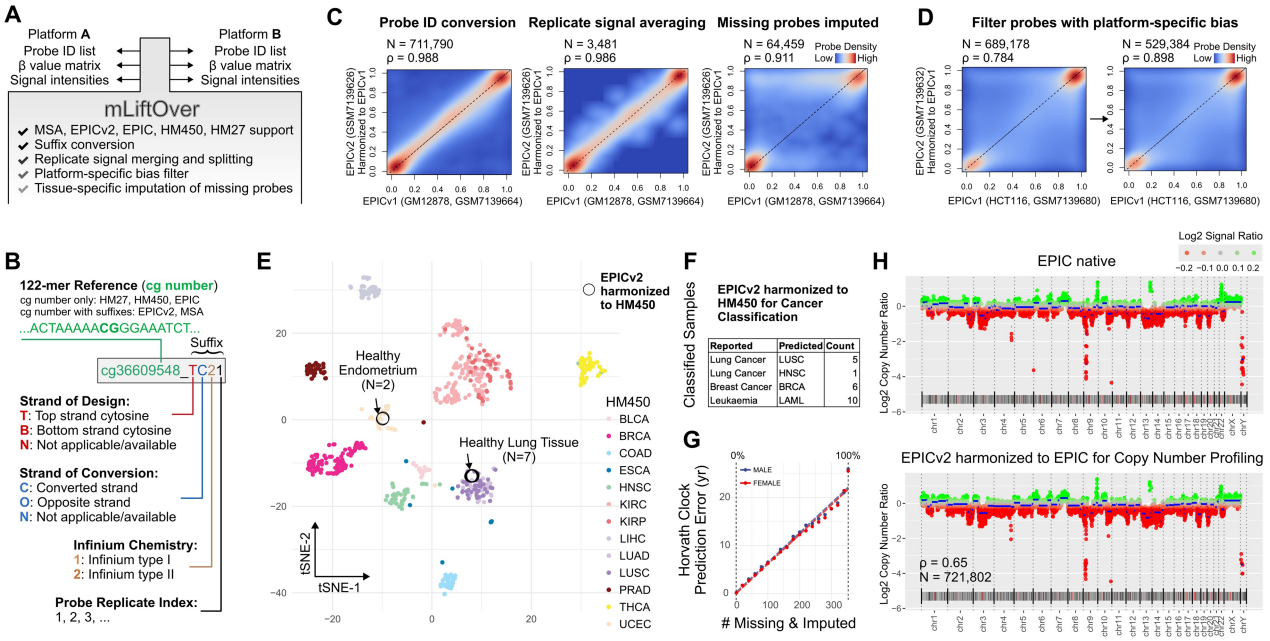

Figure 1. mLiftOver harmonizes Infinium DNA methylation BeadChip data across array platforms. (A) Schematic illustration of the core features and workflow of mLiftOver from data input to harmonized output. (B) Depiction of the probe naming convention employed in the EPICv2 and MSA arrays. (C) The accuracy of mLiftOver was evaluated using the GM12878 cell line data, contrasting measurements from EPICv1 and EPICv2. The panel is divided into three sub-panels, demonstrating (i) direct probe ID translation, (ii) signal averaging across replicates, and (iii) imputation of missing probe readings (excluding those with methylation level standard deviation >0.08). Spearman's correlation coefficients are displayed atop each subpanel, with all correlations being significant (*P*-value <1E-6). (D) Removal of platform-specific biases (tested on a pair of HCT116 cell line data that did not participate in the platform-specific bias analysis), *P*-value <1E-6. (E) Illustrates the integration process of mLiftOver for primary healthy tissue data and TCGA tumor-adjacent normal tissue data, showcasing its utility in harmonizing diverse datasets for tissue classification. (F) Demonstrates the application of cancer classification models, initially trained on HM450 data using a random forest framework, to primary tumor datasets harmonized from EPICv2 data through mLiftOver. (G) Plot relating the number of missing probes and prediction error of Horvath's pan-tissue clock, stratified by sex. (H) Compares copy number variation profiles obtained from native EPIC data and profiles harmonized from EPICv2 data, showing the consistency of mLiftOver in signal data conversion.

(Spearman $\rho = 0.82$), albeit lower than in the probe sets of direct probe conversion (Supplementary Fig. S1E), but higher than alternative imputation strategies based on genomic neighbors (Supplementary Fig. S1F and G). The Spearman's correlation remains at 0.977 for converted measurements and imputed values combined (Supplementary Fig. S1H). Filtering out 86 678 probes with higher methylation variation (SD > 0.08) in the public datasets reduces the number of imputed readings but brings the overall correlation to 0.91 (Fig. 1C, third panel). Lastly, we tested the filtering of platform-specific biases (Fig. 1D). We first examined five experiment pairs on three cell lines (GM12878, K562, and LNCaP). We defined a set of high-confidence mapping as those with methylation levels whose differences were no greater than 0.05 in four experiment pairs (see "Data Availability"). This yielded a mapping of 542 491 EPICv2 probes with 539 513 EPIC probes. mLiftOver then uses this mapping to convert unpaired EPIC and EPICv2 experiments on the HCT116 cell lines grown from different labs (Kaur *et al.* 2023). The conversion with the empirical filter yielded a slightly higher correlation (0.898 versus 0.784) with the native data than without filtering platform-specific bias (Fig. 1D).

To demonstrate the utility of mLiftOver in integrating Infinium data across multiple platforms (example function calls in Supplementary Fig. S2A), we applied it to integrate EPICv2 and HM450 data that profiled primary healthy tissue samples. We downloaded two healthy endometrium tissue methylomes and seven lung tissue methylomes (Noguera-Castells *et al.* 2023). We co-clustered the mLiftOver-converted methylomes with HM450 datasets of tumor-adjacent normal tissues from TCGA. As shown in Fig. 1E, the EPICv2-originated datasets correctly cluster with the corresponding lung and endometrium tissue samples. This suggests that mLiftOver faithfully maintained these biological samples' epigenetic identities.

Next, we evaluated whether predictive models trained on HM450 data can be used on mLiftOver-harmonized methylomes. We downloaded 22 primary tissue methylomes of the lung, breast cancer, and leukemia for cancer classification (Noguera-Castells *et al.* 2023). We applied a random forest classifier trained on 33 TCGA cancer types (Fig. 1F). The HM450-based classifier accurately predicts the cancer types of these methylomes except one, leading to an accuracy of 95%. We further evaluated the robustness of Horvath clock age (Horvath 2013) prediction by the degree of missing data imputation. As expected, data missingness is associated with loss of clock accuracy (Fig. 1G). Imputing 20 of 353 features led to a deviation of 1.71 years (Supplementary Fig. S2B).

Lastly, we tested the functionality of mLiftOver in converting signal intensities. Infinium array signal intensities are extensively used in discovering copy number aberrations. We benchmarked this functionality on EPIC and EPICv2 datasets profiling the K562 cell lines, a leukemia cell line associated with a characteristic copy number gain at chromosome 22 and loss of chromosome 9p (Zhou *et al.* 2019a). As expected, mLiftOver can produce consistent copy number profiles from EPICv1 native and EPICv2-harmonized data, capturing this hallmark structural variation (Fig. 1H).

Collectively, we demonstrate that mLiftOver enabled the integration of recent Infinium data with legacy data and allowed for legacy predictive models to be continuously used on data from updated platforms.

4 Discussion

The Infinium DNA methylation BeadChip has evolved significantly since its inception, progressing from the HM450 to the EPIC and EPICv2 array. While later versions often preserve a substantial portion of probes from a previous version, challenges persist in predictive modeling or longitudinal studies, where comparative analyses with historical data and models require identical probe IDs. This study introduces a user-friendly tool to streamline data harmonization across three dimensions: probe names, β values (methylation levels), and signal intensities.

The direction of platform harmonization (e.g. HM450 to EPIC or EPIC to HM450) should be guided by the analysis goal(s), such as the need for certain probe readings by a prediction model or the error tolerance level. Additionally, one should prioritize minimizing data imputation operations based on cohort composition [e.g. platform(s) used] and the number of platform-specific probes between the two platforms (see Supplementary Fig. S2C for probe overlap between existing platforms).

The necessity for imputing missing probe readings has arisen with the introduction of new probes and the removal of others in the newer platform iterations. Our tool, *mLiftOver*, addresses this need by harnessing publicly available data, primarily focusing on tissue-specific differences, which have been identified as principal influencers of DNA methylation patterns in various studies, including our own (Zhou *et al.* 2022, Ding *et al.* 2023). However, we acknowledge that other factors, such as age, sex, cellular malignancy, and mitotic history, have not been incorporated into our model. Moreover, our approach only supports target platforms with enough available data, and tissues with uncharacterized methylomes are absent from our reference database, posing a potential limitation. One possible solution is to utilize the methylation correlation structure, for instance, inferring methylation levels in genomic proximity, to aid in imputing missing data. This approach could exploit co-methylated regions identified in comprehensive genome-wide methylome analyses (Sofer *et al.* 2013). The feasibility of imputation could inform the design of future Infinium arrays. It is important to note that while DNA methylation levels can be imputed, the imputation of signal intensities for absent probes is not yet supported, potentially impacting the analysis of copy number alterations in converted versus native datasets. Nonetheless, mLiftOver addresses the problem of probes missing completely between array platforms by utilizing a large database of publicly available DNA methylation array data across multiple tissues and leveraging the variability in methylation levels to assess the imputation accuracy. Our imputation solution for entirely missing probe values can be helpful for predictive models requiring specific probe values, where the alternative would be a missing value. In sum, mLiftOver provides user-friendly functionality for projects seeking to analyze DNA methylation data using different versions of Infinium arrays.

Acknowledgements

The authors thank the reviewers for their constructive feedback.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

W.Z. received BeadChips from Illumina Inc. for research.

Funding

This work was supported by the National Institute of Health/National Institute of General Medical Sciences [R35-GM146978].

References

Aref-Eshghi E, *et al.* Evaluation of DNA methylation episignatures for diagnosis and phenotype correlations in 42 mendelian neurodevelopmental disorders. *Am. J. Hum. Genet.* 2020;**106**:356–70.

Arneson A, Haghani A, Thompson MJ *et al.* A mammalian methylation array for profiling methylation levels at conserved sequences. *Nat Commun* 2022;**13**:783.

Battram T, Yousefi P, Crawford G *et al.* The EWAS catalog: a database of epigenome-wide association studies. *Wellcome Open Res* 2022;**7**:41.

Bibikova M, *et al.* High-throughput DNA methylation profiling using universal bead arrays. *Genome Res.* 2006;**16**:383–93.

Bibikova M, Le J, Barnes B *et al.* Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics* 2009;**1**:177–200.

Capper D, Jones DTW, Sill M *et al.* DNA methylation-based classification of Central nervous system tumours. *Nature* 2018;**555**:469–74.

Ding W, Kaur D, Horvath S, *et al.* Comparative epigenome analysis using infinium DNA methylation BeadChips. *Brief Bioinformatics* 2023;**24**:bbac617.

Goldberg DC, Cloud C, Lee SM *et al.* MSA: scalable DNA methylation screening BeadChip for high-throughput trait association studies. bioRxiv, 2024, preprint: not peer reviewed. https://doi.org/10.1101/2024.05.17.594606.

Haghani A, Li CZ, Robeck TR *et al.* DNA methylation networks underlying mammalian traits. *Science* 2023;**381**:eabq5693.

Hawe JS, Wilson R, Schmid KT *et al*; MuTHER Consortium. Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. *Nat Genet* 2022;**54**:18–29.

Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013;**14**:R115.

Kaur D, Lee SM, Goldberg D *et al.* Comprehensive evaluation of the infinium human MethylationEPIC v2 BeadChip. *Epigenetics Commun* 2023;**3**:6.

Lee SM, Loo CE, Prasasya RD *et al.* Low-input and single-cell methods for Infinium DNA methylation BeadChips. *Nucleic Acids Res* 2024;**52**:e38.

Li M, *et al.* EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res.* 2019;**47**:D983–D988.

Li H-T, Xu L, Weisenberger DJ *et al.* Characterizing DNA methylation signatures of retinoblastoma using aqueous humor liquid biopsy. *Nat Commun* 2022;**13**:5523.

Maden SK, Thompson RF, Hansen KD *et al.* Human methylome variation across Infinium 450K data on the gene expression omnibus. *NAR Genom Bioinform* 2021;**3**:lqab025.

Mannens MMAM, Lombardi MP, Alders M *et al.* Further introduction of DNA methylation (DNAm) arrays in regular diagnostics. *Front Genet* 2022;**13**:831452.

Mazumder R, Hastie T, Tibshirani R *et al.* Spectral regularization algorithms for learning large incomplete matrices. *J Mach Learn Res* 2010;**11**:2287–322.

Min JL, Hemani G, Hannon E, *et al*; BIOS Consortium. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet* 2021;**53**:1311–21.

Neiman D, Moss J, Hecht M *et al.* Islet cells share promoter hypomethylation independently of expression, but exhibit cell-type-specific methylation in enhancers. *Proc Natl Acad Sci USA* 2017; **114**:13525–30.

Noguera-Castells A, García-Prieto CA, Álvarez-Errico D *et al.* Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. *Epigenetics* 2023;**18**:2185742.

Pasolli E, Schiffer L, Manghi P *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* 2017;**14**:1023–4.

Sofer T, Schifano ED, Hoppin JA *et al.* A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics* 2013;**29**:2884–91.

Thompson M, *et al.* Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. *NPJ Genom. Med.* 2022;**7**:50.

Xiong Z, Li M, Yang F *et al.* EWAS data hub: a resource of DNA methylation array data and metadata. *Nucleic Acids Res* 2020; **48**:D890–D895.

Yang X, Han H, De Carvalho DD *et al.* Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* 2014;**26**:577–90.

Zhou W, Laird PW, Shen H *et al.* Comprehensive characterization, annotation and innovative use of infinium DNA methylation BeadChip probes. *Nucleic Acids Res* 2017;**45**:e22.

Zhou W, Triche TJ, Laird PW *et al.* SeSAMe: reducing artifactual detection of DNA methylation by infinium BeadChips in genomic deletions. *Nucleic Acids Res* 2018;**46**:e123.

Zhou B, Ho SS, Greer SU *et al.* Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res* 2019a;**29**:472–84.

Zhou L, Ng HK, Drautz-Moses DI *et al.* Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. *Sci Rep* 2019b; **9**:10383.

Zhou W, Hinoue T, Barnes B *et al.* DNA methylation dynamics and dysregulation delineated by high-throughput profiling in the mouse. *Cell Genom* 2022;**2**:100144.

Bioinformatics, 2024, 40, 1–5
https://doi.org/10.1093/bioinformatics/btae423
Applications Note