nature communications

Article

https://doi.org/10.1038/s41467-026-70523-y

mist: a hierarchical Bayesian framework for detecting differential DNA methylation dynamics in single-cell data

Received: 21 April 2025

Accepted: 20 February 2026

Published online: 12 March 2026

Check for updates

Daoyu Duan[1], Wenjing Ma[2], Wen Tang[3], Hao Wu[4], Liangliang Zhang[1] ✉ & Hao Feng[3] ✉

Recent advancements in single-cell DNA methylation (scDNAm) sequencing technologies have enabled the profiling of epigenetic landscapes at unprecedented resolution, offering insights into cellular heterogeneity, differentiation and evolution. Trajectory inference, which orders cells along pseudotime, allows researchers to track genomics changes across continuous cell states and identify key loci exhibiting differential methylation. However, no methods currently exist to model methylation changes along pseudotime in scDNAm data. Here, we present a hierarchical Bayesian framework for scDNAm data analysis. Our method, named *mist* (methylation inference for single-cell along trajectory), models stage-specific biological variations, identifies genomic features with significant methylation changes along pseudotime, and performs Differential Methylation (DM) analysis across phenotypical groups. Simulations demonstrate its superior accuracy in detecting DM genes along pseudotime compared to existing methods. Applied to multi-omics datasets of mouse embryonic development and developing human brain, *mist* identifies key developmental regulators, whose methylation patterns align with lineage transitions. *mist* is publicly available as an R/Bioconductor package.

DNA methylation (DNAm) is a key epigenetic modification resulting from the addition of a methyl group to the fifth carbon position of cytosine, typically occurring within CpG dinucleotides where a cytosine is followed by a guanine on the same DNA strand. This modification plays a critical role in various biological processes. Extensive research has shown that DNAm regulates transcriptional activity[1], is stably inherited[2], correlates with aging through epigenetic clocks[3,4], is linked to stress responses[5], and acts as an important hallmark in cancer development[6–8]. These findings were made possible through advancements in methylation microarray and whole-genome bisulfite sequencing (WGBS) technologies and their associated tools[9–15].

Recent developments in single-cell DNA methylation (scDNAm) technologies have enabled profiling of the DNA methylome at single-cell resolution. Early studies used scDNAm to analyze embryonic stem cells (ESCs)[16], human embryos[17], and cancer tissues[18,19]. Large-scale scDNAm datasets are now being generated, providing unprecedented opportunities to study cellular heterogeneity and development using methylome data[20–24]. Concurrently, new computational methods have been developed to address key challenges in the analysis of scDNAm data, such as missing value imputation[25–27], dimensionality reduction[28,29], clustering[25,30], cell type annotation[31–33], and differential methylation calling[34].

The dynamic and revisable nature of DNA methylation during cell differentiation and changes in cell state[35–37], coupled with its regulatory effects on gene expression, make it highly valuable for capturing the continuity of cellular evolution through scDNAm data. Therefore, trajectory inference analysis, often referred to as cell ordering or lineage reconstruction in scRNA-seq literature, is a natural fit for

[1]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA. [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. [3]Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston, Houston, TX, USA. [4]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA. ✉e-mail: lxz716@case.edu; Hao.Feng@uth.tmc.edu

https://doi.org/10.1038/s41467-026-70523-y

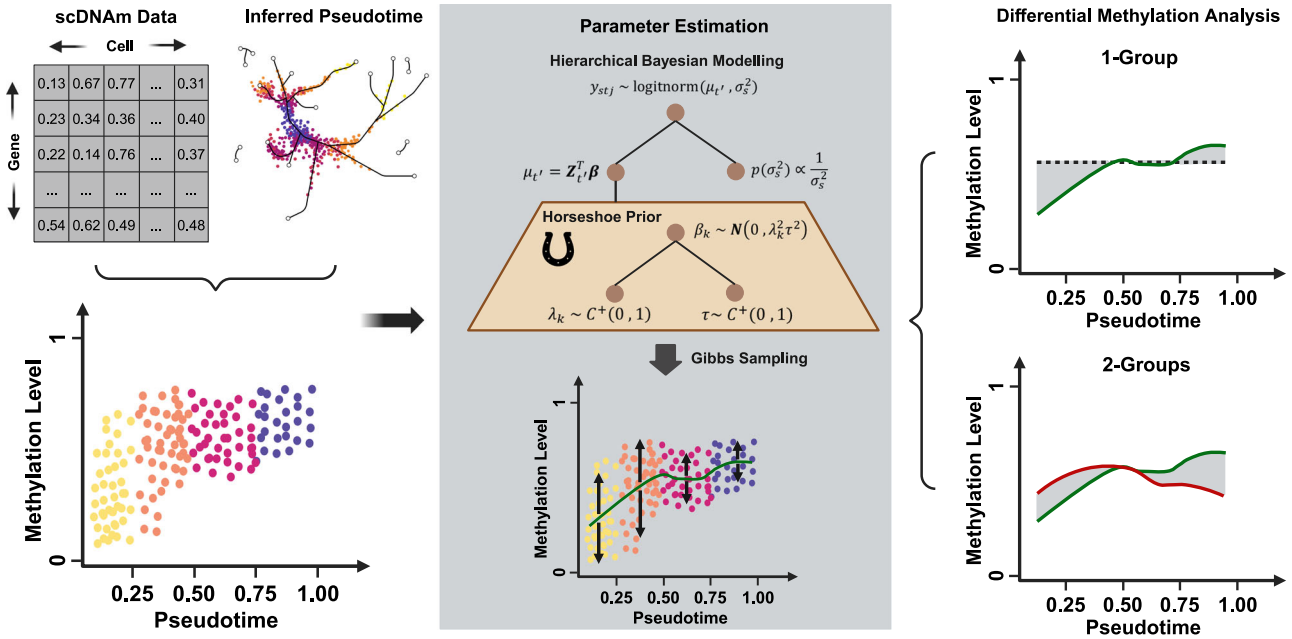

Fig. 1 | Schematic overview of *mist*. *mist* performs differential analysis for single-cell DNA methylation studies. It takes a scDNAm data matrix and a pseudotime vector as inputs, and employs a hierarchical Bayesian framework to estimate a fitted methylation trajectory. This trajectory forms the basis for differential analysis to identify genomic features exhibiting temporal methylation changes or differences between groups. Created in BioRender. Duan, D. (2026) https://BioRender.com/o33c513.

scDNAm analysis. Trajectory inference enables unbiased, epigenome-wide insight into the dynamic processes underlying cellular differentiation. By ordering cells along a "pseudotime", a temporal variable inferred from data, researchers can track changes in DNA methylation across continuous cell states, identifying key loci or regions that exhibit differential methylation over time. Despite the potential, no methods currently exist to model methylation changes along pseudotime, nor perform differential methylation (DM) analysis on pseudotime in scDNAm data.

Here, we present *mist* (methylation inference for single-cell along trajectory), a hierarchical Bayesian framework designed to model DNA methylation trajectories and conduct DM analysis along pseudotime in scDNAm data. The functionality of *mist* is three-fold: (1) it models methylation changes along pseudotime and estimates the developmental-stage-specific biological variations; (2) it performs DM analysis to identify genomic features showing drastic changes along pseudotime; (3) when there is more than one phenotypical group, it performs DM analysis across the groups to identify features that exhibit different methylation temporal patterns. *mist* employs a robust Bayesian approach, utilizing Gibbs sampling to estimate model parameters for both temporal changes and developmental-stage-specific methylation variations. *mist* is available as an R/Bioconductor package (https://bioconductor.org/packages/mist/), making it the first tool of its kind for scDNAm data analysis.

Results

Overview of *mist*

mist adopts a hierarchical Bayesian modeling framework and requires the known inputs of scDNAm data along with its estimated pseudotime (Fig. 1).

We first aggregated CpGs by genes, or other region-level features such as promoters. Within each gene, the integer number of sequencing reads and methylated reads are obtained, respectively. The ratio of the methylated read count over the total read count reflects the methylation level for each gene, and such data handling is a common practice in methylation sequencing data analysis[14,38,39]. Such region-level-feature based methylation matrix serves as the foundation input for our modeling. For pseudotime estimation, it is required as known input to *mist*, and can be easily obtained using an existing method of choice such as Monocle3[40], Slingshot[41], and STREAM[42], by feeding the algorithm with gene-by-cell methylation level matrix or other modalities such as gene expression. The estimated pseudotime values are normalized to [0, 1]. This ensures a unified pseudotime handling and interpretation regardless of pseudotime estimation methods adopted.

For each gene, *mist* fits a logit-normal model where the mean on the logit scale is parameterized as a fourth-degree polynomial in pseudotime, and the variance is allowed to vary across predefined developmental stages to capture stage-specific biological dispersion. A horseshoe prior is placed on the polynomial coefficients to adaptively shrink irrelevant terms, which is particularly useful given the sparsity of scDNAm data. Parameters are then estimated via Gibbs sampling. For differential methylation calling, *mist* computes the minimized area between fitted methylation trajectories, either between two groups or relative to a flat baseline in the one-group case. Genes are then ranked by this measure to identify those exhibiting the most pronounced methylation changes along pseudotime.

To assess how well our proposed Bayesian hierarchical model *mist* performs, we compared it with GAM and polynomial regression in identifying DM genes. For GAM, we used cubic regression splines with knots placed at pseudotime 0.25, 0.5, and 0.75. For the polynomial approach, we adopted a fourth-degree polynomial regression.

mist in parameter estimation

Figure 2 presents the overall assessment of parameter estimation accuracy for methylation dynamics using different methods. Figure 2A is an illustrative selection showcasing methylation levels along pseudotime for one selected gene, with each dot representing a cell. The *x*-axis indicates normalized pseudotime (0–1), and the *y*-axis represents methylation levels (0–1). The green curve, fitted using *mist*, captures the estimated methylation trajectory. The σ^2 values below the panel indicate the variability at each of the four stages. As a unique feature, *mist* not only fits the methylation data, but also estimates biological

https://doi.org/10.1038/s41467-026-70523-y

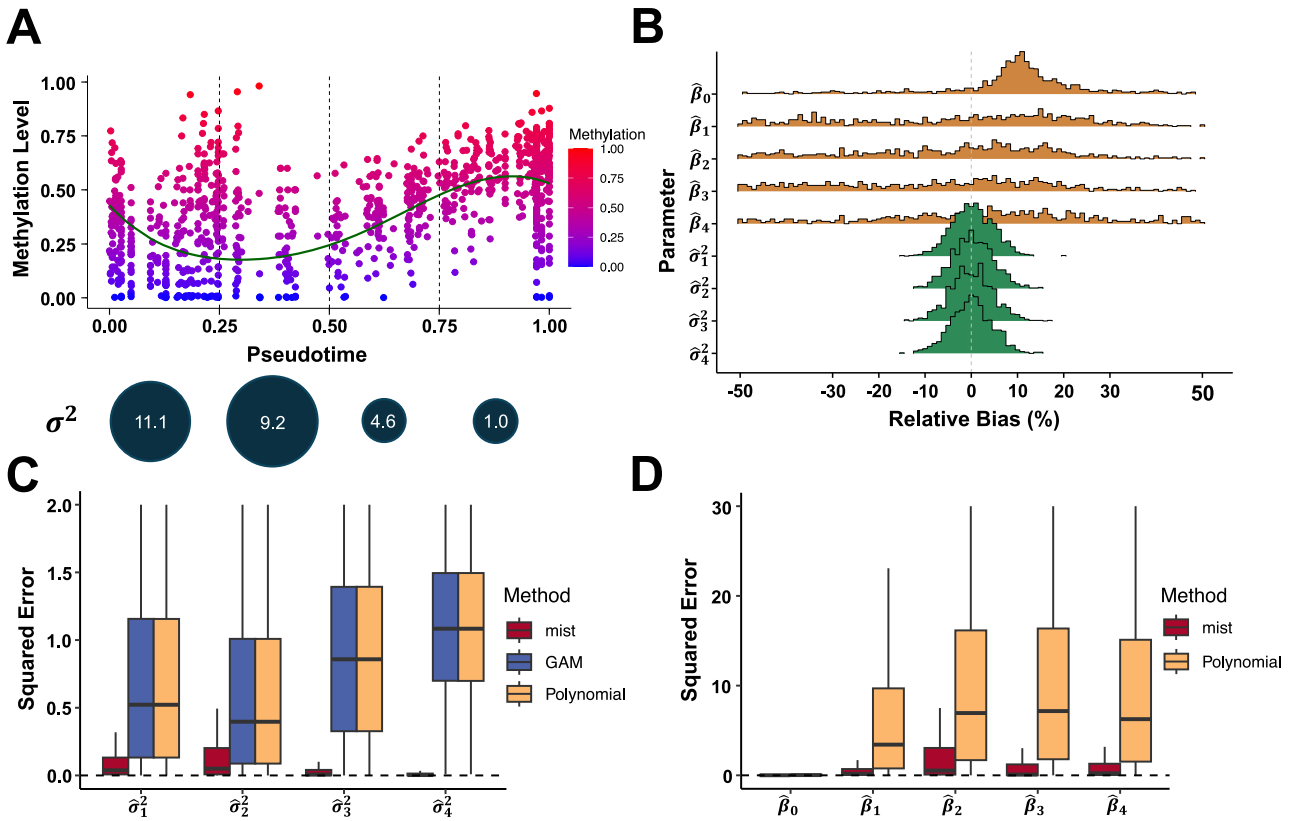

Fig. 2 | Methylation trajectory modeling for parameter estimation. **A** Methylation levels are shown along pseudotime for one selected gene, with each point colored by its methylation value. A curve fitted by the proposed Bayesian method captures the underlying methylation pattern over pseudotime. The developmental trajectory is partitioned into four intervals [0, 0.25), [0.25, 0.5), [0.5, 0.75), [0.75, 1], and the area of the circles below the plot reflects the estimated variance σ^2 at each stage. **B** Distributions of relative bias (%) for parameter estimates using the Bayesian method, truncated at −50% and 50% to exclude extreme outliers. **C**, **D** Boxplots illustrating the squared error for the estimated variance $\widehat{\sigma}_1^2$ to $\widehat{\sigma}_4^2$ and coefficients $\widehat{\beta}_0$ to $\widehat{\beta}_4$, respectively, under three modeling approaches. A total of $N = 10,000$ genes and $N = 4,000$ cells were simulated. Values beyond 2 for $\widehat{\sigma}^2$ and 30 for $\widehat{\beta}$ are truncated to avoid distortion by outliers. As GAM does not produce parameter estimates comparable to $\widehat{\beta}_0$ to $\widehat{\beta}_4$, it is not included in the β parameter comparison. The boxes represent the interquartile ranges, with the median at the center, and the whiskers extend to the limits of the distribution (excluding outliers). Source data are provided as a Source Data file.

variability at each stage along pseudotime. It achieves this by recognizing a stage-specific dispersion parameter in the Bayesian model. Figure 2B shows the bias for parameter estimates. After removing extreme outliers ($\geq \pm 50\%$), the biases for β parameters show a certain level of variability, but remain within a reasonable range. The bias for the dispersion parameter σ^2 is mostly within $\pm 10\%$, indicating highly accurate variance estimation.

We then compare parameter estimation performance by inspecting the squared errors of variance terms $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)$ and coefficient $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ parameters among three methods (Fig. 2C, D). The *mist* model consistently showed lower squared errors. In contrast, the GAM struggles in estimating variance parameters (Fig. 2C), because the constant variance assumption inherent to the Gaussian family is not well suited to capturing complex changes in variability along pseudotime. The polynomial model also showed higher errors, which suggests that the fixed polynomial form, lacking stage-specific dispersion, is less precise at capturing the true patterns compared to approaches that incorporate stage-specific biological dispersion. For the coefficient parameters (Fig. 2D), the Bayesian model again outperformed the polynomial approach, which had difficulty accurately estimating coefficients due to the model's limited flexibility. Overall, *mist* captures methylation trajectories more suitably than GAM or polynomial models. Its shrinkage effect, hierarchical structure, and flexible representation of stage-specific variability make it an improved alternative for modeling varying methylation patterns over pseudotime.

mist in DM gene detection

We assessed the *mist*'s performance in identifying DM genes in both 1-group and 2-group scenarios. Its results were benchmarked with those of GAM and polynomial regression. In the 1-group simulation (Fig. 3A–D), the Receiver Operating Characteristic (ROC) curves (Fig. 3A) illustrate that *mist* outperforms GAM and polynomial regression by a margin, particularly at lower false-positive thresholds (1 − specificity < 0.2). The precision-recall curves in Fig. 3B further support this, showing that *mist* maintains higher precision and recall values compared to the alternatives.

The True Discovery Rate (TDR) plot (Fig. 3C), defined as the proportion of true DM genes among the top-ranked genes, provides a practical measure of a method's accuracy among its top-ranking biomarkers. In clinical applications, researchers prioritize the top-ranking biomarkers for further validation, so a high TDR reflects the method's effectiveness in providing reliable biomarkers for downstream studies. Here, genes are ranked by the minimized area differences defined in Equations (3) and (5) as estimated by *mist*. For the GAM and linear models, genes are ranked by the adjusted p values obtained from the likelihood ratio test and ANOVA, respectively. The plot shows that *mist* consistently outperforms existing methods, achieving TDR values above 0.75 at various cutoffs, whereas GAM and polynomial regression struggle to exceed 0.5. The heatmap in Fig. 3D shows the robustness of *mist* across varying numbers of total cells and top-ranking genes in simulation. Notably, *mist*'s TDR increases with larger cell counts, a trend that is less evident for GAM and polynomial regression. Similarly,

https://doi.org/10.1038/s41467-026-70523-y

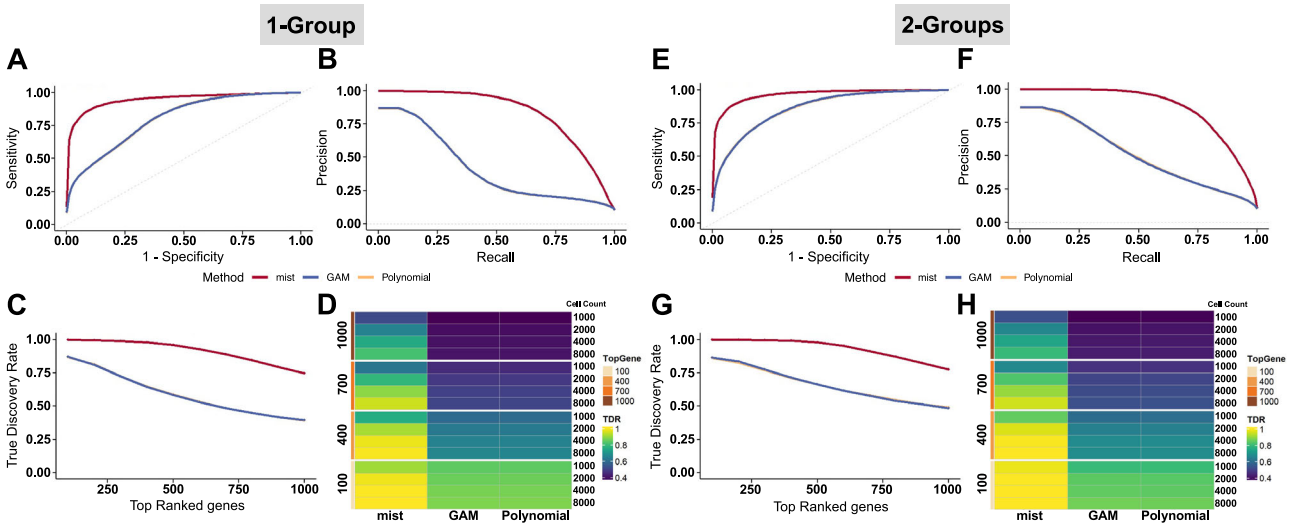

Fig. 3 | Comparison of DM detection accuracy among *mist*, GAM, and polynomial methods in detecting differentially methylated (DM) genes under both one-group and two-groups scenarios. A, **B**, **E**, **F** Receiver Operating Characteristic (ROC) curves and Precision-Recall curves of DM detection methods. **C**, **G** True Discovery Rate (TDR) along top-ranked genes identified by each method, where TDR is defined as the proportion of true DM genes among the top-ranked genes. **D**, **H** Heatmaps showing TDR across different cell counts and top-gene cutoffs (e.g., TDRs of TopGene = 400 would mean the True Discovery Rate among genes that are ranked to top 400 by their metrics.). A total of $N = 10{,}000$ genes and $N = 4{,}000$ cells were simulated under a 10% DM gene scenario. Source data are provided as a Source Data file.

in the 2-group simulations, *mist* maintains superior performance (Fig. 3E–H). *mist* shows consistently higher sensitivity, higher precision, and higher TDR values across all top-ranked genes. Although GAM and polynomial regression show slight improvements compared to the 1-group scenario, their TDR values remain significantly lower, highlighting the improvement of the Bayesian approach in DM calling at two-group comparison scenarios. Additional results under other scenarios are included in Supplementary Figs. S1–S23.

We additionally compared *mist* with scMET[34], a method that models cell-to-cell heterogeneity in single-cell DNA methylation but does not incorporate pseudotime. In simulations with 10,000 genes and 4,000 cells per group, scMET's differential variability (DV) analysis showed limited ability to detect differential trajectories compared with *mist* (Supplementary Fig. S26). This is expected, as scMET focuses on cell-to-cell variability while ignoring cell ordering along pseudotime, making it less effective for identifying dynamic methylation changes.

Real data analysis

To evaluate the performance of the proposed *mist* model in practical scenarios, we conducted real data analyses for both 1-group and 2-group designs. We utilized the dataset GSE121708[29], which provides multi-omics single-cell profiling of mouse embryos at embryonic days 4.5 to 7.5. The profiled omics modalities include RNA, DNA methylation, and chromatin accessibility, covering critical stages of embryonic early development and cell lineages.

Figure 4 summarizes the findings of the methylation trajectory analysis on the real dataset under a 1-group design. Figure 4A presents a pie chart summarizing the distribution of cells across major annotated embryonic lineages. Each segment represents a distinct lineage (e.g., ectoderm, endoderm, epiblast), with overlaid numbers indicating the count of cells in each category. Next, pseudotime inference was conducted using Monocle3 with the DNA methylation data, and the resulting UMAP under its default settings is shown in Fig. 4B. In the UMAP, each point represents an individual cell and is color-coded according to its inferred pseudotime value. The branching patterns, along with gradual shifts in color from one region to another, reflect a continuous developmental trajectory. Figure 4C validates the *mist* model's assumption of stage-specific variability by showing the distributions of dispersions (σ^2) for four developmental stages. We observe higher variability in early stages and lower variability in later stages. This pattern aligns with biological expectations, as early stages exhibit more variable methylation patterns that support diverse developmental pathways, while later stages show relatively stabilized patterns associated with cell fate commitment. Figure 4D is a Venn diagram showing the overlap among DM genes identified by *mist*, GAM, and polynomial regression. While there is a substantial amount of overlapping genes identified by all three methods, *mist* uniquely identified 3,927 DM genes. Among the top-ranked genes were several well-characterized developmental regulators whose methylation patterns aligned with known lineage transitions. For example, Fig. 4E shows that *Apela* (also known as *Elabela*), a factor essential for proper heart development and embryonic signaling[43–45], and *L1td1*, associated with the maintenance of pluripotency in embryonic stem cells[46], both exhibited methylation changes mirroring the loss of pluripotency. Additionally, *Cd3g*, a gene critical for T-cell receptor signaling and early hematopoietic development[47,48], exhibited methylation changes paralleling early lineage transitions during mouse embryogenesis. These findings demonstrate the ability of the *mist* model to uncover biologically relevant signals reflective of developmental and lineage transitions.

Building on our evaluation of the *mist* model under the 1-group design, we next applied it to a multi-group dataset, GSE213950[49], which contains single-nucleus joint profiling of chromatin conformation and DNA methylation (snm3C-seq) from human frontal cortex and hippocampus samples. This dataset spans critical developmental stages, mid-gestational, late-gestational, infant, and adult, and includes over 53,000 single-nucleus profiles.

Figure 5 demonstrates our methylation trajectory analysis across the frontal cortex and hippocampus. In Fig. 5A, a pie chart details the cellular composition among key neuronal subtypes (e.g., Exc-CA, Inh-MGE), with each segment displaying the corresponding cell count. Figure 5B presents the cell numbers in the two brain regions across four age groups (2T, 3T, infant, and adult), revealing a similar distribution that enables a direct and balanced comparison. Again, pseudotime inference is conducted using Monocle3 with the DNA methylation data. Figure 5C shows the resulting UMAP, where

https://doi.org/10.1038/s41467-026-70523-y

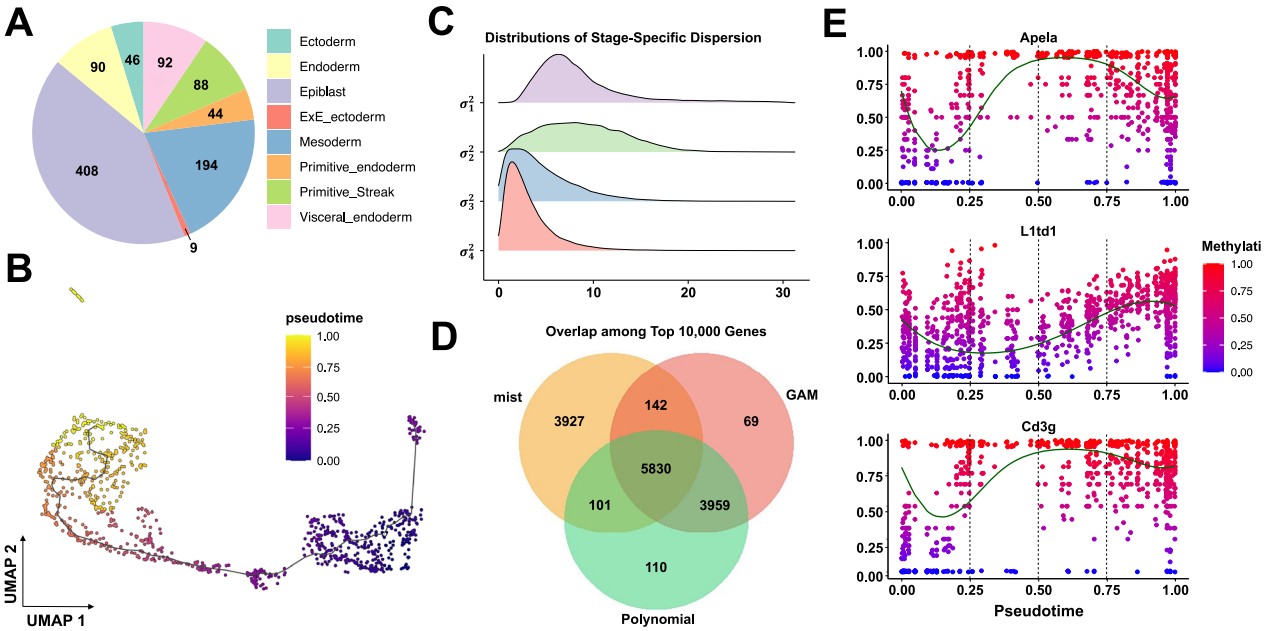

Fig. 4 | One-group differential methylation analysis on mouse embryos data using *mist*. **A** Pie chart illustrating the distribution of embryonic lineages in this dataset by corresponding number of cells. **B** UMAP projection of single cells, produced by Monocle3, and colored by normalized pseudotime. **C** Distributions of estimated dispersion at each stage across pseudotime. Here, the dispersion estimates refer to the σ^2 in the logit-normal distribution $logitnorm(\mu_t, \sigma^2)$ assumed on methylation levels. **D** Venn diagram showing the overlap among the top 10,000 genes identified by *mist*, GAM, and polynomial regression. **E** Methylation trajectories along pseudotime of three genes (*Apela, L1td1, Cd3g*), with dashed lines segmenting the 4 assumed developmental stages. Source data are provided as a Source Data file.

individual cells are plotted as points colored by their inferred pseudotime, mirroring their developmental trajectories. Finally, we applied the *mist* model to identify top DM genes between the frontal cortex and hippocampus, and generated a heatmap of methylation levels for the top 50 genes in both brain regions, thereby uncovering unique, brain region-specific epigenetic landscapes (Fig. 5D).

Our DM analysis between the two brain regions also reveals genes that exhibit distinct, brain region-dependent methylation trajectories. For instance, *CAMK2A*, a critical mediator of synaptic plasticity and memory consolidation[50,51], displays a reduction in methylation in the hippocampus as development progresses (Fig. 5E). This demethylation likely corresponds to increased gene expression, aligning with its established role in memory encoding and consolidation[52]. Additionally, *FGFR3*, which is integral to neurogenesis and cortical maturation[53,54], exhibits a gradual increase in methylation in the frontal cortex (Fig. 5F). Such hypermethylation likely corresponds with transcriptional downregulation, reflecting the transition from an early proliferative state to a more stable, long-term functional and differentiated cortical network[54]. Collectively, these findings highlight the capacity of the *mist* model to capture regulatory events that distinguish brain region-specific developmental trajectories.

Discussion

In this study, we introduced *mist*, a statistical framework for modeling DNA methylation dynamics along pseudotime in single-cell data and detecting differentially methylated (DM) features. By leveraging a Bayesian hierarchical structure and stage-specific modeling, *mist* effectively captures both non-linear methylation trajectories and stage-specific biological variation. Our results demonstrate that *mist* provides more accurate parameter estimation, and higher accuracy in identifying DM features compared to existing methods such as GAM and polynomial regression. Overall, *mist* addresses key challenges in scDNAm analysis along pseudotime.

While several pseudotime-based tools such as tradeSeq and PseudotimeDE[55,56] have been developed for scRNA-seq data, they are optimized for dense count data modeled by distributions such as the negative binomial. In contrast, single-cell DNA methylation data are highly sparse and represented as proportions. *mist* addresses these challenges by modeling logit-transformed methylation levels within a hierarchical Bayesian framework, explicitly accounting for stage-specific variability and data sparsity. In addition, *mist* performs trajectory-based differential methylation (DM) testing, enabling detection of genes whose methylation changes differ between conditions along pseudotime. This distinction underscores why RNA-focused tools are not directly transferable to methylation data and highlights *mist*'s unique contribution to single-cell epigenome analysis.

A major advantage of *mist* is its leveraging of a Bayesian framework, which inherently addresses data sparsity by incorporating prior information to stabilize parameter estimates. In this context, the horseshoe prior plays a critical role, where its global-local shrinkage properties efficiently suppress noise while allowing true biological signals to emerge, thereby enhancing the detection of subtle methylation changes. Additionally, it models developmental-stage-specific variability in DNA methylation, which allows for varying heterogeneity patterns across different biological phases. This improves the model's ability to detect functionally vital changes in biological dispersion. Furthermore, *mist* employs a differential testing strategy that minimizes the area between fitted methylation trajectories, enabling a direct detection of dynamic changes over pseudotime without being confounded by absolute methylation differences. This feature is particularly useful in removing regulatory elements that exhibit temporally correlated methylation changes during cell differentiation.

Our benchmarking experiments confirm that *mist* consistently outperforms GAM and polynomial regression in DM gene detection. Across both single-group and two-group simulations across multiple simulation scenarios, *mist* achieves higher TDR and precision-recall scores, demonstrating its robustness. This is particularly relevant given the sparsity of scDNAm data, where missing values and low CpG coverage often hinder statistical power. Furthermore, applying *mist* to a multi-omics dataset of mouse embryonic development identified key

https://doi.org/10.1038/s41467-026-70523-y

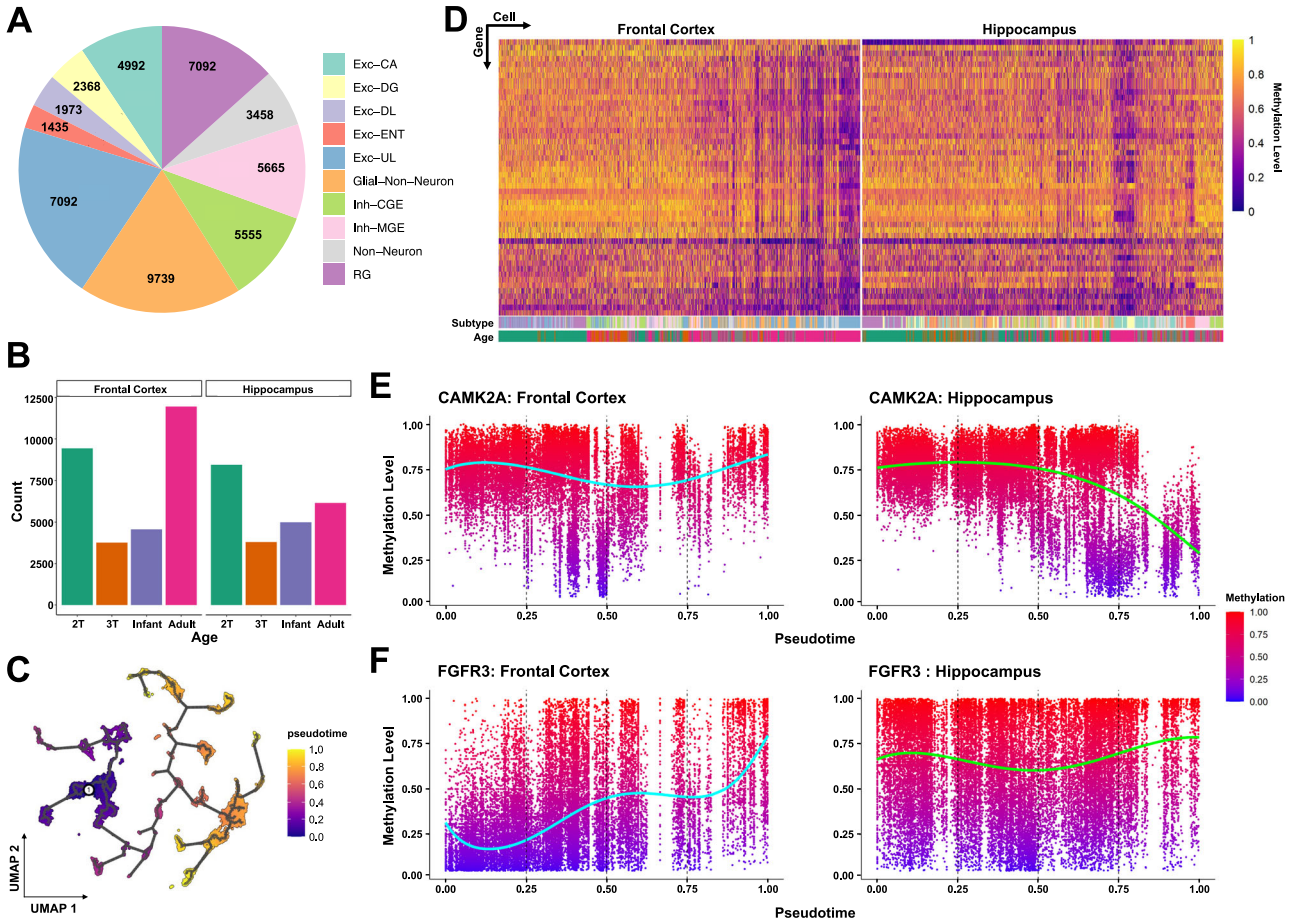

Fig. 5 | Two-group differential methylation analysis on human brain data using *mist*. **A** Pie chart showing the distribution of neuronal subtypes in this dataset. **B** Bar chart illustrating the number of cells sampled from the frontal cortex and hippocampus at different age groups. **C** UMAP projection, produced by Monocle3, and colored by pseudotime. **D** Heatmap of DNA methylation profiles for the top 50 DM genes detected by *mist* from the frontal cortex and hippocampus, with columns representing individual cells and rows corresponding to genes, and labels indicating neuronal subtypes and age groups. **E** Methylation trajectories along pseudotime of *CAMK2A* in the frontal cortex and hippocampus. **F** Methylation trajectories along pseudotime of *FGFR3* in the frontal cortex and hippocampus. Source data are provided as a Source Data file.

regulators such as *Apela* and *L1td1*, whose methylation patterns align with lineage transitions. These results highlight *mist*'s potential in uncovering epigenetic mechanisms underlying cellular differentiation and development. By mapping these regulators onto a continuous pseudotime trajectory, *mist* not only confirms their established roles but also reveals how their epigenetic regulation evolves across developmental stages. This perspective complements existing transcriptomic and genetic studies, offering a more comprehensive view of how coordinated methylation dynamics contribute to lineage specification and tissue maturation.

Despite its advantages, *mist* has limitations. This framework requires predefined developmental stages, and the default use of four stages may not accurately capture the complexity of all developmental trajectories or biological contexts. Because *mist* models DNA methylation as a function of user-provided pseudotime, inaccuracies in trajectory inference can affect its results. Most pseudotime tools were developed for transcriptomic data, so we recommend applying *mist* to multi-modal datasets where pseudotime can be inferred from gene expression or chromatin accessibility. Leveraging these modalities can improve cell ordering, and users should be mindful of parameters such as the choice of root cell, which may strongly influence results. Importantly, using a separate modality for pseudotime estimation can also help avoid the issue of double-dipping[57], where the same data are used for both trajectory inference and differential testing. We reanalyzed the data using STREAM-inferred pseudotime from transcriptomic data and found greater overlap among methods, while *mist* continued to emphasize stronger DM signals (Supplementary Figs. S29, S30). While *mist* delivers improved accuracy, its reliance on Gibbs sampling increases computational cost primarily with the number of features. For genome-wide studies using gene-level features, the cost remains manageable (Supplementary Fig. S24). However, if the basic unit is shifted to smaller genomic regions, the computational burden could become significantly higher. This issue can be partly mitigated through the adoption of parallel computing in the *mist* software package implementation. Moreover, *mist* currently models methylation at the gene or region level rather than individual CpG sites. Extending the framework to capture CpG site-specific dynamics or integrating additional multi-omics layers, such as chromatin accessibility and transcriptional activity, could provide a more comprehensive understanding of epigenetic regulation. Finally, *mist* relies on sufficient sequencing coverage to achieve robust inference. Our simulations indicate that performance is sensitive to very low coverage (10–20 reads), where technical noise dominates and power is reduced. Stable estimation is generally achieved at 50 reads or higher (Supplementary S25), which may limit applicability in ultra-sparse single-cell datasets.

More broadly, *mist* also shares limitations common to pseudotime-based analyses. Pseudotime represents cellular

progression using a single scalar ordering, which can oversimplify complex biological processes involving branching, multi-lineage differentiation, or concurrent changes across multiple cellular states. As a result, pseudotime often captures a dominant gradient in the data rather than progression along a single biological process. Recent advances such as graph-based trajectory inference, RNA velocity, and optimal-transport or probabilistic transition models aim to capture richer cellular dynamics by modeling branching structures, directionality, or cell-state transition probabilities[41,58–64]. Integrating these complementary frameworks with *mist* will be an important direction for improving the biological fidelity and interpretability of inferred trajectories in future work.

Methods

Hierarchical Bayesian model

Here we illustrate the model on one gene (or one genomic region), thus gene index is dropped. For one gene, the estimated pseudotime for each cell ranges from 0 to 1, which we divide into S discrete stages, representing distinct biological phases or developmental stages. For example, the pseudotime is segmented into four stages: stage 1 spans the interval [0, 0.25), stage 2 covers [0.25, 0.5), stage 3 covers [0.5, 0.75), and stage 4 spans [0.75, 1]. Segmenting pseudotime into discrete stages is essential because it captures the unique patterns of biological variability that characterize distinct phases of cellular development. By defining these stages, we can more accurately model how DNA methylation dynamics evolve over time, thereby revealing stage-specific regulatory mechanisms and key transitions in cell differentiation.

Within each stage s, we further cut it into T timepoints, representing the moment in pseudotime that marks the progression within a given stage. At each stage s, s = 1, 2, 3, ..., S and t^{th} timepoint, t = 1, 2, 3, ..., T, there are n_{st} cells indexed by j. The methylated count and total sequencing coverage for each cell j, j = 1, 2, 3, ..., n_{st}, are denoted as X_{stj} and N_{stj}, respectively, with the methylation level represented as $y_{stj} = X_{stj}/N_{stj}$.

We assume the methylation levels follow a logit-normal distribution:

$$y_{stj} \sim \text{logitnorm}(\mu_{t'}, \sigma_s^2), \tag{1}$$

where $\mu_{t'} = \beta_0 + \boldsymbol{Z}_{t'}\boldsymbol{\beta}$, with $\boldsymbol{Z}_{t'}$ representing the polynomial basis terms of t' up to degree 4, chosen for its balance between flexibility and avoiding overfitting, and $\boldsymbol{\beta}$ being the coefficients. The variance parameter σ_s^2 is stage-specific, allowing biological dispersion estimations for each stage. We refer to this as stage-specific dispersion/variance throughout this work. This modeling approach captures both the nonlinear trends in methylation dynamics across pseudotime and the variability specific to different biological stages, leading to a more realistic representation of the biological processes underlying methylation changes.

To enable full Bayesian inference at a later stage, we first assign a horseshoe prior[65] for the coefficients $\boldsymbol{\beta}$:

$$\beta_k \sim N(0, \lambda_k^2\tau^2),\ \lambda_k \sim C^+(0,1),\ \tau \sim C^+(0,1). \tag{2}$$

Here, $C^+(0, 1)$ represents a half-Cauchy distribution with a scale parameter of 1, and λ_k and τ are local and global shrinkage parameters, respectively. The horseshoe prior is used for β to provide shrinkage, allowing important coefficients to remain large while shrinking others close to zeros. Unlike traditional priors such as normal or Laplace, it is well suited for sparse signals because it adaptively suppresses irrelevant coefficients while preserving significant ones, making it appropriate when not all predictors are expected to be influential. In addition, we assign a normal prior centered at zero for β_0 ($\beta_0 \sim N(0, 3^2)$), a non-informative prior for σ_s^2 ($p(\sigma_s^2) \propto 1/(\sigma_s^2)$).

The parameters of interest include the intercept term β_0, the coefficient vector $\boldsymbol{\beta}$, and the stage-specific variance σ_s^2. These reflect the methylation level at pseudotime 0, the dynamics over pseudotime, and the biological dispersion within each stage, respectively.

Parameters estimation

Following the derivation of the conditional distributions for each parameter (Supplementary S1), we implement Gibbs sampling to iteratively update each parameter by sampling from its conditional distribution based on the current values of the other parameters. In addition, to sample the hyperparameters λ_k^2 and τ^2 associated with the horseshoe prior, we employ a slice sampling method[66] that introduces latent variables to enable their sampling (Supplementary S1.1). This procedure provides the posterior distributions of the parameters of interest β_0, β, and σ_s^2, which are used in the subsequent differential analysis or can be interpreted directly.

Differential methylation analysis

For each gene, we assess differential methylation between two groups by comparing their fitted methylation trajectories over pseudotime. Specifically, we calculate the minimized area between the two curves (Fig. 1). This approach captures the true difference in how methylation changes over time while adjusting for differences in absolute methylation levels. By finding the vertical shift C that minimizes the area, we effectively control for baseline differences and isolate the dynamics of the methylation changes:

$$A = \min_C \int_0^1 \left(\hat{y}_A(x) - (\hat{y}_B(x) + C)\right)^2 dx, \tag{3}$$

where C is chosen to minimize the area difference. The optimal C value, C^*, which shifts the curves as closely as possible without altering their shape, is given by

$$C^* = \int_0^1 \left(\hat{y}_A(x) - \hat{y}_B(x)\right) dx. \tag{4}$$

Here $\hat{y}_A(x)$ and $\hat{y}_B(x)$ represent the estimated methylation curves for each group. By focusing on the minimized area rather than the absolute difference, we capture the change in methylation dynamics over pseudotime, providing a more meaningful measure of differential methylation (DM).

For the one-group scenario, we simply calculate the area between the fitted curve and a constant C:

$$A = \min_C \int_0^1 \left(\hat{y}(x) - C\right)^2 dx, \tag{5}$$

where the optimal C, C^*, is given by

$$C^* = \int_0^1 \hat{y}(x) dx. \tag{6}$$

Finally, we rank these minimized area differences across all genes to identify those with the largest deviations in methylation dynamics, either between two groups or relative to a constant line within a single group.

Our hierarchical Bayesian model provides a framework for analyzing single-cell DNA methylation data, capturing both the trajectory along pseudotime and the variability across stages. This differential testing approach detects both large methylation changes within a single group and differences between groups, offering insight into epigenetic alterations along pseudotime.

Simulation settings

We conducted simulations to benchmark the accuracy of DM gene detection using our proposed method in comparison with two alternative approaches, the linear model and Generalized Additive Model (GAM). For the linear model, we considered a degree-4 additive model (i.e., $\beta_0 + \beta_1 t' + \beta_2 t'^2 + \beta_3 t'^3 + \beta_4 t'^4$) to capture high-order trends in methylation dynamics, providing flexibility in modeling while retaining some interpretability. Alternatively, GAMs are widely used in scRNA-seq analyses, particularly for trajectory inference, where they model gene expression as a non-linear function of pseudotime to effectively capture complex trends[55,56]. This flexibility has made GAMs a key component of tools like pseudotimeDE and tradeSeq[55,56] for detecting differential expression along trajectories. Both methods were evaluated in one-group and two-group scenarios to assess their accuracy in detecting differential methylation dynamics. The one-group scenario has only a single phenotypic group, where the objective is to identify genes exhibiting the most significant methylation changes over time. Such analysis provides insights into dynamic regulatory processes along pseudotime within one group of cells. The two-group scenario examines distinct phenotypical groups, such as different cell types, health statuses, or experimental conditions. Here, the goal is to identify biomarkers with the most pronounced differential methylation changes between the two groups.

We conducted simulations with DM genes at 5%, 10%, and 15% among a total of 10,000 genes. A total of 20 simulation iterations were conducted. To closely mimic real biological data, we first applied our Bayesian framework to estimate distributions of regression coefficients for each gene from a real single-cell DNA methylation dataset[29]. Genes were ranked using the previously described metric of minimum area difference. For one-group simulations, true DM genes parameters were sampled from the top 1000 to 3000 ranked genes, while parameters for non-DM genes were drawn from the subsequent 3000 genes. In the two-group scenario, genes were randomly sampled to generate 10,000 gene pairs. Then a similar approach was adopted to simulate DM and non-DM genes. Using these parameters in logit-normal distribution simulations (Equation (1)), we generated single-cell methylation data and pseudotime in both one-group and two-group settings. Detailed simulation procedure is included in Supplementary S2.

Statistics and reproducibility

No statistical method was used to predetermine sample size. The sample sizes used were consistent with those in similar single-cell DNA methylation studies and were sufficient to detect differential methylation signals. For simulations, we conducted 20 iterations with 10,000 genes and varying cell counts (including 4000 cells per group in two-group scenarios) to assess method performance across multiple scenarios. No data were excluded from the analyses. Statistical analyses were performed in R (v4.3.1 or later) unless otherwise specified. For benchmark comparisons, Generalized Additive Models (GAM) used likelihood ratio tests and polynomial regression employed ANOVA with adjusted p values for differential calling. Pseudotime inference was performed using Monocle3 v1.3.7. All computational analyses are fully reproducible using the code archived at https://doi.org/10.5281/zenodo.18451940[67], with detailed parameters and workflows described in the Methods section and package vignettes.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data for simulations and real-data analyses are available at https://doi.org/10.5281/zenodo.18451940[67]. The mouse embryos data is available under accession codes GSE121708 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121708[29]. The human brain data is available through https://cells.ucsc.edu/?ds=brain-epigenome[49]. Source data are provided with this paper.

Code availability

mist is publicly available as a R/Bioconductor package at: https://bioconductor.org/packages/mist/. The package includes detailed vignettes and example datasets to guide users through typical workflows. Source code is hosted at https://github.com/dxd429/mistand is released under the MIT License. The specific version of the code associated with this publication is archived in https://doi.org/10.5281/zenodo.18451940[67].

References

1. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).
2. Dor, Y. & Cedar, H. Principles of {DNA} methylation and their implications for biology and medicine. *Lancet* **392**, 777–786 (2018).
3. Jiang, S. & Guo, Y. Epigenetic clock: DNA methylation in aging. *Stem Cells Int.* **2020**, 1047896 (2020).
4. Horvath, S. & Raj, K. {DNA} methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
5. Matzke, M. A. & Mosher, R. A. {RNA}-directed {DNA} methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**, 394–408 (2014).
6. Kulis, M. & Esteller, M. {DNA} methylation and cancer. *Adv. Genet.* **70**, 27–56 (2010).
7. Klutstein, M., Nejman, D., Greenfield, R. & Cedar, H. Dna methylation in cancer and aging. *Cancer Res.* **76**, 3446–3450 (2016).
8. Koch, A. et al. Analysis of {DNA} methylation in cancer: location revisited. *Nat. Rev. Clin. Oncol.* **15**, 459–466 (2018).
9. Beck, S. & Rakyan, V. K. The methylome: approaches for global {DNA} methylation profiling. *Trends Genet. TIG* **24**, 231–237 (2008).
10. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for {Bisulfite}-{Seq} applications. *Bioinformatics* **27**, 1571–1572 (2011).
11. Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13**, 1–10 (2012).
12. Jaffe, A. E. et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* **41**, 200 (2012).
13. Akalin, A. et al. Methylkit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87 (2012).
14. Feng, H., Conneely, K. N. and Wu, H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.* **42**(8), https://doi.org/10.1093/nar/gku154 (2014).
15. Sun, D. et al. {MOABS}: model based analysis of bisulfite sequencing data. *Genome Biol.* **15**, R38 (2014).
16. Guo, H. et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**, 2126–2135 (2013).
17. Guo, H. et al. The {DNA} methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014).
18. Li, Q. et al. Heterogeneous {DNA} methylation status in same-cell subpopulations of ovarian cancer tissues. *Tumour Biol. J. Int. Soc. Oncodev. Biol. Med.* **39**, 1010428317701650 (2017).
19. Hou, Y. et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26**, 304–319 (2016).

20. He, Y. et al. Spatiotemporal {DNA} methylome dynamics of the developing mouse fetus. *Nature* **583**, 752–759 (2020).
21. Yao, Z. et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature* **598**, 103–110 (2021).
22. Liu, H. et al. {DNA} methylation atlas of the mouse brain at single-cell resolution. *Nature* **598**, 120–128 (2021).
23. Luo, C. et al. Single nucleus multi-omics identifies human cortical cell regulatory genome diversity. *Cell Genom.* **2**, 100106,100107 (2022).
24. Zhang, Z. et al. Epigenomic diversity of cortical projection neurons in the mouse brain. *Nature* **598**, 167–173 (2021).
25. Kapourani, C.-A. & Sanguinetti, G. Melissa: {Bayesian} clustering and imputation of single-cell methylomes. *Genome Biol.* **20**, 61 (2019).
26. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. {DeepCpG}: accurate prediction of single-cell {DNA} methylation states using deep learning. *Genome Biol.* **18**, 67 (2017).
27. Zou, L. S. et al. {BoostMe} accurately predicts {DNA} methylation values in whole-genome bisulfite sequencing of multiple human tissues. *BMC Genom.* **19**, 390 (2018).
28. Farlik, M. et al. Single-cell {DNA} methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.* **10**, 1386–1397 (2015).
29. Argelaguet, R. et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).
30. de Souza, C. et al. Epiclomal: {Probabilistic} clustering of sparse single-cell {DNA} methylation data. *PLoS Comput. Biol.* **16**, e1008270 (2020).
31. Johnson, K. C. et al. Single-cell multimodal glioma analyses identify epigenetic regulators of cellular plasticity and environmental stress response. *Nat. Genet.* **53**, 1456–1468 (2021).
32. Luo, C. et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
33. Hernando-Herraez, I. et al. Ageing affects {DNA} methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. *Nat. Commun.* **10**, 4361 (2019).
34. Kapourani, C. A., Argelaguet, R., Sanguinetti, G. & Vallejos, C. A. scMET: Bayesian modeling of DNA methylation heterogeneity at single-cell resolution. *Genome Biol.* **22**, 1–21 (2021).
35. Jones, P. A. & Taylor, S. M. Cellular differentiation, cytidine analogs and {DNA} methylation. *Cell* **20**, 85–93 (1980).
36. Razin, A. et al. Variations in {DNA} methylation during mouse cell differentiation in vivo and in vitro. *Proc. Natl. Acad. Sci. USA* **81**, 2275–2279 (1984).
37. Parry, A., Rulands, S. & Reik, W. Active turnover of {DNA} methylation during cell fate decisions. *Nat. Rev. Genet.* **22**, 59–66 (2021).
38. Dolzhenko, E. & Smith, A. D. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinform.* **15**, 215 (2014).
39. Amat, M., Guignard, G., Llor, N. & Bosch, J. Access to enantiopure 4-substituted 1,5-aminoalcohols from phenylglycinol-derived δ-lactams: synthesis of {Haliclona} alkaloids. *J. Org. Chem.* **79**, 2792–2802 (2014).
40. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
41. Street, K. et al. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genom.* **19**, 1–16 (2018).
42. Chen, H. et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with {STREAM}. *Nat. Commun.* **10**, 1903 (2019).
43. Chng, S. C., Ho, L., Tian, J. & Reversade, B. ELABELA: a hormone essential for heart development signals via the apelin receptor. *Dev. Cell* **27**, 672–680 (2013).
44. Ho, L. et al. ELABELA is an endogenous growth factor that sustains hESC self-renewal via the PI3K/AKT pathway. *Cell Stem Cell* **17**, 435–447 (2015).
45. Pauli, A. et al. Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* **343**. https://doi.org/10.1126/SCIENCE.1248636 (2014).
46. Iwabuchi, K. A. et al. ECAT11/L1td1 is enriched in ESCs and rapidly activated during iPSC generation, but it is dispensable for the maintenance and induction of pluripotency. *PloS ONE* **6**, https://doi.org/10.1371/JOURNAL.PONE.0020461 (2011).
47. Rowe, J. H. et al. Patients with CD3G mutations reveal a role for human CD3γ in Treg diversity and suppressive function. *Blood* **131**, 2335–2344 (2018).
48. Garcillán, B. et al. CD3G or CD3D knockdown in mature, but not immature, T lymphocytes similarly cripples the human TCRαβ complex. *Front. Cell Dev. Biol.* **9**. https://doi.org/10.3389/FCELL.2021.608490 (2021).
49. Heffel, M. G. et al. Temporally distinct 3D multi-omic dynamics in the developing human brain. *Nature* **635**. https://doi.org/10.1038/S41586-024-08030-7 (2024).
50. Rhein, C. et al. Association of a CAMK2A genetic variant with logical memory performance and hippocampal volume in the elderly. *Brain Res. Bull.* **161**, 13–20 (2020).
51. Yasuda, R., Hayashi, Y. & Hell, J. W. CaMKII: a central molecular organizer of synaptic plasticity, learning and memory. *Nat. Rev. Neurosci.* **23**, 666–682 (2022).
52. Lisman, J., Yasuda, R. & Raghavachari, S. Mechanisms of CaMKII action in long-term potentiation. *Nat. Rev. Neurosci.* **13**, 169 (2012).
53. Thomson, R. E. et al. Fgf receptor 3 activation promotes selective growth and expansion of occipitotemporal cortex. *Neural Dev.* **4**. https://doi.org/10.1186/1749-8104-4-4 (2009).
54. Huang, J. Y. et al. Enhanced FGFR3 activity in postmitotic principal neurons during brain development results in cortical dysplasia and axonal tract abnormality. *Sci. Rep.* **10**. https://doi.org/10.1038/S41598-020-75537-0 dec (2020).
55. Song, D. & Li, J. J. {PseudotimeDE}: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell {RNA} sequencing data. *Genome Biol.* **22**, 124 (2021).
56. den Berge, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).
57. Neufeld, A., Dharamshi, A., Gao, L. L., Witten, D. & Bien, J. Discussion of "Data fission: splitting a single data point". *J. Am. Stat. Assoc.* **120**, 151–157 (2025).
58. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
59. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
60. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 1–9 (2019).
61. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. methods* **13**, 845–848 (2016).
62. Lange, M. et al. CellRank for directed single-cell fate mapping. *Nat. methods* **19**, 159–170 (2022).
63. Weiler, P., Lange, M., Klein, M., Pe'er, D. & Theis, F. CellRank 2: unified fate mapping in multiview single-cell data. *Nat. Methods* **21**, 1196–1205 (2024).
64. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943.e22 (2019).
65. Carvalho, C. M., Polson, N. G. & Scott, J. G. The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480 (2010).

66. Neal, R. M. Slice sampling. **31**, 705–767. https://doi.org/10.1214/AOS/1056562461 (2003).
67. dxd429. dxd429/mistdata: Data_code_mist_manuscript. https://doi.org/10.5281/zenodo.18451940 (2026).

Acknowledgements

We thank Dr. Chongyuan Luo for insightful discussions and for publicly sharing the snm3C-seq data from his lab. This work was supported by grants from the National Institutes of Health awarded [NIGMS R35GM154862 to H.F.]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Funding to pay the Open Access publication charges for this article was provided by NIH [R35GM154862].

Author contributions

L.Z. and H.F. conceived the study. D.D. and W.M. performed in-silico simulation studies and real data analysis. W.T. acquired and processed real scDNAm datasets. D.D., L.Z., and H.F. wrote the manuscript. H.W. supervised the research and led discussions. All authors helped edit the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-70523-y.

Correspondence and requests for materials should be addressed to Liangliang Zhang or Hao Feng.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.