

FUSE: data-driven functional segmentation of DNA methylation data

Susanna Holmström^{1, }, Antti Häkkinen², Kari Lavikka^{1, }, Giovanni Marchi^{1, }, Sampsa Hautaniemi^{1, },
Alexandra Lahtinen^{1,*, }

¹Research Program in Systems Oncology, Research Programs Unit, Faculty of Medicine, University of Helsinki, Helsinki, 00014, Finland

²Computational Health Informatics Program, Boston Children's Hospital, Harvard Medical School, Boston, MA, 02215, United States

*Corresponding author. Research Program in Systems Oncology, Research Programs Unit, Faculty of Medicine, Haartmaninkatu 8, P.O. Box 63, FI-00014 University of Helsinki, Helsinki, Finland. E-mail: alexandra.lahtinen@helsinki.fi.

Associate Editor: Björn Grüning

Abstract

Summary: DNA methylation (DNAm) of neighbouring CpG sites is highly correlated, making DNAm function in terms of blocks. DNAm patterns and functionality are linked to both chromatin structure of DNA and gene regulation. Defining biologically meaningful DNA methylation blocks from whole-genome bisulfite sequencing (WGBS) data remains challenging, as most existing methods rely on fixed genomic windows rather than the observed methylation pattern. We present FUSE, a data-driven segmentation method that captures intrinsic methylation segments directly from WGBS data by jointly analyzing multiple samples. FUSE identifies spatially homogeneous methylation blocks shared across the input cohort while allowing different methylation states across samples. Applied to 61 WGBS samples from the ENCODE database, FUSE identified segments which overlap significantly with promoters, enhancers, and repetitive elements. FUSE was able to recover the true segment breakpoints in synthetic data with high sensitivity under increased levels of noise. As such, FUSE facilitates post hoc methylation analyses by aggregating coherent CpG sites into candidate segments for downstream differential methylation testing or other comparative studies.

Availability and implementation: FUSE is implemented as an R-package *methFuse*, available at <https://github.com/holmsusa/methFuse> and <https://cran.r-project.org/package=methFuse>. A GenomeSpy visualization of the data is available at https://csbi.ltdk.helsinki.fi/p/fuse_encode_gs/

Introduction

DNA methylation (DNAm) is an epigenetic mechanism involved in gene expression, cell differentiation, genomic imprinting, and repression of retroviral elements (Smith et al. 2025). In animal genomes, DNAm predominantly occurs as 5-methylcytosine at CpG dinucleotides (Feng et al. 2010). Aberrant methylation patterns are strongly linked to cancer progression and drug resistance (Song et al. 2025).

CpG methylation levels are shaped by local CpG density and nearby methylation states (Haerter et al. 2014). Adjacent CpG sites show strong correlations in methylation levels, particularly in CpG-dense regions, and the DNAm pattern is stably maintained through cell division (Affinito et al. 2020). Thus, methylation acts at the level of blocks rather than individual sites, making segment-level analyses more informative than single-CpG approaches.

The most prominent block-like structures are CpG islands (CGIs), commonly found in promoters of housekeeping genes (Gardiner-Garden and Frommer 1987) and linked to chromatin features such as H3K4me3 (Thomson et al. 2010). Promoters and enhancers are regulated by chromatin structure and spatial DNA conformation (Whalen et al. 2016), as well as by DNA methylation, which therefore encodes spatial and regulatory information.

Existing segment-based DNAm methods focus largely on differential methylation and cell-type specific marks, but a method to extract blocks capturing the global intrinsic patterns is lacking. Instead, studies often rely on fixed windows or predefined annotations as proxies for functional regions. To address this gap, we introduce FUSE, a data-driven segmentation method that extracts intrinsic methylation blocks from whole-genome bisulfite sequencing (WGBS) data by jointly analyzing multiple samples.

Received: 24 November 2025. Revised: 23 February 2026. Accepted: 16 April 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

FUSE identifies blocks of CpG sites that capture spatially coherent methylation patterns while allowing variation in methylation states across cell types. The resulting segmentation provides a cohort-level representation of methylation domains that can be explored at multiple resolutions and serves as a framework for downstream analyses, such as segment-level differential methylation testing or other comparative studies across samples. We show that coherent FUSE segments obtained using Bayesian Information Criterion (BIC)-based model selection show expected significant enrichment for enhancers, promoters, H3K4me3, and repetitive elements.

Implementation

Modeling methylation

Bisulfite treatment of DNA converts unmethylated cytosines to thymines, while methylated ones are preserved. The number of methylated counts at a CpG site can be modeled using a binomial distribution, where the read depth is the number of independent trials. Let $\mathbf{T} \in \mathbb{N}^{m \times n}$ and $\mathbf{C} \in \mathbb{N}^{m \times n}$ denote matrices of unmethylated and methylated read counts, respectively, with m CpG sites and n samples.

For each site $i = 1, \dots, m$ and sample $j = 1, \dots, n$, we define the methylation counts $X_{ij} \in \mathbb{N}$ as

$$X_{ij} \sim \text{Binomial}(t_{ij} + c_{ij}, p_{ij}), \quad (1)$$

where t_{ij} and c_{ij} are the unmethylated and methylated read counts, respectively, and $p_{ij} \in [0, 1]$ is the methylation rate of site i in sample j .

The maximum likelihood estimate (MLE) of the methylation rate for site i in sample j is

$$\hat{p}_{ij} = \frac{c_{ij}}{t_{ij} + c_{ij}}, \quad (2)$$

which corresponds to the site's β -value. For a subset of CpG sites $S \subseteq \{1, \dots, m\}$, the aggregated methylation rate for sample j is

$$\hat{p}_{S,j} = \frac{\sum_{i \in S} c_{ij}}{\sum_{i \in S} (t_{ij} + c_{ij})}. \quad (3)$$

The binomial likelihood for the observed counts in the subset S of sites is

$$\mathcal{L}_S(\{p_{S,j}\}) = \prod_{i \in S} \prod_{j=1}^n \binom{t_{ij} + c_{ij}}{c_{ij}} p_{S,j}^{c_{ij}} (1 - p_{S,j})^{t_{ij}}. \quad (4)$$

Evaluated at the MLE $\hat{p}_{S,j}$, the log-likelihood becomes

$$\log \hat{\mathcal{L}}_S = \sum_{i \in S} \sum_{j=1}^n [c_{ij} \log \hat{p}_{S,j} + t_{ij} \log (1 - \hat{p}_{S,j})]. \quad (5)$$

Hierarchical clustering

FUSE employs agglomerative hierarchical clustering to obtain a multiresolution segmentation of the genome. While hierarchical clustering does not guarantee a globally optimal solution, it avoids imposing directional transition assumptions or specific parametric constraints on segment lengths, as in commonly used hidden Markov models (HMMs) for genomic segmentation and yields segmentations at multiple resolutions in a computationally efficient manner.

An overview of the clustering is shown in Fig. 1a. The dissimilarity between two sets of sites $\mathcal{A}, \mathcal{B} \subseteq \{1, \dots, m\}$ is quantified by the negative log-likelihood ratio (nLLR):

$$\text{nLLR}_{\mathcal{A}\mathcal{B}} := -\log \hat{\mathcal{L}}_{\mathcal{A} \cup \mathcal{B}} + \log \hat{\mathcal{L}}_{\mathcal{A}} + \log \hat{\mathcal{L}}_{\mathcal{B}}, \quad (6)$$

where each log-likelihood is evaluated at the corresponding MLE $\hat{p}_{S,j}$ for each sample. To favor clustering of closely located CpG sites while penalizing long distances, a heavy-tailed genomic distance penalty following the Student's t -distribution with one degree of freedom is introduced.

The complete dissimilarity $d(\mathcal{A}, \mathcal{B})$ is defined as

$$d(\mathcal{A}, \mathcal{B}) = \text{nLLR}_{\mathcal{A}\mathcal{B}} + \begin{cases} -\log \left[T\left(\frac{\delta(\mathcal{A}, \mathcal{B})}{100}; 1\right) \right], & \text{if } \text{chr}(\mathcal{A}) = \text{chr}(\mathcal{B}), \\ \infty, & \text{otherwise.} \end{cases} \quad (7)$$

Here, $\delta(\mathcal{A}, \mathcal{B}) = \min_{a \in \mathcal{A}, b \in \mathcal{B}} |g_a - g_b|$, where g_a is the genomic coordinate of site a , denotes the shortest genomic distance between the two sets $\mathcal{A}$ and $\mathcal{B}$ and $T(x; 1)$ is the probability density function of the Student's t -distribution with one degree of freedom.

Clustering is performed agglomeratively producing a hierarchy of segments from fine to coarse resolution. The number of segments can be selected at the desired resolution, with Bayesian Information Criterion (BIC)-based selection (Schwarz 1978) implemented as the default in the R package and used throughout this manuscript. Segmentation is performed independently for each chromosome to ensure scalability to whole-genome WGBS datasets, and results are subsequently merged.

The R-based implementation of the algorithm outputs both a segment-level summary (including genomic coordinates, length, number of CpGs, average methylation across samples) and a segment-by-sample methylation matrix. This preserves within-segment variation across samples and enables downstream analyses, such as differential methylation testing, variability assessment, entropy-based characterization, or stratification by segment properties.

Coherently methylated segments

Regions with highly coherent methylation patterns are likely actively maintained and might be connected to biological functionality. To test this hypothesis, segments are divided into coherent and non-coherent groups.

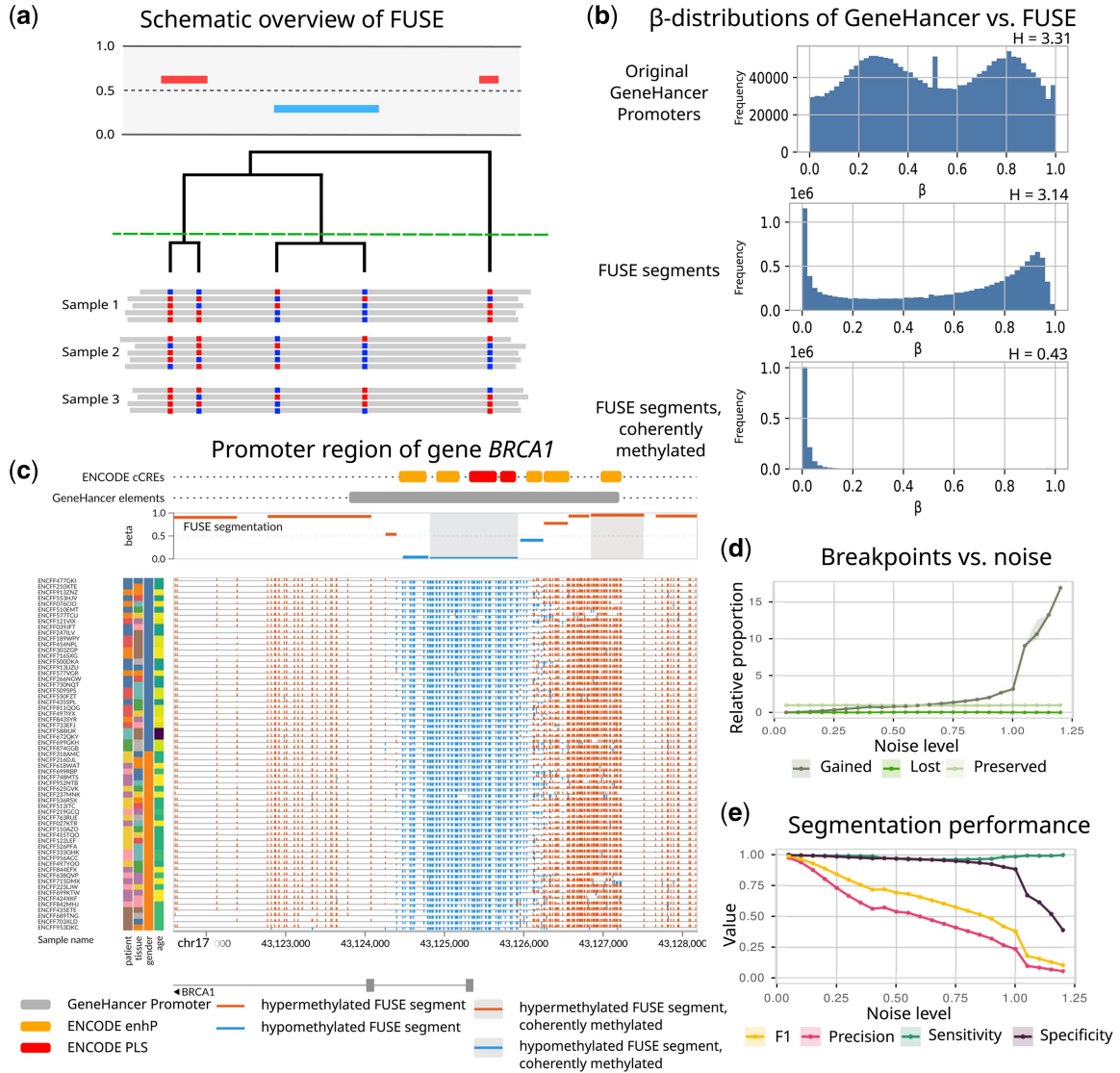


Figure 1 (a) Schematic overview of the FUSE-clustering. CpG sites are clustered hierarchically, and cutting the tree results in segments. (b and c) Results of the whole-genome segmentation of the ENCODE data set described in the text. (b) Comparison of beta distributions in promoter regions from GeneHancer. The upper plot shows the promoters as they are annotated, the middle plot shows all FUSE-segments overlapping with these promoters, and the lower plot shows all coherently methylated FUSE segments overlapping with these promoters. Entropy (H) decreases from top down, indicating a reduction in uncertainty within the data. (c) Snapshot of a GenomeSpy [Lavikka et al. 2024] WGBS visualization showing the promoter area of the gene *BRCA1*. While the promoters from GeneHancer (grey) and ENCODE (red) are not capturing the data patterns, the coherently methylated segments found by FUSE (blue and red with shadows) are. (d) The proportion of gained, lost, and preserved breakpoints between FUSE segments under different levels of noise. (e) The F1, precision, sensitivity, and specificity curves of FUSE segment breakpoints with respect to the ground truth.

To assess the internal coherence of each segment, we test whether all CpG sites in the segment share a common methylation level. The pattern over samples is not constrained and can be heterogeneous. For segment $R \subseteq \{1, \dots, m\}$ containing r CpG sites we compare two models:

- Common model M_0 : all sites in R share a common methylation rate $\hat{p}_R$, as given in Eq. 3.
- Separate model M_1 : each site $r_i \in R$ has its own methylation rate $\hat{p}_{r_i}$, as given in Eq. 2.

The corresponding log-likelihoods, $\log \hat{L}_R(M_0)$ and $\log \hat{L}_R(M_1)$ are computed using Eq. 5. Under the null hypothesis of a common

methylation level, the likelihood ratio test statistic approximately follows a χ^2 distribution with $r - 1$ degrees of freedom. Segments with no significant difference between the two models ($p > 0.05$) are considered *coherently methylated*. The coherence-flag is added to the segment-summary output of the R-implementation.

Code availability

FUSE is implemented as an R package, *methFuse*. The package is available at <https://cran.r-project.org/package=methFuse> and at <https://github.com/holmsusa/methFuse> together with examples and documentation.

Results

Enrichment of regulatory and repetitive elements among coherently methylated segments

To investigate whether biologically interesting regions co-occur with coherent and likely highly maintained methylation patterns, we evaluated the enrichment of regulatory and repetitive elements within coherently methylated segments. FUSE was applied to 61 WGBS samples from healthy donors obtained from the ENCODE portal (Luo et al. 2020) (Supplementary S1.1, available as [supplementary data](#) at *Bioinformatics* online). This resulted in 2931872 segments in chromosomes 1–22, of which 43539 were coherently methylated. Genomic annotations for candidate Cis-Regulatory Elements were retrieved from the ENCODE (ENCSR800VNX), promoter and enhancer regions from GeneHancer (Fishilevich et al. 2017), and repetitive elements from the UCSC Genome Browser RepeatMasker track (Kent et al. 2002) (Smit, Hubley & Green, RepeatMasker Open-4.0. 2013–2015 <http://www.repeatmasker.org>).

Overlap counts between coherently methylated FUSE segments and each annotation class were determined, and enrichment was assessed using Fisher's exact test with Benjamini-Hochberg correction. Significant enrichment ($FDR < 0.05$) was found for three of five ENCODE classes (PLS, pELS, and DNASE_H3K4me3), both GeneHancer categories (promoter and enhancer), and two of twenty-four repetitive element types (retroposon_sva and satellite_telo) (Supplementary S2.1, Table S3, available as [supplementary data](#) at *Bioinformatics* online). These results show that coherently methylated FUSE segments overlap known regulatory and repetitive regions.

Analysis of β -value distributions provided a descriptive characterization of the identified segments: as expected by construction, β -value entropy within coherently methylated segments was consistently lower than in the original data across all significantly enriched classes. This reflects the reduced variability that defines coherent segments. Promoter regions from GeneHancer are compared with genome-wide FUSE-segments in Fig. 1b and at the 5'-end of *BRCA1* in Fig. 1c, with analogous plots for other element classes in Figure S3, available as [supplementary data](#) at *Bioinformatics* online. An interactive GenomeSpy (Lavikka et al. 2024) visualization of the ENCODE WGBS samples, FUSE-segments, and the functional regions used in this study is available at https://csbi.ltdk.helsinki.fi/p/fuse_encode_gs/.

Validation of robustness shows high sensitivity in noisy data

To assess the robustness of FUSE against technical variation, we evaluated its ability to recover consistent methylation segments at increasing levels of synthetic noise. Since the segments identified by FUSE are designed to reflect the underlying biological methylation patterns, they are expected to remain largely invariant even when noise is introduced.

A baseline, noise-free dataset $D_0 = \{\mathbf{C}_0, \mathbf{T}_0\}$ and a collection of noisy data sets

$$\mathcal{D}_\eta = \{D_\eta^{(k)} = \{\mathbf{C}_\eta^{(k)}, \mathbf{T}_\eta^{(k)}\} | k = 1, 2, \dots, 100\} \quad (8)$$

of noise amplitudes $\eta = \{5, 10, \dots, 120\}$ % were constructed as described in [Supplementary 3.1](#), available as [supplementary data](#) at *Bioinformatics* online.

FUSE was independently applied to each $D_\eta^{(k)} \in \mathcal{D}_\eta$ across all noise levels, and the resulting segmentation was compared to the ground truth segments in D_0 to identify preserved, lost, and gained breakpoints. For each noise level, the relative proportions of these breakpoint categories were summarized in Fig. 1d. Performance metrics including sensitivity, specificity, precision, and the F1 score were computed, and their 95% confidence intervals were estimated using bootstrapping, see Fig. 1e.

FUSE maintained high sensitivity in detecting true breakpoints across all noise levels (Fig. 1d), indicating that the underlying methylation pattern could be recovered even under substantial noise. As the noise increased to over 100%, a spike in false positive breakpoints was observed (Fig. 1d). In a setting where the noise is 100% or more, most CpG sites have more noise than signal, and the segmentation becomes completely random. At noise levels between 60% and 100% there are already more gained breakpoints than original ones, meaning some over-segmentation of contiguous regions is happening. In practical applications, the hierarchical structure of FUSE allows exploration of multiple resolutions: segments can be analyzed individually or aggregated into broader regions for downstream biological analyses, ensuring flexibility in assessing regulatory domains of varying size.

Performance evaluation

Computational benchmarking on the 61-sample ENCODE WGBS dataset demonstrated that full-genome segmentation can be completed in approximately 61 minutes with peak memory usage of 15 GB when processed chromosome-wise (Supplementary S1.2, available as [supplementary data](#) at *Bioinformatics* online).

We compared FUSE against established tools [Bumphunter (Jaffe et al. 2012, Aryee et al. 2014), DSS (Wu et al. 2013, Feng et al. 2014, Wu et al. 2015, Park and Wu 2016), MethSeg (Akalın et al. 2012)] and a fixed-window baseline in terms of runtime, memory usage, and segmentation accuracy (Supplementary S3.2, available as [supplementary data](#) at *Bioinformatics* online). FUSE achieved the highest sensitivity in breakpoint recovery, while runtime and memory usage were comparable across tools.

Conclusions

In this study, we introduced FUSE, a data-driven segmentation framework for capturing spatially homogeneous methylation blocks in WGBS data. FUSE derives methylation segments over the sites directly from the observed DNAm pattern while allowing different methylation states over samples, providing a cohort-specific representation of the underlying co-methylated structure. The hierarchical clustering approach allows the recovery of segments at any resolution, enabling flexible exploration of methylation patterns across scales. As with any data-driven

consensus segmentation, the resulting segment boundaries reflect the biological context of the given study population. If sample-specific boundaries are of interest, FUSE can be applied to individual samples or subsets of samples to enable direct comparison of segment architectures across conditions.

We showed that segments with coherent methylation patterns, likely reflecting highly maintained methylation, coincide with known regulatory and repeated regions, which can be further investigated in the provided GenomeSpy visualization.

FUSE is implemented as an easy-to-use R package, *methFuse*, with optimized C-based core functions for computational efficiency. Beyond segmentation, FUSE can serve as a foundation for downstream applications, such as differential methylation analysis, integration with chromatin-state maps, or epigenetic biomarker discovery.

Future work may add more focus on the heterogeneous segments and characterize their biological characteristics. Another direction would be to extend FUSE to multi-omics contexts, allowing segmentation informed by both methylation and chromatin accessibility data, or to non-CpG methylation patterns in diverse organisms. Overall, FUSE provides a robust and interpretable framework for uncovering biologically meaningful methylation domains, advancing our ability to study epigenetic regulation at the genome scale.

Author contributions

Susanna Holmström (Data curation [lead], Formal analysis [lead], Investigation [lead], Software [equal], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]), Antti Häkkinen (Conceptualization [lead], Methodology [lead], Software [equal], Supervision [equal], Writing—review & editing [supporting]), Kari Lavikka (Visualization [supporting], Writing—review & editing [supporting]), Giovanni Marchi (Data curation [supporting], Writing—review & editing [supporting]), Sampsa Hautaniemi (Funding acquisition [lead], Project administration [lead], Resources [lead], Supervision [equal], Writing—review & editing [supporting]), and Alexandra Lahtinen (Funding acquisition [supporting], Supervision [equal], Validation [supporting], Writing—original draft [supporting], Writing—review & editing [supporting])

Supplementary material

[Supplementary material](#) is available at *Bioinformatics* online.

Conflict of interests

None declared.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation programme [965193] (DECIDER); the Sigrid Jusélius Foundation; and the iCANDOC Precision Cancer Medicine pilot. A.L. received funding from the Finnish Cancer Institute.

References

- Affinito O, Palumbo D, Fierro A *et al.* Nucleotide distance influences co-methylation between nearby cpG sites. *Genomics* 2020;**112**:144–50. <https://doi.org/10.1016/j.ygeno.2019.05.007>
- Akalin A, Kormaksson M, Li S *et al.* MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* 2012;**13**:R87. <https://doi.org/10.1186/gb-2012-13-10-r87>
- Aryee MJ, Jaffe AE, Corrada-Bravo H *et al.* Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014;**30**:1363–9. <https://doi.org/10.1093/bioinformatics/btu049>
- Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res* 2014;**42**:e69–e69. <https://doi.org/10.1093/nar/gku154>
- Feng S, Cokus SJ, Zhang X *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* 2010;**107**:8689–94. <https://doi.org/10.1073/pnas.1002720107>
- Fishilevich S, Nudel R, Rappaport N *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database 2017;**2017**:bax028. <https://doi.org/10.1093/database/bax028>
- Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol* 1987;**196**:261–82. [https://doi.org/10.1016/0022-2836\(87\)90689-9](https://doi.org/10.1016/0022-2836(87)90689-9)
- Haerter JO, Lökvist C, Dodd IB *et al.* Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states. *Nucleic Acids Res* 2014;**42**:2235–44. <https://doi.org/10.1093/nar/gkt1235>
- Jaffe AE, Murakami P, Lee H *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* 2012;**41**:200–9. <https://doi.org/10.1093/ije/dyr238>
- Kent WJ, Sugnet CW, Furey TS *et al.* The human genome browser at UCSC. *Genome Res* 2002;**12**:996–1006. <https://doi.org/10.1101/gr.229102>
- Lavikka K, Oikonen J, Li Y *et al.* Deciphering cancer genomes with genomespy: a grammar-based visualization toolkit. *Gigascience* 2024;**13**. <https://doi.org/10.1093/gigascience/giae040>
- Luo Y, Hitz BC, Gabdank I *et al.* New developments on the encyclopedia of DNA elements (encode) data portal. *Nucleic Acids Res* 2020;**48**:D882–D889. <http://dx.doi.org/10.1093/nar/gkz1062>
- Park Y, Wu H. Differential methylation analysis for bs-seq data under general experimental design. *Bioinformatics* 2016;**32**:1446–53. <http://dx.doi.org/10.1093/bioinformatics/btw026>
- Schwarz G. Estimating the dimension of a model. *Ann Statist* 1978;**6**:0090–5364. <http://dx.doi.org/10.1214/aos/1176344136>
- Smith ZD, Hetzel S, Meissner A *et al.* DNA methylation in mammalian development and disease. *Nat Rev Genet* 2025;**26**:7–30. <http://dx.doi.org/10.1038/s41576-024-00760-8>
- Song J, Yang P, Chen C *et al.* Targeting epigenetic regulators as a promising avenue to overcome cancer therapy resistance. *Signal Transduct Target Ther* 2025;**10**:219. <http://dx.doi.org/10.1038/s41392-025-02266-z>

- Thomson JP, Skene PJ, Selfridge J *et al.* CpG islands influence chromatin structure via the CpG-binding protein cfp1. *Nature* 2010;**464**:1082–6. <http://dx.doi.org/10.1038/nature08924>
- Whalen S, Truty RM, Pollard KS. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* 2016;**48**:488–96. <http://dx.doi.org/10.1038/ng.3539>
- Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics* 2013;**14**:232–43. <http://dx.doi.org/10.1093/biostatistics/kxs033>
- Wu H, Xu T, Feng H *et al.* Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res* 2015;**43**:e141. <http://dx.doi.org/10.1093/nar/gkv715>