Base-Resolution and Single-Cell Analysis of Active DNA Demethylation Using Methylase-Assisted Bisulfite Sequencing

Base-Resolution and Single-Cell Analysis of Active DNA Demethylation Using Methylase-Assisted Bisulfite Sequencing

A dissertation presented

by

Xiaoji Wu

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

March 2018

Dissertation Advisor: Professor Yi Zhang Xiaoji Wu

Base-Resolution and Single-Cell Analysis of Active DNA Demethylation Using Methylase-Assisted Bisulfite Sequencing

Abstract

In mammals, DNA methylation in the form of 5-methylcytosine (5mC) can be actively reversed to unmodified cytosine through ten-eleven translocation (TET) dioxygenase-mediated oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), followed by replication-dependent dilution or base excision repair. This process, known as active DNA demethylation, is present in many biological contexts including development and diseases. To investigate the mechanism and function of active DNA demethylation, various methods have been developed to analyze the genomic distribution of the demethylation intermediates 5hmC, 5fC and 5caC. However, previous methods suffer from limitations including low resolution, non-specificity and incompatibility with single-cell studies. To overcome these limitations, I developed methylase-assisted bisulfite sequencing (MAB-seq), a method capable of profiling 5fC and 5caC (5fC/5caC) at single-base resolution and to the single-cell level.

In my dissertation, I present the development of MAB-seq method, its applications in different biological contexts, and the insights revealed by the studies. I first established MAB-seq and applied it to mouse embryonic stem cells, uncovering unique features of active DNA demethylation including the processivity of TET-mediated oxidation and its

correlation with chromatin accessibility. I then further modified the protocol to make it compatible with low-input and single-cell studies, termed liMAB-seq and scMAB-seq, respectively. To demonstrate the application of liMAB-seq, I analyzed paternal genome demethylation in mouse zygotes, revealing the dynamics of 5mC/5hmC and 5fC/5caC at individual genomic regions. To demonstrate the utility of scMAB-seq, I applied the method to perform cell type classification, genomic mapping of sister chromatid exchange and reconstruction of cellular lineage during preimplantation development.

Taken together, this work not only establishes novel methods for studying active DNA demethylation at base-resolution and single-cell level, but also demonstrates the utilities of these methods in different biological contexts and provides insights into the demethylation process.

Table of Contents

Acknowledgement

My PhD has been a rewarding journey full of memorable moments. Though 6 years have passed, I still clearly remember the excitement when I passed my PhD Qualifying Exam, when my papers got accepted by top journals, and when I received the first job offer in my life from a top management consulting firm. I would not have accomplished these without the support from my mentors, colleagues, friends and families.

To my PhD advisor Dr. Yi Zhang. Thank you for guiding me along this journey and supporting me without reservation, for inspiring me with your love for science and showing me how to conduct first-class research, and for creating a perfect environment with the best resource and the best people. I will never forget your encouragement to focus on significant topics, and to do things that I am good at and enjoy doing. Though there are always uncertainties ahead, I will keep exploring with your suggestions in mind. Thank you!

To my mentor in the Zhang Lab, Dr. Hao Wu. When I started in the lab five years ago, I was really lucky to have the opportunity to work side by side with you, one of the best scientists I have ever met, on a great topic that translated into multiple publications. I always admire your wisdom, passion, and incredible patience to help other people out. Good luck with your career at UPenn. I cannot wait to see more of your great work!

To former and current Zhang Lab members, Li, Azusa, Tsukasa, Renchao, Lan, Falong, Yuting, Nadhir, Luis, Xudong, Qiangzong, Zhiyuan, Chunxia and all others. It has been

an honor for me to work with such a talented and supportive group of people. Li and Falong, thank you for being the go-to guys for any question I have. Azusa and Tsukasa, I will never accomplish the sc/liMAB-seq project without your hard work. Renchao and Lan, I am proud of our joint efforts in the hypothalamus project. Yuting and Nadhir, thank you for making my life easier by sharing your codes with me. Xudong and Zhiyuan, I would not be able to enjoy my vacations without you watching over my cells. Special thanks to our lab manager Jennifer Ballew for relieving lots of administrative work from us and making our lab life interesting. Thank you all!

In addition to Zhang Lab members, many people have helped me in various ways during my 6-year academic career at Harvard. Drs. Matthew Waldor and Peter Sicinski kindly offered me the opportunities to rotate in their labs, showing me what top-notch research in microbiology and cancer biology looks like. Drs. Simon Ringgaard and Wojciech Michowski mentored me during my rotations, and they are truly the best mentors one can imagine. I would also like to thank Drs. Raul Mostoslavsky, Rosalind Segal and Danesh Moazed for serving as my PQE committee, Drs. Evan Rosen, Jesse Gray, Alexander Meissner and Danesh Moazed for serving on my DAC, and Drs. Jesse Gray, Mitzi Kuroda, Mary Gehring and Eric Greer for serving as my Dissertation Defense Exam Committee. Thank you for taking your precious time to discuss my projects with me and providing insightful suggestions.

To two of my best friends providing me with continuous support during all these years, Xiuyuan Li and Sizun Jiang. To Xiuyuan, our friendship has never faded since we first

met in high school. You show me how to live an interesting and meaningful life, and I wish you always maintain that spirit. To Sizun, you never cease to amaze me with your smartness, passion, and your sense of humor. More importantly, you are someone that I can count on. Let us work together to turn our dreams into reality!

To my fellow PhD friends at Harvard, Ximei, Yijie, Zecai, Shiwei, Sining, Sharon, Chen-Hao, Qin, Kevin, Charles, and many others, I will always remember the drinks and fun we had, and I have no doubt that you will all do great in the career path you choose. To my friends from PKUAA Dragon Boat Team, Yunan, Yunfei, Jingyu, Mingxing, Guangyan, Hong, Wei, Mingxin, Ruosi, Xingchao, Xiao, Lijun, Chunmei, An, Xiaowei, Zhen and many others, being with you is like living in a big, warm family, and I will never forget the incredible races, trips and parties we enjoyed together. Special thanks to Wenyue. Thank you for making my life colorful and inspiring me to be a better self.

Finally, to mom and dad, I cannot say enough to thank you for your love and support during my entire life. Our time spent together has been so limited in the past six years, but knowing that you always have my back is enough for me to carry on this journey. I dedicate this dissertation to you, and I hope that I have accomplished something that you are proud of.

Chapter 1:

Introduction

Establishment and reversal of DNA methylation

In mammals, 5-methylcytosine (5mC) is the major form of DNA modification with important roles in development and diseases[1,2]. About 60–80% of the CpG sites in the mammalian genome are modified by 5mC[1]. The major functions of 5mC include mediating genomic imprinting and X-chromosome-inactivation, repressing transposable elements and regulating transcription[3].

5mC is both chemically and genetically stable. Chemically, the methyl group is connected to the 5-position of cytosine base through a stable carbon–carbon bond, creating a barrier for direct removal of the methyl group. Genetically, upon its establishment by the de novo DNA methyltransferases DNMT3A and DNMT3B[4], 5mC is maintained by the maintenance methyltransferase DNMT1, which recognizes hemi-methylated CpG dyads through its functional partner UHRF1[5-7]. This maintenance mechanism is crucial as it ensures faithful re-establishment of 5mC on the newly synthesized strand after DNA replication[8,9].

Despite its stability, mammalian 5mC can still be reversed to its unmodified state in two ways. First, a lack of functional DNA methylation maintenance machinery can result in dilution of 5mC during DNA replication, a process known as passive DNA demethylation. Passive DNA demethylation is only present in certain biological contexts, including preimplantation development and primordial germ cell (PGC) development[10]. Second, ten-eleven translocation (TET) dioxygenases can mediate the iterative oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine

(5caC), and replication-dependent dilution of these oxidized forms of 5mC or thymine DNA glycosylase (TDG)-mediated excision of 5fC and 5caC coupled with base excision repair (BER) will also result in demethylation (**Figure 1.1**)[10]. This process, known as active DNA demethylation, is the focus of this work.

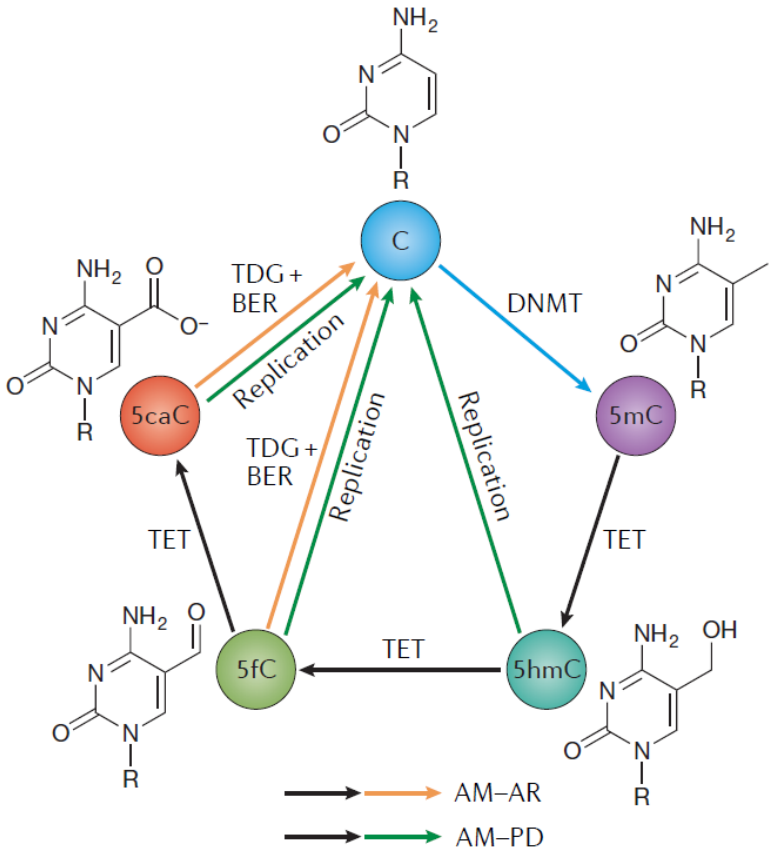

Figure 1.1 | The cycle of active DNA demethylation. DNMTs convert unmodified cytosine to 5mC. 5mC can be converted back to unmodified cytosine by TET-mediated oxidation to 5hmC, 5fC and 5caC, followed by TDG-mediated excision of 5fC or 5caC coupled with BER (a process known as active modification - active removal (AM–AR)), or replication-dependent dilution of 5hmC, 5fC or 5caC (a process known as active modification - passive dilution (AM–PD).

Discovery of the TET-TDG pathway

In mammals, earlier work suggests that 5mC can be rapidly erased from the genome in a manner that cannot be fully explained by replication-dependent dilution[11-13]. In 2009, two groundbreaking papers discovered that 5hmC accumulates to a significant level in certain mouse tissues[14], and that human TET1 is capable of converting 5mC to 5hmC[15]. The other two members of the TET family, TET2 and TET3, also possess 5mC to 5hmC oxidizing activity[16]. Further studies suggest that TET proteins also catalyze the oxidization of 5hmC to 5fC and 5caC[17,18]. Mechanistically, TET proteins are members of Iron(II)/α-ketoglutarate (Fe(II)/α-KG)-dependent dioxygenases, and the oxidation reactions require oxygen and α-KG as substrates and Fe(II) as a cofactor to generate CO_2 and succinate[19,20].

Following TET-mediated oxidation, the restoration of unmodified cytosine can be achieved in two ways. First, after 5mC is oxidized to 5fC or 5caC, TDG-mediated excision of 5fC or 5caC and BER-dependent repair of the abasic site can restore unmodified cytosine[17,21,22]. This process is defined as active modification - active removal (AM–AR) and is independent of DNA replication (**Figure 1.1**)[10,19].

In addition to AM–AR, DNA replication can also lead to the dilution of the oxidized 5mC, a process known as active modification - passive dilution (AM–PD) (**Figure 1.1**)[10,19]. During DNA replication, unmodified cytosine is incorporated into the newly synthesized strand, creating hemi-modified CpG dyads. Unlike 5mC:C dyad, 5hmC:C, 5fC:C and 5caC:C cannot be efficiently processed by UHRF1-DNMT1 machinery, resulting in a lack

of methylation maintenance at these sites[23-25]. Through multiple rounds of DNA replication, a 5hmC-, 5fC- or 5caC-modified CpG site can become demethylated at cell population level.

Tissue distribution of oxidized forms of 5mC

The amount of oxidized 5mC in the genome can reflect the extent of TET-mediated oxidation. Mass spectrometry analyses suggest that unlike 5mC, the levels of oxidized 5mC are highly variable in different tissues[18,26-30]. In adult mice, 5hmC is present at a high level in the central nervous system (CNS)[14,18,29,30]. For example, in mouse cerebellar Purkinje neurons, 5hmC abundance is nearly 40% of that of 5mC[14]. Some somatic tissues such as kidney and heart have medium levels of 5hmC (25–50% of that of CNS tissues), whereas some others, such as spleen and thymus, have low levels of 5hmC (5–15% of that of CNS tissues)[18,28,29]. 5hmC is also present in embryonic tissues. For example, in mouse embryonic stem cells (ESC), the amount of 5hmC is about 1.3×10^3 in every 10^6 cytosines, a level comparable to that of non-CNS somatic tissues[18].

5fC and 5caC are much less abundant compared with 5hmC, either because TDG is highly efficient at removing these two bases, or because the conversion of 5hmC to 5fC and 5caC is less efficient, or both. In wild-type mouse ESCs, the amount of 5fC and 5caC is about 20 and 3 in every 10^6 cytosines, respectively[18]. Consistent with efficient removal by TDG, depletion of TDG in mouse ESCs results in a 5.6-fold increase in 5fC and 8.4-fold increase in 5caC levels[31]. Beyond ESCs, 5fC can be readily detected in various somatic tissues in postnatal mice[18,27].

In addition to mass spectrometry analyses, immunostaining using antibodies against oxidized 5mC has also revealed TET-mediated oxidation in several biological contexts, including zygotes[32-34] and PGCs[35].

Genomic profiling of oxidized forms of 5mC

One way to understand the mechanism and function of active DNA demethylation is to profile the genomic distribution of the demethylation intermediates 5hmC, 5fC and 5caC. In recent years, various methods have been developed to serve this purpose, and can be separated into base-resolution and non-base-resolution categories.

Non-base-resolution methods generally involve affinity enrichment of oxidized 5mC followed by high-throughput sequencing. In the case of DNA-immunoprecipitation (DIP), 5-hydroxymethylcytosine (5hmC)[36-40], 5-formylcytosine (5fC)[31] and 5-carboxylcytosine (5caC)[31] can be enriched using antibodies directly recognizing these modifications. In other cases, e.g. CMS-seq (5hmC)[41], GLIB-seq (5hmC)[41], hMe-Seal (5hmC)[42] and fC-Seal (5fC)[43], 5fC-DP (5fC)[44,45], oxidized 5mC can be converted to other forms that allow enrichment. These methods capture DNA fragments containing the oxidized 5mC but do not determine where in the fragment the modification is, hence they provide low spatial resolution. Other caveats and disadvantages include potential contaminants of pulldown, CpG-density-related bias and lack of absolute quantification.

Most base-resolution methods are based on bisulfite sequencing (BS-seq) (**Figure 1.2**)[46].

After sodium bisulfite treatment, 5mC and 5hmC are sequenced as cytosine whereas unmodified cytosine (unmodified C), 5fC and 5caC are sequenced as thymine. The abundance of 5mC and 5hmC (5mC/5hmC) at a CpG site or genomic region can thus be quantified as (C/(T+C))%[17,47,48]. Altering the behavior of these different forms of cytosine during bisulfite conversion allows the decoding of individual modifications, either directly as in the case of TAB-seq (5hmC)[49], MAB-seq (5fC+5caC)[46,50-52] and caMAB-seq (5caC)[46] (the latter two methods will be the focus of this work), or indirectly after comparison with regular BS-seq data as in the case of oxBS-seq (5mC and 5hmC)[48], redBS-seq (5fC)[53], fCAB-seq (5fC)[43,54] and caCAB-seq (5caC)[55,56].

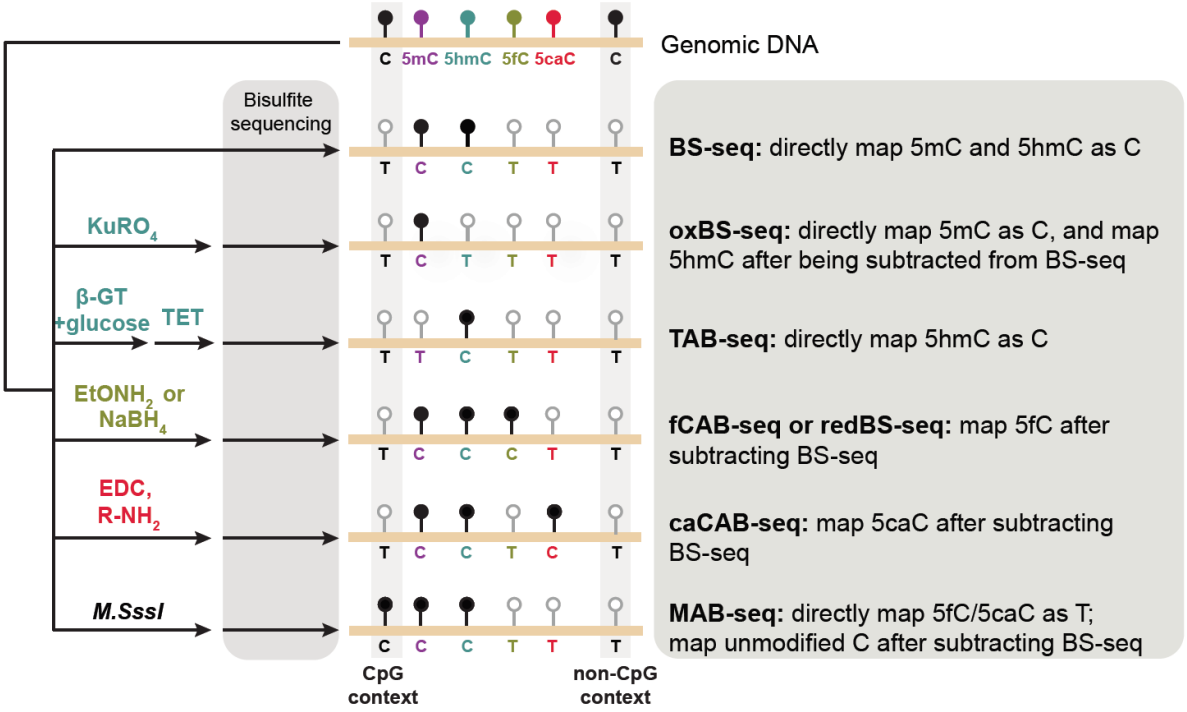

Figure 1.2 | Base-resolution methods for profiling oxidized forms of 5mC. Several BS-seq-based methods have been developed to map 5mC (BS-seq), 5hmC (oxBS-seq and TAB-seq), 5fC (fCAB-seq and redBS-seq), and 5caC (caCAB-seq).

Base-resolution and single-cell analysis of 5fC and 5caC by methylase-assisted bisulfite sequencing

Though 5hmC, 5fC and 5caC are all products of TET-mediated oxidation, they may have distinct biological meanings. 5hmC cannot be directly processed by TDG-BER machinery and can be relatively stable if further oxidation by TET and dilution from DNA replication are limited (**Figure 1.1**)[28]. Comparatively, 5fC and 5caC, being efficiently excised by TDG in most biological contexts, are better markers of ongoing DNA demethylation[31,43,46]. In fact, the accumulation of 5fC or 5caC upon TDG depletion can support the existence of ongoing TET-mediated oxidation and TDG-mediated excision[17,31,43-45,57]. As a result, to better understand the mechanism and function of active DNA demethylation, it will be necessary to perform genomic profiling of 5fC and 5caC.

In 2013 when this work started, two papers reported the genome-wide distribution of 5fC and 5caC in mouse ESCs using affinity enrichment-based methods[31,43]. TDG depletion in mouse ESCs, either through knockdown or knockout, results in accumulation of 5fC and 5caC at distal regulatory elements and bivalent promoters, supporting ongoing DNA demethylation activity at these regions[31,43]. However, as discussed above, enrichment-based methods have limitations including low genomic resolution (typically a few hundred base pairs), lack of quantification and potential non-specificity.

Because no existing methods at that time were capable of profiling 5fC and 5caC at single-base resolution, at genome-wide scale and in a quantitative manner, we started to develop a novel method named methylase-assisted bisulfite sequencing (MAB-seq) to

achieve these goals. Compared with regular BS-seq for profiling 5mC/5hmC, MAB-seq has one extra step of treatment of DNA by *M.SssI*, a bacterial DNA methyltransferase, resulting in conversion of unmodified C in CpG context to 5mC before bisulfite treatment. As a result, C/5mC/5hmC are sequenced as C in a MAB-seq experiment while 5fC/5caC are sequenced as T (**Figure 1.2**).

After the successful development of MAB-seq and a proof-of-principle study in mouse ESCs[46], we decided to further modify the original protocol (termed regular MAB-seq, using ~1μg purified genomic DNA as starting material) to make it compatible with studies using small numbers of cells or single cells. The resulting methods, termed low-input MAB-seq (liMAB-seq; starting from ~100 cells) and single-cell MAB-seq (scMAB-seq), significantly broaden the application of MAB-seq, because many biological contexts with active DNA demethylation only have limited number of cells for analysis. In addition, scMAB-seq also enables other applications such as high-resolution mapping of sister chromatid exchanges (SCE) and cellular lineage reconstruction.

Specific Aims

Aim 1: Development of MAB-seq and base-resolution analysis of active DNA demethylation in mouse ESCs

To analyze active DNA demethylation at high resolution and in a quantitative manner, Aim 1 of this work is to develop a sequencing method capable of mapping 5fC and 5caC (5fC/5caC) at single-base resolution. In Chapter 2, I discuss how we combined BS-seq with *M.SssI* methylase treatment to achieve this goal, and how we used various technical

and biological controls to validate our method, termed MAB-seq. By applying MAB-seq to mouse ESCs, we revealed several novel features of active DNA demethylation, for example the processivity of TET-mediated oxidation and its correlation with chromatin accessibility[10,46]. By combining MAB-seq with sodium borohydride ($NaBH_4$) treatment, we also developed a method named caMAB-seq, capable of profiling 5caC alone (**Figure 1.2**).

Aim 2: Development of liMAB-seq and scMAB-seq

The original version of MAB-seq, as discussed in Aim 1 and Chapter 2, requires ~1μg DNA (corresponding to ~0.2 million cells) to begin with, making it incompatible with studies using a small number of cells or single cells. A low-input protocol capable of analyzing a few hundred cells is crucial, because many biological contexts with active DNA demethylation, for example zygotes and PGCs, only have a limited amount of material for analysis. Single-cell protocol is also needed, given that the heterogeneity of active DNA demethylation is not well understood. To broaden the application scope of MAB-seq, Aim 2 is to develop low-input and single-cell versions of this method. In Chapter 3 of this work, I introduce how we established one-tube protocols for MAB-seq to minimize material loss during library preparation process and achieve the goal of profiling 5fC/5caC using ~100 cells (liMAB-seq) or single cells (scMAB-seq)[58].

Aim 3: liMAB-seq analysis of paternal genome demethylation in mouse preimplantation embryos

With the successful development of liMAB-seq in Aim 2, Aim 3 is to apply this method to

analyze 5fC/5caC dynamics during paternal genome demethylation of mouse preimplantation embryos. This process is the first example of active DNA demethylation but was not fully understood previously due to limited sample availability. Two aspects of this process were examined by liMAB-seq, and the results are presented in Chapter 4: 1) the relationship between 5fC/5caC changes and 5mC/5hmC changes, and 2) the relationship between TET processivity and chromatin accessibility[58].

Aim 4: scMAB-seq analysis of cellular heterogeneity and SCE

Aim 4 is to demonstrate the utility of scMAB-seq by testing whether this method can capture cell-to-cell heterogeneity of active DNA demethylation, and whether single-cell analysis can reveal information that is not captured by analyzing a population of cells. In Chapter 5, I present scMAB-seq analysis of single mouse ESCs and single blastomeres from mouse 2-cell-stage and 4-cell-stage embryos. scMAB-seq successfully captured the heterogeneity of 5fC/5caC distribution between different cell types and among single cells of the same cell type. scMAB-seq also captured the strand-biased distribution of 5fC/5caC resulted from DNA replication, confirming that 5fC/5caC are not directly maintained during DNA replication. Using the strand bias of 5fC/5caC, we can infer the genomic locations of sister chromatid exchange (SCE), a type of genomic rearrangement associated with genomic instability. This information can be further utilized to perform lineage reconstruction during mouse preimplantation development[58].

Chapter 2:

Base-resolution analysis of active DNA demethylation in mouse embryonic stem cells using methylase-assisted bisulfite sequencing (MAB-seq)

Abstract

Active DNA demethylation in mammals can be achieved through TET-mediated iterative oxidation of 5mC to 5hmC, 5fC and 5caC, followed by TDG-mediated removal of 5fC and 5caC (5fC/5caC). However, methods for quantitatively analyzing the generation and excision of 5fC/5caC at single-base resolution are lacking. In Chapter 2, I describe methylase-assisted bisulfite sequencing (MAB-seq), a method capable of directly mapping 5fC/5caC at base-resolution and in a quantitative manner. Genome-wide MAB-seq allows systematic identification of 5fC/5caC in TDG-depleted mouse embryonic stem cells, thereby generating a base-resolution map of active DNA demethylome. A comparison of 5fC/5caC and 5hmC distribution maps indicates that TET-mediated oxidation has distinct processivity at different CpG sites, and local chromatin accessibility positively correlates with TET processivity. MAB-seq also reveals strong strand asymmetry of active demethylation within palindromic CpG sites. Integrating MAB-seq with other base-resolution mapping methods allows separate quantification of 5fC and 5caC.

Background

In mammals, active DNA demethylation is achieved through TET-mediated oxidation of 5mC to 5hmC, 5fC and 5caC, followed by replication-dependent dilution of 5hmC/5fC/5caC (active modification - passive dilution (AM-PD) pathway) or TDG-mediated removal of 5fC/5caC (active modification - active removal (AM-AR) pathway)[10]. The AM-AR pathway involving generation and excision of 5fC/5caC is of particular interest, because it may take place in a wide range of somatic cell types including postmitotic cells. Interestingly, only TDG, but not other members of the uracil DNA glycosylase superfamily, possesses robust 5fC/5caC excision activity and is indispensable for embryonic development, implicating AM-AR pathway in regulating tissue-specific gene expression and development[59,60].

The observation that 5hmC can accumulate to a relatively high level in diverse cell types, particularly in adult neurons, raises the possibility that TET proteins tend to stall at 5hmC and that further oxidation of 5hmC to 5fC/5caC is a rate-limiting step[10]. Therefore, identifying cytosines that are committed to active DNA demethylation requires methods that permit quantitative measurement of TDG-mediated excision of 5fC/5caC at high resolution. Previous profiling studies using either 5fC/5caC-specific antibodies or chemical tagging of 5fC showed that TDG depletion results in ectopic accumulation of 5fC and 5caC at specific genomic regions such as distal regulatory elements, suggesting that these regions are undergoing TET/TDG-mediated demethylation. However, 5fC and 5caC maps generated by these affinity enrichment-based methods have low spatial resolution (several hundred base-pairs (bp)), represent only relative enrichment and lack

strand distribution information[31,43]. Though several base-resolution methods for mapping 5fC or 5caC have been described (**Figure 1.2**), they all require subtraction of BS-seq signals to indirectly map 5fC or 5caC, which significantly increases sequencing effort and introduces additional variation.

To address these limitations, we have developed a novel method named MAB-seq, which allows quantitative mapping of 5fC/5caC at single-base resolution. Furthermore, our genome-wide MAB-seq analysis of mouse embryonic stem cells (ESCs) has provided new insights into the processivity and strand asymmetry of TET-mediated oxidation.

Results

Experimental strategy and validation of MAB-seq

Recognizing the limitations of affinity enrichment-based methods[31,43] and several previously developed base-resolution methods (**Figure 1.2**)[53,56,61], we developed MAB-seq, an approach that aims to achieve direct and simultaneous measurement of 5fC and 5caC at single-base resolution (**Figure 2.1**). In standard BS-seq, C/5fC/5caC reacts with sodium bisulfite and are efficiently deaminated to uracil (C/5fC) or 5caU (5caC), both of which are sequenced as thymine (T), whereas 5mC and 5hmC are resistant to this chemical conversion and sequenced as C (**Figure 2.1**). In MAB-seq, genomic DNA is first treated with the bacterial DNA CpG methyltransferase *M.SssI*, an enzyme that is originally isolated from *Spiroplasma sp.* strain MQ1 and is known to efficiently methylate cytosines within CpG dinucleotides[62]. Bisulfite conversion of *M.SssI*-treated DNA may therefore only deaminate 5fC and 5caC; originally unmodified C within CpGs is protected as 5mC. Subsequent sequencing would reveal 5fC and 5caC as T, whereas C/5mC/5hmC would be sequenced as C (**Figure 2.1**). Notably, MAB-seq is unable to distinguish 5fC/5caC from unmodified C within a non-CpG context due to the poor activity of *M.SssI* toward C within a non-CpG context. This limitation does not affect the application of this technique, because TET-mediated oxidation predominantly happened in the CpG context based on genomic analysis[49,63] and structural analysis[64,65].

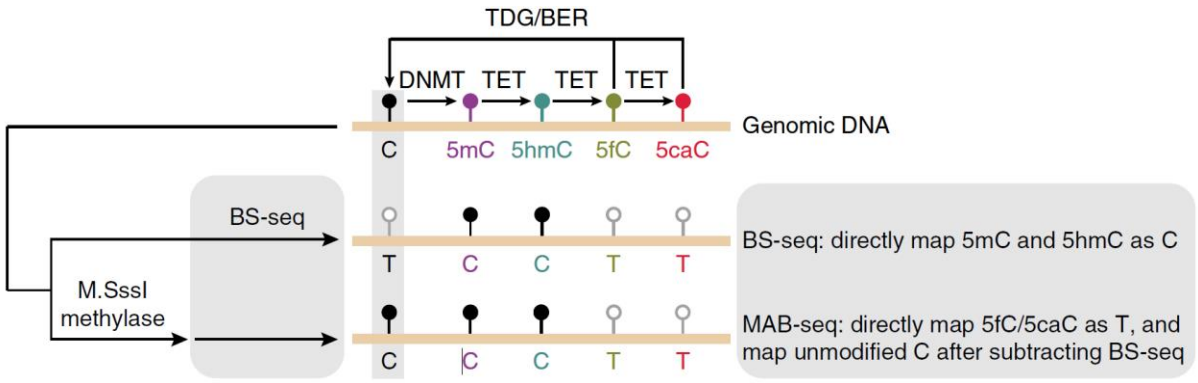

Figure 2.1 | Schematic diagram of MAB-seq. In BS-seq, unmodified C, 5fC and 5caC are sequenced as T after bisulfite conversion. In MAB-seq, an additional step of *M.SssI* treatment converts unmodified C to 5mC before bisulfite conversion, allowing 5fC and 5caC to be sequenced as T. The level of unmodified C can also be quantified when MAB-seq and BS-seq results are combined.

Successful detection of 5fC/5caC using MAB-seq requires complete conversion of C to 5mC by *M.SssI*. We identified an optimal *M.SssI* reaction condition and performed MAB-seq analysis of the unmethylated lambda phage genome (6,224 CpGs within 48,502 bp and free of any DNA modification). In MAB-seq, 97.96% of the originally unmodified CpGs are sequenced as C, in contrast to a nearly complete C-to-T conversion in a regular BS-seq (0.13% sequenced as C). We analyzed the sequences immediately flanking 67 CpG sites that are not efficiently methylated by *M.SssI*, and found that they are not associated with any specific sequences (**Supplemental Figure S2.1a**), suggesting that *M.SssI* has minimal sequence preference for catalyzing CpG methylation reactions. Next, we performed BS-seq and MAB-seq analysis of synthetic double-stranded DNA (dsDNA) containing CpG sites with specific cytosine modifications (5hmC/5fC/5caC), and confirmed that unmodified C in asymmetrically modified CpG dyads (5hmC/5fC/5caC:C)

can also be efficiently methylated by *M.SssI* methylase (**Supplemental Figure S2.1b**). Therefore, unmodified C, either in completely unmodified CpG dyad or hemi-modified CpG dyad, can be protected by *M.SssI* treatment.

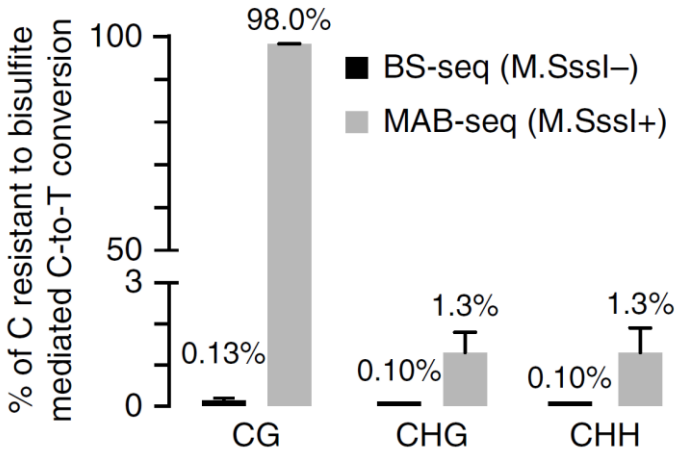

Figure 2.2 | Efficient methylation of unmodified C in CG context by *M.SssI* methylase. *M.SssI* methylase activity was measured by MAB-seq analysis of unmethylated lambda DNA. Standard BS-seq confirmed nearly complete conversion of unmethylated C to T at CpG and non-CpG sites (99.9%). Comparatively in MAB-seq, only 2.04% of unmodified C at CpG sites are read as T. Error bars, s.e.m.

In addition to efficient protection of unmodified C, efficient bisulfite conversion of 5fC/5caC is also important. Consistent with previous reports[53], our analysis showed that 5fC (84.7%) and 5caC (99.5%), but not 5hmC (3.3%), are efficiently deaminated by bisulfite treatment and read as T (**Supplementary Figure S2.1c**).

Locus-specific MAB-seq analysis of selected genomic loci in mouse ESCs

To test MAB-seq in analyzing mammalian genomic DNA, we applied this method to examine four 5fC/5caC-enriched loci (*Tbx5*, *Vps26a*, *Ace* and *Slc2a12*) that were

previously identified in TDG-depleted mouse ESCs by affinity enrichment-based methods (**Figure 2.3**, **Supplemental Figure S2.2**). Using mouse ESCs deficient for both TET1 and TET2 (Tet1/2 DKO; largely absent of 5fC/5caC) as a negative control to correct background signals[66], locus-specific MAB-seq can detect and quantify 5fC/5caC at these loci. Importantly, TDG depletion (shTdg) results in accumulation of 5fC/5caC, demonstrating that the 5fC/5caC signals detected are specific. In addition, treatment of TDG-depleted cells by vitamin C (VC), a cofactor capable of boosting TET catalytic activity[67-69], resulting in a further increase of 5fC/5caC, again confirming that the MAB-seq signals we observed are real 5fC/5caC (**Figure 2.3**, **Supplementary Figure S2.2**). Finally, similar 5fC/5caC distribution patterns were detected for biologically independent replicates, demonstrating the reproducibility and robustness of MAB-seq (**Supplemental Figure S2.3**).

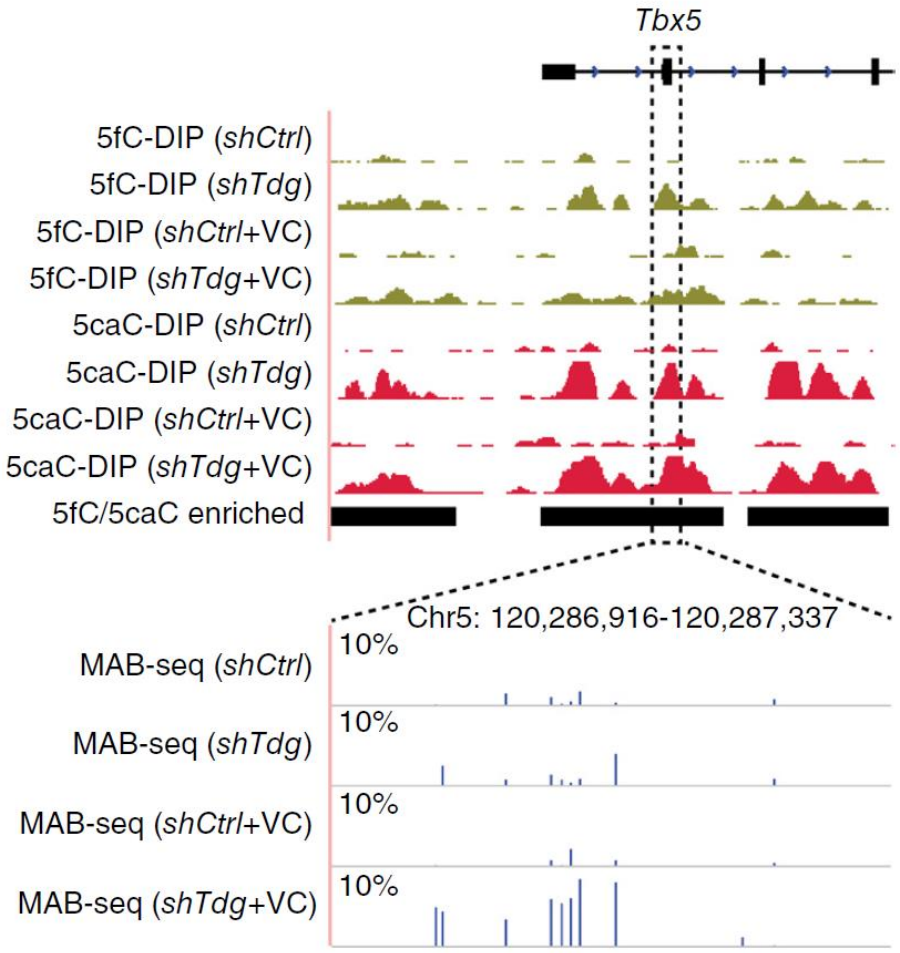

Figure 2.3 | Locus-specific MAB-seq analysis of 5fC/5caC at *Tbx5* locus in mouse ESCs. Bottom panel shows 5fC/5caC quantified by MAB-seq. The level of 5fC/5caC (only Watson strand shown) is displayed as the percentage of total C modified as 5fC/5caC, and background signals detected in Tet1/2−/− mouse ESCs were subtracted. For comparison, upper panel shows 5fC/5caC peaks identified by DIP-seq. DIP-seq tracks are represented in normalized read density (reads per 10 million reads) and the vertical axis range of all DIP-seq tracks is from 1 to 25. Black horizontal bars denote 5fC/5caC-enriched peaks called by DIP-seq. shCtrl: control cells; shTdg: TDG-depleted cells; +VC: with vitamin C treatment.

Genome-scale MAB-seq analysis of mouse ESCs

Having validated MAB-seq using locus-specific analysis, we next applied MAB-seq to identify cytosines undergoing active DNA demethylation at the genome scale. We performed whole-genome MAB-seq (WG-MAB-seq) of TDG-depleted mouse ESCs (shTdg) cultured in the presence of VC, and sequenced the sample to an average depth of 28.4× per CpG dyad (covering 95.2% of all CpG dyads). Using mouse ESCs deficient in DNMT enzymes (Dnmt1/3a/3b TKO) or TET1/2 proteins (Tet1/2 DKO) as a negative control for estimating empirical false discovery rate (FDR), we identified a total of 675,325 5fC/5caC-modified CpGs (out of 24,872,637 CpGs (N ≥ 10)) in shTdg+VC mouse ESCs with an empirical FDR of 5%. 5fC/5caC-modified CpGs in TDG-depleted cells correlate well with 5fC/5caC-enriched regions identified by DNA immunoprecipitation (DIP) approach (region 1 in **Figure 2.4**). Compared to random controls, 5fC/5caC-modified CpGs significantly overlapped with 5fC/5caC DIP-seq peaks (7.9 times as many as expected by chance, Z-score = 579.3). Furthermore, 83.6% of DIP-seq peaks (n = 50,923 out of 60,912 covered by WG-MAB-seq) overlapped with at least one 5fC/5caC-modified CpGs. By contrast, as exemplified by region 2 in **Figure 2.4**, 80.7% of 5fC/5caC-modified CpGs are outside 5fC/5caC DIP-seq peaks (region 2 in **Figure 2.4**), suggesting that MAB-seq has a markedly increased sensitivity compared with DIP-seq.

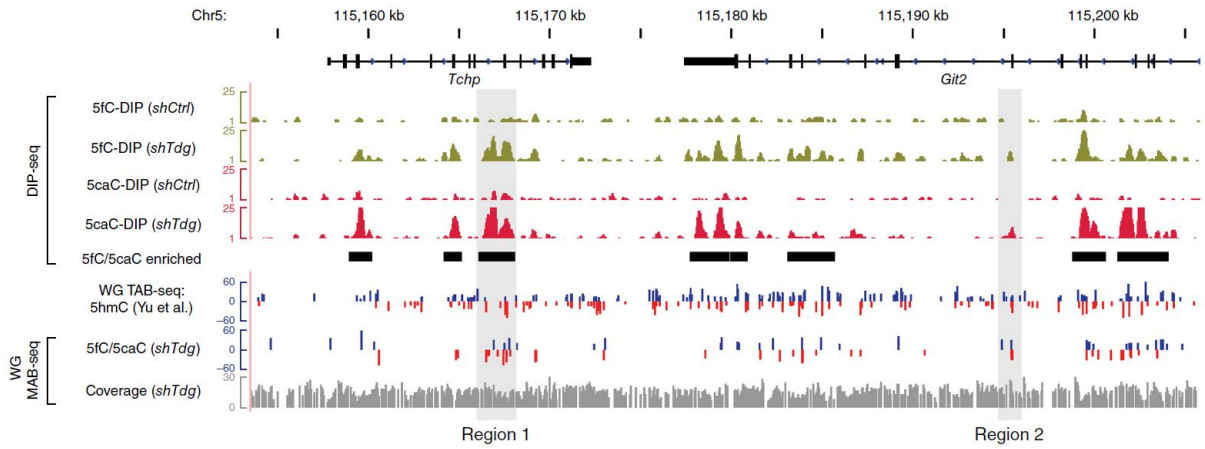

Figure 2.4 | Genome-scale MAB-seq analysis of mouse ESCs. Shown is a snapshot of Tchp-Git2 locus with affinity enrichment-based 5fC/5caC maps (DIP-seq of wild-type (shCtrl) or TDG-depleted (shTdg) mouse ESCs), base-resolution 5hmC map (TAB-seq of wild-type ESCs, published dataset by Yu et al.; see methods) and base-resolution 5fC/5caC maps (MAB-seq of shTdg+VC cells). For comparison, 5fC/5caC-enriched regions (shTdg-specific) identified by DIP-seq methods are highlighted by black horizontal bars. For base-resolution maps, positive values (blue) indicate cytosines on the Watson strand (positive strand), whereas negative values (red) indicate cytosines on the Crick strand (negative strand). For base-resolution maps of 5hmC and 5fC/5caC, the vertical axis limits are −60% to +60%. CpGs associated with statistically significant level of 5hmC (FDR = 5%) and 5fC/5caC (FDR = 5%) are shown. Sequencing coverage for WG-MAB-seq experiments is shown in gray.

Genomic distribution of active DNA demethylation activity

Previous studies using affinity enrichment-based methods have shown that 5fC/5caC are enriched at enhancers, bivalent promoters, and gene bodies[31,43]. Consistently, 5fC/5caC-modified CpG sites identified by MAB-seq are more enriched at DNase I hypersensitive

sites (DHS), ESC-specific enhancers, pluripotency factor binding sites (Oct4, Nanog and Sox2), H3K4me1-modified regions and 5fC/5caC DIP-seq peaks (**Figure 2.5;** see methods for published datasets).

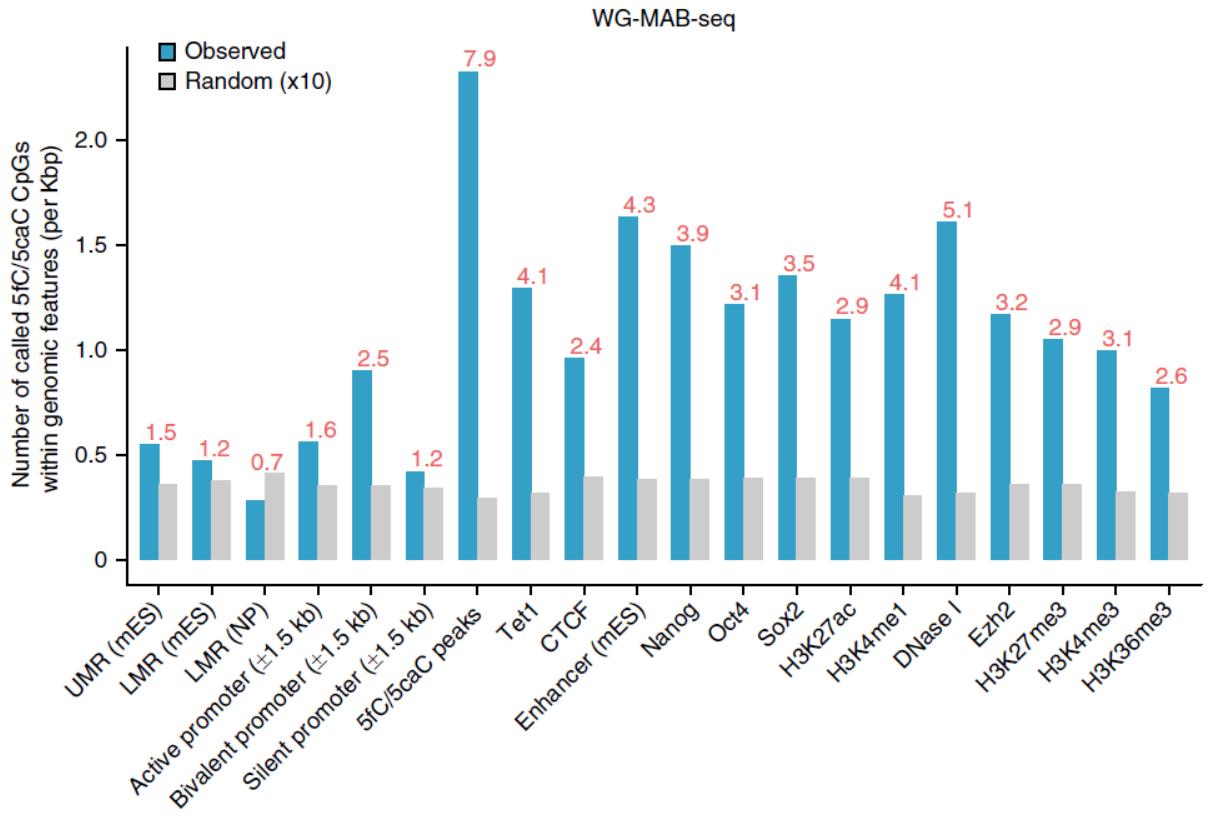

Figure 2.5 | Enrichment of 5fC/5caC-modified CpG sites in different genomic features. The relative enrichment of 5fC/5caC-modified CpGs at specific gene regulatory regions (blue) and corresponding randomly shuffled control regions (gray) is shown. 5fC/5caC peaks: 5fC/5caC-enriched regions identified by DIP-seq; UMR: unmethylated region, typically marking transcriptionally active CpG-rich promoters; LMR: low methylation region; mES: mouse ESCs; NP: neural progenitors.

Distinct processivity of TET enzymes at different CpGs

TET-mediated demethylation involves three iterative oxidation reactions (5mC → 5hmC → 5fC → 5caC). Due to low spatial resolution and lack of quantification, affinity enrichment-based profiling cannot determine whether 5mC at a CpG site tend to undergo two or three oxidation reactions to reach 5fC/5caC, or instead only undergo one reaction before stalling at 5hmC. In other words, the processivity of TET enzymes (the ability to further oxidize 5hmC to 5fC/5caC) was unknown[10]. The capability of MAB-seq in identifying individual CpG sites oxidized to 5fC/5caC offers an unique opportunity for investigating TET processivity.

To address this question, we focused on CpG sites sufficiently covered by both our 5fC/5caC map (of shTdg+VC cells) and published base-resolution 5hmC map (generated by TAB-seq[49]) (**Figure 2.6**). Comparative analysis of TAB-seq and WG-MAB-seq results (depth $\geqslant$10) allows us to estimate the number of CpG sites associated with 5fC/5caC-alone (5hmC−, 5fC/5caC+; n = 508,261), 5hmC-alone (5hmC+, 5fC/5caC−; n = 1,454,388), and both (5hmC+, 5fC/5caC+; n = 117,327). Notably, 5hmC and 5fC/5caC largely exist at different CpG sites (**Figure 2.6a**). Only 7.5% of 5hmC-modified CpGs (depth $\geqslant$ 10) also had significant levels of 5fC/5caC (corresponding to 18.8% of all 5fC/5caC-modified CpGs, FDR = 5%), whereas the majority of 5hmC-modified CpGs appeared to represent 5hmC stably accumulated without being further oxidized by TET proteins to 5fC/5caC.

Given that 5fC/5caC are preferentially enriched at active and/or poised gene regulatory

regions where chromatin is generally more accessible, we reasoned that local chromatin structure may influence TET processivity. We first analyzed DNase I hypersensitivity (measured by DNase-seq[70]) at genomic regions immediately flanking 5fC/5caC-alone, 5hmC-alone and dual modified CpGs (±50 bp). This analysis reveals that 5fC/5caC-alone and dual-modified CpGs are associated with markedly higher level of DNase I hypersensitivity signals than 5hmC-alone CpGs (**Figure 2.6b**). In addition, 5fC/5caC-modified CpG sites are also associated with higher levels of benzonase (BNase) sensitivity signals[71], histone variants (H2A.Z and H3.3) known to destabilize nucleosome structure[71,72], occupancy of TET1[73], and binding of pluripotency-related transcription factors[74] (**Figure 2.6b-c**). These results suggest that TET-mediated oxidation tend to stall at the 5hmC step at most CpGs, but exhibit higher processivity at CpGs associated with a more accessible chromatin state.

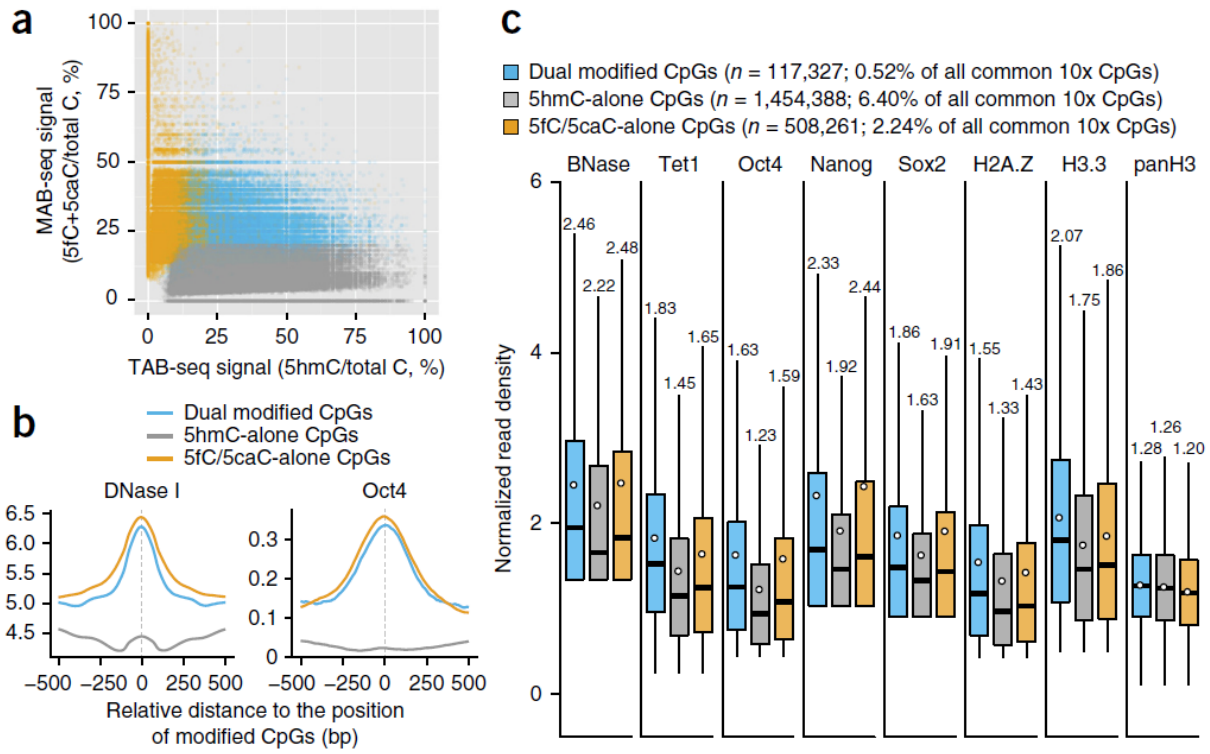

Figure 2.6 | Identification of CpG sites with distinct TET processivity. (a) Comparative analysis of base-resolution 5fC/5caC and 5hmC maps identifies CpGs associated only with 5fC/5caC (5fC/5caC-alone; n = 508261, FDR = 5%), 5hmC (5hmC-alone; n = 1,454,388, FDR = 5%) and both (dual modified with 5hmC and 5fC/5caC; n = 117,327, FDR = 5%). (b) Averaged read density of DNase I hypersensitivity signals and Oct4 ChIP-seq around 5hmC-alone, 5fC/5caC-alone and dual-modified CpGs (±500 bp). (c) The levels of chromatin accessibility signals (BNase), TET1 occupancy, pluripotency transcriptional factors (Nanog, Oct4 and Sox2) and histone variants (H2A.Z and H3.3) around the genomic position (±50 bp) of indicated groups of modified CpGs. The black bars and white circles in boxplots denote median and mean of normalized read density (reads per 10 million reads). The mean of each boxplot is shown on the top.

Strand asymmetry of 5fC/5caC at palindromic CpG dyads

Although cytosine methylation in palindromic CpG dyads is generally symmetric and exhibits very high heritability upon DNA replication[3], previous studies suggest that >80% of steady-state 5hmC in human and mouse ESCs are asymmetrically modified[49]. This prompted us to examine whether 5fC/5caC-modified CpGs also show strand asymmetry. Focusing on CpG dyads with both strands sufficiently covered by WG-MAB-seq (n = 9,261,306 CpG dyads and depth ≥ 10 on both Watson and Crick strands), we found that only 4.97% of called CpG dyads (22,590 out of 454,400 called CpG dyads, FDR = 5%) are symmetrically modified with 5fC/5caC. Further analysis focusing on CpG dyads with higher sequencing depth (≥20 on both strands) reached a similar conclusion (7.92% symmetrically modified with 5fC/5caC). Consistent with genome-scale analysis, locus-specific MAB-seq shows that the majority of 5fC/5caC-modified CpG dyads (86.7% in *shTdg* and 60% in *shTdg* + VC) exhibit strand asymmetry (**Figure 2.7**). Therefore, both genome-scale and locus-specific analyses suggest that TET-mediated generation of 5fC/5caC happens in a largely asymmetric manner at palindromic CpG dyads.

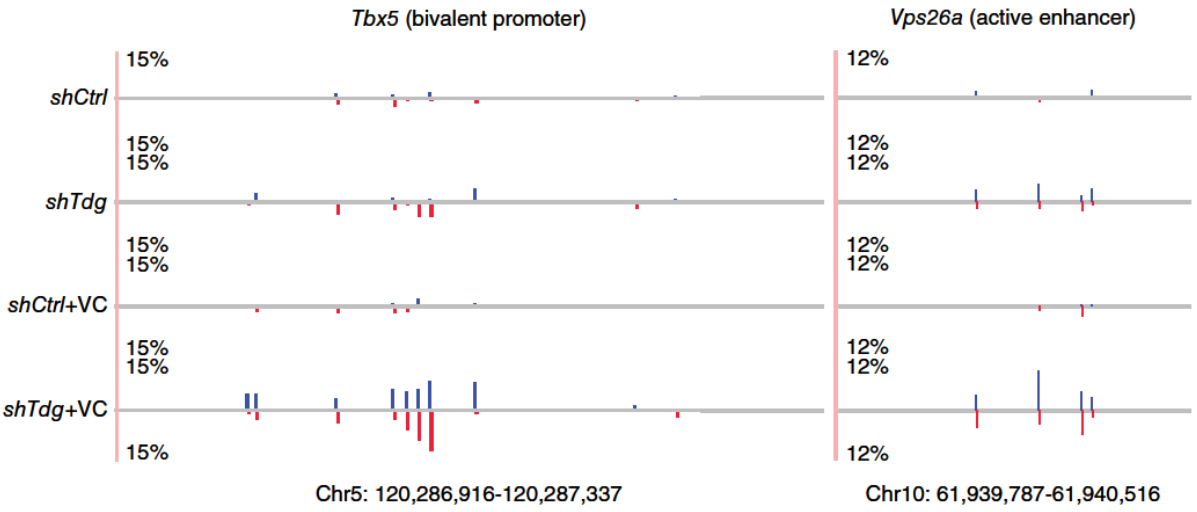

Figure 2.7 | Strand asymmetry of TET/TDG-dependent active DNA demethylation. Locus-specific MAB-seq analysis of two representative loci is shown. 5fC/5caC levels for the Watson strand are in blue, whereas those for the Crick strand are in red. A proportion of CpG dyads display different levels of 5fC/5caC between the two strands.

Base-resolution mapping of 5fC and 5caC separately

Base-resolution mapping of 5fC has been demonstrated using subtraction-based, chemical modification-assisted BS-seq methods such as fCAB-seq[43] and redBS-seq[53]. Because MAB-seq profiles 5fC and 5caC together, combining MAB-seq with these methods can allow separate quantification of 5fC and 5caC. To test this idea, we combined MAB-seq with the 5fC mapping method, redBS-seq[53] **(Figure 2.8a)**. After validating redBS-seq using synthetic oligonucleotides (**Supplemental Figure S2.4a**), we performed locus-specific redBS-seq and BS-seq to quantify 5fC at selected genomic loci enriched for 5fC and/or 5caC (*Tbx5* and *Vps26a*). The abundance and position of 5caC at these sites can then be determined by subtracting 5fC signals (derived from the

difference between redBS-seq and BS-seq) from the levels of 5fC+5caC measured by MAB-seq (**Figure 2.8b**, **Supplemental Figure S2.4c**). Using this integrative approach to map 5fC and 5caC separately, we observed that 5fC and 5caC displayed largely distinct distribution patterns at individual CpGs within these loci. For instance, within exon2 of *Tbx5*, a region enriched for both 5fC and 5caC, some CpG sites were only modified with 5fC (CpG #1 at *Tbx5* in **Figure 2.8b**), whereas some others were 5caC-only (CpG #2 at *Tbx5* in **Figure 2.8b**). More strikingly, as exemplified by CpG #3 at *Tbx5* in **Figure 2.8b**, CpG sites associated with both 5fC and 5caC may exhibit nonoverlapping, strand-specific 5fC/5caC distribution. This conclusion is further supported by analysis of an active enhancer within the *Vps26a* gene (**Supplemental Figure S2.4c**). Such diverse distribution patterns of 5fC and 5caC indicate the distinct processivity of TETs and/or substrate preference of TDG (5fC versus 5caC) at individual CpGs.

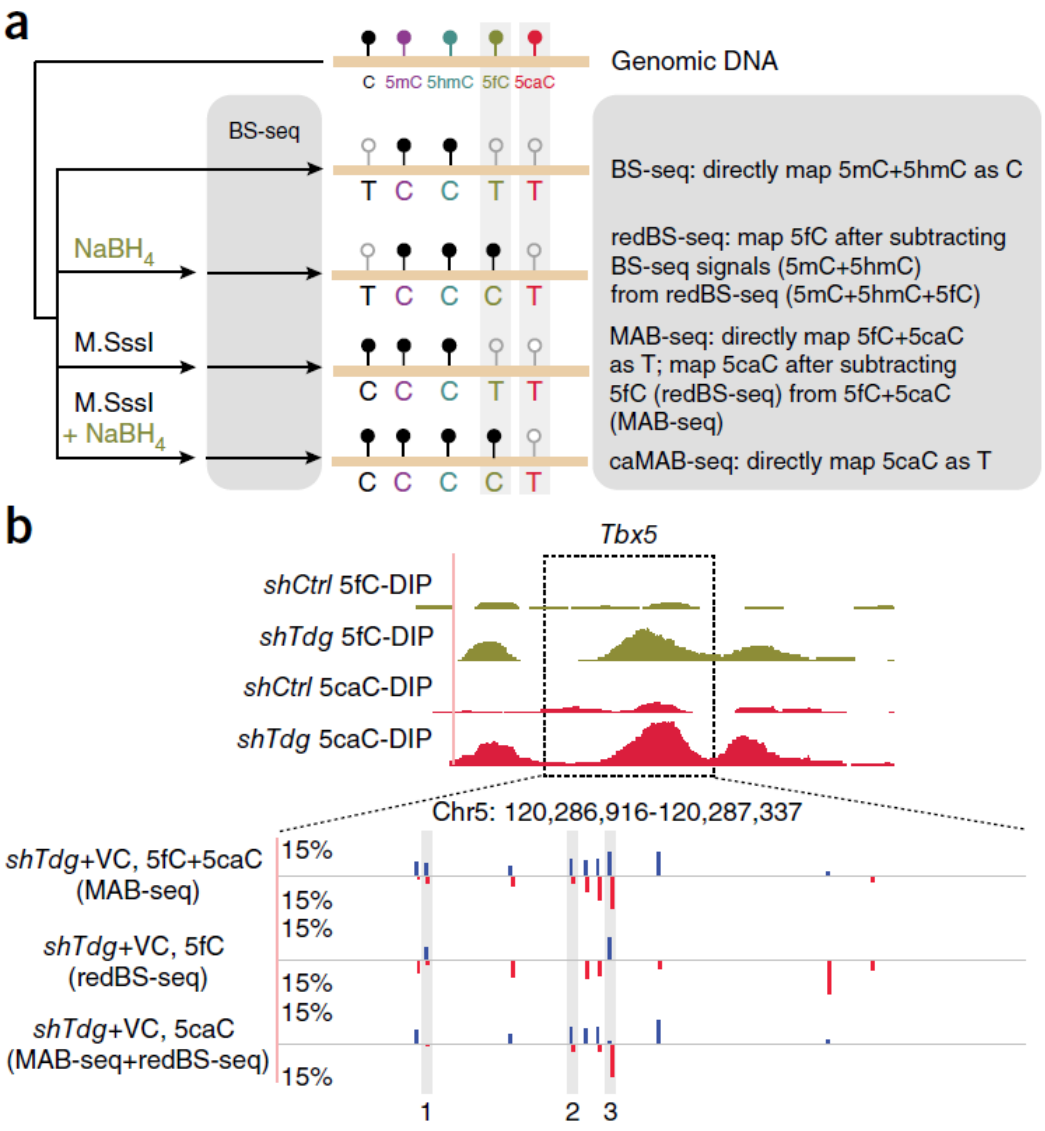

Figure 2.8 | Separate mapping of 5fC and 5caC. (a) Schematic diagram of BS-seq, redBS-seq, MAB-seq and caMAB-seq. (b) Locus-specific analysis of 5fC and 5caC at the *Tbx5* locus. For comparison, affinity enrichment-based 5fC and 5caC maps are shown on the top. Base-resolution maps of 5fC+5caC (measured by MAB-seq), 5fC (measured by redBS-seq) and 5caC (subtraction between MAB-seq and redBS-seq) are shown. Signals for the Watson strand are in blue, whereas those for the Crick strand are in red.

5caC mapping through integrating MAB-seq and redBS-seq requires two rounds of subtractions, representing a technical challenge for genome-scale analysis. Therefore, we explored a subtraction-independent 5caC mapping strategy by taking advantage of the fact that 5fC can be selectively reduced by sodium borohydride ($NaBH_4$) to 5hmC[43,53]. In this modified version of MAB-seq (termed caMAB-seq), a step of $NaBH_4$ incubation is added in addition to *M.SssI* treatment so that only 5caC is read as T after bisulfite conversion (**Figure 2.8a**). Analyses of defined sequences (lambda DNA or synthetic oligonucleotides) or 5caC-enriched genomic loci (e.g. *Tbx5* and *Vps26a*) demonstrated the validity of caMAB-seq in direct mapping of 5caC at base-resolution (**Figure 2.9**, **Supplemental Figure S2.4b and d**). Therefore, caMAB-seq can serve as a simple and effective approach to profile 5caC.

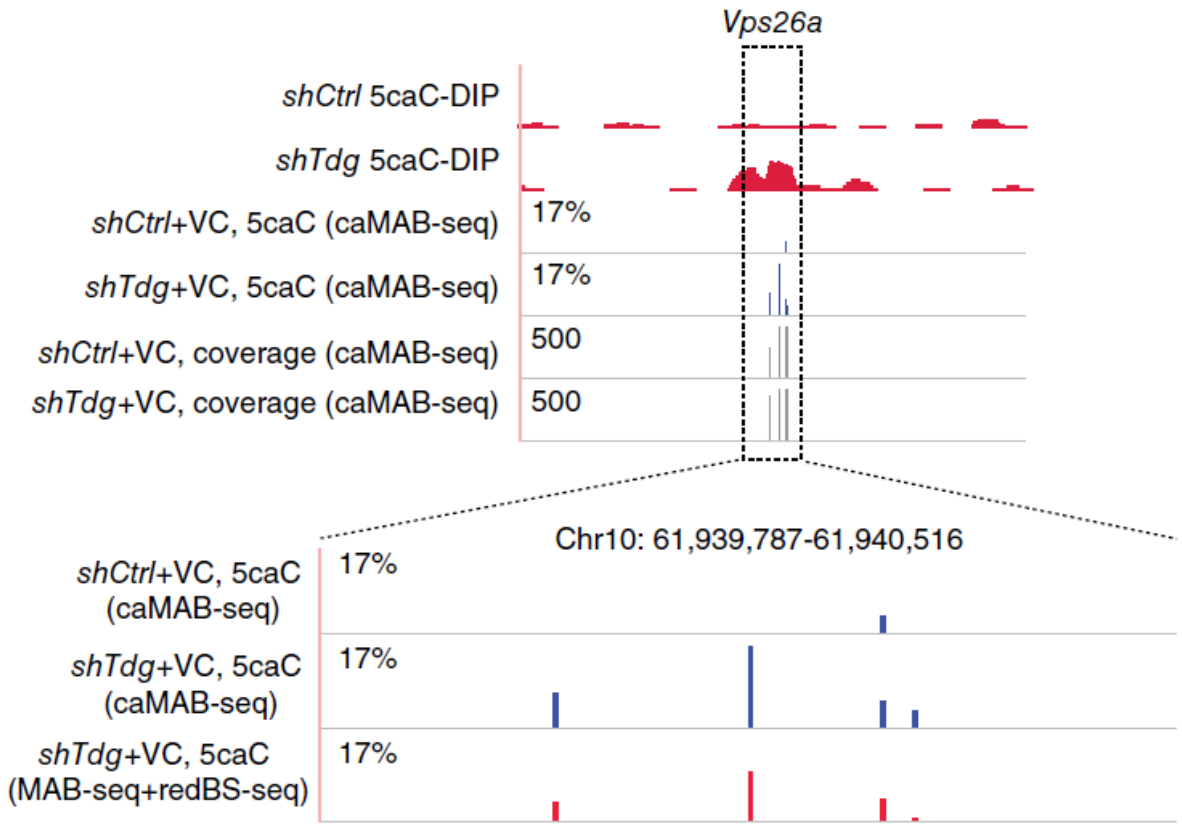

Figure 2.9 | Analysis of 5caC at *Vps26a* locus by caMAB-seq. Two base-resolution methods are compared: indirect mapping through subtraction between MAB-seq and redBS-seq (red); direct mapping by caMAB-seq (blue). For comparison, also shown are DIP-seq based maps of 5caC in control and TDG-depleted mouse ESCs. The level of 5caC (only Watson strand shown) is displayed as the percentage of total C modified as 5caC. Corresponding caMAB-seq sequencing depth (vertical axis limits are 0 to 500) at each CpG was also shown.

Discussion

In this chapter, I introduce MAB-seq, a method for quantitative mapping of 5fC/5caC at base-resolution, providing a tool that has broad application in the study of active DNA demethylation. Compared with 5fC or 5caC mapping methods based on affinity enrichment[31,43], MAB-seq allows quantitative detection of 5fC/5caC with higher spatial resolution and increased sensitivity. Compared with previously developed subtraction-based base-resolution methods for mapping 5fC or 5caC (redBS-seq[53], fCAB-seq[43] and caCAB-seq[56]), a major advantage of MAB-seq is the ability to directly determine the location and abundance of 5fC and 5caC in a single experiment, significantly reducing sequencing efforts and simplifying data analysis procedure. When combined with TDG depletion, simultaneous mapping 5fC and 5caC enables quantitative assessment of the generation and excision of 5fC/5caC, providing a direct readout of the TET/TDG-mediated active DNA demethylation activity (active modification - active removal (AM-AR) pathway).

Application of WG-MAB-seq analysis to mouse ESCs allowed us to investigate the genomic architecture and dynamics of active DNA demethylation activity at single-base resolution across the genome of this cell type. Although affinity enrichment-based 5fC and 5caC mapping methods suggest that 5hmC- and 5fC/5caC-enriched regions largely overlap[31,43], comparative analysis of base-resolution 5fC/5caC map and 5hmC map reveals that 5fC/5caC and 5hmC are largely not overlapped at individual CpG level. Additional analyses support a model in which TET proteins tend to stall at the 5hmC step at most CpGs, but exhibit higher processivity to further oxidize 5hmC to 5fC/5caC at CpGs with higher chromatin accessibility. Moreover, integrating MAB-seq with 5fC-mapping

(e.g., redBS-seq[53]) or 5caC-mapping (e.g., caMAB-seq) method not only provides a base-resolution approach to map 5fC and 5caC separately, but also reveals that 5fC and 5caC frequently do not overlap at individual CpG sites. These findings suggest the possibility that TET exhibits distinct processivity depending on local chromatin accessibility or other yet-to-be-identified regulatory processes. It is also possible that TDG may exhibit distinct activity or substrate preference at different CpG sites. Furthermore, strand-specific analysis of 5fC/5caC reveals that more than 90% of palindromic CpG dyads are asymmetrically modified by these modifications. This result suggests that TET/TDG-dependent active DNA demethylation activity preferentially targets palindromic CpG dyads in an asymmetric manner, supporting the recently proposed asymmetric base-flipping model[64,65].

In addition to a robust base-resolution mapping method for 5fC/5caC, we provide a whole-genome base-resolution map of TET/TDG-dependent active DNA demethylation activity in the mammalian genome. Given the simplicity and cost effectiveness of MAB-seq, we anticipate that genome-scale MAB-seq analysis can be applied to analyze diverse cell types in future studies, providing new insights into the mechanism and function of the DNMT-TET-TDG/BER cytosine-modifying cascade. Therefore, the 5fC/5caC mapping technology described in this chapter sets the stage for systematic investigation of the functional significance of active DNA demethylation in mammalian development and human diseases.

Methods

Mouse ESC cultures, lentiviral knockdown of Tdg and vitamin C treatment

V6.5 (control and Tdg knockdown), E14Tg2A (control, Tdg knockdown and Tet1/2−/−), and J1 (Dnmt1/3a/3b−/−) mouse ESC lines were cultured in feeder-free gelatin-coated plates in Dulbecco's Modified Eagle Medium (DMEM) (GIBCO, 11995) supplemented with 15% FBS (GIBCO), 2 mM L-glutamine (GIBCO), 0.1 mM 2-mercaptoethanol (Sigma), nonessential amino acids (GIBCO), 1,000 units/ml LIF (Millipore, ESG1107). The culture was passaged every 2 to 3 days using 0.05% Trypsin (GIBCO). Lentivirus-mediated Tdg knockdown in mouse ESCs were performed as previously described[31]. For vitamin C (Sigma, A8960) treatment, control and Tdg knockdown mouse ESCs were treated with 100 μg/ml of vitamin C for 60 h.

Whole-genome MAB-seq (WG-MAB-seq)

Genomic DNA was extracted from mouse ESCs using the DNeasy Blood & Tissue Kit (Qiagen 69504). 1 μg nonenriched genomic DNA was first spiked-in with unmethylated lambda DNA (1:400) and was then treated with *M.SssI*. In *M.SssI* treatment, DNA was first incubated with 1.0 unit/μl *M.SssI* methylase (New England Biolabs, M0226M) for 4 h in 25-μl reaction (1.25 μl of 20 unit/μl *M.SssI* and 0.5 μl of 32 mM SAM (final concentration: 640 μM)), and additional 25-μl containing same concentration of *M.SssI* (1.0 unit/μl) and SAM (640 μM) was supplemented to treat DNA for another 8 h (in 50 μl). *M.SssI*-treated genomic DNA (in 50 μl) was fragmented to an average size of 300–400 bp with Covaris M220 (20% duty factor, 200 cycles per burst, 80 s × 2). Sheared DNA was purified (1.2× AMPure XP beads), end-repaired and ligated to methylated adapters (forward: 5'-

ACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3′; reverse: 5′-/5Phos/GATCGGA AGAGCACACGTCTGAACTCCAGTC-3'; the asterisk denotes phosphorothioate bond). Methylated-adapter-ligated DNA was treated with *M.SssI* again and purified by sequential phenol/chloroform/isoamyl alcohol (PCI, 25:24:1) extraction and ethanol precipitation. Bisulfite conversion was performed using Qiagen EpiTect DNA Bisulfite Kit following manufacturer's instructions, except that the thermal cycle was run twice. Bisulfite-treated DNA was amplified using KAPA HiFi Uracil+ HotStart ReadyMix (Kapa Biosystems, KK2801) with indexed and universal primers from NEBNext Multiplex Oligos for Illumina. Amplified DNA is purified with 1.2× AMPure XP beads.

Data processing of whole-genome MAB-seq

Raw sequencing reads were trimmed for low-quality bases and adaptor sequences using Trimmomatic[75], and the data quality was examined with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The trimmed reads were mapped against the mouse genome (mm9 build) with Bismark[76]. PCR duplicates were removed using the Picard program (http://picard.sourceforge.net/). Cell line-specific SNPs overlapping with CpGs in the mouse genome (mm9) were filtered out with BisSNP[77]. All programs were performed with default setting. For MAB-seq analysis, raw signals were calculated as % of T/(C+T) at each called CpG.

Statistical calling of 5fC/5caC and assessing FDR of whole-genome MAB-seq

For each cytosine within CpG dinucleotides, we counted the number of "T" bases from MAB-seq reads as 5fC/5caC (denoted N_T) and the number of "C" bases as other forms

of cytosines (C/5mC/5hmC; denoted N_C). Next, we used the binomial distribution (N as the sequencing coverage (N_T + N_C) and p as the error rate (2.04%) of *M.SssI* methylase) to assess the probability of observing N_T or greater by chance. To estimate empirical FDR of calling 5fC/5caC-modified CpGs, we performed the procedure above on genome-scale MAB-seq signals of merged negative control sample (Dnmt1/3a/3b−/− and Tet1/2−/−) in which 5fC/5caC is largely absent[46], and on the WG-MAB-seq sample of TDG-depleted and VC-treated mouse ESCs. We then picked a P-value cutoff for FDR estimation. The FDR for a given P-value cutoff of the binomial distribution is the number of called CpG sites in the negative control sample divided by the number in the shTdg+VC sample.

Validation of MAB-seq by synthetic 38-bp oligonucleotides

To monitor the behavior of C and modified C in MAB-seq, 38-bp single-stranded DNA oligos were synthesized (forward strand: 5′-AGCCXGXGCXGXGCXGGTXGAGXGGC XGCTCCXGCAGC-3′, reverse (complementary) strand: 5′-GCTGXGGGAGXGGCXGC TXGACXGGXGXGGXGXGGGCT-3′, in which X is either unmodified C, 5hmC, 5fC or 5caC). To test whether *M.SssI* treatment alters the behavior of 5hmC, 5fC and 5caC during bisulfite sequencing, forward strands containing 5hmC, 5fC and 5caC were annealed to reverse strands containing the same modified Cs and ligated to methylated adaptors. The resulting oligos were treated by *M.SssI* using the same protocol used for locus-specific MAB-seq (described below), followed by bisulfite conversion and deep sequencing to determine their behavior. To test whether *M.SssI* functions at hemi-5hmC, 5fC and 5caC CpGs, top strands containing 5hmC, 5fC and 5caC were annealed to bottom strand containing unmodified C. The same experimental procedures were

undertaken to assess the efficiency of *M.SssI* in these contexts.

Locus-specific MAB-seq

1 µg genomic DNA was treated by *M.SssI* in a 50 µl reaction for four rounds. During each round of treatment, DNA was first treated by *M.SssI* for 2 h (1.5 µl *M.SssI* and 1 µl SAM), and additional *M.SssI* and SAM were supplemented to treat DNA for another 4 h (0.5 µl *M.SssI*, 1 µl SAM), increasing the total concentration of the enzyme to 0.8 unit/µl. DNA was purified by PCI after each round of treatment. After *M.SssI* treatment, bisulfite conversion was performed as described above. Selected loci were amplified by PCR using KAPA HiFi Hotstart Uracil+ DNA polymerase followed by sonication by Bioruptor (Diagenode), library preparation using NEBNext DNA Library Prep Master Mix Set (New England Biolabs) and deep sequencing by Illumina HiSeq 2500 sequencer. Alternatively, PCR-amplified DNA was cloned into TOPO vectors (Zero Blunt TOPO Cloning Kit, Invitrogen) for standard Sanger sequencing. Primer sequences for all locus-specific MAB-seq experiments were summarized in **Supplemental Table S2.1**.

Locus-specific redBS-seq

redBS-seq was performed as previously described[53]. 5 µl freshly made sodium borohydride aqueous solution (1M) was added to 250 ng DNA diluted in 15 µl water. The reaction was placed in darkness for an hour, quenched by 10 µl sodium acetate (0.75 M, pH = 5) and purified by PCI extraction. Bisulfite conversion was then performed, followed by PCR amplification of specific loci and deep sequencing analysis of PCR amplicons.

Locus-specific caMAB-seq

To perform locus-specific caMAB-seq, genomic DNA was first treated by *M.SssI* as described above and purified through PCI extraction. *M.SssI*-treated DNA was then reduced by $NaBH_4$ as described above in the redBS-seq protocol, followed by bisulfite conversion using Qiagen Epitect Bisulfite Kit. Selected loci were then amplified by PCR, and PCR amplicons were sequenced using Illumina deep sequencing. To examine whether unmodified C and 5fC are read as C during caMAB-seq, unmodified lambda DNA and synthesized 5fC oligo were also treated using the same protocol and analyzed by Illumina deep sequencing. In caMAB-seq, 5caC is read as T while C, 5mC, 5hmC and 5fC are read as C. To calculate the absolute level of 5caC, the background signal detected in Tet1/2−/− sample was subtracted from the raw signal detected in a tested sample.

Published datasets

We used following published datasets: 5hmC[49], Tet1[78], H3K4me3, H3K36me3, and H3K27me3[79], H3K4me1[80], Ezh2[81], Oct4, Nanog, Sox2, bivalent/active/silent promoters[82], LMR and UMR[83], H3K27ac and p300[84], CTCF and mouse ESC-specific enhancers[85], and DNase I hypersensitive sites (DHS, ENCODE project at genome.ucsc.edu). The text and figures in this chapter were adapted from our paper published on *Nature Biotechnology*[46].

Chapter 3:

Development of low-input and single-cell MAB-seq (liMAB-seq and scMAB-seq)

Abstract

To understand mammalian active DNA demethylation, various methods have been developed to map the genomic distribution of the demethylation intermediates 5fC and 5caC. However, the majority of these methods require a large number of cells to begin with. In Chapter 3, I present the development of low-input MAB-seq (liMAB-seq) and single-cell MAB-seq (scMAB-seq), capable of profiling 5fC and 5caC at genome scale using ~100 cells and single cells, respectively. To allow low-input or single-cell analysis, we modified the original MAB-seq protocol (termed regular MAB-seq) and integrated multiple steps of library preparation into a single reaction tube, minimizing the loss of DNA during the experimental process. Using mouse ESCs, we demonstrated that liMAB-seq starting from ~100 cells is comparable to regular MAB-seq starting from ~1 μg DNA, in terms of 5fC/5caC signals detected and CpG sites covered. We also showed that scMAB-seq captures the similarity and heterogeneity of 5fC/5caC among single cells.

Background

To understand the mechanism and function of active DNA demethylation, various methods have been developed to map the genomic distribution of 5fC and 5caC[31,43,44,53,56,86,87]. However, the majority of these methods require a large amount of input DNA (typically hundreds of nanograms or more, corresponding to several hundred thousand cells) and thus cannot be readily applied to biological processes with limited cell availability, such as mammalian preimplantation development and PGC development. To understand the role of active DNA demethylation in these biological processes, low-input methods capable of profiling the demethylation intermediates using a small number of cells are needed. In addition to low-input methods, single-cell methods are also required to reveal cell-to-cell heterogeneity of active DNA demethylation, which may contribute to the diversity of cell function.

When performing genomic and epigenomic studies using a small number of cells or single cells, DNA purification-associated loss of material during sequencing library construction presents a major challenge. Most of the current methods used for profiling oxidized forms of 5mC involve chemical treatment or bead-based enrichment and thus require multiple rounds of DNA purification, making them incompatible with studies where cell availability is limited. In Chapter 2, I introduce MAB-seq, an enzyme-based method capable of mapping genomic 5fC and 5caC together at single-base resolution[46]. MAB-seq is based on bisulfite sequencing (BS-seq), a commonly used method for mapping DNA methylation. In regular BS-seq, 5mC and 5hmC (abbreviated as 5mC/5hmC) are sequenced as C while unmodified C, 5fC and 5caC are sequenced as T. In MAB-seq, an additional step

of *M.SssI* methyltransferase treatment converts unmodified C in CpG context to 5mC prior to bisulfite conversion, thereby allowing 5fC and 5caC to be directly read out as T (**Figure 2.1**). Because *M.SssI* treatment is carried out in normal enzymatic conditions[88], it is possible that *M.SssI* treatment can be combined with other steps of library preparation without introducing additional rounds of DNA purification.

Currently, two different library preparation strategies, reduced representation bisulfite sequencing (RRBS)[89] and post bisulfite adaptor tagging (PBAT)[90], have been implemented to perform low-input or single-cell BS-seq[91-93]. In RRBS strategy, the genome is fragmented at fixed locations by a restriction enzyme recognizing and cutting a CG-contained sequence independent of CG-modification status (in our case, we used TaqαI which recognizes and cuts TCGA)[18]. Subsequent adaptor tagging, bisulfite conversion and library amplification allows the analysis of a fixed set of CpG sites (those within and in proximity to TCGA) with a relatively small sequencing effort. In PBAT strategy, the genome is randomly sheared during bisulfite conversion and tagged by adaptors in a random manner. Therefore, the majority, if not all, of the CpG sites in the genome can be analyzed in a relatively unbiased manner. Both RRBS and PBAT can be adapted for low-input or single-cell studies, because multiple steps can be effectively integrated into a single reaction tube. Since these two strategies each have unique advantages, we decided to develop liMAB-seq and scMAB-seq based on both of them (**Figure 3.1**).

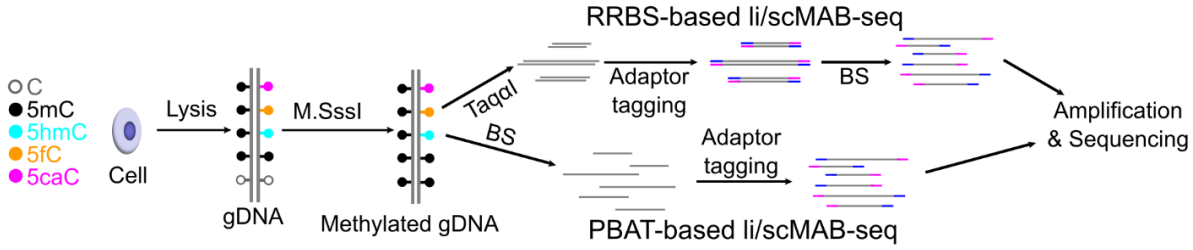

Figure 3.1 | Schematic illustration of liMAB-seq and scMAB-seq based on RRBS or PBAT strategy. For both strategies, the steps from cell lysis to bisulfite conversion are integrated into a single reaction tube without DNA purification in between.

Results

Development of RRBS-based liMAB-seq

To examine the feasibility of liMAB-seq, we first tested the combination of MAB-seq with RRBS by integrating cell lysis, *M.SssI* treatment and RRBS library preparation into a single-tube reaction, starting with 100 mouse diploid cells (**Figure 3.1**)[91]. To assess 5fC/5caC detection, we used Tet1-3 triple knockout (Tet TKO) mouse embryonic stem cells (ESC) as a negative control and TDG-depleted (shTdg) mouse ESCs treated by vitamin C as a positive control[31,46,66]. We started by testing different concentrations of *M.SssI*. While insufficient *M.SssI* may lead to incomplete methylation of unmodified C[46], a concentration of *M.SssI* used in regular MAB-seq for bulk DNA unexpectedly reduced library quality, as shown by loss of large DNA fragments in the final library (**Supplemental Figure S3.1a**), low mapping efficiency and poor CpG coverage. We therefore reduced *M.SssI* concentration and determined 3.2 units (U) / reaction as an optimized condition that allows efficient methylation of unmodified C without compromising library quality **(Figure 3.2a, Supplemental Figure S3.1a)**.

Comparative analyses suggest that the data quality of our method starting with 100 cells, termed low-input MAB-seq (liMAB-seq), is comparable to that of the regular MAB-seq starting with 1 μg DNA (~$2x10^5$ cells) in terms of unmodified C conversion rate (**Figure 3.2a**), mapping efficiency (**Supplemental Figure S3.1b**), CpG coverage (**Supplemental Table S3.1**, **Supplemental Figure S3.1c**), 5fC/5caC level in different genomic features (**Figure 3.2b**, **Supplemental Figure S3.1d**) and individual genomic regions (**Figure 3.2c,d**). The results are also comparable between biological replicates (**Supplemental**

Figure S3.1e). Notably, the method is widely applicable for analyzing samples with varying amounts of unmodified C, as Dnmt triple knockout (Dnmt TKO) mouse ESCs (with all Cs unmethylated) can be efficiently methylated (**Supplemental Figure S3.1d**).

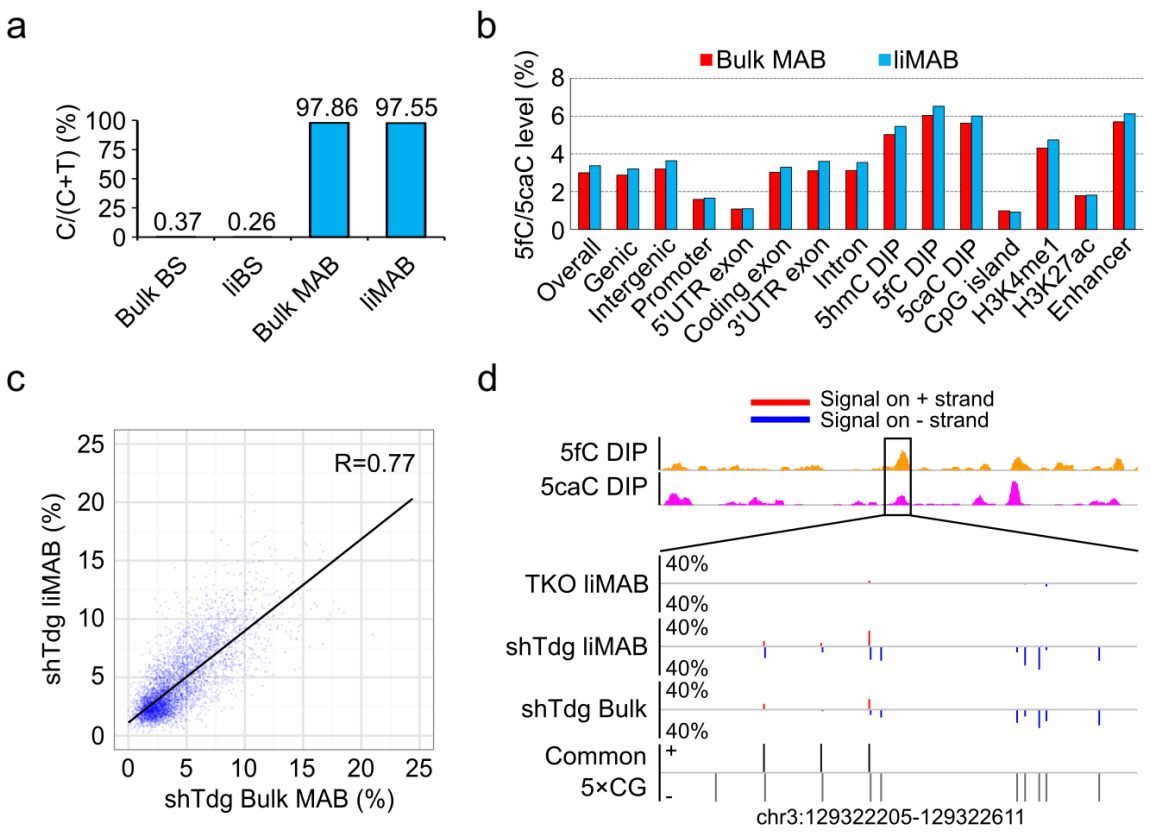

Figure 3.2 | Validation of liMAB-seq based on RRBS strategy. (a) Efficient methylation of unmodified lambda DNA by liMAB-seq. liMAB-seq is comparable with regular MAB-seq starting from bulk DNA. (b) 5fC/5caC level quantified by liMAB-seq is comparable to that quantified using regular MAB-seq (bulk). (c) Correlation between liMAB-seq and regular MAB-seq (bulk). Each dot represents a 2 kb genomic bin with at least 12 20×CGs, and 5fC/5caC level within the bin is calculated as sum(T)/(sum(C)+sum(T)). (d) Representative locus showing 5fC/5caC detected by liMAB-seq. 5fC and 5caC DIP: 5fC and 5caC profiles obtained through DNA immunoprecipitation. TKO: Tet TKO mouse ESCs. shTdg: TDG-depleted mouse ESCs. Red and blue: 5fC/5caC signals detected on top (+) and bottom (-) strand, respectively. Common 5×CG: CpG sites covered by at least five times by sequencing and shared among all samples.

Although 5fC and 5caC are relatively rare modifications, two factors allow MAB-seq to confidently identify genomic regions modified by 5fC/5caC. Firstly, despite the average level of 5fC/5caC across all the CpG sites is low, the 5fC/5caC level at individual CpG sites being modified is high enough to be distinguished from the background signal[46]. Secondly, real 5fC/5caC signals tend to cluster together at individual genomic regions, while background signals should distribute across the genome in a largely random manner. Therefore, we undertook a binning approach (100bp bin) and applied a numeric cutoff (>=10%) to call 5fC/5caC-modified regions in liMAB-seq datasets. Using this strategy, 1.3% and 12% of the regions covered in Tet TKO negative control and TDG-depleted ESCs were called as modified, respectively, representing a false discovery rate (FDR) of 11% (**Supplemental Figure S3.1f**). FDR is even lower for zygotic paternal pronuclei (FDR=3.6%), which have a higher 5fC/5caC level (see below), using a larger bin size or applying a more stringent numeric cutoff. Consistent with previous reports, 5fC/5caC-modified regions called in TDG-depleted ESCs are enriched at enhancers, H3K4me1 ChIP-seq peaks and 5hmC/5fC/5caC DIP-seq peaks (**Supplemental Figure S3.1f**)[31,43,46,54,86,87]. Among the regions covered by RRBS analysis and overlapped with 5fC/5caC DIP-seq peaks, 27.1% are called as modified in TDG-depleted cells while only 1.0% are wrongly called as modified in Tet TKO negative control, further confirming the effectiveness of our approach in distinguishing real 5fC/5caC signals from background. Therefore, liMAB-seq can confidently identify 5fC/5caC-modified regions.

Development of RRBS-based scMAB-seq

The success of liMAB-seq prompted us to test the feasibility of single-cell MAB-seq

(scMAB-seq). By further reducing the *M.SssI* concentration to 0.8 U / reaction, we succeeded in performing scMAB-seq of single mouse ESCs (**Supplemental Figure S3.1g**). The mean mapping efficiency of scMAB-seq libraries is 33.4% while negative control starting with no cell has a mapping efficiency of 0.2%, demonstrating a minimum degree of contamination (**Supplemental Figure S3.1b**). When sequencing depths are comparable (around 30 million reads per library), the number of CpG sites covered in scMAB-seq is 20-50% of that of liMAB-seq or regular MAB-seq while the distribution of the covered sites in different genomic features are comparable (**Supplemental Figure S3.1c**). The number of CpG sites covered can be further improved by increasing the length of sequencing (**Supplemental Table S3.2**) or performing pair-end sequencing (**Supplemental Table S3.3**). Although further reducing *M.SssI* concentration may increase the number of CpG sites covered, methylation efficiency will be compromised (**Supplemental Figure S3.1h**). Therefore, we chose 0.8U / reaction for scMAB-seq in order to balance CpG coverage and methylation efficiency.

scMAB-seq analyses reveal that the relative abundance of 5fC/5caC in different genomic features is largely consistent between single cells and bulk profile, while the absolute level of 5fC/5caC in individual cells displays heterogeneity (**Figure 3.3a**). At individual CpG sites, 5fC/5caC level displays a digital pattern of 0%, 50% or 100% as expected for single-cell datasets **(Figure 3.3b**, **Supplemental Figure S3.1i)**. Comparing a shared set of CpG sites, individual cells display heterogeneity of 5fC/5caC, reflecting the transient nature of the demethylation process (**Figure 3.3b**).

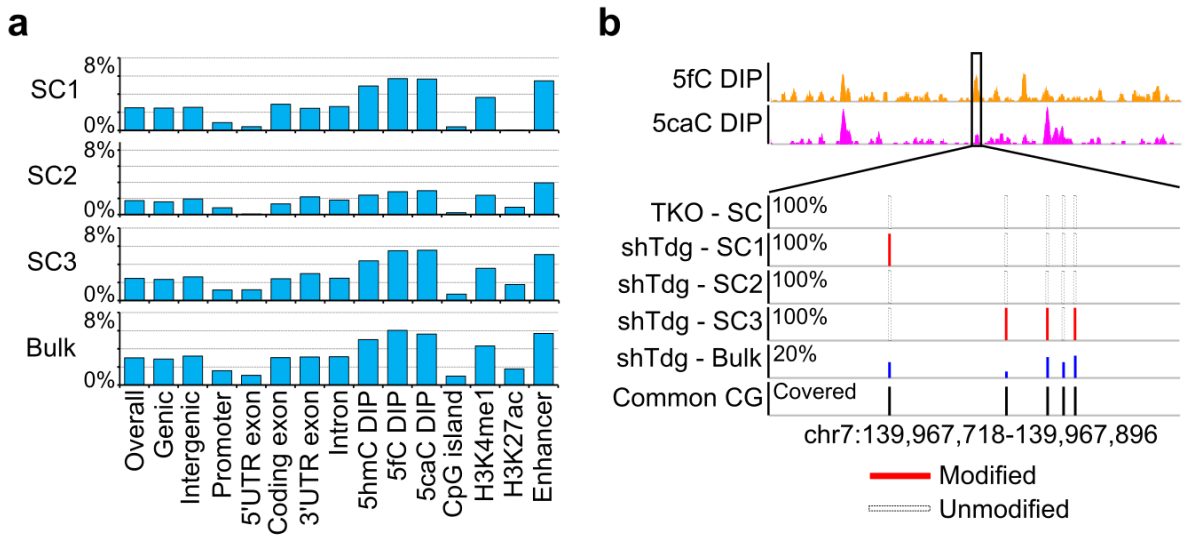

Figure 3.3 | Validation of scMAB-seq based on RRBS strategy. (a) Mean levels of 5fC/5caC in different genomic features. For each sample, background signals detected in Tet TKO control were subtracted from the raw signals to obtain the actual level of 5fC/5caC. SC1 to 3 are three TDG-depleted mouse ESCs. Different single cells have variable absolute levels of 5fC/5caC, but the trends (relative abundance) of 5fC/5caC in different genomic features are comparable. (b) A representative locus showing 5fC/5caC detected in single cells by scMAB-seq. Five common 5×CGs in the four single cells are shown. 5fC/5caC is not detected in Tet TKO negative control and displays heterogeneity and digital pattern in the three TDG-depleted mouse ESCs.

Development of PBAT-based liMAB-seq and scMAB-seq

As a proof of principle, we also tested the possible combination of *M.SssI* treatment with PBAT by modifying a published protocol (**Figure 3.1**)[92]. Our results demonstrate that PBAT-based liMAB-seq and scMAB-seq can also be successfully performed to detect 5fC/5caC in TDG-depleted mouse ESCs (**Supplemental Figure S3.2a**). With around 250

million reads, PBAT-based liMAB-seq starting from 100 cells can cover around 25 million CpG sites, representing around 60% of all the CpG sites in the mouse genome (1 CpG dyad is counted as 2 different CpG sites here). With around 30 million reads, 1.7-6.1 million CpG sites can be covered by PBAT-based scMAB-seq. The mapping efficiency is around 40% for liMAB-seq and ranges from 9.0% to 35.5% for scMAB-seq, while the mapping efficiency for the negative controls (no cell) is around 0.2% (**Supplemental Figure S3.2b**, **Supplemental Table S3.4**). The distribution of the covered CpG sites and the relative abundance of 5fC/5caC in different genomic features are comparable between PBAT-based scMAB-seq, liMAB-seq and the published whole-genome bulk MAB-seq datasets (**Figure 3.4**, **Supplemental Figure S3.2c**)[46].

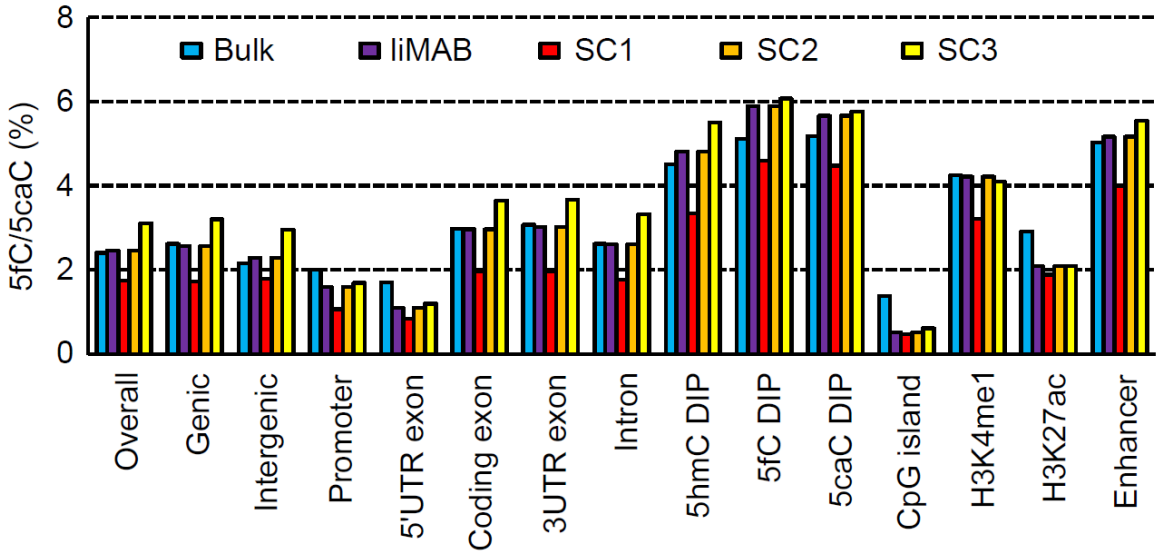

Figure 3.4 | 5fC/5caC quantification by PBAT-based liMAB-seq and scMAB-seq. 5fC/5caC level of different genomic features measured by the published whole-genome MAB-seq (bulk) and PBAT-based liMAB-seq and scMAB-seq. Three single cells (SC1-3) analyzed by PBAT-based scMAB-seq are shown. Background signals detected in TKO negative controls have been subtracted.

Notably, the distribution of the covered CpG sites and the level of 5fC/5caC in different genomic features are also largely comparable between PBAT and TaqαI-based RRBS (**Figure 3.2b, 3.3a, 3.4**, **Supplemental Figure S3.1c,d, 3.2c**). Further analyses presented in later chapters also demonstrate that both strategies lead to the same conclusions. For consistency and comparability, we chose RRBS as our major strategy for both liMAB-seq and scMAB-seq unless otherwise specified.

Discussion

To facilitate the study of active DNA demethylation in biological contexts with limited cell numbers, we developed liMAB-seq and scMAB-seq, the first methods capable of profiling 5fC/5caC using approximately 100 cells and single cells, respectively.

We have shown that liMAB-seq is comparable to regular MAB-seq starting from around 1 µg DNA in terms of *M.SssI* efficiency, CpG coverage, 5fC/5caC quantification and 5fC/5caC mapping. Possible biological systems to apply liMAB-seq include preimplantation embryos, PGCs, neurons and other scenarios with limited number of cells. In Chapter 4, I present our analysis of paternal genome demethylation in mouse preimplantation embryos using liMAB-seq.

By further modifying *M.SssI* treatment protocol, we developed scMAB-seq. Number of CpG sites covered by scMAB-seq is generally lower compared to liMAB-seq or regular MAB-seq, probably due to inevitable loss of material during library preparation. However, by using a binning approach (presented in Chapter 5), most genomic regions of interest can still be investigated. In Chapter 5, I introduce multiple applications of scMAB-seq to demonstrate its utility.

liMAB-seq and scMAB-seq can be performed based on either RRBS or PBAT strategy, providing researchers with the flexibility to choose a suitable protocol based on the biological questions to address and the resources available. In the case of liMAB-seq, RRBS version of the method will allow genome-scale analysis of a fixed subgroup of total

CpG sites with a relatively small sequencing effort, while PBAT version of the method can interrogate most, if not all, CpG sites in the genome. In the case of scMAB-seq, RRBS version of the method can produce results directly comparable to RRBS-based liMAB-seq or regular MAB-seq, while PBAT version of the method can cover a lot more CpG sites to enable high-resolution analysis of 5fC/5caC.

Methods

ES cell culture

Tet TKO and shTdg mouse ESCs (E14TG2a) were cultured in DMEM (GIBCO, 11995) with 15% FBS (GIBCO), 2 mM L-glutamine (GIBCO), 0.1 mM 2-mercaptoethanol (Sigma), nonessential amino acids (GIBCO) and 1,000 units/ml LIF (Millipore). shTdg ESCs were treated by vitamin C for 60 hours before collection to stimulate 5fC/5caC generation[46]. For liMAB-seq, 100 cells were obtained through dilution. For scMAB-seq, single cells were collected through mouth pipetting.

RRBS-based liMAB-seq

For liMAB-seq, we collected around 100 diploid cells in <=1 μL 0.01% PBS-BSA through dilution (mouse ESCs) or picking (2-cell embryos). In the case of zygotic paternal pronuclei, we started from around 150 haploid pronuclei. 5 μL lysis mix containing the following components is then added: 4.4 μL water, 0.1 μL 100×TE, 0.1 μL 1M KCl, 0.15 μL 10% Triton X-100 and 0.25 μL QIAGEN Protease (20 mg/mL, QIAGEN 19155). The reaction was incubated at 50 °C for 3 hours and 75 °C for 30 minutes. Without purifying the reaction, 14 μL *M.SssI* mix was added: 10.2 μL water, 2 μL Cutsmart buffer, 1 μL unmethylated lambda DNA (1% w/w), 0.4 μL *M.SssI* (4 U/μL, M0226S) and 0.4 μL SAM (32 mM). The reaction mix was incubated at 37 °C for 2 hours, followed by supplementation of another 0.4 μL *M.SssI* (4 U/μL) and 0.4 μL SAM (32 mM) and incubation at 30 °C for 6 hours. After the reaction mix was heat-inactivated at 65 °C for 20 minutes, 10 μL digestion mix (0.5 μL TaqαI (20 U/μl), 1 μL NEB Cutsmart buffer and 8.5 μL water) was added directly to the reaction mix, followed by incubation at 65 °C for

3 hours and 80 °C for 20 minutes. End preparation was then performed as described for regular MAB-seq. Without purifying the reaction after end preparation, 1.25 µL water, 0.5 µL NEB Cutsmart buffer, 0.25 µL 100 mM ATP, 1.0 µL 0.75 µM methyalted adaptor and 1 µL T4 ligase (2,000U/µl, New England BioLabs, cat. no. M0202M) were added, and ligation was performed as described for regular MAB-seq. Without purifying the reaction, bisulfite conversion was set up as bellow: reaction mix after ligation (~36 µL), 19 µL DNA protection buffer and 85 µL bisulfite conversion mix. The remaining steps are the same as those for regular MAB-seq, with the exception that 16 to 17 PCR cycles were used to amplify the library.

RRBS-based scMAB-seq

For scMAB-seq, all steps are the same as liMAB-seq with three modifications. Firstly, *M.SssI* (4 U/µl) was diluted 4 times (1 U/µl; 0.8 U for the total reaction) to perform MAB-seq. Next, methylation efficiency was assessed by single Tet TKO cells instead of lambda DNA spike-in, due to the consideration that lambda DNA spike-in at trace amount (~5fg per sample at 1% w/w) may not allow accurate estimation of *M.SssI* efficiency. Therefore, performing scMAB-seq of single Tet TKO cells in parallel with the samples of interest is a better strategy for quality control. Thirdly, 20 to 21 PCR cycles were used to amplify the library.

PBAT-based liMAB-seq and scMAB-seq

For PBAT-based liMAB-seq and scMAB-seq, cell lysis and *M.SssI* treatment were performed in a single-tube reaction the same as RRBS-based liMAB-seq and scMAB-seq.

The steps later are modified from a published PBAT protocol[92]. In brief, 20 μL mix after *M.SssI* treatment was used for bisulfite conversion by Zymo EZ DNA Methylation Direct kit (D5020) following manufacturer's instructions. To the 10 μL bisulfite-converted DNA eluted by 10mM Tris-HCl (pH 8.5), we added 9.5 μL H2O, 2.5 μL 10× blue buffer (Enzymatics P7010-HC-L), 1 μL 10 μM oligo 1 (CTACACGACGCTCTTCCGA TCTNNNNNNNNN) and 1 μL 10 mM dNTPs. After the 24 μL mix was incubated at 65 °C for 3 minutes and cooled to 4 °C, 1 μL Klenow exo$^-$ (50 U/μL, Enzymatics P7010-HC-L) was added. For the primer extension of oligo 1, we incubated the 25 μL mix at 4 °C for 5 minutes, increased the temperature to 37 °C in a ramp rate of +1 °C per 15 seconds (or the slowest ramp rate available for the PCR machine) and incubated at 37 °C for 30 minutes. The reaction mix was denatured at 95 °C for 1 minute, cooled down to 4 °C in PCR machine and transferred immediately to ice. 2.5 μL Klenow-oligo1 mix containing 0.5 μL Klenow exo$^-$ (50U/μL), 0.1 μL 10 mM dNTPs, 1 μL 10 μM oligo 1, 0.25 μL 10× blue buffer and 0.65 μL water was added to the 25 μL mix, followed by the same primer extension program from 4 °C to 37 °C as described above. The denaturing and primer extension process were repeated for another three rounds without purification, bringing the reaction mix to 35 μL. Without purification, 63 μL water and 2 μL E. coli Exonuclease I (NEB M0293S, 20U/μL) were added, followed by incubation at 37 °C for 1 hour, and purification by SPRI beads (0.8×). To the 20 μL DNA eluted by 10mM Tris-HCl (pH 8.5), we added 20 μL water, 5 μL 10× blue buffer, 2 μL 10 μM oligo 2 (CAGACGTGT GCTCTTCCGATCTNNNNNNNNN) and 2 μL 10 mM dNTPs. The mix was incubated at 95 °C for 1 minute, cooled down to 4 °C by PCR machine and transferred immediately to ice. 1 μL Klenow exo$^-$ (50 U/μL) was then added. For the primer extension of oligo 2, we

incubated the 50 μL mix at 4 °C for 5 minutes, increased the temperature to 37 °C in a ramp rate of +1 °C per 15 seconds (or the slowest ramp rate available for the PCR machine) and incubated at 37 °C for 90 minutes. The mix was then purified by SPRI beads (0.8×) and amplified using KAPA HiFi Uracil+ ReadyMix and NEBNext index primer, followed by SPRI (0.8×) purification of the final library. 7 and 11 cycles were used for liMAB-seq and scMAB-seq, respectively.

Illumina sequencing and data analysis

The majority of the samples were sequenced by Illumina HiSeq 2500 using 100 bp single-end mode, but some samples were sequenced using 100 bp pair-end mode or 250 bp single-end mode for comparison. For analysis other than those shown in Supplemental Table S2 and S3, all reads were trimmed to 100 bp single-end. Adaptor trimming and quality trimming were performed using Trim galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). For RRBS-based samples, the parameter is: --three_prime_clip_R1 2 --length 35. For PBAT-based samples, the parameter is: --clip_R1 9 --three_prime_clip_R1 9 --length 35. Sequencing alignment was performed using Bismark[76]. Mouse mm9 genome was used as the reference genome for sperm, zygotic paternal pronuclei and 2-cell embryos, while an mm9-based E14 genome was used for ESCs[94]. Specifically for PBAT-based samples, PCR duplicates were excluded using Picard (http://picard.sourceforge.net). For downstream analysis, known SNPs reported by Sanger Institute were excluded[95]. Specifically for RRBS-based samples, CpG sites with an abnormally high level of 5fC/5caC in multiple regular or liMAB-seq samples (>20% in both replicates of Tet TKO mouse ESCs, or >20% in both sperm and zygotic

paternal pronuclei) were also excluded. For all RRBS-based samples, CpG sites covered for at least five times (5×CG) were used for analysis and description unless otherwise specified. For each condition with biological replicates, the 5×CGs commonly covered in the replicates are merged by summing up the C and T counts, generating one dataset for downstream analyses. For PBAT-based samples, 1×CGs were used for analysis and description unless otherwise specified.

Identification of 5fC/5caC-modified regions by liMAB-seq and estimation of FDR

To call 5fC/5caC-modified regions (**Supplemental Figure S3.1f**), 5×CGs commonly covered in the sample of interest (shTdg mouse ESCs or zygotic paternal proneuclei) and the negative control (Tet TKO mouse ESCs) were extracted. Level of 5fC/5caC signal is estimated as T/(C+T) for all 100bp genomic bins with at least 2 5×CGs. To call bins modified by 5fC/5caC, a numeric cutoff of 5fC/5caC level >=10% is applied. FDR for the sample of interest is estimated as (number of called bins in Tet TKO control) / (number of called bins in sample of interest), which is 11% for shTdg mouse ESCs and 3.6% for zygotic paternal pronuclei. FDR can be further reduced when larger bin sizes are used.

Identification of 5fC/5caC-modified CpG sites and regions by scMAB-seq

For scMAB-seq data, the possible level of 5fC/5caC for a CpG site is 0%, 50% or 100% (25% and 75% are less likely but theoretically possible if the cell is undergoing DNA replication). In reality, PCR error or biased amplification may lead to an observed level slightly deviating from these three possibilities. Therefore, a digital transformation of the raw data was performed to convert CpG sites with a 5fC/5caC level of <=20% to 0%, >=80%

to 100% and any level in between to 50%. After the transformation, CpG sites that have 50% or 100% 5fC/5caC were regarded as modified. The digitally transformed data were also used to calculate the mean level of scMAB-seq signals in different genomic features (**Figure 3.3a**).

Published datasets

Sources for published datasets are the same as discussed in Chapter 2. The text and figures in this chapter were adapted from our paper published on *Genes & Development*[58].

Chapter 4:

liMAB-seq reveals insights into paternal genome demethylation in mouse preimplantation embryos

Abstract

In mouse and human, the zygotic genome undergoes global DNA demethylation shortly after fertilization. The highly methylated paternal genome inherited from the sperm is demethylated through a combination of passive dilution of 5mC and TET3-mediated generation of 5hmC, 5fC and 5caC coupled with replication-dependent dilution. Though this process was discovered two decades ago, quantification and mapping of 5fC/5caC have been difficult due to a lack of methods capable of profiling small numbers of cells from mouse early embryos. In Chapter 4, I present low-input MAB-seq (liMAB-seq) analysis of paternal genome demethylation. First, we analyzed the dynamics of 5mC/5hmC and 5fC/5caC by profiling mouse sperm and zygotic paternal pronuclei using BS-seq and MAB-seq, revealing that 5mC/5hmC decrease is accompanied by 5fC/5caC increase, and that a proportion of 5mC/5hmC is processed to unmodified cytosine. Second, we profiled 5fC/5caC in mouse 2-cell-stage embryos and revealed that TET-mediated oxidation also displays distinct processivity at different genomic regions. In addition, TET processivity correlates with chromatin accessibility.

Background

Shortly after fertilization, mouse and human zygotes undergo extensive epigenetic reprogramming, including global DNA demethylation[96-98]. Both the paternal and maternal copies of the genome are demethylated through a combination of passive dilution of 5mC (passive DNA demethylation) and TET3-mediated generation of 5hmC, 5fC and 5caC coupled with replication-dependent dilution (active modification - passive dilution (AM-PD) pathway), leading to an increased level of unmodified cytosine after rounds of cell division. Compared with the maternal genome, paternal genome undergoes TET3-mediated oxidation to a much higher extent and has been used as a model to understand the regulation and outcome of TET-mediated oxidation (**Figure 4.1**)[10]. Though paternal genome demethylation was discovered two decades ago, many questions remain to be addressed.

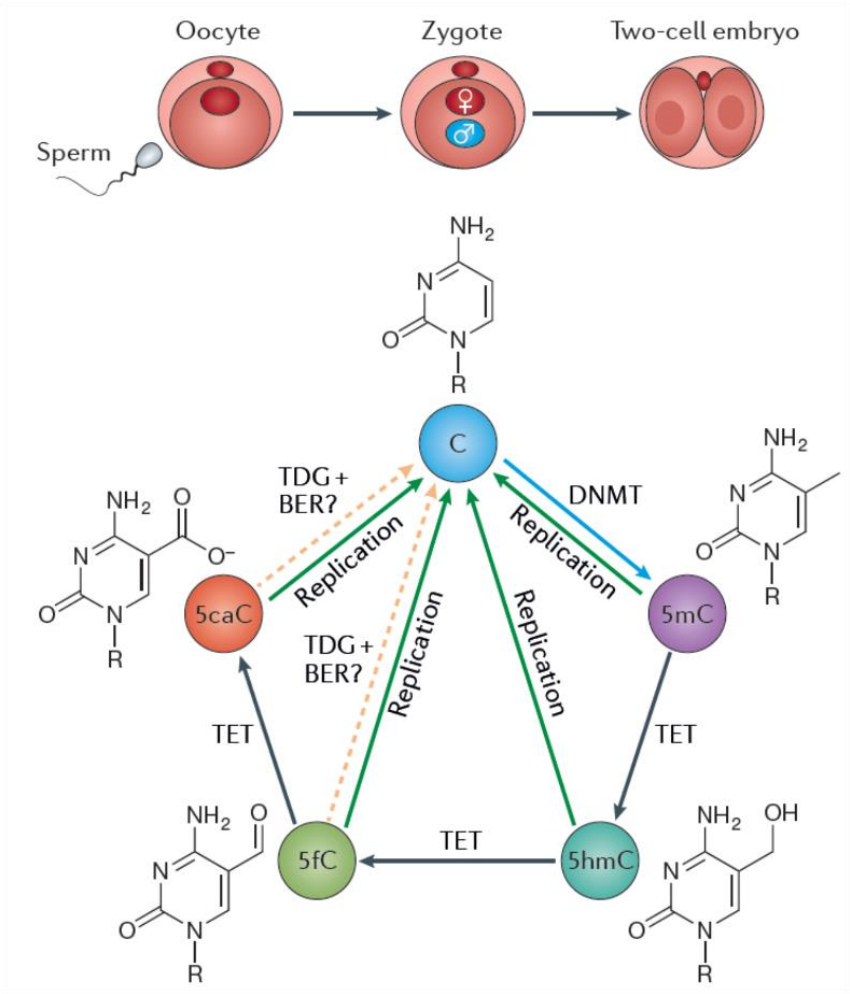

Figure 4.1 | DNA demethylation in the paternal genome of the zygote. 5‐Methylcytosine (5mC) is oxidized by TET3 to 5‐hydroxymethylcytosine (5hmC), 5‐formylcytosine (5fC) and 5‐carboxylcytosine (5caC). DNA replication occurs at around the same time as oxidation and results in the dilution of all modified forms of cytosine, including 5mC. Dashed lines denote steps that require more supporting evidence.

First, many of the molecular events, including loss of 5mC and gain of 5hmC/5fC/5caC, as well as the replication-dependent dilution of 5mC/5hmC/5fC/5caC, have been primarily investigated by immunostaining, leaving open the questions where in the genome and to

what extent 5mC is being converted to its oxidized forms. Sequencing analysis of this process has been historically difficult because of the limited amount of material available from early embryos. The successful development of liMAB-seq, as described in Chapter 3, makes it possible to analyze 5fC/5caC starting from ~100 cells, creating the opportunity to address this question.

Second, with the first question addressed, we can further explore an unsolved question in the field: is there a replication-independent and TDG-independent demethylation mechanism in zygotes (**Figure 4.1**)? Unlike ESCs and most other cell types, oocytes and zygotes have very low levels of *Tdg* mRNA, implying that restoration of unmodified cytosine (unmodified C) through TDG (active modification - active removal (AM-AR) pathway) may not occur[99]. Therefore, in the presence of replication inhibition, it is expected that unmodified C should not be generated. Surprisingly, locus-specific analysis of changes in 5mC+5hmC (quantified by BS-seq) and 5fC+5caC (quantified by MAB-seq) revealed a lower-than-expected gain of 5fC+5caC, indicating that a proportion of 5mC is processed to unmodified C[51]. This observation, together with other evidence showing BER activation and SSB generation accompanying TET3-mediated oxidation[100-102], suggest that an unknown mechanism involving DNA repair and independent of TDG and replication might restore unmodified cytosine (**Figure 4.1**). Given that locus-specific analysis may suffer from technical variation, it will be necessary to examine the dynamics of 5mC/5hmC and 5fC/5caC at genome-scale to determine whether unmodified C can still be generated despite replication inhibition and negligible level of *Tdg*. liMAB-seq in combination with low-input BS-seq (liBS-seq) can serve this purpose.

Thirdly, the property of active DNA demethylation during this process, for example the processivity of TET, is not clear[46]. 5hmC map[103] and chromatin accessibility profile[104] have been generated recently, and a comparative analysis of these datasets together with 5fC/5caC profile generated by liMAB-seq will provide more insights into the regulation of TET processivity.

Results

5mC/5hmC decrease in paternal demethylation is coupled with 5fC/5caC increase

In the mouse zygote, the paternal genome is packed in the paternal pronucleus before its fusion with the maternal genome and thereby can be isolated for analysis. TET3-mediated oxidation of paternal genome starts around PN3 stage of the zygote and overlaps with DNA replication in timing, but these two processes appear to be independent from each other[32,34,105]. To answer some of the questions mentioned above, we collected mouse sperm and paternal pronuclei from PN5 stage zygotes and performed RRBS-based MAB-seq and BS-seq to monitor changes in 5fC/5caC and 5mC/5hmC, respectively. To rule out the effect of replication-dependent dilution and focus solely on TET-mediated oxidation, zygotes were treated with replication inhibitor Aphidicolin before collection[51,106].

liMAB-seq revealed that the global 5fC/5caC level increases from 1.35% in sperm to 7.25% in paternal pronuclei, and that this increase is particularly evident for certain genomic features such as SINE repeats, where the 5fC/5caC level increases from 1.41% to 13.20% (**Figure 4.2a**). The increase of 5fC/5caC correlates very well with the decrease of 5mC/5hmC at different genomic features (**Figure 4.2a**), as well as representative genomic loci (**Figure 4.2b**, **Supplemental Figure S4.1a**). For genomic features with little 5fC/5caC increase (e.g. promoters, 5'UTR exons, low-complexity repeats and CpG islands), 5mC/5hmC level remains unchanged or slightly increased (**Figure 4.2a**). When individual genomic regions are classified into demethylated, unchanged, or de novo methylated based on BS-seq, 5fC/5caC increase is most evident at demethylated regions **(Supplemental Figure S4.1b)**. When individual CpG sites are classified into different

groups according to the degree of 5mC/5hmC loss, 5fC/5caC gain is higher for the groups with greater 5mC/5hmC loss (**Figure 4.2c**). Conversely, CpG sites with greater 5fC/5caC gain also exhibit greater 5mC/5hmC loss (**Figure 4.2d**).

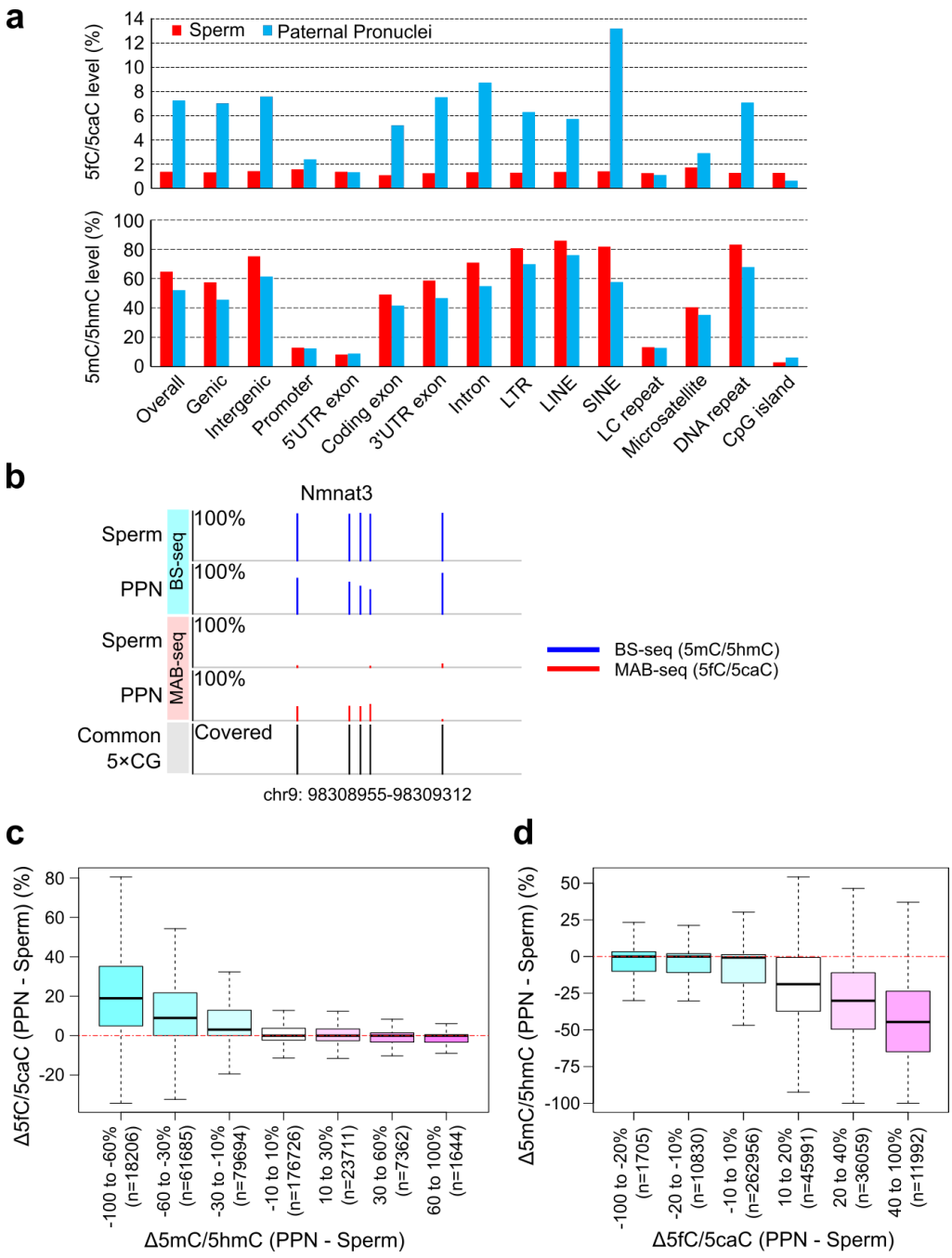

Figure 4.2 | 5mC/5hmC decrease is coupled with 5fC/5caC increase. (a) 5fC/5caC and 5mC/5hmC levels at genomic features. (b) Changes of 5mC/5hmC and 5fC/5caC from sperm to paternal pronuclei (PPN) at *Nmnat3* locus. (c) 5fC/5caC change in 7 groups of CpG sites categorized according to 5mC/5hmC change. (d) 5mC/5hmC change in 6 groups of CpG sites categorized according to 5fC/5caC changes.

Unmodified C generation accompanies 5fC/5caC generation

Despite that 5mC/5hmC decrease correlates with 5fC/5caC increase, we also noticed that the absolute amount of 5fC/5caC increase is lower than that of 5mC/5hmC decrease (**Figure 4.2c**). Because unmodified C can be quantified as 100% - (5mC+5hmC)% - (5fC+5caC)%, this observation suggests that the abundance of unmodified C increases from sperm to paternal pronuclei despite replication inhibition. To better quantify how 5fC/5caC increase matches with 5mC/5hmC decrease in absolute amount, we calculated a match index, -[Δ(5fC+5caC))/(Δ(5mC+5hmC)], for all the demethylated regions. This index will be 100% if 5fC/5caC increase completely matches 5mC/5hmC decrease. Otherwise, the index will be smaller than 100% if certain amount of unmodified C is generated. While the majority (70.7%) of the demethylated regions have gained 5fC/5caC to some degree (match index >=20%), many of them have a match index between 20% to 80% (**Figure 4.3a**). Therefore, unmodified C generation accompanies 5fC/5caC generation at many demethylated regions. The match index is largely comparable for different genomic features, but some genomic features display small but significant difference when compared with baseline (**Figure 4.3b**).

Because DNA replication was inhibited in our experiments, these results imply that a replication-independent mechanism may be responsible for restoring unmodified C during paternal genome demethylation. In addition, this mechanism is more active for specific genomic regions such as CpG island, given the lower match index of these regions. Though *Tdg* mRNA level is negligible in this context[99], it is still possible that a small amount of TDG protein exists and mediates the generation of unmodified C. Alternatively,

an unknown TDG-independent mechanism might be responsible for this result.

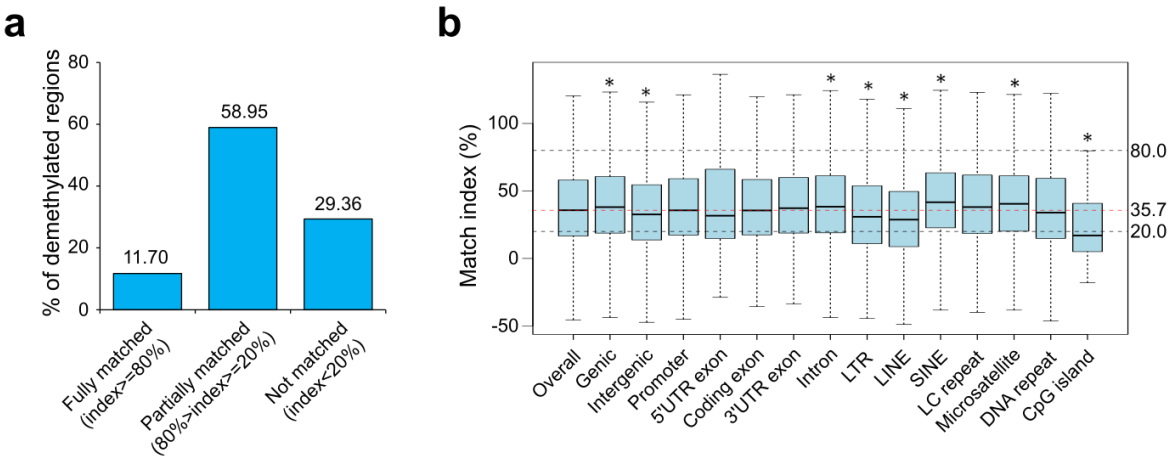

Figure 4.3 | Unmodified C generation accompanies 5fC/5caC generation at many demethylated regions. (a) Proportion of demethylated regions (250bp bins) where 5mC/5hmC decrease is fully, partially or not matched by 5fC/5caC increase. Fully, partially and not matched are defined as having a match index of >=80%, 20-80% and <20%, respectively. (b) Match index for different genomic features. For each genomic feature, a box plot is presented to summarize the match index of all the demethylated regions (250bp bins) overlapped with this feature. 35.7% is the median match index for all demethylated regions identified. 80% and 20% are the thresholds for determining whether a region is fully, partially or not matched. Mann–Whitney U test was used to determine whether a genomic feature has a higher-than-baseline match index.

Positive correlation between TET processivity and chromatin accessibility

We have previously shown that the processivity of TET enzymes positively correlates with chromatin accessibility in mouse ESCs[46]. To determine whether a similar relationship between TET processivity and chromatin accessibility exists in preimplantation embryos,

we performed liMAB-seq to map the genomic location of 5fC/5caC in 2-cell embryos. We chose to profile 2-cell embryos because the genomic distribution of 5hmC has been mapped in 2-cell embryos by TAB-seq, thereby enabling the comparative analysis of 5hmC and 5fC/5caC distribution[103]. Similar to the observation in mouse ESCs, some genomic regions in 2-cell embryos are only modified by 5hmC but not 5fC/5caC, indicating low TET processivity, while some other regions are 5fC/5caC only, indicating high TET processivity (**Figure 4.4a**). Importantly, when the genomic distribution of 5hmC and 5fC/5caC were compared with the DNase I hypersensitivity profile of the 2-cell embryos, we found that the genomic regions containing 5fC/5caC have higher DNase-seq signals compared to those of 5hmC-only regions (**Figure 4.4b**)[46,104]. These results suggest that TET processivity in 2-cell embryos is positively correlated with chromatin accessibility.

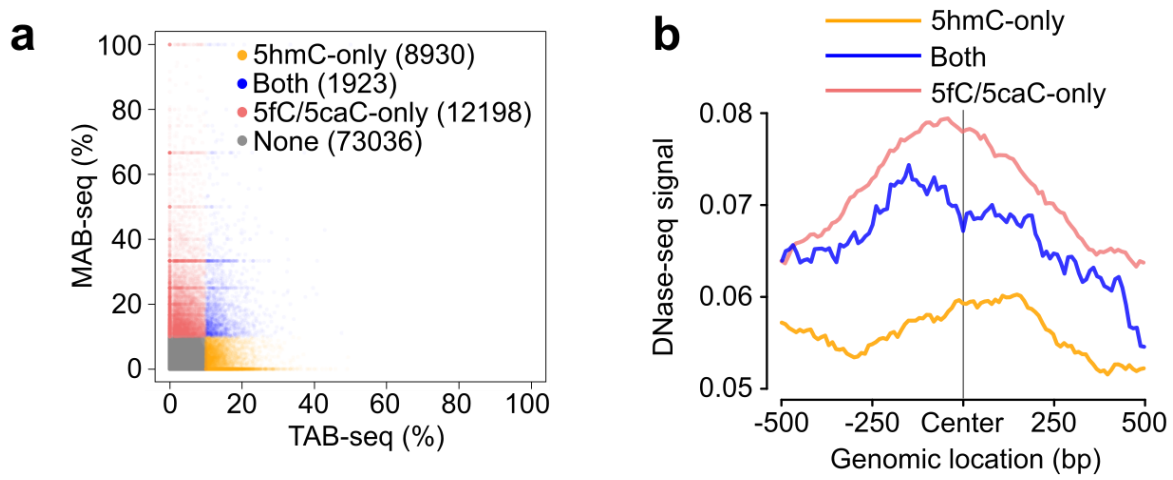

Figure 4.4 | TET processivity positively correlates with chromatin accessibility. (a) 5mC oxidation states are different in different genomic regions (100 bp bins) in 2-cell embryos. Bins with at least 3 common 5×CGs between TAB-seq and MAB-seq were analyzed, and >=10% is required to be regarded as modified. The number in the parentheses refers to the number of 100 bp bin within the indicated group. (b) TET processivity positively correlates with chromatin accessibility. Mean DNase-seq signals centering on the center of the indicated groups of 100 bp bins were plotted. DNase-seq signal is higher in regions with 5fC/5caC as compared to regions with 5hmC only.

Discussion

To demonstrate the utility of liMAB-seq and address unanswered questions of zygotic paternal genome demethylation, we generated the first genome-scale 5fC/5caC profiles of mouse sperm, zygotic paternal pronuclei and 2-cell embryos. Our study shows that a high level of 5fC/5caC is generated at genomic features such as SINE repeats. In addition, by comparing BS-seq and MAB-seq results, we provided the first sequencing-based evidence that 5mC/5hmC decrease and 5fC/5caC increase are coupled at individual genomic regions. Interestingly, we observed that despite replication inhibition, the absolute amount of 5fC/5caC increase is lower than that of 5mC/5hmC decrease, suggesting that unmodified C is also generated. Two possibilities may explain this observation. Firstly, even though *Tdg* mRNA level is low in zygotes, it is possible that a small amount of TDG protein exists and mediates the restoration of unmodified C by excising 5fC/5caC. Secondly, it is also possible that an unknown mechanism independent of TDG and replication is responsible for restoring unmodified C. To distinguish between these two possibilities, the best way is to perform MAB-seq and BS-seq profiling of Tdg knockout sperm and paternal pronuclei in the presence of replication inhibition. If unmodified C is still generated in this setup, then the second possibility is supported and worth further pursuing. Finally, by analyzing 5fC/5caC, 5hmC and DNase I hypersensitivity profiles of 2-cell embryos, we revealed that TET a positive correlation between TET processivity and chromatin accessibility.

Methods

Collection of paternal pronuclei from aphidicolin-treated zygotes

All animal studies were performed in accordance with guidelines of the Institutional Animal Care and Use Committee at Harvard Medical School. Adult B6D2F1/J females were superovulated by injecting 7.5 I.U of PSMG (Millipore) and hCG (Millipore) followed by mating with B6D2F1/J males. At 18 hrs post-hCG injection, PN0-PN1 stage (G1 phase) zygotes that did not have visible pronuclei were collected. They were cultured in KSOM containing 3 μg/ml aphidicolin (Sigma-Aldrich) in a humidified atmosphere of 5% CO2/95% air at 37.8°C. Seven hours later, zygotes reached the PN5 stage and were then transferred into M2 media containing 10 μM cytochalasin B (Sigma-Aldrich). Zona pellucidae were cut by a Piezo impact-driven micromanipulator (Prime Tech Ltd., Ibaraki, Japan). The paternal pronuclei were isolated from the zygotes and washed with PBS containing 0.2% BSA followed by collection into 0.2 ml PCR tubes. The samples were stored in -80C until use. The remaining cytoplasms containing maternal pronuclei were immunostained with anti-H3K9me3 antibody to confirm that the remaining pronuclei were maternal. The paternal pronuclei were distinguished from the maternal counterpart by the distance from the second polar body and by the pronuclear size.

RRBS-based liMAB-seq and regular MAB-seq

Experimental procedure and data analysis were performed as described in Chapter 3.

Published datasets

We used the following published datasets of mouse 2-cell-stage embryos: DNase-seq[104]

and TAB-seq[103]. The text and figures in this chapter were adapted from our papers on *Genes & Development*[58] and *Nature Reviews Genetics*[10].

Chapter 5:

scMAB-seq allows analysis of cell-to-cell heterogeneity of 5fC/5caC distribution, mapping of sister chromatid exchanges, and lineage reconstruction of mouse preimplantation embryos

Abstract

In the genome, 5fC and 5caC are undergoing active turnover resulted from TET-mediated generation, TDG-mediated excision and replication-dependent dilution. Therefore, compared with the relatively stable DNA methylome, DNA demethylome varies between cell types and among single cells. With the successful development of single-cell MAB-seq (scMAB-seq), it is now possible to analyze 5fC/5caC profile at single-cell level to reveal information that is not captured by non-single-cell studies. In Chapter 5, I present scMAB-seq analysis of single mouse ESCs and single blastomeres from mouse 2-cell and 4-cell-stage embryos. We first compared 5fC/5caC profiles of single cells from different cell types and showed that individual cells display cell-type-specific distribution of 5fC/5caC. We then analyzed the strand distribution of 5fC/5caC and revealed that 5fC/5caC is not maintained during DNA replication. The replication-induced biased-distribution of 5fC/5caC between the template strand and the newly synthesized strand allowed us to map the genomic locations of sister chromatid exchange (SCE), a type of genomic rearrangement associated with genomic instability. Finally, the pattern of SCE revealed by scMAB-seq can also be used for lineage reconstruction.

Background

In the genome, 5fC and 5caC are constantly undergoing turnover resulted from TET-mediated generation, TDG-mediated excision and replication-dependent dilution. Understanding cell-to-cell heterogeneity of 5fC/5caC can help us understand how active DNA demethylation is regulated at the single cell level. scMAB-seq, as validated in Chapter 3, can potentially provide insights from the following perspectives:

First, MAB-seq analysis of a population of cells (Chapter 2 and 4) suggests that different cell types have distinct patterns of 5fC/5caC distribution. As a result, it is expected that single cells of different cell types should possess some of these cell-type-specific features. It will be interesting to see whether this type of heterogeneity resulted from the difference of cell types can be captured by scMAB-seq, and whether cell type classification based on scMAB-seq profiles is feasible.

Second, though immunostaining suggests that 5fC/5caC is not maintained during DNA replication[34], sequencing-based confirmation is lacking. If this model is correct, during DNA replication, the newly synthesized strand should have less 5fC/5caC compared with the old template strand (**Figure 5.1a**). Because MAB-seq is a strand-specific method capable of distinguishing 5fC/5caC signals from Watson (top) and Crick (bottom) strands, scMAB-seq should be able to capture this biased-distribution of 5fC/5caC between strands.

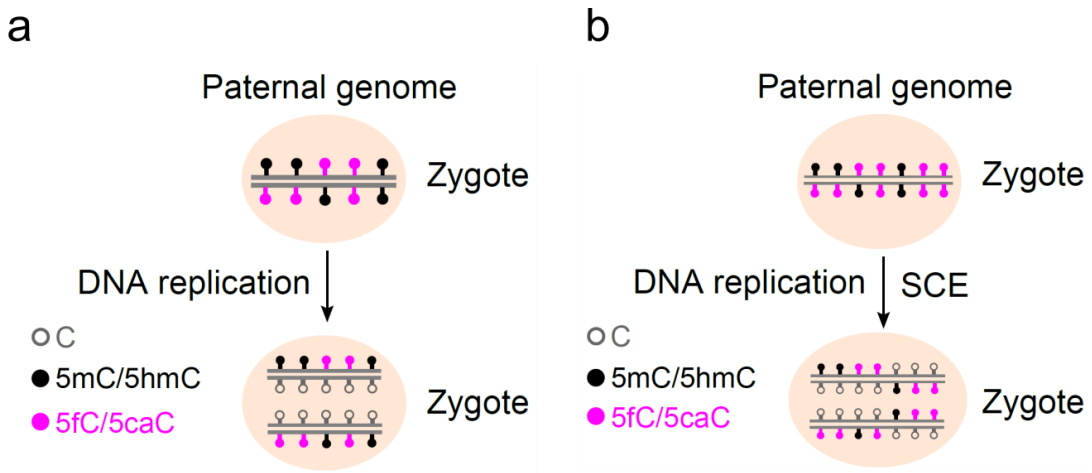

Figure 5.1 | Replication-dependent dilution of 5fC/5caC and SCE in mouse preimplantation development. (a) DNA replication results in biased-distribution of 5fC/5caC between the two DNA strands. (b) SCE results in a switch of 5fC/5caC distribution from one strand to another.

Third, immunostaining of 5fC/5caC in preimplantation embryos revealed that after replication, 5fC/5caC signals can sometimes switch from one DNA strand to another (**Figure 5.1b**)[34]. This switch reflects a phenomenon called sister chromatid exchange (SCE). SCE is caused by homologous recombination happening between the two sister chromatids during DNA replication, and has been observed in abnormally high frequency in diseases associated with genomic instability, for example cancer and Bloom Syndrome[107]. Current methods for analyzing SCE suffer from limitations including reagent-induced artifacts and low spatial resolution, limiting our understanding of the cause, consequence and biological significance of this type of genomic rearrangement[107]. If scMAB-seq can successfully capture the strand-biased distribution of 5fC/5caC resulted from DNA replication, it should also be able to map SCE at high resolution, facilitating the study of SCE.

Finally, if SCE can be successfully mapped by scMAB-seq, lineage reconstruction based on the pattern of SCE will be possible. In developmental biology and cancer biology, a crucial question is where cells of distinctive features originate from and evolve to. To address this question, different techniques have been employed to track cell lineage, but many of these methods rely on markers that are experimentally delivered[108]. On the other hand, SCE can serve as an endogenous marker for lineage tracing, because cells with closer relationships will share more SCEs. Therefore, single-cell 5fC/5caC profile generated by scMAB-seq can be used to reconstruct the lineage relationship of a group of cells.

Results

scMAB-seq captures cell-to-cell heterogeneity of 5fC/5caC resulted from the difference of cell types

With the successful development of scMAB-seq, we decided to first examine cell-to-cell heterogeneity of 5fC/5caC distribution between different cell types. To this end, we performed scMAB-seq analysis of 15 TDG-depleted ESCs (4 PBAT-based and 11 RRBS-based) and 26 blastomeres from 18 2-cell embryos (all are RRBS-based; 8 of which had both blastomeres sequenced).

To test whether scMAB-seq can capture the cell-to-cell heterogeneity resulted from the difference of cell types, we first pooled single 2-cell blastomeres and TDG-depleted ESCs for unsupervised principle component analysis (PCA) based on their 5fC/5caC profiles. In PCA, the majority of the cells are clustered according to their cell identity (**Figure 5.2a**), and PC1 largely distinguishes the two cell types (**Figure 5.2a,b**). We also analyzed the data in a supervised manner by inspecting ESC-specific and 2-cell-specific 5fC/5caC-modified regions identified by liMAB-seq, revealing that all single cells display cell-type specific 5fC/5caC patterns (**Figure 5.2c, Supplemental Figure S5.1a**). Therefore, scMAB-seq is capable of capturing the variable patterns of active DNA demethylation in different types of single cells and can be used for cell type classification. For both analyses, scMAB-seq datasets obtained from RRBS and PBAT strategies were analyzed together and supported the same conclusion (**Figure 5.2c, Supplemental Figure S5.1a**).

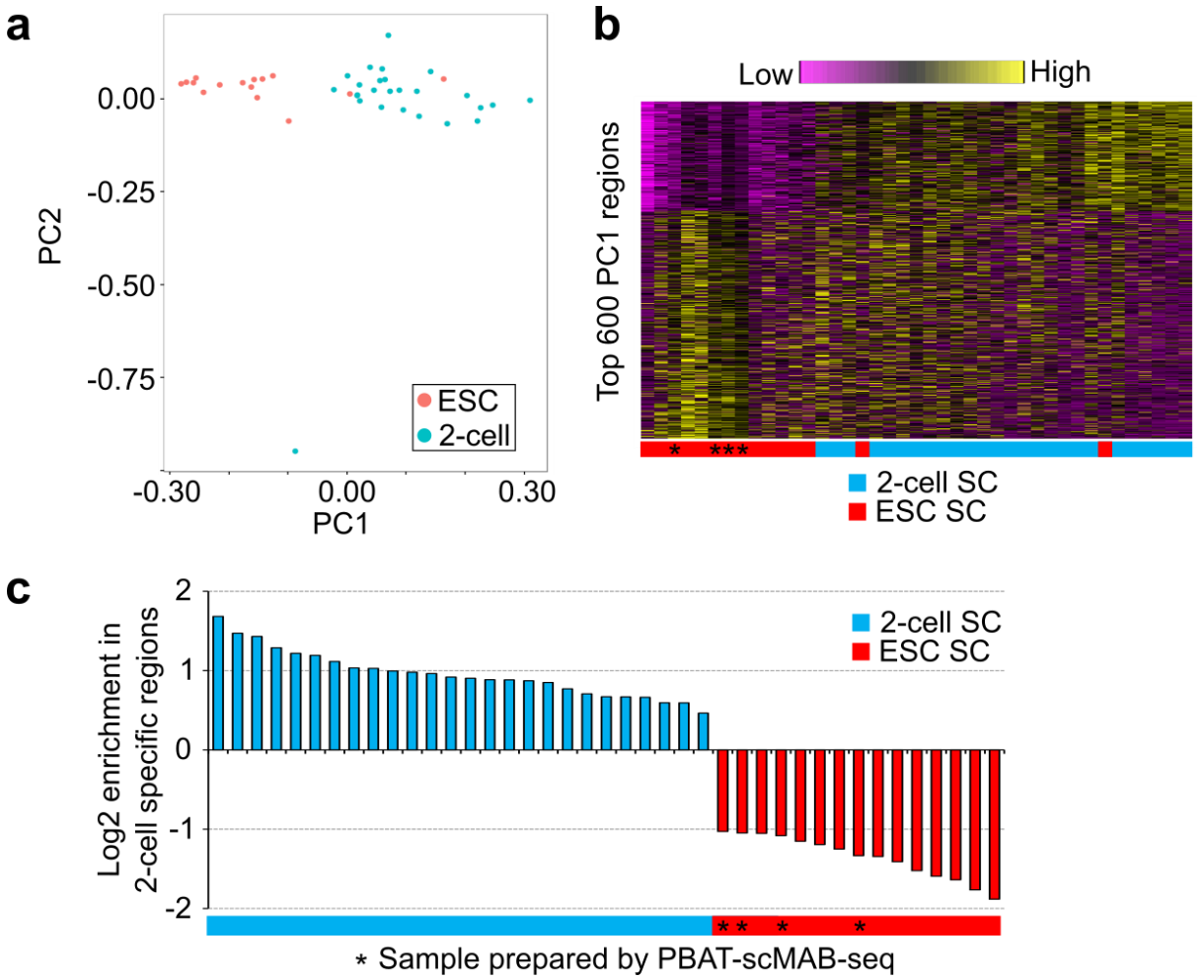

Figure 5.2 | Cell-to-cell heterogeneity of 5fC/5caC in single cells of different cell types. (a) PCA of 15 TDG-depleted mouse ESCs and 26 single 2-cell blastomeres. PC1 and 2 are shown. (b) Heatmap of 5fC/5caC of the top 600 PC1 regions. The vertical axis plots the 600 regions while the horizontal axis plots the 36 single cells. Cells are ranked based on PC1 score. (c) 5fC/5caC-modified regions (2kb genomic bin) in single cells are enriched in the corresponding cell-type-specific 5fC/5caC-modified regions identified by liMAB-seq. For each single cell, 5fC/5caC-modified regions were called, and the log2 enrichment of these regions in 2-cell specific 5fC/5caC-modified regions relative to ES cell-specific regions was calculated and plotted.

scMAB-seq captures cell-to-cell heterogeneity in 5fC/5caC resulted from DNA replication

In addition to cell type-specific patterns, cell-to-cell heterogeneity in 5fC/5caC distribution can also be introduced by DNA replication. In zygotic paternal genome, following the first round of DNA replication, the majority of the CpG sites on the newly synthesized strand are unmodified, creating a biased distribution of 5fC/5caC towards the template strand. After cell division, the two blastomeres of the 2-cell embryo should have completely complementary 5fC/5caC strand distribution (**Figure 5.3a**). To determine whether this replication-driven heterogeneity can be captured by scMAB-seq, we analyzed the strand distribution of 5fC/5caC in single 2-cell blastomeres. We divided the genome into bins (2MB or 10MB bins in **Figure 5.3b**) and calculated the strand bias of 5fC/5caC for each bin as the difference of MAB-seq signals ((T/(C+T))%) between the top and bottom strands. This analysis reveals the strand-biased distribution of 5fC/5caC at the chromosomal level, confirming replication-dependent dilution (**Figure 5.3b**). Importantly, two blastomeres from the same 2-cell embryo display complementary 5fC/5caC patterns as expected (**Figure 5.3b** and **Supplemental Figure S5.1b**). Furthermore, when 2-cell blastomeres from different embryos were pooled together for clustering analysis based on anti-correlation of 5fC/5caC pattern, two blastomeres from the same 2-cell embryo always cluster together due to their total complementary patterns (**Figure 5.3c**). These results provide the first sequencing-based evidence at the single-cell level that 5fC and 5caC are diluted by DNA replication, creating a complementary 5fC/5caC pattern in the two daughter cells.

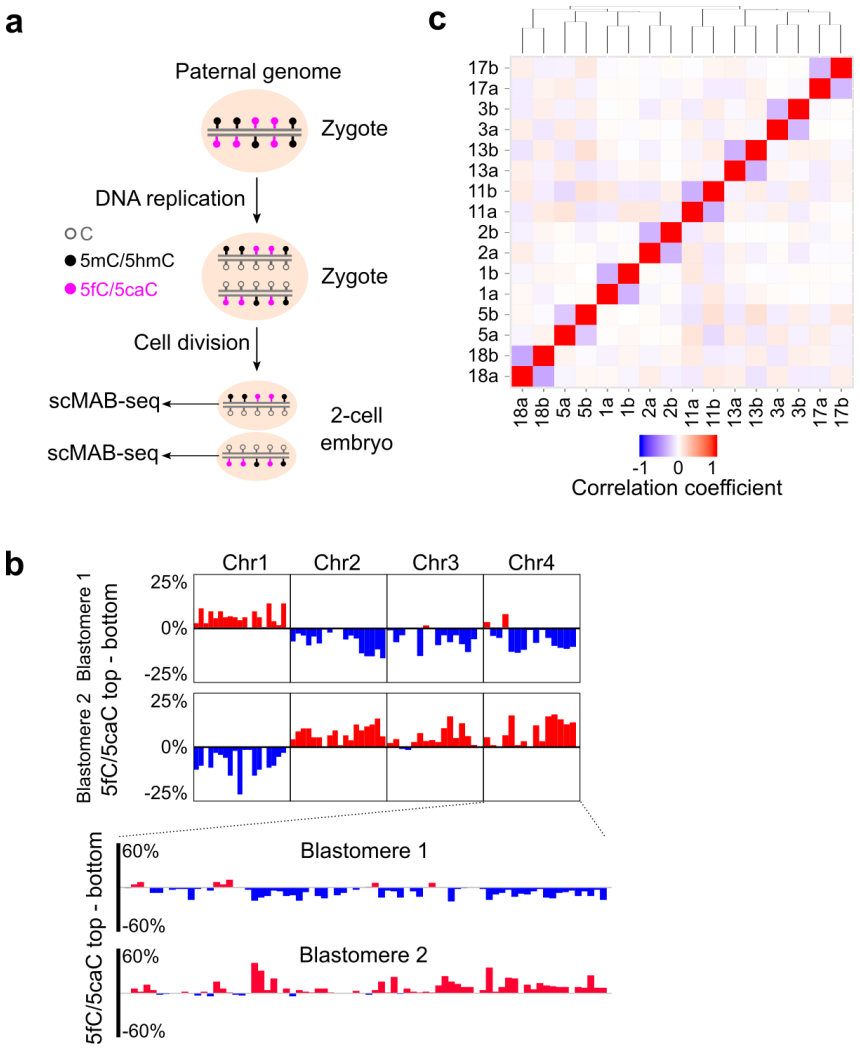

Figure 5.3 | Cell-to-cell heterogeneity of 5fC/5caC resulted from DNA replication. (a) Replication-dependent dilution of 5fC/5caC. (b) Complementary 5fC/5caC pattern observed in a pair of 2-cell blastomeres. Strand-specific MAB-seq signals were binned into 10 MB bins (top panel) and 2 MB bins (bottom panel), and the strand bias (the difference of signals between top and bottom strands) is calculated for each bin. (c) Clustering of 2-cell blastomeres based on anti-correlation of 5fC/5caC strand bias. Two blastomeres from the same 2-cell embryo are labeled as “number + a/b”. For any two blastomeres, a correlation coefficient (R) was calculated based on the strand bias of 1MB bins. Hierarchical clustering was then performed using (R+1)/2 as distance.

scMAB-seq allows mapping of SCE in mouse 2-cell embryos

In 2-cell embryos, another interesting phenomenon observed by 5fC/5caC immunostaining is sister chromatid exchange (SCE)[34]. SCE, resulted from homologous recombination taking place between the two sister chromatids during DNA replication, has been associated with genomic instability and diseases[107]. Currently, the most commonly used analysis method for SCE is bromodeoxyuridine (BrdU)-incorporation followed by staining. However, BrdU treatment itself can induce SCE, and staining provides a low-resolution metric regarding the genomic location of SCE[107]. Consequently, despite the discovery of SCE decades ago, its cause, genomic location, consequence, and biological significance are not fully understood.

In zygotes, the vast majority of 5fC/5caC are generated from the paternal genome[34]. A replication-coupled SCE happened on the paternal genome will lead to the switching of overall 5fC/5caC distribution from top (+) strand to bottom (-) strand or vice versa in the two daughter cells (**Figure 5.4a** and **Supplemental Figure S5.2a**), making the mapping of the genomic distribution of naturally-occurring SCE by scMAB-seq possible. Indeed, by analyzing the strand distribution of 5fC/5caC of single 2-cell blastomeres, we observed SCEs at the same location in the two blastomeres from one embryo (**Figure 5.4b** and **Supplemental Figure S5.1b**). When the region surrounding an SCE is sufficiently covered by sequencing and modified by 5fC/5caC, the SCE can be fine mapped to a small 30 kb genomic region (**Figure 5.4c**). In general, the analysis of both blastomeres from one 2-cell embryo or one of the two blastomeres by RRBS-based scMAB-seq can map SCE to a median resolution of 700 kb and 1250 kb, respectively (**Supplemental**

Figure S5.2b-c). These resolutions are much higher than the conventional BrdU immunostaining which has a resolution of a few megabases or worse.

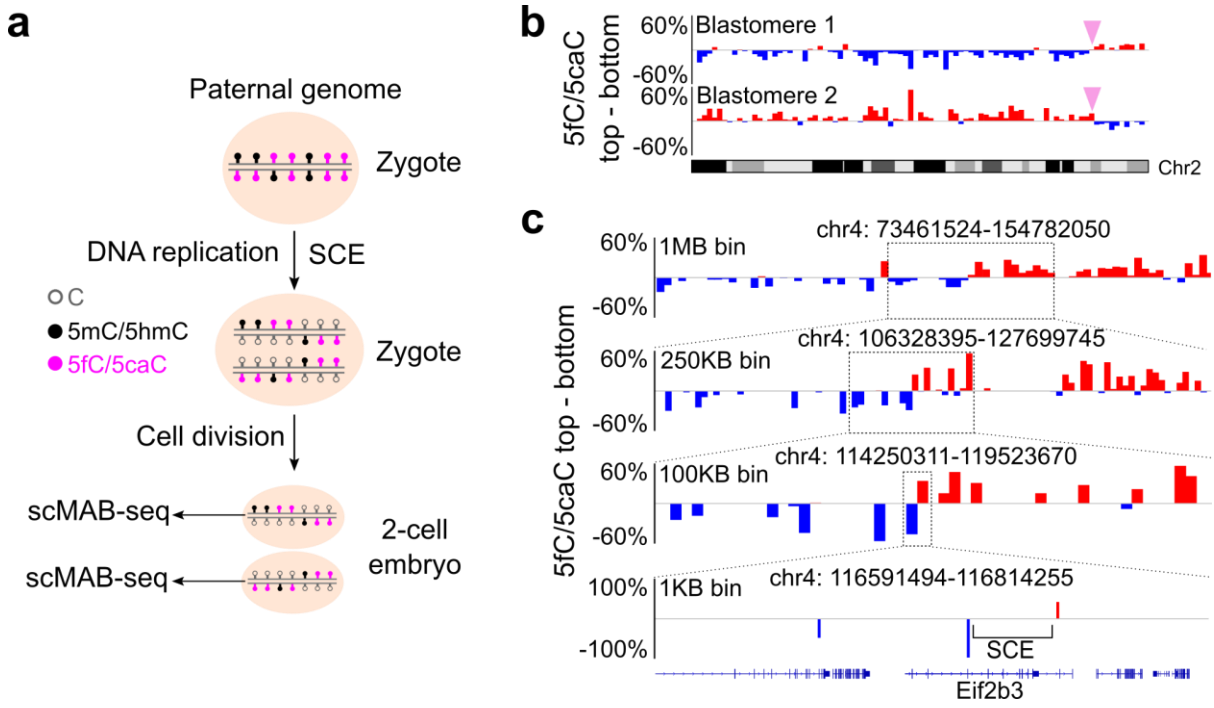

Figure 5.4 | scMAB-seq allows mapping of SCE in mouse 2-cell embryos. (a) Schematic representation of SCE in preimplantation embryos. (b) SCE in a representative pair of blastomeres from one 2-cell embryo. Each bar represents a 2 MB bin. For each bin, the difference of 5fC/5caC between the top (+) strand and the bottom (-) strand was calculated. The switch of strand bias from one strand to another suggests that an SCE has occurred, and the switch is at the same genomic position in the two daughter cells as expected. (c) High-resolution mapping of SCE. The rough location of SCE can be identified through binning the 5fC/5caC signals into 1MB bins and calculating the strand bias. By using smaller bins (250kb, 100kb and 1kb) for the calculation and zooming in further, SCE can be fine mapped to a 30kb region within the Eif2b3 gene.

We identified a total of 75 paternal autosomal SCEs in the 18 2-cell embryos analyzed (4.17 per embryo, **Figure 5.5a**). This frequency is comparable to that reported in ESCs[109]. The number of paternal autosomal SCEs in individual embryos ranges from 1 to 9, with the majority of the cells having 3 to 4 paternal autosomal SCEs (**Figure 5.5b**). The distribution of the SCEs is generally random (**Figure 5.5a**), but specific chromosomes (for example chromosome 14; **Figure 5.5c**) or genomic regions (arrowheads in **Figure 5.5a**) might have a higher SCE frequency.

Interestingly, we also observed two artificial SCEs on chromosome 10 and 14 (**Figure 5.5d** and **Supplemental Figure S5.2d**). These two artificial SCEs were observed at the same locations in every embryo examined, a frequency that is abnormally high compared with other SCEs identified (**Supplemental Figure S5.2e**). Further analysis and literature search suggested that they are not real SCEs, and the observed switch of strand bias is caused by two disoriented contigs in the mouse mm9 genome assembly[109]. In mouse mm10 genome assembly, the disoriented contig on chromosome 10 has already been corrected while the disoriented contig on chromosome 14 still remains (**Figure 5.5d** and **Supplemental Figure S5.2d**). Therefore, scMAB-seq allows the mapping of SCE at high resolution and can be used for identifying assembly errors of a reference genome.

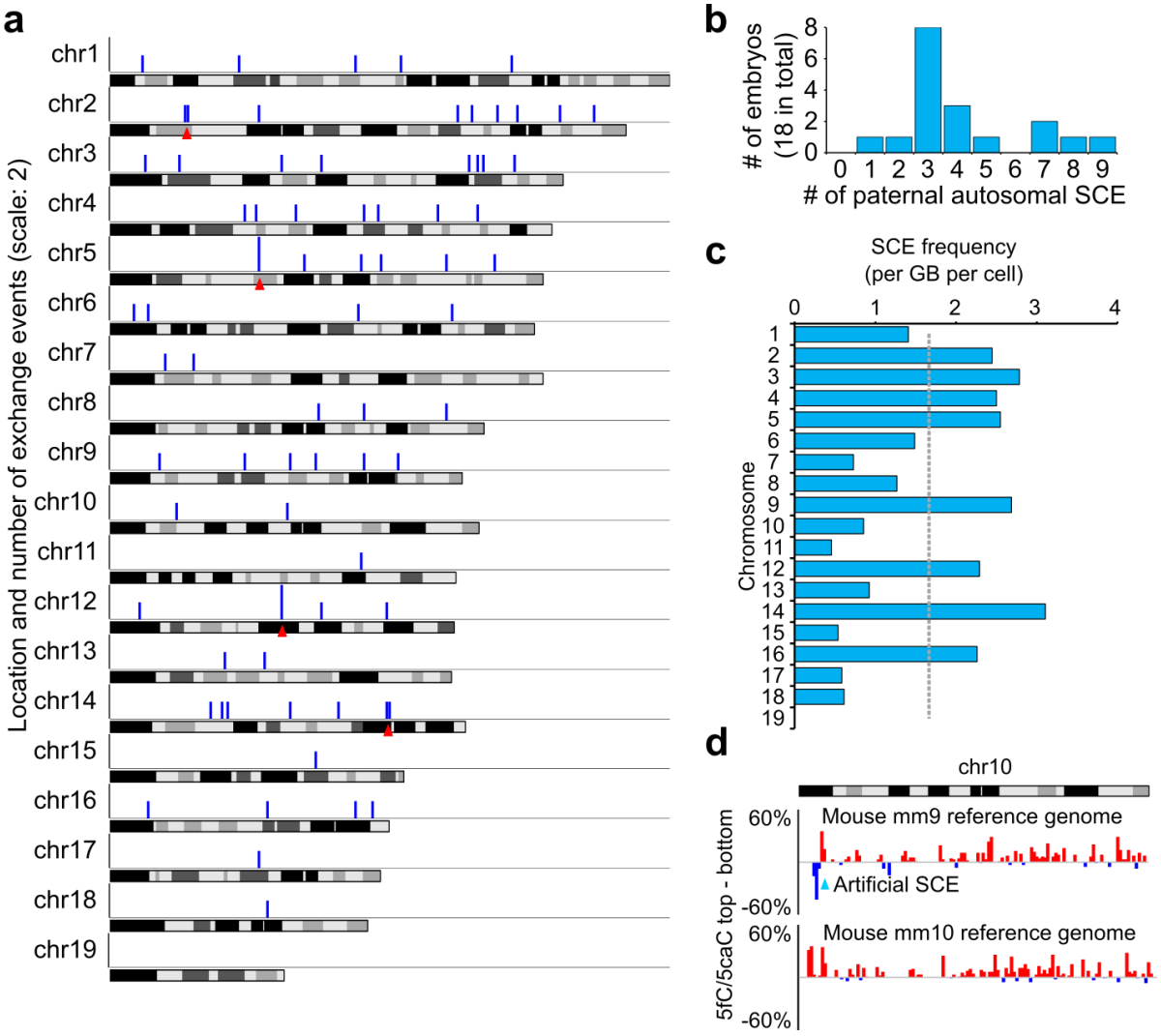

Figure 5.5 | Distribution of SCE in in mouse 2-cell embryos. (a) Chromosome distribution of the detected SCE events. Each blue bar represents a 1MB bin with SCE, and the height represents the number of SCE observed in this 1MB bin (scale: 2). (b) Histogram of the number of paternal autosomal SCE in each embryo. (c) Paternal SCE frequency for each chromosome. Frequency is calculated as number of paternal SCE per GB per cell. (d) Artificial SCE identified in mm9 chr5: 7,000,000 – 8,000,000. This artificial SCE is caused by an error in mm9 genome assembly, and the error has been fixed in mm10 genome assembly. Each bar represents the strand bias calculated for a 1MB bin.

scMAB-seq allows lineage reconstruction of mouse 4-cell embryos

After cell division, SCE occurred during the last cell cycle will pass down to the two daughter cells. When multiple rounds of cell division occur, the daughter cells will carry SCE information from multiple cell cycles, with older SCEs shared by more cells and the youngest SCEs from recent cell cycle will be shared only by the two daughter cells of the same mother cell (**Figure 5.6**). In other words, cells with closer lineage relationship should share more SCEs. To test the possibility of using SCE pattern to reconstruct cell lineage, we performed scMAB-seq of 12 blastomeres from three 4-cell stage mouse embryos (4 by PBAT-based protocol and 8 by RRBS-based protocol).

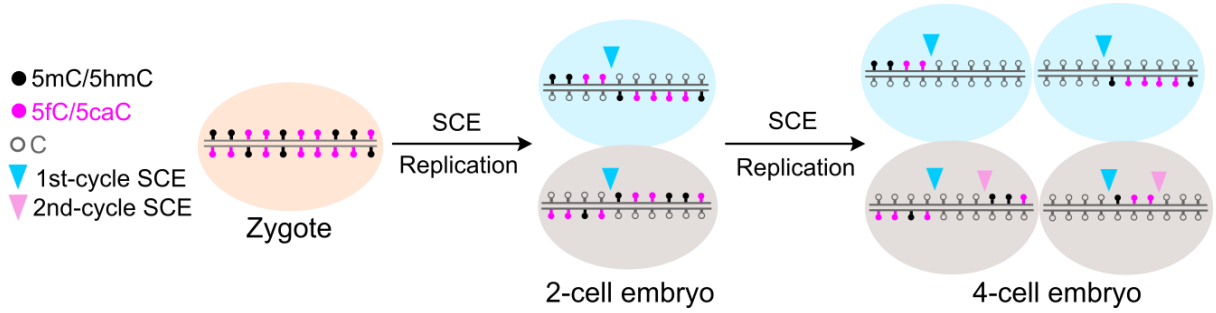

Figure 5.6 | Pattern of SCE can be used for reconstructing lineage relationship among single cells. 4-cell blastomeres carry SCE information from multiple cell cycles. SCEs occurred during the first cell cycle will be shared by all 4 blastomeres while those occurred during the second cell cycle will only be shared by the 2 blastomeres from the same mother cell.

In 4-cell embryos, 5fC/5caC has undergone two rounds of replication-dependent dilution. As a result, paternal SCE is expected to cause a shift of 5fC/5caC from biased towards one strand to largely unbiased (**Figure 5.6**). Indeed, SCE can be readily mapped based on this principle using either RRBS-based (**Figure 5.7a**) or PBAT-based (**Figure 5.7b**)

scMAB-seq.

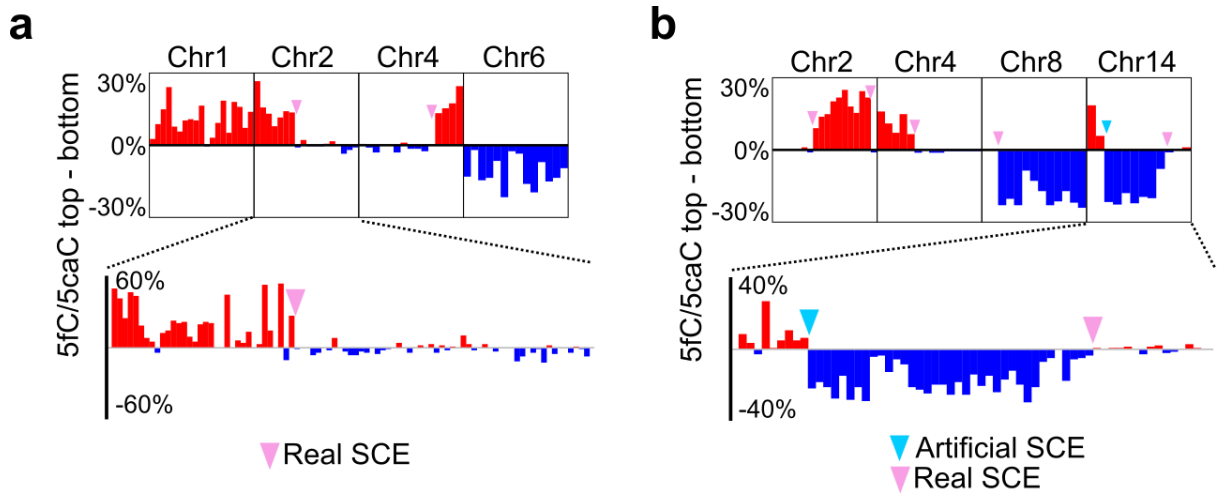

Figure 5.7 | SCE mapping in mouse 4-cell blastomeres. (a) SCE identification in a 4-cell blastomere by RRBS-based scMAB-seq. Real SCEs occurred on chromosome 2 and 4 were identified as a switch of 5fC/5caC distribution from biased towards one strand to largely unbiased. Top panel: 10 MB bin. Bottom panel: 2 MB bin. (b) SCE identification in a 4-cell blastomere by PBAT-based scMAB-seq. Real SCEs occurred on chromosome 2, 4, 8 and 14 were identified as a switch of 5fC/5caC distribution from biased towards one strand to largely unbiased. The artificial SCE on chromosome 14 was identified as a switch of 5fC/5caC from biased towards the top to the bottom. Top panel: 10 MB bin. Bottom panel: 2 MB bin.

With the SCE mapped, we then pooled the 12 4-cell blastomeres for clustering analysis based on the genomic location of SCE. The analysis shows that blastomeres from the same 4-cell embryo not only cluster together but also form clear lineage relationship (**Figure 5.8a**). Inspection of individual chromosomes further confirmed this lineage relationship (**Figure 5.8b**). SCEs took place during the first or second cell cycles (shared

by all 4 cells or only 2 cells, respectively) are also evident (**Figure 5.8b**). Therefore, SCE history established by scMAB-seq analysis can be used for cell lineage reconstruction.

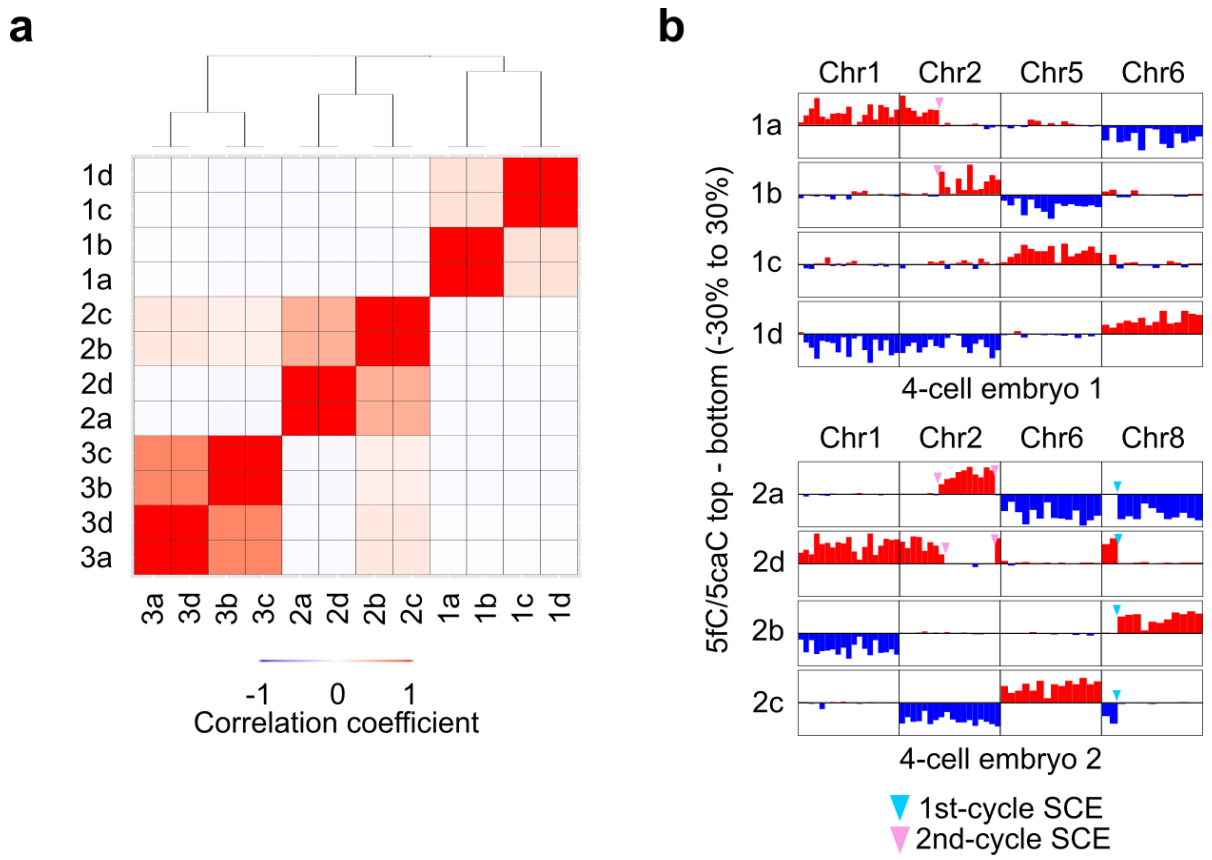

Figure 5.8 | Lineage reconstruction of mouse 4-cell blastomeres. (a) Clustering of 12 4-cell blastomeres from 3 embryos by SCE pattern. The first number denotes the embryo. For any two blastomeres, a correlation coefficient (R) was calculated by comparing the SCE pattern. Hierarchical clustering was then performed using (1-R)/2 as distance. Four blastomeres from the same 2-cell embryo cluster together and display lineage relationship. (b) Inspection of individual chromosomes further supports the lineage relationship established by clustering. Two embryos, one prepared by RRBS (top) and one by PBAT (bottom), are shown. For embryo 1, 1a-1b and 1c-1d are two pairs of daughter cells. For embryo 2, 2a-2d, 2b-2c are two pairs of daughter cells. SCEs occurred during the first cell cycle will be shared by all 4 blastomeres, while those occurred during the second cell cycle will only be shared by the 2 blastomeres from the same mother cell.

scMAB-seq reveals the strand bias of 5fC/5caC in mouse ESCs

The mouse preimplantation embryo is a unique system for studying 5fC/5caC strand bias and mapping SCE because 5fC and 5caC are mainly present on the paternal genome and largely absent in the newly synthesized strand after DNA replication[34]. However, it has been unclear whether scMAB-seq can capture the strand bias of 5fC/5caC and identify SCE in more complex biological contexts such as mouse ESCs, where both the paternal and maternal copies of the genome are modified, and where both methylation and demethylation machineries are functional to maintain DNA methylation homeostasis after DNA replication. To address this question, we analyzed 5fC/5caC strand distribution of single TDG-depleted ESCs, taking into consideration that both paternal and maternal copies of the genome are modified (**Figure 5.9**).

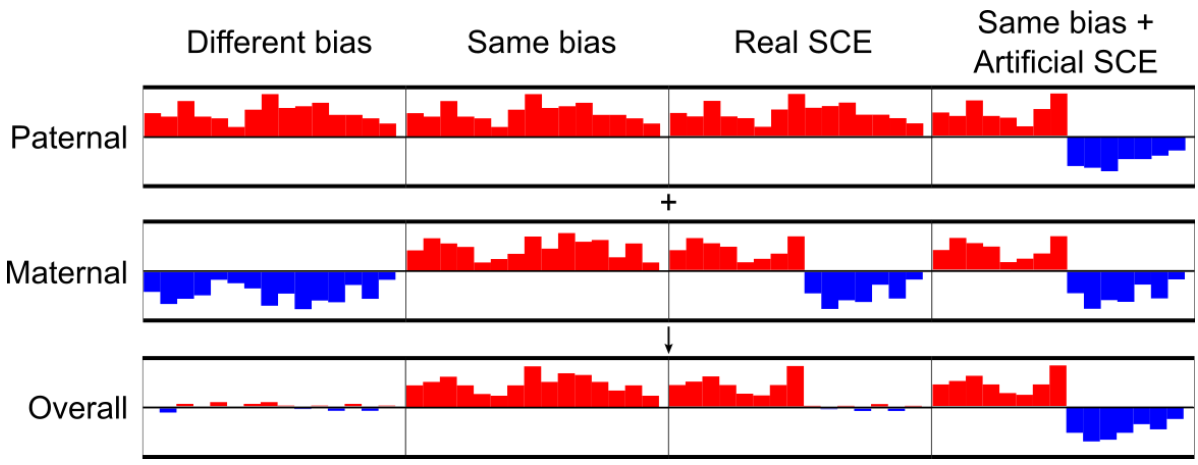

Figure 5.9 | Schematic diagram illustrating strand-biased distribution of 5fC/5caC in mouse ESCs. In mESCs, both the paternal and maternal genomes are modified by 5fC/5caC with strand-biased distribution. When the paternal bias and maternal bias are in opposite directions, the overall bias detected by sequencing is close to 0 (largely unbiased). When the two biases are towards the same direction, the overall bias should be towards that direction. When a real SCE happens, a switch from biased towards one strand to largely unbiased will be observed. In the case of artificial SCEs, if both the paternal and maternal copies are biased towards the same direction, the overall profile should be a transition from top to bottom or vice versa.

Our analysis suggested that 10 out of 15 TDG-depleted ESCs exhibit different degrees of 5fC/5caC strand bias (**Figure 5.10a)**. This high proportion of cells with strand bias suggests that after DNA replication, the re-establishment of 5fC/5caC on the newly synthesized strand is a relatively slow process, despite the presence of functional methylation and demethylation machineries.

Similarly, the strand bias of 5fC/5caC allows us to map SCE and identify disoriented

contigs (**Figure 5.10b,c**). Importantly, the strand bias, SCE and disoriented genome assembly can be identified by both RRBS-based (**Figure 5.10b**) and PBAT-based **(Figure 5.10c, Supplemental Figure S5.3)** scMAB-seq, confirming that both strategies reach a similar conclusion.

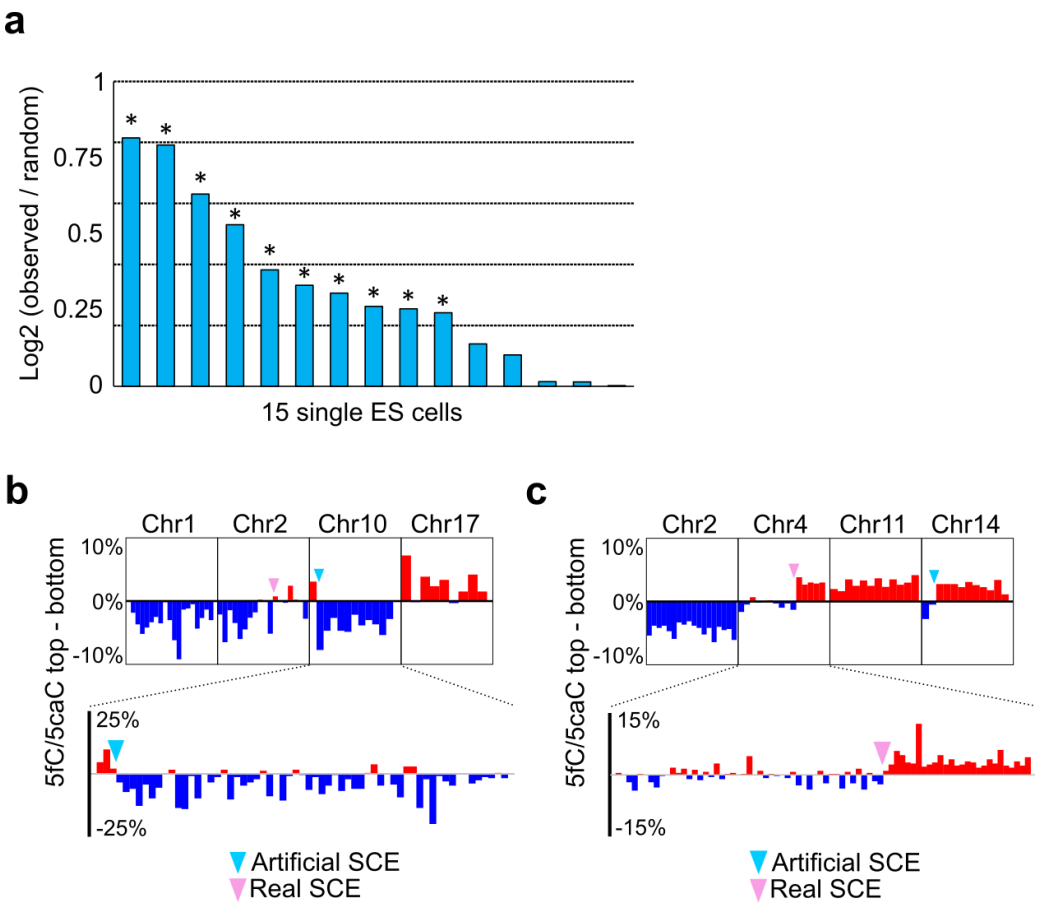

Figure 5.10 | scMAB-seq captures strand bias of 5fC/5caC and SCE in mouse ESCs. (a) Individual ESCs display different degrees of strand bias of 5fC/5caC. For each single cell, the degree of strand bias is shown as log2 (observed bias / mean bias from 500 random samplings). Asterisk denotes statistically significant bias. (b) RRBS-based scMAB-seq captures 5fC/5caC strand bias and SCE in mouse ESCs. Chromosome 2 is an example for a real SCE as shown by the transition of 5fC/5caC from biased towards the bottom to center around 0. Chromosome 10 is an example for an artificial SCE where 5fC/5caC transits from biased towards the top to the bottom. Top panel: 10 MB bin. Bottom panel: 2 MB bin. (c) PBAT-based scMAB-seq captures 5fC/5caC strand bias and SCE in mouse ESCs. Chromosome 4 and 14 serve as examples for real and artificial SCEs, respectively. Top panel: 10 MB bin. Bottom panel: 2 MB bin.

Discussion

To demonstrate the utility of scMAB-seq and address questions that cannot be answered by non-single-cell techniques, we analyzed single mouse ESCs and single blastomeres of mouse 2-cell and 4-cell embryos. Our analysis has revealed various biological insights.

First, we uncovered the heterogeneity of 5fC/5caC distribution between different cell types and among single cells of the same cell type. Compared to scBS-seq, which captures the relatively stable 5mC/5hmC, scMAB-seq captures active DNA demethylation events in individual cells, providing a powerful tool for understanding how TET activity might be regulated in individual cells. In addition, single-cell 5fC/5caC profile may be used for cell type classification, for example categorizing a cell population that are viewed as homogenous by other single-cell methods.

Second, using our scMAB-seq datasets, we analyzed the strand distribution of 5fC/5caC in singe mouse ESCs and single blastomeres from mouse 2-cell and 4-cell embryos, revealing that 5fC/5caC is not maintained during DNA replication. The biased distribution of 5fC/5caC between the template strand and the newly synthesized strand allowed us to map the genomic locations of SCE, a type of genomic rearrangement associated with genomic instability. Compared with conventional methods such as BrdU incorporation followed by staining, scMAB-seq-based SCE mapping yields a much higher genomic resolution in a BrdU-free manner.

Third, the pattern of SCE revealed by scMAB-seq can also be used for lineage tracing. In

the future, a combination of scMAB-seq based lineage tracing with functional analyses may provide insights into the mechanism and function of lineage specification.

Two methods, scAba-seq and nano-5hmC-Seal, that respectively map 5hmC using single cells or around 1000 cells, have been recently published[110,111]. Compared with these two methods which enrich 5hmC-modified regions through restriction enzyme digestion or chemical labeling, liMAB-seq and scMAB-seq are based on BS-seq and do not depend on enrichment, allowing a more quantitative analysis of the modifications and a direct comparison with BS-seq data. Moreover, while 5hmC can be a relatively stable epigenetic mark[28], 5fC and 5caC, being efficiently excised by TDG in most biological contexts, are markers of ongoing DNA demethylation[31,43,46]. With these unique features and advantages, scMAB-seq and liMAB-seq can complement scAba-seq and nano-5hmC-Seal to reveal different layers of information on DNA methylation dynamics.

Methods

Unsupervised PCA of single-cell 5fC/5caC profile

Raw scMAB-seq data were digitally transformed as described before in Chapter 2. 5fC/5caC profiles of individual cells were obtained through a sliding window approach (2 MB window and 500 kb step). For each window, the level of MAB-seq signal was calculated as the weighted mean of all CGs in the window, with sequencing coverage as the weight and without considering strand information. Background signals detected in single Tet TKO ESCs were then subtracted from MAB-seq signals of samples of interest to obtain the absolute levels of 5fC/5caC. The data were square root-transformed, and PCA was performed using the R package Seurat[112].

Analysis of strand bias of 5fC/5caC

Raw scMAB-seq data were first digitally transformed as described before in Chapter 2. The genome was segregated into strand-specific 1 MB bins (top (+) and bottom (-)). For bins with at least 10 5×CGs, the level of 5fC/5caC was calculated as the weighted mean of all 5×CGs in the bin, with sequencing coverage as the weight. For bins with less than 10 5×CGs, which indicate poor sequencing coverage, the level of 5fC/5caC was regarded as 0. For each 1 MB genomic window, the strand bias was calculated as the level of 5fC/5caC on the top bin minus that of the bottom bin. Other bin sizes were also used for visualizing the strand bias or mapping SCE at high-resolution. When using 500 kb bins, we only calculated the level of 5fC/5caC for bins with at least 5 5×CGs and regarded that of the other bins as 0. When using 100kb bins, the cutoff was at least 2 5×CGs. When using bins larger than 1MB, the cutoff was at least 10 5×CGs. For single mouse ESCs,

we also calculated an overall strand bias for each cell to represent the overall difference of the level of modifications between the old and new strands. For each chromosome, we calculated a mean strand bias for all 5 MB windows and used the absolute value of this mean as the chromosome strand bias. For a cell without strand bias, this value should be closed to 0. The overall strand bias for a cell is calculated as the mean of the chromosome strand bias multiplied by 2. The reason to multiply by 2 is that for a cell with strand bias, there is 50% probability that the paternal and maternal bias for a chromosome will be towards different directions and cancel each other out. In order to test whether a cell is significantly biased or not, we randomly shuffled the strand bias of 5MB windows and recalculated an overall strand bias. This random shuffling process was performed 500 times to establish a sample distribution for performing a two-tailed permutation test. In **Figure 5.10a**, $\log_2$ (observed bias / mean of random bias) was presented to show the extent of the bias. For all the analysis of strand bias, sex chromosomes were excluded.

RRBS-based and PBAT-based scMAB-seq

Experimental procedure and raw data analyses are the same as described in Chapter 3.

Adaptation of published work

The figures in this chapter were adapted from our paper on *Genes & Development*[58].

Chapter 6:

Conclusions and future directions

Development of MAB-seq for base-resolution, quantitative analysis of 5fC/5caC

To facilitate the study of active DNA demethylation, we developed methylase-assisted bisulfite sequencing (MAB-seq) capable of profiling 5fC and 5caC (5fC/5caC) at single-base resolution and in a quantitative manner. By combining bisulfite sequencing (BS-seq) with *M.SssI* treatment, 5fC and 5caC can be identified as T during sequencing, while unmodified C, 5mC and 5hmC are read as C (**Figure 2.1**).

Compared with affinity enrichment-based methods for mapping 5fC or 5caC[31,43], MAB-seq has the following advantages. First, the method has higher spatial resolution (single-base resolution in MAB-seq versus a few hundred base-pairs in enrichment-based methods), allowing extra layers of information to be revealed, for example the processivity of TET-mediated oxidation and the asymmetric pattern of 5fC/5caC at CpG dyads. Second, the method is quantitative, as the level of 5fC/5caC at a particular CpG site across a population of cells can be calculated as (T/(C+T))%. Quantification is especially important for understanding the biological meaning of 5fC and 5caC, which are relatively rare compared with 5mC and 5hmC[18]. A low abundance of 5fC/5caC at a genomic region indicates that 5fC and 5caC are more likely to represent intermediates of active DNA demethylation at this region, while a high abundance suggests that 5fC and 5caC might serve as stable epigenetic marks.

Compared with previous methods for mapping 5fC or 5caC at base-resolution, for example fCAB-seq[43], redBS-seq[53] and caCAB-seq[56] (**Figure 1.2**), MAB-seq has the following advantages. First, MAB-seq directly reads 5fC/5caC out as T in a single

experiment, while fCAB-seq, redBS-seq and caCAB-seq all require an additional BS-seq experiment for subtraction-based quantification of 5fC or 5caC, significantly increasing sequencing efforts and creating extra variations. In addition, subtraction-based methods are intrinsically not compatible with single-cell analysis, because the same single cell cannot be measured twice. Second, MAB-seq is an enzyme-based method performed in a buffer system that is compatible with most other enzymes used for sequencing library preparation. Therefore, it is possible to develop low-input and single-cell versions of the method by integrating multiple steps of library preparation into one reaction tube.

In addition to the above-mentioned advantages, MAB-seq also differs from other existing methods in that 5fC and 5caC are mapped and quantified together. Biologically speaking, both 5fC and 5caC are efficiently recognized and excised by TDG. As a result, MAB-seq directly assesses TDG-mediated active DNA demethylation in a single experiment. On the other hand, the ability to distinguish between 5fC and 5caC might be helpful to address certain biological questions. To achieve this goal, we developed a novel method termed caMAB-seq, adding one step of $NaBH_4$ incubation to MAB-seq protocol. Because $NaBH_4$ reduces 5fC to 5hmC before bisulfite conversion, only 5caC is read as T in bisulfite sequencing (**Figure 1.2**). Therefore, a combination of MAB-seq with caMAB-seq (or other methods mapping 5fC or 5caC alone, such as redBS-seq) will allow separate quantification of 5fC and 5caC.

Comparative analysis of 5hmC and 5fC/5caC maps reveals the processivity of TET-mediated oxidation

Affinity enrichment-based methods suggest that in TDG-depleted mouse ESCs, 5hmC-enriched regions and 5fC/5caC-enriched regions highly overlap[31,43,46]. However, due to the low spatial resolution and lack of quantification of enrichment-based methods, it is unknown whether at individual CpG sites 5hmC and 5fC/5caC overlap. Using MAB-seq, we established the first genome-wide base-resolution 5fC/5caC map of TDG-depleted mouse ESCs and compared the data with the published base-resolution 5hmC map of mouse ESCs (established by TAB-seq)[49]. Our analysis shows that the majority of 5fC/5caC-modified CpGs are not modified by 5hmC. Similarly, the majority of 5hmC-modified CpGs are not modified by 5fC/5caC either. This observation suggests that TET-mediated oxidation sometimes stalls at 5hmC (low processivity) while in other cases goes further to 5fC/5caC (high processivity).

Difference in TET processivity may result in different biological functions. Unlike 5fC/5caC which are constantly being excised by TDG, 5hmC can be relatively stable if further oxidation by TET is restricted. Therefore, it is possible that 5hmC can serve as a modification that is recognized by reader proteins. Subsequently, modulation of TET processivity may directly determine the role of 5hmC as a stable epigenetic mark versus as an intermediate of active DNA demethylation. This issue is especially important for certain cell types such as neurons, of which 5hmC level is high while 5fC/5caC level is low.

Further analysis suggests that in mouse ESCs, TET processivity positively correlates with chromatin accessibility, an observation that also holds true for mouse 2-cell embryos. The causal relationship is unclear but can be potentially examined by modulating TET-mediated oxidation or chromatin accessibility in either global or targeted manner[113-115].

Combination of MAB-seq with other methods for profiling all five states of cytosine

Cytosine can exist in five different states, unmodified C, 5mC, 5hmC, 5fC and 5caC. For individual CpG sites, the population levels of these five states sum up to 100%. Therefore, with the development of MAB-seq and caMAB-seq, it is now possible to quantify all five states with a minimum of four sequencing experiments. For example, one can perform BS-seq, TAB-seq, MAB-seq and caMAB-seq and calculate the level of each state as below (**Table 6.1**):

Table 6.1 Quantification of all five states of cytosine

Modification state	Calculation
Unmodified C	100% - (C in BS-seq)% - (T in MAB-seq)%
5mC	(C in BS-seq)% - (C in TAB-seq)%
5hmC	(C in TAB-seq)%
5fC	(T in MAB-seq)% - (T in caMAB-seq)%
5caC	(T in caMAB-seq)%

Comprehensive analysis of the five states, or some of the states, can provide a better understanding of how DNMT-mediated establishment of 5mC, TET-mediated oxidation

of 5mC/5hmC/5fC and TDG/BER-mediated erasure of 5fC/5caC coordinate together. For example, an integrated analysis of BS-seq and MAB-seq in mouse ESCs has revealed that TDG depletion results in an unexpected decrease of 5mC/5hmC and an increase of unmodified C, contradicting the expectation that DNA demethylation will be impaired by removing a component of the pathway[46]. Similarly, the combination of BS-seq and MAB-seq also allows the analysis of unmodified C dynamics during paternal genome demethylation of the zygote (Chapter 3).

Development of liMAB-seq and scMAB-seq for analyzing ~100 cells and single cells, respectively

Active DNA demethylation happens in various biological contexts. Two well known examples of global active DNA demethylation are preimplantation development and primordial germ cell (PGC) development[10]. Before the establishment of MAB-seq, it is difficult to assess active DNA demethylation in these two scenarios by sequencing, due to the limited number of cells available. To solve this problem, we developed liMAB-seq capable of profiling ~100 cells.

A technical barrier for low-input and single-cell sequencing methods is the inevitable loss of material during steps involving DNA purification. A solution is to integrate multiple steps of library preparation into a single reaction tube, bypassing DNA purification. This has been shown to be possible for scBS-seq based on either RRBS or PBAT strategy. To establish liMAB-seq, we optimized *M.SssI* treatment condition, allowing it to be compatible with single-cell RRBS or PBAT protocol. The resulting protocol is comparable

to regular MAB-seq starting from bulk DNA (~1 μg).

On the basis of liMAB-seq, we further modified the protocol and established scMAB-seq. The number of CpG sites covered by scMAB-seq is 20-50% of that by liMAB-seq or regular MAB-seq. By adopting a binning approach, we can call 5fC/5caC-modified regions in single cells.

liMAB-seq reveals insights into paternal genome demethylation of the zygote

To demonstrate the utility of liMAB-seq, we analyzed paternal genome demethylation in preimplantation development. One question we focused on is whether 5mC can be actively processed to unmodified C in this process, just like in mouse ESCs. This question is important because it is currently unclear whether TDG-mediated restoration of unmodified C (AM-AR) is present in zygotes. To address this question, we treated zygotes with replication inhibitor Aphidicolin to inhibit the incorporation of unmodified C through replication. We then performed BS-seq and MAB-seq of sperm and zygotic paternal pronuclei, quantifying unmodified C as 100% - (C in BS-seq)% - (T in MAB-seq)%. Our analysis revealed that in addition to 5fC/5caC generation, unmodified C is also actively generated during paternal genome demethtylation. Given that *Tdg* mRNA is expressed at a negligible level[99] and that maternal TDG knockout did not lead to 5fC+5caC accumulation at selected genomic loci[51], there might be a TDG-independent and BER-coupled AM-AR mechanism for restoring unmodified C. If this is indeed the case, MAB-seq coupled with BS-seq can be a good readout for testing candidate mechanisms.

In the future, it will also be interesting to apply liMAB-seq to other biological systems, for example PGC development or specific subtypes of neurons.

scMAB-seq captures cell-to-cell heterogeneity among single cells of different cell types

To demonstrate the validity of scMAB-seq, we sequenced single mouse ESCs and single blastomeres from mouse 2-cell embryos and pooled these two types of cells for principal component analysis based on 5fC/5caC distribution. Our analysis shows that scMAB-seq captured the cell type-specific patterns of 5fC/5caC distribution, suggesting that it may be used for cell type classification. In recent years, various single-cell techniques, including epigenome sequencing methods[116] and RNA-seq methods[117], have been developed. scMAB-seq can complement these methods in cell type classification, given that 5fC/5caC is located at regulatory elements in a highly tissue-specific manner.

scMAB-seq captures the strand bias of 5fC/5caC resulted from DNA replication

Unlike 5mC, 5fC/5caC is not directly maintained during DNA replication. As a result, the newly synthesized strand will have a lower level of 5fC/5caC compared with the old template strand, creating a strand-biased distribution that can be detected by scMAB-seq. In cell types with impaired DNMT or TET machinery such as preimplantation embryos, the strand bias will be maintained. In cell types with functional DNMT and TET machinery such as mouse ESCs, 5fC/5caC can be gradually restored on the newly synthesized strand as time goes by, reducing the bias. Computational modeling of strand bias can help to quantify DNMT/TET activity relative to DNA replication rate[118].

scMAB-seq allows SCE mapping and lineage reconstruction based on SCE history

In scMAB-seq datasets of mouse ESCs and single mouse blastomeres, SCE can be readily identified as a switch of strand bias of 5fC/5caC. By using a binning approach, we can narrow down the location of SCE to a small genomic region, greatly improving the spatial resolution of SCE identification. The number of cells sequenced in the current study is relatively small, but we could already see a trend that certain chromosomes or genomic regions may have higher SCE frequency. In the future, it will be interesting to establish a comprehensive SCE map by increasing the sample size. With an SCE map established, we could try to understand what factors correlates with higher SCE frequency, for example replication timing and histone modification status.

On the basis of SCE mapping, we can also establish the lineage relationships among a group of cells based on the history of SCE. During cell division, two daughter cells from the same mother cell will share the same set of SCE. As a result, if two cells are closer in terms of lineage relationship, their SCE patterns will be more similar. As a proof of principle, we analyzed mouse 4-cell blastomeres and successfully established the lineage relationship. In the future, it will be interesting to try this strategy on more complex scenarios, for example 16-cell or 32-cell embryos. There are several advantages to use scMAB-seq to trace lineage. First, this strategy does not involve the introduction of exogenous factors that might complicate result interpretation. Second, in normal cells, SCE frequency is around 7-8 times per cell division (**Figure 5.5b** and REF 109), which is high enough to ensure a resolution to each cell division, while low enough to allow a

simple data analysis. Third, in one single experiment, both 5fC/5caC profile and cellular lineage are obtained.

Combination of scMAB-seq with functional readouts

scMAB-seq data provides multiple layers of information including 5fC/5caC map, strand bias, SCE profile and cellular lineage, and it will be interesting to see how they translate to functional outcomes. Therefore, a combination of scMAB-seq with other functional readouts, for example single-cell transcriptome or expression of marker proteins, will provide more insights. Previously, scBS-seq has been successfully combined with single-cell RNA-seq (scRNA-seq). In the future, a similar combination of scMAB-seq with scRNA-seq can provide a better understanding of how 5fC/5caC distribution and strand bias affect gene expression.

Appendix:

Supplementary figures and tables

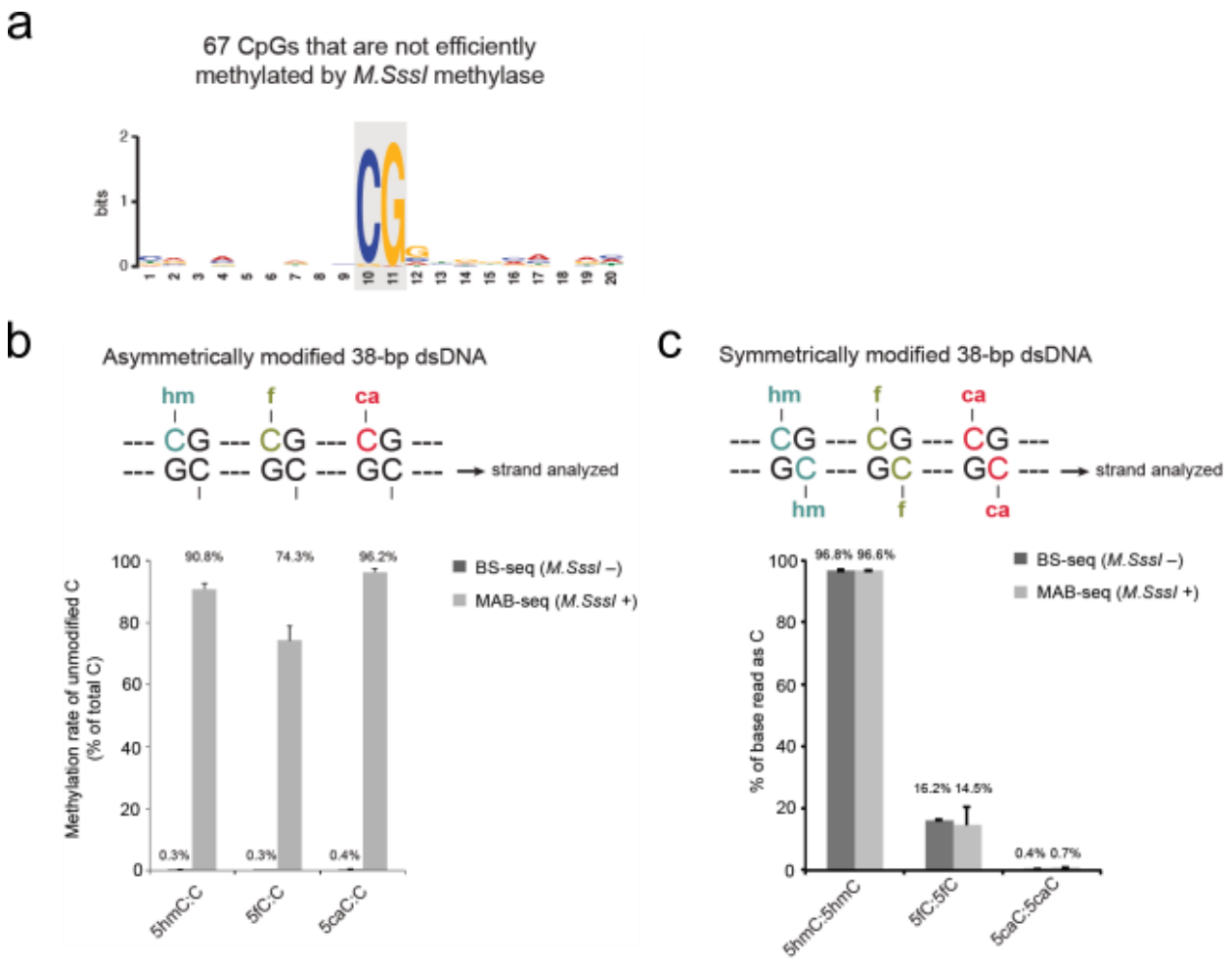

Figure S2.1 | Validation of MAB-seq. (a) Sequence context ±9 bp (Watson strand) around 67 CpGs that failed to be completely methylated by *M.SssI* (P<0.001). (b) Illumina deep sequencing of asymmetrically modified 38-bp dsDNA oligonucleotides demonstrated that *M.SssI* is capable of efficiently methylating unmodified C in hemi-modified CpGs. Error bars represents s.e.m of three experiments. (c) Illumina deep sequencing of symmetrically modified 38-bp double stranded DNA (dsDNA) oligonucleotides demonstrated that 5fC and 5caC, but not 5hmC, are efficiently converted to T.

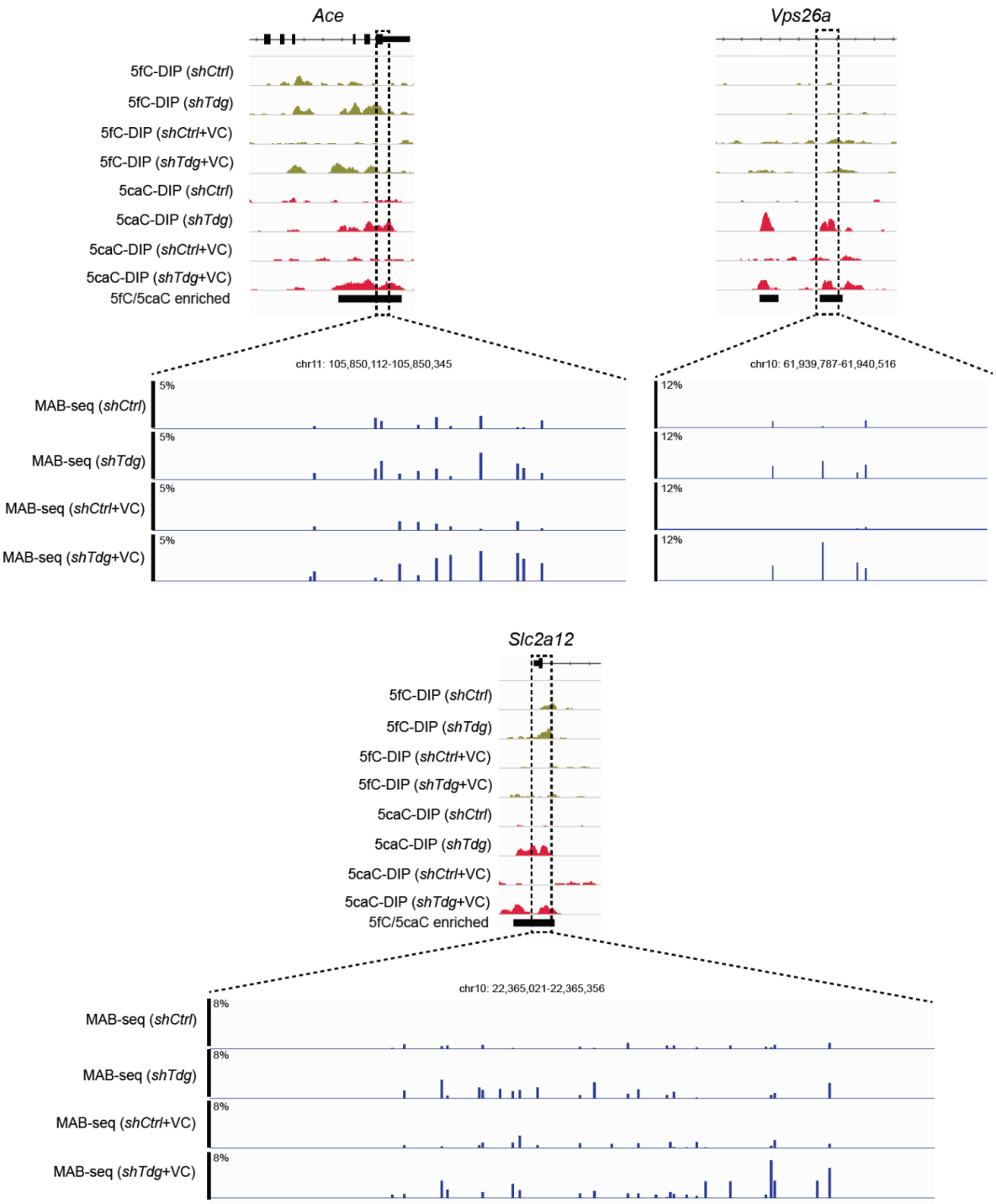

Figure S2.2 | Locus-specific MAB-seq analysis of representative genomic loci. Detection of 5fC/5caC at *Ace*, *Vps26a* and *Slc2a12* loci by Illumina deep sequencing-based locus-specific MAB-seq. The top nine tracks depict genome-wide DIP-seq analysis

Figure S2.2 (Continued)

of 5fC and 5caC. The other tracks display levels of 5fC/5caC measured by MAB-seq (corrected for background MAB-seq signals detected in Tet DKO).

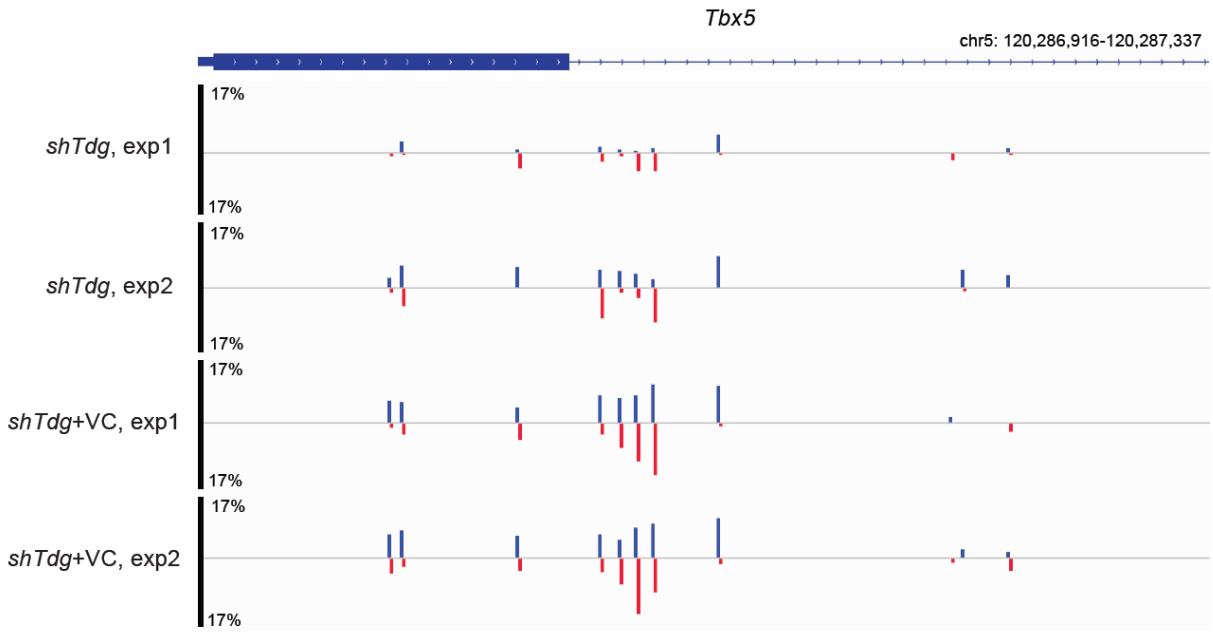

Figure S2.3 | Biological replicates showing the reproducibility of MAB-seq at a representative locus. Exp1 and exp2 represent two biologically independent replicates examined by locus-specific MAB-seq analysis. Blue and red bars indicate 5fC/5caC on top (Watson) and bottom (Crick) strands.

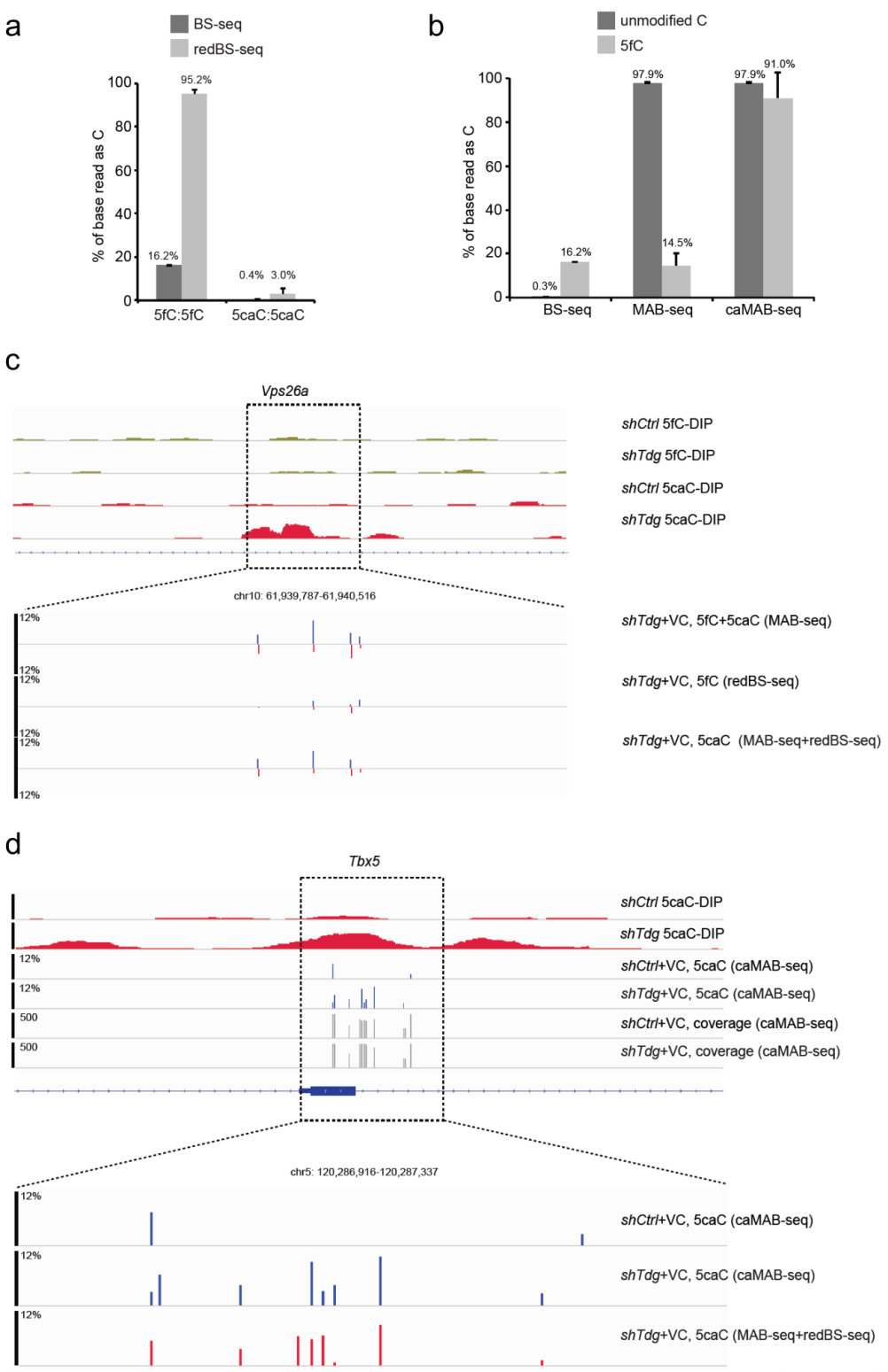

Figure S2.4 | Base-resolution mapping of 5caC by combining redBS-seq and MAB-seq. (a) Testing redBS-seq on synthetic 38-bp oligonucleotides with specific cytosine

Figure S2.4 (Continued)

modifications. Shown is the behavior of 38-bp oligonucleotides containing 5fC or 5caC during BS-seq and redBS-seq. 5fC is read as T in BS-seq but is read as C in redBS-seq, while 5caC is read as T in both BS-seq and redBS-seq. (b) Illumina deep sequencing of unmodified lambda DNA and synthesized 38bp 5fC oligo demonstrated that unmodified C and 5fC are read as C in caMAB-seq. For comparison, behavior of unmodified C and 5fC in BS-seq and MAB-seq are also provided. (c) Locus-specific analysis of 5fC and 5caC at the *Vps26a* locus indicates that 5fC and 5caC are largely non-overlapping and both modifications exhibit strong strand asymmetry. For comparison, affinity-based 5fC and 5caC maps are shown on the top. In enlarged views, base-resolution maps of 5fC+5caC (measured by MAB-seq), 5fC (measured by redBS-seq) and 5caC (subtraction between MAB-seq and redBS-seq) are shown. Signals for the Watson strand are in blue, whereas those for the Crick strand are in red. (d) Comparative locus-specific analysis of 5caC at the *Tbx5* locus by two base-resolution methods (indirect 5caC mapping through subtraction between MAB-seq and redBS-seq (red); direct 5caC mapping by caMAB-seq (blue)). For comparison, also shown are DIP-seq based maps of 5caC in control and TDG-depleted mouse ESCs. The level of 5caC (only Watson strand shown) is displayed as the percentage of total C modified as 5caC. Corresponding caMAB-seq sequencing depth (vertical axis limits are 0 to 500) at each CpG was also shown.

Supplemental Table S2.1 | Bisulfite sequencing primers for locus-specific analysis. F and R represent a primer pair. Complementary represents primers for bottom (Crick) strand.

Primer	Sequence
Ace – F	TTGGGTTTGTATTTGGAGTTATAGTAG
Ace – R	ACCAACAAAATCACCTCAAAAATAT
Slc2a12 – F	TAATTTGTTGAATTAGAAGGGGAGA
Slc2a12 – R	ATACATAAAATCCCAAAAAAACTTC
Tbx5 – F	TTTGGAGTTTGATTTTAAAGATAGGTTTTG
Tbx5 – R	AATTAAAAACTCTCCAATAATATAAATAAATTTCT
Vps26a – F	AGGGTTATTTATGGTGATTGAATTA
Vps26a – R	AAACCTTCAAAAAAACAAACAATAC
Tbx5 – complementary – F	ATAAGGTATATTTAGGGTAGTTGGGAGAGAATTAT
Tbx5 – complementary – R	CTAAAACCTAATTCCAAAAACAAATCTTAC
Vps26a – complementary - F	TGTTGGGAATTGAATTTAGGATTTTTAGAA
Vps26a – complementary - R	AAAAAAATATACATACCACAAATAAACCAC

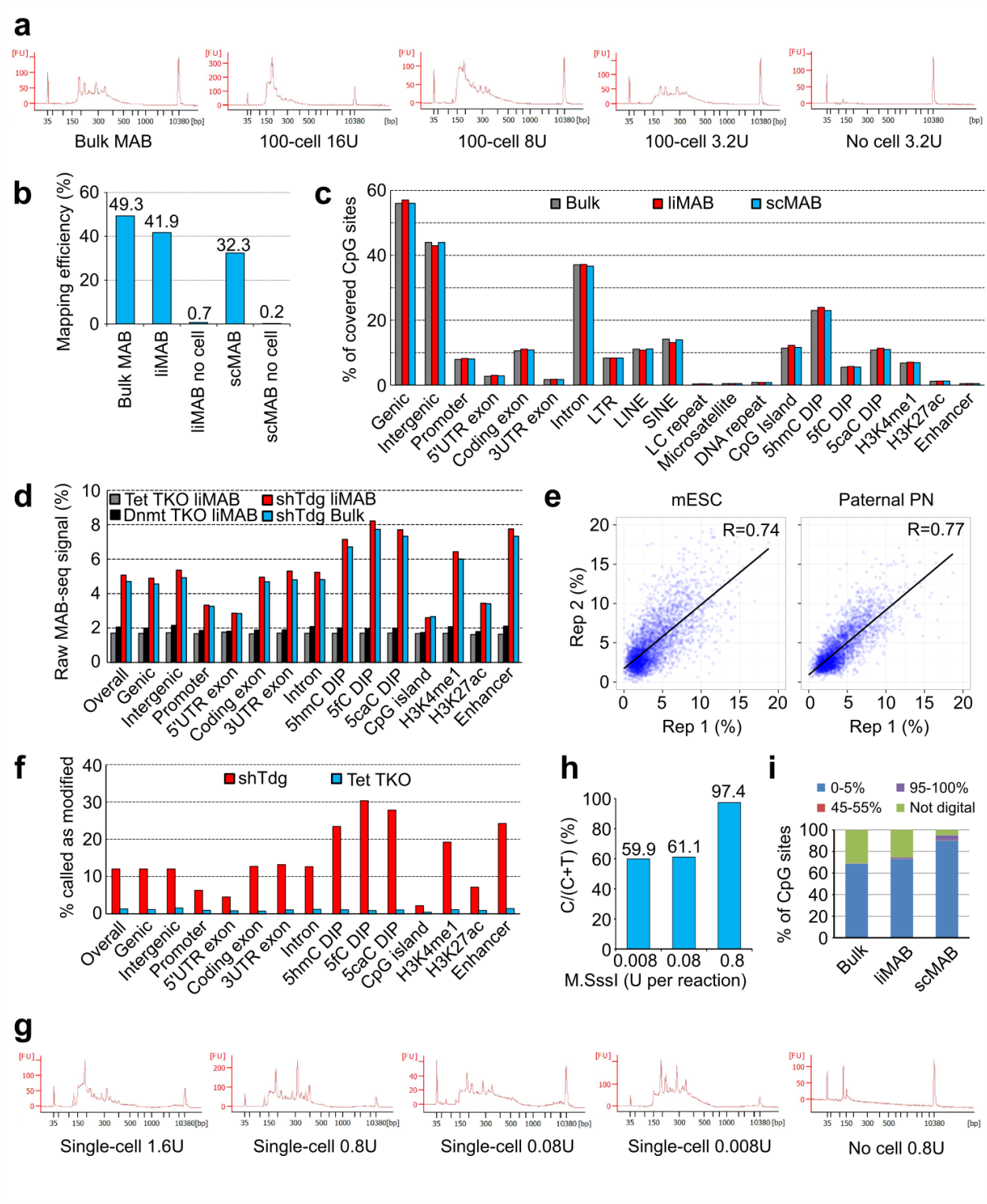

Figure S3.1 | Optimization of RRBS-based liMAB-seq and scMAB-seq. (a) Bioanalyzer traces of RRBS-based regular MAB-seq and liMAB-seq libraries prepared

Figure S3.1 (Continued)

using different concentrations of *M.SssI*. Bulk is a MAB-seq library prepared from 1μg DNA using a previous protocol. 16 U is the *M.SssI* concentration used in regular MAB-seq but results in a poor-quality library when starting from 100 cells, as indicated by a significant loss of large DNA fragments. 3.2 U is the final concentration chosen for liMAB-seq as the bioanalyzer trace shape is comparable with that from a regular MAB-seq (bulk). A negative control starting from no cell shows no signals except for adaptor dimers. (b) Mean mapping efficiency of RRBS-based regular (bulk), low-input and single-cell MAB-seq. While liMAB-seq and scMAB-seq have a mean mapping efficiency of 42.2% and 33.4%, respectively, the corresponding negative controls starting with no cell have a mapping efficiency closed to 0. (c) Distribution of the CpG sites covered by RRBS-based regular (bulk), low-input and single-cell MAB-seq. Proportion of the covered 5× CpG sites in different genomic features is shown. Samples with 20-40 million reads were compared. (d) Raw MAB-seq signals detected by liMAB-seq are comparable to those detected by regular MAB-seq (bulk) in different genomic features. The levels shown are raw signals without subtracting the background signals detected in Tet TKO negative control. The background signal, defined by that in TKO cells, is also shown. (e) Correlation plots comparing two biological replicates of liMAB-seq. liMAB-seq of mouse ESCs (mESC) and zygotic paternal pronuclei (Paternal PN) are shown. Each dot represents a 5 kb genomic bin with at least 15 20×CGs, and 5fC/5caC level within the bin is calculated as sum(T)/(sum(C)+sum(T)). (f) Proportion of 100bp bins called as 5fC/5caC-modified for individual genomic features. CpG sites were merged into 100bp genomic bins for calling 5fC/5caC-modified regions (modified is defined as T/(C+T)>=10%; bins with at least two

Figure S3.1 (Continued)

5×CGs were analyzed). For each genomic feature, the proportion of the overlapping 100bp bins called as modified is shown. liMAB-seq results of Tet TKO and TDG-depleted ESCs were compared. At global scale (overall), 1.2% of the regions covered in Tet TKO negative control were called as 5fC/5caC-modified (false positive), while 11.9% in TDG-depleted ESCs were called as modified (FDR=10.4%). (g) Bioanalyzer traces of scMAB-seq libraries prepared using different concentrations of *M.SssI*. While 1.6U enzyme per reaction leads to a poor quality library, 0.8U and below are suitable for library preparation. (h) Methylation of single Tet TKO cells by different concentrations of *M.SssI*. While 0.008U and 0.08U are not sufficient for methylating unmodified C in Tet TKO cells, 0.8U (the concentration used for scMAB-seq) allows efficient methylation. (i) Proportion of CpG sites displaying a digital pattern in regular (bulk), low-input and single-cell MAB-seq. A CpG site is defined as being digital when 5fC/5caC level is either 0-5% (blue), 45%-55% (red) or 95-100% (purple). For scMAB-seq datasets, over 95% of CpG sites display a digital pattern as expected for single-cell data.

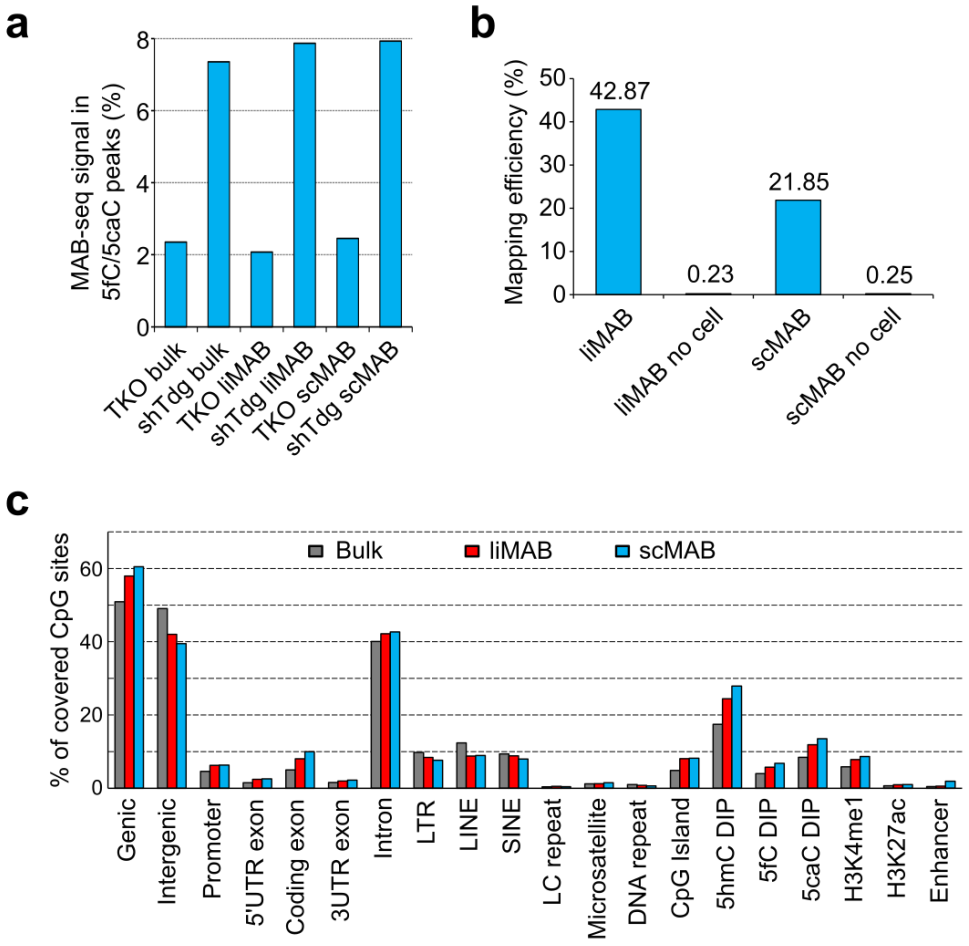

Figure S3.2 | Optimization of RRBS-based liMAB-seq and scMAB-seq. (a) Raw MAB-seq signals detected in 5fC/5caC DIP-seq peaks by regular (bulk) MAB-seq, PBAT-based liMAB-seq and PBAT-based scMAB-seq. Signals are low in TKO negative controls (2-2.5%) but high in shTdg mouse ESCs (7-8%), supporting the validity of the methods in detecting real 5fC/5caC signals. (b) Mean mapping efficiency of PBAT-based liMAB-seq and scMAB-seq. The negative controls have a mapping efficiency closed to 0, demonstrating a minimal extent of contamination. (c) Distribution of the CpG sites covered by the published whole-genome MAB-seq (bulk) and PBAT-based liMAB-seq and scMAB-seq. Proportion of the covered 5× CpG sites in different genomic features is shown.

Table S3.1 | Mapping statistics of RRBS-based regular, low-input and single-cell MAB-seq. For regular MAB-seq, 1 μg genomic DNA was used as the starting material. For low-input MAB-seq (liMAB-seq), around 100 diploid cells were used (in the case of the paternal pronuclei which are haploid, around 150 paternal pronuclei were used). For single-cell MAB-seq (scMAB-seq), a single cell was used. 1×CG and 5×CG refer to CpG sites covered for at least 1 time and 5 times by sequencing, respectively. For each method, the samples were sorted by the number of 5×CG.

Method	Experiment	PF reads after trimming	Uniquely mapped reads	Mapping efficiency	Number of 1×CG	Number of 5×CG
RRBS-based regular MAB-seq (1 μg DNA)	Sperm rep1	28,738,090	16,606,280	0.578	2,598,268	966,519
	Sperm rep2	26,917,843	13,455,072	0.500	1,380,938	966,095
	shTdg mESC	27,165,723	12,502,602	0.460	1,125,162	758,336
RRBS-based liMAB-seq (~100 diploid cells)	2-cell embryos	31,952,632	13,907,478	0.4353	1,659,538	995,323
	shTdg mESC rep2	29,205,704	11,078,150	0.3793	2,031,985	934,280
	Paternal pronuclei rep1	30,873,848	12,527,439	0.4058	1,417,107	891,785
	Tet TKO mESC rep2	27,913,538	13,345,063	0.4781	1,466,108	878,947
	Paternal pronuclei rep2	33,039,317	10,703,204	0.3240	1,008,455	765,432
	Tet TKO mESC rep1	19,201,923	9,053,976	0.4715	1,934,988	764,281
	shTdg mESC rep1	21,636,674	10,265,147	0.4744	1,889,543	756,502
	Dnmt TKO mESC rep2	22,497,650	9,832,898	0.4371	1,251,116	713,976

Table S3.1 (Continued)

RRBS-based liMAB-seq (~100 diploid cells)	Dnmt TKO mESC rep1	27,535,402	10,020,554	0.3639	1,012,523	596,356
RRBS-based scMAB-seq	Tet TKO mESC	19,334,066	7,423,872	0.384	480,210	421,649
	shTdg mESC	29,328,097	13,013,857	0.444	439,729	384,426
	shTdg mESC	14,544,392	6,177,465	0.425	360,136	284,590
	shTdg mESC	16,574,750	6,577,015	0.397	437,599	272,443
	Tet TKO mESC	27,866,527	9,911,731	0.356	285,858	199,704
	shTdg mESC	14,100,383	5,341,181	0.379	217,691	193,710
	shTdg mESC	13,310,417	4,981,658	0.374	206,623	179,259
	shTdg mESC	15,074,769	6,290,954	0.417	237,457	178,104
	shTdg mESC	13,601,420	4,004,439	0.294	245,814	176,566
	shTdg mESC	14,497,019	4,943,628	0.341	238,184	165,249
	2-cell blastomere	23,668,009	8,553,326	0.361	211,703	157,679
	4-cell blastomere	8,482,794	2,171,010	0.256	210,281	156,341
	shTdg mESC	30,492,794	11,475,900	0.376	222,725	153,755
	shTdg mESC	14,293,291	3,732,004	0.261	163,951	144,859
	2-cell blastomere	15,514,778	5,713,036	0.368	175,121	140,792
	2-cell blastomere	14,835,647	4,744,635	0.320	170,154	140,466
	4-cell blastomere	7,889,004	2,495,648	0.316	188,420	139,933
	2-cell blastomere	14,075,671	5,145,670	0.366	157,082	139,850

Table S3.1 (Continued)

RRBS-based scMAB-seq	4-cell blastomere	7,770,324	2,183,668	0.281	175,480	136,719
	2-cell blastomere	12,952,144	3,143,347	0.243	151,410	128,455
	2-cell blastomere	15,725,860	6,220,382	0.396	152,497	125,772
	2-cell blastomere	21,307,958	3,545,059	0.166	153,881	125,095
	Tet TKO mESC	14,763,476	5,788,015	0.392	184,302	119,621
	2-cell blastomere	11,368,138	3,037,307	0.267	131,870	111,172
	Tet TKO mESC	12,776,248	3,470,479	0.272	151,070	111,063
	2-cell blastomere	20,542,784	8,059,263	0.392	139,317	110,726
	2-cell blastomere	20,678,581	5,994,340	0.290	139,794	108,510
	2-cell blastomere	15,847,026	4,698,879	0.297	156,016	108,436
	2-cell blastomere	16,081,885	4,761,718	0.296	162,213	108,271
	2-cell blastomere	10,047,141	3,174,735	0.316	149,664	105,990
	2-cell blastomere	10,953,254	3,251,724	0.297	118,926	104,215
	2-cell blastomere	11,058,523	3,893,824	0.352	150,970	102,633
	2-cell blastomere	11,045,878	3,861,295	0.350	148,250	102,491
	2-cell blastomere	19,497,636	6,627,878	0.340	136,881	101,833
	4-cell blastomere	6,272,960	1,398,458	0.223	124,601	101,821
	2-cell blastomere	13,949,330	4,177,662	0.299	161,713	100,019
	2-cell blastomere	19,317,452	5,567,865	0.288	136,795	98,721
	2-cell blastomere	10,301,424	2,390,977	0.232	148,443	95,988
	2-cell blastomere	12,827,855	4,486,767	0.350	129,737	93,895
	2-cell blastomere	12,615,030	3,949,754	0.313	142,998	92,229

Table S3.1 (Continued)

RRBS-based scMAB-seq	2-cell blastomere	11,862,856	4,587,690	0.387	131,842	91,148
	2-cell blastomere	12,635,338	4,187,961	0.331	141,729	90,497
	2-cell blastomere	15,418,582	5,679,811	0.368	139,748	89,739
	2-cell blastomere	9,670,289	3,349,756	0.346	131,093	86,595
	4-cell blastomere	9,176,965	456,875	0.050	151,555	85,174
	4-cell blastomere	8,454,203	2,692,643	0.318	118,029	82,300
	4-cell blastomere	7,466,731	2,195,684	0.294	106,305	81,643
	4-cell blastomere	7,808,799	2,579,086	0.330	98,145	71,902
	2-cell blastomere	10,528,203	3,259,547	0.310	93,634	56,231

Table S3.2 | Increased CpG coverage using longer sequencing length. Four RRBS-based scMAB-seq libraries were sequenced in single-end 250 bp (SE 250) mode. The SE 250 data were compared with the same data trimmed to 100 bp (SE 100). The analysis suggested that when necessary, longer sequencing length can help recover more CpG sites.

scMAB-seq sample	Number of reads	Number of 1×CG		Number of 5×CG		Number of 10×CG	
		SE 100	SE 250	SE 100	SE 250	SE 100	SE 250
1	14183724	411425	590004	386658	502038	325798	438787
2	19982722	237457	349383	178104	249946	170509	236627
3	18112141	151410	215121	128455	178726	121910	167205
4	19389934	131870	191898	111172	157666	104029	145112

Table S3.3 | Increased CpG coverage using pair-end sequencing. Four RRBS-based scMAB-seq libraries were sequenced in pair-end 100 bp (PE 100) mode. The PE 100 data were compared with the same data using only read 1 (SE 100). The analysis suggested that when necessary, performing pair-end sequencing can help recover more CpG sites.

scMAB-seq sample	Number of reads	Number of 1×CG		Number of 5×CG		Number of 10×CG	
		SE 100	PE 100	SE 100	PE 100	SE 100	PE 100
1	17566012	217691	452392	193710	326490	186341	296856
2	18911106	163951	354656	144859	243562	138520	219834
3	9131030	149555	315158	135070	222066	126754	198910
4	8305919	113003	244364	99968	198794	92499	146575

Table S3.4 | Mapping statistics of PBAT-based low-input and single-cell MAB-seq. For low-input MAB-seq (liMAB-seq), around 100 diploid cells were used. For single-cell MAB-seq (scMAB-seq), a single cell was used. 1×CG refers to CpG sites covered for at least 1 time by sequencing. For each method, the samples were sorted by the number of 1×CG.

Method	Experiment	PF reads after trimming	Uniquely mapped reads	Mapping efficiency	PCR duplication rate	Number of 1×CG
PBAT-based liMAB-seq (~100 diploid cells)	shTdg mESC	251,815,145	106,783,752	0.424	0.127	25,253,604
	TKO mESC	27,429,466	11,884,783	0.433	0.034	8,196,246
PBAT-based scMAB-seq	Tet TKO mESC	30,033,906	10,624,469	0.354	0.135	6,086,412
	shTdg mESC	30,274,887	7,211,236	0.238	0.111	4,339,008
	Tet TKO mESC	19,720,899	5,344,165	0.271	0.077	3,636,782
	shTdg mESC	28,765,221	5,018,444	0.174	0.084	3,183,118
	Tet TKO mESC	18,329,827	4,259,691	0.232	0.067	2,987,072
	shTdg mESC	27,449,771	3,943,879	0.144	0.122	2,335,982
	4-cell blastomere	15,586,920	3,812,399	0.245	0.177	2,267,937
	4-cell blastomere	12,816,582	2,909,310	0.227	0.151	1,834,911
	4-cell blastomere	14,572,877	3,184,182	0.219	0.191	1,834,503
	shTdg mESC	25,785,216	2,329,210	0.090	0.077	1,674,503
	4-cell blastomere	9,249,967	1,940,987	0.210	0.122	1,374,069

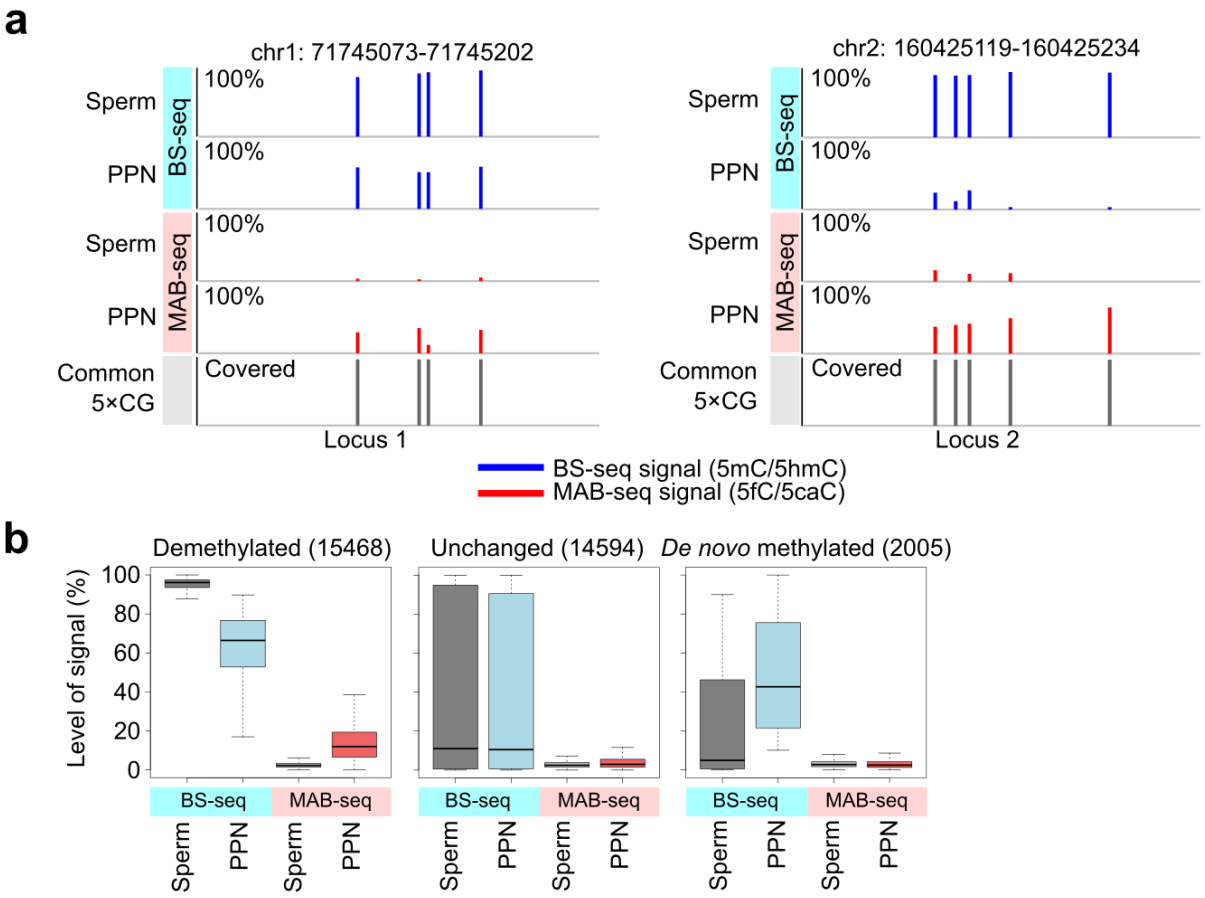

Figure S4.1 | 5mC/5hmC decrease is coupled with 5fC/5caC increase. (a) Two representative loci showing the changes of 5mC/5hmC (measured by BS-seq) and 5fC/5caC (measured by MAB-seq) from sperm to paternal pronuclei (PPN). Common 5×CGs are shown. (b) Change of 5mC/5hmC and 5fC/5caC in demethylated, unchanged and de novo methylated regions (250bp bins). Using the BS-seq data of sperm and paternal pronuclei, 250bp bins with at least 4 5×CGs are classified into the three groups based on Fisher's exact test. 5fC/5caC increase is evident for demethylated regions but not evident for de novo methylated regions.

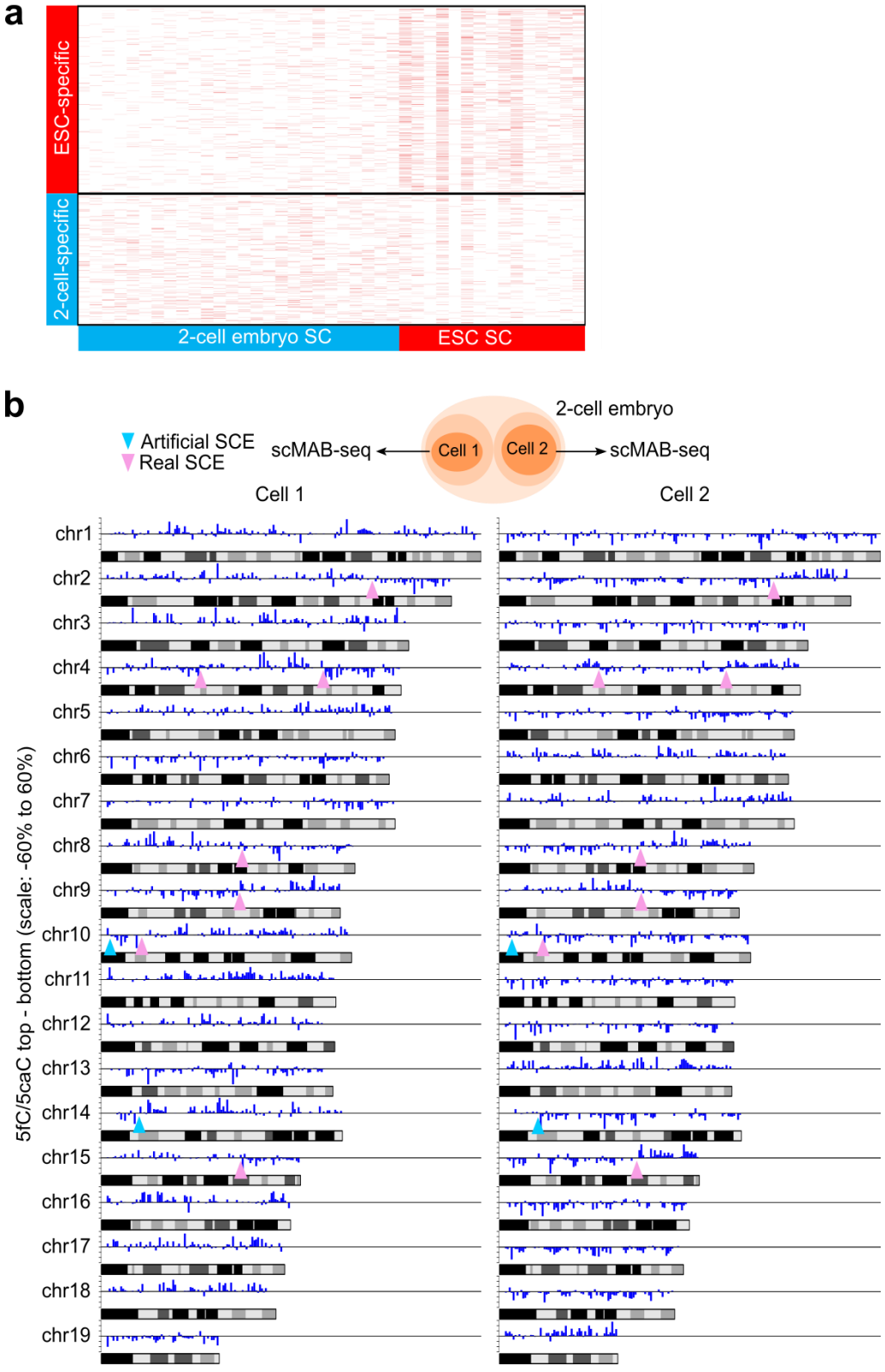

Figure S5.1 | scMAB-seq captures cell-to-cell heterogeneity of 5fC/5caC. (a) 5fC/5caC signals detected in single cells are enriched at 2-cell specific (blue) and ESC-

Figure S5.1 (Continued)

specific 5fC/5caC regions (red) identified by liMAB-seq. Each red bar indicates a 2 kb bin called as modified in a single cell. The cells are displayed in an order determined by the degree of enrichment (**Figure 5.2c**). (b) Karyogram showing 5fC/5caC distribution in two blastomeres of one 2-cell embryo. Each bar is a 1MB bin, and the difference of 5fC/5caC between the top and bottom strands are calculated. The two cells display complementary 5fC/5caC pattern. Artificial SCE (blue arrowheads) and real SCE (red arrowheads) were observed.

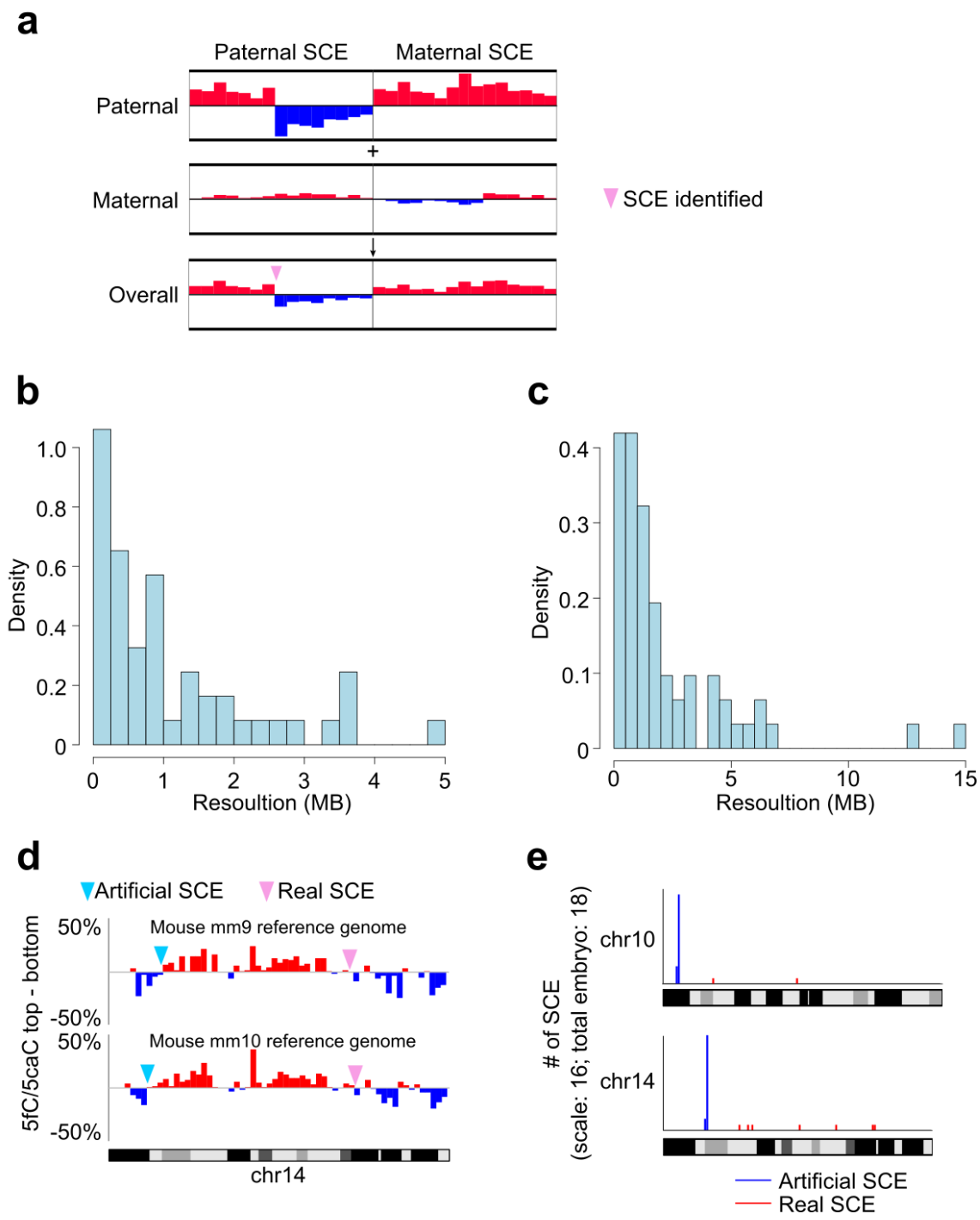

Figure S5.2 | Mapping SCE by scMAB-seq. (a) Diagram showing the expected 5fC/5caC pattern in 2-cell blastomeres. Because the majority of 5fC/5caC is contributed by the paternal genome, the overall 5fC/5caC profile in the diploid blastomere will largely

Figure S5.2 (Continued)

resemble the paternal 5fC/5caC profile. A paternal SCE will cause a switch of 5fC/5caC from biased towards one strand to the other strand. A maternal SCE is not sufficient to cause a switch of strand bias and won't be identified in this case. (b) Histogram showing the resolution of SCE mapping using a pair of blastomeres from one 2-cell embryo. The median resolution is 700 kb. (c) Histogram showing the resolution of SCE mapping using one blastomere from one 2-cell embryo. The median resolution is 1250 kb. (d) Artificial SCE (blue arrowheads) identified at mm9 chr14: 20,000,000~21,000,000. This artificial SCE comes from an error of mouse mm9 genome assembly, and the error hasn't been fixed in mm10 genome assembly. (e) The two artificial SCEs (blue) were observed in all of the 18 embryos analyzed. The red bar denotes 1MB bin with artificial SCEs. The blue bar denotes 1MB bin with real SCEs. The height of the bar indicates the number of SCE observed in that bin (scale: 16). The two artificial SCEs were observed in abnormally high frequency (every single embryo) compared to real SCEs.

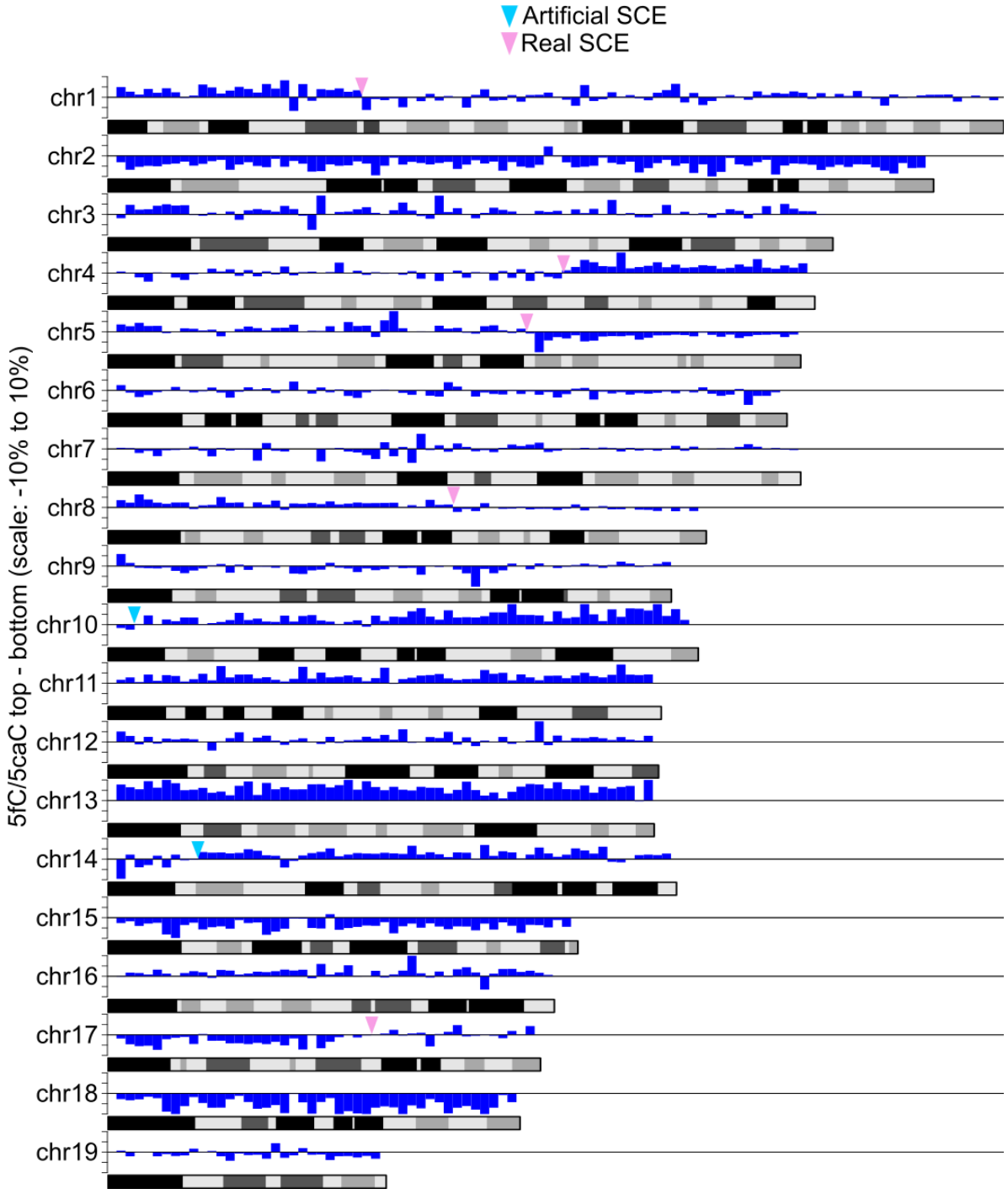

Figure S5.3 | Identification of strand bias and SCE in mouse ESCs. Karyogram showing the strand distribution of 5fC/5caC in a single mouse ES cell depleted of TDG and analyzed by PBAT-based scMAB-seq. The genome was segmented into 2MB bins, and the difference of 5fC/5caC between the top and bottom strands was calculated for

Figure S5.3 (Continued)

each bin. Real SCEs (red arrowheads) can be identified by a switch of 5fC/5caC distribution from biased toward one strand to largely unbiased or less biased. Artificial SCE (blue arrowheads) can be identified by a switch of 5mC/5hmC distribution from biased toward one strand to completely biased toward another strand.

References

1. Smith, Z.D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**, 204-20 (2013).

2. Li, E. & Zhang, Y. DNA methylation in mammals. *Cold Spring Harb Perspect Biol* **6**, a019133 (2014).

3. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**, 6-21 (2002).

4. Okano, M., Bell, D.W., Haber, D.A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-57 (1999).

5. Hermann, A., Goyal, R. & Jeltsch, A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *J Biol Chem* **279**, 48350-9 (2004).

6. Bostick, M. *et al.* UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* **317**, 1760-4 (2007).

7. Sharif, J. *et al.* The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450**, 908-12 (2007).

8. Holliday, R. & Pugh, J.E. DNA modification mechanisms and gene activity during development. *Science* **187**, 226-32 (1975).

9. Riggs, A.D. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet* **14**, 9-25 (1975).

10. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat Rev Genet* **18**, 517-534 (2017).

11. Oswald, J. *et al.* Active demethylation of the paternal genome in the mouse zygote. *Curr Biol* **10**, 475-8 (2000).

12. Hajkova, P. *et al.* Epigenetic reprogramming in mouse primordial germ cells. *Mech Dev* **117**, 15-23 (2002).

13. Mayer, W., Niveleau, A., Walter, J., Fundele, R. & Haaf, T. Demethylation of the zygotic paternal genome. *Nature* **403**, 501-2 (2000).

14. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929-30 (2009).

15. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-5 (2009).

16. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129-33 (2010).

17. He, Y.F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-7 (2011).

18. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300-3 (2011).

19. Kohli, R.M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472-9 (2013).

20. Lu, X., Zhao, B.S. & He, C. TET family proteins: oxidation activity, interacting molecules, and functions in diseases. *Chem Rev* **115**, 2225-39 (2015).

21. Maiti, A. & Drohat, A.C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem* **286**, 35334-8 (2011).

22. Weber, A.R. *et al.* Biochemical reconstitution of TET1-TDG-BER-dependent active DNA demethylation reveals a highly coordinated mechanism. *Nat Commun* **7**, 10806 (2016).

23. Hashimoto, H. *et al.* Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res* **40**, 4841-9 (2012).

24. Otani, J. *et al.* Cell cycle-dependent turnover of 5-hydroxymethyl cytosine in mouse embryonic stem cells. *PLoS One* **8**, e82961 (2013).

25. Ji, D., Lin, K., Song, J. & Wang, Y. Effects of Tet-induced oxidation products of 5-methylcytosine on Dnmt1- and DNMT3a-mediated cytosine methylation. *Mol Biosyst* **10**, 1749-52 (2014).

26. Pfaffeneder, T. *et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew Chem Int Ed Engl* **50**, 7008-12 (2011).

27. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat Chem Biol* **11**, 555-7 (2015).

28. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem* **6**, 1049-55 (2014).

29. Globisch, D. *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* **5**, e15367 (2010).

30. Munzel, M. *et al.* Quantification of the sixth DNA base hydroxymethylcytosine in the brain. *Angew Chem Int Ed Engl* **49**, 5375-7 (2010).

31. Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692-706 (2013).

32. Wossidlo, M. *et al.* 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Commun* **2**, 241 (2011).

33. Iqbal, K., Jin, S.G., Pfeifer, G.P. & Szabo, P.E. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proc Natl Acad Sci U S A* **108**, 3642-7 (2011).

34. Inoue, A., Shen, L., Dai, Q., He, C. & Zhang, Y. Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. *Cell Res* **21**, 1670-6 (2011).

35. Yamaguchi, S. *et al.* Dynamics of 5-methylcytosine and 5-hydroxymethylcytosine during germ cell reprogramming. *Cell Res* **23**, 329-39 (2013).

36. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402 (2011).

37. Wu, H. *et al.* Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev* **25**, 679-84 (2011).

38. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-8 (2011).

39. 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).

40. Jin, S.G., Wu, X., Li, A.X. & Pfeifer, G.P. Genomic mapping of 5-hydroxymethylcytosine in the human brain. *Nucleic Acids Res* **39**, 5015-24 (2011).

41. Pastor, W.A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394-7 (2011).

42. Song, C.X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* **29**, 68-72 (2011).

43. Song, C.X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678-91 (2013).

44. Raiber, E.A. *et al.* Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol* **13**, R69 (2012).

45. Iurlaro, M. *et al.* In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome Biol* **17**, 141 (2016).

46. Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat Biotechnol* **32**, 1231-40 (2014).

47. Huang, Y. *et al.* The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One* **5**, e8888 (2010).

48. Booth, M.J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934-7 (2012).

49. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368-80 (2012).

50. Hu, X. *et al.* Tet and TDG mediate DNA demethylation essential for mesenchymal-to-epithelial transition in somatic cell reprogramming. *Cell Stem Cell* **14**, 512-22 (2014).

51. Guo, F. *et al.* Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell* **15**, 447-58 (2014).

52. Neri, F. *et al.* Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics. *Cell Rep* (2015).

53. Booth, M.J., Marsico, G., Bachman, M., Beraldi, D. & Balasubramanian, S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat Chem* **6**, 435-40 (2014).

54. Lu, X. *et al.* Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Res* **25**, 386-9 (2015).

55. Schomacher, L. *et al.* Neil DNA glycosylases promote substrate turnover by Tdg during DNA demethylation. *Nat Struct Mol Biol* **23**, 116-24 (2016).

56. Lu, X. *et al.* Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. *J Am Chem Soc* **135**, 9315-7 (2013).

57. Wheldon, L.M. *et al.* Transient accumulation of 5-carboxylcytosine indicates involvement of active demethylation in lineage specification of neural stem cells. *Cell Rep* **7**, 1353-61 (2014).

58. 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).

59. Cortazar, D. *et al.* Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature* **470**, 419-23 (2011).

60. Cortellino, S. *et al.* Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell* **146**, 67-79 (2011).

61. Szulwach, K.E. *et al.* 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat Neurosci* **14**, 1607-16 (2011).

62. Renbaum, P. *et al.* Cloning, characterization, and expression in Escherichia coli of the gene coding for the CpG DNA methylase from Spiroplasma sp. strain MQ1(M.SssI). *Nucleic Acids Res* **18**, 1145-52 (1990).

63. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).

64. Hu, L. *et al.* Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell* **155**, 1545-55 (2013).

65. Hashimoto, H. *et al.* Structure of a Naegleria Tet-like dioxygenase in complex with 5-methylcytosine DNA. *Nature* **506**, 391-5 (2014).

66. Lu, F., Liu, Y., Jiang, L., Yamaguchi, S. & Zhang, Y. Role of Tet proteins in enhancer activity and telomere elongation. *Genes Dev* **28**, 2103-19 (2014).

67. Yin, R. *et al.* Ascorbic acid enhances Tet-mediated 5-methylcytosine oxidation and promotes DNA demethylation in mammals. *J Am Chem Soc* **135**, 10396-403 (2013).

68. Blaschke, K. *et al.* Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature* **500**, 222-6 (2013).

69. Minor, E.A., Court, B.L., Young, J.I. & Wang, G. Ascorbate induces ten-eleven translocation (Tet) methylcytosine dioxygenase-mediated generation of 5-hydroxymethylcytosine. *J Biol Chem* **288**, 13669-74 (2013).

70. Consortium, E.P. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**, e1001046 (2011).

71. Hu, G. *et al.* H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* **12**, 180-92 (2013).

72. Banaszynski, L.A. *et al.* Hira-dependent histone H3.3 deposition facilitates PRC2 recruitment at developmental loci in ES cells. *Cell* **155**, 107-20 (2013).

73. Wu, H. *et al.* Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473**, 389-93 (2011).

74. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-19 (2013).

75. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-20 (2014).

76. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-2 (2011).

77. Liu, Y., Siegmund, K.D., Laird, P.W. & Berman, B.P. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol* **13**, R61 (2012).

78. Wu, H. *et al.* Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473**, 389-93 (2011).

79. Mikkelsen, T.S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-60 (2007).

80. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-70 (2008).

81. Ku, M. *et al.* Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* **4**, e1000242 (2008).

82. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521-33 (2008).

83. Stadler, M.B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490-5 (2011).

84. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-6 (2010).

85. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116-20 (2012).

86. Sun, Z. *et al.* A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. *Mol Cell* **57**, 750-61 (2015).

87. Xia, B. *et al.* Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat Methods* **12**, 1047-50 (2015).

88. Wu, H., Wu, X. & Zhang, Y. Base-resolution profiling of active DNA demethylation using MAB-seq and caMAB-seq. *Nat Protoc* **11**, 1081-100 (2016).

89. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**, 5868-77 (2005).

90. Miura, F., Enomoto, Y., Dairiki, R. & Ito, T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res* **40**, e136 (2012).

91. 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**, 2126-35 (2013).

92. Smallwood, S.A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* **11**, 817-20 (2014).

93. Farlik, M. *et al.* Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep* **10**, 1386-97 (2015).

94. Incarnato, D., Krepelova, A. & Neri, F. High-throughput single nucleotide variant discovery in E14 mouse embryonic stem cells provides a new reference genome assembly. *Genomics* **104**, 121-7 (2014).

95. Keane, T.M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289-94 (2011).

96. Saitou, M., Kagiwada, S. & Kurimoto, K. Epigenetic reprogramming in mouse pre-implantation development and primordial germ cells. *Development* **139**, 15-31 (2012).

97. Lee, H.J., Hore, T.A. & Reik, W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14**, 710-9 (2014).

98. Wu, H. & Zhang, Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell* **156**, 45-68 (2014).

99. Tang, F. *et al.* Deterministic and stochastic allele specific gene expression in single mouse blastomeres. *PLoS One* **6**, e21208 (2011).

100. Wossidlo, M. *et al.* Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. *EMBO J* **29**, 1877-88 (2010).

101. Ladstatter, S. & Tachibana-Konwalski, K. A Surveillance Mechanism Ensures Repair of DNA Lesions during Zygotic Reprogramming. *Cell* (2016).

102. Hajkova, P. *et al.* Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. *Science* **329**, 78-82 (2010).

103. Wang, L. *et al.* Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**, 979-91 (2014).

104. Lu, F. *et al.* Establishing Chromatin Regulatory Landscape during Mouse Preimplantation Development. *Cell* **165**, 1375-88 (2016).

105. Inoue, A. & Zhang, Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* **334**, 194 (2011).

106. Shen, L. *et al.* Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. *Cell Stem Cell* **15**, 459-70 (2014).

107. Wilson, D.M., 3rd & Thompson, L.H. Molecular mechanisms of sister-chromatid exchange. *Mutat Res* **616**, 11-23 (2007).

108. Woodworth, M.B., Girskis, K.M. & Walsh, C.A. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat Rev Genet* **18**, 230-244 (2017).

109. Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat Methods* **9**, 1107-12 (2012).

110. Han, D. *et al.* A Highly Sensitive and Robust Method for Genome-wide 5hmC Profiling of Rare Cell Populations. *Mol Cell* **63**, 711-9 (2016).

111. 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* (2016).

112. 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).

113. Maeder, M.L. *et al.* Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat Biotechnol* **31**, 1137-42 (2013).

114. Therizols, P. *et al.* Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science* **346**, 1238-42 (2014).

115. Liu, X.S. *et al.* Editing DNA Methylation in the Mammalian Genome. *Cell* **167**, 233-247 e17 (2016).

116. Wen, L. & Tang, F. Single cell epigenome sequencing technologies. *Mol Aspects Med* **59**, 62-69 (2018).

117. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* **18**, 35-45 (2018).

118. 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**, 852-6 (2016).