nature biotechnology

Article

https://doi.org/10.1038/s41587-024-02148-9

Simultaneous single-cell analysis of 5mC and 5hmC with SIMPLE-seq

Received: 24 August 2022

Accepted: 18 January 2024

Published online: 09 February 2024

Check for updates

Dongsheng Bai[1], Xiaoting Zhang[1], Huifen Xiang[2,3], Zijian Guo[4], Chenxu Zhu[5,6] ✉ & Chengqi Yi[1,7,8] ✉

Dynamic 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) modifications to DNA regulate gene expression in a cell-type-specific manner and are associated with various biological processes, but the two modalities have not yet been measured simultaneously from the same genome at the single-cell level. Here we present SIMPLE-seq, a scalable, base resolution method for joint analysis of 5mC and 5hmC from thousands of single cells. Based on orthogonal labeling and recording of 'C-to-T' mutational signals from 5mC and 5hmC sites, SIMPLE-seq detects these two modifications from the same molecules in single cells and enables unbiased DNA methylation dynamics analysis of heterogeneous biological samples. We applied this method to mouse embryonic stem cells, human peripheral blood mononuclear cells and mouse brain to give joint epigenome maps at single-cell and single-molecule resolution. Integrated analysis of these two cytosine modifications reveals distinct epigenetic patterns associated with divergent regulatory programs in different cell types as well as cell states.

Dynamic chemical modifications to the genome, including histone tails and DNA bases, regulate gene expression by facilitating or inhibiting the binding of transcriptional factors to them[1,2]. DNA 5-methylcytosine (5mC) is a major epigenetic modification that can be deposited or eliminated in a cell-type-specific manner, driven by DNA methyltransferases and replication-dependent or replication-independent demethylation processes[3]. The replication-independent active DNA demethylation is mediated by Ten-Eleven Translocation (TET) family proteins, generating 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) sequentially, followed by base excision repair of the latter two forms to reinstall unmodified cytosines[4]. Both 5mC and 5hmC are shown to regulate diverse biological processes, including stem cell pluripotency, development, aging and tumorigenesis, in mammals[5–8].

Various molecular assays, including DNase-seq/ATAC-seq[9,10], ChIP-seq[11] and Bisulfite-seq[12], were developed to identify the location and states of regulatory elements in many cell types and species[13]. Recent methods developed in single-cell genomics have revolutionized the study of gene regulatory networks by accessing gene expression[14–16], chromatin high-order organizations[17], chromatin accessibilities[18–21], histone modifications[22–29] and DNA base modifications[30–34], one modality at a time or together[35–40], at single-cell resolution from complex biological systems. In particular, pioneering single-cell methylome analyses have revealed the heterogeneity and cell-type-specific patterns of 5mC in different cellular environments[30,34,41–44]. However, the bisulfite treatment-dependent methods did not resolve 5hmC from 5mC[31], obscuring the functional interpretation of their individual roles from the mixed outputs. Using the glucosylated 5hmC (5ghmC)-dependent restriction endonuclease AbaSI, methods were recently developed to map 5hmC in single cells[31,45]. Combining chemical and enzymatic approaches enabled discrimination between 5hmC and 5mC at

[1]State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing, China. [2]Department of Obstetrics and Gynecology, First Affiliated Hospital of Anhui Medical University, Anhui, China. [3]NHC Key Laboratory of Study on Abnormal Gametes and Reproductive Tract, Anhui Medical University, Anhui, China. [4]State Key Laboratory of Coordination Chemistry, Coordination Chemistry Institute, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, China. [5]New York Genome Center, New York, NY, USA. [6]Department of Physiology and Biophysics, Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. [7]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China. [8]Department of Chemical Biology and Synthetic and Functional Biomolecules Center, College of Chemistry and Molecular Engineering, Peking University, Beijing, China. ✉e-mail: czhu@nygenome.org; chengqi.yi@pku.edu.cn

single-base resolution, starting to provide insights into the distribution patterns and biological functions of 5hmC[46–53] as well as hinting the potential significance of its relationships with other epigenetic layers, particularly 5mC. However, measuring these two modifications one at a time can capture only the averaged outcome from different cell populations while potential heterogeneous relationships, such as the DNA methylation equilibrium contributed by generation of both 5mC and 5hmC in dynamic cellular systems, may be lost. In line with this notion, the 'six-letter seq' method was developed to probe 5mC and 5hmC simultaneously from cell-free DNA or genome DNA molecules isolated from bulk cell populations[54]. However, dissecting the relationships and combined functional effects of 5mC and 5hmC in different cell types from complex systems, such as development and diseases, requires measuring them at single-cell resolution.

A major challenge for single-cell joint profiling of 5mC and 5hmC is to orthogonally record the two modalities from the same DNA molecule. To address this, the DNA template should be minimally disrupted; hence, recently developed bisulfite-free chemical labeling approaches—for instance, TET-assisted pyridine borane sequencing (TAPS) and chemical-assisted C-to-T conversion of 5hmC sequencing (hmC-CATCH)—that are mild and specific for 5mC or 5hmC provide a potential opportunity[47,55] (Extended Data Fig. 1a). However, the two approaches in their current form do not allow orthogonal recording of the two modifications and are, thus, not directly compatible with each other: blocking the 5hmC modality in TAPS[43], which ensures the specificity of 5mC detection, will incapacitate 5hmC for further labeling and detection.

Here we present SIMPLE-seq (simultaneous profiling of epigenetic cytosine modifications by sequencing) for joint analysis of 5mC and 5hmC at single-cell and single-base resolution. Based on a combination of bisulfite-free chemical labeling reactions[47,55], SIMPLE-seq introduces the 'C-to-T' mutation signals for 5mC and 5hmC sequentially and detects the locations and types of the modifications from the same DNA molecule in single cells. We applied SIMPLE-seq to mouse embryonic stem cells (mESCs), human peripheral blood mononuclear cells (PBMCs) and mouse brain samples. Integrated analysis of the joint 5mC and 5hmC maps at single-cell and single-molecule levels revealed divergent epigenetic programs for different cell states and regulatory elements.

Results

Overview of SIMPLE-seq

We proposed a strategy to sequentially label and record the two modalities from the same DNA molecule (Fig. 1a). Considering that the average abundance of 5hmC is approximately 10% or less of 5mC[56] and the specificity of the two labeling approaches, it is optimal to perform 5hmC labeling before the 5mC reaction for high specificity. We also introduced a primer extension step right after 5hmC labeling to record its signals. After 5mC labeling reactions, both 5hmC-derived and 5mC-derived signals can be amplified from the same reaction mixture for sequencing. We first performed a proof-of-concept experiment on a synthesized model oligonucleotide containing single 5mCG and 5hmCG sites (T5MH; Supplementary Table 1 and Extended Data Fig. 1b,c) by Sanger sequencing. After ruthenate (VI) oxidation of 5hmC to 5fC[47,57] and indanedione labeling of the newly generated 5fC[32,58], a 'C-to-T' signal is specifically generated at the 5hmC site after polymerase chain reaction (PCR) amplification but not at C or 5mC sites (Extended Data Fig. 1b). Both the labeling reactions are mild, without notable degradation (Extended Data Fig. 1d,e). The endogenous 5fC (0.24–1.52% of 5hmC) derived from 5hmC oxidation by TET enzymes[59] will also be labeled; but, considering its sparsity, we do not expect 5fC to significantly interfere with 5hmC analysis. Primer extension was introduced to record the '5hmC-to-T' on the newly synthesized complementary strand to the original template. Next, TET-mediated oxidation was carried out to convert 5mC on the original template to 5caC, followed by borane reduction to DHU[55], to give the second 'C-to-T' signal for the 5mC site in the same molecule; the unmodified cytosine and other bases remained unchanged (Extended Data Fig. 1b). In addition, we observed low background conversion rates from high-throughput sequencing results of spike-in lambda DNA sequencing library (0.06% (from 0.04% to 0.08%, at 95% interval) for 5hmC treatment-derived reads and 0.57% (from 0.49% to 0.65%, at 95% interval) for 5mC treatment-derived reads)—measured from 'C-to-T' mutation of all unmodified Cs on the model DNA sequences, T1M (Extended Data Fig. 1f–h). These results suggest that the two labeling strategies can be integrated with our sequential recording approach for specific 5hmC and 5mC detection.

Another major challenge is to decode the two modalities jointly from thousands of single cells. Because both TAPS[43] and hmC-CATCH[43] read out cytosine modifications via 'C-to-T' mutations, their signals must be deduced from the recording products of the same DNA templates. Although resolving the two signals in one single cell could be relatively straightforward, obtaining such information from thousands of single cells is very challenging but is required for understanding methylation dynamics in complex biological systems. We designed a primer pre-deposited with a 5caC base to record 5hmC signals in the extension products; the 5caC base will be converted to a 'T' signal during the subsequent 5mC reaction, and this signal will be used to distinguish amplification products derived from the two modalities (Fig. 1a). Therefore, in the SIMPLE-seq procedure, cells were first fixed with formaldehyde and permeabilized, followed by nucleosome depletion via weak SDS treatment to allow unbiased capture of the genome[43] (Fig. 1a and Extended Data Fig. 1i,j). To reduce the potential barcode collision, a brief sonication was carried out to prevent nuclei clumping (Extended Data Fig. 1i). Next, Tn5 tagmentation reaction was carried out to attach an adaptor to genome DNA, followed by nuclei barcoding with a ligation-based combinatorial indexing strategy[60]. To increase genomic coverage, we benchmarked multiple enzyme concentrations and buffer conditions (Extended Data Fig. 1k). The nuclei were then lysed, and genomic DNA was purified, followed by sequential chemical labeling (Fig. 1a). After sequencing, the cellular barcodes ligated to DNA fragments were used to assign each read to individual cells, combined with the 5caC signal pre-deposited on the primer to deduce 5mC and 5hmC at single-base resolution for each cell (Fig. 1a and Methods). We also minimized the purification and transferring steps during the whole procedure to reduce the risk of material loss (Methods).

Joint analysis of 5mC and 5hmC from single cells

In the SIMPLE-seq procedure, the first round of nuclei barcoding by tagmentation was used as a sample multiplexing barcode (to label cells from different conditions) for a single experiment (Fig. 1a and Methods). The C-to-T mutation rates for 5mC and 5hmC are estimated as 86.9% and 85.6% from spiked-in model double-stranded DNA (dsDNA) containing 5mC or 5hmC sites, suggesting a reasonably high sensitivity for identifying modified cytosines. Notably, both modifications from the same molecule can be efficiently detected (T1M, T2H and T3MH; Supplementary Table 1 and Extended Data Fig. 1h). To normalize the detected base modification levels, we performed oligonucleotide mixing experiments by mixing modified and unmodified DNA oligos at different ratios to establish the standard curves to offset the observed C-to-T conversion efficiencies (T4C, T4M and T4H; Supplementary Table 1 and Extended Data Fig. 1l–n). In addition, SIMPLE-seq detects 5mC and 5hmC based on the positive mutational signals instead of negatively selecting the unconverted Cs as in bisulfite-based methods, showing high selectivity for both 5mC and 5hmC (Extended Data Fig. 1b,h,o). The species-mixing experiment suggested an expected barcode collision rate of approximately 5.1%, and the fraction of reads mapped to human and mouse genomes are highly concordant between 5mC and 5hmC modalities, suggesting successful recording of both 5mC and 5hmC status for the same single cells (Fig. 1b and Extended Data Fig. 1p,q).

https://doi.org/10.1038/s41587-024-02148-9

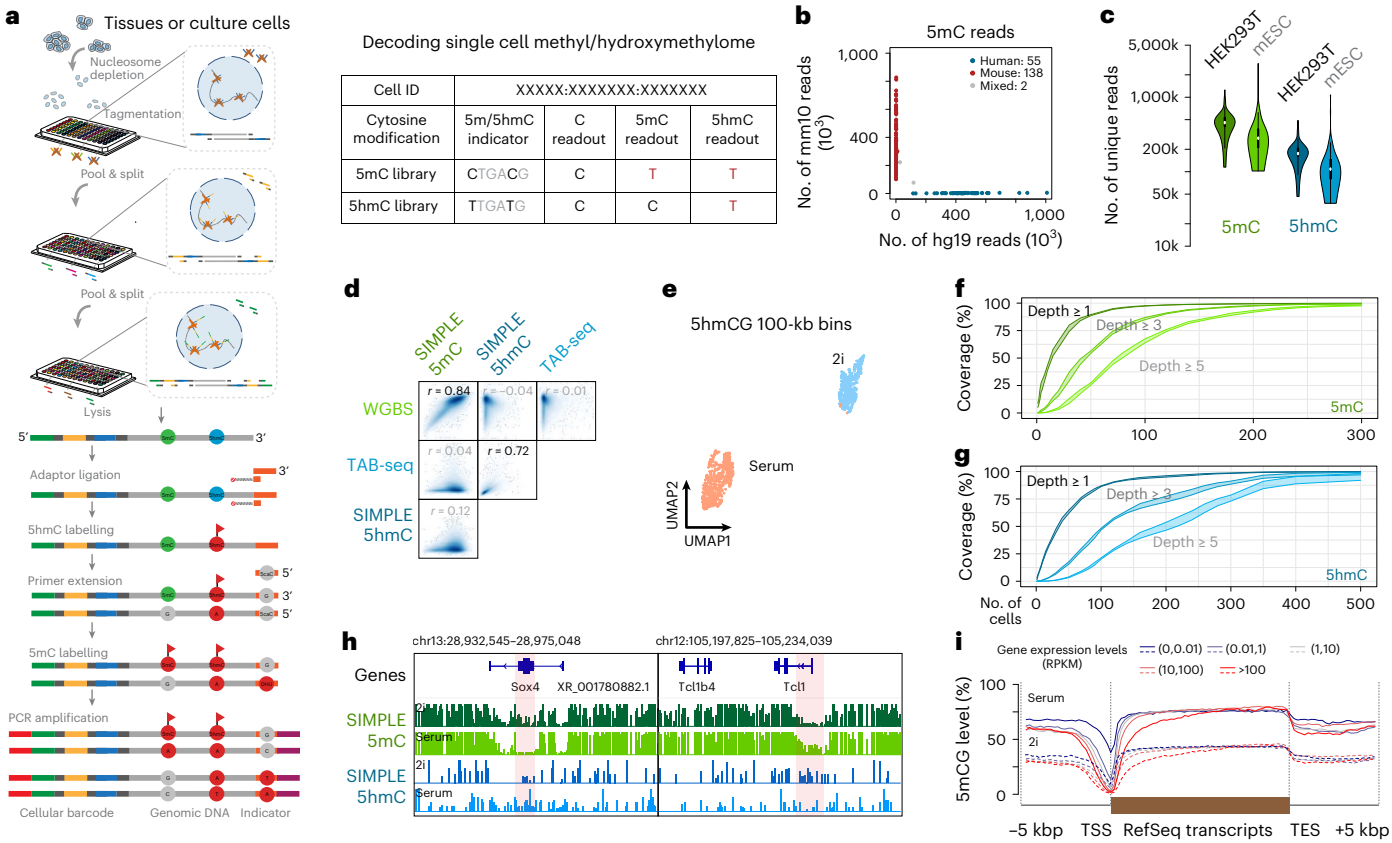


Fig. 1 | Base resolution, joint analysis of 5mC and 5hmC in single cells. **a**, Schematics of SIMPLE-seq. Tn5 tagmentation reaction is carried out on crosslinked and nucleosome-depleted nuclei, followed by nuclei barcoding via combinatorial indexing. The nuclei were then lysed, and the barcoded genomic DNA was extracted, followed by sequential chemical labeling and recording of labeled products. The 5mC and 5hmC sites in single cells can be identified at base resolution from the 'C-to-T' signal after sequencing. **b**, Scatter plot showing the number of reads mapped to human and mouse genome in each cell from the species-mixing experiment. **c**, Violin plots showing the number of unique reads per cell assigned to 5mC and 5hmC in human HEK293T and mESCs. Cell number $n = 55$ (HEK293) and $n = 138$ (mESC). Data are presented as 485,854 ± 25,334 (HEK293T, 5mC), 339,706 ± 24,745 (mESC, 5mC), 188,233 ± 9,699 (HEK293T, 5hmC) and 128,736 ± 9,250 (mESC, 5hmC) (mean ± s.e.m.). **d**, Scatter plots showing the genome-wide 5mCG and 5hmCG modification levels (in 25-kb non-overlapping bins) between 5mC and 5hmC profiles generated by different assays (5mC: WGBS and SIMPLE-seq; 5hmC: TAB-seq and SIMPLE-seq). Pearson's correlations between different datasets are indicated in the boxes. **e**, UMAP embedding showing cells based on their 5hmCG levels (in 100-kb non-overlapping bins). Each dot represents a single cell and is colored according to its original identity. **f**,**g**, Line plots showing the cumulated CG site coverages of different depths for 5mC (**f**) and 5hmC (**g**) from different numbers of single cells. The shadowed area shows the error ranges from five randomly sampled cell sets. **h**, Genome browser showing the CG modification levels at *Sox4* (serum mESC highly expressed) and *Tcl1* (2i mESC highly expressed) loci. The differentially (hydroxyl)methylated regions for the two genes are indicated with pink boxes. **i**, Line plots showing the 5mCG levels around genic regions of genes with different expression levels. RPKM, reads per kilobase per million mapped reads.

We carried out the SIMPLE-seq experiment on mouse embryonic stem cells (mESCs) cultured in two different states—2i-cultured mESCs and serum-cultured mESCs—and sequenced approximately 1,500 mESCs (~6.7% of the nuclei recovered from one SIMPLE-seq experiment) to moderate depth (average PCR duplicate rate: 19.6%) and recovered 1,095 cells after filtering out cells with low sequencing coverage. SIMPLE-seq datasets showed high mapping rates: an average of 93.0% of reads in each cell can be mapped to the reference mouse genome (GRCm38.p6); among them, 94.6% can be assigned to 5mC (68.4%) and 5hmC (26.2%) labeling-derived reads, respectively (Extended Data Fig. 1r,s). For the 5mC modality, we recovered a median number of 313,261 unique mapped reads per cell with an average genomic coverage of 1.96% (from 0.97% to 3.56%, at 95% interval), corresponding to 1.78-fold of coverage with only approximately 1/5 of sequencing depth compared to the previous sci-MET method[43] (average sequenced reads per cell: 438,591 for SIMPLE-seq and 2,485,858 for sci-MET; Fig. 1c, Extended Data Fig. 1t and Supplementary Table 2). For the 5hmC modality, we recovered a median number of 150,372 unique mapped reads per cell, corresponding to 0.79% average genomic coverage for single cells (from 0.38% to 1.43%, at 95% interval; Fig. 1c, Extended Data Fig. 1u and Supplementary Table 2). We calculated the molecular complexities of SIMPLE-seq libraries and predicted that the single-cell genomic coverage of 5% could be obtained by sequencing to 1.5 million reads per-cell depth (10.2% if sequenced to saturation)[61]. Compared to scBS-seq[41], scRRBS[30] and scAba-seq[31], SIMPLE-seq recovers similar or higher numbers of CG sites per cell at the same sequencing depth (Extended Data Fig. 1v,w). Aggregated signal of SIMPLE-seq datasets shows reasonably well agreement on the genome-wide modification levels of 5mC and 5hmC with published whole-genome bisulfite sequencing (WGBS) and Tet-assisted bisulfite sequencing (TAB-seq) datasets from the same cell line (Fig. 1d and Extended Data Fig. 2a–d)[46,55]. In total, we identified 19,968,022 5mCG sites from 184.9 million reads and 1,057,466 5hmC sites from 68.3 million reads of the 2i-cultured mESCs, and, among them, 622,323 were shared (Extended Data Fig. 2e,f). Compared to the 5mC–5hmC shared sites, the 5hmC-only sites are more enriched in regulatory elements, including active and poised enhancers and active promoters associated with the H3K4me3 histone mark (Extended Data Fig. 2g,h).

We then summarized the 5mC and 5hmC modification levels in 100 kilobase (kb) size non-overlapping bins, followed by dimensional

reduction with principal component analysis (PCA) and visualization by uniform manifold approximation and projection (UMAP), which can clearly separate mESCs of the two different states (Fig. 1e, Extended Data Fig. 2i–m and Methods). SIMPLE-seq can analyze thousands of single cells in a single experiment, and high-coverage datasets can be recovered by aggregating signals from single cells in the same cluster: at the single CG site level, 80% of genomic coverage (no fewer than three reads) can be obtained from approximastely 100 cells (5mC) and approximately 250 cells (5hmC) (Fig. 1f,g); at 10-kb non-overlapping bins level, more than 95% of the mappable genomic regions can be recovered (from no fewer than three reads) by aggregating 20 and 40 cells for 5mC and 5hmC, respectively (Extended Data Fig. 2n). Consistent with previous observations[62], our aggregated signal for the same cell groups showed a higher level of 5mC in serum-cultured mESCs compared to mESCs in 2i, with a negative correlation between gene expressions and 5mCG levels at the transcriptional start site (TSS) proximal regions and weakly positive correlations between gene expression and TSS-proximal 5hmCG levels in genes with higher expression levels (Fig. 1i and Extended Data Fig. 2o–q). These results support SIMPLE-seq as a scalable method that can simultaneously generate high-quality 5mC and 5hmC profiles from single cells.

DNA methylation and active demethylation dynamics in different mESC states

Extensive genome-wide de novo cytosine methylation and active DNA demethylation occur during the conversion of 2i state mESCs to cells maintained in serum conditions. To investigate the (hydroxyl)methylome dynamics during the 2i-to-serum mESC transition, we constructed the pseudotime trajectory based on the 5mCG status of single cells (Fig. 2a). It has been suggested that 5hmC offers a different platform upon which transcription factors may bind or 5mC-specific binding proteins may be excluded[63,64]. Thus, understanding the regulatory functions of these two modifications requires joint analysis of their levels on binding motifs of potential transcriptional factors (TFs). Based on high-quality 5mC and 5hmC profiles from single cells, ChromVAR[65] motif analysis results revealed both concordant and disconcordant changes of enrichment levels of TF binding motifs on 5mC and 5hmC sites across the trajectory (Fig. 2b and Extended Data Fig. 3a). We also analyzed the expression level of TFs, assuming that the abundant TFs are likely functional. We classified TFs into different groups, depending on the relationship between TF gene expression levels and motif enrichment levels on 5mCG–5hmCG sites: group 1 and group 2 of TFs (27 and 23 out of 93) show concordant 5mC–5hmC dynamics and display negative and positive relationships, respectively, whereas group 3 and group 4 of TFs (28 and 15 out of 93) show disconcordant 5mC–5hmC dynamics and exhibit negative and positive relationships, respectively (Fig. 2b and Extended Data Fig. 3a). For example, in group 1, genomic regions containing the Myc binding motifs showed both lower 5mCG and 5hmCG levels in cells cultured with serum condition compared to those with 2i condition, and *c-Myc* has increased expression in serum-cultured mESCs[66]. Another example of group 1 is *Hes1*, which expresses higher in 2i condition compared to serum condition and is known to delay embryonic stem differentiation to neural cells[67]. For disconcordant motifs, representative TFs from group 3 include: motifs recognized by lineage specifier Sox3 (neurogenesis)[68] show high 5hmCG and low 5mCG levels in mESCs of serum conditions; on the other side, high 5mCG and low 5hmCG level regions in serum conditions are enriched for motifs of self-renewal factor Klf4 (ref. 69) (Fig. 2b). Thus, we observed similar numbers of TFs from both concordant and disconcordant motif groups displaying positive or negative relationships with TF gene expression levels, suggesting distinct and complex impacts of 5mC and 5hmC on TF binding and the regulatory output.

Interestingly, we found that mESCs of these two states can be more distinctly separated from each other based on 5hmC than on 5mC (Fig. 1e and Extended Data Fig. 2i–k,m); a previous study also showed that the 5hmC levels increased to reach the plateau faster during the state transition as well as showed a clearer difference between the two states[70]. To explore why less homogeneous 5mCG states were observed during the mESC state transition, we grouped the cells into five groups according to the pseudotime (split into five equal portions according to pseudotime score ranks) and compared the 5mCG level changes from 'early' to 'late' pseudotime points by aggregating the single-cell signals. We found that, for the third group ('midpoint' when the cells connected the two cell populations, corresponding to the cell population with less similarity in 5mCG state to other serum-cultured cells after removal of 2i), the promoter of proliferation genes, such as *Plk1* and *Ctnnb1*, showed increased methylation levels, and promoter of cell cycle repressor genes, such as *Cdkn2a* and *Cdkn2c*, showed a decreased methylation level, compared to the other cell groups (Fig. 2c and Extended Data Fig. 3b); thus, the snapshot of such an intermediate cell state could represent a possible delayed cell cycle or senescence state during mESC state transition.

The stark heterogeneity differences of 5mCG and 5hmCG across cells prompted us to analyze the 5mCG–5hmCG relationships in single cells. To quantify the 5mC–5hmC relationship at the cell level, we calculated the cytosine modification entropy (or modification entropy) for each single cell from the percentages of its 5hmCGs with different neighboring 5mCG and 5hmCG numbers (surrounded by only 5hmCGs, by only 5mCGs, by both 5mCGs and 5hmCGs and without surrounding 5mCG/5hmCG). Cells with complex 5mCG–5hmCG distribution relationships tend to have high modification entropy values and vice versa (Fig. 2d and Methods). Interestingly, we found that the cells resembling the intermediate state between 2i and serum mESC populations showed higher modification entropy values compared to other cells (Fig. 2e and Extended Data Fig. 3c). The increased 5hmCG–5mCG complexity could be explained by ongoing active DNA demethylation and re-methylation that did not yet reach the static states, and, thus, the modification entropy could be used to identify the transient reprogramming events by identifying the 'intermediate' cells that may be missed from single-modality analyses.

A type of 5mC-associated 5hmC site is correlated with active chromatin state

5hmC sites are generated by oxidation of 5mC by TET enzymes[71], which could be either intermediate products of active DNA demethylation or stable epigenetic modifications with regulatory function. Previous knowledge suggests that TET-mediated demethylation exhibits different degrees of processivities: TET enzymes may oxidize all 5mC sites within a range of the genome or selectively target individual loci[4]. However, whether 5hmC generated from these two different modes is present during 2i-to-serum transition and, if yes, may have different roles remained unclear. To this end, we performed the CG site-level analysis of 5mC–5hmC relationships. We used the sequenced genomic fragments with both 5mC and 5hmC modality captured. Under moderate sequencing depth, an average of 12.3% of 5hmC reads in each cell have the paired 5mC modality of the same loci captured, which recovered 17.9% of all detected 5hmCG sites (190,276 sites covered by at least five cells) (Extended Data Fig. 3d,e). Among them, we identified 27.6% of 5hmCG sites that co-exist with 5mCG in the same regions (type 2, 5mCG-associated 5hmCG sites); for the rest of 5hmCG sites, we did not detect such 5mC–5hmC relationship in single cells (type 1, basal level) (Fig. 2f, Supplementary Table 3 and Methods). Consistent with the previous observation[62], 5hmCG levels (of both types 1 and 2) are generally higher in serum-cultured mESCs than in 2i-cultured cells. Interestingly, the modification levels of type 2 5hmCG sites first increased but then dropped during the 2i-to-serum transition, with the highest levels in the intermediate cells connecting the two distinct populations. Such distinct patterns suggest potential differential roles of two 5hmCG types associated with the two modes of TET processivities (Fig. 2f and Extended Data Fig. 3f).

https://doi.org/10.1038/s41587-024-02148-9

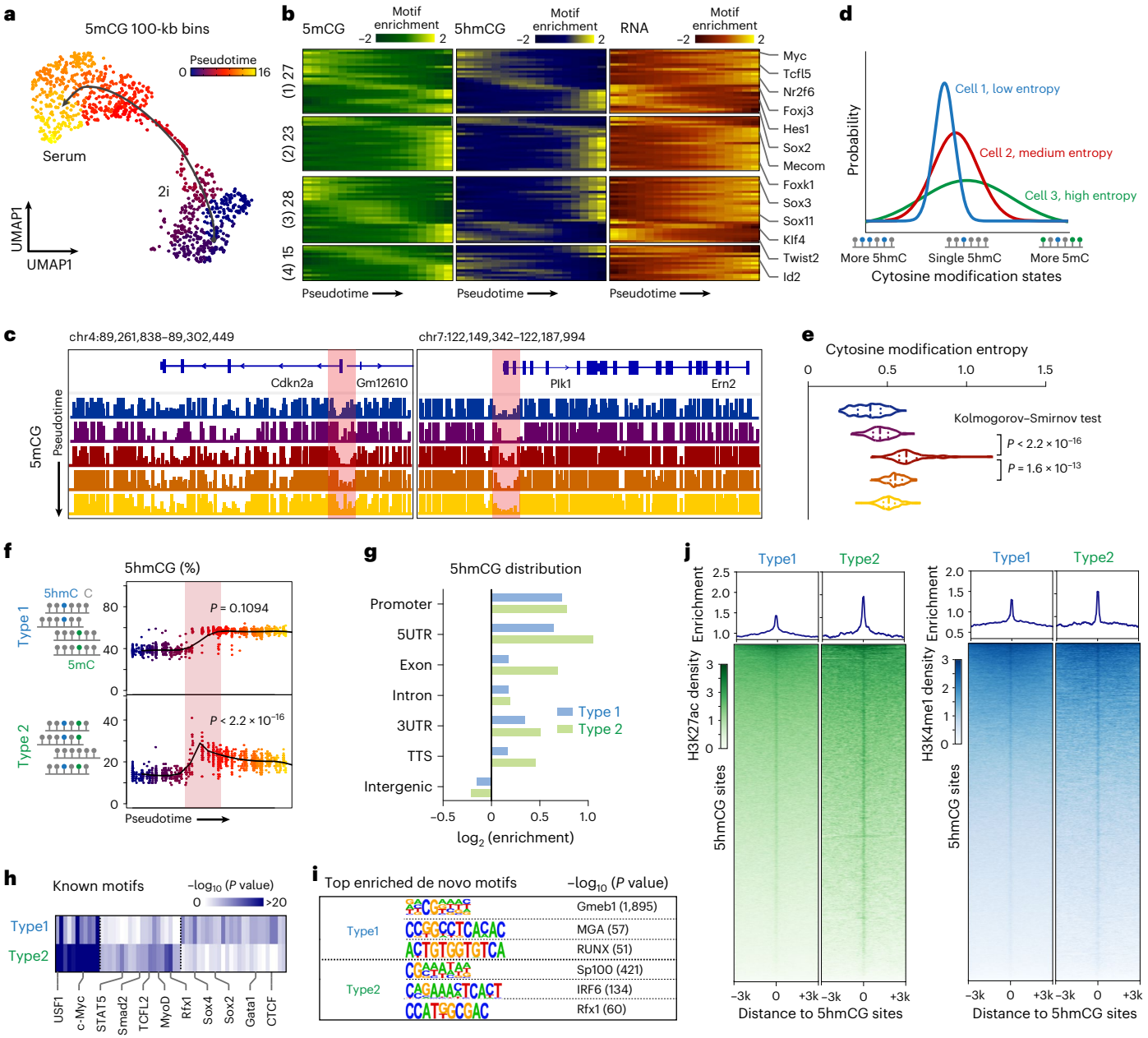


Fig. 2 | Analysis of 5mC and 5hmC from the same molecules revealed multiple 5hmC types associated with active chromatin. a, UMAP showing the single-cell trajectory of 2i-cultured mESCs to serum-cultured mESC transition. Each dot represents a single cell and is colored according to its assigned pseudotime score. **b**, Heat maps showing the 5mCG and 5hmCG TF motif enrichments and TF gene expression during mESC 2i to serum state transition. **c**, Genome browser view showing the 5mCG levels at promoters of proliferation gene *Plk1* and cell cycle repressor gene *Cdkn2a* from cells of different pseudotime groups. **d**, The distribution of 5hmCG sites with different 5hmCG–5mCG relationships can be quantified by cytosine modification entropy, and the cells with more unified 5hmCG states have lower modification entropy and vice versa. **e**, Violin plots showing the 5hmCG state entropies of cells in different pseudotime groups. *P* value, Kolmogorov–Smirnov test. **f**, Schematics showing the classification of 5hmCG sites: 5hmCG sites with neighboring 5mCG sites in a significant fraction of cells are grouped as type 2 sites. Average modification levels of two types of 5hmCG sites in single cells. Each dot represents a single cell and is colored and ordered according to its pseudotime score. *P* value, two-sided Wilcoxon test of modification level differences between intermediate cells and the rest of the cells. **g**, Bar plots showing the relative enrichment of two types of 5hmCG sites in different genomic regions. **h**, Heat map showing the known motif enrichment for 5hmCG types. **i**, Top enriched de novo motifs for 5hmCG types. *P* value, one-sided Fisher's exact test. **j**, Histograms and heat maps showing the relationship between two types of 5hmCG with mESC H3K27ac and H3K4me1 ChIP-seq signals from ENCODE (ENCSR000CGQ and ENCSR000CGN).

To further understand the two 5hmCG types, we compared the genomic distributions of 5mCG-associated 5hmCG sites (type 2) with the rest of the sites (type 1): both types of 5hmCG sites are enriched in TSS-proximal and genic regions, in which type 2 sites are particularly enriched for exons and transcriptional termination sites (TTSs) (Fig. 2g). Genomic Regions Enrichment of Annotations Tool (GREAT)[72] analysis shows that type 2 5hmCG sites are enriched for developmental processes of diverse lineages, suggesting that active demethylation of these 5hmC sites may be associated with mESC priming for differentiation (Extended Data Fig. 3g). We then performed TF binding motif enrichment analysis for 5hmCG sites, and both shared and differential motifs were identified for the two 5hmCG types (Fig. 2h,i). For example, the binding motif of pluripotency factor c-Myc is enriched for both type 1 and type 2 5hmCG, whereas motifs for Sox2 and CTCF are more

enriched in basal-level 5hmC. Among them, a previously identified 5mC reader[73], *Rfx1*, was identified from both enriched for known motifs and de novo motifs analysis of type 2 5hmCG sites (Fig. 2h,i). In addition, compared to basal-level type 1 5hmCG, the type 2 sites are more enriched for active regulatory elements marked by H3K27ac (Fig. 2j and Extended Data Fig. 3h–j). The different patterns of the two 5hmCG types suggested the potential distinct regulatory functions of the two TET processivity modes. These results demonstrated the utility of SIMPLE-seq in revealing different 5hmCG types associated with distinct regulatory functions during 2i state to serum state transition of mESCs.

SIMPLE-seq recovered major cell types from PBMCs

To test the ability of SIMPLE-seq in resolving cell-type-specific (hydroxyl)methylome and methylome from in vivo systems, we applied SIMPLE-seq to snap-frozen PBMC samples. We sequenced approximately 3,000 PBMCs with an average number of approximately 450,000 sequenced reads per cell and recovered 2,110 cells after filtering low-coverage cells: more than 95% of reads can be mapped to the reference human genome (GRCh38), and, among them, 64.5% and 30.9% of reads were assigned as 5mC and 5hmC reads, respectively (Supplementary Table 4). After removing the 32.1% PCR-duplicated reads of the 5mC modality, we obtained the median number of 136,835 unique mapped reads per cell with an average genomic coverage of 0.91% (from 0.13% to 2.73%, at 95% interval). For 5hmC, we recovered a median number of 150,372 unique mapped reads per cell at PCR duplicate rates of 25.6%—corresponding to 0.48% average genomic coverage for each single cell (from 0.07% to 1.43%, at 95% interval; Supplementary Table 4).

We then performed a similar dimensional reduction on PBMC datasets based on 5mC and 5hmC in CG, and CHG contents and single cells formed distinct groups on the UMAP plots (Fig. 3a,b and Extended Data Fig. 4a–g). Louvain clustering[74] on 5mCG levels grouped the cells into five major clusters, which can be annotated as $CD4^+$ and $CD8^+$ T cells, B cells, natural killer (NK) cells and monocytes based on their promoter 5mCG modification levels of the representative marker genes (Fig. 3c and Extended Data Fig. 4a,h). For example, *TCF7* expression marks self-renewal $CD4^+$ T cells[75], and its promoter showed a hypomethylation pattern specifically in $CD4^+$ T cells; the promoter region of B cell marker *CD19* (ref. 76) is also hypomethylated in the corresponding track of aggregated single-cell signals (Fig. 3c). Interestingly, we found, for dimensional reduction and visualization with 5hmC, that the cell groups showed less distinct separation, but the identities of single cells were in good concordance with the cell types identified from 5mCG levels (Fig. 3b and Extended Data Fig. 4b–d). This is different from the observations in mESCs where our focus is the dynamic epigenetic changes of the two different states, in which the 5hmC sites mainly represent regions where epigenetic reprogramming occurs. For differentiated cell types (such as PBMC) where cell state changes are less common, the cell identities were defined mainly by methylome, and, thus, 5hmC has lower resolution in revolving the static differences. By aggregating the 5hmC signal from single cells in groups defined by 5mC-based clustering, we generated these cell-type-specific, paired 5mC and 5hmC profiles from heterogeneous and low-5hmC-level cellular populations.

Next, we calculated the differential modified regions (DMRs) for both 5mCG and 5hmCG pair-wisely for the immune cell types (see Methods for details) and identified from 416 to 1,360 5mCG DMRs and from 476 to 4,946 5hmCG DMRs (Fig. 3d–g, Extended Data Fig. 4i–m and Supplementary Tables 5 and 6). Motif enrichment analysis of 5mCG DMRs revealed regulators for these immune cell types. For example, STAT3 is an important determinant of whether the naive T cell differentiates into regulatory or inflammatory T cell lineage[77], whose binding motifs are hypomethylated in $CD4^+$ T cells and hypermethylated in monocytes (Extended Data Fig. 4k). The transcriptional factor Bach2 is another key regulator that controls the formation and functions of multiple T cell lineages[78], whose binding motifs are highly hydroxymethylated in $CD4^+$ T cells (Fig. 3e). GREAT analysis revealed that both 5mC and 5hmC DMRs are involved in immune-related processes, such as T cell co-stimulation and interferon-gamma-mediated signaling pathway (Fig. 3g and Extended Data Fig. 4m). These results suggest that SIMPLE-seq can generate matched cell-type-specific 5mC and 5hmC maps from heterogeneous PBMC samples.

Cell-type-specific analysis of cytosine modification states

To further study the relationship between 5mC and 5hmC at the immune cell types, we adopted the ChromHMM algorithm[79], which is designed to integrate multiple chromatin datasets to discover de novo the major re-occurring combinations of 5mC and 5hmC modification patterns. We grouped the genome into 10 cytosine states, including 5mCG-marked regions (E3 and E4), 5hmCG-marked regions (E5), regions marked by both 5mCG and 5hmCG (E1 and E2), 5mCHG/5mCHH-enriched regions (E6), regions enriched for 5hmCHG (E7) or 5hmCHH (E8), regions with both 5mC and 5hmC in CHG/CHH contents (E9) and hypomethylated regions (E10) (Fig. 3h and Extended Data Fig. 5a). Among them, E3 and E5 (modifications in CG sites) are depleted from TSSs, and E5 and E8 (5hmC in CG and CHH sites) are depleted from transcriptional end sites (TESs) (Extended Data Fig. 5b). To associate the different cytosine states with functional genomics regions, we overlapped the cytosine states with the ENCODE candidate *cis*-regulatory elements from the same representative cell type (Fig. 3i). E3 state showed the strongest association with transcription (intragenic regions), agreeing with the positive correlation of abundant gene body methylation with gene expression[80]. E5 state is enriched in active enhancers, and E7 and E8 states are associated with weak enhancers, consistent with the association of 5hmC with enhancers[81]. Interestingly, we also found that E5 is highly enriched for heterochromatin, which is similar to mESCs in that a significant fraction of 5hmC sites previously identified with TAB-seq[46] is enriched for the H3K9me3-associated heterochromatin regions ($\log_2$(observed/expected) = 2.68).

Next, we compared the conserved and differential cytosine states across different immune cell types. E2–4 and E9 states are the most stable groups among the immune cell types, and E1 and E5 are likely to be changed from one state to another between different cell types (Fig. 3j). We then grouped the regions of different states into conserved and differential regions, according to their differences between cell types, and analyzed their status of multiple histone modifications (Fig. 3k and Extended Data Fig. 5c–e). Interestingly, we found that, for the E3 state, the differential compartment shows a higher fraction of regions that are marked by H3K4me1 ($P = 2.4 \times 10^{-5}$) and H3K27me3 ($P = 3.8 \times 10^{-4}$), corresponding to distal regulatory elements, whereas the average of 60.4% of conserved E3 regions is marked by H3K36me3, corresponding to gene bodies of transcribing genes (Fig. 3k). For the 5hmCG-enriched E5 state, the differential groups have a higher fraction of regions marked by repressive H3K27me3 mark ($P = 0.005$) while less marked by H3K4me1 ($P = 0.004$); similarly, the differential group of E7 state (5hmCHG) also has a higher fraction of regions marked by H3K27me3, suggesting that the differential 5hmCG and 5hmCHG across different cell types are likely to exist in bivalent or repressed regulatory elements (Extended Data Fig. 5c,d). For E9 states, we observed a decreased or increased fraction of region associated with H3K9me3 or H3K36me3, respectively (Extended Data Fig. 5e). We did not observe a significant difference in their relationships with histone marks for the rest of the cytosine states. GREAT analysis revealed that the differential groups of E3, E5 and E7 states (more likely to be associated with distal regulatory elements) are enriched for biological processes of specific cell types—for example, alpha-beta T cell differentiation for $CD8^+$ T cells (E3) and phagocytosis/engulfment for monocytes (E7) (Fig. 3l and Extended Data Fig. 5f). These results suggest that integrated analysis of 5mC and 5hmC at single-base resolution from the same cells can help elucidate their regulatory roles in complex heterogeneous cellular populations.

https://doi.org/10.1038/s41587-024-02148-9

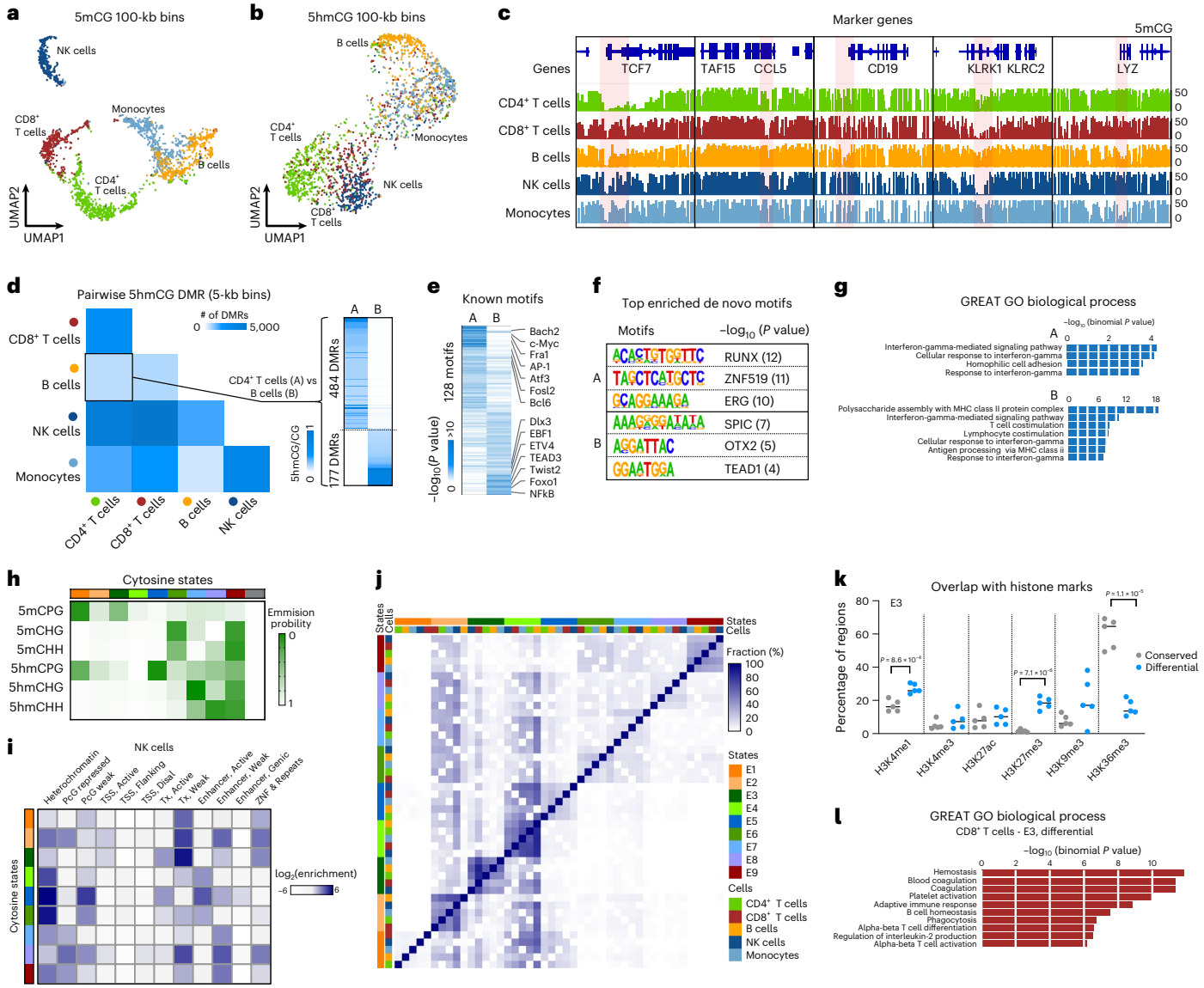


Fig. 3 | Cell-type-specific 5mC and 5hmC landscapes in human PBMCs. **a**,**b**, UMAP embedding showing the single-cell clustering based on 5mCG (**a**) and 5hmCG (**b**) levels (in 100-kb non-overlapping bins) from PBMCs. Each dot represents a single cell and is colored according to its annotation based on 5mCG. **c**, Genome browser view showing the 5mCG levels in several representative marker gene loci. The differentially methylated TSS-proximal regions are shadowed in pink. **d**, The heat map left-sided showing the numbers of pairwise differential 5hmCG regions across cell types (in 5-kb non-overlapping bins). The heat map right-sided showing the hydroxymethylation levels of 5hmCG DMRs of a representative group (CD4$^+$ T cells and B cells). **e**–**g**, The enrichment analysis of known motifs (**e**), top enriched de novo motifs (**f**) and top enriched GREAT GO terms (**g**). *P* value, one-sided Fisher's exact test. **g**, For 5hmCG, DMRs of this representative group are also shown. *P* value, one-sided Fisher's exact test. **h**, Heat map showing the emission probability of 5mC and 5hmC in CG, CHG and CHH content in different cytosine states. **i**, Heat map showing the relative enrichment of different cytosine states overlapped with ENCODE chromatin states in NK cells. **j**, Heat map showing the fraction of genome regions overlapped between cell types and cytosine states. Cell types and cytosine states are indicated with colored boxes. **k**, Scatter plot showing the fraction of genome regions overlapped with peaks of different histone marks in conserved and differential E3 state regions. *P* value, one-sided Fisher's exact test. **l**, Top enriched GREAT GO terms for differential E3 state regions in CD8$^+$ T cells. *P* value, one-sided Fisher's exact test. MHC, major histocompatibility complex.

Single-cell 5mC and 5hmC landscapes in the mouse brain

Previous studies revealed that 5hmC is highly abundant in the brain tissues[71], although analyzing the cell-type-specific 5hmC maps from this heterogeneous tissue type is challenging due to few available high-throughput single-cell 5hmC sequencing methods. We applied SIMPLE-seq to snap-frozen mouse cerebral cortex tissue collected from 8-week-old male mice. We sequenced approximately 6,000 cells to an average of 341,464 (combined 5mC and 5hmC) sequenced reads per cell and recovered 4,767 cells after filtering low-quality cells (79.5% recovery rate) (Supplementary Table 7). SIMPLE-seq recovers 63% (5mC) and 67% (5hmC) more unique reads per cell than the recent Joint-snhmC-seq method if sequenced to the same depth (Extended Data Fig. 6a,b)[82]. Next, we performed the joint clustering with the 'weighted nearest neighbors' approach using both 5mC and 5hmC modality (average modification levels in 100-kb non-overlapping bins across the genome) from the same cells and revealed 11 major brain cell types, including two excitatory neuron cell types (hypomethylated on *Snap25*, *Neurod6* and *Slc17a7* promoters), one inhibitory neuron cell type (*Snap25* and *Gad1/2*), two astrocyte cell groups (*Apoe* and *Gfap*), three oligodendrocyte cell groups (*Mbp* and *Mobp*), oligodendrocytes (*Pdgfra*), microglia cells (*Csf1r*) and endothelial cells (*Flt1*) (Fig. 4a–c). On the contrary, cell clustering based on 5mCG alone could resolve the three neuron cell

https://doi.org/10.1038/s41587-024-02148-9

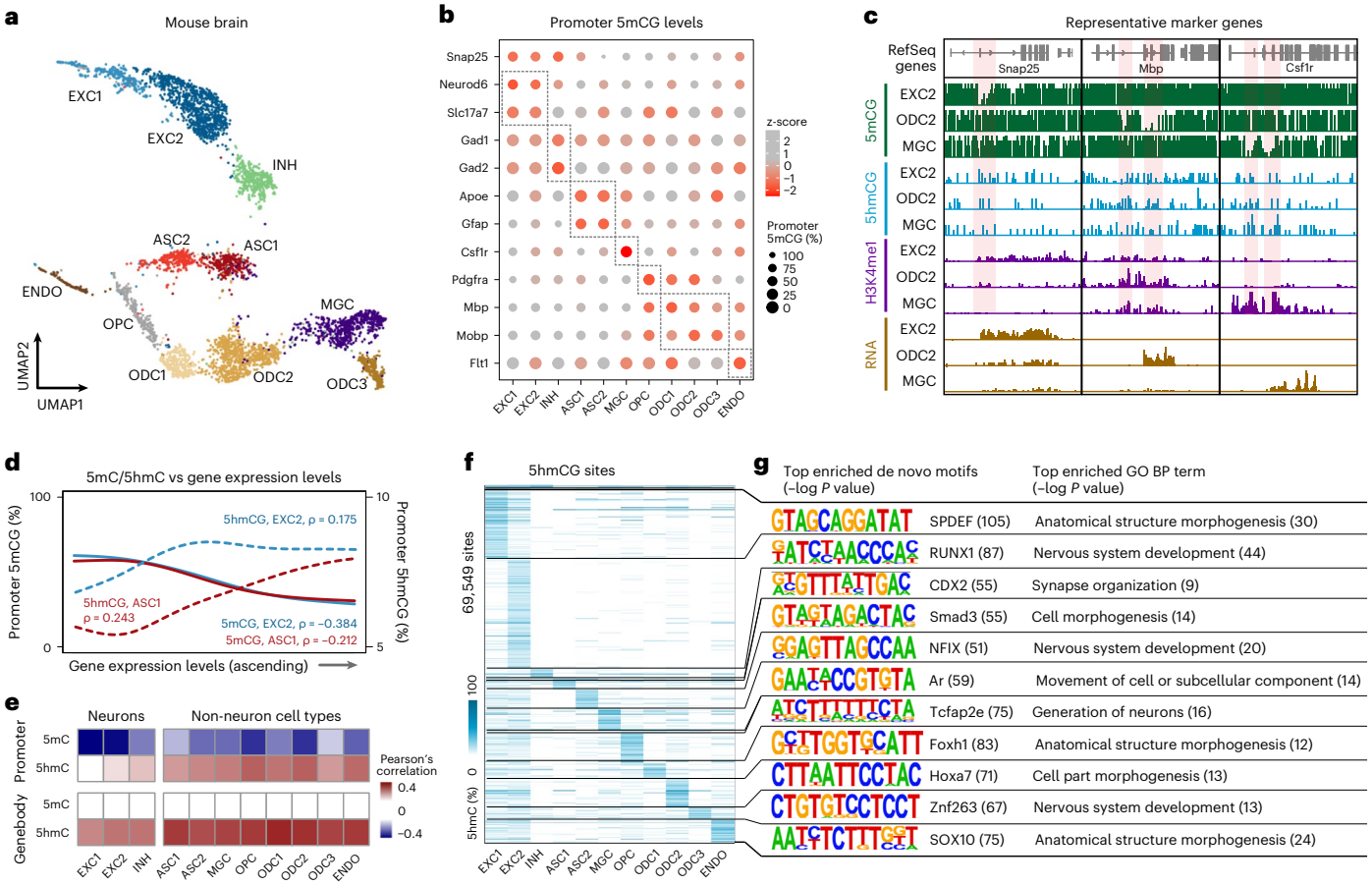


Fig. 4 | Single-cell joint analysis of 5mC and 5hmC from the mouse brain. **a**, UMAP embedding showing the single-cell clustering based on 5mC and 5hmC profiles from the mouse brain. Each dot represents a single cell and is colored according to its annotation based on promoter methylation levels of marker genes. **b**, Dot plots showing the promoter 5mCG levels (−1,500 bp to +500 bp) of representative marker genes in the detected cell types. **c**, Genome browser showing the CG modification levels at *Snap25* (EXC2 highly expressed), *Mbp* (ODC2 highly expressed) and *Csf1r* (MGC highly expressed) loci. The differentially (hydroxyl)methylated regions for the two genes are indicated with pink boxes. **d**, Line plots showing the correlations between promoter cytosine modification levels and gene expression levels in EXC2 and ASC1 cell types. **e**, Heat map showing the Pearson's correlation coefficients between promoter and gene body 5mC–5hmCG levels with gene expression in different cell types. **f**, Heat map showing the 5hmCG modification levels of cell-type-specific 5hmCG across the 11 cell types. **g**. Top enriched de novo motifs (left) and top enriched GO terms (right) for each cell type are also shown. *P* value, one-sided Fisher's exact test. BP, Biological Process.

types while lacking the resolution for the two astrocyte subgroups; similarly, 5hmCG-based clustering does not resolve endothelial cells well (Extended Data Fig. 6c–e). These results suggest that the bi-modal profiles from SIMPLE-seq can be used to identify the major brain cell types and generate the cell-type-specific 5mC and 5hmC profiles without the need for external reference maps.

Next, we aggregated the signal from single cells in the same group as defined by joint clustering and found that 5mCG is enriched for repressive histone mark H3K27me3, whereas 5hmCG is enriched for histone mark H3K4me1 (ref. 83) (Extended Data Fig. 6f,g). We then analyzed the relationships between 5mCG and 5hmCG modification levels with gene expression (Fig. 4d,e and Extended Data Fig. 6h). Promoter 5mCG levels showed negative correlations with gene expression levels in both neuron and non-neuron cell types, whereas promoter 5hmCG levels showed weak positive correlations with gene expression in neuron cells but higher correlations in non-neuron cell types (Fig. 4e). This is different from mESCs where a large fraction (58.8% of 5hmCG sites overlapped with 5mCG) of 5hmCG are intermediates of active DNA methylation, and the promoter 5hmCG is weakly correlated with gene expression levels (Extended Data Fig. 2p). In addition, 5hmCG levels on the gene body regions are also positively correlated with gene expression, whereas 5mCG levels are weakly correlated (Fig. 4e). Next, we performed differential 5mCG and 5hmCG site analysis and identified 264,918 and 69,549 sites for the 11 cell groups ($\log_2$(fold change) > 1; Methods) (Fig. 4f, Extended Data Fig. 6i and Supplementary Table 8). We found that the two excitatory neuron cell types have the highest abundance of 5hmCG sites (19.5% and 30.5% of total 5hmCG sites) and 5mCG sites (16.8% and 24.7% of total 5mCG sites). Motif and Gene Ontology (GO) term enrichment analysis revealed that these sites are possibly involved in gene regulation during neuron development processes. For example, the binding motif for RUNX1 is enriched for EXC2 5hmCG sites, and this gene is known to be involved in the development of cholinergic and motor neurons[84]. The astrocyte-enriched motif is predicted to be recognized by NFIX, which is a regulator driving astrocytic maturation[85] (Fig. 4g and Extended Data Fig. 6j). By generating the cell-type-specific 5mC and 5hmC maps for mouse brain cell types, our data suggest that 5mC and 5hmC may have distinct regulatory roles in neuron and non-neuron cell types of mouse brain tissue.

Discussion

We report a high-throughput method, SIMPLE-seq, for joint analysis of 5mC and 5hmC at single-base and single-cell resolution. During the revision of this paper, several other exciting methods for single-cell joint analysis of 5mC and 5hmC were published[82,86,87]. In comparison,

scDARESOME and scDyad-seq are based on restriction enzyme cutting and, thus, detect only one site per DNA fragment and could not analyze multiple neighboring 5mC and 5hmC sites in short ranges[86,87]. Compared to Joint-snhmC-seq, SIMPLE-seq converts only modified cytosines and provides higher fractions of mappable reads (>90% versus ~60%) (Extended Data Fig. 6a)[82]. Without the requirement to split genomic materials into two halves for 5mC and 5hmC library preparation, SIMPLE-seq could detect the two modifications from the same molecules with higher genomic coverages (Extended Data Fig. 6b). Unlike these methods that label individual cells in tubes or plates, SIMPLE-seq analyzes 10^4–10^5 and is preferred for heterogeneous samples with larger cell numbers. Here we demonstrated SIMPLE-seq by applying it to cultured mESCs, heterogeneous human primary PBMC samples and mouse brain tissues, giving genome-wide 5mC and 5hmC maps at cell type resolutions for these heterogeneous cellular populations with different 5hmC dynamics and modification levels. By jointly analyzing the two epigenetic modalities at single-cell and single-molecule levels, we found that the DNA methylation dynamics vary across cells of different states; in addition, we identified different groups of regulatory elements with distinct epigenetic patterns.

By evaluating DNA methylation and active demethylation dynamics in single cells, we showed that, even for the seemly homogeneous mESC populations, distinct distributions of 5mC–5hmC relationships exist in different cells: the cells with more disordered 5mC–5hmC relationships tend to reprogram and proliferate slower compared to the other cells with more regulated epigenome programs. Such epigenetic dynamics analysis provides an approach to identify potential 'intermediate' or 'declined' cell populations for the discovery of candidate transient epigenetic reprogramming events that may be associated with disease progression or aging from complex tissues. By identifying 5hmCG sites that are selectively generated from methylated regions, we found that 5hmCG sites flanked by 5mCG sites are associated with higher permissive activities. Interestingly, these 5hmC sites also showed the highest modification levels during state transition of mESCs, suggesting a possible role of TET-dependent de-repression of master regulators during mESCs' epigenetic reprogramming. In addition, whether a similar mechanism exists and can be used for other differentiation systems—for example, identifying master regulators for efficient generation of pancreatic cells for the treatment of diabetes[88]—requires further investigations. By integrated analysis of TF gene expression with motif enrichments on 5mC and 5hmC sites, we revealed that TF binding is likely to be repressed by hypermethylation, whereas TF binding motifs on 5hmC-modified regions are positively correlated with TF gene expression. However, further biochemical and cellular analyses are desired to further demonstrate the biological functions of these DNA modifications on regulatory elements[73]. We also showed that 5mC and 5hmC have varying abilities in resolving cell types for different cellular environments. For dynamic systems, such as state transition of mESCs, 5hmC is more able to separate different cell populations; for relatively static differentiated systems, such as immune cell types in PBMCs, more distinct signatures were observed for 5mC but not for 5hmC; whereas, for tissues with abundant 5hmC levels (such as the mouse brain), both 5mC and 5hmC have similar resolution in resolving cell types. Thus, SIMPLE-seq is a versatile tool to identify cell types and states for multiple systems by simultaneously obtaining information on both of the two epigenetic modifications.

The current SIMPLE-seq protocol relies on plate-based combinatorial barcoding[19,60,83] while retaining the compatibility with droplet-based combinatorial indexing platforms[89,90] for fast and ultra-high-throughput single-cell epigenomics profiling. We provided an example of how the convertible index (5caC pre-deposited in 5hmC recording primers) could be used for decoding multiple readouts from the same mixture without the need for physical separation; this strategy could also be integrated with approaches for single-molecule multiplexed detection of other modalities, such as different types of RNA modifications. In the present study, we focused only on 5mC and 5hmC, two major cytosine epigenomic marks; with additional modification, 5mC and 5hmC should be able to be further jointly monitored with other molecular layers, including transcriptome, three-dimensional genome structure, chromatin states and protein abundances[36,83,91–97] for larger-scale single-cell multimodal integration[98]. The base calling of 5mC and 5hmC is dependent on C-to-T mutational signals, and existing base mutation and unblocked 5fC could result in false-positive detection (0.24–1.52% according to 5fC:5hmC ratios)[59]; on the other hand, the incomplete converted signal (currently ~87%) may become a source of false-negative detection and could be further optimized in the future to improve sensitivity. Taking advantage of the throughput of SIMPLE-seq, we addressed such 'dropouts' by dimension reduction and imputation from closely related cells[99]; future optimizations, including increasing labeling efficiency[100], pre-blocking 5fC bases and reducing amplification and sequencing errors[54,101] will further improve the detection. Compared to scRNA-seq and scATAC-seq, SIMPLE-seq showed lower resolution in identifying minor cell types due to the high percentage of missing values; future development in combining SIMPLE-seq with scRNA-seq will help to address this by identifying a cell type's identity from the transcriptome-based analysis and reconstructing 5mC and 5hmC landscapes at cell type resolution. Compared to enzymatic 5hmC detection approaches[31,45], SIMPLE-seq allowed the detection of multiple 5hmC and 5mC sites from the same fragments. Besides single-cell analysis, the non-destructive sequential labeling strategy described here is also amenable for the measurement of low-input, highly fragmented materials, such as cell-free DNA (cfDNA). Recent focuses on measuring 5mC[102] and 5hmC[103] from cfDNA highlighted their potential in identifying types and stages of tumors, and measuring these two modifications from the same molecules jointly is expected to further improve the dissection of their comprehensive relationships and distinct signatures in different cancer types.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-024-02148-9.

References

1. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: recording the past and predicting the future. *Science* **358**, 69–75 (2017).
2. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* **17**, 551–565 (2016).
3. Bhutani, N., Burns, D. M. & Blau, H. M. DNA demethylation dynamics. *Cell* **146**, 866–872 (2011).
4. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat. Rev. Genet.* **18**, 517–534 (2017).
5. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: in the right place at the right time. *Science* **361**, 1336–1340 (2018).
6. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
7. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
8. Parry, A., Rulands, S. & Reik, W. Active turnover of DNA methylation during cell fate decisions. *Nat. Rev. Genet.* **22**, 59–66 (2021).
9. Crawford, G. E. et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16**, 123–131 (2006).

10. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
11. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
12. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
13. Consortium, E. P. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
14. Tang, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
15. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
16. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
17. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
18. Jin, W. et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142–146 (2015).
19. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
20. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
21. Preissl, S. et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).
22. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).
23. Harada, A. et al. A chromatin integration labelling method enables epigenomic profiling with lower input. *Nat. Cell Biol.* **21**, 287–296 (2019).
24. Kaya-Okur, H. S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
25. Carter, B. et al. Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (ACT-seq). *Nat. Commun.* **10**, 3747 (2019).
26. Ku, W. L. et al. Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. *Nat. Methods* **16**, 323–325 (2019).
27. Wang, Q. et al. CoBATCH for high-throughput single-cell epigenomic profiling. *Mol. Cell* **76**, 206–216 (2019).
28. Ai, S. et al. Profiling chromatin states using single-cell itChIP-seq. *Nat. Cell Biol.* **21**, 1164–1172 (2019).
29. Grosselin, K. et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060–1066 (2019).
30. Guo, H. et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**, 2126–2135 (2013).
31. Mooijman, D., Dey, S. S., Boisset, J. C., Crosetto, N. & van Oudenaarden, A. Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nat. Biotechnol.* **34**, 852–856 (2016).
32. Zhu, C. et al. Single-cell 5-formylcytosine landscapes of mammalian early embryos and ESCs at single-base resolution. *Cell Stem Cell* **20**, 720–731 (2017).
33. Wu, X., Inoue, A., Suzuki, T. & Zhang, Y. Simultaneous mapping of active DNA demethylation and sister chromatid exchange in single cells. *Genes Dev.* **31**, 511–523 (2017).
34. Luo, C. et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
35. Lake, B. B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
36. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
37. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
38. Zhu, C. et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).
39. Luo, C. et al. Single nucleus multi-omics identifies human cortical cell regulatory genome diversity. *Cell Genom.* **2**, 100107 (2022).
40. Xie, Y. et al. Droplet-based single-cell joint profiling of histone modifications and transcriptomes. *Nat. Struct. Mol. Biol.* **30**, 1428–1433 (2023).
41. Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
42. Farlik, M. et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.* **10**, 1386–1397 (2015).
43. Mulqueen, R. M. et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
44. Nichols, R. V. et al. High-throughput robust single-cell DNA methylation profiling with sciMETv2. *Nat. Commun.* **13**, 7627 (2022).
45. Wangsanuwat, C., Chialastri, A., Aldeguer, J. F., Rivron, N. C. & Dey, S. S. A probabilistic framework for cellular lineage reconstruction using integrated single-cell 5-hydroxymethylcytosine and genomic DNA sequencing. *Cell Rep. Methods* **1**, 100060 (2021).
46. Yu, M. et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
47. Zeng, H. et al. Bisulfite-free, nanoscale analysis of 5-hydroxymethylcytosine at single base resolution. *J. Am. Chem. Soc.* **140**, 13190–13194 (2018).
48. Schutsky, E. K. et al. Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4204 (2018).
49. Sun, Z. et al. High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells. *Cell Rep* **3**, 567–576 (2013).
50. Liu, Y. et al. Subtraction-free and bisulfite-free specific sequencing of 5-methylcytosine and its oxidized derivatives at base resolution. *Nat. Commun.* **12**, 618 (2021).
51. Cohen-Karni, D. et al. The MspJI family of modification-dependent restriction endonucleases for epigenetic studies. *Proc. Natl Acad. Sci. USA* **108**, 11040–11045 (2011).
52. Vaisvila, R. et al. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res* **31**, 1280–1289 (2021).
53. Sen, M. et al. Strand-specific single-cell methylomics reveals distinct modes of DNA demethylation dynamics during early mammalian development. *Nat. Commun.* **12**, 1286 (2021).
54. Fullgrabe, J. et al. Simultaneous sequencing of genetic and epigenetic bases in DNA. *Nat. Biotechnol.* **41**, 1457–1464 (2023).

55. Liu, Y. B. et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.* **37**, 424–429 (2019).
56. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
57. Booth, M. J. et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
58. Xia, B. et al. Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat. Methods* **12**, 1047–1050 (2015).
59. Ito, S. et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
60. Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
61. Daley, T. & Smith, A. D. Predicting the molecular complexity of sequencing libraries. *Nat. Methods* **10**, 325–327 (2013).
62. Sim, Y. J. et al. 2i maintains a naive ground state in ESCs through two distinct epigenetic mechanisms. *Stem Cell Rep.* **8**, 1312–1328 (2017).
63. Hashimoto, H. et al. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res.* **40**, 4841–4849 (2012).
64. Yildirim, O. et al. Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* **147**, 1498–1510 (2011).
65. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
66. Marks, H. et al. The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**, 590–604 (2012).
67. Kobayashi, T. et al. The cyclic gene *Hes1* contributes to diverse differentiation responses of embryonic stem cells. *Genes Dev.* **23**, 1870–1875 (2009).
68. Bylund, M., Andersson, E., Novitch, B. G. & Muhr, J. Vertebrate neurogenesis is counteracted by Sox1–3 activity. *Nat. Neurosci.* **6**, 1162–1168 (2003).
69. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
70. Habibi, E. et al. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13**, 360–369 (2013).
71. Tahiliani, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
72. McLean, C. Y. et al. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
73. Spruijt, C. G. et al. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146–1159 (2013).
74. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* P10008 (2008).
75. Nish, S. A. et al. $CD4^+$ T cell effector commitment coupled to self-renewal by asymmetric cell divisions. *J. Exp. Med.* **214**, 39–47 (2017).
76. Wang, K., Wei, G. & Liu, D. CD19: a biomarker for B cell development, lymphoma diagnosis and therapy. *Exp. Hematol. Oncol.* **1**, 36 (2012).
77. Egwuagu, C. E. STAT3 in $CD4^+$ T helper cell differentiation and inflammatory diseases. *Cytokine* **47**, 149–156 (2009).
78. Tsukumo, S. et al. Bach2 maintains T cells in a naive state by suppressing effector memory-related genes. *Proc. Natl Acad. Sci. USA* **110**, 10735–10740 (2013).
79. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
80. Hellman, A. & Chess, A. Gene body-specific methylation on the active X chromosome. *Science* **315**, 1141–1143 (2007).
81. Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. & Jacobsen, S. E. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.* **12**, R54 (2011).
82. Fabyanic, E. B. et al. Joint single-cell profiling resolves 5mC and 5hmC and reveals their distinct gene regulatory effects. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-022-01652-0 (2023).
83. Zhu, C. et al. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat. Methods* **18**, 283–292 (2021).
84. Theriault, F. M., Roy, P. & Stifani, S. AML1/Runx1 is important for the development of hindbrain cholinergic branchiovisceral motor neurons and selected cranial sensory neurons. *Proc. Natl Acad. Sci. USA* **101**, 10343–10348 (2004).
85. Matuzelski, E. et al. Transcriptional regulation of *Nfix* by NFIB drives astrocytic maturation within the developing spinal cord. *Dev. Biol.* **432**, 286–297 (2017).
86. Viswanathan, R. et al. DARESOME enables concurrent profiling of multiple DNA modifications with restriction enzymes in single cells and cell-free DNA. *Sci. Adv.* **9**, eadi0197 (2023).
87. Chialastri, A., Sarkar, S., Schauer, E. E., Lamba, S. & Dey, S. S. Combinatorial quantification of 5mC and 5hmC at individual CpG dyads and the transcriptome in single cells reveals modulators of DNA methylation maintenance fidelity. Preprint at *bioRxiv* https://doi.org/10.1101/2023.05.06.539708 (2023).
88. Shahjalal, H. M., Abdal Dayem, A., Lim, K. M., Jeon, T. I. & Cho, S. G. Generation of pancreatic β cells for treatment of diabetes: advances and challenges. *Stem Cell Res. Ther.* **9**, 355 (2018).
89. Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
90. Datlinger, P. et al. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods* **18**, 635–642 (2021).
91. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
92. Li, G. et al. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat. Methods* **16**, 991–993 (2019).
93. Lee, D. S. et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat. Methods* **16**, 999–1006 (2019).
94. Mimitou, E. P. et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* **39**, 1246–1258 (2021).
95. Chung, H. et al. Joint single-cell measurements of nuclear proteins and RNA in vivo. *Nat. Methods* **18**, 1204–1212 (2021).
96. Zhang, B. et al. Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *Nat. Biotechnol.* **40**, 1220–1230 (2022).
97. Chen, A. F. et al. NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat. Methods* **19**, 547–553 (2022).
98. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
99. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 1169 (2020).
100. Xu, H. et al. Modular oxidation of cytosine modifications and their application in direct and quantitative sequencing of

5-hydroxymethylcytosine. *J. Am. Chem. Soc.* **145**, 7095–7100 (2023).
101. Schmitt, M. W. et al. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **109**, 14508–14513 (2012).
102. Chan, K. C. et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc. Natl Acad. Sci. USA* **110**, 18761–18768 (2013).
103. Song, C. X. et al. 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Res.* **27**, 1231–1242 (2017).

Methods

Cell culture and processing

2i-cultured version 6.5 mESCs were regularly maintained on the pre-gelatinized dishes without feeders, supplemented with two inhibitors (2i)—1 mM PD0325901 (Selleck Chemicals, S1036) and 3 mM CHIR99021 (Selleck Chemicals, S1263)—in the presence of 1,000 U ml^{-1} leukemia inhibitory factor (LIF, Millipore, ESG1107) and 20% FBS (Gibco, 26140087) in DMEM/F-12 (Gibco, 11320033). mESCs (serum) were cultured with DMEM/F12 with 20% FBS and 1,000 U ml^{-1} LIF. Human PBMCs were purchased form AllCells (PB003F-C). Cultured cells or PBMCs were harvested and washed with fresh no-serum media (DMEM/F-12, Gibco) and then re-suspended in 10 ml of no-serum media before proceeding to SIMPLE-seq experiments. Dry ice snap-frozen mouse cerebral cortex tissue dissected from 8-week-old male mice (C57BL/6J) was ordered from The Jackson Laboratory. Single-cell suspensions were prepared from douncing of the frozen tissue with Douncing Buffer (0.25 M sucrose (Sigma-Aldrich, S7903), 25 mM KCl (Sigma-Aldrich, P9333), 5 mM $MgCl_2$, 10 mM Tris-HCl pH 7.4, 1 mM DTT (Sigma-Aldrich, D9779), 1× protease inhibitor cocktail (Sigma-Aldrich, 11873580001), 0.5 U μl^{-1} RNaseOUT (Invitrogen, 10777019), 0.5 U μl^{-1} SUPERaseIN (Invitrogen, AM2696) and 0.1% Triton X-100 (Sigma-Aldrich, T9284)) and followed by the SIMPLE-seq procedure.

Oligonucleotide and model DNA synthesis

Oligonucleotides containing site-specific 5mC, 5hmC and 5caC (Supplementary Table 1) were synthesized on the Expedite 8909 nucleic acid synthesizer using commercially available phosphoramidites (Glen Research, 10-1510-02, 10-1564-02 and 10-1066-02). Regular oligonucleotides were purchased from Sangon Biotech. Long duplex 5mC DNAs (T1M; Supplementary Table 1) were constructed from lambda DNA, and in vitro methylation reaction was performed using S-adenosylmethionine (New England Biolabs (NEB), B9003S) and M.SssI methyltransferase (NEB, M0226S). Methylation of CpGs on DNA were validated by Sanger sequencing. Long duplex 5hmC DNAs (T2H; Supplementary Table 1) were prepared through ligation of short duplex fragments (20–40 base pairs (bp)) with sticky overhangs. Long duplex 5mC and 5hmC DNAs (T3MH; Supplementary Table 1) for next-generation sequencing were produced by the annealing and extension method.

Oxidation and malononitrile-based selective labeling of 5hmC

To achieve selective 5hmC oxidation, we used potassium ruthenate (K_2RuO_4), a ruthenium (Ru^{6+}) oxidant. Although $KRuO_4$ caused severe DNA degradation, K_2RuO_4-mediated oxidation is very mild but also complete (Extended Data Fig. 1d,e). Then, 10× potassium perruthenate (Sigma-Aldrich, 334537) solution was prepared according to the published protocol[43]. In brief, 0.15 mmol potassium perruthenate was added to 0.5 M NaOH solution (1 ml; Alfa Aesar, A18395) and vortexed to make sure all solid was dissolved, and the solution was incubated at 25 °C for 2 d to produce potassium ruthenate solution (K_2RuO_4). Before 5hmC oxidation, genomic DNA was purified with 1× AMPure XP beads (Beckman Coulter, B23319), followed by additional purification with Micro Bio-Spin P-6 SSC column (Bio-Rad, 732–6200). The DNA was then denatured in 0.05 M NaOH for 30 min at 37 °C, followed by chilling on an ice water bath for 5 min. Next, 1.5 µl of 1× oxidant was added to the denatured DNA sample (28.5 µl) and briefly mixed with tapping, followed by incubation on ice for 1 h. The oxidized DNA was purified with a Micro Bio-Spin P-6 SSC column. To label the newly generated 5fC, the reaction was carried out in 10 mM pH 7.0 Tris buffer and 150 mM malononitrile (J&K, 261700) in a total volume of 35 µl at 37 °C for 20 h and 850 r.p.m. in a thermomixer.

Oxidation and pic-borane-based selective reduction of 5mC

First, 293F cells were transfected with pCDNA3-Flag-mTET1CD plasmid using PEI (Polysciences, 23966-1). After 48 h, cells were collected and processed with the published mTET1CD purification protocol[44]. Cells were re-suspended in the lysis buffer (50 mM Tris buffer pH 7.5 (Invitrogen, 15567027), 500 mM NaCl (Invitrogen, AM9759), 1× cOmplete protease inhibitor cocktail (Sigma-Aldrich, 11873580001), 1 mM PMSF (Thermo Fisher Scientific, 36978) and 1% Triton X-100 (Sigma-Aldrich, T8787)) and incubated on ice for 20 min. Cell lysate was then spun down for 30 min at 30,000*g* and 4 °C. The supernatant was then purified with Anti-flag Affinity Gel Beads 4FF (Smart-Lifescience), and purified protein was eluted with elution buffer (20 mM HEPES pH 8.0 (Sigma-Aldrich, H3375), 150 mM NaCl, 0.1 mg ml^{-1} 3× Flag peptide (Sigma-Aldrich, F4799F4799), 1× cOmplete protease inhibitor cocktail and 1 mM PMSF). Eluted protein solution was concentrated in storage buffer (20 mM HEPES pH 8.0, 150 mM NaCl and 1 mM dithiothreitol, 30% v/v glycerol) and stored at −80 °C. mTET1CD oxidation reaction was prepared as follows: DNA up to 100 ng, 50 mM HEPES pH 8.0, 100 mM NaCl, 1 mM α-ketoglutaric acid (Sigma-Aldrich, K3752-5G), 2 mM L-ascorbic acid (Sigma-Aldrich, 95210-50G), 1.2 mM ATP (Sigma-Aldrich, A6419), 2.5 mM DTT (Fluorochem, M02712), 100 µM Fe^{2+} (Sigma-Aldrich, 09719) and 8 µM mTET1CD. After that, the oxidation reaction was carried out at 37 °C for 80 min. Then, 0.8 U of Qiagen Protease (Qiagen, 19157) was added to the oxidation reaction and incubated for 1 h at 50 °C. The reaction mixture was purified using 1.8× AMPure XP beads. Then, 2-picoline-borane (pic-borane, Sigma-Aldrich, 654213-5G) was dissolved in DMSO to give ~3.26 M solution. Next, 2.5 µl of 3 M sodium acetate solution, pH 5.2 (Sigma-Aldrich, R1181), and 12.5 µl of 3.26 M pic-borane solution were added to 10 µl of DNA sample and incubated at 70 °C in a thermocycler for 4 h. The reaction mixture was purified with Zymo Oligo & Clean Concentrator Kit (Zymo Research, D4060) or Bio-Spin P-30 Gel Column (Bio-Rad, 732-6223). The purified product was subjected to library amplification using high-fidelity uracil-tolerated DNA polymerase.

Potential degradation estimates of labeling reactions

Lambda DNA was treated with 5hmC-specific chemistry, and lambda DNA of equal amount was used as a control. Both control and treated lambda DNA were processed with the same procedures, including purification steps, except the chemical treatment. A 222-bp-specific region on lambda DNA was chosen as representative to perform qPCR to estimate the integrity of lambda DNA (Supplementary Table 1). Fragmentated lambda DNA (>400 bp) was treated with 5hmC-specific chemistry and 5mC-specific chemistry sequentially, and gel analysis was performed using the same amount of sample recovered and a control sample without treatment.

Sanger sequencing of the labeled model DNA

Single-strand synthesized DNA that contains both 5mC and 5hmC (T5MH; Supplementary Table 1) was treated with 5hmC-specific chemistry and 5mC-specific chemistry sequentially, and, after each round of labeling, PCR amplification was performed using 2× Kapa U^+ HiFi Master Mix (Kapa Biosystems, 0795900520001). PCR products were purified with DNA Clean & Concentrator Kit (Vistech) for Sanger sequencing.

Barcoded Tn5 assembly

To generate barcoded Tn5 transposase, barcoded DNA oligos (33 nucleotides (nt), which consist of 9-bp overhang for ligation and 5-nt barcode region and 19-nt mosaic pMENTs (5Phos/GCATTCGAGACGCAAGATGTGTATAAGAGACAG)) were annealed to a pMENTs oligo (5Phos/CTGTCTCTTATACACATCT/ddC; ddC, dideoxycytosine) in a thermocycler with the following program: 95 °C for 5 min, slowly cooled to 4 °C with a temperature ramp of 1 °C per minute. The transposons (9 µl, 10 µM) were then mixed with 10 µl of unloaded transposase Tn5 (Vazyme, 0.5 mg ml^{-1}, S601) and 17 µl of coupling buffer (Vazyme, S601), mixed by pipetting and quickly spun down and incubated at 30 °C for 30 min. The loaded transposases can be stored at −20 °C.

Oligo mixing experiment

DNA oligos that contain 5mC or 5hmC (T4C, T4M and T4H; Supplementary Table 1) were mixed with unmodified oligos in varying ratios (0%, 20%, 40%, 60%, 80% and 100%) and treated with the SIMPLE-seq protocol, including 5hmC treatment, primer extension, 5mC treatment and PCR amplification. PCR products were purified with 1.8× AMPure XP beads and eluted with 20 µl of nuclease-free water for high-throughput sequencing.

SIMPLE-seq procedures

Cell fixation and nucleosome depletion. To fix the cells, 406 µl of 37% formaldehyde (Sigma-Aldrich, F8775) was added into 10 ml of cell suspension and incubated at room temperature for 10 min with gentle shaking. After the incubation, 800 µl of 2.5 M glycine (Sigma-Aldrich, G7126) was immediately added and rotated up and down for five times and incubated on ice for 5 min to quench the fixation. The fixed cells were spun down at 4 °C and 550*g* for 8 min, and the cell pellet was washed with ice-cold PBS buffer (Invitrogen, 10149) and spun down again at 4 °C and 550*g* for 8 min. The cell pellet was then re-suspended in ice-cold NIB buffer (20 mM HEPES (Sigma-Aldrich, H3375), 10 mM NaCl, 3 mM $MgCl_2$ (Invitrogen, AM9530G), 0.1% Igepal (Thermo Fisher Scientific, 85124) and 1× cOmplete protease inhibitor cocktail) and incubated on ice for 20 min with gentle shaking. Nuclei were spun down again at 4 °C and 500 *g* for 5 min and re-suspended in 1× NEBuffer 2.1 (NEB, B7202) supplemented with 0.3% SDS (Invitrogen, 15553-035) and incubated at 60 °C and 850 r.p.m. for 10 min in the thermomixer. To quench the reaction, 200 µl of 10% Triton X-100 was then added, and the incubation was continued at 37 °C and 850 r.p.m. for 60 min. Nuclei were spun down at 100*g* and 4 °C for 10 min.

Species mixing. To estimate the chance of random cellular barcode collision, we performed a species-mixing experiment as described in the previous study[19] with minor modifications. In brief, human HEK293T cells and mESCs were permeabilized with ice-cold NIB buffer. Nuclei were then spun down at 4 °C and 500*g* for 5 min and re-suspended in 1× NEBuffer 2.1 supplemented with 0.3% SDS and incubated at 60 °C and 850 r.p.m. for 10 min in the thermomixer and quenched by adding 200 µl of 10% Triton X-100, and the incubation was continued at 37 °C and 850 r.p.m. for 60 min. Nuclei were spun down at 10*g* and 4 °C for 10 min and resuspended in PBS and counted. Next, 50,000 human nuclei and 100,000 mouse nuclei were mixed and spun down at 100*g* and 4 °C for 10 min, and barcoded tagmentation was then performed with Tn5 transposase pre-labeled with barcode 01.

Barcoded tagmentation. The nuclei pellet was resuspended in 1× Tn5 tagmentation buffer (Vazyme, S601), and a brief sonication was performed to break the nuclei clumps. Nuclei were passed through a 20-µm cell strainer. The nuclei were distributed into 48 wells (10,000 nuclei per well) that contained 12 well-specific barcoded Tn5 transposases (each of four wells sharing the same Tn5 barcode) and were incubated at 55 °C and 300 r.p.m for 30 min in a thermomixer. The reaction was quenched by adding 15 mM EDTA (Invitrogen, AM9260G), and nuclei were pooled together and passed through a 20-µm cell strainer. Nuclei were spun down at 1,000*g* and 4 °C for 10 min and resuspended in 1 ml of 1× NEBuffer 3.1 (NEB, B7203) and transferred to Ligation Mix (2,262 µl of UltraPure water (Invitrogen), 500 µl of 10× T4 DNA Ligase Buffer, 50 µl of 10 mg ml^{-1} BSA, 100 µl of 10× NEBuffer 3.1 and 100 µl of T4 DNA Ligase (NEB, M0202L)).

Ligation-based combinatorial barcoding. Next, 40 µl of the ligation mix was distributed to BC Plate 01 and incubated in a thermomixer at 37 °C and 300 r.p.m. for 30 min. After that, 10 µl of R01-Blocking-Solution (264 µl of 100 µM Blocker-R01 oligo, 250 µl of 10× T4 DNA Ligase Buffer and 486 µl of UltraPure water) was then added to each well, and the reaction was continued for 30 min. All nuclei were pooled together and centrifuged at 1,000*g* for 10 min at 4 °C. The second round of ligation was carried out similarly as the first round of ligation, except using BC Plate 02 and Blocker-R02 oligo instead of the reagents used above. After 30 min of the ligation reaction, Termination Solution (264 µl of 100 µM Blocker-R02, 250 µl of 0.5 M EDTA and 236 µl of UltraPure water) was added to quench the reaction. Typically, about 100,000 nuclei could be tagged after two rounds of ligation-based barcoding. Nuclei were resuspended in PBS buffer, counted and separated to sub-libraries containing optimal ~2,000 nuclei per tube, and each sub-library was diluted to 35 µl by PBS buffer. Then, 5 µl of 4 M NaCl, 5 µl of 10% SDS and 5 µl of 10 mg ml^{-1} Protease K (NEB, P8107S) were added and incubated in the thermomixer at 55 °C for 2 h and 850 r.p.m. The samples were cooled to room temperature, purified with 1× AMPure XP beads and eluted with 30 µl of nuclease-free water.

Second adaptor tagging. Genomic DNA fragments were incubated at 72 °C for 5 min and then chilled on an ice bath for 3 min. Next, 1× Exonuclease I reaction buffer and 1 µl of Exonuclease I (NEB, M0293S) were added to DNA solution and incubated at 37 °C for 60 min to remove excess Linker_R02. After 1.8× AMPure XP beads purification, DNA solution was incubated at 95 °C for 1 min in a thermocycler with a heated lid and then chilled on an ice bath. For the 3′-end ligation, Bridge Adapter (final concentration 10 µM), 1× T4 DNA Ligase Buffer, PEG8000 (7.5% w/v), 30 U of T4 DNA ligase (Thermo Fisher Scientific, EL0013) and nuclease-free water were added to the mixture on ice, to a total volume of 30 µl. The reaction was performed at 20 °C for 16 h, purified with 1× AMPure XP beads and eluted with 26 µl of nuclease-free water.

5hmC transition and 5hmC-to-T recording. Then, the genomic DNA was spiked with a 10-pg mixture of adaptor-ligated T1M, T2H and T3MH and subjected to 5hmC-specific oxidation and malononitrile-based selective labeling. To obtain the 5hmC profiles from the single cells, the 5hmC-labeled DNA was subjected to primer extension to record the 5hmC-to-T transition signal. Primer extension mix (10 µl of 5× KAPA buffer GC, 1 µl of 10 mM dNTPs, 2 µl of 10 µM P7-indicator primer and 0.8 µl of KAPA2G Robust HS DNA polymerase (Roche, KK5023)) was added, and primer extension was performed with the following program: step 1: 95 °C × 5 min; step 2: 65 °C × 60 s and 68 °C × 8 min and repeat step 2 an additional 15 times; and step 3: 72 °C × 10 min and hold at 4 °C. Then, 3 µl of 10× Exonuclease I reaction buffer and 1 µl of Exonuclease I were added to the mix, and the primer digestion reaction was incubated at 37 °C for 60 min and purified using 1.8× AMPure XP beads and eluted with 10 µl of nuclease-free water.

5mC transition and indexing PCR. After the primer extension reaction, the purified DNA was subjected to 5mC-specific oxidation and pic-borane-based selective labeling immediately. After the reaction, the final library was amplified by PCR amplification mix (25 µl of 2× Kapa U^+ HiFi Master Mix, 2.5 µl of 10 µM Universal primer and 2.5 µl of 10 µM Index primer), and the reaction was performed with the following program: step 1: 98 °C × 10 s; step 2: 98 °C × 10 s, 68 °C × 15 s and 72 °C × 1 min and repeat step 2 an additional 3–5 times; and step 3: 72 °C × 10 min and hold at 4 °C. The product was purified with 1.8× AMPure XP beads and eluted with 20 µl of nuclease-free water.

Sequencing. The purified libraries were sequenced on an MGISEQ-2000 sequencer (MGI) with the following read lengths: PE 200 +6 + 100 (Read 1 + Index 1 + Read 2); or on a NovaSeq sequencer (Illumina) with the following read lengths: PE 150 + 8 + 8 + 150 (Read 1 + Index 1 + Index 2 + Read 2).

Data analysis procedures

Pre-processing of SIMPLE-seq data. Cellular barcode extraction was carried out as previously described[83] with minor modifications. In brief, (1) unique molecular identifiers (UMIs) were extracted from

the first 10 bp of Read 2; (2) linker sequences were then identified to locate the first base of BC1 (11–15), BC2 (48–52) and BC3 (79–83); (3) barcode sequences were appended and mapped with Bowtie[104] to the reference with all possible combinations to assign the Cell ID for each read: reads with more than one mismatch and can be assigned to more than one cell were discarded (>85% of reads were retained in this study); (4) adaptor sequences were trimmed with Trim Galore[105], and low-quality reads (minimal read length $L = 30$, minimal base-calling quality $Q = 30$) were excluded from downstream analyses; (5) reads were mapped to the mouse GRCm38 or the human GRCh38 reference genome with Bowtie2 (ref. 106); and (5), finally, 5mC and 5hmC reads were split based on indicator bases to generate separate BAM files.

Reads splitting and modified site calling. Mapped 5mC and 5hmC reads were first split according to the indicator sequence of each read. To split the reads, we assign the reads with both 5caC sites in the indicator sequence converted (5hmC) or unconverted (5mC); the assignment accuracy is estimated as 99.94% based on measured non-converted 5caC sites. In addition, the efficiency of the primer extension steps could also contribute to the imbalance of 5mC–5hmC reads (Extended Data Fig. 1p).

For modified site calling at the single-cell level, we consider all cytosines with at least one read covered in the given single cell, and cytosine sites with more than 50% of C-to-T mutation rates were called modified cytosine. For modified base calling at the pseudo-bulk level, the following criteria were used: (1) ≥10 reads covered, (2) ≥10% C-to-T mutational rates and (3) Fisher's exact test $P < 0.05$, compare with non-specific C-to-T mutation of the genome.

Calculation of 5mC and 5hmC conversion levels on spike-in DNAs. Mapped 5mC and 5hmC reads were first split according to the indicator sequence of each read. The conversion rate of 5mC and 5hmC of known positions is averaged from all the detected reads. The non-specific conversion rate of 5hmC reads and 5mC reads is also averaged from all the C-to-T mutation of unmodified cytosines with known positions on spike-in dsDNA.

Normalization of 5mC and 5hmC modification levels. We performed normalization steps to adjust the estimated 5mC/5hmC modification levels at region-scale and global-scale of single cells. First, the non-mutational signals from the complementary strand were adjusted by multiplexing the average C-to-T rates by 2 (the fractions of reads derived from the complementary strand without base modification is expected to be 50% from population level). Next, the modification levels were adjusted according to the standard curve based on the oligo mixing experiments in Extended Data Fig. 1l,m. For site-level normalization, only the standard curve was applied.

Cell clustering with 5mC and 5hmC. Alignment files were converted to the cell-to-abundance matrix with cells as columns and 5mC–5hmC modification levels as rows (average abundance in 100-kb non-overlapping bins). Cells with fewer than 10,000 bins covered and bins with less than 50% of cells covered were removed, and missing values were imputed from the average levels for all cells with values. The cell-to-abundance matrix was then converted to a cell-to-cell similarity matrix by calculating the pair-wise cosine distance for each cell, followed by visualization with UMAP[107] and clustering with the Louvain algorithm[74] in Seurat[108] software. For joint clustering, the weighted nearest neighbor analysis framework in Seurat[108] software was used.

Calculation of cytosine modification entropy. To quantify the distributions of 5hmC states across single cells, we calculated the normalized Shannon entropy[109] of individual cells' disorder of 5hmC–5mC relationships. For each 5hmCG site, the numbers of 5hmCG sites and 5mCG sites within the ±100-bp range in the same cell were counted. The 100-bp range is determined by the sequencing read length, and scanning a longer range could increase the false-positive detection from different alleles. Next, the frequency of 5hmCG sites with the same number of flanking 5hmCG and the same number of flanking 5mCG sites was calculated by dividing the number of these 5hmCG sites (x_i) by the total number of 5hmCG sites captured in this cell (n_{5hmCG}): $p_i = x_i / n_{5hmCG}$. The cytosine modification entropy was then calculated using the formula: $H = -\sum_{i=1}^{n} p_i \log(p_i)$. To estimate the background noise, we shuffled the cell barcodes of detected 5mCG sites and calculated the background level cytosine modification entropy; the shuffling was performed for 10 times to obtain the average background levels for each cell.

Classification of 5hmC sites. To identify 5hmCG sites flanking by 5mCG, we first filtered fragments covered by both 5mC and 5hmC detection reads in each cell and identified 5hmCG sites detected in at least five cells from these 'dual-omics' fragments. Next, for each 5hmCG site, we scanned CG sites within ±100-bp range for each individual cell, and the numbers of cells without flanking 5mCG (n_i) and the numbers of cells with flanking 5mCG sites (n_d) were counted. To generate the background model, we shuffled the Cell IDs for the 'dual-omics' fragments and scanned the CG sites within ±100-bp range for each individual shuffled cell. To ensure reproducibility, the Cell IDs were shuffled 10 times, and the average numbers of shuffled cells without flanking 5mCG (n_{is}) and the average numbers of cells with flanking 5mCG sites (n_{ds}) were counted. To identify the 5hmCG sites enriched for cells with flanking 5mCG sites (type 2) from basal level 5hmCG sites (type 1), a cutoff (r) was applied to select 5hmCG sites with higher fractions of 5mCG flanking cells ($n_i / n_d > r$ and $n_{is} / n_{ds} > r$ in the shuffled group). We calculated false-positive detection rates (FDRs) based on the fraction of averaged false detected sites from shuffled groups in total detected sites, and FDR = 0.0467 was selected.

Cytosine state analysis. Cytosine state analysis was carried out with ChromHMM[79] software to integrate 5mC and 5hmC modification levels in different cell types. To prepare input for ChromHMM, average modification levels were calculated in 5-kb non-overlapping bins for 5mC (CG, CHG and CHH) and 5hmC (CG, CHG and CHH). We tested varying numbers (2–12) of states and chose the 10-state model because it recapitulated all modification combinations in the rest of the models. To identify the conserved and differential regions, we compared the cytosine states of the same genomic regions across different cell types: genomic regions which were classified as the same cytosine state in all five cell types were grouped as 'conserved regions'; regions which were classified into at least two different states in the five cell types were grouped as 'differential regions'.

Pseudotime analysis. Pseudotime analysis was carried out with Monocle3 (ref. 110) software based on dimensional reduction of 5mCG modification levels in 100-kb non-overlapping bins. ChromVAR[65] software was then used to calculate the TF binding motif enrichment score for 5mC and 5hmC sites (extended 500 bp for both upstream and downstream directions) in single cells. To identify potential functional TF binding motifs, motifs with positive and negative correlations (Spearmen's correlation > 0.2 or < −0.2) between 5mC and 5hmC modification levels along the pseudotime were retained.

Differential (hydroxyl)methylated regions and site analysis. To identify differential methylated and hydroxymethylated regions, average CG site modification levels were first calculated for all cells of the same cell type for each 5-kb non-overlapping bins. For pair-wise cell type differential modified regions, bins with at least 10% of maximum modification levels (5mCG and 5hmCG) and at least four-fold (5mCG)

or eight-fold (5hmCG) of modification level difference (A versus B) and Fisher's exact test $P < 0.05$ were considered as differentially modified regions. Motif enrichment analysis was carried out with Homer[111] software, and GO enrichment analysis was carried out with DAVID[112] with default parameters. To identify cell-type-specific 5mCG and 5hmCG sites in the mouse brain dataset, single-cell reads from the same cell type were first aggregated, and CG sites with fewer than 10 reads covered were removed. Next, 5mCG and 5hmCG sites with more than 10% C-to-T mutational rates were preserved. CG sites with modification levels at least two-fold compared to the average modification levels of the rest of the cell types were identified as cell-type-specific 5mCG or 5hmCG sites.

Reporting summary

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

Data availability

Raw sequencing and processed data generated in this study are available from the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE197740 (ref. 113). Other external datasets were downloaded from the GEO with the following accession numbers: WGBS and TAPS of mESC[55] (GSE112520), RNA sequencing of mESC (2i and serum)[66] (GSE23943), TAB-seq of mESC[46] (GSE36173), Joint-snhmC-seq of mouse brain[82] (GSE236798), Paired-Tag of mouse brain[83] (GSE152020); ArrayExpress with the following accession numbers: scRNA-seq of mESC[114] (E-MTAB-2600); 10x Genomics website: scRNA-seq of PBMCs (https://www.10xgenomics.com); ENCODE with the following accession numbers: DNase-seq ChIP-seq of E14 mESC (ENCSR000CMW), H3K4me1 ChIP-seq of E14 mESC (ENCSR000CGN), H3K27ac ChIP-seq of E14 mESC (ENCSR000CGQ), H3K4me1 ChIP-seq of immune cells (ENCSR777RWW ($CD4^+$ T cell), ENCSR631BPS ($CD8^+$ T cells), ENCSR214VUB (B cells), ENCSR963TKB (NK cells) and ENCSR400VWA (monocytes)), H3K4me3 ChIP-seq of immune cells (ENCSR263WLD ($CD4^+$ T cells), ENCSR231FDF ($CD8^+$ T cells), ENCSR269OVV (B cells), ENCSR570AUC (NK cells) and ENCSR796FCS (monocytes)), H3K27ac ChIP-seq of immune cells (ENCSR546SDM ($CD4^+$ T cells), ENCSR835OJV ($CD8^+$ T cells), ENCSR191ZQT (B cells), ENCSR391EQV (NK cells) and ENCSR012PII (monocytes)), H3K27me3 ChIP-seq of immune cells (ENCSR043SBG ($CD4^+$ T cells), ENCSR797GOJ ($CD8^+$ T cells), ENCSR522EGW (B cells), ENCSR939JZW (NK cells) and ENCSR080XUB (monocytes)), H3K9me3 ChIP-seq of immune cells (ENCSR453GNY ($CD4^+$ T cells), ENCSR905SHH ($CD8^+$ T cells), ENCSR295PSK (B cells), ENCSR021FSY (NK cells) and ENCSR236JVK (monocytes)), H3K36me3 ChIP-seq of immune cells (ENCSR828WZG ($CD4^+$ T cells), ENCSR694CDP ($CD8^+$ T cells), ENCSR789RGI (B cells), ENCSR519SOC (NK cells) and ENCSR244XWL (monocytes)) and ChromHMM states of NK cells (ENCSR972ZND).

Code availability

Custom scripts used for analyzing SIMPLE-seq datasets are available from GitHub (https://github.com/cxzhu/SIMPLE-seq)[115].

References

104. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
105. Krueger, F. Trim Galore. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (2019).
106. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
107. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2008).
108. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
109. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
110. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
111. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
112. Sherman, B. T. et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, W216–W221 (2022).
113. Bai, D., Zhu, C. & Yi, C. Single-cell joint analysis of 5-methylcytosine and 5-hydroxymethylcytosine. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE197740 (2023).
114. Kolodziejczyk, A. A. et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**, 471–485 (2015).
115. Bai, D., Zhu, C. & Yi, C. Custom scripts and pipeline for SIMPLE-seq data analysis. https://github.com/cxzhu/SIMPLE-seq (2023).

Acknowledgements

We thank Y. Zhuang, J. Song, H. Zeng, C.-G. Ji, H.-W. Meng and Z.-R. Xu for technical assistance; P. Du and F.-C. Tang (Peking University) for providing mESCs; G.-L. Xu (Chinese Academy of Sciences) for providing mTET1CD plasmid; J.-Y. Xiao (Peking University) for protein purification assistance; and Z. Xi (Nankai University) for discussion. We thank the National Center for Protein Sciences at Peking University for technical help. We carried out data analysis on the High-Performance Computing Platform at the School of Life Sciences, Peking University. This study is supported by the Ministry of Science and Technology of China (no. 2023YFC3402200, no. 2019YFA0110900 and no. 2019YFA0802201 to C.Y.), the Beijing Natural Science Foundation (no. Z220013 to C.Y.) and the National Natural Science Foundation of China (no. 91953201 and no. 21825701 to C.Y.).

Author contributions

D.B., C.Z. and C.Y. conceived and designed the study and wrote the paper. D.B. developed and optimized the SIMPLE-seq protocol and generated the data. C.Z. performed pilot labeling experiments, with help from D.B., and data analysis, with help from D.B. and X.Z. All authors discussed the results and edited the paper. C.Y. supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

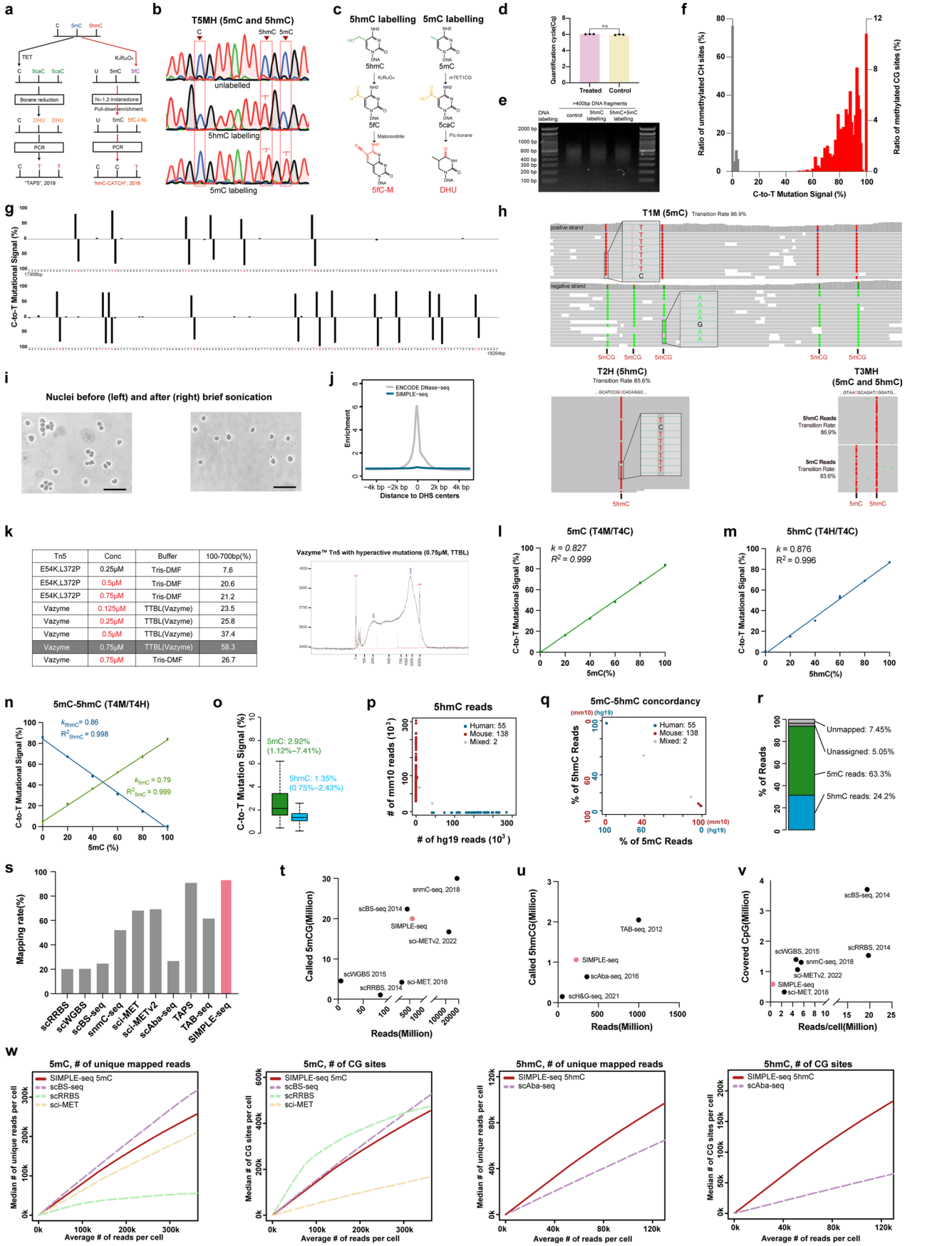
Extended data is available for this paper at https://doi.org/10.1038/s41587-024-02148-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-024-02148-9.

Correspondence and requests for materials should be addressed to Chenxu Zhu or Chengqi Yi.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

https://doi.org/10.1038/s41587-024-02148-9



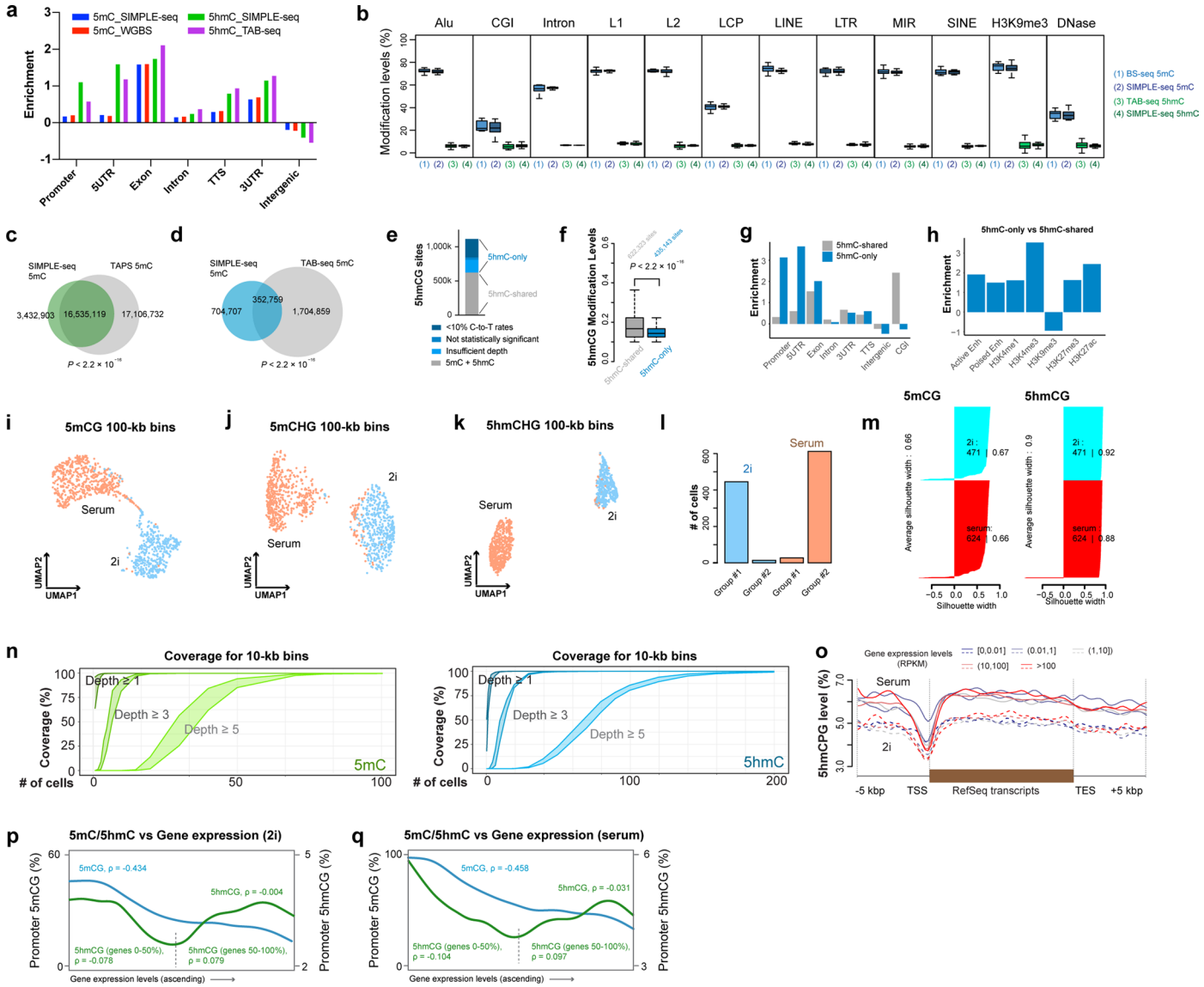
Tn5	Conc	Buffer	100-700bp(%)
E54K,L372P	0.25μM	Tris-DMF	7.6
E54K,L372P	0.5μM	Tris-DMF	20.6
E54K,L372P	0.75μM	Tris-DMF	21.2
Vazyme	0.125μM	TTBL(Vazyme)	23.5
Vazyme	0.25μM	TTBL(Vazyme)	25.8
Vazyme	0.5μM	TTBL(Vazyme)	37.4
Vazyme	0.75μM	TTBL(Vazyme)	58.3
Vazyme	0.75μM	Tris-DMF	26.7

Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41587-024-02148-9

Extended Data Fig. 1 | Sequential chemical labeling enables simultaneous detection of 5mC and 5hmC bases on the same molecules. a, Overview of TAPS and hmC-CATCH. **b**, Sanger sequencing results showing the 'C-to-T' conversion signal from model oligonucleotide sequence (T5MH) before treatment, after 5hmC labeling, and after both 5hmC and 5mC labeling. **c**, Schematics of chemical labeling for 5hmC and 5mC. **d**, qPCR result of lambda DNA (25,086bp-25,308bp) before and after potassium ruthenate (K_2RuO_4) treatment; n = 3 (Treated), n = 3 (Control). Data are presented as 6.020 ± 0.006 (Treated) and 5.957 ± 0.026 (Control). **e**, Agarose gel images of dsDNA of fragmented lambda DNA treated with 5hmC-labelling reaction only (left panel) and sequential 5hmC and 5mC-labelling (right panel). Experiment was performed once. **f**, Barplot showing the distribution of C-to-T mutation rates for unmethylated CH sites and methylated CG sites. **g**, C-to-T mutation signals on both strands of T1M spike-in model DNA, symmetric methylated CG sites are indicated in red on the sequences below. **h**, Genome browser view showing the sequenced reads aligned to spike-in model DNA (upper: T1M with 5mCG, positive and negative strands are separately displayed, bottom left: T2H with a single 5hmCG site, bottom right: T3MH with both 5mC site and 5hmCG site at known position). C-to-T is colored in red for positive strands, and G-to-A is colored in green for negative strands, respectively. Conversion rates are estimated from all the modified cytosines of spike-in model DNA. **i**, Nuclei clumps and resolved single nuclei suspension after brief sonication under bright field microscope. Experiment was performed once. Scale bars, 100 µm. **j**, Enrichment analysis of genome coverage by SIMPLE-seq and DNase-seq on DHS (DNase I hypersensitive sites). **k**, A table showing tagmentation efficiency under different Tn5 reaction conditions, including Tn5 with hyperactive mutations, working concentrations and reaction buffers; right-sided showing the fragments analysis result under optimal condition. **l-n**, Standard curves for mixed oligo DNA with (**l**) 5mC and C, (**m**) 5hmC and C, (**n**) 5mC and 5hmC, were plotted based on gradient mixing ratios (0:10; 2:8; 4:6; 6:4; 8:2; 10:0). **o**, C-to-T mutation rate estimated from 10-kb non-overlap bins across the whole genome. For both boxplots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median, whiskers with maximum 2× interquartile range (IQR). For 5mC, minima = 1.12%, maxima = 1.47%; for 5hmC, minima = 0.75%, maxima = 2.43%. 5mC, n = 156,755; 5hmC, n = 159,962. **p** and **q**, Scatter plot showing (**p**) the number of 5hmC reads mapped to human and mouse genome and (**q**) the fraction of 5mC and 5hmC reads mapped to human and mouse genome in each cell from the species-mixing experiment with twin axes. **r**, Stacked barplot showing the fraction of reads mapped to the reference genome and assigned to 5mC, 5hmC, or cannot be assigned. **s-v**, Comparisons between SIMPLE-seq and published single-cell and bulk 5mC and 5hmC sequencing methods: (**s**) fraction of mappable reads, (**t**) the number of 5mCG sites detected and total sequenced reads in each study, and (**u**) the number of 5hmCG sites detected and total sequenced reads in each study, (**v**) Dot plot showing the average number of covered CGs and sequenced reads for each cell in each study. **w**, Line plots showing the number of unique mapped reads or CG dinucleotides with at different sequenced read depths per cell in each study.

https://doi.org/10.1038/s41587-024-02148-9

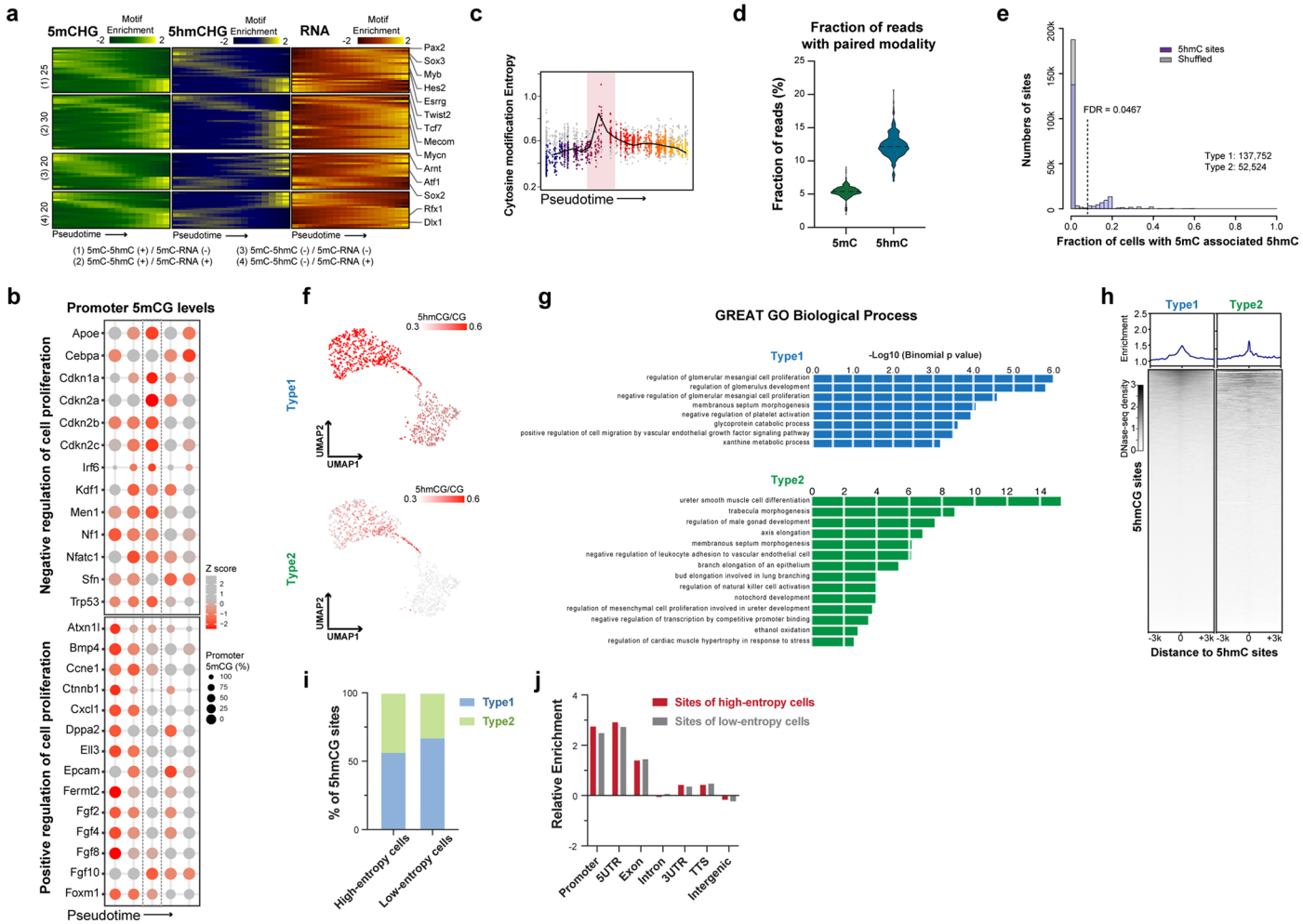


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41587-024-02148-9

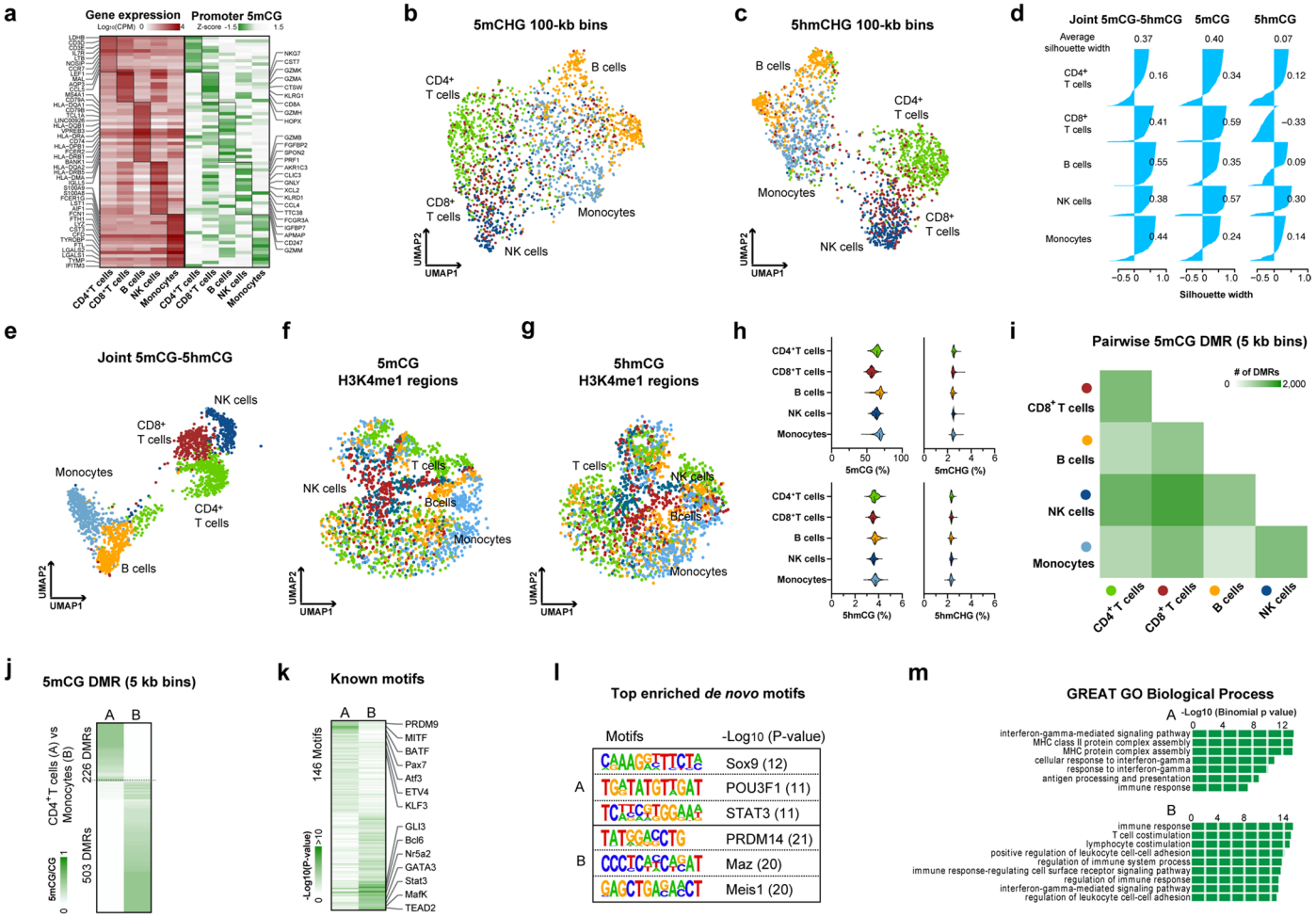
Extended Data Fig. 2 | Joint analysis of 5mC and 5hmC from single cells. **a**, Barplot showing the enrichment of 5mCG and 5hmCG sites detected by SIMPLE-seq, WGBS and TAB-seq over different genomic regions. **b**, Boxplots showing the 5mC or 5hmC modification levels on different genomic regions from bisulfite sequencing, TAB-seq and SIMPLE-seq. For all boxplots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median, whiskers with maximum 2× IQR. The minima/maxima/numbers of elements of all boxplots: 60.76%/74.60%/157,324 (Alu,(1)), 58.52%/68.61%/137,869 (Alu,(2)), 3.71%/10.71%/37,810 (Alu,(3)), 0.00%/10.91%/8,332 (Alu,(4)), 14.71%/67.84%/5,828 (CGI,(1)), 12.36%/54.72%/3,662 (CGI,(2)), 1.96%/5.87%/1,059 (CGI,(3)), 1.45%/5.24%/304 (CGI,(4)), 6.02%/72.62%/45,014 (Intron,(1)), 8.48%/79.35%/30,294 (Intron,(2)), 4.19%/5.01%/7,333 (Intron,(3)), 3.41%/4.85%/1,669 (Intron,(4)), 70.19%/75.69%/124,777 (L1,(1)), 64.68%/94.66%/103,247 (L1,(2)), 6.89%/10.32%/22,350 (L1,(3)), 4.90%/8.53%/5,414 (L1,(4)), 67.40%/75.88%/26,295 (L2,(1)), 68.26%/87.71%/18,819 (L2,(2)), 1.86%/7.30%/4,342 (L2,(3)), 2.20%/6.31%/875 (L2,(4)), 64.16%/73.22%/909 (LCP,(1)), 57.34%/79.78%/739 (LCP,(2)), 2.88%/10.32%/171 (LCP,(3)), 2.56%/7.71%/62 (LCP,(4)), 70.15%/75.10%/194,933 (LINE,(1)), 65.93%/81.40%/177,226 (LINE,(2)),3.79%/7.49%/34,376 (LINE,(3)), 3.99%/5.61%/9,412 (LINE,(4)), 68.59%/75.74%/189,543 (LTR,(1)), 60.01%/72.15%/170,132 (LTR,(2)), 2.85%/7.75%/32,147 (LTR,(3)), 3.24%/5.59%/10,073 (LTR,(4)), 68.60%/74.58%/47,771 (MIR,(1)), 66.88%/85.86%/30,355 (MIR,(2)), 2.88%/9.75%/7,707 (MIR,(3)), 0.00%/7.62%/1,520 (MIR,(4)), 62.68%/74.58%/325,867 (SINE,(1)), 58.25%/73.48%/302,881 (SINE,(2)), 4.93%/9.37%/56,712 (SINE,(3)), 0.00%/9.52%/19,715 (SINE,(4)), 62.82%/79.57%/4,651 (H3K9me3,(1)), 48.15%/94.31%/3,514 (H3K9me3,(2)), 3.79%/12.09%/862 (H3K9me3,(3)), 0.00%/11.97%/226 (H3K9me3,(4)), 14.28%/69.99%/21,711 (DNase,(1)), 12.70%/70.84%/15,340 (DNase,(2)), 0.00%/10.11%/3,961 (DNase,(3)), 0.00%/6.10%/973 (DNase,(4)). **c**, Venn plot showing the 5mCG sites overlap between SIMPLE-seq and TAPS. P-value, two-sided Fisher's exact test. **d**, Venn plot showing the 5hmCG sites overlap between SIMPLE-seq and TAB-seq. P-value, two-sided Fisher's exact test. **e**, Stacked barplot showing the fraction of called 5hmC sites overlapped with 5mC (grey, 5hmC-shared) and 5hmC-sites did not overlapped with a called 5mC sites (blue, 5hmC-only). **f**, Boxplot showing the 5hmC modification levels of 5mC-5hmC shared sites and the 5hmC-only sites. For both boxplots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median, whiskers with maximum 2× IQR. For 5mC-5hmC shared sites, minima = 0.01, maxima = 1.00, sites number n = 622,323, and for 5hmC-only sites, minima = 0.01, maxima = 1.00, sites number n = 435,143. P-value, two-sided Fisher's exact test. **g**, Barplot showing the enrichment of 5mC-5hmC shared sites and the 5hmC-only sites over different genomic regions. **h**, Barplot showing the relative enrichment of 5hmC-only sites over 5mC-5hmC shared sites on different genomic regions. **i-k**, UMAP embedding showing cells based on their (**i**) 5mCG, (**j**) 5mCHG and (**k**) 5hmCHG levels (in 100-kb non-overlapping bins). Each dot represents a single cell and is colored according to its original identity. **l**, Assignment of 2i mES cells and serum mES cells into two distinct clusters grouped by unsupervised clustering. **m**, Silhouette plot to evaluate the degree of separation of the clusters based on 5mC or 5hmC. **n**, Line plots showing the cumulated coverages of 10-kb non-overlapping bins with different depths for 5mC (green) and 5hmC (blue) from different numbers of single cells. The shadowed area showing the error ranges from 5 randomly sampled cell sets. **o**, Smoothed line plots showing the 5hmCG levels around genic regions of genes with different expression levels (using the *smooth.spline* function with parameter *df* = 30). **p-q**, Line plots showing the relationships between promoter 5mCG and 5hmCG modification levels with gene expression levels in (**p**) 2i mES cells, (**q**) serum mES cells.

https://doi.org/10.1038/s41587-024-02148-9



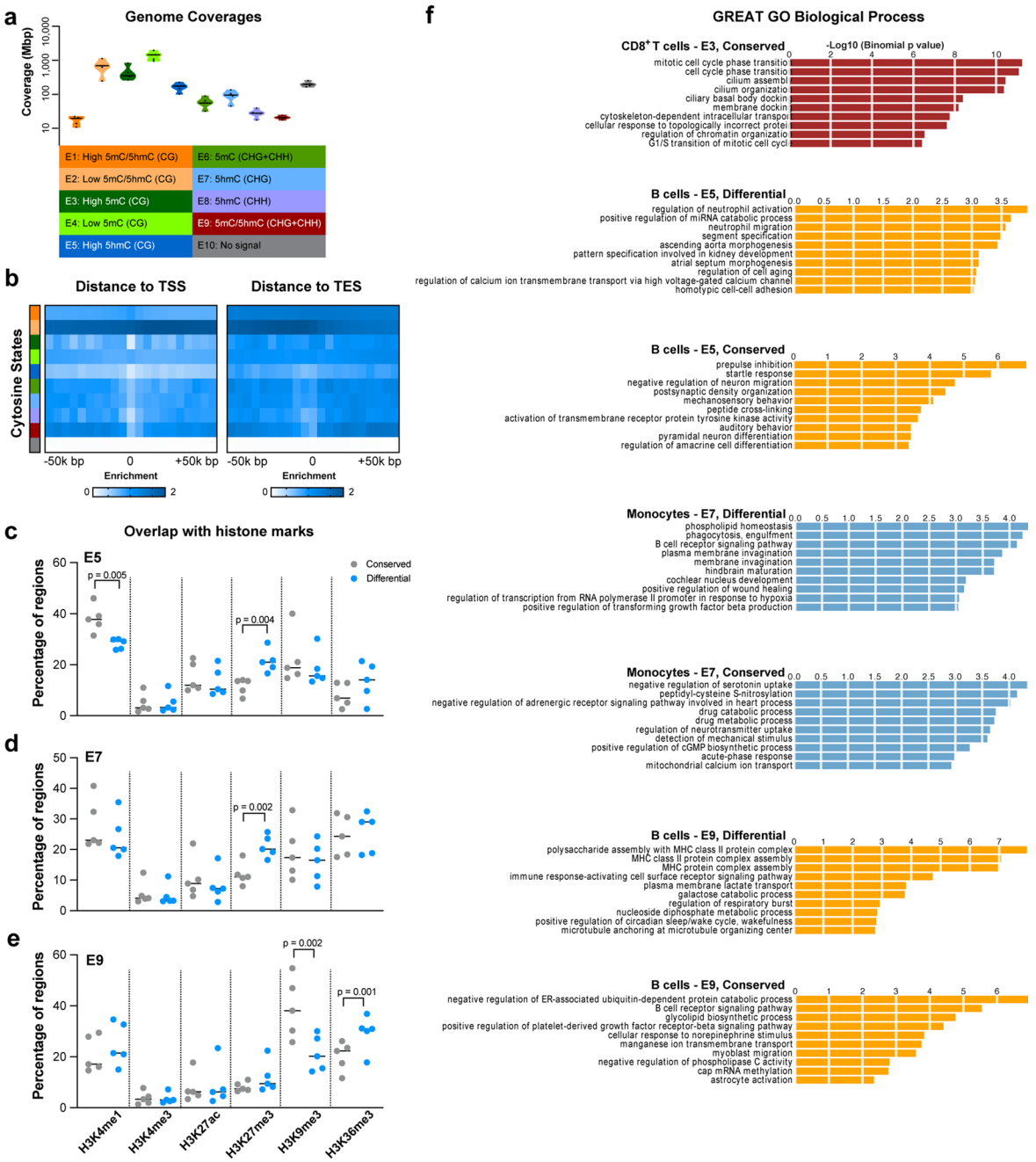
Extended Data Fig. 3 | Analysis of 5mC and 5hmC from the same molecules revealed multiple 5hmC types. a, Heat maps showing the 5mCHG and 5hmCHG TF motif enrichments along with TF expression level during mESC 2i state to serum state transition. **b**, Dotplots showing promoter methylation levels for genes associated with cell proliferation during mESC 2i to serum state transition. **c**, Cytosine modification entropies in single cells. Each dot represents a single cell and is colored and ordered according to its pseudotime score. Grey dots are background entropy levels estimated by shuffled the cell barcodes of called modification sites. **d**, Violin plots showing the fraction of 5mC and 5hmC reads with paired modality in single cells. For both Violin plots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median, whiskers with maximum 2× IQR. For 5mC modality, minima = 1.94%, maxima = 9.12%; for 5hmC modality, minima = 6.93%, maxima = 20.75%; n = 300 randomly sampled cells. **e**, Barplot showing the distributions of 5hmCG sites with different fractions of cells with detected 5mCG-associated 5hmCG sites. False positive detection rates (FDR) based on the fraction of averaged false detected (Type 2) sites from shuffled groups in total detected sites (FDR = 0.0467 was selected). **f**, UMAP embedding showing 5hmC levels of different types in single cells. Each dot represents a single cell and is colored according to the average 5hmC level. **g**, Top enriched GREAT GO terms for different types of 5hmCG sites. P-value, one-sided Fisher's exact test. **h**, Histograms and heatmaps showing the relationship between two types of 5hmCG with mESC DNase-seq signals from ENCODE (ENCSR000CMW). **i**, Fraction of Type 1 and Type 2 5hmCGs in detected from low-entropy (0–75%) or high-entropy (75%-100%) cells. **j**, Barplot showing the enrichment of 5hmCG sites detected in low-entropy or high-entropy cells over different genomic regions.

https://doi.org/10.1038/s41587-024-02148-9



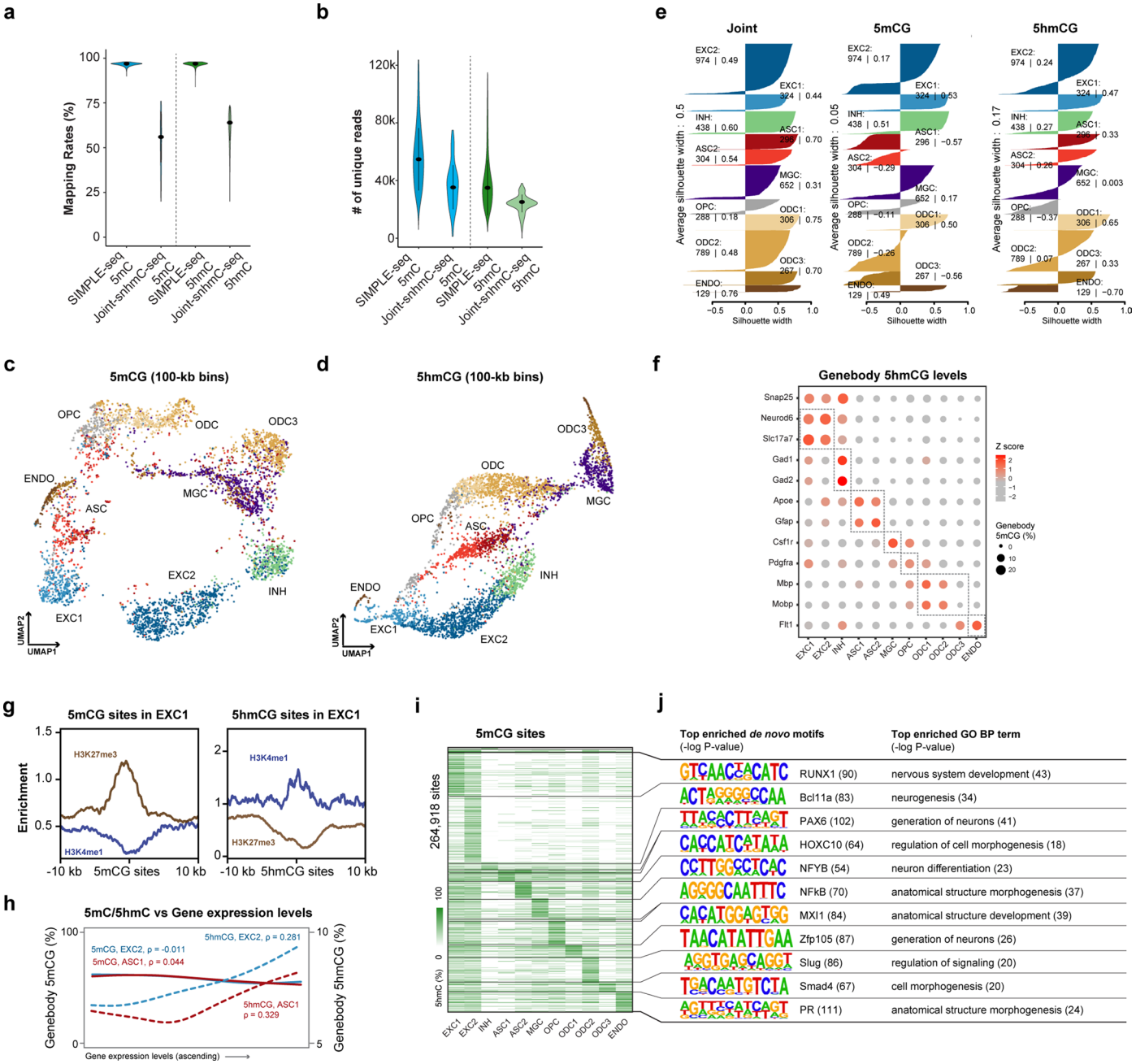
Extended Data Fig. 4 | SIMPLE-seq generates cell-type-specific 5mC and 5hmC profiles from human PBMC. a, Heatmap showing the gene expression and promoter methylation levels of marker genes for the five major cell types. **b** and **c**, UMAP embedding showing the single-cell clustering based on (**b**) 5mCHG and (**c**) 5hmCHG levels (in 100-kb non-overlapping bins) from PBMC. Each dot represents a single cell and is colored according to its annotation based on 5mCG.**d**, Silhouette plot showing the degree of separation of the PBMC clusters based on 5mC, 5hmC or joint 5mCG-5hmCG. **e**, UMAP embedding showing the single-cell clustering based on joint 5mCG-5hmCG levels (in 100-kb non-overlapping bins) from PBMC. Each dot represents a single cell and is colored according to its annotation based on 5mCG. **f** and **g**, UMAP embedding showing the single-cell clustering based on (**f**) 5mCG levels of H3K4me1 regions (**g**) 5hmCG levels of H3K4me1 regions from PBMC. Each dot represents a single cell and is colored according to its annotation based on 5mCG. **h**, Violin plots showing the modification levels of 5mCG, 5hmCG, 5mCHG and 5hmCHG in different cell types. **i**, The heatmap showing the numbers of pairwise differentially methylated (5mCG) regions across cell types (in 5-kb non-overlapping bins). **j**, Heatmap showing the methylation levels of 5mCG DMRs of a representative group ($CD4^+$ T cells and Monocytes). **k**-**m**, The enrichment analysis of (**k**) know motifs, (**l**) top enriched de novo motifs, P-value, one-sided Fisher's exact test, and (**m**) top enriched GREAT GO terms, P-value, one-sided Fisher's exact test.

https://doi.org/10.1038/s41587-024-02148-9



Extended Data Fig. 5 | Comparison of conserved and differential cytosine states among immune cells. a, Violin plots showing the genomic coverages of the 10 cytosine states. **b**, Heatmap showing the enrichment of different cytosine state regions around TSS and TES sites. **c**-**e**, Scatter plot showing the fraction of genome regions overlapped with peaks of different histone marks in conserved and differential (**c**) E5 (**d**) E7 and (**e**) E9 state regions. P-value, two-sided t-test. **f**, Top enriched GREAT GO terms for conserved and differential regions of E5, E7 and E9 in representative cell types. P-value, two-sided t-test.

https://doi.org/10.1038/s41587-024-02148-9



Extended Data Fig. 6 | SIMPLE-seq generates cell-type-specific 5mC and 5hmC profiles from mouse brain. a, Violin plots showing the fraction of reads mapped to reference mouse genome for 5mC and 5hmC in SIMPLE-seq (this study) and Joint-snhmC-seq (GSE236798) datasets. **b**, Violin plots showing the numbers of unique reads per cell for 5mC and 5hmC in SIMPLE-seq (this study) and Joint-snhmC-seq (GSE236798) datasets. For fair comparisons, Joint-snhmC-seq dataset was down sampled to the same per-cell depth in SIMPLE-seq. Data are presented as 57,563 ± 4,260 (SIMPLE-seq, 5mC), 35,283 ± 552 (joint-snhmC-seq, 5mC), 39,374 ± 4,553 (SIMPLE-seq, 5hmC), 20,538 ± 565 (joint-snhmC-seq, 5hmC). Cell number n = 4,767 (SIMPLE-seq, 5mC and 5hmC), n = 552 (joint-snhmC-seq, 5mC and 5hmC). **c** and **d**, UMAP embedding showing the single-cell clustering based on (**c**) 5mCG and (**d**) 5hmCG levels (in 100-kb non-overlapping bins) from mouse brain cells. Each dot represents a single cell and is colored according to its annotation based on joint 5mCG-5hmCG clustering. **e**, Silhouette plot showing the degree of separation of the mouse brain cell clusters based on 5mC, 5hmC or joint 5mCG-5hmCG. **f**, Dot plots showing the genebody 5hmCG levels of representative marker genes in the detected cell types. **g**, The distribution of H3K4me1 and H3K27me3 reads densities around the 5mCG sites or 5hmCG sites from EXC1. **h**, Line plots showing the relationships between gene body 5mCG and 5hmCG levels with gene expression levels in EXC2 and ASC1 cell types **i**, Heatmap showing the 5mCG modification levels of cell type-specific 5mCG across the 11 cell types. **j**, Top enriched de novo motifs (left) and top enriched GO terms (right) for each cell type were also shown. P-value, one-sided Fisher's exact test.

nature portfolio

Corresponding author(s): Chengqi Yi

Last updated by author(s): Jan 17, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Next-generation sequencing data was collected and demultiplexed by BGI MGI-2000 platform and Illumina Hiseq.
Data analysis	Bowtie2(2.2.9),TrimGalore(0.6.5), HOMER(4.10), Seurat(3.0), ChromHMM(1.23), ChromVAR(3.12), Louvain (iGraphl.2.6), Monocle3(0.2.3), Samtools(1.14), deepTools(3.5.1), IGV(2.15.4), bedtools(2.27.1), David(2021)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw sequencing and processed data generated in this study are available from GEO (http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE197740. Other external datasets were downloaded from NCBI GEO with the following accession numbers: WGBS and TAPS of mESC56 (GSE112520), RNA-seq of mESC (2i and serum)67 (GSE23943), TAB-seq of mESC47 (GSE36173), Joint-snhmC-seq of mouse brain84 (GSE236798), Paired-Tag of mouse brain85 (GSE152020); ArrayExpress

with the following accession numbers: scRNA-seq of mESC115 (E-MTAB-2600); 10x Genomics website: scRNA-seq of PBMC (https://www.10xgenomics.com); ENCODE with the following accession numbers: DNase-seq ChIP-seq of E14 mESC (ENCSR000CMW), H3K4me1 ChIP-seq of E14 mESC (ENCSR000CGN), H3K27ac ChIP-seq of E14 mESC (ENCSR000CGQ), H3K4me1 ChIP-seq of immune cells (ENCSR777RWW (CD4+ T cell), ENCSR631BPS (CD8+ T cells), ENCSR214VUB (B cells), ENCSR963TKB (NK cells), and ENCSR400VWA (Monocytes)), H3K4me3 ChIP-seq of immune cells (ENCSR263WLD (CD4+ T cell), ENCSR231FDF (CD8+ T cells), ENCSR269OVV (B cells), ENCSR570AUC (NK cells), and ENCSR796FCS (Monocytes)), H3K27ac ChIP-seq of immune cells (ENCSR546SDM (CD4+ T cell), ENCSR835OJV (CD8+ T cells), ENCSR191ZQT (B cells), ENCSR391EQV (NK cells), and ENCSR012PII (Monocytes)), H3K27me3 ChIP-seq of immune cells (ENCSR043SBG (CD4+ T cell), ENCSR797GOJ (CD8+ T cells), ENCSR522EGW (B cells), ENCSR939JZW (NK cells), and ENCSR080XUB (Monocytes)), H3K9me3 ChIP-seq of immune cells (ENCSR453GNY (CD4+ T cell), ENCSR905SHH (CD8+ T cells), ENCSR295PSK (B cells), ENCSR021FSY (NK cells), and ENCSR236JVK (Monocytes)), H3K36me3 ChIP-seq of immune cells (ENCSR828WZG (CD4+ T cell), ENCSR694CDP (CD8+ T cells), ENCSR789RGI (B cells), ENCSR519SOC (NK cells), and ENCSR244XWL (Monocytes)), ChromHMM States of NK cells (ENCSR972ZND).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. The sample size was determined based on prior published data from similar experiments (Mulqueen et.al., Nat. Biotechnol, 2018; Zhu et.al., Nat. Methods, 2021).
Data exclusions	Cells (mESC, PBMC and mouse brain cells) with less than 10,000 bins covered and bins with less than 50% of cells covered were removed.
Replication	48 technical replicates were performed for mESC, PBMC and mouse brain cells, respectively. All datasets from independent replicates showed similar results and shown in the manuscript.
Randomization	Allocation was random.
Blinding	Experiments were not blinded in our studies because experimental conditions were evident. All samples were treated identically through standard procedures.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	mouse embryonic stem cells were provided by the Laboratory of Prof. Peng Du (source: v6.5 mESC); HEK293T(CRL-3216) and NIH/3T3(CRL-158) were purchased from ATCC.
Authentication	No method of cell line authentication was used.
Mycoplasma contamination	All cell lines are tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No such cell lines listed as commonly misidentifiedl lines were used in this study.