Single-cell bisulfite-free 5mC and 5hmC sequencing with high sensitivity and scalability

Yunlong Cao[a,b,1] , Yali Bai[b,c,1] , Tianjiao Yuan[a,1] , Liyang Song[b], Yu Fan[b], Liuhao Ren[a] , Weiliang Song[a], Jiahui Peng[a], Ran An[b], Qingqing Gu[b], Yinghui Zheng[a], and Xiaoliang Sunney Xie[a,b,d,2]

Existing single-cell bisulfite-based DNA methylation analysis is limited by low DNA recovery, and the measurement of 5hmC at single-base resolution remains challenging. Here, we present a bisulfite-free single-cell whole-genome 5mC and 5hmC profiling technique, named Cabernet, which can characterize 5mC and 5hmC at single-base resolution with high genomic coverage. Cabernet utilizes Tn5 transposome for DNA fragmentation, which enables the discrimination between different alleles for measuring hemi-methylation status. Using Cabernet, we revealed the 5mC, hemi-5mC and 5hmC dynamics during early mouse embryo development, uncovering genomic regions exclusively governed by active or passive demethylation. We show that hemi-methylation status can be used to distinguish between pre- and post-replication cells, enabling more efficient cell grouping when integrated with 5mC profiles. The property of Tn5 naturally enables Cabernet to achieve high-throughput single-cell methylome profiling, where we probed mouse cortical neurons and embryonic day 7.5 (E7.5) embryos, and constructed the library for thousands of single cells at high efficiency, demonstrating its potential for analyzing complex tissues at substantially low cost. Together, we present a way of high-throughput methylome and hydroxymethylome detection at single-cell resolution, enabling efficient analysis of the epigenetic status of biological systems with complicated nature such as neurons and cancer cells.

single-cell | 5mC sequencing | 5hmC sequencing | bisulfite-free conversion

In mammals, DNA methylation mostly occurs at CpG dinucleotides in the form of 5-methylcytosine (5mC), which plays a vital role in gene regulation, cellular development, and disease formation (1, 2). Single-cell DNA methylation sequencing is required to reveal cellular heterogeneity and to study rare samples such as early embryonic development (3–7). However, bisulfite sequencing (BS-seq), the gold standard for mapping mammalian DNA methylation (8, 9), involves a harsh chemical reaction that degrades most of the double-stranded DNA (dsDNA), resulting in considerable loss of information (10). Despite huge DNA loss, bisulfite-dependent single-cell methylation sequencing has already been deployed in biological research. For example, a single-cell methylome sequencing method based on reduced representation bisulfite sequencing (RRBS) was developed and applied to probe the methylation in early mouse embryos (11). Yet, this method preferably covers regions with high CpG density and yields a genomic coverage of less than 5%. To achieve whole-genome methylation sequencing, other methods such as scBS-seq and scWGBS have been developed based on the post-bisulfite adaptor tagging (PBAT) strategy through random-primed DNA synthesis (12–15). However, multiple rounds of random priming, which are intended to achieve maximum dsDNA recovery, could introduce primer dimers and concatemers which lead to low mapping efficiency. The later reported snmC-seq and snmC-seq2 methods improved read mapping by combining random priming with adaptase reaction (16, 17). Nonetheless, these methods still all involve harsh bisulfite treatment which degrades most of the DNA and limits the overall genomic coverage. An efficient whole-genome methylation detection method at single-cell and single-base resolution is highly preferred. However, to develop such a method, several problems need to be solved.

First, DNA damage caused by bisulfite treatment should be avoided, to which end several bisulfite-free methylome sequencing techniques have been proposed. TET-assisted pyridine borane sequencing (TAPS) uses TET oxidation coupled with pyridine borane reduction to convert modified cytosines to dihydrouracil (DHU) (18). Enzymatic methyl-seq (EM-seq) combines TET2 and BGT protection of 5mC and 5-hydroxymethylcytosine (5hmC), with APOBEC deamination of C to U (19). While these methods allow for better DNA preservation by avoiding bisulfite treatment, the multiple purification steps involved in these methods lead to substantial loss of DNA during the process, limiting their application to single-cell scenarios. Similarly, techniques developed for 5hmC profiling such as 5hmC-DIP

Significance

Most current methylation profiling techniques rely on bisulfite treatment, which suffers low DNA recovery. The technique proposed in this study, named Cabernet, can be used to measure 5mC and 5hmC at single-base resolution with high genomic coverage. By using Tn5 transposome, hemi-methylation status can be measured and high-throughput methylome profiling can be achieved. Together, it provides an efficient way to analyze the epigenetic landscape of complicated biological systems.

and hMe-Seal (20, 21) which are based on affinity enrichment, cannot achieve single-base resolution. The later reported TAB-seq, oxBS-seq, and ACE-seq methods, all require high DNA input due to complicated workflows and multiple purification steps, limiting their application to single cells (22–25). To reveal 5mC and 5hmC heterogeneity in cell populations across the entire genome, a bisulfite-free DNA modification profiling approach with minimal DNA loss at single-cell level and single-base resolution is preferred.

Second, characterizing cell-type-specific methylation status in complex tissues requires a scalable single-cell-based sequencing technique. Currently established single-cell sequencing methods (12–17), where each single cell is processed in an individual preparation, are highly labor-intensive. Although the application of brute-force strategy can help scale up to a certain extent, as in snmC-seq (16), it is still a laborious method since every cell produced requires an individual reaction. On the other hand, sci-MET solves this problem by using multiple stages of indexing to label thousands of single cells for pooled library preparation, enabling high-throughput 5mC profiling (26). The employment of barcoded Tn5 transposome during combinatorial indexing is a critical step for the high throughput of this method; however, the use of bisulfite treatment limits the role of Tn5, necessitating subsequent random priming. Yet, despite high scalability, this method yields substantially low genomic coverage due to bisulfite conversion. Therefore, a high-throughput single-cell methylation sequencing method with improved DNA coverage is in need.

To address these challenges, we developed a method named Cabernet (Carrier-Assisted Base-conversion by Enzymatic ReactioN with End-Tagging), based on EM-seq. Instead of bisulfite treatment, Cabernet relies on enzymatic conversion where 5mC and 5hmC are converted to 5gmC through TET oxidation and BGT-mediated glycosylation while C is converted to uracil through APOBEC-mediated deamination, following EM-seq protocol (19). Importantly, Cabernet utilizes carrier DNA to minimize DNA loss during purification and achieves scalability through the employment of Tn5 transposome. Together, this technique enables high-throughput detection of whole methylome and hydroxymethylome at single-cell and single-base resolution with high genomic coverage.

Results

Validation of Cabernet. The process of Cabernet includes the whole methylation conversion protocol of EM-seq (19), which combines TET-mediated oxidation, BGT-mediated glycosylation, and APOBEC-mediated deamination reactions. Given that the original protocol is not applicable to single-cell settings due to huge DNA loss, the following adjustments are made. First, Tn5 insertion, instead of ligation, is used during fragmentation to minimize DNA loss from ultrasonic shearing. Second, carrier dsDNA is introduced to prevent the loss of dsDNA from irretrievable adhesion during purification. Third, library amplification is carried out directly after enzymatic conversion without ssDNA purification. As in BS-seq, Cabernet measures the combined level of 5mC and 5hmC as 5mC; however, since the abundance of 5mC is much higher than 5hmC in most circumstances, the combined level of 5mC and 5hmC can approximate to that of 5mC alone for most cell types. Additionally, omitting TET2 conversion during the first step of enzymatic treatment enables direct measurement of 5hmC, where 5hmC is specifically protected from deamination through BGT-mediated glycosylation to 5gmC and read as C while 5mC/C are deaminated and read as T, and we name this 5hmC sequencing protocol as Cabernet-H (Fig. 1A).

To evaluate the accuracy of Cabernet in profiling single-cell 5mC, we applied Cabernet to single K562 cells as compared with scBS-seq, which is known as the gold standard of DNA methylation detection. A high consistency was observed between Cabernet (single-cell or bulk) and scBS-seq in mapping genome-wide DNA methylation (Fig. 1B and *SI Appendix,* Fig. S1). Of note, 5mC, 5hmC, hemi-5mC, or hemi-5hmC level in this study is referred to as the respective modification level at CpG sites, unless specified elsewhere. To test Cabernet-H in mapping 5hmC modifications, we applied Cabernet-H to mouse embryonic stem cells (mESCs). Genome-wide 5hmC modification maps obtained by single-cell Cabernet-H and bulk Cabernet-H broadly matched previously published DIP-seq dataset (Fig. 1C and *SI Appendix,* Fig. S2) (27). As verified in K562 and mESCs, Cabernet and Cabernet-H can yield 5mC and 5hmC maps in a broad match with previously proved methods.

To further validate the accuracy of Cabernet and Cabernet-H, we tested the conversion efficiency on three standard controls, which are unmodified λ-DNA, CpG methylated pUC19 plasmid, and 5hmC modified DNA fragment. In Cabernet, up to 98.5% of 5mC were correctly called, with a false-positive rate of 0.854% (C recognized as 5mC), as verified on CpG methylated pUC19 plasmid and unmodified λ-DNA. In Cabernet-H, the recall rate of 5hmC was 99.7% as verified on 5hmC modified DNA fragment, with only 0.415% of C and 1.29% of 5mC falsely called 5hmC (Fig. 1D and *SI Appendix,* Fig. S3). Besides, Cabernet and Cabernet-H also showed substantially higher mapping rates than scBS-seq (Fig. 1E), due to the use of a specific adaptor for selective amplification in contrast to the random priming strategy used by bisulfite-based methods.

The major limitation of scBS-seq is low DNA coverage, which we aim to improve with Cabernet. To examine the advantage of Cabernet over scBS-seq regarding genome coverage, we performed deep sequencing to K562 samples prepared under both protocols. The average genome coverage obtained with Cabernet and Cabernet-H was about twice that yielded by scBS-seq. Analysis of downsampled sequencing data across different depths showed Cabernet and Cabernet-H yielded a similar coverage under the same depth, both significantly outperforming scBS-seq. Specifically, Cabernet yielded nearly 50% genome coverage, doubling that of scBS-seq (Dataset S1). Such an advantage is especially prominent at lower sequencing depths. For example, when downsampled to 20 million of reads (1×), the gap is amplified to 2.6-fold (Fig. 1F). Even at 0.01X~0.1X sequencing depths, Cabernet and Cabernet-H still demonstrate an advantage (*SI Appendix,* Fig. S4). This confirms that Cabernet and Cabernet-H can yield higher coverage, potentially due to less DNA damage and higher DNA recovery.

To assess the stability of Cabernet performance in single cells, we examined the correlation between single-cell measurements in a homologous cell population. Pearson's correlation analysis indicated a strong correlation among single cells, revealing the stability of Cabernet. Besides, the strong correlation between single-cell Cabernet and scBS-seq further validates the accuracy of Cabernet (Fig. 1G and *SI Appendix,* Fig. S5A). By contrast, Cabernet-H analysis of mESCs (Dataset S2) demonstrated prominent 5hmC heterogeneity between individual cells, probably due to intrinsic 5hmC dynamics (Fig. 1H and *SI Appendix,* Fig. S5B). Nonetheless, merged single-cell Cabernet-H data broadly matched those obtained from bulk sequencing (Fig. 1H). As shown above, Cabernet and Cabernet-H can measure 5mC and 5hmC with high genome coverage at high accuracy.

Cabernet-H Accurately Captures 5hmC. As shown above, unlike 5mC measurements, single-cell 5hmC measured by Cabernet-H does not show high consistency among single cells, probably due to intrinsic 5hmC dynamics. To further validate whether
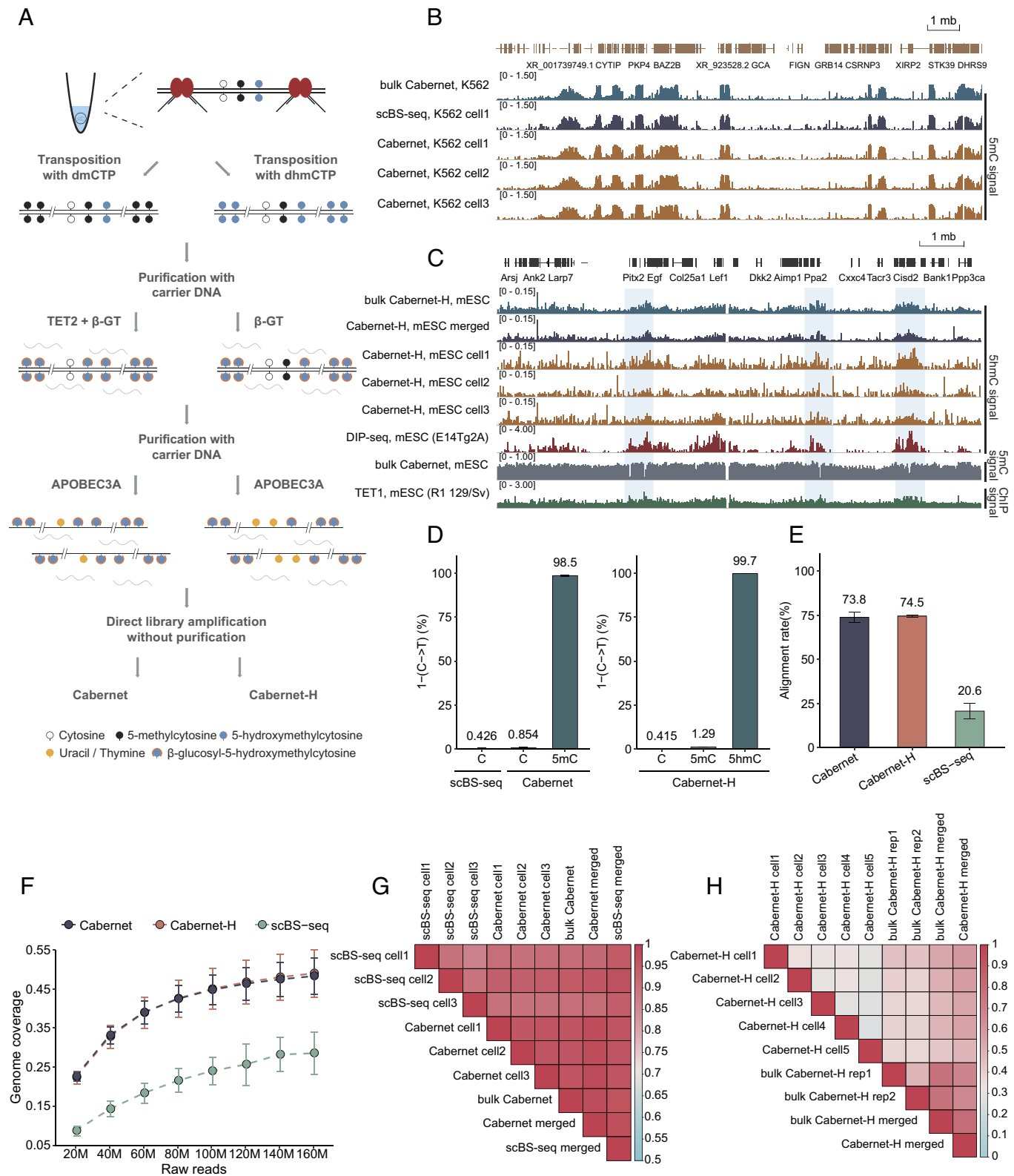
Fig. 1. Workflow and validation of Cabernet. (*A*) Schematic of Cabernet and Cabernet-H in a single cell. (*B*) IGV visualization of 5mC enrichment in K562 cell line using the Cabernet sequencing method (chr2:154009136-169479494). (*C*) IGV visualization showing 5hmC enrichment in mESC using the Cabernet-H sequencing method (chr3:126,231,297-137,134,000). DIP-seq and ChIP-seq data were downloaded from GSE46111 and GSE77453. (*D*) C-T conversion rate of 5mC in Cabernet and 5hmC in Cabernet-H; the conversion rate of unmodified C is monitored by λ-DNA spike-in. (*E*) Barplot of the genome alignment rate in Cabernet, Cabernet-H, and scBS-seq. (*F*) Genome coverage of Cabernet, Cabernet-H, and scBS-seq under different numbers of downsampled reads in K562 cells. (*G*) Heatmap of Pearson correlation between Cabernet and scBS-seq sequencing results in K562. (*H*) Heatmap of Pearson correlation between single Cabernet-H and bulk Cabernet-H in mESC.

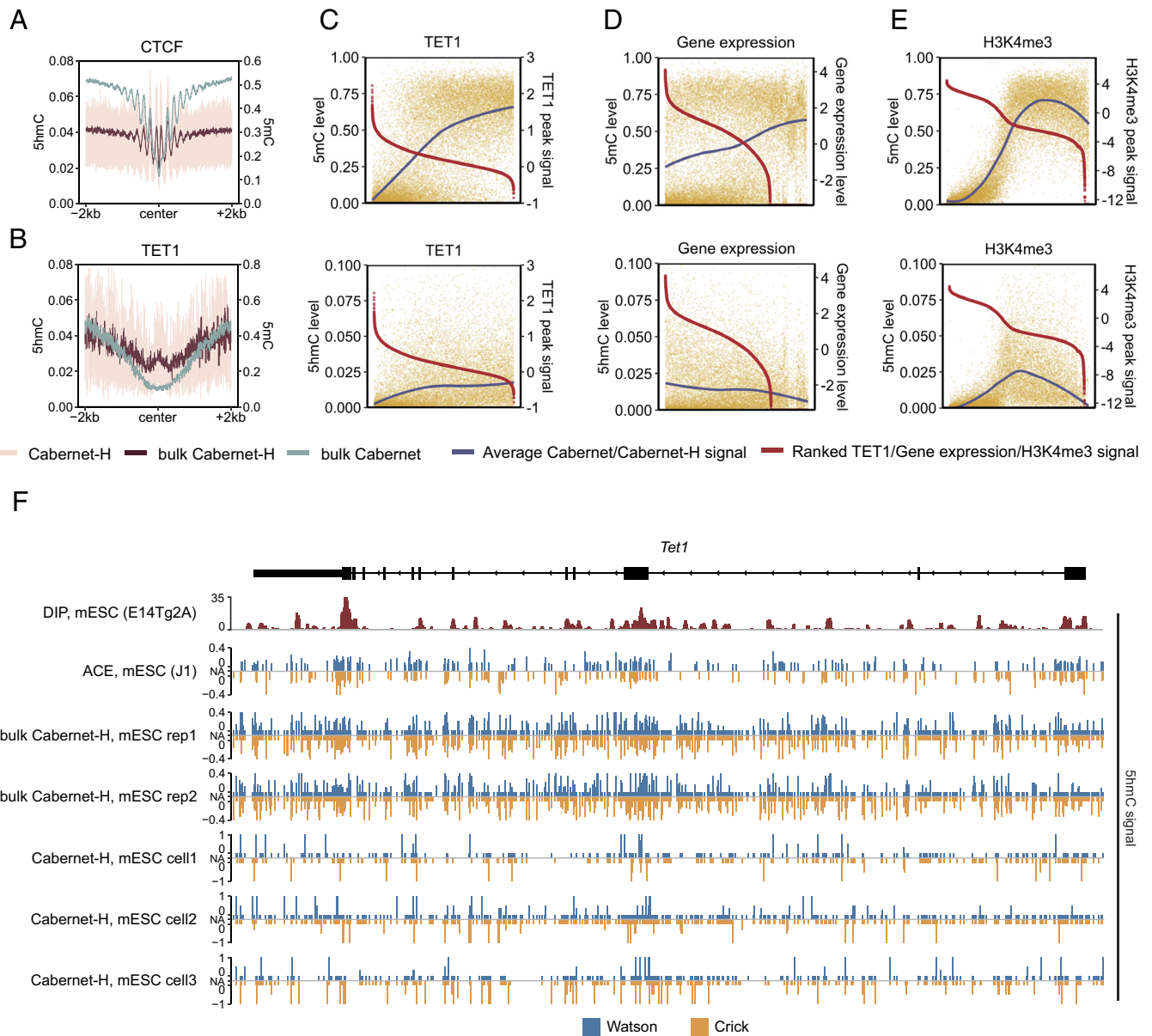**Fig. 2.** Characteristics of methylation (5mC) and hydroxy-methylation (5hmC) respectively detected by Cabernet and Caberbet-H. (*A*) Distribution of 5mC and 5hmC in the CTCF binding regions (downloaded from GSE36027, CTCF peak center ±2 kb). Pink: Cabernet-H; brown: Cabernet-H_bulk; green: Cabernet_bulk. (*B*) Distribution of 5mC and 5hmC in the TET1 binding regions (downloaded from GSE77453, TET1 peak center ±2 kb). Pink: Cabernet-H; brown: Cabernet-H_bulk; green: Cabernet_bulk. (*C*) *Top*: signal intensities of TET1 ChIP-Seq peaks within promoter regions in mESCs and 5mC levels at the corresponding peak regions. The red and blue fitting curves represent peak signal intensity and 5mC level in corresponding regions, respectively. *Bottom*: Signal intensities of TET1 ChIP-Seq peaks within promoter regions in mESCs and the 5hmC levels of the corresponding peak regions. The red and blue fitting curves represent peak signal intensity and 5hmC level in corresponding regions, respectively. The horizontal axis from left to right of each box represents the TET1 ChIP-Seq peaks, which overlapped with promoter regions, ranked by peak signal intensities from high to low. (*D*) *Top*: 5mC levels at promoter regions and the expression levels (downloaded from GSE153575) of corresponding genes in mESCs. The log10 of gene expression levels (transcripts per kilobase per million mapped reads, TPM) were calculated and presented. Red and blue curves represent gene expression levels and 5mC levels at promoters, respectively. *Bottom*: DNA 5hmC levels at promoters and the expression levels of corresponding genes in mESCs. (*E*) *Top*: Signal intensities of H3K4me3 ChIP-Seq peaks within promoter regions in mESCs (downloaded from GSE31039) and 5mC levels at the corresponding peak regions, represented by red and blue fitting curves, respectively. *Bottom*: Signal intensities of H3K4me3 ChIP-Seq peaks within promoter regions in mESCs and 5hmC levels at the corresponding peak regions, represented by red and blue fitting curves, respectively. (*F*) Snapshot of base-resolution 5hmC maps detected by ACE-Seq (GSE116016) and Cabernet-H, which compared with DIP-Seq (GSE46111) near the Tet1 locus (chr10:62802570-62882014; genome build: mm10). Blue and yellow represent the positive (Watson) and negative (Crick) strands, respectively.

Cabernet-H faithfully captures 5hmC, we examined 5hmC distribution around the binding sites of functional proteins in mESCs. Low levels of 5mCpG and 5hmCpG were observed around the binding sites of CTCF and other methylation-averse DNA-binding proteins, such as H3K4me3 and EP300 (Fig. 2*A* and *SI Appendix,* Fig. S6), consistent with the DNA-binding preferences of these proteins (28, 29). In contrast, 5hmCpG was generally enriched around the center of TET1 binding sites where 5mCpG was depleted (Figs. 1*C* and 2*B*), consistent with findings

that TET oxidizes 5mC to 5hmC. However, overly enriched TET proteins would lower the level of 5hmC in promoter regions (Fig. 2*C* and *SI Appendix,* Fig. S7*A*), in line with TET-mediated iterative oxidization of 5mC to 5hmC, 5fC and 5caC (30–33).

We also validated Cabernet and Cabernet-H based on the relations between 5mC/5hmC modification and gene expression. Consistent with previous reports (20, 34), our Cabernet and Cabernet-H analysis revealed that gene activity is negatively correlated with the 5mC level and positively correlated with the

5hmC level (Fig. 2*D* and *SI Appendix,* Fig. S7*B*). To validate the relations between 5mC/5hmC modification and histone markers, we examined 5mC and 5hmC levels in relation to the H3K4me3 peak signal. 5mC depletion was observed where the H3K4me3 peak signal was high. Intriguingly, 5hmC was depleted at both extreme ends of H3K4me3 peak signal intensity (Fig. 2*E* and *SI Appendix,* Fig. S7*C*). Of note, such analysis suffers a limitation that RNA-seq and ChIP-seq used a different batch of cell line than Cabernet-H.

To further validate the accuracy of 5hmC signals on a specific gene, we examined 5hmC signals on the Watson and Crick strand of the Tet1 gene region captured by Cabernet-H as compared with DIP-seq and ACE-seq. It was found that 5hmC-enriched regions identified with Cabernet-H were broadly aligned with DIP-seq and ACE-seq (Fig. 2*F*). Besides, compared to ACE-seq, strand asymmetry was found in both bulk Cabernet-H and single-cell Cabernet-H, especially in the latter, as previously reported (24). Together, these observations reveal that Cabernet-H faithfully captures 5hmC modifications.

5mC and 5hmC Dynamics in Early Mouse Embryos. Having validated the feasibility and accuracy of Cabernet and Cabernet-H, we then applied them to characterize 5mC/5hmC dynamics in early mouse embryos. A total of 501 individual cells from sperms, oocytes, and pre-implantation embryos of the mouse were analyzed (Dataset S3). An average of 50.4% genomic coverage was achieved for each single cell (Fig. 3*A*). We used Cabernet and Cabernet-H to map 5mC and 5hmC landscapes across different stages during mouse embryonic development. An overall decline in the 5mC level after fertilization was observed, consistent with previous findings (11, 35, 36). The level of 5hmC peaked at the late zygote and late 2-cell stages and then declined afterward (Fig. 3*B*), as reported in previous studies (37–40), revealing dynamics of TET-mediated active demethylation. We subsequently merged single-cell data from each embryonic stage and analyzed the distribution of 5mC and 5hmC across the gene body, which revealed similar dynamics across developmental stages (*SI Appendix,* Fig. S8).

Notably, since the usage of Tn5 transposome for DNA fragmentation results in different insertion sites on different alleles, the aligned reads with the same start and end sites were recognized as from the same allele (*SI Appendix,* Fig. S9). Hemi-modification profiling requires measurement of the methylation status on both two DNA strands of a same allele, and only the alleles with reads sequenced from both two strands are informative of hemi-5mC and hemi-5hmC. At 1X sequencing depths, the reads of these informative alleles account for an average of 1.05% of total reads, and this ratio increases at deeper sequencing depths (*SI Appendix,* Fig. S10).

In this way, hemi-5mC and hemi-5hmC landscapes in mouse embryos were mapped (Fig. 3*C*), which reflects passive demethylation activity (41). The abundance of hemi-5mC sharply increased after fertilization, reflecting a surge in passive demethylation; hemi-5mC abundance gradually elevated after late 2-cell when 5hmC declined (Fig. 3 *B* and *C*), suggesting that passive demethylation becomes dominant after late 2-cell. Hemi-5hmC demonstrates a dynamic pattern similar to that of 5hmC, in agreement with the strand-asymmetry pattern of 5hmC distribution.

Through SNP phasing, parental differences in demethylation were able to be observed (*SI Appendix,* Fig. S11). Demethylation in the paternal genome was faster than that in the maternal genome during early embryo development, consistent with previous studies (11, 35, 42). Studies on zygotes have found that 5hmC-related active demethylation is present in both maternal and paternal genome (38, 39, 43, 44). Similarly, we found that 5hmC abundance was substantially higher in the paternal genome than in the maternal genome, especially during the late zygote and late 2-cell stages (Fig. 3*D*). To investigate whether 5hmC-related active demethylation drives gene expression, we probed the 5hmC signal on the gene body of both active genes and silenced genes. It was found that for both paternal and maternal genome, 5hmC preferably enriched at the gene body of active genes instead of silenced genes (*SI Appendix,* Fig. S12), suggesting that active demethylation is a potential driver for gene activation in both paternal and maternal genome. Moreover, hemi-5mC abundance was higher in the paternal genome (Fig. 3*E*), indicating stronger activity of passive demethylation. However, such apparently strong passive demethylation in the paternal genome appears to have little effect on gene activation, considering that hemi-5mC abundance is similar on active and silenced genes. By contrast, hemi-5mC is enriched only on active genes (*SI Appendix,* Fig. S12), suggesting that passive demethylation is associated with maternal gene activation. Together, these findings suggest different demethylation strategies for regulating paternal and maternal gene expression.

We compared 5mC, hemi-5mC, and 5hmC levels at different genetic regions across different embryonic development stages. It was found that the 5mC level was consistently low at 5'UTR and CpG island; genic regions (promoter, exon, intron, and 3'UTR) and intergenic regions experienced a notable drop in 5mC level and an increase in hemi-5mC (Fig. 3*F*), reflecting pronounced demethylation activity after fertilization, mostly in the form of passive demethylation.

The ability of this method to measure hemi-5mC and 5hmC offers an opportunity for gaining insight into lineage differentiation during preimplantation embryonic development. Compared to 5mC, 5hmC, hemi-5mC, or hemi-5hmC profile alone, integrating 5mC and hemi-5mC profiles from one single cell for t-SNE analysis would generate a more unambiguous clustering map to reveal the lineage during pre-implantation development (Fig. 4 *A–C* and *SI Appendix,* Fig. S13). Unsupervised hierarchical clustering and t-SNE analysis showed that 5hmC is capable of clustering cells at different embryonic stages (Fig. 4*C*). Pearson's correlation analysis revealed that batch-associated variability is much smaller than true biological variability (*SI Appendix,* Fig. S14), ruling out the batch-associated confounding effects.

Based on the profiles of hemi-5mC and 5hmC, different types of cells at different developmental stages can be better classified. For example, pre- and post-replication cells can be more easily distinguished from each other based on the hemi-5mC profile than using the 5mC profile (Fig. 4 *D* and *E* and *SI Appendix,* Fig. S15*A*). Notably, 5hmC and hemi-5hmC profiles can also be used to differentiate pre- and post-replication cells (Fig. 4 *F* and *G* and *SI Appendix,* Fig. S15*B*). Stage-specific 5hmC modification in promoter and gene-body regions was identified (Fig. 4*H* and *SI Appendix,* Fig. S15*C*), indicating potential regulatory functions during embryo development. Together, these suggest the capability of Cabernet and Cabernet-H for hemi-5mC, 5hmC, and hemi-5hmC profiling could be used to better distinguish between cell types and states.

Here, we proved that Cabernet and Cabernet-H can be used to probe 5mC, 5hmC, hemi-5mC, and hemi-5hmC at single-cell single-base resolution with high genomic coverage, providing an efficient tool for the study of biological processes during embryonic development.

High-Throughput Single-Cell DNA Methylation Sequencing with sci-Cabernet. Bisulfite treatment would limit the role of Tn5 due to DNA fragmentation, necessitating subsequent random

Fig. 3. Methylation landscape across each development stage of early mouse embryos. (*A*) Experimental design. Mouse embryos at different development stages were retrieved. Each embryo was separated into single cells for further experiments with Cabernet and Cabernet-H. (*B*) Average 5mC/5hmC level at each developmental stage detected by Cabernet and Cabernet-H, respectively. (*C*) Abundance of DNA hemi-5mC (red) and hemi-5hmC (blue) across developmental stages of mouse early embryos. (*D*) *Top*: Average 5mC level of maternal (red) and paternal (blue) genome at each developmental stage. *Bottom*: Average DNA 5hmC level of maternal (red) and paternal (blue) genome across developmental stages of early mouse embryos. (*E*) *Top*: Abundance of hemi-5mC in maternal and paternal genomes. *Bottom*: Abundance of hemi-5hmC in maternal (red) and paternal (blue) genome. (*F*) *Top*: DNA 5mC level at different genome elements in early mouse embryos. *Middle*: Abundance of hemi-5mC within genome elements in early mouse embryos. *Bottom*: 5hmC level at different genomic regions across developmental stages of early mouse embryos.

priming. By avoiding bisulfite treatment, Cabernet could utilize Tn5 transposome for barcoded transposition, which allows for high-throughput single-cell indexing (sci) (Fig. 5*A*). Our technique, named sci-Cabernet, allows high-throughput single-cell methylation profiling by pooling multiple cells into one library

preparation, enabling both scalability and cost-effectiveness ($1 per cell, *SI Appendix,* Fig. S16A). With this method, we were able to construct the library for thousands of single cells in 2 d by 1 person.

When sequenced to 0.1× average depth, sci-Cabernet could achieve 2.7 ± 0.27% genome coverage (28.2 ± 2.13% of 5-kb bin
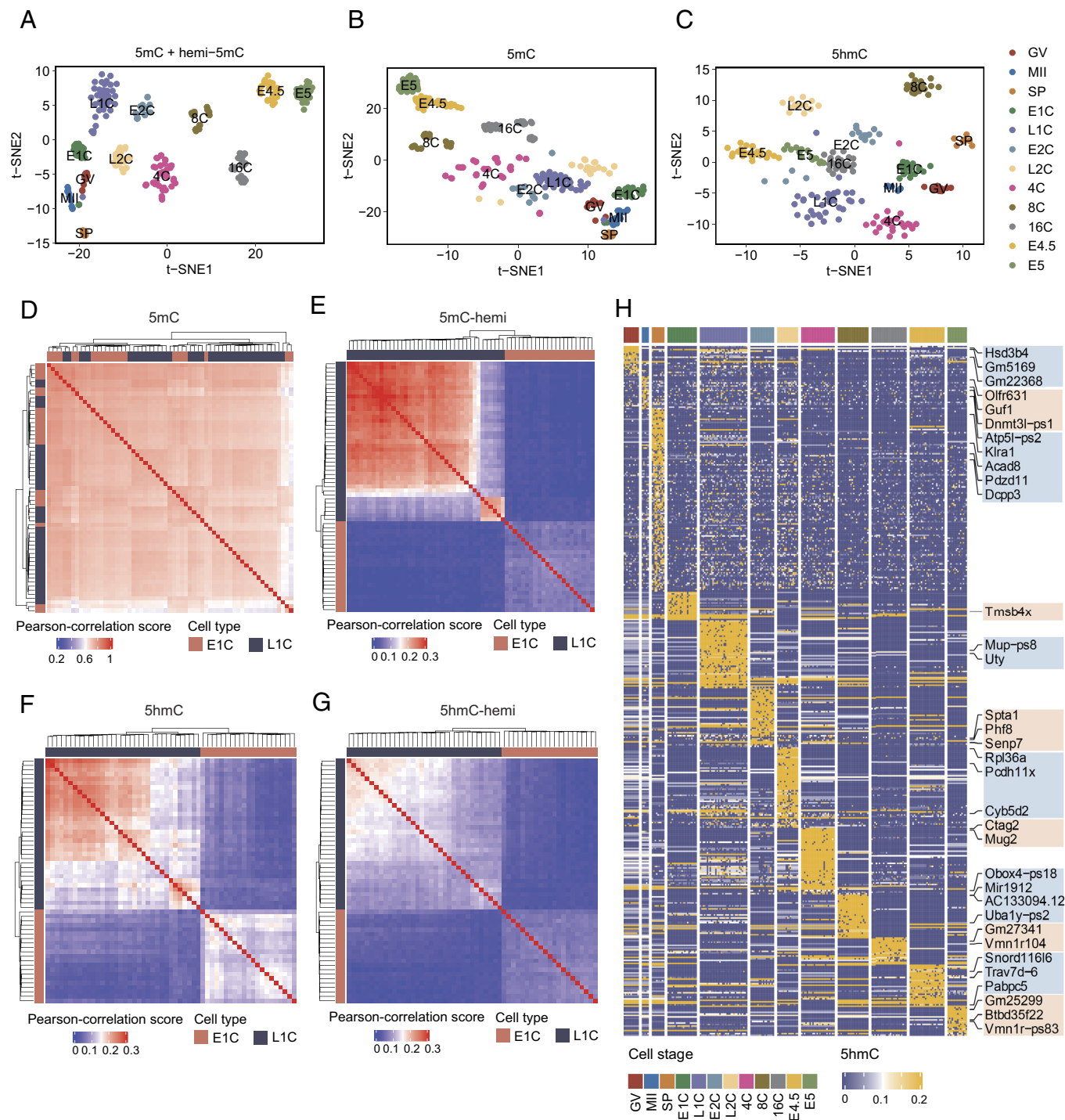
Fig. 4. Key features of DNA methylome in early mouse embryos. (*A*) t-stochastic neighbor embedding (t-SNE) plot of cells based on the level of 5mC and hemi-5mC. Colors indicate different cell stages. (*B*) t-SNE plot of cells based on the level of 5mC. (*C*) t-SNE plot of cells based on the level of 5hmC. (*D*) Clustered heatmap showing Pearson correlation between 5mC abundance of different cells at early-1-cell stage (E1C) and late-1-cell stage (L1C). (*E*) Clustered heatmap showing Pearson correlation between hemi-5mC abundance of different cells at E1C stage and L1C stage. (*F*) Clustered heatmap showing Pearson correlation between 5hmC abundance of different cells at E1C stage and L1C stage. (*G*) Clustered heatmap showing Pearson correlation between hemi-5hmC abundance of different cells at E1C stage and L1C stage. (*H*) Abundance of 5hmC modification on promoter regions at different developmental stages.

coverage), and increasing sequencing depth would lead to higher genomic coverage (Fig. 5*B*). We subsequently merged a few cells from the same population to perform further analysis. Merging 175 single cells can obtain about 75% genomic coverage at the 0.1× genome depth (Fig. 5*C*). Merged sequencing data from as few as 35 single cells broadly matched those obtained from bulk sequencing (Fig. 5*D*), making sci-Cabernet a reliable method for

scalable single-cell DNA methylation sequencing. UMAP (Uniform Manifold Approximation and Projection) analysis suggested that two distinct cell populations as K562 cells and stomach cells could be effectively clustered after being sequenced by sci-Cabernet (Fig. 5*E*).

We further tested the performance of sci-Cabernet on adult mouse cortex and E7.5 mouse embryo cells (Dataset S4). First, a
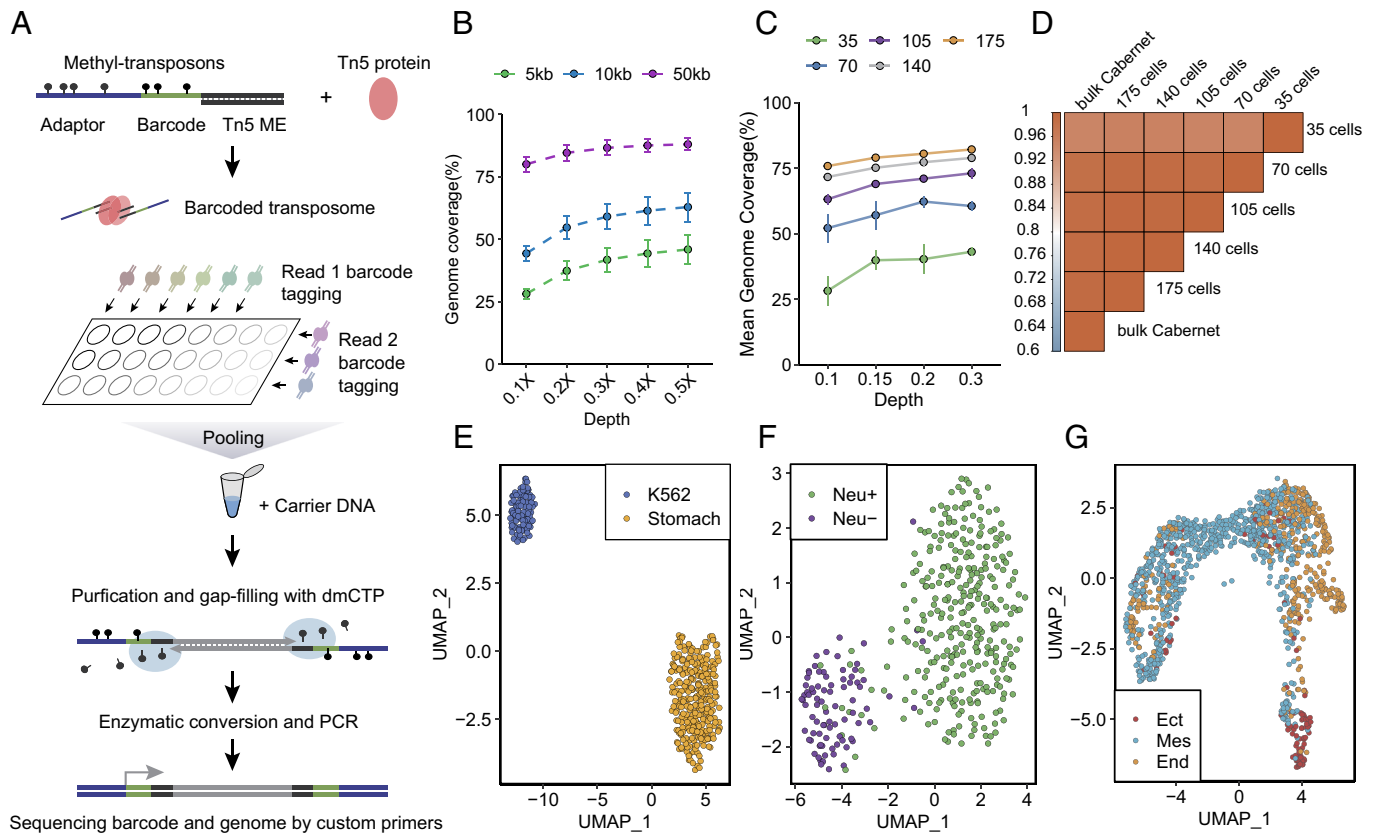
Fig. 5. Cabernet for high-throughput single-cell DNA methylation (sci-Cabernet) profiling. (*A*) Schematic diagram of sci-Cabernet workflow. (*B*) Plot showing the genome coverage of sci-Cabernet at different sequencing depths and different bin sizes in K562 cells. (*C*) Plot showing average genome coverage of sci-Cabernet in K562 cells at different sequencing depths and different cell numbers. (*D*) Heatmap showing the correlation (Pearson correlation) of average 5mC level within 10 kb using different numbers of K562 cells at 0.1× depth. (*E*) Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) figure showing the clustering of K562 cells and stomach cells after sequenced by sci-Cabernet. (*F*) UMAP showing the clustering of mouse cortical neuronal cells and non-neuronal cells after sequenced by sci-Cabernet. (*G*) UMAP visualization of E7.5 mouse embryo cells sequenced by sci-Cabernet and colored by cell-types defined by indirect annotation.

total of 480 neurons (NeuN+) and 96 non-neuronal cells (NeuN–) from the adult mouse cortex were analyzed. Based on single-cell 5mC profiles, neurons were efficiently distinguished from non-neuronal cells at high accuracy (Fig. 5*F*). Second, we also analyzed 5mC profiles of 1248 cells from E7.5 mouse embryos. Based on sci-Cabernet sequencing data, these cells were clustered into four groups (*SI Appendix,* Fig. S16*B*). To determine the cell type of each group, we used Find Transfer Anchors of Seurat to identify the anchors between sci-Cabernet sequencing data and 10× scRNA-seq data of E7.5 (GSE205117) and then transferred the annotation information from the scRNA-seq dataset to sci-Cabernet cells (*SI Appendix,* Fig. S16 *B* and *C*). Thus, scRNA-seq data were integrated with sci-Cabernet sequencing data, and scRNA-seq data were used to determine the subtypes of cells sequenced by sci-Cabernet. Specifically, cluster 0 is dominated by endoderm cells, cluster 1/2 is mostly comprised of mesoderm cells, and cluster 3 is largely composed of ectoderm cells. Although cell types in cluster 1 are slightly confounded, due to limitations of indirect scRNA-seq annotation, it is still obvious that mesoderm cells predominate in cluster 1 (Fig. 5*G*). The endoderm and mesoderm clusters are relatively adjacent to each other compared to the ectoderm cluster, which is aligned with the developmental trajectory of the three cell lineages. Together, these results demonstrate the scalability and accuracy of sci-Cabernet in cell-type determination for complex systems.

Discussion

In this study, we have developed a carrier-assisted base-conversion by enzymatic reaction with end-tagging technology called Cabernet for base-resolution profiling of DNA methylation in single cells. With high-detectability, Cabernet stands out as a promising candidate to be broadly applied to DNA methylation analysis, especially when sample DNA quantities are extremely low such as during early embryonic development. Compared with the bisulfite-based 5mC/5hmC profiling methods, Cabernet uses enzymatic conversion of cytosines in the process, leading to mild reaction condition and better DNA preservation than bisulfite conversion. Compared with other bisulfite-free methods, Cabernet entails optimized purification steps, minimizing the loss of target DNA in single cells whose sample DNA quantities are low in the first place. With these advantages, Cabernet can therefore capture more CpGs at a similar or even lower sequencing depth and better classify different cell types. It is also capable of mapping hemi-5mC/hemi-5hmC profiles based on allele and strand assignment, allowing direct readouts of replication-dependent passive demethylation.

Although bisulfite-based single-cell 5mC mapping methods have been reported in early mouse embryos (11, 35), no previous studies have reported base-resolution mapping of 5hmC in single cells during pre-implantation development. By selectively protecting 5hmC

from deamination, Cabernet-H provides an approach to effectively distinguish 5hmC from 5mC, enabling single-cell base-resolution 5hmC mapping across mouse pre-implantation development. Moreover, integrating 5mC mapping (Cabernet) with 5hmC mapping (Cabernet-H) can obtain the profiles of 5mC, 5hmC, and hemi-modifications at different embryonic stages, which in turn reveals the dynamics of active and passive DNA demethylation during pre-implantation development. In addition, robust allele phasing based on parental SNP enables allele-specific 5mC and 5hmC identification. The findings suggest that TET-mediated active DNA demethylation preferentially targets the paternal over the maternal genome, consistent with several recently reported studies (38, 39). Last, we show that integrative analysis of 5mC and hemi-5mC mapping allows more precise lineage tracing across early embryonic development and that hemi-methylation status can be used to distinguish between the early and late phases of each stage.

In addition to a base-resolution mapping method for DNA methylation at single-cell level, we also provide sci-Cabernet, which employs Tn5 transposome to achieve high-throughput methylome profiling. Given the scalability and cost-effectiveness of sci-Cabernet, thousands of single-cell libraries can be constructed within each preparation, pointing out a reliable approach for deciphering complex biological processes such as neurogenesis and cancer development. Although the resulting genome coverage for each cell is relatively low (*SI Appendix,* Fig. S16D), merging multiple cells would substantially raise coverage (Fig. 5C). The use of enzymatic conversion instead of bisulfite treatment allows better compatibility of Tn5 in Cabernet and Cabernet-H, which could enable multi-omics study that incorporates the profiling of DNA methylation and other modalities such as chromosome accessibility and 3D genome structure.

Despite its scalability, the library quality produced by sci-Cabernet may be limited by unevenness of genome coverage across different cells. The effect of high throughput was achieved by introducing a unique barcode to each individual cell via barcoded Tn5 transposomes. Tn5 transposomes with different barcodes may compete against each other during amplification, resulting in unevenness across cells. Future studies may focus on other means of introducing the barcode, such as barcode ligation.

Materials and Methods

Cell Culture. V6.5 mESCs were cultured under feeder-free conditions in gelatin-coated plates in Dulbecco's Modified Eagle Medium (DMEM) (GIBCO) supplemented with 15% FBS (GIBCO), 2 mM L-glutamine (GIBCO), 0.1 mM 2-mercaptoethanol (Sigma), nonessential amino acids (GIBCO), and 1,000 U/mL LIF (Millipore, ESG1107). The culture was passaged using 0.05% Trypsin (GIBCO) every 2 to 3 d. K562 cells were maintained in Iscove's Modified Dulbecco's Medium (IMDM) (Invitrogen, 12440) with 10% fetal bovine serum, 100 U/mL penicillin, and 100 μg/mL streptomycin (1% P/S). Cells were split 1:8 every 2 to 3 d.

Mouse Embryos Collection. C57BL/6 N females between 6 and 8 wk of age were superovulated by injection of 8 to 10 IU of PMSG, followed by injection of 8 to 10 IU of hCG 48 h later. MII oocytes were collected 14 to 15 h after hCG treatment. To collect early embryos, females were mated with DBA1 males after hCG injection. Early zygotes were collected 14 h after hCG injection and placed in the medium. Different stages of embryos were harvested at 10 h (middle zygote), 16 h (late zygote), 26 h (early 2-cell), 34 h (late 2-cell), 44 h (4-cell), 51 h (8-cell), 61 h (morula), or 76 h (blastocysts) after incubation. The embryos were digested, and single cells were sorted to individual 0.2 mL Maxymum Recovery PCR tubes. Sorted single cells were then lysed in 20 mM Tris pH 8.0, 20 mM NaCl, 0.1% Triton X-100, 15 mM DTT, 1 mM EDTA, 1.5 mg/mL Qiagen protease, 0.5 μM carrier ssDNA, and 0.1 pg spike-in control at 50 °C for 1 h, 65 °C for 1 h, and 70 °C for 15 min. Lysed cells were stored at −80 °C for later use.

Transposome Preparation. The transposons were designed as two versions respectively with 5mC and 5hmC modification (*SI Appendix,* Table S1). Homemade Nextera methyl-transposons have one strand of 5′-/Phos/-CTGTCTCTTATACACATCT-3′ and another strand of 5mC modified P5/P7 adaptor (Me-P5 adaptor or Me-P7 adaptor) (Genewiz, PAGE purification). Nextera hydroxymethyl-transposons have one strand of 5′-/Phos/-CTGTCTCTTATACACATCT-3′ and another strand of 5hmC modified P5/P7 adaptor (HydMe-P5 adaptor or HydMe-P7 adaptor) (Genewiz, PAGE purification). Each of the oligos was dissolved in 0.1× TE to a final concentration of 100 μM. The transposons were annealed by heating the plates to 95 °C for 1 min and then cooling down 0.1 °C per 3 s to reach 25 °C. The transposase was purified after being expressed from the pTXB1-Tn5 plasmid (Addgene). Comparable results can be generated with purchasable Tn5 transposase (Vazyme Biotech Co., Ltd.). Then, the transposons and transposase were dimerized into transposomes at a concentration of 1.25 μM, which was then 1:10 diluted and aliquoted for single uses and stored at −80 °C.

Preparation of Spike-In Control. For accurate identification of 5hmC conversion rate, cytosines in 87 bp DNA fragments were modified with hydroxymethylation by PCR amplification with 5-hydroxymethyl-dCTP supplement (ZYMO). The hydroxy-methylated DNA fragment was combined with equal amounts of CpG methylated pUC19 DNA (NEB) and unmethylated lambda DNA (NEB). CpG-methylated pUC19 DNA and unmethylated lambda DNA were used to measure the conversion rate of 5mC and unmodified C.

scCabernet Library Preparation. Lysate from cells was transposed in a final concentration of 10 mM Tris pH 8.0, 5 mM MgCl$_2$, 8% PEG 8000, 1:2,500 (0.5 nM dimer) homemade Nextera methyl-transposome and incubated at 55 °C for 10 min. Transposases were removed by addition of 1 μL 2 mg/mL Qiagen protease, 1 μL 0.5 M NaCl, and 75 mM EDTA and incubated at 50 °C for 40 min and 70 °C for 15 min. Gaps in the transposed DNA were filled by 4 μL gap-filling mix [3.2 μL Q5 reaction buffer (NEB), 0.07 μL 1 M MgCl2, 0.032 μL 100 mM dATP, 0.032 μL 100 mM dTTP, 0.032 μL 100 mM dGTP, 0.32 μL 10 mM 5-methyl-dCTP (NEB N0356S), 0.16 μL Q5 (NEB M0491S), 0.123 μL 120 ng/μL sonicated carrier lambda DNA] and incubated at 65 °C for 3 min. The product was then purified with 1.8× Ampure XP beads (Beckman Coulter). The DNA with end tags was oxidized in 15 μL TET2 reaction mix [3 μL TET2 reaction buffer plus reconstituted TET2 reaction buffer supplement (NEB), 0.3 μL oxidation supplement (NEB), 0.3 μL oxidation enhancer (NEB), 0.3 μL DTT (NEB), 1.2 μL TET2 (NEB E7125S), 1.5 μL 1:1,249 diluted Fe^{2+} solution (NEB), and 8.4 μL water] at 37 °C for 1 h. Then, TET2 reaction was terminated with 0.3 μL stop solution (NEB) and incubation at 37 °C for 30 min. Afterward, the oxidized DNA was purified with 1.8× Ampure XP beads and eluted in 8 μL elution buffer (NEB). The DNA was denatured by adding 2 μL 0.1 M freshly diluted NaOH and incubated at 50 °C for 10 min, followed by quick cooling with ice. The APOBEC reaction mix [2 μL APOBEC reaction buffer (NEB), 0.2 μL APOBEC (NEB), 0.2 μL BSA (NEB), 1 μL 0.12 M freshly diluted HCl, and 6.6 μL water] was added to the denatured product on ice which was then incubated at 37 °C for 3 h. Whole genome amplification was performed by addition of 20 μL 2× KAPA HiFi HotStart Uracil+ ReadyMix (Roche KK2801), 0.4 μL 100 μM Nextera indexed P5 primer and 0.4 μL 100 μM Nextera indexed P7 primer (*SI Appendix,* Table S2) and incubation at 98 °C for 45 s, followed by 14 cycles of amplification (98 °C for 15 s, 60 °C for 30 s, 72 °C for 1 min), and then 72 °C for 3 min. Size selection was performed with SPRI beads (Beckman Coulter, typically 0.8×) according to the manufacturer's instructions.

scCabernet-H Library Preparation. Lysate was transposed in a final concentration of 10 mM Tris pH 8.0, 5 mM MgCl$_2$, 8% PEG 8000, 1:2,500 (0.5 nM dimer) homemade Nextera hydroxymethyl-transposome and incubation at 55 °C for 10 min. Transposases were removed by the addition of 1 μL 2 mg/mL Qiagen protease, 1 μL 0.5 M NaCl, and 75 mM EDTA and incubation at 50 °C for 40 min and 70 °C for 15 min. Transposed DNA was gap-filled by addition of 4 μL gap-fill mix [3.2 μL Q5 reaction buffer (NEB), 0.07 μL 1 M MgCl2, 0.08 μL 40 mM dATP, 0.08 μL 40 mM dTTP, 0.08 μL 40 mM dGTP, 0.08 μL 40 mM 5-hydroxymethyl-dCTP (ZYMO), 0.16 μL Q5 (NEB M0491S), 0.12 μL 120 ng/μL sonicated carrier lambda DNA, 0.13 μL water] and incubation at 65 °C for 3 min. The product was purified with 1.8× Ampure XP beads (Beckman Coulter). The end-tagged DNA was oxidized in 15 μL BGT reaction mix [0.3 μL uridine diphosphate glucose (NEB), 1.5 μL NEBuffer 4 (NEB), 0.75 μL T4 Phage β-glucosyltransferase (NEB M0357S), and 12.45 μL water] at 37 °C for 2 h. BGT reaction was terminated by addition of 0.75 μL

Proteinase K (NEB P8107S) and incubation at 37 °C for 30 min. The product was purified with 1.8× Ampure XP beads and eluted in 8 μL elution buffer (NEB). The DNA was denatured by the addition of 2 μL 0.1 M freshly diluted NaOH and incubation at 50 °C for 10 min, followed by quick cooling with ice. The APOBEC reaction mix [2 μL APOBEC reaction buffer (NEB), 0.4 μL APOBEC (NEB), 0.2 μL BSA (NEB), 1 μL 0.12 M fresh diluted HCl, and 6.4 μL water] was added to the denatured product on ice which was then incubated at 37 °C for 3 h. Whole genome amplification was performed by the addition of 20 μL 2× KAPA HiFi HotStart Uracil+ ReadyMix (Roche KK2801), 0.4 μL 100 μM Nextera P5 index primer and 0.4 μL 100 μM Nextera P7 index primer and incubation at 98 °C for 45 s, 14 cycles of amplification (98 °C for 15 s, 60 °C for 30 s, 72 °C for 1 min), and 72 °C for 3 min. Size selection was performed with SPRI beads (Beckman Coulter, typically 0.8×) according to the manufacturer's guidelines.

Bulk Cabernet Library Preparation. For bulk samples, V6.5 mESCs and K562 genomic DNA were purified from 1 million cells using the QIAamp micro kit (QIAGEN), according to the manufacturer's instructions. Then, 3.75 ng of purified genomic DNA was used for bulk Cabernet and Cabernet-H library preparation. Bulk libraries were prepared according to the protocol above, with more transposition enzymes and fewer PCR amplification cycles.

Sci-Cabernet Library Preparation. Lysate in a 96-well plate was transposed in a final concentration of 10 mM Tris pH 8.0, 5 mM MgCl₂, 8% PEG 8000, 1:3,333(0.75 nM monomer) homemade indexed P5 Nextera methyl-transposome, and 1:3,333 (0.75 nM monomer) homemade indexed P7 Nextera methyl-transposome (together forming 12 × 8 unique dual-index, *SI Appendix*, Table S3), and incubation at 55 °C for 10 min. Transposases were removed by adding 1 μL 2 mg/mL Qiagen protease, 1 μL 0.5 M NaCl, 75 mM EDTA and incubation at 50 °C for 40 mi n, then 70 °C for 15 min. Contents in all the wells were collected and pooled together. Next, 200 ng sonicated carrier lambda DNA was added. The product was then purified with 2× 20% diluted Ampure XP beads (Beckman Coulter, diluted with PEG buffer composed of 19% PEG, 2.5 M NaCl, 10 mM Tris pH 8.0, 1 mM EDTA, and 0.5% Tween 20) and eluted by 20 μL 1 mM Tris pH 8.0. Gaps in the transposed DNA were filled by 15 μL gap-fill mix [7 μL Q5 reaction buffer (NEB), 7 μL Q5 high GC enhancer (NEB), 0.07 μL 100 mM dATP, 0.07 μL 100 mM dTTP, 0.07 μL 100 mM dGTP, 0.7 μL 10 mM 5-methyl-dCTP (NEB N0356S), 0.35 μL Q5 (NEB M0491S)] and incubated at 65 °C for 3 min. The product was purified with 1.8× SPRIselect beads (Beckman Coulter) and eluted with 15 μL 1 mM Tris pH 8.0. The DNA with end tags was oxidated by adding 15 μL TET2 reaction mix [6 μL TET2 reaction buffer plus reconstituted TET2 reaction buffer supplement (NEB), 0.6 μL oxidation supplement (NEB), 0.6 μL oxidation enhancer (NEB), 0.6 μL DTT (NEB), 2.4 μL TET2 (NEB E7125S), 3 μL 1:1,249 diluted Fe²⁺ solution (NEB)] at 37 °C for 1 h. TET2 reaction was terminated by the addition of 0.6 μL stop solution (NEB) and incubation at 37 °C for 30 min. The product was purified with 1.8× SPRIselect beads and eluted in 13.5 μL 1 mM Tris pH 8.0. The DNA was denatured by the addition of 1.5 μL 0.2 M freshly diluted NaOH and incubation at 50 °C for 10 min, followed by quick cooling with ice. The APOBEC reaction mix [3 μL APOBEC reaction buffer (NEB), 0.3 μL APOBEC (NEB), 0.3 μL BSA (NEB), 1.5 μL 0.12 M fresh diluted HCl, and 9.9 μL water] was added to the denatured product on ice which was then incubated at 37 °C for 3 h. Pre-amplification was performed by the addition of 30 μL 2× Q5U Master Mix (NEB M0597S), 0.4 μL 100 μM pre-amplification forward primer and 0.4 μL 100 μM pre-amplification reverse primer (*SI Appendix*, Table S4) and incubation at 98 °C for 30 s, eight cycles of amplification (98 °C for 10 s, 62 °C for 30 s, 65 °C for 90 s), and 65 °C for 5 min. Excessive primers were digested with the addition of 0.5 μL Exonuclease I (NEB M0293S), and incubation at 37 °C for 30 min and 72 °C for 20 min. Library construction was completed by adding 15 μL water, 15 μL 2× Q5U Master Mix (NEB M0597S), 0.4 μL 100 μM Nextera P5 index primer and 0.4 μL 100 μM Nextera P7 index primer, and incubation at 98 °C for 30 s, four cycles of amplification (98 °C for 10 s, 62 °C for 30 s, 65 °C for 90 s), and 65 °C for 5 min. Purification was then done with DNA Clean & Concentrator-5 (ZYMO D4013) according to the manufacturer's guidelines. Size selection was performed with Select-a-Size DNA Clean & Concentrator (ZYMO D4080) according to the manufacturer's instructions.

Deep Sequencing of DNA Library. PCR-amplified DNA libraries were quantified by a Qubit fluorometer (Life Technologies). The libraries were pooled and quantified between the range of 150 bp and 800 bp using a Fragment Analyzer (Agilent).

Library pools were sequenced on an Illumina HiSeq X Ten System, loaded at 0.9 pM, with a 10% PhiX spike-in to improve library complexity. Notably, sci-Cabernet libraries were sequenced with custom sequencing primers (*SI Appendix*, Table S5).

Reference Genome and Genomic Features. The ENCODE hg38 human reference genome and mm10 mouse reference genome were used in this study. Genomic blacklist regions were downloaded from ENCODE (hg38: NCFF419RSJ, mm10: ENCFF547MET). GENCODE M18 and M35 were used to annotate mouse and human genome. CGI annotation was downloaded from the UCSC genome browser. Promoters are defined as regions from −2 kb to +0.2 kb relative to the transcriptional start site (TSS). The mm10 mouse genome and hg38 human genome annotated by RepeatMasker were downloaded from the UCSC genome browser.

Downloaded RNA and ChIP-Seq Data. A full list of the public datasets used in this study is provided in Dataset S5. The readcounts file of RNA-Seq in mESC (GSE153575) was downloaded from the GEO database, and then, we calculated the TPM (Transcripts Per Kilobase Million) of each gene for further analysis. The bed and bigwig file of ChIP-Seq were downloaded from ENCODE, the other ChIP-Seq data (*Mus musculus*) which were not found in ENCODE were downloaded from GEO. First, the raw reads were processed with Trim_galore, and the cleaned data were then aligned to the reference genome; the unique mapped reads were used for further analysis. Then, we used MACS2 for narrowpeak calling with default parameters. For ATAC-Seq data (GSE116854, GSE136403, GSE66581, and GSE143658), after obtaining unique mapped reads, we used MACS2 to capture open chromatin regions (macs2 callpeak -t -n -shift -75 -extsize 150 -nomodel -B -SPMR -g mm -keep-dup all).

BS-Seq Sequencing Data Processing. Low-quality bases, short length, and adaptor sequences in all single-cell BS-seq and bulk BS-seq data were trimmed off using TrimGalore-0.4.5 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, parameters: –fastqc –paired –phred33 –retain_unpaired –clip_R1 9 –clip_R2 9). Cleaned reads were mapped to the hg38 human reference genome by Bismark with the following options: "–fastq –non_directional –unmapped –nucleotide_coverage." Reads that mapped to multiple locations were removed and PCR duplicate reads were removed with picard.jar (v-2.4.1). Methylation levels were called using Bismark methylation extractor.

Preprocessing of Cabernet and Cabernet-H Sequencing Data. Low-quality bases, short length and adaptor sequences in raw sequencing reads were trimmed off using TrimGalore-0.4.5 with default parameters (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). High-quality reads were then aligned to the reference genome (hg38 and mm10) by Bismark (v.0.19.0) (parameters: –fastq –non_directional –unmapped –nucleotide_coverage). Unmapped reads and those located in genome blacklist regions were excluded. Tn5 insertion caused 9-bp gap regions then filled by methylated cytosine was removed after alignment. In this way, we were able to reduce the reads cleaning rate and improve the alignment rate. And at the same time, the start and end points of the original fragments were marked. Fragments from the same allele but a different strand can be retained in the next step when removing PCR duplication reads by picard.jar (v.2.4.1). Methylation levels were called using Bismark methylation extractor.

Preprocessing of sci-Cabernet Sequencing Data. Before trimming and alignment, reads of different cells should be separated from each other and barcodes and Tn5 sequences are removed in the meantime. Before separation, the barcode and Tn5 were merged to form a longer barcode sequence (referred to as barcode sequence) to improve the pick-up rate. Calculate the hamming distance between the 5′ end of the read and each barcode sequence to identify the most similar barcode indicating a specific cell. If a read has a distance of over 8 from all barcodes, then it should be filtered off. At the same time, identify the Tn5 19 bp insertion site and remove the barcode sequence. Then, after the library of each single cell is established, process the single-cell reads with the above pipeline described under "Preprocessing of Cabernet sequencing data."

5hmC Methylation Calling. To call methylated sites, we counted the number of reads that supported methylation at a certain site and the number of reads that did not. These counts were then used to perform a binomial test with a probability of success equal to the conversion rate, which was determined by the fraction of

methylated reads in the lambda genome (spike-in during library construction). The false discovery rate (FDR) for a given P-value cutoff was computed using the Benjamini–Hochberg approach.

Calling of Hemi-5mC/Hemi-5hmC Abundance at CpGs. The use of Tn5 transposome for DNA fragmentation would result in different insertion sites on different alleles. If two reads are aligned to the same starting and ending sites on the reference genome, then they are regarded as from the same allele. First, select those alleles that have reads sequenced from both the Watson strand and the Crick strand. Then, measure the methylation status (5mC/5hmC) at CpG dinucleotides on both strands of the same allele, and calculate the number of hemi-methylated CpGs (hemi-5mCpGs, CpGs wherein one strand has a methylated C and the other does not) and hemi-hydroxymethylated CpGs (hemi-5hmCpGs, CpGs wherein one strand has a hydroxymethylated C and the other does not). Hemi-5mC/hemi-5hmC abundance was calculated as the ratio of hemi-5mCpGs/ hemi-5hmCpGs to the total number of CpGs detected on two strands.

Parental Phasing. The reference SNPs of C57BL/6 N and DBA/1 J were downloaded from the Mouse Genomes Project (REL-1505-SNPs_Indels) which can be accessed at https://www.sanger.ac.uk/data/mouse-genomes-project/. Only the SNPs that could distinguish the paternal (DBA/1 J) and maternal (C57BL/6 N) genomes were used in our analysis. Considering that the C->T transition cannot distinguish the allele of C and T, so in the case of C→T transition, remove the C and T appearing on the same allele. The rest of SNPs in the genome were used to classify the reads into paternal and maternal reads. The same also applies to the G→A transition. Only those SNPs with base quality >30 are included in the analysis. When the difference value (D-Values) between the ratio of SNPs on a read belonging to paternal SNPs and the ratio belonging to maternal SNPs is over 0.2, designate the reads as maternal or paternal. The aligned reads were further processed to calculate the methylation levels at CpG sites.

Correlation between Cabernet and scBS-Seq Datasets. The average methylation levels within 10-kb windows were calculated by deepTools (45) multiBigwigSummary. To profile the relationship between the results from scBS-seq and Cabernet method, the Pearson correlation coefficient (PCC) was calculated directly from the average 5mC/5hmC level within 10-kb windows of the K562 cell line detected by scCabernet, bulk Cabernet, merged scCabernet, scBS-seq and

scBS-seq merged datasets. At the same time, PCC was also calculated between scCabernet-H, bulk Cabernet-H and merged scCabernet-H with mESC cells.

Methylation Distribution across Gene Body and ChIP-Peak Center. The average methylation levels around genic regions (from 2 kb upstream of TSS through the gene body to 0.5 kb downstream of the transcription end sites) were computed for each gene in each dataset by computeMatrix scale-regions. Similarly, the average methylation levels around the ChIP-peak regions, which is between 2 kb upstream and downstream of the peak center, were computed by the computeMatrix reference-point. Values obtained from cells of the same stage or cell types were combined to plot the average levels around the genomic and ChIP-peak regions.

Identifying Highly Methylated Regions and Dimension Reduction Clustering. The average 5mC/5hmC levels within promoter regions (−2 kb, TSS, +0.2 kb) and gene body regions (from TSS to TES) were computed for each dataset by the computeMatrix BED-file as methylation features. Seurat packages with the FindAllMarkers function were used to find the specific highly methylated regions during each stage. The heat map of these highly methylated regions was drawn by the ComplexHeatmap (46). The top 8,000 features calculated by the Seurat FindVariableFeatures function were used to perform dimension reduction clustering by Rtsne (47).

Correlation between Cell Stages. Pearson correlation coefficient (PCC) was calculated between cells of E1C and L1C based on methylation levels on promoter regions. First, calculate the average methylation (5mC/5hmC/hemi-5mC/hemi-5hmC) levels at promoter regions (−2 kb, TSS, +0.2 kb) for each cell of zygote stage (E1C and L1C). Then, calculate PCC between each two different cells and draw the heat map using the ComplexHeatmap.

Data, Materials, and Software Availability. Raw sequencing data have been deposited in GSA (CRA005812 and HRA001740) (48, 49). Code is available via GitHub (https://github.com/yali-bai/CabernetMethylPip) (50).

1. P. A. Jones, Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
2. Z. D. Smith, A. Meissner, DNA methylation: Roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
3. I. C. Macaulay, T. Voet, Single cell genomics: Advances and future perspectives. *PLoS Genet.* **10**, e1004126 (2014).
4. S. A. Smallwood et al., Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat. Genet.* **43**, 811–814 (2011).
5. Z. D. Smith et al., A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339–344 (2012).
6. Z. D. Smith et al., DNA methylation dynamics of the human preimplantation embryo. *Nature* **511**, 611–615 (2014).
7. H. Guo et al., The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014).
8. S. J. Cokus et al., Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
9. R. Lister et al., Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523–536 (2008).
10. E. E. Holmes et al., Performance evaluation of kits for bisulfite-conversion of DNA from tissues, cell lines, FFPE tissues, aspirates, lavages, effusions, plasma, serum, and urine. *PLoS One* **9**, e93933 (2014).
11. H. Guo 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).
12. S. A. Smallwood et al., Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
13. M. Farlik et al., Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.* **10**, 1386–1397 (2015).
14. C. Angermueller et al., Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).
15. S. Bian et al., Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science* **362**, 1060–1063 (2018).
16. C. Luo et al., Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
17. C. Luo et al., Robust single-cell DNA methylome profiling with snmC-seq2. *Nat. Commun.* **9**, 3824 (2018).
18. Y. Liu et al., Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.* **37**, 424–429 (2019).
19. Z. Sun et al., Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome Res.* **31**, 291–300 (2021).
20. C. X. Song et al., Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.* **29**, 68–72 (2011).
21. K. Williams et al., TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343–348 (2011).
22. M. J. Booth et al., Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
23. E. K. Schutsky et al., Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat. Biotechnol.* **36**, 1083–1090 (2018).
24. M. Yu et al., Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
25. J. Füllgrabe et al., Simultaneous sequencing of genetic and epigenetic bases in DNA. *Nat. Biotechnol.* **41**, 1457–1464 (2023), 10.1038/s41587-022-01652-0.
26. R. M. Mulqueen et al., Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
27. L. Shen et al., Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692–706 (2013).
28. H. Wang et al., Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22**, 1680–1688 (2012).
29. M. Weber et al., Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39**, 457–466 (2007).
30. Y. F. He et al., Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
31. S. Ito et al., Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129–1133 (2010).
32. S. Ito et al., Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
33. M. Tahiliani et al., Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
34. G. Ficz et al., Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398–402 (2011).
35. F. Guo et al., Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* **27**, 967–988 (2017).

36. Y. Zhang *et al.*, Dynamic epigenomic landscapes during early lineage specification in mouse embryos. *Nat. Genet.* **50**, 96–105 (2018).
37. T. P. Gu *et al.*, The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477**, 606–610 (2011).
38. K. Iqbal, S. G. Jin, G. P. Pfeifer, P. E. Szabó, Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 3642–3647 (2011).
39. A. Inoue, Y. Zhang, Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* **334**, 194 (2011).
40. H. Wu, Y. Zhang, Reversing DNA methylation: Mechanisms, genomics, and biological functions. *Cell* **156**, 45–68 (2014).
41. M. Sen *et al.*, Strand-specific single-cell methylomics reveals distinct modes of DNA demethylation dynamics during early mammalian development. *Nat. Commun.* **12**, 1286 (2021).
42. L. Wang *et al.*, Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**, 979–991 (2014).
43. L. Shen *et al.*, Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. *Cell Stem Cell* **15**, 459–471 (2014).
44. F. Guo *et al.*, Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell* **15**, 447–459 (2014).
45. F. Ramírez *et al.*, deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
46. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics (Oxford, England)* **32**, 2847–2849 (2016).
47. L. J. P. van der Maaten, G. E. Hinton, Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
48. Y. Cao, A high coverage bisulfite-free single-cell methylome sequencing method. National Genomics Data Center (NGDC). https://download.cncb.ac.cn/gsa2/CRA005812/. Deposited 24 December 2021.
49. Y. Cao, A high coverage bisulfite-free single-cell methylome sequencing method. National Genomics Data Center (NGDC). https://download.cncb.ac.cn/gsa-human/HRA001740/. Deposited 24 December 2021.
50. Y. Bai, CabernetMethylPip. GitHub. https://github.com/yali-bai/CabernetMethylPip. Deposited 27 February 2023.