Genome Biology

METHOD Open Access

scMET: Bayesian modeling of DNA methylation heterogeneity at single-cell resolution

ChantriolInt-Andreas Kapourani[1,2†], Ricard Argelaguet[3†], Guido Sanguinetti[2,4*] and Catalina A. Vallejos[1,5*]

*Correspondence:
gsanguin@sissa.it;
catalina.vallejos@igmm.ed.ac.uk
†ChantriolInt-Andreas Kapourani and Ricard Argelaguet contributed equally to this work.
[1]MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
[2]School of Informatics, University of Edinburgh, Edinburgh, UK
Full list of author information is available at the end of the article

Abstract

High-throughput single-cell measurements of DNA methylomes can quantify methylation heterogeneity and uncover its role in gene regulation. However, technical limitations and sparse coverage can preclude this task. scMET is a hierarchical Bayesian model which overcomes sparsity, sharing information across cells and genomic features to robustly quantify genuine biological heterogeneity. scMET can identify highly variable features that drive epigenetic heterogeneity, and perform differential methylation and variability analyses. We illustrate how scMET facilitates the characterization of epigenetically distinct cell populations and how it enables the formulation of novel hypotheses on the epigenetic regulation of gene expression. scMET is available at https://github.com/andreaskapou/scMET.

Keywords: DNA methylation, Single-cell, Epigenetic heterogeneity, Hierarchical Bayes

Background

DNA methylation (DNAm) at cytosine residues plays an important role in the regulation of gene expression [1]. It is also critical for a broad range of biological processes, including X-chromosome inactivation, genomic imprinting, and cancer [2–4]. The gold standard approach to profile DNAm at single-base resolution is to treat DNA with sodium bisulphite, which efficiently converts unmethylated cytosines to uracils, while leaving methylated cytosines unmodified [5]. Although bulk bisulphite sequencing (BS-seq) experiments have paved the way for mapping the methylome landscape across different tissues, they fall short of explaining the inter-cellular methylation heterogeneity and quantifying its dynamics in a variety of biological contexts [6].

More recently, advances in sequencing technologies have enabled the development of protocols that profile DNAm with single-cell resolution (e.g., scBS-seq, [7, 8]) and multiplexing protocols offer scalability to thousands of cells in a single experiment [9, 10]. In contrast to gene expression signatures from scRNA-seq experiments, which are influenced by the environment, DNA methylation profiles are highly distinct between cell types and stable across individuals and over the life span [11, 12]. Moreover, while scRNA-seq assays might fail to capture information about genes with moderate expression levels, cell-level measurements of DNAm offer a more complete coverage across genomic regions [9]. However, due to the small amounts of initial genomic DNA and the destructive nature of bisulphite on nucleic acids, the output data are often noisy and extremely sparse; that is, a large proportion of CpG dinucleotides is not observed (ranging from 80 to 95%). While tailored computational imputation methods such as Melissa [13] and DeepCpG [14] might ameliorate the sparsity problem, disentangling genuine epigenetic variability from technical biases remains a formidable problem.

Here, we present scMET, a Bayesian framework that addresses the statistical challenges associated with sparse scBS-seq data and provides novel functionality that is tailored to single-cell-level datasets. To overcome sparsity, scMET aggregates the input data within regions (hereafter also referred to as genomic *features*): either by combining CpG information in a sliding window approach or by using pre-annotated contexts, such as promoter regions or enhancers [7, 15]. To dissect genuine epigenetic variability from the many confounding technical biases, scMET adopts a hierarchical model specification which shares information across cells and genomic features, while incorporating feature-level characteristics (e.g., CpG density). Critically, scMET introduces residual *overdispersion* estimates as a measure of DNAm variability that is not confounded by mean methylation. These estimates can be used to perform differential DNAm variability testing among groups of cells, embracing the cellular resolution of the data to provide novel insights which are not possible using traditional differential mean tests on bulk data [16]. scMET can also identify highly variable features (HVFs) which, among others, can be used as input for unsupervised clustering analyses.

scMET scales readily to thousands of cells and features, making it a powerful tool for large-scale single-cell epigenetic studies. Our results both on simulated and real datasets demonstrate that it can accurately and robustly quantify DNAm heterogeneity. Results on two recent large-scale datasets show that scMET detects biologically relevant highly variable features which result in improved clustering performance. In addition, we show that scMET can facilitate the interrogation of single-cell multi-omics assays, yielding novel biological hypotheses on the role of epigenetic variability in gene regulation in early development.

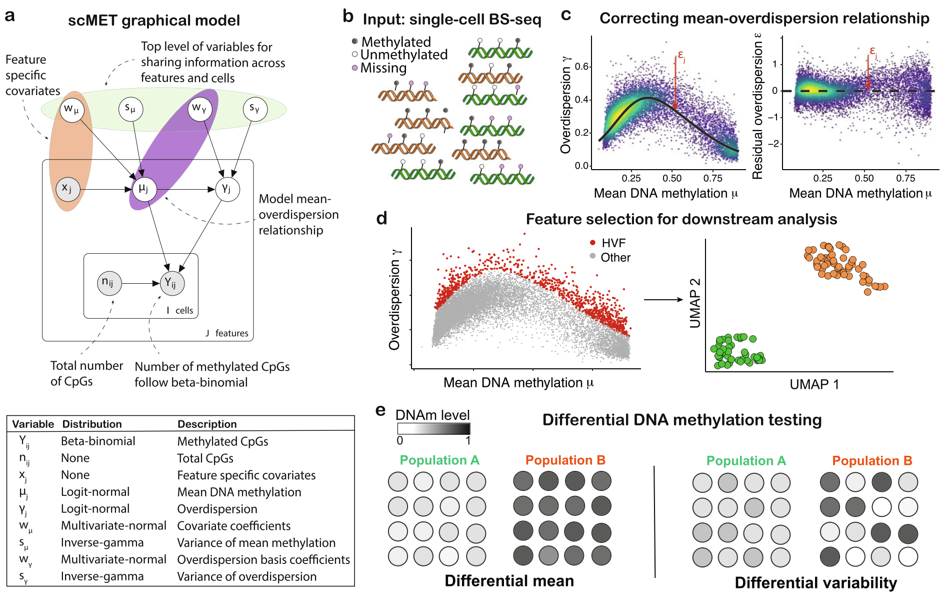
Results

Quantifying cell-to-cell DNAm heterogeneity with scMET

Standard statistical models for count data, such as the Poisson and binomial distributions, do not always capture the properties of data generated by high-throughout sequencing assays (e.g., RNA sequencing, bisulphite sequencing). In such cases, the data typically exhibit higher variance than what is predicted by these models — this is often referred to as *overdispersion* [17, 18]. This overdispersion may relate to *technical variation* (e.g., differences in sequencing depth) or to *biological variation* between the units of interest (e.g., cells or subjects) that is linked to genetic, environmental, or other factors. Disentangling these sources of variation is a major challenge in computational biology.

To disentangle technical from biological variability and overcome data sparsity, scMET couples a hierarchical beta-binomial (BB) model with a generalized linear model (GLM) framework (Fig. 1a, b). For each cell i and feature j, the input for scMET is the number of CpG sites that are observed to be methylated (Y_{ij}) and the total number of sites for which methylation status was recorded (n_{ij}), see Additional file 1: Figure S1. The BB model uses feature-specific mean parameters μ_j to quantify overall DNAm across all cells and biological *overdispersion* parameters γ_j as a proxy for cell-to-cell DNAm heterogeneity. The latter capture the amount of variability that is not explained by binomial sampling noise, which would only account for technical variation (see Additional file 1: Section S2.1). Hence, γ_j is akin to the overdispersion term used in negative binomial models for RNA-seq data (e.g., [19]). Although BB models have been developed for bulk DNAm data (e.g., [20, 21]), they typically use data from individual CpG sites as input, a strategy prone to fail for the highly sparse scBS-seq data.

The GLM framework is incorporated at two levels. Firstly, to introduce feature-specific covariates $\mathbf{x}_j$ (e.g., CpG density) that may explain differences in mean methylation μ_j across features. Secondly, similar to [22], we use a non-linear regression framework to capture the mean-overdispersion trend that is typically observed in high-throughput sequencing data, such as scBS-seq (Fig. 1c). Critically, this trend is used to derive *residual*



Variable	Distribution	Description
Y_{ij}	Beta-binomial	Methylated CpGs
n_{ij}	None	Total CpGs
x_j	None	Feature specific covariates
μ_j	Logit-normal	Mean DNA methylation
γ_j	Logit-normal	Overdispersion
w_μ	Multivariate-normal	Covariate coefficients
s_μ	Inverse-gamma	Variance of mean methylation
w_γ	Multivariate-normal	Overdispersion basis coefficients
s_γ	Inverse-gamma	Variance of overdispersion

Fig. 1 Graphical outline for scMET. **a** Overview of the scMET probabilistic graphical model. The random variables and data that form the model, along with the distributional assumptions, are shown. Input values are denoted by gray circles. Model parameters are denoted by white circles. **b** scMET uses single-cell DNAm data as input. The data could consist of measurements obtained from different groups of cells, such as experimental conditions or cell types (represented by green and orange colors in the diagram). For each region of interest (e.g., promoters), the input data is recorded in terms of the number of CpG sites for which a valid measurement was recorded and, among those, the number of methylated CpG sites. Note that many CpG sites will not be covered by a read (denoted by red color), leading to sparse information per genomic region. **c** By combining a hierarchical beta-binomial specification with a generalized linear model framework, scMET captures the mean-overdispersion relationship (left) that is typically observed in bisulphite sequencing readouts and derives residual overdispersion estimates that are not confounded by mean methylation (right). **d** scMET can be used to identify HVFs that drive epigenetic heterogeneity within a cell population. For example, these could be used as the input of dimensionality reduction techniques or clustering analyses. **e** scMET uses a probabilistic decision rule to perform differential methylation analysis: to identify features that show differences in mean methylation (left) and/or methylation variability (right) between pre-specified groups of cells

overdispersion parameters ϵ_j — a measure of cell-to-cell variability that is not confounded by mean methylation. Feature-specific parameters are subsequently used for (i) feature selection, to identify highly variable features (HVFs) that drive cell-to-cell epigenetic heterogeneity (Fig. 1d), and (ii) differential methylation testing, to highlight features that show differences in DNAm mean or variability between specified groups of cells (Fig. 1e).

By using a Bayesian formulation, scMET infers the posterior distribution for all model parameters (Methods). Moreover, a variational Bayes scheme [23] permits scalable analysis to thousands of cells and features (Additional file 1: Figure S2), while having comparable posterior inference performance when compared to a Markov Chain Monte Carlo implementation (Additional file 1: Figure S3 and S4). As in [24], the output generated by scMET is used to implement a probabilistic decision rule to enable HVF selection and differential methylation testing. The decision rule is calibrated to control the expected false discovery rate (EFDR, [25]). A more detailed description of the model specification and its implementation is provided in the "Methods" section. scMET is implemented as an R package and is available at https://github.com/andreaskapou/scMET.

scMET can be combined with different choices for the input set of features. As a default, we recommend these to be defined by existing pre-annotated contexts (e.g., enhancers), as these can facilitate downstream interpretation. However, scMET can also be used for de novo annotation of regulatory regions by using sliding windows as input features (see the "Methods" section).

Benchmarking scMET on synthetic data

First, we benchmark the performance of scMET using synthetic data. To mimic the properties observed in real scBS-seq data, we simulated features with rich and poor CpG density (see the "Methods" section for details about the simulation settings). We compared mean and overdispersion estimates obtained by scMET with respect to BB maximum likelihood estimates (BB MLE), which were obtained separately for each feature. As expected, mean parameters μ_j are easier to infer and estimates were comparable across both methods (Additional file 1: Figure S5). However, scMET outperformed BB MLE when inferring overdispersion parameters γ_j, particularly for small numbers of cells (Additional file 1: Figure S6).

To assess whether the shrinkage introduced by scMET improves overdispersion estimates in real data, we performed down-sampling experiments based on a subset of the dataset introduced by [9]. For this analysis, we focused on 424 inhibitory neurons (a more detailed description is provided in the "Methods" section). We compared estimates obtained using the full and down-sampled datasets (Additional file 1: Figure S7). We observed scMET posterior estimates to be more stable than BB MLE as the sample size decreased, suggesting that scMET leads to more robust inference. This is particularly important for rare cell populations or during early development, where large numbers of cells are difficult to obtain. In combination with the simulation study described above, this showcases the benefits of using a Bayesian hierarchical framework to share information across cells and genomic features.

When comparing scMET to existing DNA methylation analysis tools (primarily developed for bulk assays), it is important to emphasize fundamental differences with respect to the functionality provided by scMET. These are summarized in Additional file 1: Table S1 and in the "Methods" section. In particular, most approaches focus on differences in mean methylation. Among others, these are often based on Fisher's exact test or on BB models (e.g., DSS; [20]). As seen in Additional file 1: Figure S5, both scMET and BB MLE estimates for mean methylation parameters are comparable. As it can be expected, our comparison with respect to DSS also led to similar results (Additional file 1: Figure S8). When comparing with respect to Fisher's exact test, we did not observe substantial differences in performance either: scMET was better in terms of the F1-measure and type I error control, but led to more conservative results (Additional file 1: Figure S9 and S10 and Additional file 1: Section S2.4). Therefore, we do not aim to claim overall superiority of scMET for differential mean methylation testing. If a user is *only* interested in differential *mean* methylation, existing approaches such as DSS could be used. Instead, the main advantage of scMET is the ability to robustly perform differential variability (DV) testing and HVF selection (benchmarks for our HVF approach are shown in the next section).

In terms of DV testing, our simulations showed that for small effect sizes we would need more than 200 cells to achieve 50 to 80% power, whereas for features with larger effect sizes we would need around 50 cells per group to achieve 80% power (Additional file 1: Figure S11 and Section S2.4). To assess coverage requirements, we grouped features according to the average coverage CpGs across all cells (Additional file 1: Figure S12). For each feature, we computed the posterior standard deviation for the residual overdispersion parameter ϵ_j. Large values (e.g., above 0.5) indicate higher estimation uncertainty and may affect the robustness of downstream analyses. Based on this analysis, we suggest to exclude regions that have less than 3 CpGs covered (on average across all cells) for relatively large sample sizes (>100 cells), and a higher threshold of 5 CpGs for small datasets (<50 cells).

scMET improves feature selection for unsupervised analysis of single-cell methylomes from the mouse frontal cortex

To demonstrate the performance of scMET in real data, we considered a dataset where DNA methylation was profiled in 3069 cells isolated from the frontal cortex of young adult mice [9]. To date, this is one of the largest and most heterogeneous publicly available scBS-seq datasets. The main source of heterogeneity in this dataset is due to two broad classes of neurons: excitatory ($I = 2645$) and inhibitory ($I = 424$). Within each class, a hierarchy of sub-populations can be identified according to the cortical depth (Fig. 2a), where excitatory neurons progress from deep layers (mDL-1, mDL-2, mDL-3, mL6-1, mL6-2) to middle (mL5-1, mL5-2, mL4) and superficial layers (mL2/3). These groups were validated in the original study and thus can be used as a benchmark for clustering analyses.

We applied scMET to genomic features from three different putative regulatory elements: gene promoters within 2 kb around transcription start site ($J = 12{,}774$), distal H3K27ac peaks ($J = 17{,}284$), and H3K4me1 peaks ($J = 30{,}374$). As expected, scMET captured the mean-overdispersion relationship within each genomic context, and estimates for residual overdispersion parameters ϵ_j were not confounded by mean DNAm (Additional file 1: Figure S13).

Here, we illustrate scMET as a feature selection tool, using residual overdispersion estimates to identify HVFs that can be used as input for unsupervised analyses, such as clustering. For each genomic context, we selected HVFs (Additional file 1: Figure S14a) and performed a clustering analysis with varying numbers of HVFs (ranked by

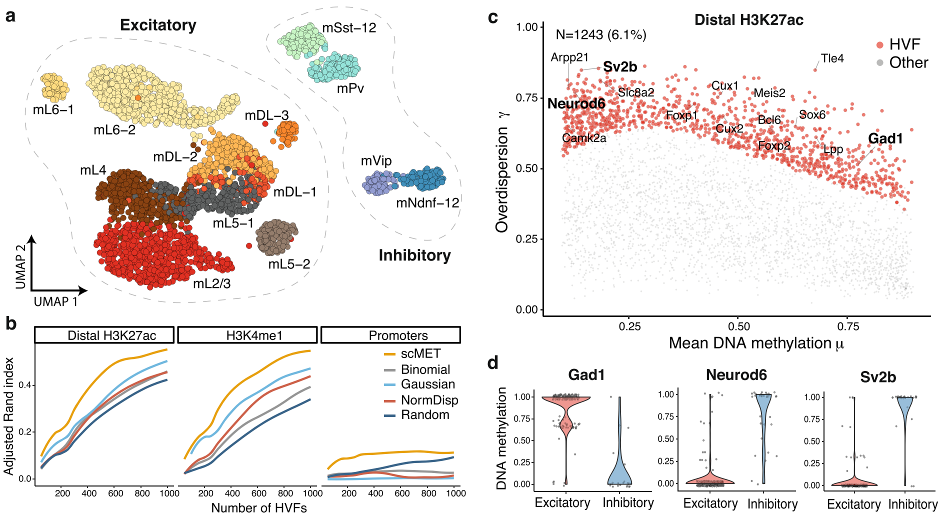


Fig. 2 Feature selection using scMET characterises cellular heterogeneity on the mouse frontal cortex. **a** UMAP [26] representation of neuron sub-populations present in the mouse frontal cortex dataset by combining the top 4000 HVFs identified by scMET across distal H3K27ac and H3K4me1 genomic contexts. Cells are colored according to the original cell sub-population assignments obtained by Luo et al. **b** Clustering performance in terms of adjusted Rand index (ARI), for varying number of HVFs. HVF selection was based on scMET's residual overdispersion parameters ϵ_j (yellow), binomial variance (gray), Gaussian variance (cyan), normalized dispersion (NormDisp, red), and random selection (blue). A finite grid of HVFs was used for ARI evaluation and non-parametric regression was used to obtain a smoothed interpolation across all values (Methods). **c** Identifying HVFs for the distal H3K27ac genomic context. Red points correspond to features being called as HVF (EFDR = 10% and percentile threshold $\delta_E = 90\%$). To ease interpretation, each element is linked to its nearest gene. Labels highlight known neuron marker genes that were used in the [9] study to define the different cell populations. **d** Example distal H3K27ac HVFs whose methylation patterns distinguish the two broad neuronal populations. Panel titles correspond to the nearest genes

decreasing values of their associated tail posterior probabilities) as input. More concretely, we performed dimensionality reduction followed by k-means clustering (Methods) and used the adjusted Rand index (ARI, [27]) to quantify agreement with respect to the sub-populations validated by Luo et al. As a comparison, we also evaluated four alternative HVF selection strategies based on a random choice (Random), normalized dispersion values (NormDisp, [28, 29]), Gaussian, and binomial models (Methods). As expected, the clustering performance improved steadily with increasing number of HVFs for all methods. However, scMET consistently led to better clustering performance (Fig. 2b and Additional file 1: Figure S14b, as well as Additional file 1: Figure S15 and S16 for visual inspection in a low-dimensional space). We could already separate inhibitory from excitatory neurons using only the top 100 HVFs obtained by scMET, and generally resulted in more distinct cell sub-populations. In all cases, promoters were less able to disentangle the neuronal sub-populations. This is consistent with the lower overdispersion levels observed in this genomic context (Additional file 1: Figure S14c).

To gain a better understanding about the HVF selection implemented by each method, an additional comparison is provided in Additional file 1: Figure S17a. Similar to scMET, the NormDisp approach selects HVFs that are not confounded by the mean-overdispersion relationship. However, NormDisp HVFs exhibit poorer clustering performance. This potentially occurs due to the normalized dispersion point estimates being more unstable. For instance, when analyzing the discrepancies among the top 200 HVFs selected by NormDisp and scMET, we observe that the top three HVFs called by NormDisp (but not called by scMET) appear to be driven by outliers (Additional file 1: Figure S17b). In contrast, the top HVFs called by scMET (but not called by NormDisp) show a bimodal distribution and are able to better recapitulate the underlying population structure (Additional file 1: Figure S17c).

To facilitate interpretation for the HVFs highlighted by scMET, we linked genomic features to genes by overlapping the genomic coordinates allowing for a maximum distance of 20 kb from the transcription start site in the case of distal elements. We explored whether features identified as HVF (red points in Fig. 2c and Additional file 1: Figure S18a) were enriched for neuronal markers identified in the [9] study (Additional file 2: Table S1). This enrichment was observed for distal H3K27ac and H3K4me1 marks, but not for promoters (Additional file 1: Figure S18b). As representative examples, we display three distal H3K27ac elements among the HVFs that are located proximal to known gene markers of each neuron class: *Gad1* for inhibitory and *Neurod6* and *Sv2b* for excitatory (Fig. 2d).

Finally, we also explored the use of 20-kb sliding windows as input features for scMET (Additional file 1: Section S2.5). After quality control, this led to $J = 82,308$ features. HVFs selected among these windows led to similar results as when using pre-annotated H3K27ac and H3K4me1 features as input (Additional file 1: Figure S19a and S19e). This suggests that the selected windows can distinguish most of the different cellular sub-populations. However, the slightly lower ARI values suggest that pre-annotated features can better recapitulate the underlying population structure. Among the windows with the largest residual overdispersion, we identified putative regulatory elements that discriminate cell types and that are located near neuronal marker genes identified by [9] (Additional file 1: Figure S19b-S19d). As an example, chr1:56940001-56960001 is located within the intronic region of *Satb2*, a marker of excitatory neurons. As expected, this region is unmethylated in most excitatory neurons and highly methylated in inhibitory neurons. In contrast, chr1:68780001-68800001 is located within the intronic region of *Erbb4*, a marker of inhibitory neurons. This region is highly methylated in all excitatory neurons and only partially methylated in inhibitory neurons. These results suggests that scMET can provide a meaningful characterization for the role of previously un-annotated genomic regions.

scMET enables differential methylation testing between cellular sub-populations

To showcase scMET as a differential methylation tool, we applied it on the same mouse frontal cortex dataset [9], after separating the cells in excitatory and inhibitory groups. Initially, we applied scMET to characterize differential methylation (DM), i.e., changes in mean methylation. Across all genomic contexts, we observed a substantially larger fraction of features being hyper-methylated in inhibitory compared to excitatory neurons (Fig. 3a and Additional file 1: Figure S20a). Within distal H3K27ac peaks, for instance, scMET identified 5242 features to have higher methylation levels in inhibitory neurons, compared to only 935 features showing higher methylation in the excitatory group (Fig. 3a, b). After mapping features to their nearest gene, we observed that DM hits were enriched for known marker genes that differentiate inhibitory and excitatory neurons (Additional file 1: Figure S20b).

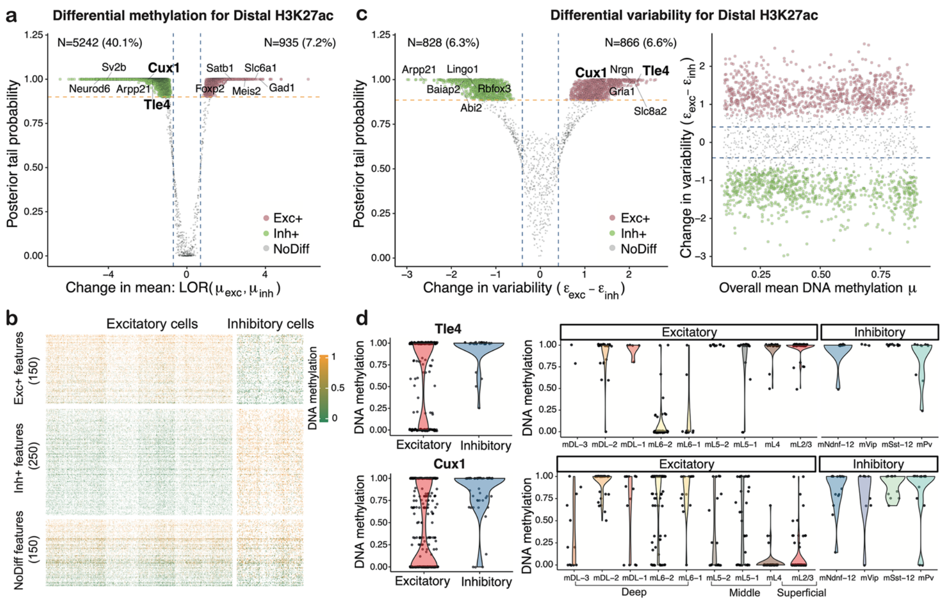


Fig. 3 Summary of changes in methylation patterns (mean and variability) for inhibitory and excitatory neurons. **a** Identifying differentially methylated (DM) features for the distal H3K27ac genomic context. Green and pink points correspond to features showing higher mean methylation in inhibitory (Inh+) and excitatory (Exc+) neurons, respectively. Labels highlight neuron marker genes that were used in [9] to define the different cell populations. Blue vertical dashed lines correspond to log-odds ratio threshold $\psi_M = \pm \log(2)$. Yellow horizontal dashed line is located at posterior evidence probability cut-offs defined by EFDR = 5%. **b** Representative heatmap of methylation rates (Y_{ij}/n_{ij}) across cells (columns) and features (rows) for the distal H3K27ac genomic context. Cells are grouped in excitatory and inhibitory classes. A set of randomly selected features is displayed. These are grouped according to the DM analysis output as: Exc+, Inh+, and no mean methylation difference. The color code represents features with low (0, green) and high (1, yellow) mean methylation level. Features with no CpG coverage are denoted with white color. **c** Identifying differentially variable (DV) features for the distal H3K27ac genomic context. Green and pink points correspond to features showing higher methylation variability in inhibitory and excitatory neurons, respectively. Blue dashed lines correspond to log-odds ratio threshold $\psi_E = \pm \log(1.5)$. Yellow dashed line is located at posterior evidence probability cut-offs defined by EFDR = 5%. For each feature, posterior estimates for the change in residual overdispersion parameter ϵ_j between excitatory and inhibitory neurons is plotted against the posterior tail probability of calling a feature as DV (left). For each feature, posterior estimates for mean methylation parameter μ_j is plotted against posterior estimates for the change in residual overdispersion parameter ϵ_j between excitatory and inhibitory neurons (right). **d** Example features that are called as being more variable in excitatory neurons. Left subplots show broad differences in methylation patterns. Right subplots show methylation patterns separately within each broad neuronal class. Each data point represents the methylation rate for a cell

Besides DM testing, the primary focus of the scMET differential test is to identify changes in cell-to-cell methylation variability. In principle, differential variability (DV) testing could be based on feature-specific overdispersion parameters γ_j, but these results would be confounded by the mean-overdispersion trend (Fig. 1c). Hence, meaningful DV analysis based on γ_j would need to be restricted to non-DM features. Instead, we propose to perform DV analysis based on residual overdispersion parameters ϵ_j. For the mouse frontal cortex dataset, we identified a large number of DV features across genomic contexts, except from promoter regions which showed tighter methylation patterns across inhibitory and excitatory neurons (Fig. 3c and Additional file 1: Figure S21a). Critically, the procedure for calling DV features was not confounded by mean methylation levels (Fig. 3c and Additional file 1: Figure S21b).

As representative examples, we show two distal H3K27ac peaks that are located proximal to neuronal markers and exhibit higher variability in excitatory neurons (Fig. 3d). Both features show substantial variation across the different sub-types of excitatory neurons: the first is mostly unmethylated in the mL6 cortical layer, and the second is mostly unmethylated in the superficial cortical layer. These patterns are consistent with previously reported spatial expression for *Tle4* (mostly expressed in the deep cortical layer, see [30, 31]) and *Cux1* (which shows expression specificity for the superficial layer, see [30, 32]). It should be noted that these features would be excluded from DV analysis based on γ_j, since they are also DM between the two broad classes (Fig. 3a). In summary, these findings demonstrate the ability of scMET to identify potential markers that drive between and within cell population heterogeneity.

Exploring the relationship between transcriptional and DNAm variability using single-cell multi-omics data

As a second use case, we considered a single-cell multi-omics dataset where scNMT-seq [33] was employed to profile RNA expression, DNAm, and nucleosome occupancy at single-cell resolution, spanning multiple time points from the exit from pluripotency to primary germ layer specification [34]. The multi-modal nature of this dataset provides a unique opportunity to link cell-to-cell variation between DNAm and transcription across individual cells. Here, we used scMET to quantify DNAm variability at promoter elements, which we subsequently contrasted to RNA expression heterogeneity for the corresponding genes. For this analysis, we exclusively used promoter elements as, unlike distal regulatory elements, they can be unambiguously matched to their respective genes.

For each gene, we quantified transcriptional heterogeneity using the residual overdispersion estimates generated by BASiCS (Additional file 1: Figure S22a, [22]). Promoter DNAm variability was calculated using the residual overdispersion estimates inferred using scMET (Additional file 1: Figure S22b). More details about these analyses and the associated data pre-processing steps are described in the “Methods” section.

When comparing residual overdispersion estimates for RNA expression and promoter DNAm, there was no clear genome-wide association (Fig. 4a). However, when restricting to genes that display high levels of transcriptional variability, two main groups can be identified. The first category corresponds to genes with low levels of promoter DNAm residual overdispersion, and it includes differentiation and germ layer markers such as *Mesp1*, *Lefty2*, and *Id3* (mesoderm) and *Cldn6*, *Cer1*, and *Krt8* (endoderm). The second category is characterized by genes with high promoter DNAm residual overdispersion and includes known pluripotency markers such as *Dppa5a*, *Zfp42*, *Spp1*, and *Peg3*. Representative examples for these genes are displayed in Fig. 4b.

These results suggest the presence of two modes of regulation. On the one hand, downregulation of pluripotency genes is associated with high promoter DNAm heterogeneity, linked to a pronounced increase in promoter DNA methylation throughout the embryonic stages. On the other hand, upregulation of differentiation genes is not linked to high levels of promoter DNAm variability. This suggests that other genomic contexts or molecular layers might be responsible for their activation [34]. Finally, we also find genes with low RNA expression variability that display high levels of promoter DNAm heterogeneity (Additional file 1: Figure S22c), suggesting that the coupling between

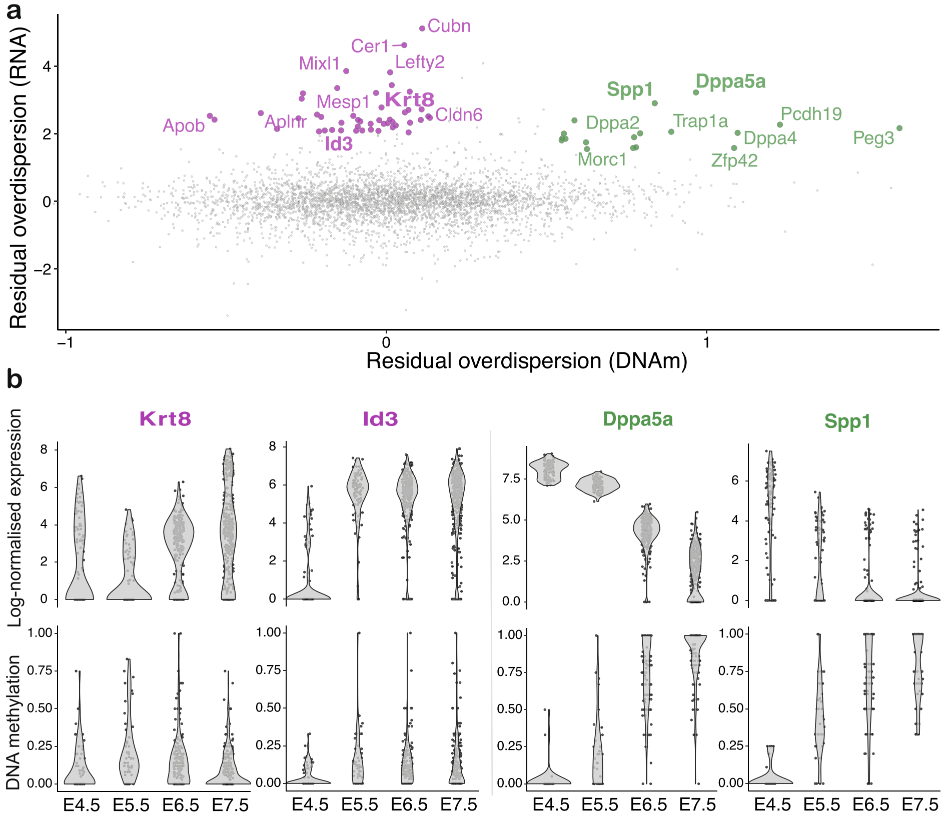


Fig. 4 scMET applied to the multi-omics scNMT-seq gastrulation dataset reveals a complex linkage between promoter DNAm and RNA expression during embryonic development. **a** Scatter plot displays posterior median estimates for residual DNAm overdispersion parameters ϵ_j in gene promoters (*x*-axis) versus RNA residual overdispersion of the corresponding genes (*y*-axis). Among the genes with high levels of RNA heterogeneity, green and pink points correspond to promoters showing high and low levels of DNAm variability, respectively. **b** Representative examples of DNAm and RNA expression patterns across developmental stages for genes with high transcriptional heterogeneity and low (left, pink) or high (right, green) DNAm heterogeneity. The *y*-axis shows BASiCS log-normalized gene expression (in a $\log(x+1)$ scale) (top) and promoter DNAm rate (bottom). Cells are stratified by embryonic stage (*x*-axis)

promoter DNAm and transcriptional activity is more complex than previously acknowledged during embryonic development stages [35].

Discussion

Single-cell DNAm assays can currently profile hundreds to thousands of DNA methylomes, with increasingly complex experimental designs. The high resolution of these measurements enables us to measure cell-to-cell epigenetic variability, as well as uncover the regulatory features that modulate it [36]. However, the noise and biases intrinsic to such technologies create a need for computational frameworks that can systematically interrogate the data generated, dissecting genuine variability and quantifying uncertainties.

In this study, we introduced scMET, a statistical framework for modeling DNA methylation heterogeneity from scBS-seq data. Using a hierarchical Bayesian framework to borrow information across cells and features, scMET robustly quantifies genuine cell-to-cell variability. Our results demonstrated the ability of scMET in highlighting genomic features that drive cell-to-cell heterogeneity across neuronal sub-populations in a large dataset of single-cell methylomes from the mouse frontal cortex. Furthermore, scMET can be used as a quantitative tool to interrogate changes in DNAm patterns between pre-specified cell populations. Unlike common approaches that only detect changes in mean methylation levels [20, 37], scMET can also identify features with differences in DNAm variability between populations. Importantly, the differential variability estimates are quantified through residual overdispersion parameters, thus accounting for the known confounding relationship between mean and overdispersion in scBS-seq datasets.

Recently, complementary approaches have been proposed to analyze single-cell DNAm data. These include Melissa [13] and Epiclomal [38], which are probabilistic clustering and imputation methods based on a hierarchical mixture model. In addition, MAPLE [39] uses a supervised learning approach to detect correlations between RNA expression and DNAm, which enables cell type assignments by transferring labels from scRNA-seq to single-cell DNAm datasets. In all cases, scMET could be incorporated in the analysis pipeline, by selecting a relevant set of input features defined by identifying regions with high DNAm variability which can better characterize the underlying population structure.

scMET uses a GLM framework to explicitly model known biases in the data in the form of additional covariates, such as CpG content. The flexibility of the GLM approach enables it to easily incorporate additional features, such as DNA motifs, which could be important to elucidate the role of sequence or chromatin state in modulating DNA methylation. Additionally, the framework could readily be extended to model joint variability in multiple molecular layers (such as transcriptome and methylome), opening a path to new methodologies in integrative, single-cell multi-omics analyses. Given the increasing prominence of such studies, we expect scMET to become an important tool in the extraction of biological signals from DNAm datasets of increasing complexity.

Conclusions

We presented scMET — a Bayesian hierarchical approach to robustly quantify epigenetic heterogeneity using high-throughput single-cell DNAm datasets. We extensively evaluated the performance of scMET using synthetic data and recent large-scale assays in the context of the mouse brain cortex and during early development. In particular, we have shown how feature selection by scMET can be incorporated in single-cell DNAm analysis pipelines, improving the characterization of epigenetically distinct cell types. Moreover, we introduced a novel differential methylation framework which, unlike existing approaches, fully exploits the cellular resolution of the data to identify changes in epigenetic variability. In combination, these findings demonstrate how the integrative approach implemented in scMET can overcome data sparsity and improve the quantification of genuine epigenetic cell-to-cell heterogeneity.

Methods

The scMET model

Let Y_{ij} represent the number of methylated CpGs out of the n_{ij} CpGs for which DNAm was measured for genomic feature $j \in \{1, \ldots, J\}$ in cell $i \in \{1, \ldots, I\}$ (see Additional file 1: Figure S1). These genomic features could be defined by pre-annotated regions (e.g., enhancers) or other regions of interest. To capture data overdispersion, scMET assumes a beta-binomial (BB) hierarchical formulation (see also Additional file 1: Section S2.1):

$$Y_{ij} \mid \theta_{ij} \sim \text{Binomial}(n_{ij}, \theta_{ij}), \qquad \theta_{ij} \mid \mu_j, \gamma_j \sim \text{Beta}(\mu_j, \gamma_j). \tag{1}$$

In Eq. (1), the beta distribution is parameterized such that $\mathbb{E}[\theta_{ij}] = \mu_j$ and $\text{Var}[\theta_{ij}] = \gamma_j^2 \mu_j \left(1 - \mu_j(1-\gamma_j) - \gamma_j\right)(1-\gamma_j)^{-1}$, with $\mu_j \in (0,1)$ and $\gamma_j \in (0,1)$. If $\gamma_j = 0$, the model in Eq. (1) reduces to a binomial model with parameters n_{ij} and μ_j. After integrating out the random effects θ_{ij}, it can be seen that μ_j corresponds to the mean methylation across all cells for feature *j* and that γ_j controls the *overdispersion* that is not captured by binomial sampling. In fact, the BB variance can be written as:

$$\text{Var}[Y_{ij} \mid \mu_j, \gamma_j] = \underbrace{n_{ij}\mu_j(1-\mu_j)}_{\text{technical variation}} + \underbrace{n_{ij}\mu_j(1-\mu_j)(n_{ij}-1)\gamma_j}_{\text{additional (biological) variation}}. \tag{2}$$

Parameters μ_j and γ_j can be inferred via maximum likelihood estimation. However, due to the high sparsity and noise present in single-cell DNAm data, these estimates can be unstable, especially for overdispersion parameters γ_j (Additional file 1: Figure S6 and S7). To overcome this, we use a Bayesian framework with a hierarchical prior specification for μ_j and γ_j, sharing information across sets of similar types of genomic features (e.g., enhancers). Our approach is flexible and can incorporate feature-specific covariates $\mathbf{x}_j$ that explain differences in mean methylation across features. For instance, features with high CpG density tend to have lower methylation levels. These covariates are introduced within a generalized linear model (GLM) framework through the prior on the mean methylation parameters μ_j:

$$\mu_j \mid \mathbf{w}_\mu, s_\mu, \mathbf{x}_j \sim \text{Logit}\mathcal{N}\left(f_\mu(\mathbf{x}_j; \mathbf{w}_\mu), s_\mu\right), \quad \text{where } f_\mu(\mathbf{x}_j; \mathbf{w}_\mu) = \mathbf{w}_\mu^\top \mathbf{x}_j. \tag{3}$$

In Eq. (3), Logit$\mathcal{N}$ denotes a logit-normal distribution, $\mathbf{w}_\mu$ is a vector of regression coefficients, and s_μ is the standard deviation for logit(μ_j). Throughout our analyses, we assume $\mathbf{x}_j = (1, C_j)$, where C_j denotes the CpG density for feature *j*. However, scMET is flexible and users can introduce other feature-specific covariates.

Our prior specification is also designed to capture the mean-overdispersion relationship that is typically observed in the data generated by high-throughput sequencing assays, such as scBS-seq (Fig. 1c). Here, we follow the approach in [22], introducing a non-linear regression model through an informative prior for γ_j:

$$\gamma_j \mid \mu_j, \mathbf{w}_\gamma, s_\gamma \sim \text{Logit}\mathcal{N}(f_\gamma(\mu_j; \mathbf{w}_\gamma), s_\gamma), \quad \text{where } f_\gamma(\mu_j; \mathbf{w}_\gamma) = w_{\gamma 1} + \sum_{l=2}^{L} w_{\gamma l} g_l(\mu_j). \tag{4}$$

Here, $f_\gamma(\mu_j; \mathbf{w})$ can be interpreted as the overdispersion (logit scale) that is predicted by mean methylation levels μ_j (fitted black line in Fig. 1c), $g_l(\mu_j)$ represent radial basis function kernels (defined as in [40]), and $w_{\gamma 1}, \ldots, w_{\gamma, L}$ are regression coefficients. Unless otherwise stated, we use $L = 4$ throughout our analyses. The remaining elements of the prior are described in Additional file 1: Section 2.2.

The prior distribution in Eq. (4) can be rewritten as a non-linear regression model

$$\text{logit}(\gamma_j) = f_\gamma(\mu_j; \mathbf{w}_\gamma) + \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, s_\gamma), \tag{5}$$

where ϵ_j corresponds to a feature-specific *residual overdispersion* parameter that captures deviations from the overall trend. Hence, a feature that exhibits positive ϵ_j values has more variation than expected for features with similar mean methylation. Accordingly, negative ϵ_j values indicate less variation than expected for features with similar mean methylation.

Implementation

The posterior distribution for the model parameters in scMET is not amenable to analytic solutions. Hence, we resort to variational Bayes (VB, [23]) and Markov Chain Monte Carlo (MCMC, [41]) implementations using the Stan probabilistic programming language [42]. scMET is publicly available as an R package at https://github.com/andreaskapou/scMET and will be shortly submitted to Bioconductor.

Identifying highly variable features

Residual overdispersion parameters ϵ_j can be used to label highly variable features (HVFs) within a population of cells. Our decision rule is based on tail posterior probabilities [24] associated to whether ϵ_j exceed a pre-specified threshold ϵ_0:

$$\pi_j^E(\epsilon_0) \equiv P(\epsilon_j > \epsilon_0 \mid \text{data}), \tag{6}$$

As a default choice, we define ϵ_0 based on the distribution of posterior estimates for residual overdispersion parameters ϵ_j across all features. In particular, we define ϵ_0 to match the δ_E-th percentile of the distribution. Unless otherwise stated, we set as default $\delta_E = 0.9$.

The probabilities in Eq. (6) can be estimated by counting the proportion of posterior draws (obtained by VB or MCMC) for which the chosen criteria are met [43]. scMET labels as HVFs those for which their associated tail posterior probabilities are above a given posterior evidence threshold α_H $(0.6 < \alpha_H < 1)$, where α_H is calibrated via the expected false discovery rate (EFDR; [25]), see also Additional file 1: Section S2.3.

Differential testing

scMET provides a similar probabilistic rule to label differentially methylated (DM) and differentially variable (DV) features across experimental conditions or cell types (Fig 1e). Here, we define DM features as those for which mean methylation varies across the groups of cells under study. More concretely, let μ_j^A and μ_j^B be the mean methylation parameters associated with feature *j* in groups A and B. We quantify differences in mean methylation as the log-odds ratio (LOR):

$$\text{LOR}(\mu_j^A, \mu_j^B) = \text{logit}(\mu_j^A) - \text{logit}(\mu_j^B). \tag{7}$$

Similar to the HVF analysis, our decision rule for DM testing is defined as:

$$\pi_{jAB}^M(\psi_M) \equiv P(|\text{LOR}(\mu_j^A, \mu_j^B)| > \psi_M \mid \text{data}) > \alpha_M, \tag{8}$$

where α_M $(0.6 < \alpha_M < 1)$ is a posterior evidence threshold chosen to match a desired EFDR level and ψ_M is a LOR threshold which can be interpreted as a minimum effect size to be detected by the test. As default, we use $\psi_M = \log(2)$, i.e., a two-fold change in odds ratio.

Beyond highlighting DM features, scMET embraces the cellular resolution of scBS-seq data to perform differential variability (DV) analyses, identifying changes in cell-to-cell DNAm variability across groups. Although overdispersion parameters γ_j could be used as the input for the DV test, the results would be confounded by the mean-overdispersion relationship that is typically observed within each genomic context (Fig. 1c). Instead, we propose to perform DV analysis based on ϵ_j — a measure of cell-to-cell DNAm variability that is not confounded by differences in mean methylation. Let γ_j^A and γ_j^B denote the overdispersion parameters linked to feature *j* in groups A and B. To label DV features based on residual overdispersion, we make use of Eq. (5), and decompose the LOR between γ_j^A and γ_j^B parameters as:

$$\mathrm{LOR}(\gamma_j^A, \gamma_j^B) = \underbrace{f_\mu^A(\mathbf{x}_j; \mathbf{w}_\mu^A) - f_\mu^B(\mathbf{x}_j; \mathbf{w}_\mu^B)}_{\text{mean contribution}} + \underbrace{\epsilon_j^A - \epsilon_j^B}_{\text{residual change}} \quad . \tag{9}$$

In Eq. (9), the first term captures the changes in overdispersion that are explained by mean methylation and the second term captures residual overdispersion changes after accounting for the mean methylation. This residual change is used to identify features with statistically significant differences in residual overdispersion. For a given posterior evidence threshold α_E ($0.6 < \alpha_E < 1$) and tolerance threshold ψ_E, the following rule is used to identify DV features:

$$\pi_{jAB}^E(\psi_E) \equiv P(|\epsilon_j^A - \epsilon_j^B| > \psi_E \mid \text{data}) > \alpha_E. \tag{10}$$

As default, we set $\psi_E = \log(1.5)$, i.e., 50% change in overdispersion LOR between the groups. As above, the posterior evidence threshold α_E is calibrated via the EFDR, see Additional file 1: Section 2.3.

Alternative methods

When comparing scMET to existing DNAm analysis tools (primarily developed for bulk assays), it is important to emphasize fundamental differences with respect to the functionality provided by scMET. These are summarised in Table S1 and briefly described below.

Parameter estimation

The BB MLE method corresponds to estimating the parameters of the beta-binomial model in Eq. (1) independently per feature using maximum likelihood. The `VGAM` package was used for parameter estimation [44].

Differential mean methylation analyses

When comparing methylation profiles between experimental conditions or cell types, bulk DNA methylation analysis tools primarily focus on differences in mean methylation. Existing methods are often based on some version of a *t*-test (e.g., DMRcate [45], BSmooth [37]); a Fisher test on the methylation counts (e.g., MethylKit [46], RnBeads [47]); negative binomial models (often adapted from the bulk RNA sequencing literature, e.g., edgeR [48]); or on beta-binomial models (e.g., DSS [20], RADMeth [21]).

For differential mean methylation testing on the synthetic datasets, we compared scMET with Fisher's exact test. Features with log-odds ratio > log(1.5) between the two groups and FDR <10% (Benjamini-Hochberg procedure) were called as differentially methylated.

Differential methylation variability analyses

In the context of microarray DNAm, DiffVar [49] was proposed to perform DNAm variability analyses, with an empirical Bayes framework to stabilize the t-statistics when contrasting the variance of the continuous methylation rates. This approach can also be applied to bulk bisulphite data, as methylation rates can be confidently estimated when having a large number of observations. However, this is not appropriate for sparse measurements (i.e., in single-cell data), as there is high uncertainty in their estimation.

Instead, scMET directly models methylation read-counts, using a hierarchical framework to address data sparsity and to propagate statistical uncertainty. It is also important to note that single-cell bisulphite data displays a strong dependency between mean and variability estimates (see, e.g., Additional file 1: Figure S13 and S23). Hence, direct downstream analyses of methylation variability would be confounded by mean methylation. Instead, scMET explicitly derives a residual measure of variability (residual overdispersion) that is not confounded by mean methylation and that can be directly employed in downstream tasks such as HVF selection and differential variability.

HVF selection

Feature selection has traditionally been done either using the binomial variance or the Gaussian variance as variability estimates [33, 50]. The binomial model is where features are ranked according to binomial variance given by $1/I \sum_i \theta_{ij}(1-\theta_{ij})$, where $\theta_{ij} = Y_{ij}/n_{ij}$ is the methylation rate for feature j in cell i. The Gaussian model on methylation rates $\theta_{ij} \sim \mathcal{N}(\mu_j, \sigma_j)$ is where features are ranked according to σ_j. The normalized dispersion (NormDisp) method is widely applied for scRNA-seq data [28, 29]. Briefly, we calculated the mean and a dispersion measure (variance/mean) for each gene across all single cells and placed genes into 20 bins based on their mean expression. Normalized dispersion is calculated as the absolute difference between dispersion and median dispersion of the expression mean, normalized by median absolute deviation within each bin [29]. Additionally, we also include a random selection approach in which HVFs are selected at random among all input features. Finally, we note that, to our knowledge, none of the beta-binomial or negative binomial methods discussed above implements a strategy to perform HVF selection.

Scalable strategy for genome-wide sliding windows

By pooling information across cells and features, scMET leads to more robust parameter estimates than simpler methods which analyze each feature independently. We addressed the increased computational complexity by adopting a variational Bayes algorithm which scales linearly with the number of features, enabling the analysis of large-scale data sets (Additional file 1: Figure S2). However, additional challenges arise due to the increased dimensionality introduced by using a sliding window approach on a genome-wide scale. For example, for the mouse genome, using sliding windows of 20 kb with a step size of 20 kb yields approximately 130,000 features. This may preclude the use of scMET within a practical timeline for large-scale applications. In the spirit of divide-and-conquer schemes [51], we bypass this problem via a parallelization strategy in which we apply scMET separately to each chromosome. Feature-specific estimates obtained for each chromosome can be combined post hoc when performing HVF selection and differential analyses.

We suggest careful consideration when selecting the input feature set while using the sliding window approach. First, it is critical to apply a quality control step to remove windows with very low coverage (number of CpGs whose methylation status was observed), for which inference is less reliable. Second, the choice of window size can restrict the type of regions to be included in the analysis. For example, regulatory elements such as enhancers tend to be relatively small, spanning from a hundred to a few thousand base pairs [52]. Such regions would be excluded when using large window sizes. In contrast, small windows may fail to have sufficient coverage. Data-driven approaches could be used to identify an optimal window size. For example, [53] identified that 2-kb windows are sufficient to capture regions with high CpG-to-CpG concordance. An alternative solution would be to iteratively refine the window size while performing downstream analyses. This was implemented in *metilene* [54] in the context of differential mean methylation testing for bulk DNAm sequencing data. Implementation of such iterative strategy for differential variability testing and HVF selection while taking into account the sparsity of single-cell DNAm data is a very interesting direction for future research.

Simulation study

We simulated J = 300 features for varying number of cells ranging from I = 20 up to I = 1000. To mimic the properties observed in real scBS-seq data, we assume that for each feature we have coverage for a subset of cells given by $I_j \sim \text{Binomial}(I, p_j)$, where $p_j \sim \text{Uniform}(0.4, 0.8)$ to generate diverse I_j across features. We also simulate three alternative regions that have rich (N = 50), moderate (N = 15), and poor (N = 8) CpG density. That is, the number of CpGs (n_{ij}) is simulated from $\text{Binomial}(N, q_j)$, where $q_j \sim \text{Uniform}(0.4, 0.8)$ to generate a broad range of CpG coverage across features. Next, for each feature, we generate mean methylation parameters $\mu_j \sim \text{Logit}\mathcal{N}(\mathbf{w}_\mu^\top \mathbf{x}_j, 1)$, where $\mathbf{w}_\mu = (-0.5, -1.5)$ and $\mathbf{x}_j = (1, C_j)$ are feature-specific covariates, where C_j denotes the CpG density. The negative weight on $\mathbf{w}_\mu$ is used to simulate the known negative association between mean methylation and CpG density. Next, we simulated feature-specific overdispersion parameters $\gamma_j \sim \text{Logit}\mathcal{N}(\mathbf{w}_\gamma^\top \mathbf{g}_j(\mu_j), 0.25)$ to mimic the mean-overdispersion relationship. We set $\mathbf{w}_\gamma = (-1.2, -.3, 1.1, -.9)$ and $\mathbf{g}_j(\mu_j)$ is a vector of basis functions with methylation level μ_j. Finally, we simulated the number of methylated CpGs from BB distribution, using the `VGAM` package, that is, $Y_{ij} \sim \text{BB}(n_{ij}, \mu_j, \gamma_j)$.

For differential testing analysis, we used the above approach to generate cells from the first group (group A). For DM analysis, 15% of features were randomly selected and their corresponding μ_j were randomly increased or decreased by three different LOR thresholds: 2, 3, and 5, to generate cells from the second group (group B). Similarly, for DV analysis, 15% of features were randomly selected from the first group and their corresponding γ_j were randomly increased or decreased by three different LOR thresholds: 2, 3, and 5, to generate cells from the second group.

Mouse frontal cortex dataset

Data processing

The dataset is available from the Gene Expression Omnibus repository under accession number GSE97179. Details on quality control and data pre-processing can be found in [9]. Additional file 3: Table S2 contains metadata for the 3069 cells, such as cell type annotations and cortical layer information. We aggregated closely related cellular subpopulations with less than 25 cells, following the hierarchy established in [9]. DNA methylation was quantified using mCG dinucleotides over three genomic contexts: (1) gene promoters (±2-kb windows around the transcription start sites of genes extracted from ENSEMBL version 87, [55]), (2) Distal H3K27ac ChIP-seq peaks, and (3) H3K4me1 ChIP-seq peaks. The latter was based on two ChIP-seq datasets that were profiled in adult (8 weeks) mouse cortex as part of the ENCODE project (see Additional file 4: Table S3).

For each genomic feature $j \in \{1, \ldots, J\}$ and cell $i \in \{1, \ldots, I\}$, the following censoring procedure was applied: we recorded Y_{ij} as a missing value if methylation coverage was available in less than 3 CpGs (i.e., $n_{ij} < 3$). The purpose of this censoring step was to exclude observations with very low coverage for which DNAm quantification is less reliable. Subsequently, we removed features that did not have CpG coverage in at least 15 cells. In addition, we excluded features that had mean methylation across cells lower than 0.1 or higher than 0.9; the rationale being that fully (un)methylated features do not drive methylation heterogeneity and will not provide information for identifying cell sub-populations.

Down-sampling experiment

Using the characterized sub-populations from [9], we performed down-sampling experiments on 424 inhibitory neurons. Both scMET and BB MLE methods were run once on the full dataset (424 cells) to generate pseudo-ground truth parameter estimates. Subsequently, 20, 50, 100, and 200 cells were randomly down-sampled from the full population prior to parameter estimation. This procedure was repeated 5 times for each sample size. The same censoring step as described above was applied. Moreover, due to smaller sample sizes, we filtered genomic features that did not have CpG coverage in at least 5 cells.

HVF analysis

HVF analysis was applied on 12,774 gene promoters, 17,284 distal H3K27ac peaks, and 30,374 H3K4me1 peaks. To model the mean-overdispersion relationship, we used $L = 4$ radial basis function kernels and kept default hyper-parameter values (Additional file 1: Section S2.2). The total number of iterations was set to 50,000 and convergence was attained when the evidence lower bound difference between two consecutive iterations was less than 1e−04.

Differential analysis

For differential analysis between excitatory and inhibitory neurons, we only included features with CpG coverage in at least 15 cells, in both sub-populations. This resulted in 12,611 gene promoters, 13,075 distal H3K27ac peaks, and 20,212 H3K4me1 peaks. To model the mean-overdispersion relationship, we used $L = 4$ radial basis function kernels and kept default hyper-parameter values (Additional file 1: Section S2.2). The total number of iterations was set to 50,000 and convergence was attained when the evidence lower bound difference between two consecutive iterations was less than 1×10^{-4}.

Dimensionality reduction

Dimensionality reduction was applied using a Bayesian Factor Analysis algorithm, as implemented in the `MOFA2` package [56]. The motivation for this method, as opposed to the conventional Principal Component Analysis, is to handle the large presence of missing values without the need for imputation. A second (non-linear) dimensionality reduction step was applied using UMAP (as implemented in the `umap` package) to project the data into a two-dimensional space (Additional file 1: Figure S15 and S16).

Clustering

A finite grid of HVFs (from 50 to 1000 with step size of 50) was selected by each of the competing methods. Subsequently, clustering analysis was performed using the k-means algorithm on the latent space defined by the `MOFA` factors (fixed to 15). The number of clusters was set to the number of cell types as characterized by [9]. We assessed clustering performance using the ARI and cluster purity. A non-parametric regression (implemented by the `loess` function) was used to obtain a smoothed interpolation across all HVF values.

scNMT-seq gastrulation dataset

Data processing

The parsed scNMT-seq gastrulation dataset was downloaded from https://ftp://ftp.ebi.ac.uk/pub/databases/scnmt_gastrulation. Raw files are available from the Gene Expression Omnibus repository under accession number GSE121708. Details on the quality control and data processing can be found in [34]. We selected all cells from E4.5 to E7.5 days after excluding the extra-embryonic visceral endoderm cells, as they display distinct DNA methylation profiles. Additional file 5: Table S4 contains sample metadata for the 848 cells retained for analysis. DNA methylation was quantified over gene promoters (±2kb windows around the transcription start sites of genes extracted from ENSEMBL version 87, [55]).

Calculation of DNA methylation and RNA expression heterogeneity

For the DNAm data, we applied the same censoring procedure and feature exclusion criteria as described in the pre-processing of the [9] dataset. This resulted in 13,785 gene promoters for downstream analysis. Residual overdispersion estimates were calculated by scMET with default parameter values using the same number of iterations and convergence criteria described above.

For the RNA expression data, we removed lowly expressed genes (no counts in less than 10 cells and average count across expressed cells less than 5). This resulted in 14,076 genes for downstream analysis. Residual overdispersion estimates were obtained using BASiCS [22]. The algorithm was run using 20,000 iterations, applying a burn in of 10,000 and thinning of 10. An empirical Bayes approach was used to derive the prior hyper-parameters associated to gene-specific mean expression parameters within BASiCS. In the comparison displayed in Fig. 4, we focused on the 10,192 genes contained in the intersection of the lists obtained above.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02329-8.

Additional file 1: Supplementary material. Supplementary figures (Section 1) and supplementary notes (Section 2).

Additional file 2: **Table S1**. Table with marker genes from the Luo2017 [9] study.

Additional file 3: **Table S2**. Table with sample metadata from the Luo2017 [9] study.

Additional file 4: **Table S3**. Table with ChIP-seq annotation information.

Additional file 5: **Table S4**. Table with sample metadata from the Argelaguet2019 [34] study.

Additional file 6: Review history.

Acknowledgements

We thank Oliver Stegle and John C. Marioni for constructive feedback during early stages of this project.

Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 6.

Authors' contributions

All authors conceived and planned the methods and experiments. CAK developed the statistical model and software implementation. CAK and RA analyzed the data and wrote the manuscript, with input from all authors. GS and CAV supervised the study. All authors contributed to the interpretation of results and approved the final manuscript.

Funding

C.A.K. is a cross-disciplinary post-doctoral fellow supported by funding from the University of Edinburgh and Medical Research Council (MC_UU_00009/2). C.A.V. is a Chancellor's Fellow funded by the University of Edinburgh. R.A. was funded by the European Molecular Biology Laboratory (EMBL) international PhD program.

Availability of data and materials

All datasets analyzed in this article are publicly available. The mouse cortex dataset [9] is available under GEO accession number GSE97179. The parsed scNMT-seq gastrulation dataset [34] was downloaded from https://ftp://ftp.ebi.ac.uk/pub/databases/scnmt_gastrulation. Raw files are available under GEO accession number GSE121708.

The scMET model is publicly available as R software released under the GNU GPL-3 license (see https://github.com/andreaskapou/scMET [57] and https://doi.org/10.5281/zenodo.4629327 [57]). The code used to process and analyze the data is available at https://github.com/andreaskapou/scMET-analysis. The following software versions were used throughout the analyses: `scMET` (0.99.1), `BASiCS` (2.0.1), `coda` (0.19.3), `MOFA2` (0.99.7), `rstan` (2.19.3), `uwot` (0.1.10), and `VGAM` (1.1.3).

Declarations

Ethics approval and consent to participate

Ethical approval was not needed for this study.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [2]School of Informatics, University of Edinburgh, Edinburgh, UK. [3]European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. [4]SISSA, International School of Advanced Studies, Trieste, Italy. [5]The Alan Turing Institute, London, UK.

Received: 22 September 2020 Accepted: 25 March 2021
Published online: 20 April 2021

References

1. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals,. Nat Genet. 2003;33(March):245–54. https://doi.org/10.1038/ng1089.
2. Avner P, Heard E. X-chromosome inactivation: counting, choice and initiation. Nat Rev Genet. 2001;2(1):59.
3. Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. Nat Rev Genet. 2011;11(10):726–34. https://doi.org/10.1038/nrc3130.
4. Reik W, Walter J. Genomic imprinting: parental influence on the genome. Nat Rev Genet. 2001;2(1):21.
5. Krueger F, Kreck B, Franke A, Andrews SR. DNA methylome analysis using short bisulfite sequencing data,. Nat Methods. 2012;9(2):145–51. https://doi.org/10.1038/nmeth.1828.
6. Schwartzman O, Tanay A. Single-cell epigenomics: techniques and emerging applications,. Nat Rev Genet. 2015;16(12):716–26. https://doi.org/10.1038/nrg3980.
7. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014;11(8): 817–20. https://doi.org/10.1038/nmeth.3035.
8. Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 20132126–35. https://doi.org/10.1101/gr.161679.113..
9. Luo C, Keown CL, Kurihara L, Zhou J, He Y, Li J, Castanon R, Lucero J, Nery JR, Sandoval JP, Bui B, Sejnowski TJ, Harkins TT, Mukamel EA, Behrens MM, Ecker JR. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex,. Science. 2017;357(6351):600–4. https://doi.org/10.1126/science.aan3351.
10. Mulqueen RM, Pokholok D, Norberg SJ, Torkenczy KA, Fields AJ, Sun D, Sinnamon JR, Shendure J, Trapnell C, O'Roak BJ, Xia Z, Steemers FJ, Adey AC. Highly scalable generation of DNA methylation profiles in single cells. Nat Biotechnol. 2018;36(5):428–31. https://doi.org/10.1038/nbt.4112.
11. Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, Urich MA, Nery JR, Sejnowski TJ, Lister R, Eddy SR, Ecker JR, Nathans J. Epigenomic signatures of neuronal diversity in the mammalian brain. Neuron. 2015;86(6):1369–84. https://doi.org/10.1016/j.neuron.2015.05.018.
12. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, Yu M, Tonti-Filippini J, Heyn H, Hu S, Wu JC, Rao A, Esteller M, He C, Haghighi FG, Sejnowski TJ, Behrens MM, Ecker JR. Global epigenomic reconfiguration during mammalian brain development. Science. 2013;341(6146):629. https://doi.org/10.1126/science.1237905.
13. Kapourani C-A, Sanguinetti G. Melissa: Bayesian clustering and imputation of single-cell methylomes. Genome Biol. 2019;20(61):1–15. https://doi.org/10.1186/s13059-019-1665-8.
14. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18(1):67. https://doi.org/10.1186/s13059-017-1189-z.

15. Gravina S, Dong X, Yu B, Vijg J. Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome. Genome Biol. 2016;17(150):2–8.
16. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21(1):1–35.
17. Cox DR. Some remarks on overdispersion. Biometrika. 1983;70(1):269–74. https://doi.org/10.2307/2335966.
18. Hinde J, Demétrio CGB, et al. Overdispersion: models and estimation. Comput Stat Data Anal. 1998;27(2):151–70. https://doi.org/10.1016/S0167-9473(98)00007-3.
19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
20. 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(8):69.
21. Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics. 2014;15(215):1–8.
22. Eling N, Richard AC, Richardson S, Marioni JC, Vallejos CA. Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. Cell Syst. 2018;7(3):284–94. https://doi.org/10.1016/j.cels.2018.06.011.
23. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: a review for statisticians. J Am Stat Assoc. 2017;112(518): 859–77. https://doi.org/10.1080/01621459.2017.1285773.
24. Bochkina N, Richardson S. Tail posterior probability for inference in pairwise and multiclass gene expression data. Biometrics. 2007;63(4):1117–25.
25. Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004;5(2):155–76.
26. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–44.
27. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2(1):193–218. https://doi.org/10.1007/BF01908075.
28. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502.
29. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
30. Zahr SK, Yang G, Kazan H, Borrett MJ, Yuzwa SA, Voronova A, Kaplan DR, Miller FD. A translational repression complex in developing mammalian neural stem cells that regulates neuronal specification. Neuron. 2018;97(3): 520–37.
31. Sorensen SA, Bernard A, Menon V, Royall JJ, Glattfelder KJ, Desta T, Hirokawa K, Mortrud M, Miller JA, Zeng H, et al. Correlated gene expression and target specificity demonstrate excitatory projection neuron diversity. Cereb Cortex. 2015;25(2):433–49.
32. Georgala PA, Manuel M, Price DJ. The generation of superficial cortical layers is regulated by levels of the transcription factor Pax6. Cereb Cortex. 2011;21(1):81–94.
33. Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC, Stegle O, Reik W. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells,. Nat Commun. 2018;9(1):1–9. https://doi.org/10.1038/s41467-018-03149-4.
34. Argelaguet R, Clark SJ, Mohammed H, Stapel LC, Krueger C, Kapourani CA, Imaz-Rosshandler I, Lohoff T, Xiang Y, Hanna CW, Smallwood S, Ibarra-Soria X, Buettner F, Sanguinetti G, Xie W, Krueger F, Göttgens B, Rugg-Gunn PJ, Kelsey G, Dean W, Nichols J, Stegle O, Marioni JC, Reik W. Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature. 2019;576(7787):487–91. https://doi.org/10.1038/s41586-019-1825-8.
35. Anastasiadi D, Esteve-Codina A, Piferrer F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics Chromatin. 2018;11(1):37.
36. Eling N, Morgan MD, Marioni JC. Challenges in measuring and understanding biological noise. Nat Rev Genet. 2019;20(9):536–48.
37. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions,. Genome Biol. 2012;13(10):83. https://doi.org/10.1186/gb-2012-13-10-r83.
38. de Souza CPE, Andronescu M, Masud T, Kabeer F, Biele J, Laks E, Lai D, Ye P, Brimhall J, Wang B, et al. Epiclomal: probabilistic clustering of sparse single-cell DNA methylation data. PLOS Comput Biol. 2020;16(9):1008270.
39. Uzun Y, Wu H, Tan K. Predictive modeling of single-cell DNA methylome data enhances integration with transcriptome data. Genome Res. 2021;31(1):101–9.
40. Kapourani CA, Sanguinetti G. Higher order methylation features for clustering and prediction in epigenomic studies. Bioinformatics. 2016;32(17):405–12. https://doi.org/10.1093/bioinformatics/btw432.
41. Liang F, Liu C, Carroll R. Advanced Markov Chain Monte Carlo methods: learning from past samples: Wiley; 2010. https://doi.org/10.1002/9780470669723.
42. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. Stan: a probabilistic programming language. J Stat Softw. 2017;76(1). https://doi.org/10.18637/jss.v076.i01.
43. Lewin A, Richardson S, Marshall C, Glazier A, Aitman T. Bayesian modeling of differential gene expression. Biometrics. 2006;62(1):10–8.
44. Yee TW. Vector generalized linear and additive models: with an implementation in R: Springer; 2015. https://doi.org/10.1007/978-1-4939-2818-7.
45. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, Lord RV, Clark SJ, Molloy PL. De novo identification of differentially methylated regions in the human genome. Epigenetics Chromatin. 2015;8(1):1–16.
46. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles,. Genome Biol. 2012;13(10):87. https://doi.org/10.1186/gb-2012-13-10-r87.
47. Müller F, Scherer M, Assenov Y, Lutsik P, Walter J, Lengauer T, Bock C. RnBeads 2.0: comprehensive analysis of DNA methylation data. Genome Biol. 2019;20(1):1–12.

48. Chen Y, Pal B, Visvader JE, Smyth GK. Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR. F1000Research. 2017;6. https://doi.org/10.12688/f1000research.13196.2. Accessed 28 Mar 2021.
49. Phipson B, Oshlack A. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. Genome Biol. 2014;15(9):465.
50. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood SA, Ponting CP, Voet T, Kelsey G, Stegle O, Reik W. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity,. Nat Methods. 2016;13(3):229–32. https://doi.org/10.1038/nmeth.3728.
51. Bardenet R, Doucet A, Holmes CC. On Markov chain Monte Carlo methods for tall data. J Mach Learn Res. 2017;18(47):. http://arxiv.org/abs/1505.02827.
52. Benton ML, Talipineni SC, Kostka D, Capra JA. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. BMC Genomics. 2019;20(1):1–22.
53. Hui T, Cao Q, Wegrzyn-Woltosz J, O'Neill K, Hammond CA, Knapp DJHF, Laks E, Moksa M, Aparicio S, Eaves CJ, et al. High-resolution single-cell DNA methylation measurements reveal epigenetically distinct hematopoietic stem cell subpopulations. Stem Cell Rep. 2018;11(2):578–92.
54. Jühling F, Kretzmer H, Bernhart SH, Otto C, Stadler PF, Hoffmann S. Metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Res. 2016;26(2):256–62.
55. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Nuhn M, Parker A, Patricio M, Pignatelli M, Rahtz M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Birney E, Harrow J, Muffato M, Perry E, Ruffier M, Spudich G, Trevanion SJ, Cunningham F, Aken BL, Zerbino DR, Flicek P. Ensembl 2016. Nucleic Acids Res. 2016;44(D1):710–6.
56. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):1–17.
57. Kapourani C-A, Argelaguet R, Sanguinetti G, Vallejos CA. scMET: Bayesian modelling of DNA methylation heterogeneity at single-cell resolution. Github Repository. 2021. https://doi.org/10.5281/zenodo.4629327. https://github.com/andreaskapou/scMET.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.