Genomics Proteomics Bioinformatics 21 (2023) 48–66

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

REVIEW

Computational Methods for Single-cell DNA Methylome Analysis

Waleed Iqbal [1], **Wanding Zhou** [1,2,*]

[1] *Center for Computational and Genomic Medicine, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA*
[2] *Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*

Received 31 December 2021; revised 28 April 2022; accepted 10 May 2022
Available online 17 June 2022

Handled by Ting Wang

KEYWORDS

Single-cell genomics;
Bioinformatics;
DNA methylation;
Computational tool;
Epigenetics

Abstract Dissecting intercellular epigenetic differences is key to understanding tissue heterogeneity. Recent advances in single-cell DNA methylome profiling have presented opportunities to resolve this heterogeneity at the maximum resolution. While these advances enable us to explore frontiers of chromatin biology and better understand cell lineage relationships, they pose new challenges in data processing and interpretation. This review surveys the current state of **computational tools** developed for single-cell DNA methylome data analysis. We discuss critical components of single-cell DNA methylome data analysis, including data preprocessing, quality control, imputation, dimensionality reduction, cell clustering, supervised cell annotation, cell lineage reconstruction, gene activity scoring, and integration with transcriptome data. We also highlight unique aspects of single-cell DNA methylome data analysis and discuss how techniques common to other single-cell omics data analyses can be adapted to analyze DNA methylomes. Finally, we discuss existing challenges and opportunities for future development.

Introduction

DNA methylation typically refers to the methylation of the 5-carbon of the cytosine base. It is one of the most classic epigenetic modifications in higher-order eukaryotes [1–4]. DNA methylation consolidates epigenetic states over cell replication and is extensively implicated in many biological activities: transcriptional regulation [5–7], genomic imprinting [8–11], X-chromosome inactivation [12,13], transposable element suppression [14–17], cell differentiation [18–20], and tissue and organismal development [21–25]. It has also been extensively studied and applied to track aging [26,27] and various forms of human diseases [28–34].

Many non-recurrent cellular alterations with phenotypic manifestation are inheritable over mitosis and coded into the epigenome. As a classic epigenetic mark, the DNA methylome encodes rich information that delineates the state of a cell, including its developmental lineage, cell cycle stage, transcriptional activity, mitotic age, and proliferation potential. In this review, the term "methylome" will refer to the genome-wide distribution of 5-methylcytosine in DNA. Other forms of DNA and protein methylation (*e.g.*, histone methylation) will

* Corresponding author.
E-mail: wanding.zhou@pennmedicine.upenn.edu (Zhou W).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.05.007

be specified explicitly. The most effective strategy for capturing this cell-to-cell diversity is to profile the methylomes of single cells. Like other single-cell omics assays, the single-cell DNA methylome assay has been actively developed with a rapid surge in data volume and variety over the past decade. **Table 1** highlights the limitation of bulk methylome analyses compared to their single-cell counterparts.

Compared to single-cell transcriptomic data, which carry information on the transcriptional state, DNA methylome data carry information on mechanisms of gene expression regulation, such as the involvement of specific *cis*-regulatory elements and their interactions. Although this genome-wide coverage enables detecting gene regulatory differences overlooked in single-cell transcriptomics, single-cell methylome data contain signals from hundreds of thousands to millions of CpGs in the genome, representing much higher dimensional data. In other words, the data sparsity challenge, common to all single-cell data, is particularly prominent for single-cell DNA methylome data due to the limitation of DNA (compared to RNA) material per locus per diploid cell. In addition, sodium bisulfite conversion [35,36], a technique commonly used in methylome profiling, causes DNA damage and further contributes to DNA loss and data sparsity.

DNA methylation data have an uneven genomic representation with a complex spatial correlation pattern. CpG dinucleotides, where most DNA methylation readings are collected, are unevenly distributed in the genome (non-CpG or CpH methylation is only found in some tissue and cell types such as neurons and embryonic stem cells). Different genomic features have different CpG densities. Late-replicating lamina-attached DNA is CpG-sparse [37], while CpG island DNA retains higher CpG density [38]. The CpG island lengths at different gene promoters are on a continuous spectrum. DNA methylation of neighboring CpGs is intrinsically correlated with different genomic scales depending on the underlying cellular and biochemical processes. The determination of DNA methylation at *cis*-regulatory elements [18] and promoters [39] is typically focal, and loss of methylation is often indicative of transcriptional machinery binding. However, some transcription factors (TFs) preferentially bind methylated sites [40,41]. In contrast, loss of methylation at partially methylated domains (PMDs) due to extensive cell division occurs on mega-base scales [42,43]. This complex grammar of DNA methylation determination mandates consideration of CpG distribution and its correlative structure at multiple genomic scales during bulk-tissue and single-cell DNA methylome analyses [44].

Computational and statistical techniques are emerging for single-cell methylome analysis, though they are not catching up with the development of informatics for single-cell RNA sequencing (scRNA-seq) and single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq). This lag is partly due to the challenges and complexities mentioned above. This review surveys the current computational methods developed for single-cell DNA methylome analysis (**Table 2**). As some methods share principles common to the analyses of bulk-tissue DNA methylome data and other single-cell omics data, we will focus on the unique challenges of single-cell DNA methylome analysis.

We first briefly discuss DNA methylome assay technologies and data preprocessing workflows. Then we focus on common methylome analysis tasks: quality control, dimensionality reduction, cell clustering, differential methylation analysis, cell lineage analysis, motif analysis, analysis of cell-to-cell methylation concordance, and analysis of non-CpG methylations. Although single-cell DNA methylome analysis is briefly compared with data of other modalities, we refer readers to other reviews for detailed discussions of the corresponding data types: scRNA-seq [45–48], scATAC-seq [49,50], and single-cell HiC (scHiC) [51,52]. Non-traditional uses of the DNA methylation sequencing data, such as identifying genetic mutations and copy number alterations, are also discussed. Finally, we discuss how to integrate single-cell DNA methylome data with data from other omics and how to analyze co-assay data, in which DNA methylation and other omics data were collected from the same cells.

Table 1 Comparison of typical analyses performed on single-cell and bulk methylome sequencing data

Analysis	Bulk tissue	Single cell
Tissue- or cell-specific epigenome	Lineage relationship of the major constituent cell states of tissues	Differentiation trajectory of single cells and cell states
Cell composition analysis	Top-down cell composition analysis of convoluted tissue signals	Bottom-up cell composition analysis focusing on rare cell types
DNA copy number annotation	Weighted average DNA copy number from the whole cell population	DNA copy numbers of each cell and its heterogeneity
DNA copy number on sex chromosomes	Sex inference and sex chromosome abnormality of the whole individual	Sex chromosome epigenetic mosaicisms across cell types
Primary tumor epigenetic alteration	Focus on tumor *vs.* normal epigenetic differences such as global hypomethylation	Distinguish cell-autonomous epigenetic changes in individual tumor cells
Tumor cell evolution history	Compare the epigenome of tumor at multiple sites (*e.g.*, primary *vs.* metastatic)	Resolve clonal evolution history of tumor cells
Cell cycle	Compute the fraction of cells at different cell cycle stages	Determine cell cycle stage of individual cells
Epigenome–transcriptome association	Epigenome–transcriptome associations influenced by cell type variation	Epigenome–transcriptome associations across cells of the same cell type

Table 2 Overview of analytical features of single-cell methylome analysis tools

General information			Input data			Data processing					Cell state annotation			Modeling		Multi-omics			Data presentation	
Name	Programming language	Benchmarked dataset	Epigenome	scRNA-seq integration	Data format	Filtering	Normalization	Imputation	Feature matrix	Feature engineering	Motif analysis	Cell typing	Cell trajectory	Clustering	Differential methylation	Label transfer	Gene activity scoring	Cell space	Visualization	Ref.
BPRmeth	R	scNMT	M		MC	F + V		+	RC	SVLR				+	+		+		+	[134]
DeepCpG	P	scBS-seq scRRBS	M		MC			+	RC	CNN	+			+	+				+	[139]
MELISSA	R	scM&T-seq scBS-seq	M		MC		+	+	RC	BI				+					+	[143]
Epiclomal	P + R	scBS-seq scRRBS scWGBS	M		MC	F + V	+	+	RC	BI				+						[144]
MAPLE	R	snmC-seq scM&T-seq scNMT-seq	M	+	AB					ENS						S	+	DS*		[211]
MethylStar	P + R + Sh	scBS-seq	M		FQ			MI												[141]
scMET	R	snmC-seq scNMT-seq	M		MC			+	RC	BI					+					[189]
EpiScanpy	P	snmC-seq	M/A		MC	F + V + D	+	+	RC	NG			+	+	+		A		+	[138]
coupleCoC +	ML	snmC-seq	M/A	+	MG			+	GC	ITC				+			A	DS	+	[240]
ALLCools	P	snmC-seq2	M/A	+	PY	F + V + D	+		RC	+	+	+		+	+			DS	+	[66]
MATCHER	P	scM&T-seq scGEM	M/A	+	MB								+	+	+	+		DS		[218]
LIGER	R	snmC-seq	M/A	+	MG				FC	MF		+		+	+	+		DS	+	[210]
scAI	R	scM&T-seq	M/A	+	MB	F	+		FC	MF				+	+			CA	+	[228]
MOFA +	R	scM&T-seq	M/A	+	MO		+		FC	MF				+	+			CA	+	[227]

Note: The epigenetic input of these tools is depicted by M if they are designed solely for single-cell methylation-seq or M/A if the input can be either single-cell methylation-seq or single-cell ATAC-seq. '+' indicates that the feature is supported or the functionality is through another specified software. For convenience, we have included the methylation datasets used by the original papers of these tools, for testing and training purposes; single-cell datasets not containing methylation have been omitted (scRNA-seq, scATAC-seq, or both). scRNA-seq, single-cell RNA sequencing; scATAC-seq, single-cell assay for transposase-accessible chromatin sequencing; P, Python; Sh, Shell/BASH; ML, MATLAB; M, DNA methylation; A, chromatin accessibility; MC, methylation call; AB, aligned read in BAM; FQ, FASTQ raw read; MG, methylation call by gene; PY, processed read in the YAPS MCDS format; MB, binarized methylation call; MO, MultiAssayExperiment Object; F, filtering CpGs and cells by sequencing depth or data sparsity; V, filtering CpGs by methylation variation; D, doublet detection and filtering; MI, conducted through METHImpute; RC, genomic region-by-cell matrix; GC, gene-by-cell matrix; FC, factor-by-cell matrix (factor includes shared factors and non-shared factors); SVLR, support vector linear regression; CNN, convolutional neural network; BI, Bayesian inference; ENS, ensemble machine learning (CNN + elastic net + random forest); NG, neighbor graph and graph-based clustering (Louvain, Leiden, *etc.*); ITC, information-theoretic co-clustering; MF, matrix factorization; S, conducted through Seurat; CA, integration of co-assays; DS, integration of data from different sample spaces; DS*, integration of data from different sample spaces but using co-assay data for regressor training.

Single-cell DNA methylome assay technologies

Mainstream single-cell DNA methylome assays chemically convert cytosines to other bases depending on the cytosine's methylation state [53]. This conversion is typically mediated by sodium bisulfite treatment [35,36]. However, it has recently been achieved through enzymes [54] and other chemicals [55–57]. Cytosine conversion takes place on single-strand DNA and breaks the strand complementarity. Conventional library preparation procedures first attach Y-shape adapters to the double-strand DNA before bisulfite conversion. For example, the single-cell reduced representation bisulfite sequencing (scRRBS) method [58] follows this paradigm. However, bisulfite conversion damages the adapter-ligated library and causes significant DNA loss. Therefore, modern single-cell bisulfite sequencing methods attach either the second or both adapters after bisulfite conversion. An example is the post-bisulfite adapter tagging (PBAT) method [59] (**Figure 1**). Current single-cell methylome sequencing technologies can be classified by whether they cover the whole methylome, *e.g.*, single-cell whole-genome bisulfite sequencing (scWGBS) [60–66], or specific subsets, *e.g.*, scRRBS [58,67,68] and its variants [69,70], which primarily target CpG-dense genomic regions. Differences among these methodologies are reflected in their choice of genome fragmentation method. Conventionally, fragmented library DNA was generated using DNA sonication, mechanical shearing, and random priming-based pre-amplification. Alternatively, restriction enzyme digestion [58,69] and transposase-mediated "tagmentation" [71] allow for multiplex barcoding [67] and combinatorial barcoding [64]. These methods have higher cell number throughput and can profile hundreds to thousands of cells in one experiment. Cytosine conversion-based DNA methylome profiling methods are compatible with co-assays of other omics data that may be profiled through affinity binding, genome fragmentation, and other base conversions. Such technologies include: single-cell triple-omics sequencing (scTrio-seq) [72,73], single-cell methylation and transcriptome sequencing (scM&T-seq) [74], single-cell nucleosome occupancy and methylome sequencing (scNOMe-seq) [75], single-cell chromatin overall omic-scale landscape sequencing (scCOOL-seq) [76], single-cell nucleosome, methylation and transcription sequencing (scNMT-seq) [77], 'switching mechanism at the end of the 5′-end of the RNA transcript' sequencing combined with RRBS (SmartRRBS) [68], methyltransferase treatment followed by single-molecule long-read sequencing (MeSMLR-seq) [78], single-cell methyl-HiC sequencing (scMethylHiC) [79], and single-nucleus methyl-3C sequencing (sn-m3C-seq) [80].

Other conversion-free methylation-specific restriction enzyme (MSRE)-based methods, *e.g.*, single-cell genome-wide CpG island (CGI) methylation sequencing (scCGI-seq) [81] and epigenomics and genomics of single cells analyzed by restriction (epi-gSCAR) [82], have also been developed. However, these methods do not profile the whole methylome at the base resolution. Recently, long-range sequencing [83] and imaging-based technologies have yielded equivalent single-molecule data without single-cell isolation/barcoding [84–87]. These methods are promising for studying locus-specific genomic regions in single cells [54] or potentially small genomes when single reads can cover the whole chromosome or genome. We refer the readers to previous reviews to compare these single-cell methylome assays [53,88,89].

Single-cell DNA methylome data preprocessing

Read processing and library quality control

Traditional read trimmers (*e.g.*, TrimGalore!) and mappers (*e.g.*, Bismark [90], BSmap [91], and BSseeker [92–94]) were first designed for bulk-tissue analysis but can be applied to

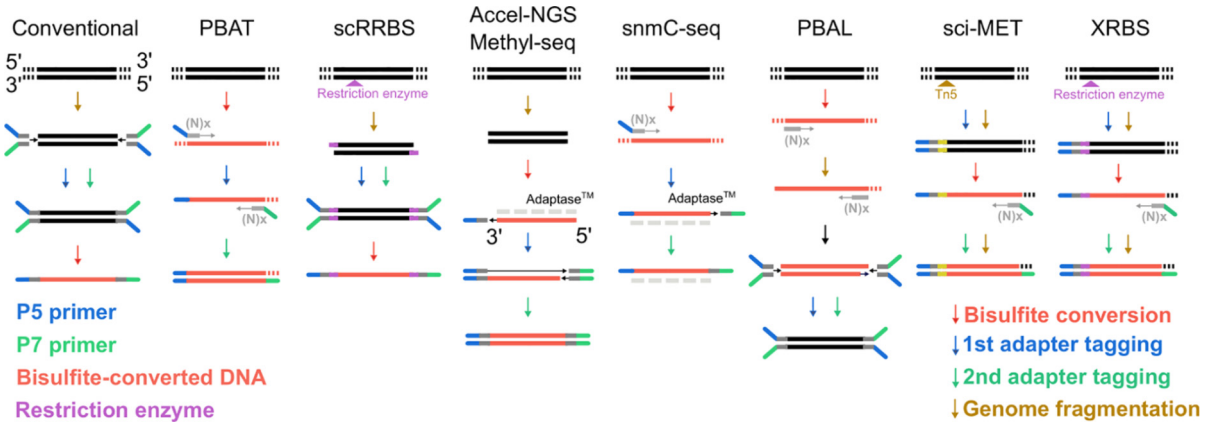

Figure 1 Library preparation strategies of single-cell bisulfite sequencing assays
These assays can be differentiated based on having either bisulfite conversion (red arrow) or genome fragmentation (golden arrow) as the first step, followed by subsequent differences in adapter tagging. Blue arrows represent the first adapter tagging, and green arrows represent the second adapter tagging. Gray dashed lines denote the missing complementary strand for single-stranded DNA. Additionally, the bisulfite-converted DNA is shown in red, the insertion from the restriction enzyme is shown in purple, and the P5 and P7 primers are shown in blue and green, respectively. PBAT, post-bisulfite adaptor tagging; scRRBS, single-cell reduced representation bisulfite sequencing; Accel-NGS Methyl-seq, accel next-generation methylation sequencing; snmC-seq, single nucleus methylome sequencing; PBAL, post-bisulfite adapter ligation; sci-MET, single-cell combinatorial indexing for methylation analysis; XRBS, extended representation bisulfite sequencing.

single-cell data analysis with some necessary adaptations. Single-cell bisulfite sequencing library preparation often has one or multiple rounds of random priming-based pre-amplification [65]. The first sequencing adapter may tag both the bisulfite-converted strand and the daughter strand; mappers need to be aware of this change so that reads can be mapped to all four-strand types [60]. Lower mapping rates, higher percentage of adapter dimers, and smaller insert sizes are common issues seen in single-cell methylome data (**Figure 2**). Mappers that do not support sub-read mapping require proper trimming of leading and trailing subsequences that may be artificial due to overhang end-repair, incomplete conversion at the 5′ end, and low sequencing quality at the 3′ end [95]. To enhance mapping efficiency, mate reads in a pair may be mapped separately as single-end reads, with or without first mapping them in pairs, to accommodate potential microhomology-mediated chimerism [60,76,96]. The scBS-map package increases the mapping rate of such chimeric reads through local alignment [97]. Additionally, certain regions are recommended to be excluded in DNA methylation quantification due to low mappability in the bisulfite-converted genome [98].

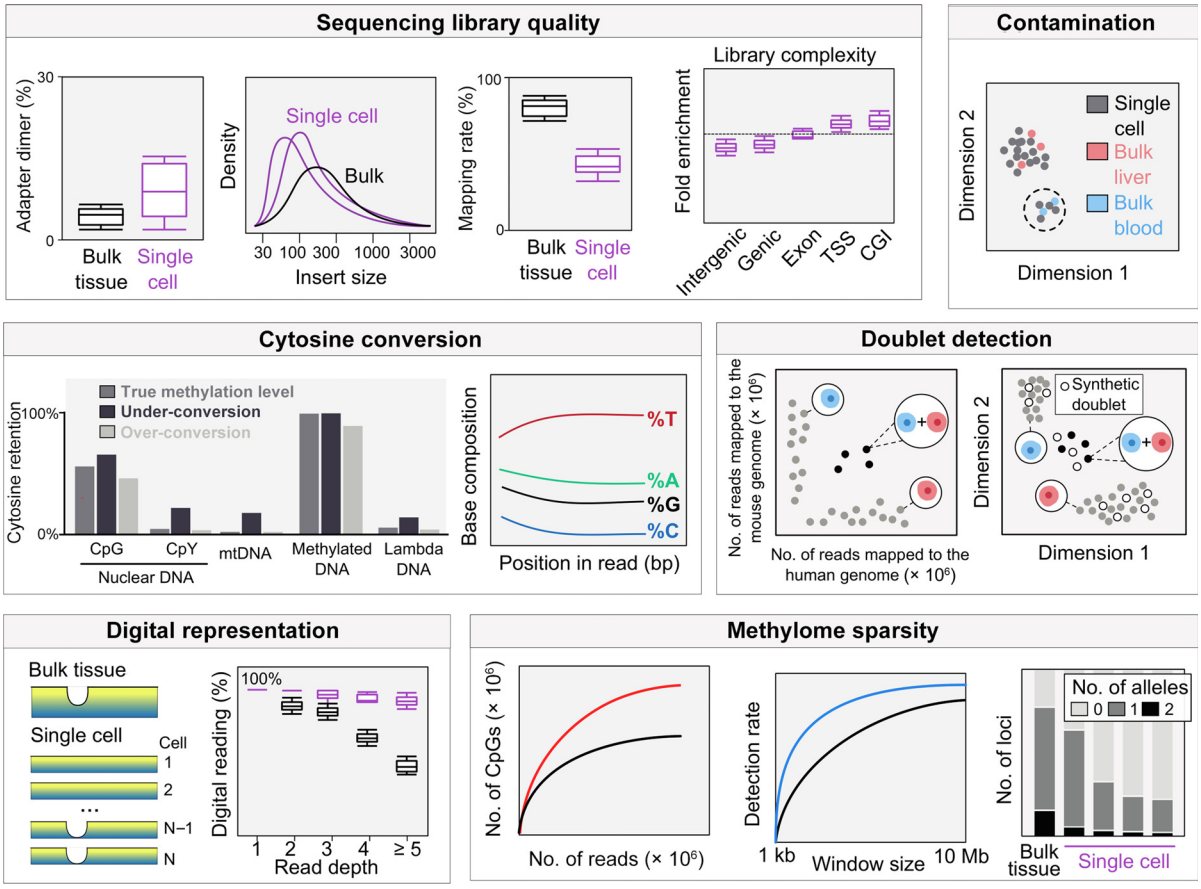

Figure 2 Schematic representation of single-cell methylome data quality control

Quality control consists of sequence library assessment, sample contamination detection, doublet detection, cytosine conversion control, and filtering cells and samples based on data sparsity. Sequencing library quality: sequencing library quality can be assessed by comparing metrics such as adapter dimer (%), adapter insert size, and mapping rate (%) between bulk tissues (black) and single cells (purple); library complexity can be analyzed based on fold enrichment of different genomic regions. Contamination control: single-cell samples can be compared with their bulk counterparts to check for potential contamination. Cytosine conversion: (1) cytosine context retention [C followed by G (CpG), C followed by C or T (CpY), mitochondrial DNA (mtDNA), methylated DNA, and spike-in lambda phage DNA] and examples of expected percentages according to true methylation levels (dark gray), under-conversion (black), and over-conversion (light gray) are shown; (2) base composition percentage according to read position after bisulfite treatment [T will be present at the highest percentage (unmethylated Cs are converted to T due to bisulfite treatment, increasing total T percentage after conversion), and C will be present at the lowest percentage (only methylated Cs are retained)]. Doublet detection: (1) cross-species mixing experiment can be used to identify and determine the doublet rates in single-cell assays (*e.g.*, in a human-mouse mixing experiment, samples with high mapping rates to both human and mouse genomes are doublets); (2) clustering actual samples with synthetic doublets can also be used to filter doublets (samples that cluster close to known synthetic doublets could be marked and removed). Digital representation: (1) comparing overall methylation patterns in bulk tissues and single cells; (2) comparing read depth and digital reading (0 or 1) percentage (%) in bulk tissues (black) and single cells (purple). Methylome sparsity: (1) comparing the number of CpGs covered as more reads are sequenced (red and black lines represent schematic examples of two different experiments or sequencing methodologies); (2) comparing the detection rate as the smoothing window size increases (blue and black lines represent schematic examples of two different experiments or sequencing methodologies); (3) single-cell data may have fewer allele-specific methylations as compared to bulk data.

Contamination control

Quality control is more relevant for single-cell methylome analysis than its bulk-tissue counterpart (Figure 2). Single-cell methylome data contain common technical confounders such as sample contamination, cytosine conversion artifacts, and the presence of doublets. To check for contamination and sample mislabeling, one can confirm donor sample identity by extracting single nucleotide polymorphism (SNP) information [99] and copy number alterations [100] from bisulfite sequencing data; cells from the same donor should have almost identical genotypes. We recommend mapping sequencing reads against potential contaminants and co-clustering single-cell data with existing bulk methylome data so that one can discriminate contaminating cells [100] (Figure 2) and evaluate the signal over noise ratio. For example, single-cell methylome data are supposed to form clusters with bulk data of similar tissue types rather than with different tissues assayed with the same technology.

Conversion control

Proper cytosine conversion can be challenging for DNA methylome assays since it is subject to the influence of bisulfite treatment duration [101,102], incubation temperature [103], and choice of the polymerase for subsequent amplification [104,105]. One could use internal or external controls to gauge proper cytosine conversion. Intrinsic controls include cytosines with assumed methylation states. Cytosines in the mitochondrial genome [65], at CpC and CpT sites [106], and at CpCpC sites [63,66] may be used as internal controls for incomplete conversion because they often lack biological DNA methylation [107]. Although universally methylated cytosines are more difficult to find, a similar strategy can detect over-conversion. Promising targets may include CpGs at certain transposable elements whose methylation is critically maintained for cell viability [108]. Since the assumed methylation states of internal controls are subject to rare exceptions [109], one can also use external controls such as spiked-in unmethylated lambda DNA [100] or amplified and unmethylated DNA for detecting incomplete conversion, as well as M.SssI-treated fully methylated DNA for detecting over-conversion (Figure 2). Based on these principles, we recommend excluding incompletely converted reads (*e.g.*, by removing reads with three or more consecutive non-converted CpH cytosines [110]) or whole-cell samples (*e.g.*, requiring mCpCpC level < 0.03, mCpG level > 0.5, and mCpH level < 0.2 [66]).

Doublet detection

Due to the technical challenge of single-cell isolation [111,112], the inclusion of more than one cell in a single-cell experiment can occur. The exclusion of doublets and samples with more cells is critical to single-cell methylation and ploidy analyses (*e.g.*, one based on lambda DNA spike-ins [76]). To our knowledge, no dedicated method has been developed for doublet detection for single-cell methylome data. Several strategies have been proposed and used in practice. For example, doublets typically feature an unusually high read number [64,66] after PCR duplicate marking [90,113]. Doublets can also have a higher number of CpGs with non-0-1 methylation levels. Most human and murine cells are diploid. Polyploidy may arise in cell types such as hepatocytes [114], cardiomyocytes [115], megakaryocytes [116], cells in the S and G2 phases, and cancer cells with somatic copy number changes [117]. A popular generic strategy developed originally for scRNA-seq (*e.g.*, DoubletFinder [118] and Scrublet [119]) and lately for scATAC-seq (*e.g.*, ArchR [120] and SnapATAC [121]) is to compare with synthetic doublets created by mixing signals from different cells. Cells that demonstrate similarity to these synthetic doublets are labeled for removal. The doublet rate of a single-cell sequencing method can be estimated using a cross-species pooling experiment where cells from different species (*e.g.*, human and mouse) are pooled before being sequenced. The single-cell data are then mapped to the two genomes separately [100]. High read mapping rates in both species indicate the presence of more than one cells [60,122] (Figure 2).

Discretization of DNA methylation fractions

Since DNA methylation fractions can only be 0%, 50%, and 100% in diploid cells (Figure 2), Hui et al. considered only CpGs with 0 and 1 in methylation fraction and ignored the other CpGs [62]. One can also discretize the observed cytosine retention rate, *e.g.*, taking fractions between 0 and 0.1 to 0, between 0.9 and 1 to 1, and everything else to 0.5 [65,123]. Smallwood et al. [65] and Argelaguet et al. [96] modeled each read as a Bernoulli random variable and inferred the methylation rate.

Handling data sparsity

Data sparsity is a defining challenge common to all single-cell data [124–126]. This challenge is exacerbated in single-cell DNA methylomes due to the limited DNA material per cell [88]. For scRNA-seq, scATAC-seq, and scHiC data, biological and nonbiological zeros are hard to discriminate [127,128]. Single-cell methylome signal is based on chemical conversion and is uniquely exempt from having this problem. For single-cell methylome data, missing data are explicitly reflected on the read depth, attributable to the dropout of one or both alleles. These two scenarios can be hard to discriminate when one allele is lost and the other gets amplified in the experiment. The dropout of one allele could lead to misinterpretation of the methylation signal at mono-allelically methylated sites. Some downstream analyses may require a complete data matrix without missing values. Here we discuss three common strategies to address the data sparsity challenge (**Figure 3**).

Cell and feature filtering

Filtering cells with suboptimal genome coverage or mapped read counts can effectively reduce data sparsity (Figure 3). However, single-cell methylome often has distinct technical characteristics and targets different biology; therefore, different thresholds need to be applied for filtering low-quality data. For example, Hernando-Herraez et al. discarded cells with less than 1,000,000 reads or 500,000 CpGs covered and downsampled each cell to 1,000,000 reads per cell to avoid sequencing depth bias [129]. Liu et al. filtered cells with less than 500,000 reads [66]. Gaiti et al. filtered cells with less than 50,000 CpGs covered using scRRBS [68]. In addition, one can also exclude genomic regions sparsely represented in the data and regions flagged for other quality issues [130].

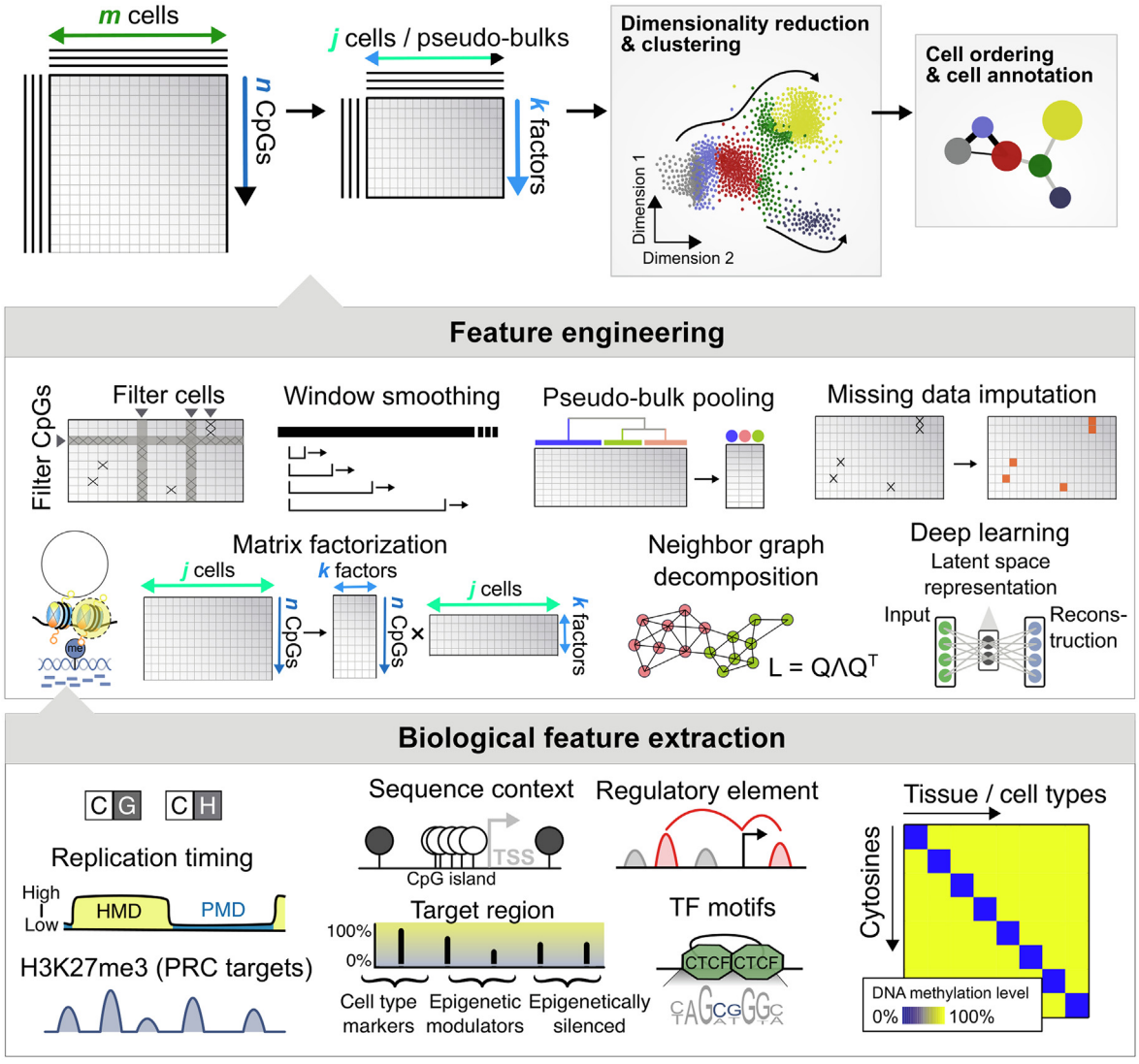

Figure 3 Schematic representation of data-driven and biological feature selections

After quality control and data preprocessing, the major components of most bioinformatics tools are feature engineering (middle panel) and biological feature extraction (bottom panel), which are performed before downstream analyses such as cell clustering, annotation, and trajectory analysis. Feature engineering: (1) filtering CpGs and samples (based on read depth, data sparsity, and variance); (2) window smoothing (based on defined window sizes and step sizes); (3) pseudo-bulk creation; (4) imputation of missing values; (5) matrix factorization; (6) neighbor graph decomposition; and (7) deep learning latent space representation. Biological feature extraction: (1) methylation extraction from CpG and CpH (where H represents A, C, or T), respectively; (2) identifying replication timings based on CpG methylations in HMDs and PMDs; (3) DNA methylation on PRC targets with histone H3K27me3 marks; (4) sequence context-specific methylations [an example describing unmethylated CpG islands (black and white circles denote methylated and unmethylated cytosines, respectively) at specific gene promoters]; (5) methylations in regulatory element sites and TF binding motifs (an example showing CCCTC-binding factor, CTCF, and binding); and (6) cell type- and tissue-specific methylation signatures. HMD, highly methylated domain; PMD, partially methylated domain; PRC, polycomb repressive complex; TSS, transcription start site; TF, transcription factor.

Cell and feature aggregation

To further enhance the signals from sparse data, one could aggregate signals from multiple CpGs to reduce data dimensionality and alleviate data sparsity without explicitly recovering the missing data for each CpG. Genome tiling or window smoothing is a simple but effective method (Figure 3). Methylation fraction averaging can be done with sliding windows of different lengths and step sizes. However, the choice of the window size is often constrained by cost and experimental limitations. Windows range from 1 kb-size non-overlapping windows [100], 3 kb-size 600 bp-stepping windows [61], to 100 kb-size non-overlapping windows [63,66]. Choice of window size affects the quantification of cell-to-cell similarity. For example, large window size of 100 kb may dilute focal lineage-specific enhancer signals and diminish cell lineage discrimination. Alternatively, one can perform analyses based on shared genomic and epigenomic features where methylation readings are aggregated and cells are compared on genome-wide aggregates [74,100]. Most targeted genomic features are gene-centric. They include gene promoters, gene body with flanking regions

[66], proximal and distal regulatory elements (from, *e.g.*, FANTOM5 [131], the Ensembl regulatory build [132], and ENCODE [133]), and accessible chromatin identified from bulk and scATAC-seq [96]. Genomic features can also be extracted from data using unsupervised methods such as by maximizing methylation variation across cell samples [96], Variational Bayes [134], or models based on the curated annotation of epigenetic markers [63].

Besides feature aggregation, one could aggregate signals from similar cells and produce a single pseudo-bulk methylome to represent a cell group (Figure 3). The underlying rationale of this approach is that when cells from the same group are sufficiently similar, it is safe enough to ignore the residual heterogeneity in exchange for higher genome coverage. The subsequent analysis focuses on analyzing the pseudo-bulk profiles rather than individual single cells [63,64,66,72]. One may also build different pseudo-bulk samples to mimic sample replicates for analysis that requires replicates. This pseudo-bulk pooling can be done by sampling cells with or without overlaps [120]. Accurate clustering of cells with similar epigenomes is key to data aggregation with minimum information loss.

Data imputation

Several data imputation methods have also been proposed to handle missing bulk and single-cell DNA methylome data (Figure 3). Zhang et al. proposed a random forest method to impute missing DNA methylation data from bulk methylome data using neighboring CpG methylation levels and epigenomic annotations [135]. BoostMe used XGBoost, a gradient boosting algorithm, for imputing low-quality WGBS data [136]. Farlik et al. imputed missing CpG methylations using the impute R package with 5-nearest neighbor averaging [137]. Luo et al. [63] and EpiScanpy [138] imputed CpH methylations by replacing the missing bin values with the average methylation across all cells for those bins. DeepCpG used a convolution neural network (CNN) framework to impute DNA methylation levels in single cells [139]. MethylStar is a pipeline that wraps in METHimpute, a hidden Markov model-based method [140], for single-cell methylome analysis [141]. Hui et al. used Bsmooth [142] to estimate methylation levels at all CpG sites in the genome [62]. Melissa [143] and Epiclomal [144] used Bayesian mixture models that jointly cluster sparse single-cell methylomes and impute missing values. LightCpG [145] and CaMelia [146] both used gradient boosting to train classifiers with reduced training time and increased accuracy, leveraging positional, structural information, and other bulk methylome data. CpG Transformer adapted the transformer neural network architecture to allow parallel imputation and better scaling on large datasets [147].

Non-CpG methylation

Although cytosine methylations predominantly occur at CpG dinucleotides, they can also be found at non-CpG (referred to as CpH or CH) sites in specific tissues and cell types [109] (Figure 3). Non-CpG methylations primarily occur at CpA sites [106] in embryonic stem cells (ESCs) [37], neurons [63], and oocytes [148]. While gene activity is associated with CpG methylation at gene promoters and enhancers, non-CpG methylation over gene bodies was found to be more predictive of gene expression in neurons [149]. It can negatively correlate with transcriptional activity in single-cell methylomes [63,66]. Non-CpG cytosines can be classified into CHG and CHH cytosines [148], whose methylation is usually highly correlated in mammalian cells [106].

Data techniques customized for single-cell methylome analysis

Manifold learning and dimensionality reduction

DNA methylation data have a higher dimension than gene expression data ($\sim$ 28,000,000 CpGs *vs.* $\sim$ 20,000 genes for human genome). To understand cell-to-cell similarities and facilitate data visualization, one can perform dimensionality reduction techniques that project cells from the original data space into a lower-dimensional (*e.g.*, 2D) feature space (Figure 3). One can also partition the whole-cell population into similar sub-populations using clustering analysis. These two analyses are closely related, and both rely on learning the cell-to-cell distances (or similarities) in the original data manifold.

Besides direct feature aggregation, most dimensionality reduction techniques are based on matrix factorization, neighbor graph decomposition, and generative model inference. Matrix factorization methods can be classified into linear methods such as principal component analysis (PCA) [150], non-negative matrix factorization (NMF) [151], and nonlinear methods such as multi-dimensional scaling (MDS) [152]. Notable examples of neighbor graph-based methods are t-distributed stochastic neighbor embedding (tSNE) [153] and uniform manifold approximation and projection (UMAP) [154]. Generative model-based methods include Bayesian nonparametric models [155] and deep generative models. These methods are often used in combinations to avoid the dominance of data sparsity in downstream analyses and reduce computational load. For example, Farlik et al. first aggregated methylations using 1 kb tiles and then applied MDS to project each cell's methylome to 2D [100]. The same team also applied PCA to methylation aggregation according to TF binding sites in a later study [137]. Luo et al. first aggregated methylations using 100 kb tiles before applying tSNE [63]. Hui et al. performed MDS using a global dissimilarity measure that averages the absolute methylation difference [62]. Mulqueen et al. first performed NMF to extract latent variables before using tSNE to project the data to the 2D space [64]. MOFA and MOFA+ adapted a group factor analysis framework and applied mean-field Variational Bayes inference to factor matrices with single-cell data from multiple modalities, including DNA methylomes [156]. The same authors co-opted MOFA for data visualization by restraining the number of latent variables to 2 [96]. Liu et al. first grouped cells on three hierarchical levels, *i.e.*, cell class, major type, and subtype, and applied UMAP and tSNE on each level [66].

Clustering

Cell clustering [157] often serves as a primary step in analyzing cell population structures [48]. The resulting cell clusters can be used in data imputation, pseudo-bulk methylome construction [66], trajectory inference [158], and cell type refinement (Figure 3). Clustering methods used in previous single-cell DNA methylome studies include k-mean clustering [159], k-medoid clustering [160], density-based clustering (*e.g.*,

mean-shift [161] and density-based spatial clustering of applications with noise (DBSCAN) [162]), hierarchical clustering [163], nonparametric Bayesian methods (*e.g.*, Dirichlet mixture models [164]), affinity propagation [165], and neighbor graph-based clustering (*e.g.*, community detection [166,167], spectral clustering [168], modularity [169], Louvain clustering [170], Infomap [171], and Leiden clustering [172]). Most clustering algorithms share their mathematical roots with dimensionality reduction methods. Choosing the optimal clustering method depends on the geometric shape of the expected clusters and practical computational constraints; see [48] for a detailed discussion. There are many direct adaptations of the standard clustering methods for single-cell methylome data analysis. Of these methods, hierarchical clustering is one of the most widely used [65,72,74,137] and adapted (*e.g.*, by BackSPIN [173] and PDclust [62]). Mulqueen et al. used DBSCAN after dimensionality reduction [64]. Liu et al. used the Leiden algorithm for clustering [66]. Melissa [143] and Epiclomal [144] used probabilistic mixture models to assign cells to clusters while simultaneously imputing missing data. Epiclomal employed a Bernoulli mixture model [144] but used density-based clustering and hierarchical clustering methods for model initialization. To reconcile different clustering results or accommodate randomness in the clustering algorithms, consensus clustering can be used on top of multiple clustering assignment matrices [174], which might be derived from running multiple clustering algorithms or the same stochastic clustering algorithm multiple times. Liu et al. treated each clustering assignment from multiple Leiden clustering runs as the new feature for each cell and performed DBSCAN to achieve consensus clustering [66].

Cell ordering and lineage reconstruction

To capture the continuity of cell state evolution, one often orders cells or cell groups to quantify cell state transitions such as differentiation and cell cycle changes (Figure 3). Here we will use the term of cell ordering (total or partial order). The readers may find similar analyses referred to as pseudo-time ordering, trajectory inference, lineage reconstruction, and taxonomy reconstruction in the literature. Methods are available to accommodate different trajectory topologies ranging from linear, bifurcating, multifurcating, tree-like to acyclic and cyclic graphs (see Saelens et al. [175] for a survey and benchmarking).

Strategies used in previous single-cell methylome analyses include visual inspection, neighbor graph-based methods, and model-based methods. Farlik et al. studied the role of methylation in lineage differentiation of mouse ESCs by comparing 5-azacytidine-treated and untreated cells [100]. 5-azacytidine causes a global decrease in methylation due to its inhibition of DNA methyltransferases (DNMTs). Cells are ordered in a lineage plot spanned by the positive and negative methylation differences. EpiScanpy [138] uses PAGA [158] for cell ordering. PAGA was initially developed to infer trajectory from scRNA-seq data. It first constructs a k-nearest neighbor-like graph within UMAP and then partitions the graph using the Louvain method. Finally, a new abstraction graph is constructed, treating cell groups as nodes in the new graph. The connectivity between nodes in the new graph quantifies the ratio of observed over expected number of inter-group edges in the original graph [158]. Clark et al. used the diffusion pseudo-time (DPT) method [176] to infer developmental trajectory from the gene expression component of scNMT-seq data [77]. DPT constructs a weighted neighbor graph before performing a random walk to model cell state transitions. Popular alternative approaches to infer trajectory from neighbor graphs include the use of a minimum spanning tree (*e.g.*, in Monocle [177]) and the shortest path algorithms (*e.g.*, in Wanderlust [178] and Wishbone [179]). These methods remain to be tested on single-cell DNA methylome data.

Besides neighbor graph-based methods, one can also use model-based statistical inference methods, *e.g.*, phylogenetic reconstruction techniques, to infer lineage relationships among cells. Unlike neighbor graph-based methods, phylogenetic methods assume that observed cells are at the leaves of the phylogeny (akin to extant species), and the methylation states of the internal nodes will be inferred. To account for DNA methylation heterogeneity at different genomic regions, Gaiti et al. [68] first selected a base-substitution model using a model-selection procedure that incorporates rate heterogeneity across sites [180]. The authors then applied IQ-TREE [181], a maximum likelihood-based phylogenetic tree searching algorithm, in which branch support is estimated using the ultrafast bootstrap (UFBoot) method [182] combined with a Shimodaira-Hasegawa-like approximate likelihood ratio test and an approximate Bayes test [68]. The robustness of the tree construction is further evaluated by holding out specific chromosomes or CpGs [68]. This maximum likelihood-based approach is over 3-fold more robust than maximum parsimony-based reconstruction. Furthermore, the branch lengths (the cumulative methylation change) can be further translated to the number of cell divisions using rates previously calibrated in colorectal cancer [183].

Supervised cell annotation

DNA methylome-based cell type annotation often leverages data of known cell types and markers (CpGs or genes) with known cell type associations (Figure 3). Farlik et al. applied an elastic net-regularized generalized linear model [184] with training data labels derived from data pooled from cells of known identity [137]. Model features were extracted from regulatory regions defined in the BLUEPRINT Ensembl regulatory build. Luo et al. [63] and Liu et al. [66] annotated neuronal and other brain cells using global CpH methylation and locus-specific methylation at marker genes with known methylation–expression correlation. Luo et al. first clustered cells using the BackSPIN method [173] and annotated clusters based on the depletion of CpH methylation at genes (whole gene bodies ± 2 kb) whose expression levels are known to mark neuron subtypes [63]. Liu et al. first classified cells at three hierarchical levels and then carried out cell type annotations within levels to maximize the power of subtype discrimination [66]. Aside from cell type information, one can also annotate other aspects of the cell state. For example, Guo et al. annotated cell ploidy using controlled lambda DNA spike-in [76]. Hernando-Herraez et al. annotated the epigenetic age of single-cell methylomes using a linear model fitted on bulk-tissue data [129]. Recently Trapp et al. introduced a percentile-based approach to address coverage discrepancy from single cells to track the cellular aging process using

single-cell methylomes [185]. Johnson et al. [186] annotated the tumor type from scRRBS data using the MolecularNeuropathology classification tool [187].

Differential methylation modeling and motif analysis

DNA methylation differences across cells, cell clusters, and genomic regions can be inferred using standard statistical tests or methods previously developed for bulk methylome analyses (**Figure 4**). Statistical significance, effect size, and multiple test correction are three commonly used filtering criteria in calling differentially methylated CpGs (DMCs), differentially methylated windows (DMWs), differentially methylated CpG islands (DM-CGIs), and differentially methylated regions (DMRs). For example, Gravina et al. used two-sided *z*-tests and *t*-tests for DMW detection [61]. Hui et al. first calculated DMCs using the z-score method before merging them into DMRs [62]. Hou et al. used Fisher's exact test for detecting DM-CGIs and required $P < 0.05$ and a minimum methylation difference > 0.3 between two subpopulations [72]. Farlik et al. first identified consistently methylated regions and then used a *t*-test with a *P* value cutoff of 0.01 to assess whether a region is a DMR [100]. Zhu et al. [123] and Bian et al. [73] used a false discovery rate (FDR)-adjusted Student's *t*-test to detect 300 bp DMWs with FDR > 0.05 and polarized methylation levels ≥ 0.8 in the higher group and ≤ 0.2 in the lower group. Gaiti et al. defined DMRs based on the absolute weighted methylation difference (> 0.3) and a two-sided nonparametric permutation test ($P < 0.05$) [68]. Luo et al. [63] and Liu et al. [66] used the DMRfind function in methylpy [109] to calculate DMRs across subtypes. Luo et al. further merged neighboring DMRs into larger DMRs at least 1 kb apart. To detect regions that lack methylation and are potentially subject to regulatory machinery binding, Li et al. used MethylSeekR to call lowly methylated regions (LMRs) in human ESCs [188]. scMET models the number of methylated CpGs using a beta-binomial model before using a generalized linear model framework to test differential methylation [189]. This framework explicitly accounts for sequence feature-specific factors (*e.g.*, CpG density) to disentangle overdispersion of biological causes from technical variations.

DMCs, DMRs, and LMRs may be further scanned for the presence/enrichment of sequence motifs using tools such as hypergeometric optimization of motif enrichment

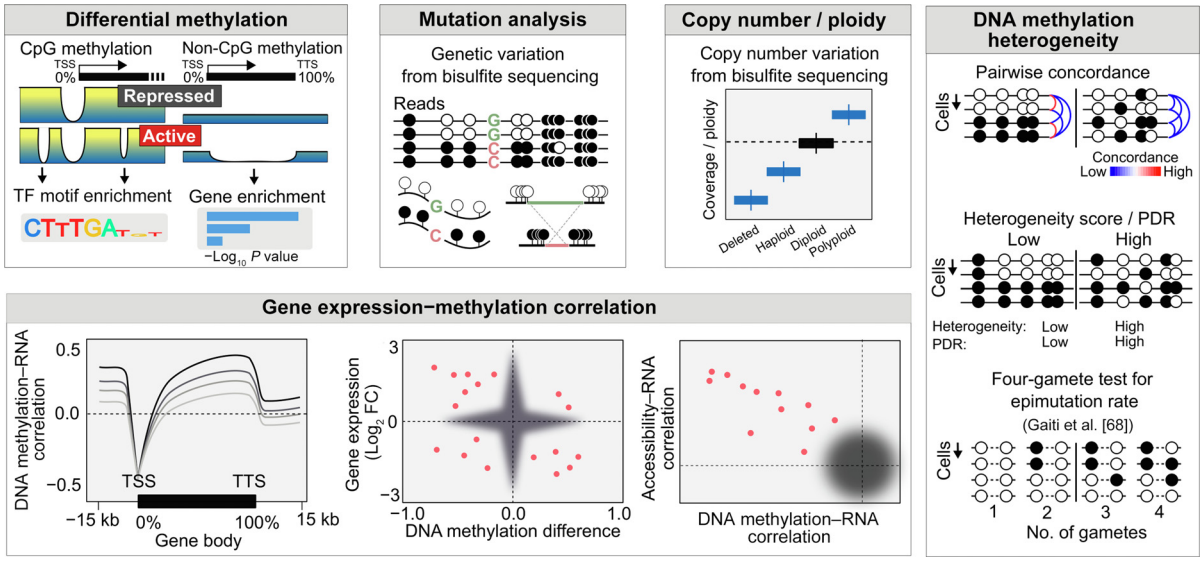

Figure 4 Schematic representation of DNA methylation variation and its association with mutational, transcriptional, and chromatin accessibility signals

Differential methylation: (1) CpG methylation differences at regulatory elements (these differences can be associated with TF binding and can be enriched for binding motifs); (2) non-CpG (CpH) methylation differences at gene bodies (these differences can predict gene transcriptional states and can be used for pathway enrichment analysis). Mutation analysis: mutation calling from bisulfite sequencing (white circles represent unmethylated reads, and black circles represent methylated reads). Copy number / ploidy: genome coverage signal can be used to determine ploidy using bisulfite sequencing. Gene expression−methylation correlation: (1) correlation of DNA methylation with RNA expression varies across genomic regions (*e.g.*, TSS methylations may be anti-correlated or uncorrelated with RNA expression, while gene body methylations may be positively correlated with RNA expression); (2) comparison of RNA expression change with DNA methylation change in single cells; (3) comparison of correlations between DNA methylation and RNA expression and between chromatin accessibility and RNA expression. Open accessibility regions are positively correlated with gene expression, while DNA methylation (*e.g.*, at enhancers) is generally negatively correlated with gene expression. The pink dots represent outlier samples profiled for RNA expression and DNA methylation (middle panel) or RNA expression, chromatin accessibility, and DNA methylation (right panel). DNA methylation heterogeneity: (1) pairwise concordance of cells (high concordances are marked by red lines, and low concordances are marked by blue lines) can be determined based on methylation patterns (white circles indicate unmethylated cytosines and black circles represent methylated cytosines); (2) methylation states (white circles indicate unmethylated cytosines and black circles represent methylated cytosines) in cells are used for heterogeneity scoring based on PDR; (3) determining epimutation rate using the four-gamete test (figure adapted from Gaiti et al. [68]). TTS, transcription termination site; PDR, proportion of discordant reads.

(HOMER) [190] or the multiple expectation maximizations for motif elicitation (MEME) suite [191]. The identified motifs can then be matched against TF binding motifs in well-curated databases based on motif–motif similarity (Tomtom [192]). Commonly used motif databases include TRANSFAC [193], UniPROBE [194], JASPAR [195], Cis-BP [196], TFClass [197], and HOCOMOCO [198]. For testing enrichment, the choice of background is critical. For example, analyses testing for the enrichment of TF binding motifs often use H3K27ac sites as the background [199] to look for differential methylations associated with enhancer regions. H3K27ac peaks are generally associated with all transcriptionally activated TF bindings and can be used as the null distribution when testing for the binding of specific TFs.

Mutation and copy number analysis

Like many omics data, single-cell methylome data harbor rich information regarding the DNA itself, such as genetic variation [200] and copy number alterations [201] (Figure 4). Zhu et al. used a binomial test and Bis-SNP [99] to call SNPs from bisulfite sequencing data to identify parental-specific methylations [123]. Similarly, Li et al. identified parental SNPs from scCOOL-seq data and linked neighboring genomic sites to identify parental allele-specific methylations and chromatin accessibility [188]. Farlik et al. compared the methylome coverage of HL60 and K562 cells against HL60- and K562-specific copy number alterations to verify cell identity [100]. Hou et al. [72] inferred copy numbers from the scTrio-seq data by using a hidden Markov model (HMM) [202]. Guo et al. used HMMcopy [203] to infer copy numbers using normalized read counts from scCOOL-seq data [76]. Bian et al. [73] and Johnson et al. [186] used Ginkgo [204] to infer copy number alterations from single-cell methylomes.

Stochastic DNA methylation variation and heterogeneity

Measuring DNA methylome at the single-cell resolution allows us to distinguish stochastic methylation changes (epigenetic drift) from coordinated methylation changes by studying consistency in the local methylation patterns within and across cells (Figure 4). With bulk DNA methylome data, this was done mainly through read-level analyses using metrics such as epi-polymorphisms [205], proportion of discordantly methylated reads (PDR) [206], methylation entropy, and methylated haplotype load (MHL) [207]. Smallwood et al. computed the cell-to-cell variance in methylation from single-cell methylomes and found that CpGs associated with active enhancer elements have significantly higher variances of methylation than CpGs in CGIs and intracisternal A-particle repeat DNA [65]. Farlik et al. tracked the pairwise Euclidean distance among single cells before and after 5-azacytidine and vitamin D treatment. They identified a temporary surge in variability as cells individually transition in response to treatment [100]. Hernando-Herraez et al. developed a normalized methylation heterogeneity score to detect the hallmarks of aging and to identify DNA regions influenced by epigenetic drift [129,208]. The score is based on Hamming distance and Shannon entropy and accounts for the dependency of methylation variances on the methylation means [129]. Gaiti et al. measured epimutation rates based on the four-gamete test that allows for the calculation of methylation heterogeneity in CpG-sparse regions [68]. Johnson et al. used the PDR to conclude that glioma cells have a higher epigenetic heterogeneity in comparison to normal cells [186].

Integration with scRNA-seq data in different sample spaces

To facilitate the integration of DNA methylation with gene expression data from different sample spaces, one can leverage cytosine methylations correlated with gene expression to predict gene transcriptional activity in a computational procedure sometimes referred to as gene activity scoring [134,209–211]. Despite the extensive study of DNA methylation as a regulator of gene expression [5–7,212], it is often nontrivial to precisely characterize the methylation–expression relationship, which often depends on the cell type and genomic context of the cytosines. For example, although promoter CpG methylations were thought to be negatively associated with the gene expression, it is no longer considered a general genome-wide rule for most genes. Farlik et al. found that only a small number of differentially expressed genes showed differences in methylation [137]. This is in part due to the orthogonal gene expression regulatory process, *e.g.*, silencing via the polycomb repressive complexes, and other methylome-impinging processes, such as the cell division effect on late replicating DNA. In fact, 70%−80% of genomic CpGs were estimated to have stable methylation states, and only $\sim$ 15%−21% have dynamic methylation patterns in adults, depending on the tissue types [109,213]. As such, feature selection can be critical to the success of predicting gene expression based on DNA methylation. Luo et al. [63] and Liu et al. [66] leveraged the association between gene body non-CpG cytosine methylations and gene expression activity to find methylation markers for neuron subtypes. However, this approach can be limited to a subset of cells with high non-CpG cytosine methylations, such as neurons [109]. CpG methylations at *cis*-regulatory elements are more commonly used for other cell types to inform gene expression regulation [18,137].

Several strategies have been adopted to predict gene expression from DNA methylation, including using support vector linear regression (SVLR) by BPRmeth [134] and ensemble machine learning by MAPLE [211], which leverage previously generated co-assay data as training [74,77,96,129,189] (**Figure 5**). These methods are designed to be used in combination with specific feature engineering methods such as fixed-size window smoothing [211] and Bayesian clustering [134]. The predicted synthetic gene expression matrices can then be co-analyzed with single-cell gene expression datasets using canonical correlation analysis (*e.g.*, by Seurat [214]), mutual nearest neighbor (MNN) analysis (*e.g.*, by Scanorama [215]), weighted nearest neighbor analysis (WNN) (*e.g.*, by Seurat [216]), and the batch-balanced k-nearest neighbor method (BBKNN [217]). For example, EpiScanpy employed BBKNN to co-embed data from different sources [138].

Alternatively, one can also integrate DNA methylation and RNA-seq data without explicit gene expression prediction (Figure 5). MATCHER projects cells from both the DNA methylation and RNA-seq datasets to a "master" pseudotime scale (0 to 1) using a Gaussian process latent variable

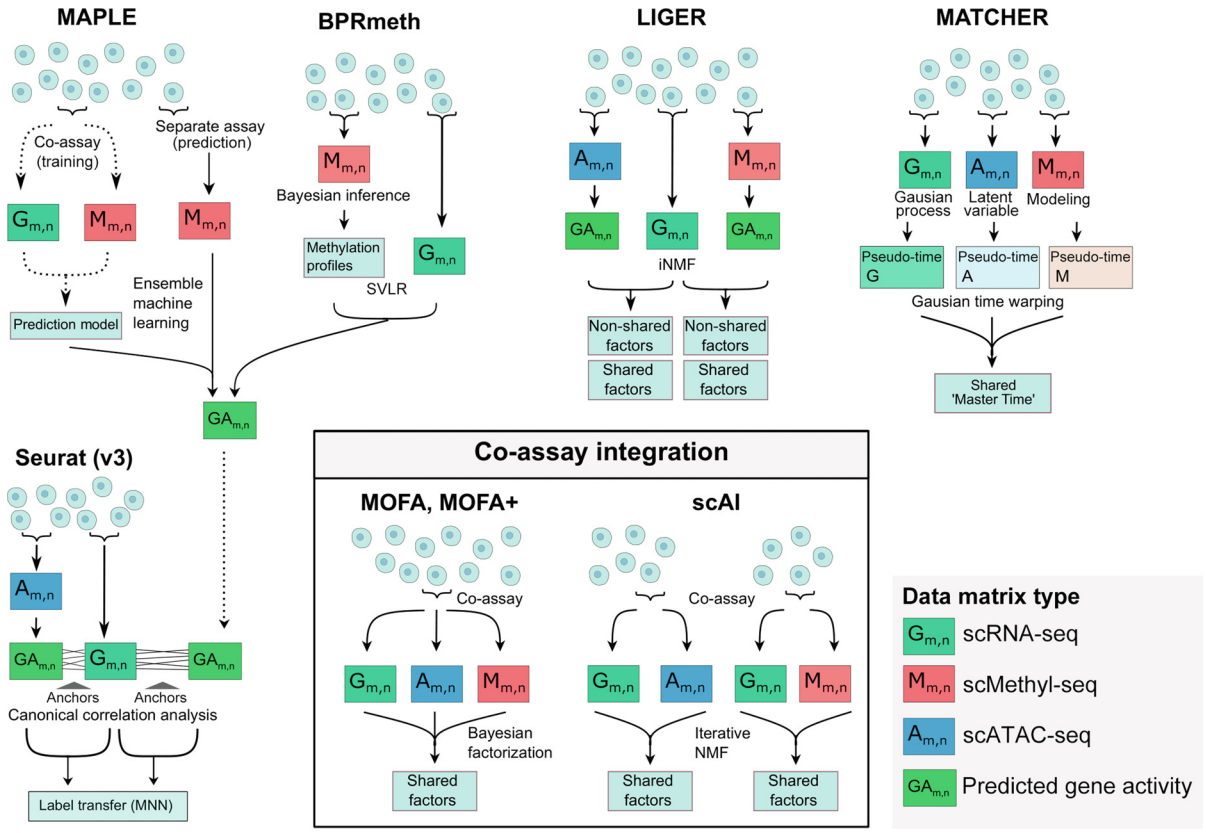

Figure 5 Overview of bioinformatics tools that integrate single-cell methylome data with other omics datasets

The top panel includes the bioinformatics tools that can incorporate data from different cells for multi-omics integration. MATCHER projects each data modality to a pseudo-time scale. Multi-omics integration can also be achieved through integrating scRNA-seq data (as by LIGER and Seurat) with gene activity matrix predicted from scATAC-seq and scMethyl-seq data (as by LIGER, BPRmeth, and MAPLE). In Seurat, 'anchors' between datasets are identified using canonical correlation analysis followed by label transfer using the MNN method. In LIGER, iNMF identifies shared and non-shared factors between datasets. BPRmeth first learns methylation features or profiles of a given cell type and then uses SVLR to predict gene expression corresponding to the methylation profiles. MAPLE produces the gene activity matrix by ensemble machine learning. MOFA and scAI integrate scMethyl-seq data with the other modalities co-assayed in the same cells in the bottom panel. scRNA-seq, single-cell RNA sequencing; scMethyl-seq, single-cell methylation sequencing; scATAC-seq, single-cell assay for transposase-accessible chromatin sequencing; MNN, mutual nearest neighbor; NMF, non-negative matrix factorization; iNMF, integrative non-negative matrix factorization; SVLR, support vector linear regression; scAI, single-cell aggregation and integration.

model (GLVM) [176]. The two projections can then be combined to establish cell-to-cell correspondence through manifold alignment [218]. One can also perform joint matrix factorization using this alignment (*e.g.*, by coupled NMF [219], LIGER [210], CSMF [220]). Notably, LIGER [221] and its online adaptation [222] employ integrative NMF (iNMF) [223] for joint dimensionality reduction of single-cell expression and methylome data. Cells can then be represented on a shared factor neighborhood---a low-dimensional space spanned by factor loadings. This analysis can be semi-supervised by biologically-guided anchor cells and anchoring features (see reviews of integrating multiple omics datasets [46,224] for the principle of this approach).

Multi-omics co-assays including DNA methylome

Single-cell multi-omics co-assay technologies profile multiple omics data in the same cells and bypass the challenge of aligning cells from single-cell methylomes to other data modalities [225]. Notable single-cell co-assay methods include ones that combine DNA methylation with gene transcription (scM&T-seq [74], scTrio-seq [72], and smart-RRBS [70]), chromatin accessibility (scNOMe-seq [75] and ATAC-Me [226]), both gene transcription and chromatin accessibility (scCOOL-seq [76] and snNMT-seq [77]), and 3D genome conformation (methyl-HiC [79] and sn-methyl-3C [80]). One can study the relationship of DNA methylation with other molecular phenotypes directly in the same cells. Hou et al. [72] and Bian et al. [73] verified the known negative association of gene expression with promoter methylations and positive association with gene body methylations in single cells using scTrio-seq. Gaiti et al. used Smart-RRBS and identified a negative correlation between gene expression and promoter DNA methylations in normal B cells and more prominently in lymphocytic leukemic cells [68]. Applying scNMT-seq to mouse gastrulation cells, Argelaguet et al. identified 125 genes with expression significantly correlated with promoter methylations [96]. Based on a Bayesian group factor analysis framework, the authors also developed and applied MOFA [156] and MOFA +

[227] to perform dimensionality reduction and integration in the reduced latent factor space (Figure 5). Similarly, scAI also took a unified matrix factorization approach to analyze single-cell multi-omics co-assay data that included the DNA methylome [228] (Figure 5).

Ongoing challenges and future directions

Many methods have been developed for analyzing scRNA-seq and scATAC-seq data. However, single-cell DNA methylome analysis methods remain relatively limited, with comprehensive software tool suites only emerging recently [138,229]. This is partly attributable to the DNA methylome being high in dimensionality and low in copy number. The probability of missing DNA methylation information per CpG is higher than that of missing a transcriptional signal in a scRNA-seq experiment of similar sequencing depth. As a result, compared to scRNA-seq data, analyzing single-cell DNA methylome data is often more challenging and demands more advanced feature selection.

The challenge of single-cell DNA methylome analysis can also be attributed to the complex grammar of DNA methylation determination, coordinated by the biochemical and cellular processes that deposit, dilute, and erase the methylation marks on the cytosine bases. First, these processes often operate at different genomic scales. For example, replication timing-associated DNA methylation change takes place on mega-base-pair-scale domains [42,43,230]. Non-CpG cytosine methylations are correlated with gene transcription at gene bodies in neurons [66]. TF binding dictates more focal DNA methylation patterns [231,232]. The difference in these scales of representation requires feature selection and analysis to be performed at different genomic scales [44]. Second, supervised analysis of focal lineage-specific or disease-specific methylation alterations needs to account for processes that impact the DNA methylation level globally to avoid confounding effects. These processes include the action of methylation readers, writers, and other cellular phenotypes, such as the cell cycle stage, which is not usually included in the bulk methylome analysis. Supervised single-cell annotation methodology based on grammar learned from bulk DNA methylomes is yet to be developed.

An implication of the multi-factor determination of DNA methylation is the relative lack of data that capitulate single-cell DNA methylome under diverse biological conditions. Although we are witnessing an explosive increase in the single-cell DNA methylome data volume [88], most data are assayed in individuals of one age, one genetic background, and one pathological state. For example, state-of-the-art tools like LIGER [210] and MAPLE [211] were trained using a limited number of real datasets and synthetic datasets; therefore, whether their performance can be extended to other biological and technical scenarios remains to be verified. Nevertheless, several ongoing single-cell multi-omics data consortia have started to include DNA methylation and will generate more detailed DNA methylome references under more diverse conditions [233].

Aside from the complex biological grammar, current public single-cell methylome data are often obtained using different assay technologies [53], which challenges their integration. For example, combinatorial barcoding usually yields more cells at lower sequencing depth and is more susceptible to allelic dropout than plate-based methods [64,234]. The discrepancy of different assay technologies in genomic coverage can also cause systematic bias when the window-smoothed signal is compared. Whether solutions to correct batch and platform-specific effects in scRNA-seq data [235] can be applied to DNA methylome data is yet to be validated. The discrepancies among assay technologies also require more general and customizable computational methods. A comprehensive benchmarking of single-cell methylome assay technologies and analysis tools (like existing benchmark efforts for the scRNA-seq data [127,175]) is a pressing unmet need. Methods that can adequately integrate data from different assay technologies and provide a standard cell-specific methylome reference need to be better developed.

The prevalence of single-cell transcriptome and chromatin accessibility data [236–239] presents the question of how DNA methylome data can be co-analyzed with scRNA-seq and chromatin accessibility. The added value provided by single-cell DNA methylome data needs to be better quantified. Multiple methods have been proposed to analyze single-cell methylome data and scRNA-seq data in different [210,211] and same sample spaces [156,227,228]. Besides integration with scRNA-seq data, integration with scATAC-seq data is also feasible. For example, Li et al. used scCOOL-seq to report a negative correlation between chromatin accessibility and DNA methylation in embryonic cells [188]. Integration with single-cell chromatin accessibility data remains an active direction of development. The coupleCoC+ tool uses an information-theoretic co-clustering framework for integrating multimodal single-cell genomic datasets, including single-cell methylome data [240].

The rich resource of informatics tools for scRNA-seq analysis [241] also raises the question of whether these tools can be effectively applied to single-cell DNA methylome data. Many scRNA-seq analysis strategies, *e.g.*, lineage construction, share the same principle as DNA methylome analyses. Therefore, repurposing these scRNA-seq tools for DNA methylome analyses can be highly feasible. For example, Liu et al. [66] extensively used Scanpy [242] for selecting variable methylations, dimensionality reduction, and nearest neighbor graph construction. EpiScanpy [138] used PAGA [158] for lineage inference and cell type abstraction. Complete repurposing of other tools for single-cell DNA methylome data requires coordinated efforts from experts in different omics fields.

Conclusion

The increasingly widespread use of single-cell DNA methylome profiling for tracking cell identity and understanding gene regulation in biomedical research and clinical applications has been producing single-cell methylome data on scales of increasing magnitude. The increase in the volume of single-cell methylome data demands the development of highly efficient, flexible, versatile, and easy-to-use computational tools for its analysis. Advances in analytical tools can help precipitate the adoption of modern high-throughput single-cell DNA methylome profiling technologies. This focus on single-cell methylome profiling will unveil the complexity of

intercellular interactions and gene regulation heterogeneity in broad biological and translational contexts.

Competing interests

Neither of the authors has any competing interest to declare.

CRediT authorship contribution statement

Waleed Iqbal: Writing – original draft, Writing – review & editing, Visualization. **Wanding Zhou:** Conceptualization, Writing – original draft, Writing – review & editing, Visualization. Both authors have read and approved the final manuscript.

Acknowledgments

The work was supported by the grants from the Children's Hospital of Philadelphia (CHOP) New Investigator Startup Funding (to WZ) and the FOXO Technologies Inc Research Sponsorship (to WZ). We thank Diljeet Kaur for proofreading the manuscript.

ORCID

ORCID 0000-0002-1478-2969 (Waleed Iqbal)
ORCID 0000-0001-9126-1932 (Wanding Zhou)

References

[1] Doskocil J, Sorm F. Distribution of 5-methylcytosine in pyrimidine sequences of deoxyribonucleic acids. Biochim Biophys Acta 1962;55:953–9.
[2] Riggs AD. X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet 1975;14:9–25.
[3] Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 2010;328:916–9.
[4] Li E, Zhang Y. DNA methylation in mammals. Cold Spring Harb Perspect Biol 2014;6:a019133.
[5] Ben-Hattar J, Jiricny J. Methylation of single CpG dinucleotides within a promoter element of the *Herpes simplex* virus *tk* gene reduces its transcription *in vivo*. Gene 1988;65:219–27.
[6] Watt F, Molloy PL. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. Genes Dev 1988;2:1136–43.
[7] Iguchi-Ariga SM, Schaffner W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. Genes Dev 1989;3:612–9.
[8] Ferguson-Smith AC, Sasaki H, Cattanach BM, Surani MA. Parental-origin-specific epigenetic modification of the mouse *H19* gene. Nature 1993;362:751–5.
[9] Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. Nature 1993;366:362–5.
[10] Bartolomei MS, Webber AL, Brunkow ME, Tilghman SM. Epigenetic mechanisms underlying the imprinting of the mouse *H19* gene. Genes Dev 1993;7:1663–73.
[11] Stöger R, Kubicka P, Liu CG, Kafri T, Razin A, Cedar H, et al. Maternal-specific methylation of the imprinted mouse *Igf2r* locus identifies the expressed locus as carrying the imprinting signal. Cell 1993;73:61–71.
[12] Mohandas T, Sparkes RS, Shapiro LJ. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. Science 1981;211:393–6.
[13] Lock LF, Takagi N, Martin GR. Methylation of the *Hprt* gene on the inactive X occurs after chromosome inactivation. Cell 1987;48:39–46.
[14] Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet 1998;20:116–7.
[15] Estécio MRH, Gallegos J, Vallot C, Castoro RJ, Chung W, Maegawa S, et al. Genome architecture marked by retrotransposons modulates predisposition to DNA methylation in cancer. Genome Res 2010;20:1369–82.
[16] Zhou W, Liang G, Molloy PL, Jones PA. DNA methylation enables transposable element-driven genome expansion. Proc Natl Acad Sci U S A 2020;117:19359–66.
[17] Deniz Ö, Frost JM, Branco MR. Regulation of transposable elements by DNA modifications. Nat Rev Genet 2019;20:417–31.
[18] Reizel Y, Sabag O, Skversky Y, Spiro A, Steinberg B, Bernstein D, et al. Postnatal DNA demethylation and its role in tissue maturation. Nat Commun 2018;9:2040.
[19] Reizel Y, Morgan A, Gao L, Schug J, Mukherjee S, García MF, et al. FoxA-dependent demethylation of DNA initiates epigenetic memory of cellular identity. Dev Cell 2021;56:602–12.e4.
[20] Luo C, Hajkova P, Ecker JR. Dynamic DNA methylation: in the right place at the right time. Science 2018;361:1336–40.
[21] Greenberg MVC, Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol 2019;20:590–607.
[22] Zhou F, Wang R, Yuan P, Ren Y, Mao Y, Li R, et al. Reconstituting the transcriptome and DNA methylome landscapes of human implantation. Nature 2019;572:660–4.
[23] Gkountela S, Zhang KX, Shafiq TA, Liao WW, Hargan-Calvopiña J, Chen PY, et al. DNA demethylation dynamics in the human prenatal germline. Cell 2015;161:1425–36.
[24] Guo H, Zhu P, Yan L, Li R, Hu B, Lian Y, et al. The DNA methylation landscape of human early embryos. Nature 2014;511:606–10.
[25] Guo F, Yan L, Guo H, Li L, Hu B, Zhao Y, et al. The transcriptome and DNA methylome landscapes of human primordial germ cells. Cell 2015;161:1437–52.
[26] Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet 2018;19:371–84.
[27] Horvath S. DNA methylation age of human tissues and cell types. Genome Biol 2013;14:R115.
[28] Zhong J, Agha G, Baccarelli AA. The role of DNA methylation in cardiovascular risk and disease: methodological aspects, study design, and data analysis for epidemiological studies. Circ Res 2016;118:119–31.
[29] Dalgaard K, Landgraf K, Heyne S, Lempradl A, Longinotto J, Gossens K, et al. Trim28 haploinsufficiency triggers bi-stable epigenetic obesity. Cell 2016;164:353–64.
[30] Balnis J, Madrid A, Hogan KJ, Drake LA, Chieng HC, Tiwari A, et al. Blood DNA methylation and COVID-19 outcomes. Clin Epigenet 2021;13:118.
[31] Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature 2017;541:81–6.
[32] Jones PA, Issa JPJ, Baylin S. Targeting the cancer epigenome for therapy. Nat Rev Genet 2016;17:630–41.
[33] Shen H, Laird PW. Interplay between the cancer genome and epigenome. Cell 2013;153:38–55.
[34] Baylin SB, Jones PA. Epigenetic determinants of cancer. Cold Spring Harb Perspect Biol 2016;8:a019505.

[35] Robert S, Servis RE, Marvin W. Reactions of uracil and cytosine derivatives with sodium bisulfite. J Am Chem Soc 1970;92:422–4.
[36] Hayatsu H, Wataya Y, Kazushige K. The addition of sodium bisulfite to uracil and to cytosine. J Am Chem Soc 1970;92:724–6.
[37] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009;462:315–22.
[38] Cohen NM, Kenigsberg E, Tanay A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell 2011;145:773–86.
[39] Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. N Engl J Med 2003;349:2042–54.
[40] Héberlé É, Bardet AF. Sensitivity of transcription factors to DNA methylation. Essays Biochem 2019;63:727–41.
[41] Luo X, Zhang T, Zhai Y, Wang F, Zhang S, Wang G. Effects of DNA methylation on TFs in human embryonic stem cells. Front Genet 2021;12:639461.
[42] Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet 2012;44:40–6.
[43] Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet 2018;50:591–602.
[44] Knijnenburg TA, Ramsey SA, Berman BP, Kennedy KA, Smit AFA, Wessels LFA, et al. Multiscale representation of genomic signals. Nat Methods 2014;11:689–94.
[45] Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods 2017;14:565–71.
[46] Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet 2019;20:257–72.
[47] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol 2019;15:e8746.
[48] Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 2019;20:273–82.
[49] Sinha S, Satpathy AT, Zhou W, Ji H, Stratton JA, Jaffer A, et al. Profiling chromatin accessibility at single-cell resolution. Genomics Proteomics Bioinformatics 2021;19:172–90.
[50] Baek S, Lee I. Single-cell ATAC sequencing analysis: from data preprocessing to hypothesis generation. Comput Struct Biotechnol J 2020;18:1429–39.
[51] Pal K, Forcato M, Ferrari F. Hi-C analysis: from data generation to integration. Biophys Rev 2019;11:67–78.
[52] Zhou T, Zhang R, Ma J. The 3D genome structure of single cells. Annu Rev Biomed Data Sci 2021;4:21–41.
[53] Ahn J, Heo S, Lee J, Bang D. Introduction to single-cell DNA methylation profiling methods. Biomolecules 2021;11:1013.
[54] Sun Z, Vaisvila R, Hussong LM, Yan B, Baum C, Saleh L, et al. Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Genome Res 2021;31:291–300.
[55] Liu Y, Siejka-Zielińska P, Velikova G, Bi Y, Yuan F, Tomkova M, et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat Biotechnol 2019;37:424–9.
[56] Liu Y, Cheng J, Siejka-Zielińska P, Weldon C, Roberts H, Lopopolo M, et al. Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. Genome Biol 2020;21:54.
[57] Liu Y, Hu Z, Cheng J, Siejka-Zielińska P, Chen J, Inoue M, et al. Subtraction-free and bisulfite-free specific sequencing of 5-methylcytosine and its oxidized derivatives at base resolution. Nat Commun 2021;12:618.
[58] 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 2013;23:2126–35.
[59] Miura F, Enomoto Y, Dairiki R, Ito T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res 2012;40:e136.
[60] Clark SJ, Smallwood SA, Lee HJ, Krueger F, Reik W, Kelsey G. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). Nat Protoc 2017;12:534–47.
[61] 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.
[62] Hui T, Cao Q, Wegrzyn-Woltosz J, O'Neill K, Hammond CA, Knapp DJHF, et al. High-resolution single-cell DNA methylation measurements reveal epigenetically distinct hematopoietic stem cell subpopulations. Stem Cell Rep 2018;11:578–92.
[63] Luo C, Keown CL, Kurihara L, Zhou J, He Y, Li J, et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science 2017;357:600–4.
[64] Mulqueen RM, Pokholok D, Norberg SJ, Torkenczy KA, Fields AJ, Sun D, et al. Highly scalable generation of DNA methylation profiles in single cells. Nat Biotechnol 2018;36:428–31.
[65] Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods 2014;11:817–20.
[66] Liu H, Zhou J, Tian W, Luo C, Bartlett A, Aldridge A, et al. DNA methylation atlas of the mouse brain at single-cell resolution. Nature 2021;598:120–8.
[67] Guo H, Zhu P, Guo F, Li X, Wu X, Fan X, et al. Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. Nat Protoc 2015;10:645–59.
[68] Gaiti F, Chaligne R, Gu H, Brand RM, Kothen-Hill S, Schulman RC, et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. Nature 2019;569:576–80.
[69] Shareef SJ, Bevill SM, Raman AT, Aryee MJ, van Galen P, Hovestadt V, et al. Extended-representation bisulfite sequencing of gene regulatory elements in multiplexed samples and single cells. Nat Biotechnol 2021;39:1086–94.
[70] Gu H, Raman AT, Wang X, Gaiti F, Chaligne R, Mohammad AW, et al. Smart-RRBS for single-cell methylome and transcriptome analysis. Nat Protoc 2021;16:4004–30.
[71] Adey A, Shendure J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. Genome Res 2012;22:1139–43.
[72] Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res 2016;26:304–19.
[73] Bian S, Hou Y, Zhou X, Li X, Yong J, Wang Y, et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. Science 2018;362:1060–3.
[74] Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods 2016;13:229–32.
[75] Pott S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. ELife 2017;6:e23203.
[76] Guo F, Li L, Li J, Wu X, Hu B, Zhu P, et al. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. Cell Res 2017;27:967–88.
[77] Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of

chromatin accessibility DNA methylation and transcription in single cells. Nat Commun 2018;9:781.
[78] Wang Y, Wang A, Liu Z, Thurman AL, Powers LS, Zou M, et al. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. Genome Res 2019;29:1329–42.
[79] Li G, Liu Y, Zhang Y, Kubo N, Yu M, Fang R, et al. Joint profiling of DNA methylation and chromatin architecture in single cells. Nat Methods 2019;16:991–3.
[80] Lee DS, Luo C, Zhou J, Chandran S, Rivkin A, Bartlett A, et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nat Methods 2019;16:999–1006.
[81] Han L, Wu HJ, Zhu H, Kim KY, Marjani SL, Riester M, et al. Bisulfite-independent analysis of CpG island methylation enables genome-scale stratification of single cells. Nucleic Acids Res 2017;45:e77.
[82] Niemöller C, Wehrle J, Riba J, Claus R, Renz N, Rhein J, et al. Bisulfite-free epigenomics and genomics of single cells through methylation-sensitive restriction. Commun Biol 2021;4:153.
[83] Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol 2021;39:1348–65.
[84] Sharim H, Grunwald A, Gabrieli T, Michaeli Y, Margalit S, Torchinsky D, et al. Long-read single-molecule maps of the functional methylome. Genome Res 2019;29:646–56.
[85] Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, et al. Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods 2017;14:411–3.
[86] Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods 2017;14:407–10.
[87] Song CX, Diao J, Brunger AT, Quake SR. Simultaneous single-molecule epigenetic imaging of DNA methylation and hydroxymethylation. Proc Natl Acad Sci U S A 2016;113:4338–43.
[88] Karemaker ID, Vermeulen M. Single-cell DNA methylation profiling: technologies and biological applications. Trends Biotechnol 2018;36:952–65.
[89] Evrony GD, Hinch AG, Luo C. Applications of single-cell DNA sequencing. Annu Rev Genomics Hum Genet 2021;22:171–97.
[90] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 2011;27:1571–2.
[91] Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 2009;10:232.
[92] Chen PY, Cokus SJ, Pellegrini M. BS Seeker: precise mapping for bisulfite sequencing. BMC Bioinformatics 2010;11:203.
[93] Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics 2013;14:774.
[94] Huang KYY, Huang YJ, Chen PY. BS-Seeker3: ultrafast pipeline for bisulfite sequencing. BMC Bioinformatics 2018;19:111.
[95] Lin X, Sun D, Rodriguez B, Zhao Q, Sun H, Zhang Y, et al. BSeQC: quality control of bisulfite sequencing experiments. Bioinformatics 2013;29:3227–9.
[96] Argelaguet R, Clark SJ, Mohammed H, Stapel LC, Krueger C, Kapourani CA, et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature 2019;576:487–91.
[97] Wu P, Gao Y, Guo W, Zhu P. Using local alignment to enhance single-cell bisulfite sequencing data efficiency. Bioinformatics 2019;35:3273–8.
[98] Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Res 2018;46:e120.
[99] Liu Y, Siegmund KD, Laird PW, Berman BP. Bis-SNP: combined DNA methylation and SNP calling for bisulfite-seq data. Genome Biol 2012;13:R61.
[100] Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Rep 2015;10:1386–97.
[101] Holmes EE, Jung M, Meller S, Leisse A, Sailer V, Zech J, et al. Performance evaluation of kits for bisulfite-conversion of DNA from tissues, cell lines, FFPE tissues, aspirates, lavages, effusions, plasma, serum, and urine. PLoS One 2014;9:e93933.
[102] Olova N, Krueger F, Andrews S, Oxley D, Berrens RV, Branco MR, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol 2018;19:33.
[103] Genereux DP, Johnson WC, Burden AF, Stöger R, Laird CD. Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies. Nucleic Acids Res 2008;36:e150.
[104] Warnecke PM, Stirzaker C, Melki JR, Millar DS, Paul CL, Clark SJ. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. Nucleic Acids Res 1997;25:4422–6.
[105] Warnecke PM, Stirzaker C, Song J, Grunau C, Melki JR, Clark SJ. Identification and resolution of artifacts in bisulfite sequencing. Methods 2002;27:101–7.
[106] Guo JU, Su Y, Shin JH, Shin J, Li H, Xie B, et al. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nat Neurosci 2014;17:215–22.
[107] Sirard MA. Distribution and dynamics of mitochondrial DNA methylation in oocytes, embryos and granulosa cells. Sci Rep 2019;9:11937.
[108] Tang WWC, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, et al. A unique gene regulatory network resets the human germline epigenome for development. Cell 2015;161:1453–67.
[109] Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature 2015;523:212–6.
[110] Schutsky EK, DeNizio JE, Hu P, Liu MY, Nabel CS, Fabyanic EB, et al. Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. Nat Biotechnol 2018;36:1083–90.
[111] Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 2013;14:618–30.
[112] Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. Mol Cell 2015;58:598–609.
[113] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.
[114] Celton-Morizur S, Desdouets C. Polyploidization of liver cells. Adv Exp Med Biol 2010;676:123–35.
[115] Brodsky VYA, Sarkisov DS, Arefyeva AM, Panova NW, Gvasava IG. Polyploidy in cardiac myocytes of normal and hypertrophic human hearts; range of values. Virchows Arch 1994;424:429–35.
[116] Zimmet J, Ravid K. Polyploidy: occurrence in nature, mechanisms, and significance for the megakaryocyte-platelet system. Exp Hematol 2000;28:3–16.
[117] Gordon DJ, Resio B, Pellman D. Causes and consequences of aneuploidy in cancer. Nat Rev Genet 2012;13:189–203.
[118] McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst 2019;8:329–37.e4.
[119] Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Syst 2019;8:281–91.e9.
[120] Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable software package for

integrative single-cell chromatin accessibility analysis. Nat Genet 2021;53:403–11.
[121] Fang R, Preissl S, Li Y, Hou X, Lucero J, Wang X, et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat Commun 2021;12:1337.
[122] Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell 2020;183:1103–16.e20.
[123] Zhu P, Guo H, Ren Y, Hou Y, Dong J, Li R, et al. Single-cell DNA methylome sequencing of human preimplantation embryos. Nat Genet 2018;50:12–9.
[124] Pott S, Lieb JD. Single-cell ATAC-seq: strength in numbers. Genome Biol 2015;16:172.
[125] Trapnell C. Defining cell types and states with single-cell genomics. Genome Res 2015;25:1491–8.
[126] Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature 2017;541:331–8.
[127] Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol 2020;38:737–46.
[128] Kim TH, Zhou X, Chen M. Demystifying "drop-outs" in single-cell UMI data. Genome Biol 2020;21:196.
[129] Hernando-Herraez I, Evano B, Stubbs T, Commere PH, Jan Bonder M, Clark S, et al. Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. Nat Commun 2019;10:4361.
[130] Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: identification of problematic regions of the genome. Sci Rep 2019;9:9354.
[131] Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature 2014;507:455–61.
[132] Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. Genome Biol 2015;16:56.
[133] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[134] Kapourani CA, Sanguinetti G. BPRMeth: a flexible Bioconductor package for modelling methylation profiles. Bioinformatics 2018;34:2485–6.
[135] Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol 2015;16:14.
[136] Zou LS, Erdos MR, Taylor DL, Chines PS, Varshney A, McDonnell Genome Institute, et al. BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. BMC Genomics 2018;19:390.
[137] Farlik M, Halbritter F, Müller F, Choudry FA, Ebert P, Klughammer J, et al. DNA methylation dynamics of human hematopoietic stem cell differentiation. Cell Stem Cell 2016;19:808–22.
[138] Danese A, Richter ML, Chaichoompu K, Fischer DS, Theis FJ, Colomé-Tatché M. EpiScanpy: integrated single-cell epigenomic analysis. Nat Commun 2021;12:5228.
[139] Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol 2017;18:67.
[140] Taudt A, Roquis D, Vidalis A, Wardenaar R, Johannes F, Colomé-Tatché M. METHimpute: imputation-guided construction of complete methylomes from WGBS data. BMC Genomics 2018;19:444.
[141] Shahryary Y, Hazarika RR, Johannes F. MethylStar: a fast and robust pre-processing pipeline for bulk or single-cell whole-genome bisulfite sequencing data. BMC Genomics 2020;21:479.
[142] Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol 2012;13:R83.
[143] Kapourani CA, Sanguinetti G. Melissa: bayesian clustering and imputation of single-cell methylomes. Genome Biol 2019;20:61.
[144] de Souza PE, Andronescu M, Masud T, Kabeer F, Biele J, Laks E, et al. Epiclomal: probabilistic clustering of sparse single-cell DNA methylation data. PLoS Comput Biol 2020;16:e1008270.
[145] Jiang L, Wang C, Tang J, Guo F. LightCpG: a multi-view CpG sites detection on single-cell whole genome sequence data. BMC Genomics 2019;20:306.
[146] Tang J, Zou J, Fan M, Tian Q, Zhang J, Fan S. CaMelia: imputation in single-cell methylomes based on local similarities between cells. Bioinformatics 2021;37:1814–20.
[147] De Waele G, Clauwaert J, Menschaert G, Waegeman W. CpG Transformer for imputation of single-cell methylomes. Bioinformatics 2022;38:597–603.
[148] Yu B, Dong X, Gravina S, Kartal Ö, Schimmel T, Cohen J, et al. Genome-wide, single-cell DNA methylomics reveals increased non-CpG methylation during human oocyte maturation. Stem Cell Rep 2017;9:397–407.
[149] Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, et al. Epigenomic signatures of neuronal diversity in the mammalian brain. Neuron 2015;86:1369–84.
[150] Pearson K. On lines and planes of closest fit to systems of points in space. Philos Mag Ser 1901;6:559–72.
[151] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788–91.
[152] Kruskal J, Wish M. Multidimensional Scaling. Thousand Oaks: SAGE Publications, Inc; 1978.
[153] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.
[154] McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv 2018;1802.03426.
[155] Hjort NL, Holmes C, Muller P, Walker SG. Bayesian Nonparametrics. New York: Cambridge University Press; 2010.
[156] Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol 2018;14:e8124.
[157] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Appl Stat 1991;40:486.
[158] Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol 2019;20:59.
[159] MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and Probability 1967;14:281–97.
[160] Kaufman L, Rousseeuw PJ. Partitioning around medoids (program PAM). In: Kaufman L, Rousseeuw PJ, editors. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken: John Wiley & Sons, Inc; 1990, p.68–125.
[161] Cheng YZ. Mean shift, mode seeking, and clustering. IEEE Trans Pattern Anal Mach Intell 1995;17:790–9.
[162] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining 1996;96:226–31.
[163] Johnson SC. Hierarchical clustering schemes. Psychometrika 1967;32:241–54.
[164] Ferguson TS. A bayesian analysis of some nonparametric problems. Ann Statist 1973;1:209–30.
[165] Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007;315:972–6.
[166] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys Rev E Stat Nonlin Soft Matter Phys 2004;70:066111.

[167] Mucha PJ, Richardson T, Macon K, Porter MA, Onnela JP. Community structure in time-dependent, multiscale, and multiplex networks. Science 2010;328:876–8.
[168] Chung F. Spectral Graph Theory. Providence: American Mathematical Society; 1996.
[169] Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci U S A 2006;103:8577–82.
[170] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech 2008;2008:P10008.
[171] Rosvall M, Axelsson D, Bergstrom CT. The map equation. Eur Phys J Spec Top 2009;178:13–23.
[172] Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 2019;9:5233.
[173] Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 2015;347:1138–42.
[174] Strehl A, Ghosh J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 2003;3:583–617.
[175] Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol 2019;37:547–54.
[176] Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods 2016;13:845–8.
[177] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32:381–6.
[178] Bendall SC, Davis KL, Amir EAD, Tadmor MD, Simonds EF, Chen TJ, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell 2014;157:714–25.
[179] Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol 2016;34:637–45.
[180] Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 2017;14:587–9.
[181] Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 2015;32:268–74.
[182] Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. Ufboot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 2018;35:518–22.
[183] Siegmund KD, Marjoram P, Woo YJ, Tavaré S, Shibata D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. Proc Natl Acad Sci U S A 2009;106:4828–33.
[184] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.
[185] Trapp A, Kerepesi C, Gladyshev VN. Profiling epigenetic age in single cells. Nat Aging 2021;1:1189–201.
[186] Johnson KC, Anderson KJ, Courtois ET, Gujar AD, Barthel FP, Varn FS, et al. Single-cell multimodal glioma analyses identify epigenetic regulators of cellular plasticity and environmental stress response. Nat Genet 2021;53:1456–68.
[187] Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. Nature 2018;555:469–74.
[188] Li L, Guo F, Gao Y, Ren Y, Yuan P, Yan L, et al. Single-cell multi-omics sequencing of human early embryos. Nat Cell Biol 2018;20:847–58.
[189] Kapourani CA, Argelaguet R, Sanguinetti G, Vallejos CA. scMET: Bayesian modeling of DNA methylation heterogeneity at single-cell resolution. Genome Biol 2021;22:114.
[190] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. Mol Cell 2010;38:576–89.
[191] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics 2011;27:1017–8.
[192] Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol 2007;8:R24.
[193] Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, et al. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 2000;28:316–9.
[194] Newburger DE, Bulyk ML. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. Nucleic Acids Res 2009;37:D77–82.
[195] Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res 2010;38:D105–10.
[196] Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell 2014;158:1431–43.
[197] Wingender E, Schoeps T, Haubrock M, Krull M, Dönitz J. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. Nucleic Acids Res 2018;46:D343–7.
[198] Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res 2018;46:D252–9.
[199] McLeay RC, Bailey TL. Motif enrichment analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics 2010;11:165.
[200] Onuchic V, Lurie E, Carrero I, Pawliczek P, Patel RY, Rozowsky J, et al. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. Science 2018;361:361.
[201] Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature 2011;472:90–4.
[202] Ni X, Zhuo M, Su Z, Duan J, Gao Y, Wang Z, et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proc Natl Acad Sci U S A 2013;110:21083–8.
[203] Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. Genome Res 2012;22:1995–2007.
[204] Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, et al. Interactive analysis and assessment of single-cell copy-number variations. Nat Methods 2015;12:1058–60.
[205] Landan G, Cohen NM, Mukamel Z, Bar A, Molchadsky A, Brosh R, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat Genet 2012;44:1207–14.
[206] Landau DA, Clement K, Ziller MJ, Boyle P, Fan J, Gu H, et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer Cell 2014;26:813–25.
[207] Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor

tissue-of-origin mapping from plasma DNA. Nat Genet 2017;49:635–42.
[208] Veitia RA, Govindaraju DR, Bottani S, Birchler JA. Aging: somatic mutations, epigenetic drift and gene dosage imbalance. Trends Cell Biol 2017;27:299–310.
[209] Kapourani CA, Sanguinetti G. Higher order methylation features for clustering and prediction in epigenomic studies. Bioinformatics 2016;32:i405–12.
[210] Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. Jointly defining cell types from multiple single-cell datasets using LIGER. Nat Protoc 2020;15:3632–62.
[211] Uzun Y, Wu H, Tan K. Predictive modeling of single-cell DNA methylome data enhances integration with transcriptome data. Genome Res 2021;31:101–9.
[212] Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology 2013;38:23–38.
[213] Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LTY, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature 2013;500:477–81.
[214] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–20.
[215] Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol 2019;37:685–91.
[216] Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184:3573–87.e29.
[217] Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park JE. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics 2020;36:964–5.
[218] Welch JD, Hartemink AJ, Prins JF. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol 2017;18:138.
[219] Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. Proc Natl Acad Sci U S A 2018;115:7723–8.
[220] Zhang L, Zhang S. Learning common and specific patterns from data of multiple interrelated biological scenarios with matrix factorization. Nucleic Acids Res 2019;47:6606–17.
[221] Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell 2019;177:1873–87.e17.
[222] Gao C, Liu J, Kriebel AR, Preissl S, Luo C, Castanon R, et al. Iterative single-cell multi-omic integration using online learning. Nat Biotechnol 2021;39:1000–7.
[223] Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics 2016;32:1–8.
[224] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res 2018;46:10546–62.
[225] Macaulay IC, Ponting CP, Voet T. Single-cell multiomics: multiple measurements from single cells. Trends Genet 2017;33:155–68.
[226] Guerin LN, Barnett KR, Hodges E. Dual detection of chromatin accessibility and DNA methylation using ATAC-Me. Nat Protoc 2021;16:5377–97.
[227] Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol 2020;21:111.
[228] Jin S, Zhang L, Nie Q. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. Genome Biol 2020;21:25.
[229] Uzun Y, Yu W, Chen C, Tan K. SINBAD: a flexible tool for single cell DNA methylation data. bioRxiv 2021;465577.
[230] Charlton J, Downing TL, Smith ZD, Gu H, Clement K, Pop R, et al. Global delay in nascent strand DNA methylation. Nat Struct Mol Biol 2018;25:327–32.
[231] Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science 2017;356:eaaj2239.
[232] Zhu F, Farnung L, Kaasinen E, Sahu B, Yin Y, Wei B, et al. The interaction landscape between transcription factors and the nucleosome. Nature 2018;562:76–81.
[233] Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol 2020;21:31.
[234] Mulqueen RM, Pokholok D, O'Connell BL, Thornton CA, Zhang F, O'Roak BJ, et al. High-content single-cell combinatorial indexing. Nat Biotechnol 2021;39:1574–80.
[235] Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 2018;36:421–7.
[236] Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, et al. A single-cell atlas of *in vivo* mammalian chromatin accessibility. Cell 2018;174:1309–24.e18.
[237] Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature 2019;566:496–502.
[238] Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, et al. A human cell atlas of fetal gene expression. Science 2020;370:eaba7721.
[239] Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, et al. A human cell atlas of fetal chromatin accessibility. Science 2020;370:eaba7612.
[240] Zeng P, Lin Z. coupleCoC+: an information-theoretic co-clustering-based transfer learning framework for the integrative analysis of single-cell genomic data. PLoS Comput Biol 2021;17:e1009064.
[241] Zappia L, Phipson B, Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput Biol 2018;14:e1006245.
[242] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15.