UNIVERSITY OF CALIFORNIA

Santa Barbara

Developing single cell multiomics technologies to understand mammalian development

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Chemical Engineering

by

Alex James Chialastri

Committee in charge:

Professor Siddharth S. Dey, Chair

Professor Michelle A. O'Malley

Professor Arnab Mukherjee

Professor Kenneth S. Kosik

December 2021

The dissertation of Alex James Chialastri is approved.

_____
Michelle A. O'Malley

_____
Arnab Mukherjee

_____
Kenneth S. Kosik

_____
Siddharth S. Dey, Committee Chair

November 2021

Developing single cell multiomics technologies to understand mammalian development

DEDICATION

I would like to dedicate this dissertation to my grandparents, without whom I would not be here today. I always cherish the moments we spent together.

Rita C. Donahue (mom-mom)

James F. Donahue (pop-pop)

Dorothy Chialastri (granny C.)

Dr. Augustine Chialastri (poppy)

ACKNOWLEDGEMENTS

I would like to thank Dr. Siddharth S. Dey for his guidance and providing me the ability to pursue my interests. Having arrived with little laboratory experience, under his guidance I have grown immensely as a scientist and thinker. I would like to thank my committee members, Dr. Michelle A. O'Malley, Dr. Arnab Mukherjee, and Dr. Kenneth S. Kosik, for their continued support and helpful feedback.

I am also grateful for the support of my family. I am especially grateful for my parents, Dr. Gregg Chialastri and Patricia Donahue Chialastri, for their weekly phone calls and unconditional love. I am also grateful for my brother and sister, Dr. Paul Chialastri and Kate Chialastri, for taking care of me and being my role models. I would also like to thank my fiancée Kiran Vasudevan for her constant encouragement and will always be grateful to her for uprooting her life to move out here with me. Lastly, I would like to thank all of my friends who reminded me there is more to life than my science projects.

VITA of Alex James Chialastri
November 2021

EDUCATION

B.S. in Chemical Engineering, Summa Cum Laude, June 2016 – Drexel University

M.S. in Chemical Engineering, Summa Cum Laude, June 2016 – Drexel University

Ph.D. in Chemical Engineering, Bioengineering Emphasis, 2016 – 2021 (expected) – University of California, Santa Barbara

PROFESSIONAL EMPLOYMENT

2013: **Assistant Associate Engineer***, Merck & Co* (West Point, PA)

2014: **Production Engineer/Run Plant Engineer,** *The Dow Chemical Company* (Midland MI)

2015: **Global "MAKE" Engineer,** *Johnson & Johnson–Neutrogena Plant* (Los Angeles, CA)

2016-2021: **Graduate Student Researcher: Dey Lab,** *University of California, Santa Barbara* (Santa Barbara, CA)

PUBLICATIONS

Chialastri A, Gell J, Wamaitha SE, Clark AT, Dey SS. "scMTH-seq: connecting 5-methylcytosine, 5-hydroxymethylcytosine, and the transcriptome from the same single cell reveals processes responsible for human primordial germ cell maturation and DNA methylation erasure" (In preparation)

Chialastri A, Dey SS. "scDyad&T-seq: heterogenous global demethylation in naïve embryonic stem cells results from differences in DNA methylation maintenance in single cells" (In preparation)

Chialastri A*, Wangsanuwat C*, Dey SS. "Integrated single-cell sequencing of 5-hydroxymethylcytosine and genomic DNA using scH&G-seq" *STAR Protocols* (accepted) [* denotes equal contribution]

Chialastri A, Karzbrun E, Khankhel AH, Radeke MJ, Streichan SJ, Dey SS. "Integrated single-cell sequencing reveals principles of epigenetic regulation of human gastrulation and germ cell development in a 3D organoid model" *Nat. Genet.* (Under review)

Wangsanuwat C, **Chialastri A**, Aldeguer JF, Rivron NC, Dey SS. A probabilistic framework for cellular lineage reconstruction using integrated single-cell 5-hydroxymethylcytosine and genomic DNA sequencing. *Cell Reports Methods* (2021) 1:100060

Sen M*, Mooijman D*, **Chialastri A*,** Boisset JC, Popovic M, Heindryckx B, de Sousa Lopes C, Dey SS, van Oudenaarden A. "Strand-specific single-cell

methylomics reveals distinct modes of DNA demethylation dynamics during early mammalian development" *Nat Commun.* (2021) 12:1286 [* denotes equal contribution]

Chialastri P, **Chialastri A**, Mueller T. "Does Prostatic Urethral Lift Reduce Urinary Medications? Trends in Medical Treatment Before and After Prostatic Urethral Lift" *J Endourol*. (2021) 35(5):657-662

Gell J, Liu W, Sosa E, **Chialastri A**, Hancock G, Tao Y, Wamaitha SE, Bower G, Dey SS, Clark AT. "An Extended Culture System that Supports Human Primordial Germ Cell-like Cell Survival and Initiation of DNA Methylation Erasure" *Stem Cell Reports*. (2020) 14(3):433-446

Markodimitraki CM, Rang FJ, Rooijers K, de Vries SS, **Chialastri A**, de Luca K, Lochs SJA, Mooijman D, Dey SS, Kind J. "Simultaneous quantification of protein–DNA interactions and transcriptomes in single cells with scDam&T-seq" *Nat Protoc.* (2020) 15(6):1922-1953

Jang J, Han D, Golkaram M, Audouard M, Liu G, Bridges D, Hellander S, **Chialastri A**, Dey SS, Petzold LR, Kosik KS. "Control over single-cell distribution of G1 lengths by WNT governs pluripotency" *PLoS Biol.* (2019) 17(9):e3000453

Rooijers K, Markodimitraki C, Rang F, de Vries S, **Chialastri A**, de Luca K, Mooijman D, Dey SS, Kind J. "Simultaneous quantification of protein-DNA contacts and transcriptomes in single cells" *Nat Biotechnol.* (2019) 37(7):766-772

Jang Y, Park Y, Nam J, Yang Y, Lee J, Lee K, Kang M, **Chialastri A**, Noh H, Park J, Lee J, Lim K. "Nanotopography-based engineering of retroviral DNA integration patterns" *Nanoscale* (2019). 11(12):5693-5704

AWARDS

Connie Frank Fellowship: For advances in human health and well-being (2021)

Best Seminar Talk Award: Chemical engineering seminar series, first place (2019)

Honorable Mention: Outstanding contributions to UCSB bioengineering (2017)

Johnson & Johnson Gold Encore Award: Implementation of safety improvements (2015)

Academic All-America Division I Men's At-Large Third Team: Swimming (2015)

ABSTRACT

Developing single cell multiomics technologies to understand mammalian development

by

Alex James Chialastri

Next generation sequencing has been key in unlocking the ability to detect thousands of features at single base resolution in thousands of single cells simultaneously in a single experiment. In addition to the DNA bases present, DNA sequencing libraries can contain information on RNA transcripts as well as epigenetic features central to cellular identity like DNA methylation (5mC), DNA hydroxymethylation (5hmC), and DNA accessibility. This dissertation develops multiple single cell sequencing methodologies to explore these epigenetic features simultaneously from the same cell. We first developed scMspJI-seq to detect 5mC from single cells. To investigate 5mC, DNA accessibility, and the transcriptome from the same cell we built upon scMspJI-seq to create scMAT-seq. Then by incorporating 5hmC detection into this measurement, we gained the ability to detection of all 4 features (scMATH-seq) or a subset of them (scMTH-seq). Finally, by combining these technologies with more traditional techniques we developed scDyad&T-seq to detect the transcriptome and the presence of 5mC on both strands of the same piece of DNA. Using these techniques, 4 key areas of human

development were investigated: 1. pre-implantation development, 2. gastrulation and primordial germ cell (PGC) specification, 3. PGC maturation, and 4. stem cell pluripotency.

Pre-implantation development: The global erasure of 5mC from the parental genomes during preimplantation mammalian development is critical to reset the methylome of gametes to the cells in the blastocyst, but how this process occurs remains unclear. By applying scMspJI-seq, we discover that methylation maintenance is active till the 16-cell stage followed by passive demethylation in a fraction of cells within the early mouse blastocyst. In human embryos we find slightly delayed but similar demethylation dynamics as was found in mice.

Gastrulation and primordial germ cell specification: Human gastrulation is marked by dynamic changes in cell states that are difficult to isolate at high purity, thereby making it challenging to map how epigenetic reprogramming impacts gene expression and cellular phenotypes. Applying scMAT-seq to 3D human gastruloids, we characterized the epigenetic landscape of major cell types corresponding to the germ layers and human primordial germ cell-like cells (hPGCLC). Here we find hPGCLCs are specified from progenitors which emerge from epiblast cells and show transient characteristics of both amniotic- and mesoderm-like cells. Finally, we find that during gastrulation DNA accessibility is tightly correlated to both upregulated and downregulated genes, while reorganization of gene body DNA methylation is strongly related to only genes that get downregulated.

Primordial germ cell maturation: PGC maturation is marked by global erasure of 5mC followed by transient high levels of 5hmC. Extended culture systems can achieve passive demethylation in a subset of hPGCLCs, but what initiates this

heterogenous process is unknown. By applying scMTH-seq to hPGCLCs in extended culture we observe that DND1 and SOX15 likely play a role in the initial phase of passive demethylation experienced by hPGCLCs. Additionally, we find that the hPGCLCs in this system stall in their maturation and do not accumulate high levels of 5hmC.

Stem cell pluripotency: Genome wide erasure of 5mC is associated with the acquisition of pluripotency. By applying scDyad&T-seq to different time points of mouse embryonic stem cells transitioning from a primed to a naïve state of pluripotency, we observe extreme demethylation dominated by passive processes and discover this process is highly heterogenous and delayed in some cells. By connecting RNA expression from the same cells, we detect a small set of genes directly linked to 5mC levels during this transition. Finally, we determine that regions of the genome which escape 5mC reprograming do so by retaining high levels of 5mC maintenance and are associated with specific histone modifications.

TABLE OF CONTENTS

1. Background and Motivation

A. The role of epigenetic features and their detection

1. DNA accessibility

DNA accessibility is a critical epigenetic feature that drives cell type specific gene expression. In eukaryotes, DNA is bundled into collections of nucleosomes, which are comprised of an octamer core of two of each of the four core histones: H2A, H2B, H3 and H4[1]. Each nucleosome is wrapped by 147 base pairs of DNA and separated by linker DNA[2]. The positioning of nucleosomes can significantly modify the *in vivo* DNA binding ability of transcription machinery and other DNA binding proteins, affecting gene expression, DNA repair, replication and recombination[2]. There are at least 80 known covalent modifications to the histone proteins that make up a nucleosome, in addition to this, there are numerous histone variants[2,3]. DNA is also highly bound by transcription factors, architectural proteins, and other chromatin-binding factors, which results in further complexity in assessing the state of the DNA[4,5]. To profile even a small fraction of these modifications in a single system is difficult, time-consuming, and expensive. Fortunately many of these features impart a physical characteristic on the local DNA structure, affecting how accessible the DNA is for binding[4]. Thus, obtaining a measurement of DNA accessibility or determining the absence or presence of nucleosomes at specific locations in the genome allows for the profiling of cellular state and can be used to find regulatory regions for a given biological system (Fig. 1.1).

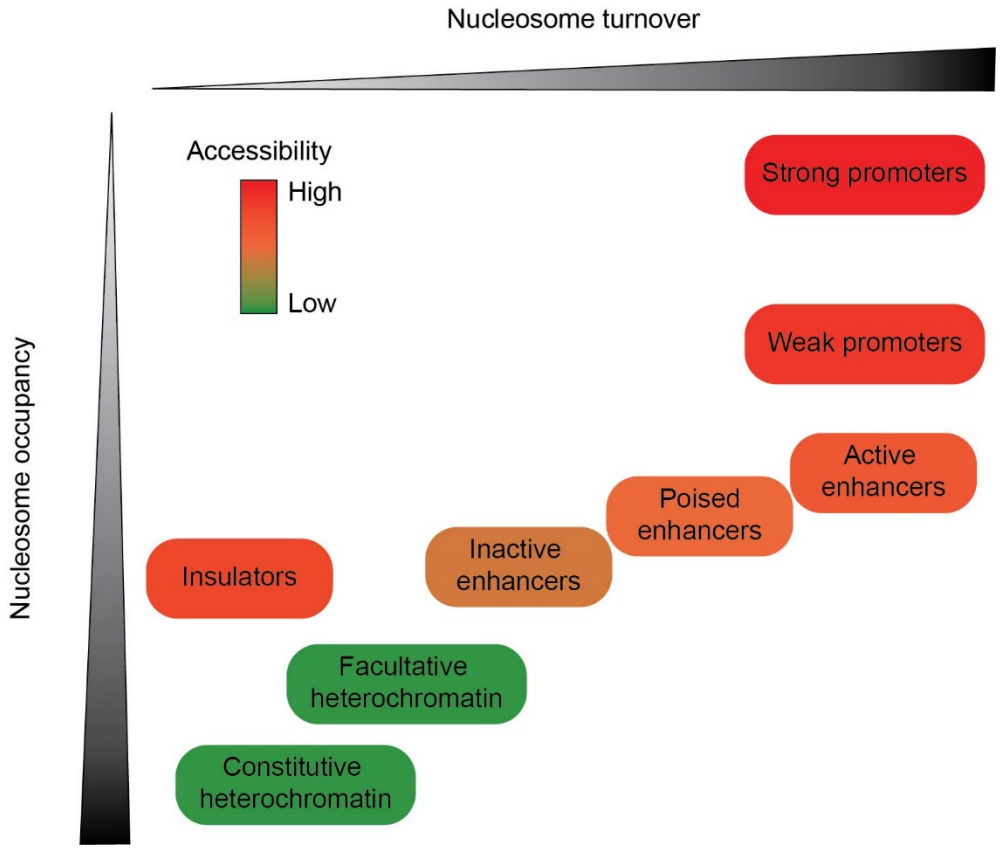


Figure 1.1 | Regulatory regions of the genome differ in nucleosome turnover and occupancy, impacting DNA accessibility.

Accessible DNA can be enriched in regions with low nucleosome occupancy and high nucleosome turnover, but it can also be high in regions with stable nucleosomes for instance insulators marked by CTCF. This figure is adapted from Klemm *et al*.[4]

DNA sequencing is a highly popular method for identifying region specific information, as it allows for both base specific and genome wide studies to be performed simultaneously. The local DNA accessibility has commonly been interrogated using DNase I digestion. The DNA fragments created are enriched in open chromatin regions and their genomic location can be identified through sequencing (DNase-seq)[6,7]. Another common technique uses micrococcal nuclease to fully digest open chromatin, leaving behind only DNA wrapped in nucleosomes to be purified and sequenced (MNase-seq)[8]. Many DNA accessibility methods rely on

similar principles, where open areas of DNA are more accessible to enzymatic digestion. These highly accessible regions are also more readily modified by DNA modifying enzymes like M.CviPI, which can methylate cytosines in a guanine followed by a cytosine (GpC) in the 5' to 3' direction context[9]. Whole genome bisulfite sequencing or a variety of other methods which will be discussed in the next section, can be used as a read out and endogenous methylation can be split from exogenous methylation based on the methylation context, revealing the locations of nucleosomes (NOMe-seq[10]). In addition to these techniques, the binding of DNA to nucleosomes can further be leveraged in DNA sequencing library preparation by simply crosslinking proteins to the DNA, fractionating the DNA, and using a phenol/chloroform extraction to isolate free DNA from DNA crosslinked to proteins for sequencing (FAIRE-seq & Sono-seq)[11,12]. All the techniques discussed so far are fairly time consuming, while due to its speed, high efficiency, and ease, a relatively new method known as ATAC-seq is now commonly used, where a hyperactive transposase (Tn5) is used to cut and insert adaptors into open DNA[13]. With these benefits, it is not surprising that ATAC-seq has been scaled down to the single cell level and many aspects have been optimized to increase scale and sensitivity. Currently, the use of Tn5 is the dominate methodology to profile DNA accessibility in single-cells (scATAC-seq)[14–17]. While less popular, most of the other methodologies discussed here are now also possible at single-cell resolution including scDNase-seq, scMNase-seq, and scNOMe-seq[18–21].

2. *5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), and other oxidized cytosine derivatives*

One of the most well-studied epigenetic marks in mammals is DNA cytosine methylation, 5-methylcytosine (5mC). It has been found that 5mC primarily occurs in mammals in the context of a cytosine followed by guanine (CpG) in the 5' to 3' direction[22]. At lower frequencies, 5mC can also occur in other contexts, for example cytosine followed by adenine (CpA) in the 5' to 3' direction[22]. Among other roles, CpG methylation is critical for maintaining stable repression of target genes, establishing parent-specific gene expression, and maintaining cell type specific identity[22–24]. The role of 5mC in maintaining cell type specific identities makes it an important epigenetic mark in the study of tissue development.

Methylation dynamics can be described through active and passive processes[25]. After DNA replication and mitosis, each daughter cells double-stranded DNA comprises of one original strand of DNA and one new strand of DNA. During DNA replication in the previous cell cycle, the DNA methyltransferase 1 (Dnmt1) protein faithfully copies 5mC in a CpG context to the new strand[26]. If Dnmt1 is downregulated and copying of 5mC to the new strand does not occur, then 5mC is passively demethylated. Active demethylation involves the oxidation of 5mC to 5-hydroxymethylcytosine (5hmC) by the Ten-Eleven Translocation (TET) protein family members (TET 1, 2, and 3)[25]. TET family members can further modify 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), both of which can be recognized by DNA repair pathways and be replaced by an unmodified cytosine (C)[25]. The DNA methyltransferase (DNMT) family of proteins catalyze the reaction of cytosine to 5mC (Fig. 1.2)[27]. As previously mentioned, DNMT1 prevents passive demethylation from occurring in a CpG context. Unlike DMNT1, other DMNT

proteins, DNMT3a and DNMT3b act throughout the cell cycle and can actively methylate cytosines in a *de novo* fashion, primarily in a CpG context[27–29].

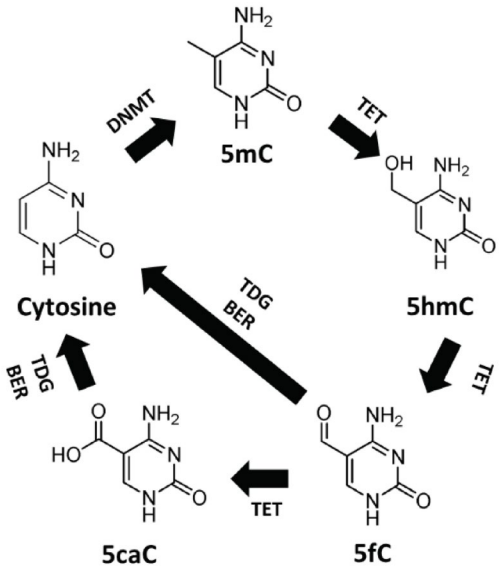


Figure 1.2 | Cytosine modification dynamics.

Cytosine can be methylated by the DNMT protein family. TET proteins then further oxidize 5mC into 5hmC. Further oxidation by TET can occur to 5fC and 5caC, both of which can be acted upon by DNA repair pathways, for example the Thymine-DNA Glycosylase (TDG) and Base Excision Repair (BER) pathways, to be replaced with an unmodified cytosine.

The only known biological route to obtain 5hmC is through the TET family of dioxygenases, which can actively oxidize 5mC (Fig. 1.2)[30]. In most cell types, oxidized forms of 5mC are present at very low levels compared to the amount of 5mC[25]. Even in embryonic stem cells which have high levels of these oxidized forms of cytosine, the levels of 5fC and 5caC are at least an order of magnitude lower than the levels of 5hmC[31,32]. Thus, to understand methylation and demethylation dynamics in different biological systems, quantifying 5mC and 5hmC simultaneously from the same cell is the most critical. Studies have shown that 5hmC is enriched in gene bodies of actively transcribed genes[33]. Obtaining base pair resolution

information of both 5mC and 5hmC combined with gene expression data from the same cell would also allow for the direct evaluation of the role of cytosine modifications on gene expression.

In addition to its biological role, a previous studies from our group have shown that 5hmC can be used to infer cellular lineages[34,35]. Lineage can be inferred based on a lack of maintenance of 5hmC during cell division, which results in the oldest DNA strands having a majority of 5hmC. This lack of maintenance means that the relative levels of 5hmC on the two strands of DNA will be anticorrelated for each chromosome between daughter cells. This framework has been extended to enable probabilistic lineage reconstruction up to the 32-cell stage of mouse embryogenesis[35].

To detect 5mC, regions of the genome containing 5mC can be enriched for prior to sequencing. 5mC enrichment-based techniques are highly similar to Chromatin immunoprecipitation sequencing (ChIP-seq) experiments, except an antibody that detects 5mC is used (MeDIP-seq) or a methyl-CpG-binding domain (MBD) coated bead is used (MethylCap-seq)[36,37]. Similar ChIP-seq based methodologies have also been developed for 5hmC (hMeDIP-seq), in addition to chemistries to tag 5hmC with biotin for enrichment (hMe-Seal)[38,39]. Unfortunately, these techniques do not give base specific resolution. To do this, there are three broad methodologies that are commonly used, nucleobase conversion-based techniques, enzymatic detection-based techniques, and amplification free direct detection with third-generation DNA sequencing.

The gold standard in detecting 5mC is nucleobase conversion though the use of sodium bisulfite, where 5mC and 5hmC are unaffected by sodium bisulfite treatment,

but unmethylated cytosine converts to uracil, resulting in a point mutation during amplifiation[40]. Using this methodology, bisulfite sequencing of 5mC can be profile on the whole genome (WGBS) or bisulfite sequencing can be performed in only CG rich regions by using restriction enzymes that cut CG rich motifs giving a cost effective reduced representative profile (RRBS-seq and XRBS)[41–43]. Additionally, bisulfite sequencing has also been adapted to investigate methylation patterns on complimentary DNA strands using hairpin-bisulfite sequencing. In hairpin-bisulfite sequencing, DNA is fragmented and ligated to a hairpin adaptor, which linearly connects the two DNA strands. After ligation, the samples are exposed to sodium bisulfite to reveal fully methylated or hemimethylated CpG dyads[44].

Although 5mC is present at much higher levels than 5hmC and typically represents only a small error in the measurement of 5mC, there have been many modifications to the bisulfite procedure to specifically measure only one of these epigenetic features[25]. By first reacting genomic DNA with potassium perruthenate ($KRuO_4$) or potassium ruthenate (K_2RuO_4), 5hmC can be converted to 5fC, afterwards bisulfite conversion results in signal only originating from 5mC (Oxy-BS)[45]. Similarly, 5hmC can be blocked by enzymatically adding a glucose moiety and TET enzymes can be added to oxidize 5mC to 5caC, afterwards bisulfite conversion results in signal only originating from 5hmC (TAB-seq)[46]. Sodium bisulfite is not the only reagent available to perform nucleobase conversion reactions. Pyridine borane and 2-methylpyridine borane (pic-borane) can be used to selectively convert 5fC and 5caC to dihydrouracil, resulting in a point mutation during amplifiation[47,48]. By using TET enzymes before chemical conversion both 5hmC and 5mC will undergo a point mutation and can be observed (TAPS)[47,48]. Instead, if

5hmC is protected prior to TET usage, only 5mC will undergo a point mutation (TAPSβ), conversely if instead $KRuO_4$ or K_2RuO_4 is initially used prior to conversion, only 5hmC will undergo a point mutation (CAPS)[47,48]. Enzymatic means of performing this nucleobase conversion are also possible using members of the AID/APOBEC family of enzymes, which can catalyze the deamination of cytosine to uracil. APOBEC3A has specifically been useful in DNA sequencing library construction because of its ability to deaminate both cytosine and 5mC but not 5hmC in benign conditions, allowing for the detection of 5hmC (LR-EM-seq)[49]. Adding TET into the protocol allows for the detection of both 5mC and 5hmC, similar to bisulfite sequencing (ACE-seq & EM-seq)[50,51]. Techniques involving nucleobase conversion have scaled down to the single cell level and currently mainly involve optimized bisulfite library preparations for low input (scBS-seq, scRRBS, scmC-seq, scmC-seq2, sci-MET, scWGBS, and scXRBS), but a APOBEC3A based 5hmC sequencing technique has also been recently developed (snhmC-seq) [43,52–58].

The detection of 5mC and 5hmC is also possible though the use of enzymatic detection-based sequencing techniques. MspI cuts DNA regardless of methylation status while isoschizomer HpaII has the same recognition sequence but cannot cut methylated DNA, using these two enzymes during library preparation allows for the detection of methylated or unmethylated sites (HELP-seq and MSCC-seq)[59–61]. This methodology can be scaled down for single-cell applications by digesting the DNA with HpaII first, ligating specific adapters and then performing a subsequent MspI digestion and adapter ligation (DARE-seq) [62]. An alternative approach is to instead use restriction enzymes with high specificity for only 5mC or 5hmC. The PvuRts1I family of restriction enzymes has high affinity for 5hmC and glucosylated 5hmC

when compared to 5mC and has been successful used to study 5hmC in single cells (scAba-seq)[34,63]. The MspJI family of enzymes have a similar trait but have high affinity for 5mC over 5hmC and only very limited affinity for glucosylated 5hmC, and as part of the work described here, we have successfully used MspJI to detect 5mC in single-cells (scMspJI-seq)[64,65]. Both scAba-seq and scMspJI-seq methodologies are used in this dissertation as the starting points for further developing single-cell multiomics technologies.

3. Detection of mRNA expression in single cells

There are three widely used methods for measuring gene expression in single cells, namely single-molecule mRNA fluorescence *in situ* hybridization (mRNA-FISH), single-cell quantitative PCR (qPCR), and single-cell mRNA sequencing[66]. mRNA-FISH is a powerful technique because it retains the spatial information of the cell in a tissue as well as its ability to accurately quantify low transcript numbers through counting of fluorescent dots under a microscope[67]. The main limitation of this technique is the limited number of genes that can be interrogated simultaneously from the same cell. Single-cell qPCR utilizes gene-specific primers to amplify and detect single transcripts enabling comparison between cells[68]. While powerful, this technique requires some prior knowledge of genes to compare between cells. Single-cell mRNA sequencing allows for the potential detection of all transcripts, giving higher dimensional data to discern variability between individual cells.

Single-cell RNA sequencing was first developed in 2009 and since that time there has been a vast increase in the number of techniques described in literature[69].

Most techniques count the detection of transcripts at the 3' end, but some are used to identify the fully length of transcripts for example SMART-seq[70]. While most techniques are relatively low throughput and rely on sorting individual cells into a reaction well, extremely high throughput methods have recently been developed including those that co-isolate cells with barcoded beads using microfluidic devices (Drop-seq, inDrop-seq, and commercially available microfluidic systems for scRNA-seq like the 10x Genomics system)[71,72]. Even higher throughput can be achieved by removing the necessity to isolate individual cells, instead using the cell as its own reaction vessel, permeabilizing it and delivering barcoded sequencings. By pooling these cells and performing multiple rounds of barcoding, most cells receive a unique barcode sequence and can be computationally demultiplex (sci-RNA-seq and SPLiT-seq)[73,74].

The work presented here is based off a single-cell mRNA sequencing technique called as CEL-Seq2 (Fig. 1.3). In this technique, mRNA from individual cells is reverse transcribed using a primer with an overhang containing a cell-specific barcode, a stretch of random nucleotides known as unique molecule identifiers (UMI), a part of the 5' Illumina sequence, and a T7 promoter. After second-strand synthesis, the cDNA from individual cells is pooled and amplified using *in vitro* transcription, producing single-stranded amplified RNA (aRNA) that does not contain the T7 promoter sequence. The aRNA is then reverse transcribed with a random hexamer primer containing a partial 3' Illumina adaptor. The resulting cDNA is then further amplified through PCR, where the full Illumina adapter sequences are introduced. CEL-Seq2 is reported to have an efficiency of 19.7%, indicating that if a cell had only one transcript of type A, then it would be detected 19.7% of the time[75].

As such, genes that are transcribed at extremely low levels can be problematic to detect, although this method has relatively high sensitivity compared to many others. Another drawback of this method is the low throughput when compared to the most recently developed droplet and split-and-pool techniques. Lastly all single-cell techniques incur the relatively high cost of Illumina sequencing. While there are some drawbacks, CEL-Seq (the unoptimized predecessor to CEL-Seq2) has previously been combined with genomic DNA sequencing in DR-seq, which enabled sequencing both genomic DNA and mRNA from the same cell without the need to physically separate the nucleic acids before amplification[76]. Additionally, the process of amplification is the same for CEL-Seq2 as it is in our enzymatic methods for quantifying 5mC and 5hmC, thereby making it potentially compatible to make multiple measurements from the same cell.

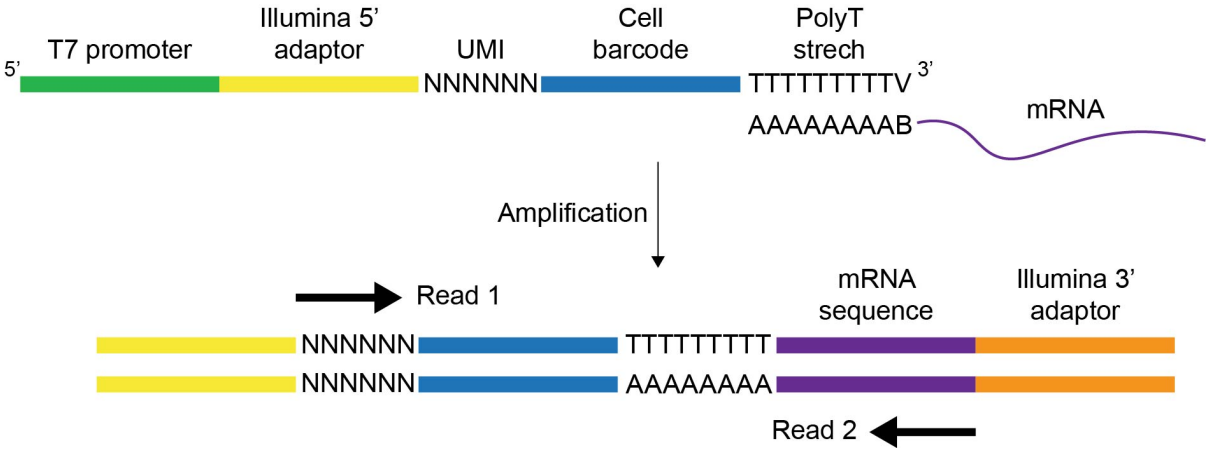


Figure 1.3 | DNA sequencing library schematic in CEL-Seq2.

mRNA is reversed transcribed into cDNA. Individual cells are then pooled and amplified with IVT followed by random priming, PCR amplification and paired end sequencing. Figure adapted from Hashimshony *et al*.[27].

4. *Single cell multiomics methods involving 5mC, 5hmC, DNA accessibility and RNA*

Since the starting material in bulk assays can be split into multiple different sequencing experiments, multiomics as discussed here applies to single cells or in systems where starting material is inherently limited. Recently, there has been an explosion in single cell multiomics techniques, specifically those addressing both DNA accessibility and RNA. These techniques mainly rely on the activity of a hyperactive transposases (Tn5) and work with nearly every technology for performing scRNA-seq, such as the Fluidigm C1 plateform (ASTAR-seq), traditional plate based (scCAT-seq), higher throughput droplet based (SNARE-seq), and ultra-high throughput split-and-pooling methodologies (sci-CAR, Paired-seq, SHARE-seq)[77–82]. Together, these technologies have shown a strong connection between DNA accessibility and RNA expression and have laid the groundwork for what multiomics methodologies can provide in terms of greater biological insight than only one modality.

Other single-cell technologies for reading DNA accessibility, have already been discussed previously, one technology scNOMe-seq, is particularly interesting because it uses a GC methyltransferase for marking open chromatin and then uses bisulfite sequencing as the readout, which results in the detection of both DNA accessibility and DNA methylation simultaneously from the same single cell[21]. This concept has been enhanced to get high detection in each individual cell and utilize the sequencing results to identify the ploidy of the cells involved (scCOOL-seq and iscCOOL-seq)[83,84]. Compared to solutions using Tn5, connecting the transcriptional output of a cell with bisulfite sequencing based single-cell techniques has proven more difficult. The most notable methods come from techniques first developed to assay DNA methylation and the transcriptome from the same cell, and utilize

magnetic beads containing a polyT sequence (scM&T-seq, scTEM-seq, and Smart-RRBS), ultra-centrifugation to separate the nucleus from the cytoplasm (scTrio-seq), or simply picking manually picking the nucleus and leaving the cytoplasm behind (scMT-seq)[85–89]. More recently, one technique has been developed to perform this feat without physical separation, which was done by incorporating methylated cytosine nucleotides into the reverse transcription, thereby allow RNA and DNA to be deconvoluted computationally after sequencing using the level of methylation observed (snmCT-seq)[90]. By incorporating these technologies with the GC methyltransferase used in scNOME-seq, it is possible to obtain 5mC, DNA accessibility and the transcriptome all from the same cell (scNMT-seq, scChaRM-seq, scNOMeRe-seq, and snmC2T-seq)[91–94]. Most of these techniques require physical separation of the nucleus and cytoplasm and are inherently low throughput, limiting processing capabilities to the order of tens of cells per experiment, thus limiting their applications to systems where known cell type enrichment techniques are available or where the number of cells involved are low. These techniques also only profile cytoplasmic RNAs, conversely, the one technique not requiring physical separation (snC2T-seq) has only been performed on single-nuclei and thus while potentially higher throughput, it only profiles nuclear RNAs. Currently, all the described techniques of this type use sodium bisulfite for detection and thus have the additional drawback of not being able to differentiate between 5mC and 5hmC. Additionally, no multiomics techniques have been developed to connect 5hmC with transcription or other epigenetic marks. As such, the field could benefit from the development of a modular single-cell technology which could provide any

combination of a measurement of 5hmC, 5mC, DNA accessibility and mRNA simultaneously from the same single cell.

B. Biological systems investigated: early mammalian systems and pluripotency

1. Stem Cells

The ability of stem cells to self-proliferate and differentiate is central to tissue development, homeostasis, and repair. Discerning how stem cells achieve this regulation is critical not only to better understand tissue development and repair but also to gain insights into dysregulation of these pathways that can lead to both aging and cancer[95]. During development, homeostasis or to repair damaged tissues, stem cells must maintain an exquisite balance between controlled differentiation that results in the spatially correct placement of specialized cells and self-proliferation for use in future events. A quantitative understanding of such systems is critical to engineer and control tissue development and repair for regenerative medicine applications where stem cells could potentially be used to replace injured or diseased tissues of patients.

Stem cells dynamically regulate self-proliferation and differentiation through two distinct modes of cell division. Symmetric cell divisions produce identical progeny, while asymmetric cell divisions result in sister cells with different identities (Fig. 1.4a-c). Both cell division strategies can produce differentiating daughter cells with a more restricted fate than the mother cell. This differentiation towards a terminal state with reduced potency and restriction in fate results from changes in the epigenome that regulates gene expression and ultimately cell identity[96].

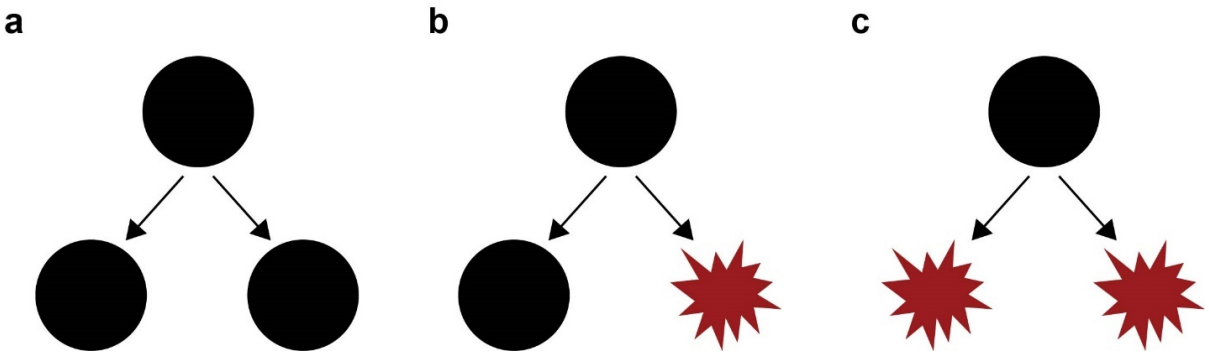


Figure 1.4 | Stem cell divisions.

(a) A symmetric self-proliferation cell division. (b) An asymmetric cell division. (c) A symmetric differentiation division. (a-c) black circle refers to a pluripotent cell, while red star indicates a more differentiated cell.

The epigenome including DNA methylation, hydroxymethylation and DNA accessibility play a central role in cell identity, cellular reprograming and disease[97]. To address how the epigenome tunes symmetric *vs.* asymmetric cell divisions during development, we would need to quantify symmetric and asymmetric cell divisions which requires delineating both the cellular lineage and cell type of cells within a tissue. While it is well understood that tissues are composed of a specific distribution of terminally differentiated cell types, it remains unclear how populations of stem cells communicate and function synergistically to regulate their cell division strategies to produce tissues of precise composition and architecture. Thus, this fundamental unit of fate decision is key to understanding how tissue composition is dynamically regulated, yet it is challenging to make these measurements quantitatively *in vivo* with current technologies. A common technique to reconstruct cell lineages uses genetically modified mice, where the expression of a fluorescent protein can be tracked[98]. While this method has been very insightful, it is primarily limited to studying clonal dynamics instead of individual cell divisions, requires the

generation of complex mouse models for each stem cell of interest and is challenging to apply to opaque tissues. Other methods of cellular lineage tracing, including viral barcoding or CRISPR-Cas9 based genome editing, suffer from precise targeting of stem cells or the need to generate transgenic animals[99]. These limitations can potentially be overcome by employing an endogenous epigenetic mark that is differentially inherited by daughter cells, as was recently shown to occur for 5hmC[34,35]. Further, classical microscopy-based methods also approximate cell types using a single fluorescent marker, which limit our ability to quantitatively discriminate between sister cells, a limitation that is overcome by single-cell mRNA sequencing. Therefore, a technology to simultaneously quantify the epigenetic features 5hmC, 5mC, and DNA accessibility along with mRNA from a single cell, would be useful in understanding how 5mC and DNA accessibility influence symmetric and asymmetric cell divisions *in vivo*.

2. Pre-implantation mammalian development

During mammalian preimplantation embryogenesis, a single fertilized egg (zygote) gives rise to a transient tissue known as a blastocyst (32- to 64-cell stage embryo in mice). The blastocyst contains two distinct cell types, the inner cell mass (ICM) that gives rise to the entire embryo and the trophectoderm that contributes to extraembryonic tissues like the placenta[100]. In several biological systems, the location and local environment around pluripotent cells have been shown to be crucial in maintaining the plasticity of these cells[95]. However, it remains unclear how cell-cell contacts and other signaling cues within the pluripotent cell niche regulate these cell division strategies. During this period of early development, it is well documented that the cellular identity changes drastically, in part due to genome wide

global demethylation (Fig. 1.5). This unique phenomenon is critical to reset the methylation status of the egg and sperm cells and renew the cycle of life. While the mechanisms regulating demethylation in this process have not been definitively determined, it is well known that the TET proteins are expressed at high levels and play an important role in actively converting 5mC to 5hmC at this time[101–103]. New technology to assess 5mC and 5hmC from single cells would allow for studying of the demethylation dynamics from the zygote to the early blastocyst stage of development. A better understanding of preimplantation embryogenesis has implications for both regenerative medicine and *in vitro* fertilization procedures. Broadly, the goal of regenerative medicine is to replace and heal damaged or diseased tissues by creating new healthy tissues. A potential source for creating healthy tissue is embryonic stem cells, which are derived from the ICM of the blastocyst[104]. In addition to regenerative medicine, understanding early embryos has implications for *in vitro* fertilization. In this procedure, eggs and sperm are collected from patients and combined *in vitro*. The resulting fertilized egg develops outside the body until the blastocyst stage, after which it is injected into the uterus[105]. While this procedure has been a scientific breakthrough for couples with infertility problems, it has been associated with an increase in major birth defects[105]. Improving our understanding of blastocyst development has the potential to reveal the mechanisms responsible for this increase in birth defects as well as provide a framework for understanding basic tissue development.

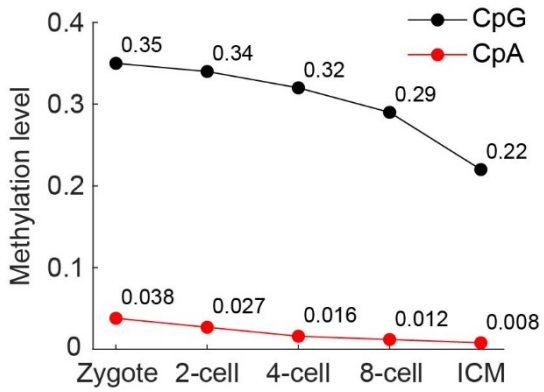


Figure 1.5 | Global methylation levels during early mouse embryogenesis.

Mean CpG (black) and CpA (red) methylation per 100 base pairs during early mouse embryogenesis. ICM indicates the inner cell mass cells of the blastocyst (approximately the 32-cell stage). Adapted from Smith *et al.*[103].

3. Post-implantation mammalian development and PGCs

Gastrulation is a critical step in human embryonic development where pluripotent cells differentiate into lineage specific precursors. However, due to a lack of access to *in vivo* tissue samples, these early stages of human development are difficult to study, limiting our understanding of specific developmental defects and certain cancers, such as germ cell tumors[106]. In mammals, much of what is known about post-implantation development comes from observations made in mouse models. In mice, after implantation the embryo continues to grow and acquires a cup shape with two layers of cells, the inner epiblast and the outer visceral endoderm[107]. After this formation, gastrulation occurs where the embryo comprised of two germ layers subsequently becomes comprised of three primary germ layers, the ectoderm, mesoderm, and endoderm. The process of gastrulation beings with the formation of the primitive streak on the posterior side of the embryo, thus defining the anterior-posterior axis[107]. To create a third layer of cells, at the primitive streak, epiblast cells undergo an epithelial to mesenchymal transition. These cells then migrate between

the epiblast and into the endoderm layer, some migrating cells fill the space between the epiblast and definitive endoderm, becoming the mesoderm layer[107]. While the gastrulation process in humans is fairly similar, primates morphology is distinct from rodents and pre-gastrulating embryos consist of a bilaminar disc, which becomes a trilaminar disc post-gatrulation[108].

Around the time of gastrulation, in mice, the extra-embryonic ectoderm provides critical signaling molecules including bone morphogenetic protein 4 (BMP4) for primordial germ cell (PGC) specification, which occurs at day 6.25 post coitus (E6.25)[109,110]. The area of high signaling is restricted to only the most posterior region of the embryo by inhibitory signals produced by the anterior visceral endoderm, thus resulting in only a handful of specified PGCs[111]. At E7.5 the specified PGCs begin to migrate and soon begin to rapidly proliferate, reaching the genital ridge at E10.5 to form the embryonic gonad, eventually giving rise to mature germ cells[112].

Germ cells are those cells that give rise to the gametes of a sexually reproducing organism. Development of germ cells follows a similar global trend to preimplantation mouse embryogenesis where a small group of highly methylated cells undergo vast epigenetic changes, including global demethylation[113,114]. In this system, demethylation is known to occur in two phases and is known to have impact on cellular identity[115–117]. The first phase occurs when a small group of nascent germ cells actively migrate to the genital ridge. This phase is thought to be dominated by passive demethylation where methylation is lost slowly with each cell division as the methylation marks are not copied over during replication[118,119]. Once the PGCs reach the genital ridge and begin colonization of the gonads, the second period of

demethylation occurs, where methylation is lost quickly through active conversion to 5hmC (Fig. 1.6)[118,119]. Immunofluorescence imaging has revealed that 5hmC is found to only occur with high frequency on one sister chromatid in E12.5 PGCs, an indication that 5hmC accumulation has slowed by this timepoint[120]. At the end of these two waves of demethylation, the genome wide DNA methylation levels are at the lowest seen in any point of development[119].

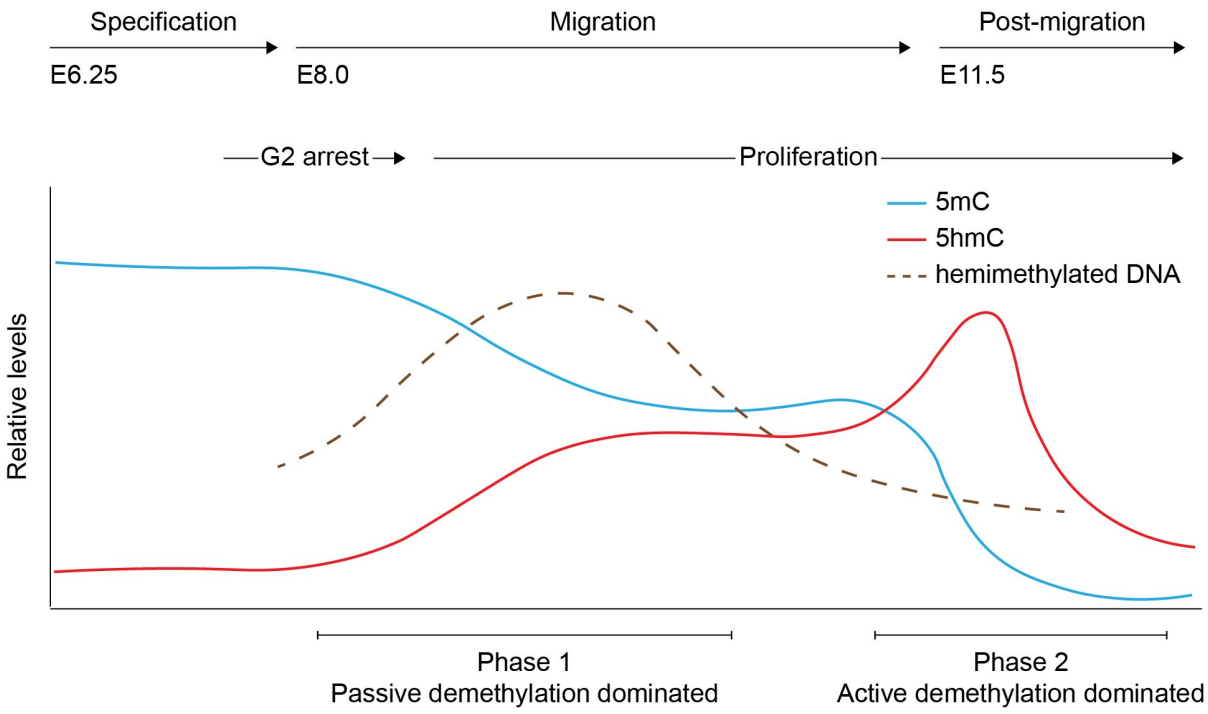


Figure 1.6 | Cytosine dynamics in PGC development.

Demethylation occurs in two phases during PGC development, first a phase dominated by passive demethylation followed by rapid active demethylation by the TET proteins. Figure adapted from Messerschmidt *et al.*[118].

The findings of some sister chromatids having low 5hmC by E12.5 suggests 5hmC could also be used for at least partial cellular lineage reconstruction[120]. Hemimethylated DNA is seen to accumulate in the first phase of demethylation giving confidence that at least partial cellular lineage reconstruction may be

achievable through single-cell 5mC sequencing as the old DNA strand remains highly methylated compared to newer strands[121]. Additionally in another system, hemimethylated DNA has been found to be heritable and stable over many cell cycles, indicating it could be more than an intermediate and instead have a functional role in PGC development[122]. Cellular lineage reconstruction during PGC development would be key in systematically understanding differences in germ cell competency between PGCs. In addition, single-cell 5hmC sequencing will allows the detection of genomic localization of 5hmC, which has been seen to transiently accumulate in centromeric regions throughout PGC development[120]. In this system, it is crucial to detect 5hmC and 5mC from the same cell to fully appreciate the dynamics and role of cytosine modifications during PGC development, and their role on cellular identity.

4. In vitro post implantation organoids and primordial germ cell like cells

Complex studies of mouse post implantation and PGC development have been performed, but these same studies are infeasible in humans[123]. Morphological divergence from mice to humans make it challenging to fully leverage these studies for human health and disorders. To complicate matters further, some critical factors of PGC development in humans differ from those in mouse. One well studied example is the transcription factor Sox17. Sox17 is well regarded for its ability to promote endodermal differentiation in mouse embryonic stem cells, but it has also been found as a critical factor for human PGC formation but is not critical in mouse PGC formation[124,125]. With such complexity and limited human material availability, *in vitro* alternatives have been developed which use human stem cells and defined growth factors to induce gastrulation like behaviors and cellular differentiation[126–132].

Some of the gastrulating organoids developed at this point can induce primordial germ cell like cell (PGCLC) formation, and it has become common to create human PGCLCs from three-dimensional disorganized aggregates[133]. Unfortunately, human PGCLCs created in this manner can only be grown for a short time and only exhibit markers similar to early PGCs[133]. Our collaborators have developed techniques to grow these human PGCLCs for significantly longer, and studies performed in collaboration with the tools developed in this thesis have shown that PGCLCs grown for an extended period begin to passively demethylated similar to their *in vivo* counterparts[134]. While there are many obstacles remaining to produce fully developed human germ cells in *in vitro*, these new systems provide attainable materials for understanding human PGC development.

C. Technological and equipment background

1. Fluorescence-activated cell sorting

Fluorescence-activated cell sorting (FACS) is a method that can be used to sort single cells into individual reaction wells and is described by Tomlinson et al[135]. Briefly, if desired, a single cell suspension is incubated with fluorescent dyes and/or fluorescently labeled antibodies. In a cell sorter, individual cells are encapsulated into liquid droplets and their fluorescence along with forward and side light scattering properties are measured. Based on these signatures from the droplet, it is either collected or not collected. This is done by electrically charging each droplet and passing them through a deflector plate. FACS is useful for rapidly sorting single cells into reaction wells but typically require starting with a large initial number of cells as many cells of interest are not captured. In cases where it is important to capture

each individual cell, for instance during early mouse embryogenesis, manual cell isolation is required.

2. Next-generation sequencing technologies

Due to Next-generation sequencing (NGS) technologies, the cost of sequencing a human genome is approximately $1,000[136]. During the infancy of NGS, many competing technologies for template amplification and base readout were common. A prevalent strategy was sequencing by synthesis (454 pyrosequencing, Ion Torrent, GeneReader, and Illumia), although sequencing by ligation (SOLid) was also used[136]. While all sequencing techniques have their merits and drawbacks, at this point, NGS has become synonymous with Illumina technology. In Illumina sequencing, individual single-stranded DNA templates containing the proper DNA sequences at both ends bind to a flow cell, where bridge amplification occurs allowing for many copies of the same template to be colocalized[136]. Specified primers along with fluorophore-labeled nucleotides are flown in and in each successive round one nucleotide is added, the flow cell is imaged, and then the fluorophores are cleaved and washed away, allowing the cycle to continue[136]. Sequencing of this type is the basis of all the methodologies derived in this dissertation.

While dominate, Illumina technology does suffer from some drawbacks, specifically it can only produce high quality data for short stretches of DNA (<300 bases) and requires short initial templates (<1000 bp). New third-generation sequencing techniques (SMRT and Nanopore sequencing) have been developed which resolve these two issues, but mass adoption has been slow due to initial lower throughput and less sequencing accuracy when compared to Illumina[136,137]. In

principle the techniques developed in this dissertation could be adapted for sequencing on these third-generation platforms, but it remains to be seen if such redesigns would be of practical use.

D. Thesis goals and organization

Much is still unknown about early mammalian development, including how epigenetics play a role in cell fate, pluripotency, and differentiation. This introduction has outlined some of what is known in these systems about the role of 5hmC, 5mC, and DNA accessibility, as well as the key involvement in some genes during this process. While the large number of technological advancements outlined in this introduction have provided key insights into early mammalian development and other biological systems, there is still a need for more advanced multiomics sequencing techniques which can be efficiently scaled in both the number of detected features and the number of cells investigated. This thesis describes the develop of new methods fitting this description, with the ability to sequencing combinations of 5hmC, 5mC, DNA accessibility and mRNA simultaneously from the same single cell. In Chapter 2, an enzymatic detection-based sequencing techniques for detecting 5mC (scMspJI-seq) in single cells and its application towards early mouse and human embryos is discussed. In Chapter 3, scMspJI-seq is expanded to also allow the detection of DNA accessibility and the transcriptome simultaneously from the same cell (scMAT-seq), and this technique is applied to investigate gastrulation and primordial germ cell development in human gastruloids. In Chapter 4, scMspJI is further built upon to provide a measurement of 5hmC and the transcriptome from the same single cell (scMTH-seq) to investigate *in vitro* primordial germ epigenetic reprograming and maturation. In Chapter 5, a novel

technique to detect hemimethylated DNA as well as RNA transcripts from the same cell (scDyad&T-seq) is developed and intricacies of 5mC dynamics during changes in pluripotency is investigated. In Chapter 6, the goal of detecting of 5hmC, 5mC, DNA accessibility and mRNA simultaneously from the same single cell is achieved. In addition to this, in this chapter the limitations of the techniques described in this dissertation and future directions of the field are discussed.

2. Strand-specific single-cell methylomics reveals distinct modes of DNA demethylation dynamics during early mammalian development

This chapter has been reproduced from its original publication in *Nature Communications* with differences[65].

A. Introduction

In mammalian systems, DNA methylation (5-methylcytosine or 5mC) is a key epigenetic modification that is typically stably inherited from mother to daughter cells[138]. This property of 5mC plays an important role in facilitating the propagation of cellular identity through cell divisions and restricting the developmental potential of terminally differentiated cells[138,139]. Consequently, during preimplantation mammalian development, DNA methylation patterns on the terminally differentiated paternal sperm and maternal egg genomes are erased post-fertilization at a genome-wide scale to revert cellular memory towards an undifferentiated state in the blastocyst[140]. Therefore, understanding the mechanisms underlying global DNA demethylation dynamics is central to understanding the emergence of pluripotent cells during early development.

Removal of 5mC can proceed through two alternate mechanisms – passive and active demethylation. Methylated cytosines, within a CpG dinucleotide context are typically copied over to the newly synthesized DNA strands during genome replication by the maintenance methyltransferase, DNMT1[141]. Passive demethylation relies on loss of 5mC through replicative dilution, in which inhibition of DNA methylation maintenance results in a reduction of 5mC levels after cell division and

can be detected through asymmetric levels of 5mC on the two DNA strands of a chromosome. Alternatively, active mechanisms of 5mC erasure occur via conversion of 5mC to 5-hydroxymethylcytosine (5hmC) and other oxidized derivatives, which are not recognized by the DNA maintenance methylation machinery and are subsequently removed by base-excision repair pathways[30,31,142]. While early immunofluorescence-based studies revealed that the paternal genome undergoes active demethylation through conversion to 5hmC in the zygote, the maternal genome was presumed to undergo passive demethylation through the lack of DNMT1 activity during replication[143–146]. Advances in biochemistry, next-generation sequencing and mass spectroscopy based studies improved upon this coarse quantification of methylation dynamics to show that the orthogonal regulation of demethylation by active and passive mechanisms for the two parental genomes was not as distinct as suggested by these early studies. For example, it was later shown that while DNMT1 is mostly cytoplasmic during these early stages of development, low levels of a Dnmt1 isoform, DNMT1s, together with UHRF1 is observed in the nucleus, raising the possibility that 5mC is maintained on the maternal genome[103,117,147–152]. However, the conclusions in these recent studies were partly based on bulk bisulfite-sequencing based methods that could not directly distinguish between active vs. passive demethylation, and therefore the relative contribution of these two mechanisms to 5mC reprogramming remains poorly understood.

B. Results

1. Strand-specific quantification of 5mC using scMspJI-seq

To distinguish between active and passive mechanisms of demethylation requires strand-specific detection of 5mC in single cells. While asymmetric levels of

5mC between two DNA strands of a chromosome would indicate passive demethylation, the global loss of methylation coupled with symmetric levels of 5mC between two DNA strands would indirectly imply active demethylation (Fig. 2.1a)[153]. Therefore, to identify the mechanisms regulating DNA demethylation dynamics, we developed a new method called scMspJI-seq to strand-specifically quantify 5mC on a genome-wide scale in single cells. Single cells are isolated into 384-well plates by fluorescence activated cell sorting or manual pipetting. All downstream steps are subsequently performed using a liquid-handling platform (Nanodrop II, BioNex Solutions). Following cell lysis and protease treatment to remove chromatin, 5hmC sites in genomic DNA (gDNA) are glucosylated using T4 phage β-glucosyltransferase (T4-βGT) (Fig. 2.1b). This modification blocks downstream detection of 5hmC and therefore, enables detection of only 5mC in scMspJI-seq. Next, the restriction enzyme MspJI is added to the reaction mixture that recognizes mCNNR sites in the genome and creates double-stranded DNA breaks 16 bp downstream of the methylated cytosines leaving a 4-nucleotide 5' overhang[64]. Thereafter, double-stranded DNA adapters containing a 4-nucleotide 5' overhang are ligated to the fragmented gDNA molecules. These double-stranded DNA adapters, similar in design to those previously developed by us, contain a cell-specific barcode, a random 3 bp unique molecule identifier (UMI) to label individual 5mC sites on different alleles, a 5' Illumina adapter and a T7 promoter[34,154]. The ligated molecules are then amplified by in vitro transcription and used to prepare Illumina libraries as described previously, enabling the processing of hundreds to thousands of single cells per day (Fig. 2.1b)[34,154].

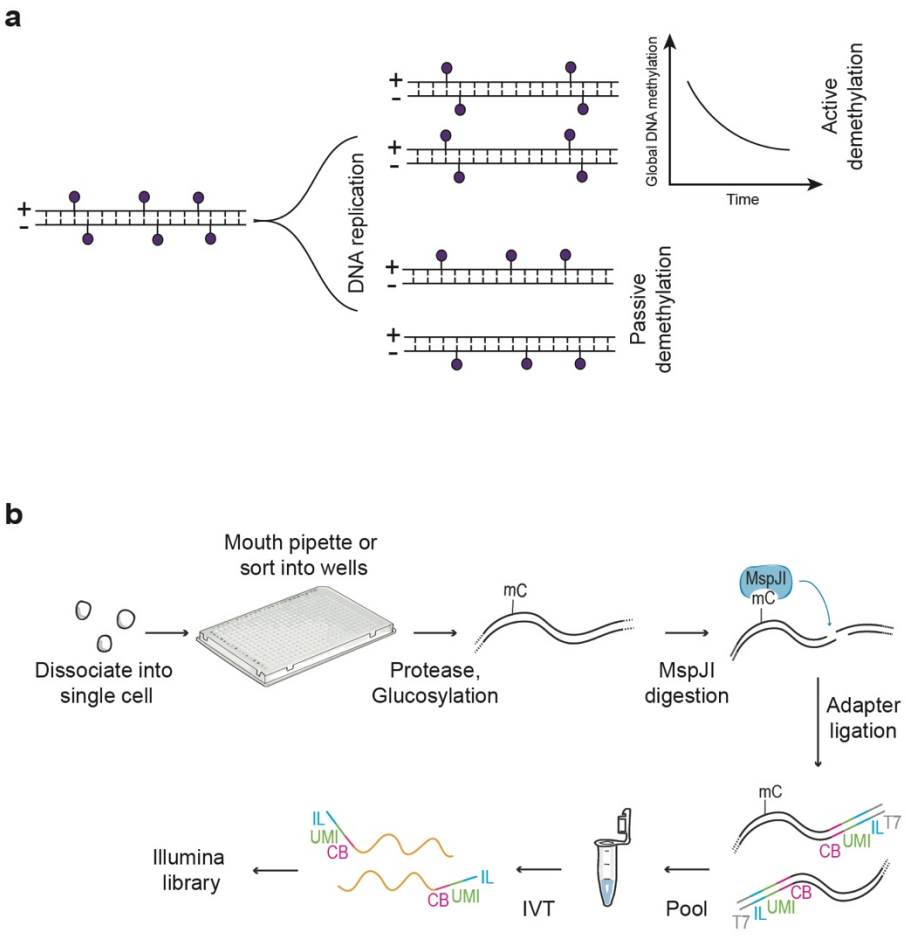


Figure 2.1 | Schematic of scMspJI-seq.

(a) DNA methylation maintenance can be probed using strand-specific quantification of 5mC in single cells. Cells displaying symmetric levels of 5mCpG on both DNA strands of a chromosome coupled with a global temporal loss of 5mCpG indicates active demethylation whereas loss of methylation maintenance with asymmetric levels of 5mCpG between the two DNA strands indicates passive demethylation. (b) Single cells isolated by FACS or manual pipetting are deposited into 384-well plates and lysed. Following protease treatment to strip off chromatin and blocking of 5hmC sites by glucosylation, MspJI is used to recognize 5mC sites and cut gDNA 16 bp downstream of the methylated cytosine. After ligating double-stranded adapters – containing a cell-specific barcode (CB, pink), a random 3 bp unique molecule identifier to label individual 5mC sites on different alleles (UMI, green), 5' Illumina adapter (IL, blue) and T7 promoter (T7, gray) – to the fragmented gDNA, molecules from all single cells are pooled and amplified by in vitro transcription. The amplified RNA molecules are used to prepare scMspJI-seq libraries and sequenced on an Illumina platform.

To validate the method, we first applied scMspJI-seq to single E14TG2a (E14) mouse embryonic stem cells (mES) cells. As reported previously, we found

that MspJI cuts gDNA 16 bp downstream of the methylated cytosine (Supplementary Fig. 2.1)[64]. We detected between 212,000 to 977,000 unique 5mC sites per cell, with a median of 484,000 5mC sites per cell (Supplementary Fig. 2.2). Further, we found that 97.2% of the 5mCpG sites detected by scMspJI-seq in single cells overlapped with methylated sites observed in bulk bisulfite sequencing of E14 gDNA (Supplementary Fig. 2.3a). Similarly, we found that averaged single-cell data from scMspJI-seq correlates well with the bulk bisulfite methylome (Pearson r = 0.84) (Supplementary Fig. 2.3b)[155]. Furthermore, while we observed that the genome-wide distribution of 5mC over different genomic elements in scMspJI-seq was similar to that observed in bisulfite sequencing, we also found that scMspJI-seq shows a slight preference for detection of 5mC sites within genomic regions that have a lower density of CpG sites (Supplementary Fig. 2.4 and 2.5). This possibly occurs as our method is dependent on the digestion of the genome around methylated cytosines, reducing the likelihood of detecting closely spaced 5mC sites. However, both scMspJI-seq and bisulfite sequencing captured similar genome-wide landscapes of 5mC at a variety of genomic elements. For example, we observed similar gene body methylome profiles as well as the expected hypomethylation of CpG islands (CGI) and transcription start sites (TSS) using both methods (Supplementary Fig. 2.6). In addition, compared to single-cell bisulfite sequencing that detects a combination of 5mC and 5hmC sites, a distinct feature of scMspJI-seq is that it can identify only 5mC in the genome by blocking detection of 5hmC sites using T4-βGT. By combining scMspJI-seq data with scAba-seq results, we were able to estimate the false-positive detection rate of 5hmC to be around 1.1% (Supplementary Fig. 2.7)[34]. Most importantly, due to the maintenance activity of DNMT1 in E14 cells, we

observed similar levels of 5mC on both DNA strands of a chromosome in single cells, as expected (Supplementary Fig. 2.8a). To quantify the strand-specific distribution of 5mCpG on each chromosome of a single cell, we defined a metric called as strand bias (denoted by f), which is the ratio of the number of 5mCpG sites detected on the plus strand divided by the total number of 5mCpG sites detected on both the plus and minus strands. Finally, to ensure that scMspJI-seq can detect differences in 5mCpG distribution between the two strands, and to confirm that the observed strand bias of 0.5 in E14 cells results from the maintenance activity of DNMT1, we used CRISPR-Cas9 to knockout Dnmt1. We observed a dramatic increase in strand bias in E14 cells without Dnmt1, strongly suggesting that our new technology provides a sensitive readout of strand-specific methylation and the ability to distinguish between passive and active demethylation (Supplementary Fig. 2.8b).

2. mES cells display heterogeneity in strand-specific 5mC

During preimplantation development, the maternal and paternal genomes display dramatically different 5mC erasure dynamics, and therefore we next wanted to test our ability to quantify strand-specific 5mC at the resolution of individual alleles. As the single-cell measurements in E14 cells did not provide allele-specific detection of 5mC for each chromosome, we applied scMspJI-seq to hybrid serum grown mES cells (CAST/EiJ x 129/Sv background)[154]. While the majority of cells displayed methylation maintenance as expected, we surprisingly observed a small population of cells that showed strong 5mC strand bias (Fig. 2.2). For example, cell 562 displayed similar levels of 5mCpG on the two DNA strands of chromosomes across both alleles (Fig. 2.2a), whereas cell 216 showed substantially different levels of 5mC on each DNA strand of a chromosome (Fig. 2.2b). Pearson correlation

coefficient (r) between the plus and minus strands of individual cells show that while a majority of cells displayed high correlation, a small subset of cells were weakly correlated, suggesting loss of methylation maintenance in these cells (Fig. 2.2c). Allele-specific 5mCpG strand bias further revealed the existence of two epigenetically distinct population of mES cells (Fig. 2.2d). Taken together with the E14 cells, these results highlight that in the absence of allele-specific measurements, strand-specific 5mC quantification is averaged across both alleles, potentially obscuring a detailed view of the methylation status of the genome. Finally, we find that these two distinct 5mC strand bias patterns are also observed at a sub-chromosomal resolution, suggesting this is genome-wide phenomenon that potentially arises from differential methylation maintenance between individual mES cells (Fig. 2.2e).

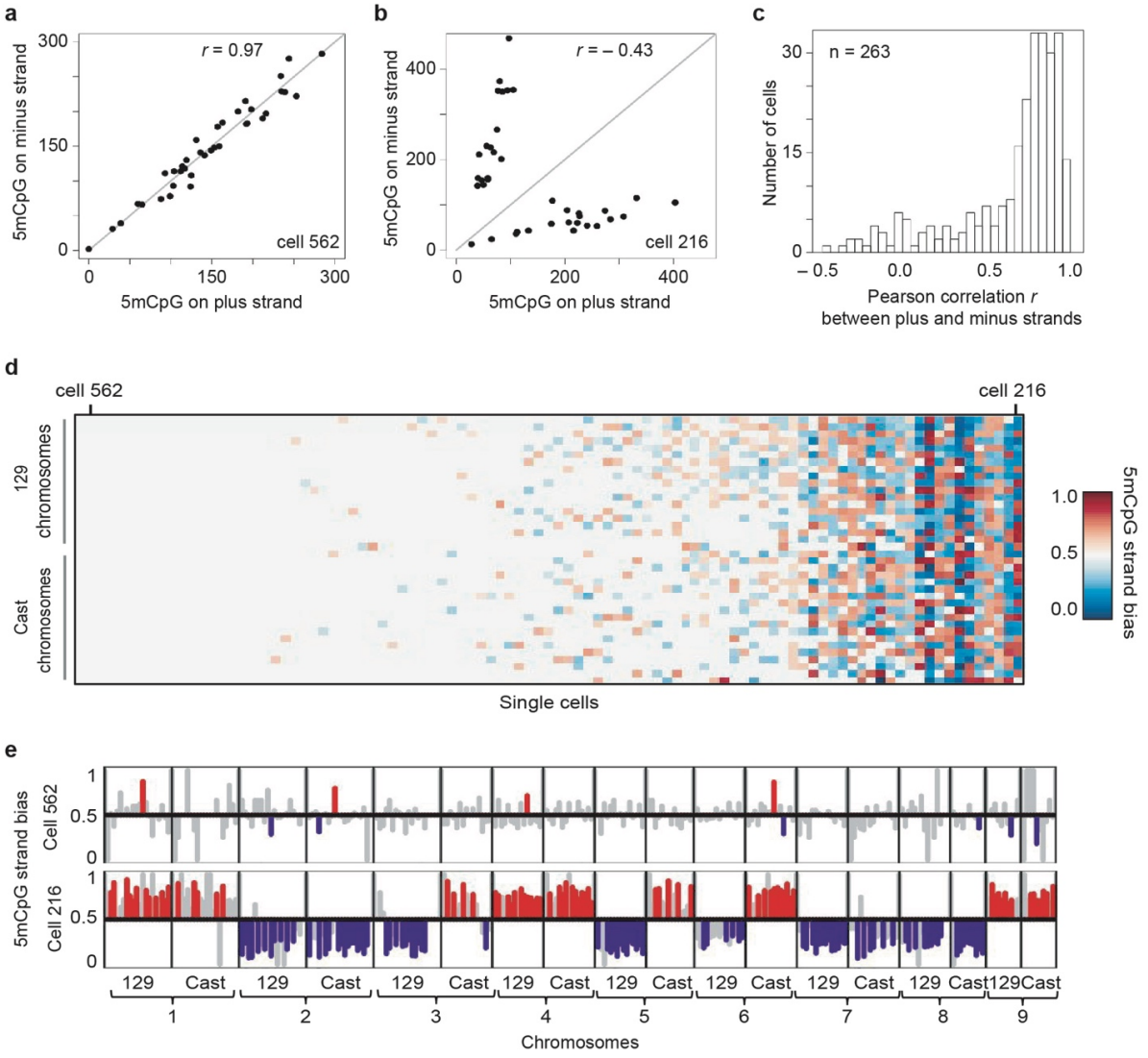


Figure 2.2 | Cell-to-cell heterogeneity in genome-wide strand-specific methylome landscapes in mES cells.

(a) An example of a mES cell (cell #562) processed by scMspJI-seq shows similar amounts of 5mCpG on both the plus and the minus strand of each chromosome. (b) Another mES cell (cell #216) with asymmetric amounts of 5mCpG between the plus and the minus strand of each chromosome. (c) Histogram of Pearson correlations between the 5mCpG levels on the plus and the minus stand over all chromosomes in a cell show that while a majority of cells have similar amounts of 5mCpG on both strands (high Pearson correlation), a small fraction of cells display unequal levels of 5mCpG between the two strands of each chromosome (low Pearson correlation). (d) Ordered heatmap showing 5mCpG strand bias per chromosome for the maternal and paternal alleles in individual mES cells. (e) 5mCpG strand bias of cell #526 (top) and cell #216 (bottom) for 10 MB bins along the first 9 chromosomes are shown with statistically significant ($P < 0.05$, likelihood ratio test) strand biases towards the plus and minus strands shown in red and blue, respectively. Strand biases of bins that are not statistically significant are shown in gray ($P > 0.05$, likelihood ratio test).

To validate this cell-to-cell heterogeneity in 5mC strand bias, we reanalyzed data from a recent study that quantified 5mC in single cells using bisulfite sequencing, a method that can potentially also be used to infer strand-specific 5mC[42,92]. In agreement with our findings using scMspJI-seq, reanalysis of the published dataset also revealed hybrid mES cells with similar levels of 5mC on the plus and minus strands, and a small fraction of cells with substantially different levels of 5mC on the two strands of a chromosome (Fig. 2.3). These results validate our previous observation of two distinct mES cell populations with and without 5mC strand bias (Fig. 2.2).

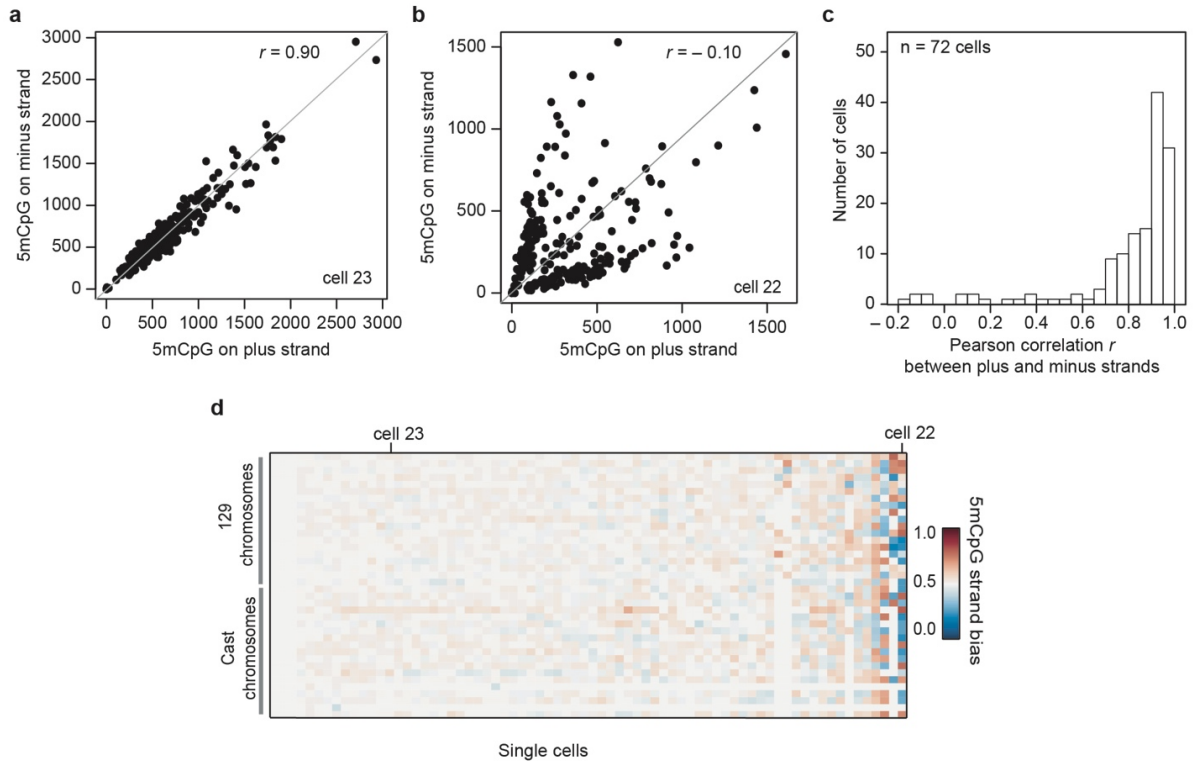


Figure 2.3 | Variability in strand-specific 5mCpG profiles in mES cells.

(a) A representative mES cell (cell #23) with similar amounts of 5mCpG within 10 MB bins on both DNA strands. (b) Another representative mES cell (cell #22) with unequal amounts of 5mCpG between the two DNA strands for 10 MB bins. (c) Histogram of Pearson correlations between the 5mCpG levels on the plus and the minus stand over the entire genome (10 MB) in a cell. (d) Ordered heatmap showing 5mCpG strand bias per chromosome for maternal and paternal alleles in individual

mES cells (n=72). The results in this figure is based on strand-specific reanalysis of single-cell bisulfite sequencing data obtained from previous work by Clark et al.[92].

3. Preimplantation embryos display distinct modes of demethylation dynamics

After establishing this new method, we next used scMspJI-seq to gain a deeper understanding of the 5mC erasure dynamics during preimplantation mouse development as the mechanistic details regulating this genome-wide reprogramming remains unclear from previous work. Early immunofluorescence-based studies showed that 5mC marks on the paternal genome are converted to 5hmC in the zygote[143–146]. As 5hmC is not maintained through cell division and can be further oxidized to be removed by cytidine deaminase and base excision repair pathways, the paternal genome is effectively demethylated from the 1-cell to early blastocyst stage (approximately E3.5 or 32-cell stage) of development[142]. These same studies also reported that the maternal genome retains 5mC in the zygote[143–146]. This observation together with reports that DNMT1 is primarily cytoplasmic during these early cell divisions, indirectly suggested that the maternal genome is passively demethylated through a lack of maintenance methylation[156–159]. However, later studies showed the existence of two isoforms of Dnmt1, with the lowly abundant DNMT1s isoform present in the nucleus of blastomeres[160–162]. Thus, it remains unclear the extent to which the maternal genome is passively demethylated during these early stages. Further, more recently, bulk 5mC and 5hmC sequencing during these early stages have shown that the maternal genome also carries 5hmC marks, suggesting that the maternal genome also undergoes partial active demethylation[117]. As the mechanisms underlying this critical process of 5mC erasure during embryonic

development remains unclear, we used strand-specific detection of 5mC in single cells to probe the dynamics of demethylation more closely.

We performed scMspJI-seq on hybrid mouse embryos (CAST/EiJ x C57BL/6 background) from the 2- to 32-cell stage of development. In contrast to previous studies that suggested passive demethylation of the maternal genome due to cytoplasmic localization of DNMT1, experiments in 2-cell hybrid mouse embryos surprisingly revealed that 5mCpG on the maternal genome shows a tight strand bias distribution centered around 0.5, implying similar amounts of the mark of both DNA strands and that DNMT1-mediated methylation maintenance is active at this stage (Fig. 2.4a and Supplementary Fig. 2.9a). To ensure that this lack of strand bias in the maternal genome at the 2-cell stage is not a technical artifact or a consequence of high de novo methylation activity of DNMT3a/3b, we quantified the levels of 5mCpA, the most abundant non-CpG methylation, in these cells. Non-CpG methylation is not a substrate for DNMT1 and is deposited on the genome as a result of the activity of the de novo methyltransferases, DNMT3a and DNMT3b[121,163,164]. In the 2-cell embryos, we found that 5mCpA on the maternal genome showed a bimodal pattern of strand bias distribution, suggesting that the lack of strand bias observed for 5mCpG is possibly a result of the maintenance activity of DNMT1 and not a consequence of high de novo methylation rates by DNMT3a/3b (Fig. 2.4b and Supplementary Fig. 2.9b). Further, we have previously shown that bimodal strand bias distributions for 5hmC in 2-cell mouse embryos arises from the slow kinetics of Tet activity and can be used to identify sister cells[34,35]. This is because 5hmC is not maintained through cell divisions and new DNA strands have lower levels of 5hmC than older strands, resulting in sister cells

exhibiting anti-correlated strand bias patterns over all the chromosomes in a cell. Similarly, as 5mCpA is not maintained through cell division, we found that the strong anti-correlation in 5mCpA between chromosomes of single cells can be used to identify sister cells (Supplementary Fig. 2.9c,d). These results further imply that at the 2-cell stage of development the kinetics of de novo methylation by DNMT3a and DNMT3b is slow (Fig. 2.4b). Taken together, these experiments provide preliminary evidence that the similar levels of 5mCpG found on both DNA strands of chromosomes in 2-cell blastomeres is a result of DNMT1 maintenance activity.

Quantifying the dynamics of demethylation beyond the 2-cell stage, we observed for both the maternal and paternal genomes that a majority of chromosomes displayed no significant 5mCpG strand bias up to the 16-cell stage (Fig. 2.4a and Supplementary Fig. 2.9a). Surprisingly, beyond the 16-cell stage, we observed a widening of the 5mCpG strand bias distribution, suggesting reduced DNMT1 maintenance activity (Fig. 2.4a and Supplementary Fig. 2.9a). These experiments suggest two distinct phases during preimplantation mouse development – an initial period of DNMT1-mediated maintenance methylation followed by passive demethylation. Finally, we observed that the 5mCpG strand bias distribution at the 32-cell stage is trimodal. Performing k-means clustering on the 5mCpG strand bias in these single cells identified two distinct groups of cells as inferred by the mean silhouette scores – a population with no strand bias and another population with a bimodal strand bias distribution (Fig. 2.4c,d). Further, within the bimodal population, we observed pairs of cells for which all chromosomes were strongly anti-correlated, suggesting that these pairs are sister cells (Fig. 2.4e and Supplementary Fig. 2.9e). These observations reveal the existence of significant cell-to-cell heterogeneity in

the genome-wide methylome landscapes of cells within the early blastocyst. Taken together, these results suggest maintenance methylation is active till the 16-cell stage and that from the 16- to 32-cell stage, a fraction of cells within the embryo show strong 5mCpG strand bias and undergo passive demethylation.

Finally, to conclusively demonstrate that the absence of 5mCpG strand bias up to the 16-cell stage arises from DNMT1 mediated maintenance methylation, we performed bulk hairpin bisulfite sequencing on non-hybrid preimplantation mouse embryos. A hallmark of DNMT1 mediated methylation is that both cytosines in a CpG dyad are symmetrically methylated and therefore we performed bulk hairpin bisulfite sequencing that enables interrogation of the methylation status of CpG dyads[165]. We observed that the fraction of symmetrically methylated CpG dyads in the genome is high up to the 16-cell stage, with a dramatic reduction at the 32-cell stage (that is matched by an increase in hemi-methylated CpG dyads at this stage), thereby demonstrating that maintenance methylation is active initially and is followed by passive demethylation at the 32-cell stage (Fig. 2.4f).

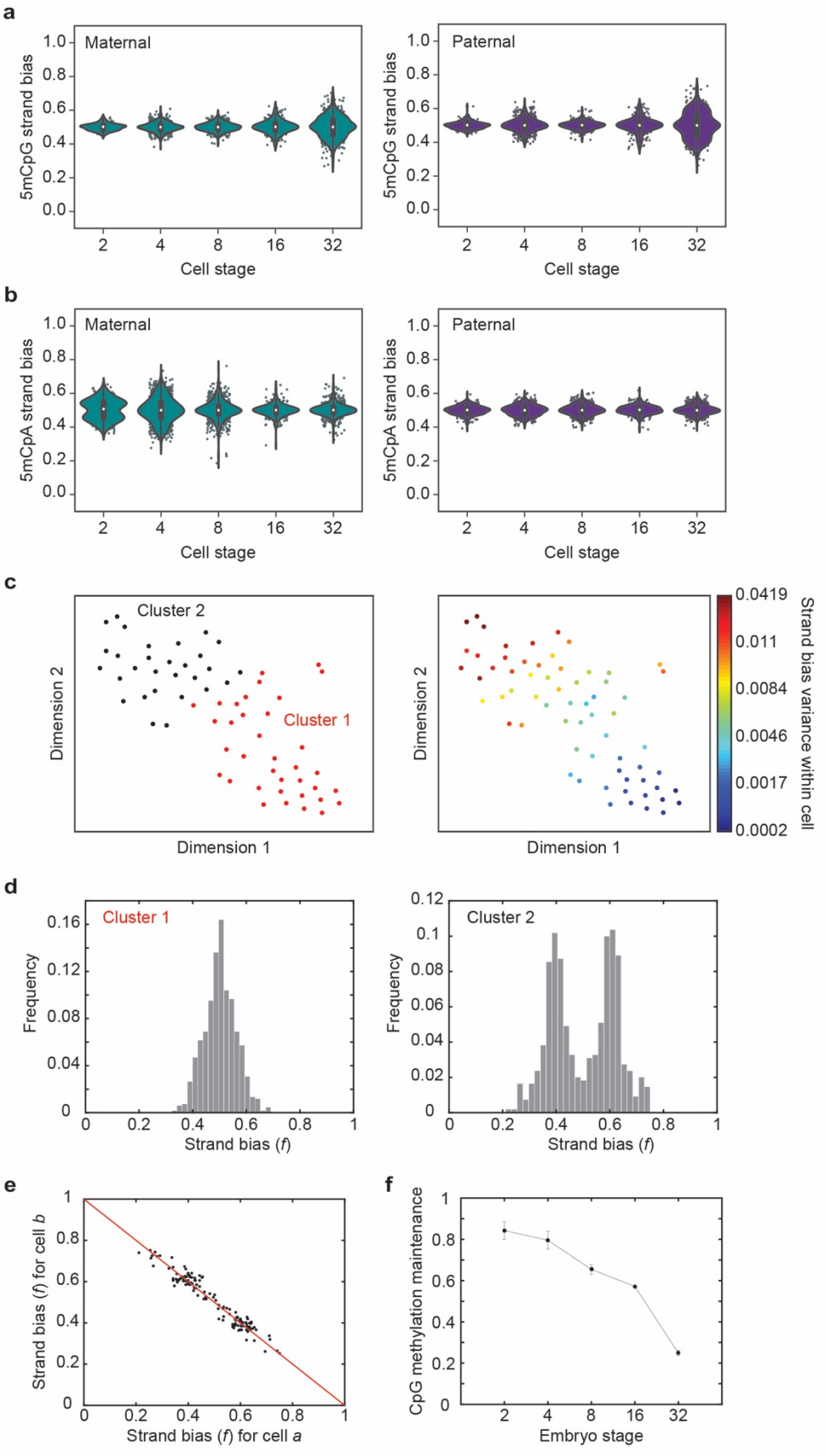

a
Maternal
Paternal
5mCpG strand bias
1.0
0.8
0.6
0.4
0.2
0.0
2
4
8
16
32
Cell stage
b
Maternal
Paternal
5mCpA strand bias
Cell stage
c
Cluster 2
Cluster 1
Dimension 2
Dimension 1
0.0419
0.011
0.0084
0.0046
0.0017
0.0002
Strand bias variance within cell
d
Cluster 1
Cluster 2
Frequency
Strand bias (f)
e
Strand bias (f) for cell b
Strand bias (f) for cell a
f
CpG methylation maintenance
Embryo stage

Figure 2.4 | DNA demethylation dynamics in preimplantation mouse embryos.

(a) Violin plots of 5mCpG strand bias for both the maternal (left) and paternal (right) genome show a tight distribution centered around f = 0.5 till the 16-cell stage and a wider distribution at the 32-cell stage of development (n=332 single cells from 42 embryos). (b) For the maternal genome (left), 5mCpA strand bias show a bimodal distribution at the 2-cell stage that moves towards a tight unimodal distribution by the 32-cell stage of development. The paternal genome (right) shows a unimodal distribution centered at f = 0.5 throughout preimplantation development till the 32-cell stage (n=332 single cells from 42 embryos). In panels a and b, the white dot indicates the median, the black bar indicates the first and third quartile, and the whiskers indicate the minima and maxima. (c) t-SNE map displaying 2 cluster of single cells. These clusters were identified by k-means clustering on the 5mCpG strand bias for all paternal chromosomes (left). The right panel shows the strand bias variance within each cell superimposed on the t-SNE map. (d) The two clusters shown in panel c display dramatically different 5mCpG strand bias distributions – one cluster (left) shows a unimodal distribution while the other cluster (right) shows a bimodal distribution implying loss of methylation maintenance. (e) Strand bias of chromosomes between anti-correlated cell pairs suggesting that these pairs are sister cells. (f) Bulk hairpin bisulfite sequencing reveals that the fraction of CpG dyads that are symmetrically methylated drops substantially from the 16- to 32-cell stage of development (n=2 biologically independent bulk samples). Error bars represent the genome-wide standard deviation from the mean methylation maintenance.

We finally extended scMspJI-seq to explore the dynamics of global demethylation in human preimplantation embryos, ranging from developmental day 2 to 7. Studies in human preimplantation embryos have shown temporally slower, yet similar developmental dynamics to mouse embryos[166]. Despite lacking allelic information, our results suggest that the mouse and human 5mCpG demethylation dynamics are similar, with an initial phase till the 16-cell stage displaying a tight 5mCpG strand bias distribution centered around 0.5, followed by an increase in strand bias in a small fraction of cells from the 32- to 128-cell stage (Fig. 2.5a and Supplementary Fig. 2.10a). This is consistent with previous immunostainings in human preimplantation embryos that show a decrease in DNMT1 protein levels between Day 5 and Day 6 blastocysts[167,168]. Further, 5mCpA strand-bias distributions of human preimplantation embryos appear to be similar to the trend observed in mouse embryos with a majority of cells till the 16-cell stage displaying 5mCpA strand bias (Fig. 2.5b and Supplementary Fig. 2.10b). Finally, upon closer

inspection of 5mCpA strand bias per cell, we observed three sister pairs in Day 3 embryos with a mirrored pattern of strand bias along the entire genome (Supplementary Fig. 2.10c).

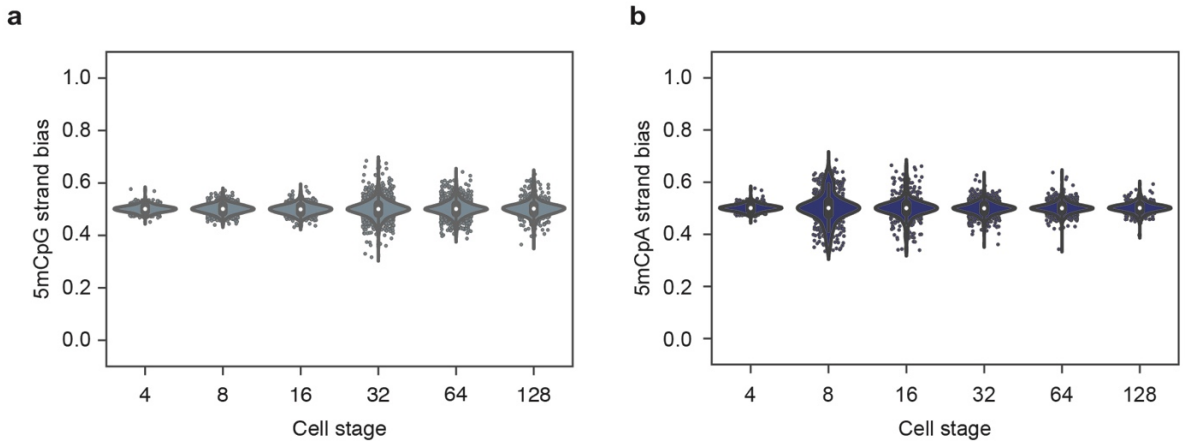


Figure 2.5 | DNA demethylation dynamics in preimplantation human embryos.

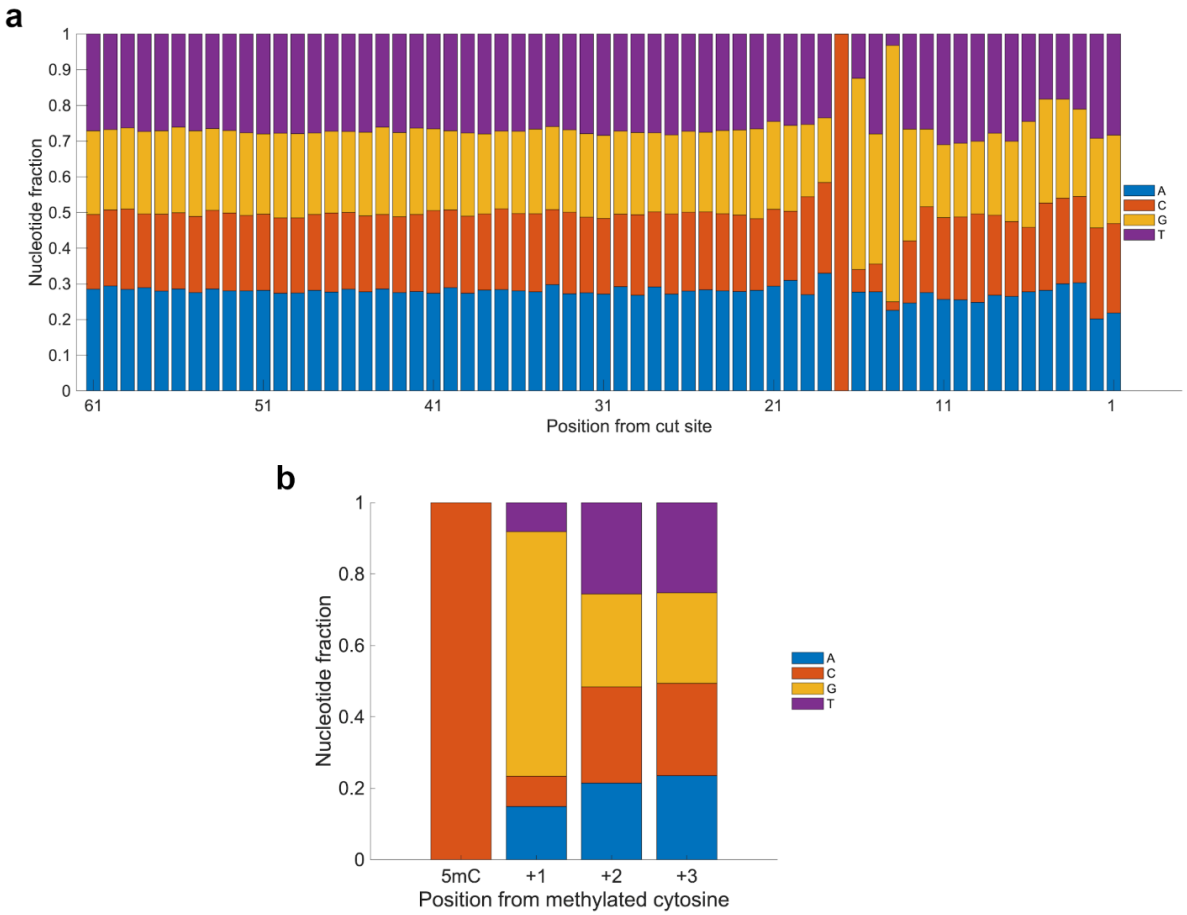
(a) Violin plots showing 5mCpG strand bias from the 4- to 128-cell stage of human embryogenesis. In the absence of allele specific information, the strand bias represents an average over both alleles. Similar to mouse embryos, human embryos initially show no 5mCpG strand bias followed by an increase at the 16-cell stage of embryogenesis. (b) Violin plots showing 5mCpA strand bias from the 4- to 128-cell stage of human embryogenesis. 5mCpA strand bias dynamics in human embryos is similar to that observed in mouse embryos in Figure 2.4b. In these panels, the white dot indicates the median, the black bar indicates the first and third quartile, and the whiskers indicate the minima and maxima.

C. Conclusion

In summary, we have developed a new cost effective and easy to implement strand-specific method that enables us to detect 5mC on a genome-wide scale in single cells. When applied to serum grown mES cells, we found substantial cell-to-cell variability in strand-specific 5mC landscapes, revealing the existence of chromosome-wide heterogeneity in the methylome of mES cells. Reanalysis of a previous single-cell bisulfite sequencing study further confirmed these results[92]. Furthermore, in addition to exploring strand-specific 5mC heterogeneity in single

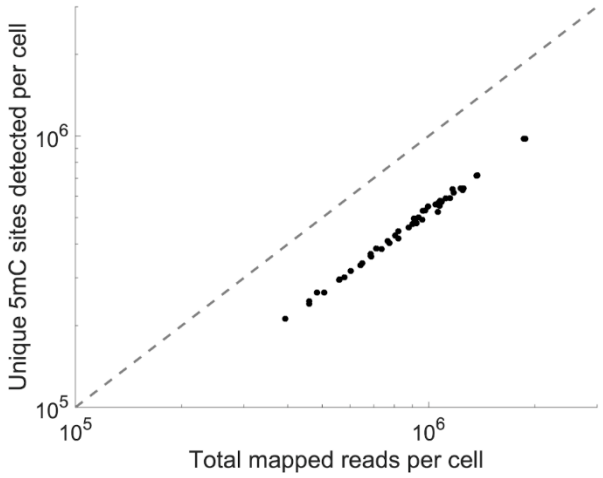
cells, scMspJI-seq also enables systematic investigation of the mechanisms regulating demethylation dynamics. In preimplantation mouse embryos, we surprisingly discovered two distinct phases of methylation dynamics – an initial phase till the 16-cell stage where methylation maintenance is active, followed by loss of maintenance in a fraction of cells within the early blastocyst at the 32-cell stage. These results further highlight the presence of strand-specific 5mC heterogeneity between individual cells during early mammalian development. In the future, we plan to explore how this genome-wide heterogeneity in the methylome regulates lineage commitment during development. Finally, despite the reduced resolution due to lack of allelic information, we found similar demethylation dynamics in preimplantation human embryos. Thus, scMspJI-seq presents a new single-cell strand-specific technology that potentially can be used to probe the dynamics of methylation during development, cancer progression, aging and in other biological systems.

D. Supplementary figures



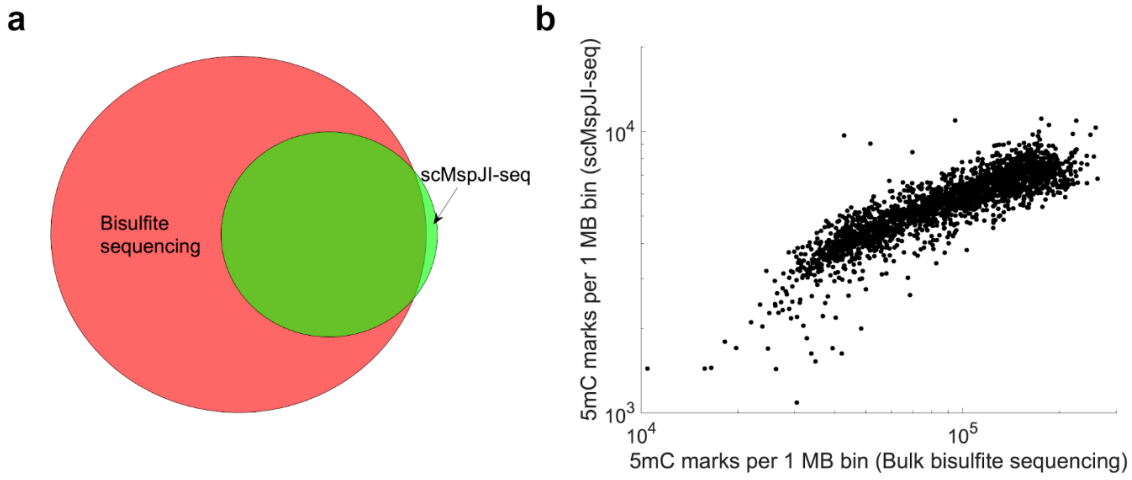
Supplementary Figure 2.1 | Nucleotide composition around methylated cytosine in scMspJI-seq and bisulfite sequencing.

(a) Panel shows the nucleotide composition that was observed downstream of the MspJI cut site. In agreement with previous reports, MspJI was found to cut gDNA 16 bp downstream of the cut site[64]. In scMspJI-seq, an average of 44.0% (with a range of 31.0% to 59.9% in individual cells) of the methylated cytosines were found in a non-CpG context. (b) Analysis of nucleotide composition downstream of the methylated site in previously published single-cell whole-genome bisulfite sequencing[52]. In the bisulfite sequencing data, an average of 31.5% (with a range of 23.4% to 53.1% in individual cells) of the methylated cytosines were found in a non-CpG context. Both scMspJI-seq and bisulfite sequencing data are for E14 mESCs.



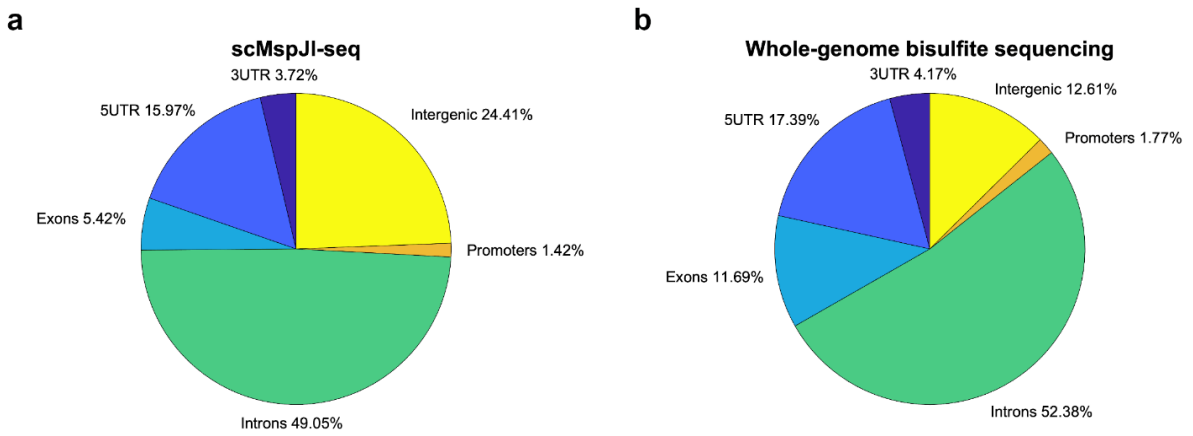
Supplementary Figure 2.2 | Number of unique 5mC sites detected in scMspJI-seq.

The figure shows the number of unique 5mC sites detected per cell as a function of the sequencing depth. The number of unique 5mC sites detected per cell ranged from 212,000 to 977,000, with a median of 484,000 5mC sites per cell. The number of unique 5mC sites detected per cell is increasing monotonically with the sequencing depth, suggesting that more unique sites could be detected per cell by sequencing the Illumina libraries deeper. The figure shows data from E14 mESCs.



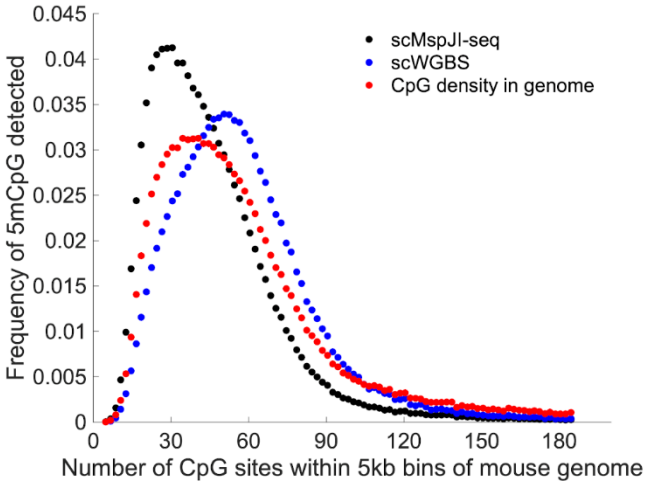
Supplementary Figure 2.3 | Comparison of scMspJI-seq to bulk bisulfite sequencing.

(a) Venn diagram shows that 97.2% of the 5mCpG sites detected in single E14 cells by scMspJI-seq is also found in bulk bisulfite sequencing of E14 gDNA. (b) Number of DNA methylation marks detected within 1 MB bins in scMspJI-seq correlates well with the bulk bisulfite sequencing methylome (Pearson r = 0.84).



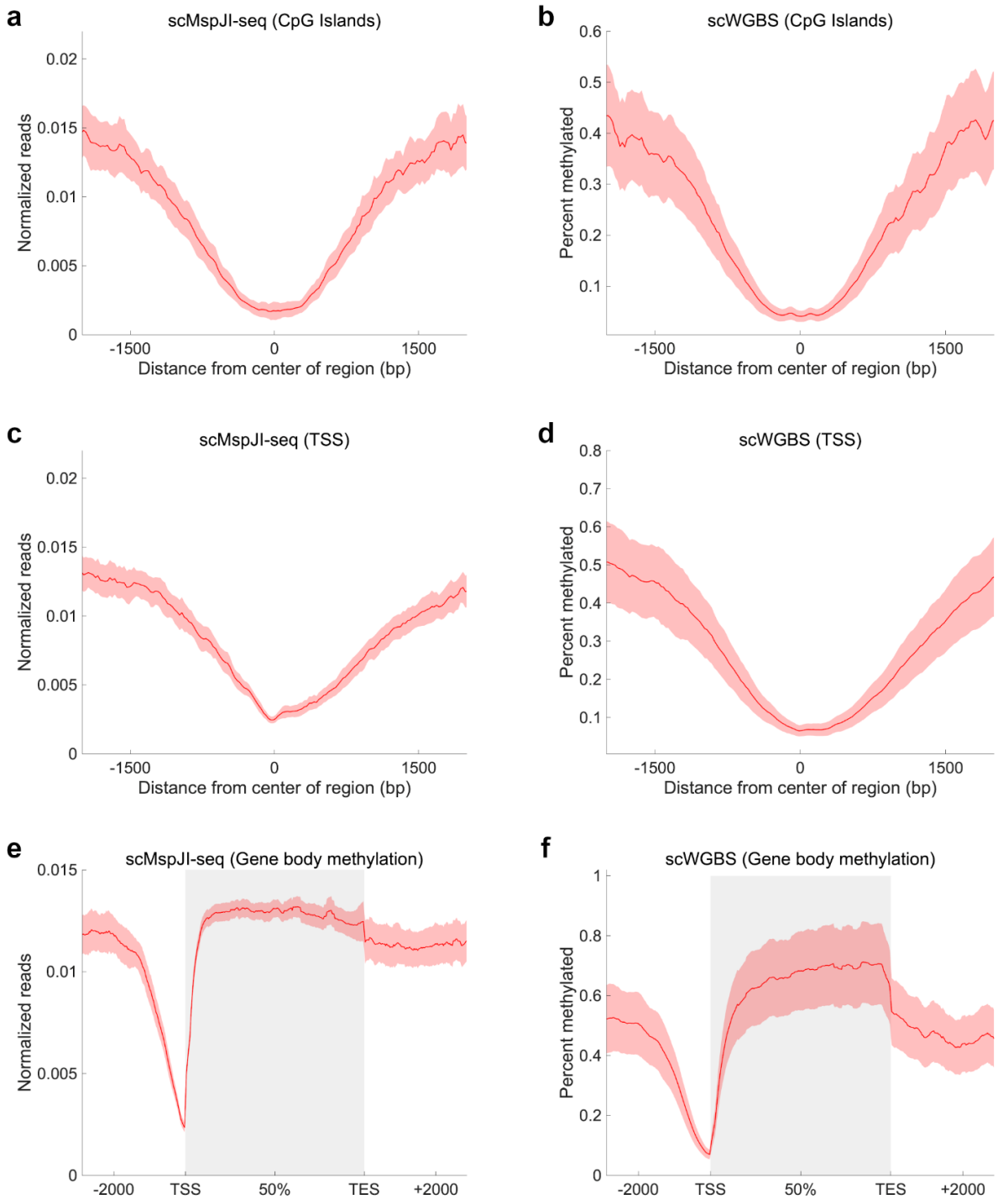
Supplementary Figure 2.4 | Distribution of 5mC over different genomic elements.

(a) Pie chart showing the distribution of 5mC sites over promoters, 5' UTRs, exons, introns, 3' UTRs, and intergenic regions in scMspJI-seq. (b) Pie chart showing the distribution of 5mC sites over the same genomic elements in whole-genome bisulfite sequencing[52]. The figure shows data from E14 mESCs.



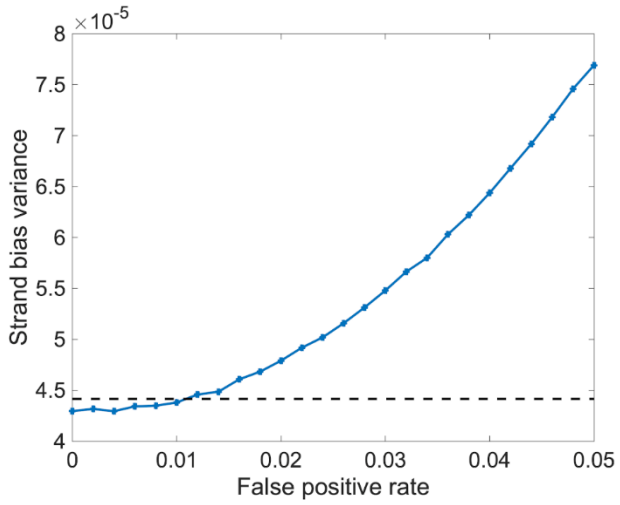
Supplementary Figure 2.5 | Distribution of 5mCpG sites over genomic regions of varying CpG density.

The red curve shows the distribution of CpG sites over 5 kb bins of the mouse genome. The black and blue curves show the distribution of 5mCpG sites that are detected in genomic bins of different CpG densities in scMspJI-seq and scWGBS, respectively[52]. scMspJI-seq is slightly biased towards the detection of 5mCpG within genomic regions that have lower CpG density. The figure shows data from E14 mESCs.



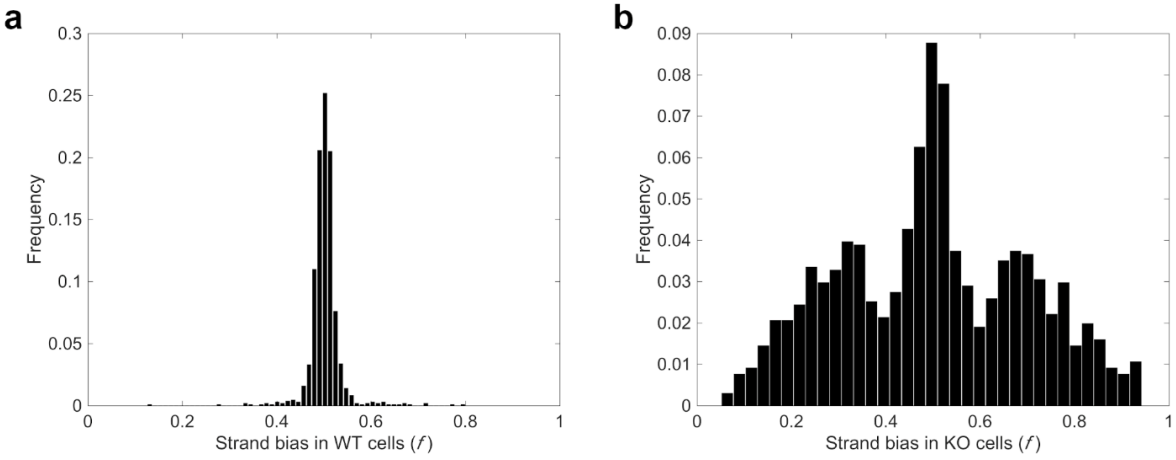
Supplementary Figure 2.6 | Genome-wide DNA methylation landscapes.

(a,b) Panels showing hypomethylation at CpG islands in scMspJI-seq and scWGBS, respectively[52]. (c,d) Panels showing hypomethylation at transcription start sites (TSS) in scMspJI-seq and scWGBS, respectively. (e,f) Gene body DNA methylation profiles obtained from scMspJI-seq and scWGBS, respectively. Shaded red regions indicate standard deviations in the distribution of 5mC. The figure shows data from E14 mESCs.



Supplementary Figure 2.7 | False positive detection rate of 5hmC in scMspJI-seq.

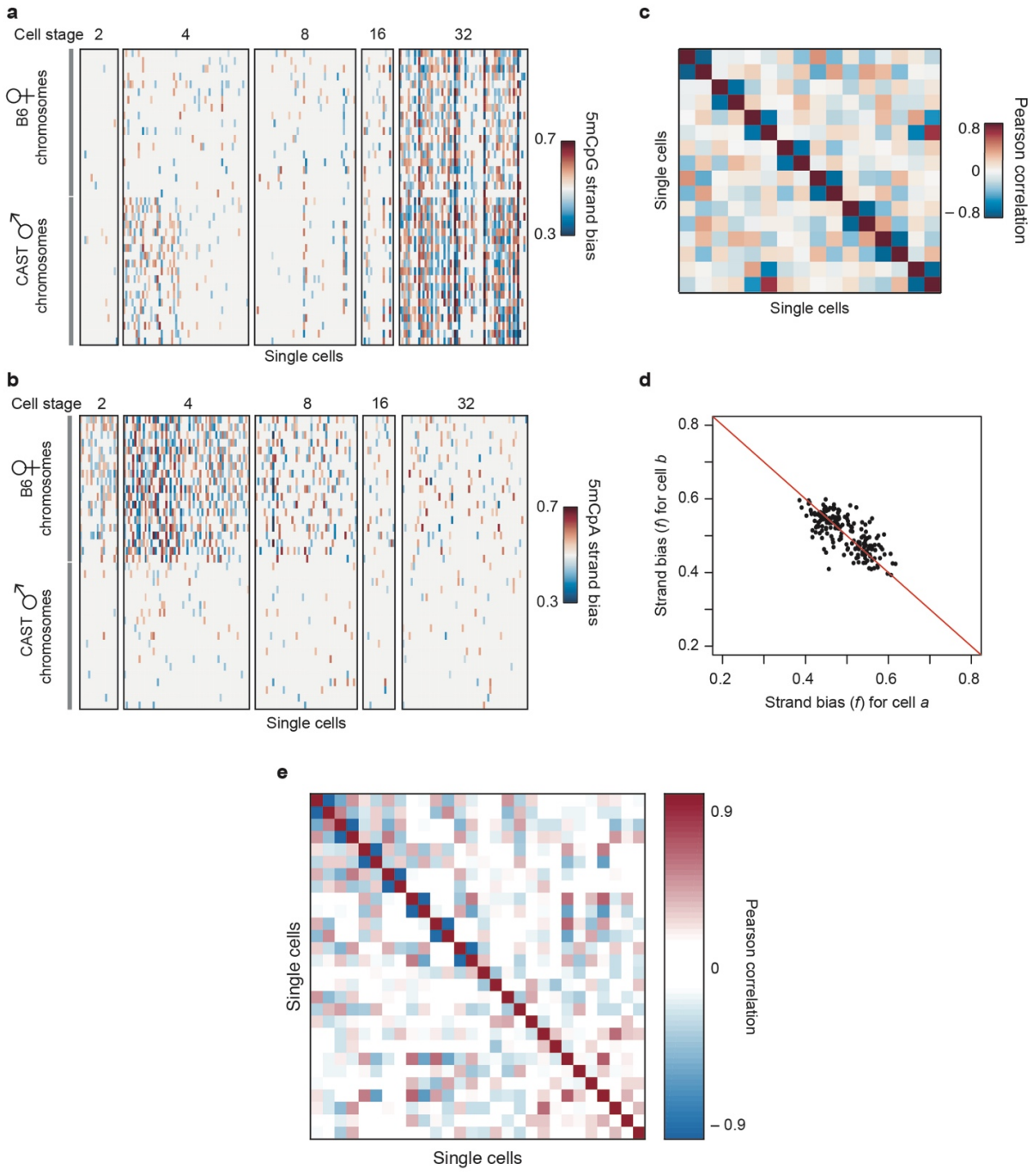
The panel shows that the variance of the simulated strand bias distribution increases with higher rates of 5hmC false positive detection in scMspJI-seq (blue line). The dashed black line indicates the experimental strand bias variance obtained from applying scMspJI-seq to E14 mESCs. Data from scMspJI-seq and scAba-seq were combined to quantify the false positive detection rate of 5hmC in scMspJI-seq[34]. For different efficiencies of 5mC vs. 5hmC detection, a mathematical model was built where 5mC and 5hmC sites were drawn from a binomial distribution and distributed on the two DNA strands of a chromosome using the strand bias distributions from scMspJI-seq and scAba-seq[34]. By comparing the variance of the experimental strand bias distribution to that obtained from the simulations, the false-positive detection rate of 5hmC was estimated to be around 1.1%.



Supplementary Figure 2.8 | Strand-specific detection of 5mC in single cells using scMspJI-seq.

(a) Chromosomes in E14 cells show a tight strand bias distribution centered around 0.5. (b) CRISPR-Cas9 mediated knockout of Dnmt1 in E14 cells results in a dramatic increase in the width of

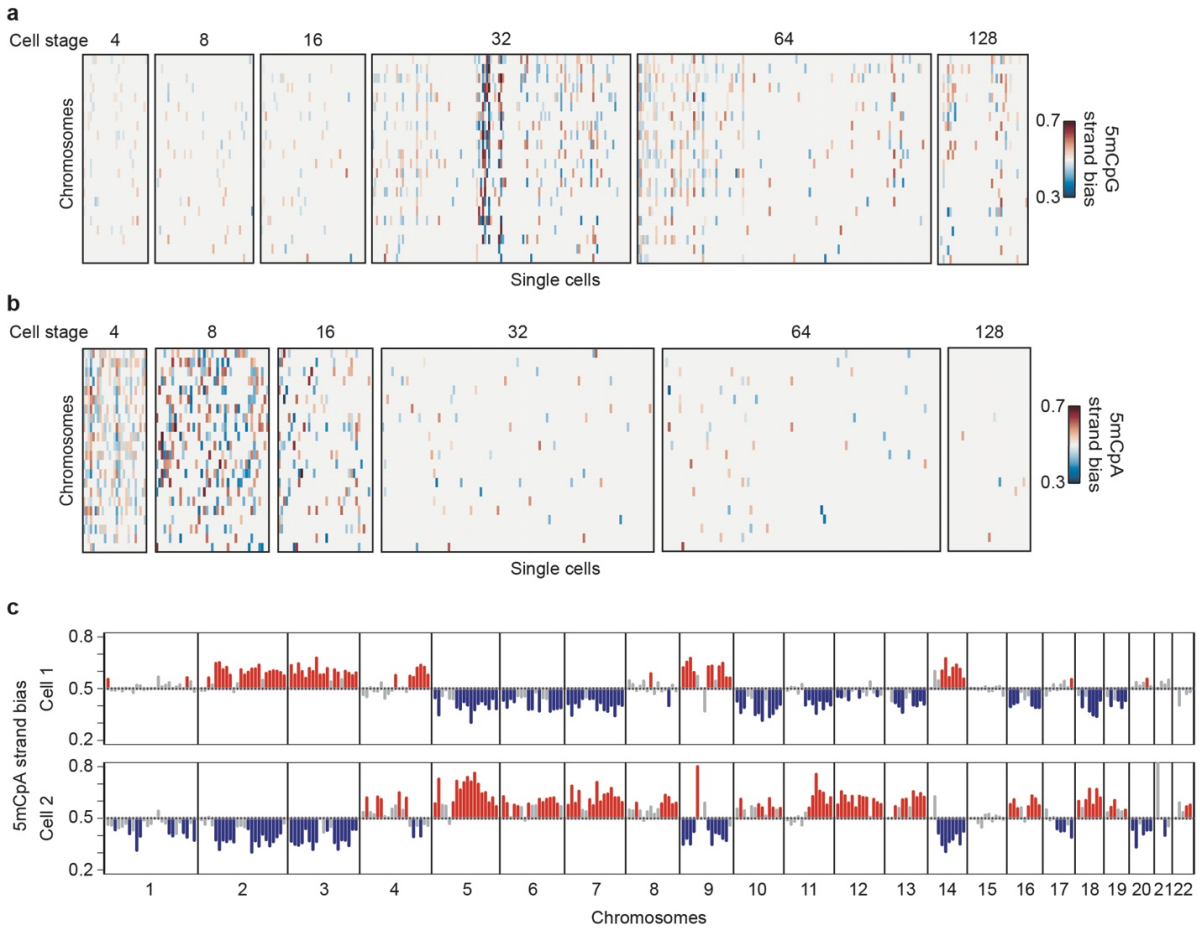
the strand bias distribution indicating loss of maintenance methylation and the ability of scMspJI-seq to quantify strand-specific 5mC in single cells.



Supplementary Figure 2.9 | DNA demethylation dynamics in preimplantation mouse embryos.

(a) Heatmap shows 5mCpG strand bias for all maternal and paternal chromosomes from the 2- to 32-cell stage of development. The data shows a dramatic increase in 5mCpG strand bias from the 16- to 32-cell stage of development. (b) Heatmap shows 5mCpA strand bias for all maternal and

paternal chromosomes from the 2- to 32-cell stage of development. For a majority of cells at the 2- and 4-cell stage, the maternal genome displays 5mCpA strand bias that deviates from 0.5. (c) Heatmap shows Pearson correlation for the maternal 5mCpA strand bias between pairs of cells at the 2-cell stage of development. (d) Pairs of cells in c that display strongly anticorrelated 5mCpA strand bias are shown here, suggesting that we can use this method to identify sister cells at the 2-cell stage of development. (e) Heatmap shows Pearson correlation for the paternal 5mCpG strand bias between pairs of cells (within the bimodal strand bias distribution) at the 32-cell stage of development. Strongly negative Pearson correlations indicate that we can identify sister cells within 32-cell stage embryos.



Supplementary Figure 2.10 | DNA demethylation dynamics in preimplantation human embryos.

(a) Heatmap shows 5mCpG strand bias for all chromosomes from the 4- to 128-cell stage of human development. (b) Heatmap shows 5mCpA strand bias for all chromosomes from the 4- to 128-cell stage of human development. 5mCpA strand bias deviates from 0.5 for a large number of chromosomes till the 16-cell stage. (c) An example of a pair of cells that display strongly anti-correlated 5mCpA strand bias along the entire genome.

3. Integrated single-cell sequencing reveals principles of epigenetic regulation of human gastrulation and germ cell development in a 3D organoid model

A. Introduction

Regulation of gene expression in mammalian systems is tightly regulated by several layers of the epigenome that ensure precise cell type specific programs[97]. Therefore, mapping the genome-wide epigenetic landscape of critical features such as DNA accessibility and DNA methylation (5-methycytosine or 5mC) is central to understanding how these regulatory factors tune complex cellular phenotypes in dynamic systems such as early post-implantation mammalian development. While the role of epigenetic reprogramming in mouse gastrulation has been extensively studied, it remains unclear how reorganization of DNA accessibility and 5mC are coupled with the emergence of cell types during human gastrulation[123]. Further, most current methods for quantifying these epigenetic features rely on the ability to isolate the desired cell type at high purity that is achieved either using cell type specific fluorescent reporters in genetically modified organisms or through access to high-quality cell type specific antibodies. However, the former approach cannot be extrapolated to humans, and many cell types also lack well-defined and unique cell surface markers and/or high-quality antibodies. Furthermore, antibodies can fail to capture transient cell states that do not present the necessary antigen, and thus these approaches are not ideally suited for studying complex systems like early human embryogenesis, that is characterized by a series of rapidly transitioning cell states[169].

B. Results

1. Tri-omic quantification using scMAT-seq in hESCs

To overcome these limitations, we present a single-cell multiomics method scMAT-seq to simultaneously quantify DNA methylation, DNA accessibility, and the transcriptome from the same cell, thereby providing a marker-free approach to map the epigenetic landscape during human gastrulation using an organoid model of development. While two recent methods have been developed to make all three measurements from the same cell, these techniques rely on the physical separation of DNA and RNA prior to amplification, resulting in low throughput and a potential loss of material, limiting its usefulness[92,123,170]. To resolve this, single cells in scMAT-seq are sorted into 384-well plates and the following two steps are performed simultaneously – mRNA is reverse transcribed using a poly-T primer with an overhang containing a cell- and mRNA-specific barcode, a unique molecule identifier (UMI), the 5' Illumina adapter and a T7 promoter; and the methyltransferase M.CviPI is used to methylate cytosines in a GpC context within open chromatin (Fig. 3.1a). Performing these two steps simultaneously is critical to minimize mRNA degradation and to ensure that the *in vivo* state of chromatin can be captured immediately after cell lysis. Next, second strand synthesis is used to generate cDNA, chromatin is stripped off gDNA using proteases, and 5-hydroxymethycytosine (5hmC) sites in the genome are glucosylated to block downstream detection by the restriction enzyme MspJI. Thereafter, MspJI is added, which recognizes methylated cytosines in the genome and creates double-stranded DNA breaks that are ligated to adapters containing a cell- and gDNA-specific barcode, a UMI, the 5' Illumina adapter and a T7 promoter[65]. Following this step, all molecules are tagged with cell- and molecule-

of-origin-specific barcodes, and contain a T7 promoter, allowing samples to be pooled and amplified by *in vitro* transcription (IVT). As described previously, Illumina libraries are then prepared, enabling simultaneous quantification of mRNA, 5mC and DNA accessibility from the same cell without requiring physical separation of the nucleic acids[34,65,171]. Finally, depending on the context of the methylated cytosine in the sequencing data, gDNA reads are either assigned to the methylome (CpG context) or DNA accessibility (GpC context) dataset for each individual cell[92,123,170,172].

To successfully implement scMAT-seq, we optimized two important steps of this method. First, we evaluated different buffer conditions (1x first strand buffer, 1x GC buffer, or 50:50 ratio of each) to ensure that the simultaneous reverse transcription of mRNA and marking of open chromatin by M.CviPI is efficient. While the number of transcripts and the total number of methylated cytosines detected were similar across all buffer conditions (Supplementary Fig. 3.1a-c), we found that the ratio of exogenous to endogenous methylated cytosines was significantly lower in the 1x first strand buffer, indicating inhibition of M.CviPI in this buffer (Supplementary Fig. 3.1d). Therefore, we used a 50:50 buffer ratio for all further experiments (Supplementary Fig. 3.1e). Next, a consequence of making three different measurements from the same cell without physical separation of nucleic acids prior to amplification is that detection of the less abundant type of molecule requires sequencing the libraries to higher depths, thereby increasing the cost of the method. For example, in our initial implementation of scMAT-seq, only 4.2% of molecules in the library derived from mRNA, requiring higher sequencing depths for cell type identification (Supplementary Fig. 3.2a), a limitation we also observed in our

previous single-cell multiomics measurements[154]. To overcome this, we developed an mRNA enrichment protocol after IVT where poly-A mRNA derived amplified RNA molecules were selectively separated from other molecules using biotinylated poly-A primers and streptavidin coated magnetic beads (Fig. 3.1a). To maximize mRNA enrichment, we tested four commercially available beads, and found that C1 and M270 beads provided the best enrichment, with a significant increase in the number of transcripts and genes detected (Supplementary Fig. 3.2b-d). Additionally, we found that library preparation could be performed directly on beads without impacting efficiency, thereby simplifying the enrichment protocol (Supplementary Fig. 3.2e,f). After enrichment, we observed a 9.4 fold increase in mRNA-derived molecules (40.0%), along with a significant increase in gene detection at levels comparable to other single-cell mRNA sequencing techniques (Fig. 3.1b and Supplementary Fig. 3.2a)[171]. Finally, the transcriptome obtained after enrichment was highly correlated (Pearson's r = 0.95) to the non-enriched transcriptome, showing that the mRNA enrichment protocol does not introduce biases in quantifying gene expression (Supplementary Fig. 3.2g).

As proof-of-concept, we first applied scMAT-seq to H9 human embryonic stem cells (hESC). Comparison with previously published DNase I hypersensitivity sites (DHS) showed that, as expected, more accessible sites were associated with larger peaks and greater domain spreading in scMAT-seq (Fig. 3.1c)[173]. Similarly, results from scMAT-seq displayed a highly monotonic relationship with DHS scores (Spearman's p = 0.99), showing that scMAT-seq faithfully reproduces a widely used technique for profiling DNA accessibility (Supplementary Fig. 3.3a). Furthermore, while DNA accessibility profiling is typically performed on fresh samples to maintain

chromatin structure, we tested if scMAT-seq could be extended to frozen samples. When compared to freshly processed cells, we found that sorted cells stored at -80°C produced similar genome-wide profiles of DNA accessibility at DHS and transcription start sites (TSS), demonstrating that scMAT-seq can be applied to a larger spectrum of cryopreserved samples (Supplementary Fig. 3.4a-c). Next, we validated how accurately scMAT-seq captures the methylome of single cells. scMAT-seq builds upon a method we recently developed (scMspJI-seq), where we showed that scMspJI-seq is an alternate approach to single-cell bisulfite sequencing[65]. As with scMspJI-seq, we find that greater than 97% of the 5mC sites detected by scMAT-seq overlapped with published bulk bisulfite sequencing (Supplementary Fig. 3.3b)[174]. In addition, we observed global demethylation at CpG islands (CGI), consistent with hypomethylation at most CGIs within mammalian genomes (Fig. 3.1d)[175]. To verify that scMAT-seq can capture the relationship between DNA accessibility and DNA methylation, and its impact on gene expression, we segregated genes based on their expression level to find that increasing levels of gene expression are associated with more open chromatin and reduced DNA methylation at TSSs (Fig. 3.1e). Interestingly, when compared directly at the single-cell level, we observed higher correlations between gene expression and DNA accessibility that is computed over both the promoter and gene body instead of just the promoter alone (Fig. 3.1f). Further, as seen previously by Clark *et al.*, the pseudo-bulk correlation between gene expression and DNA accessibility in the promoter region is higher than the average of individual cells (Pearson's *r* of 0.80 and 0.28, respectively), likely due to the small size of the promoter region which limits the detection of reads in these regions in single cells[92]. Together, these results

demonstrate that scMAT-seq can be used to quantify DNA accessibility, DNA methylation and the transcriptome from the same cell. Finally, to demonstrate that scMAT-seq can be used to identify distinct cell types and construct cell type-specific epigenetic landscapes, we performed scMAT-seq on HEK293T cells. As expected, RNA expression data from scMAT-seq could be used to easily distinguish between the two cell lines H9 and HEK293T (Fig. 3.1g). Similarly, the cell lines could be segregated by the first principal component for both DNA accessibility and DNA methylation, suggesting that scMAT-seq can successfully capture cell type-specific epigenetic profiles (Fig. 3.1h,i).

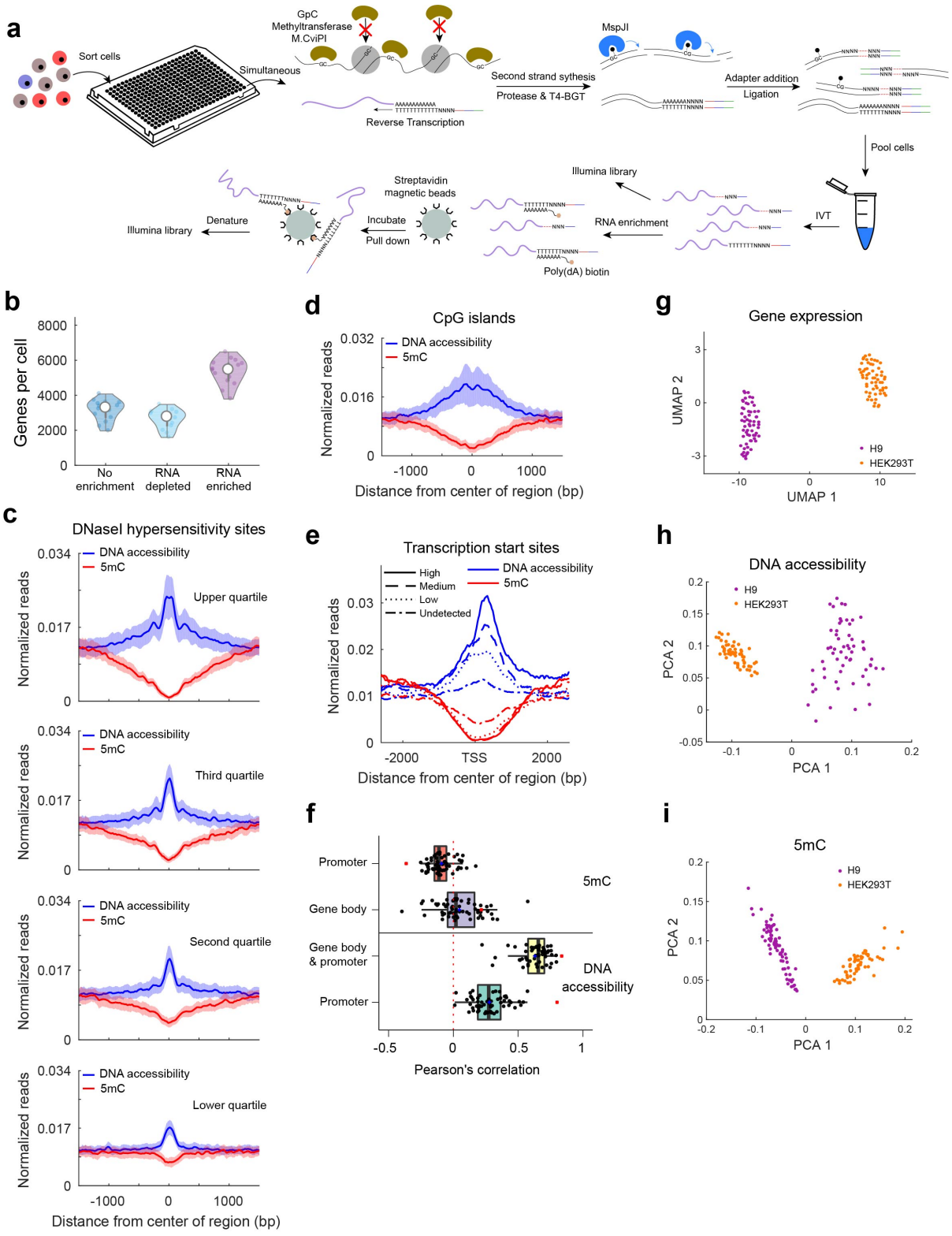


Figure 3.1 | Joint profiling of DNA methylation, DNA accessibility, and the transcriptome from the same cell using scMAT-seq.

(**a**) Workflow of scMAT-seq, where the first step involves simultaneous reverse transcription of mRNA and marking of open chromatin with M.CviPI. After second strand synthesis, protease and T4-BGT treatment, methylated cytosines in the genome are digested using MspJI. The barcoded cDNA and gDNA molecules are then pooled and amplified using IVT. This is followed by mRNA enrichment and Illumina library preparation. In the schematic, RNA is shown in purple, DNA in black, cell- and mRNA-/gDNA-specific barcodes in red, Illumina read 1 sequencing primer in blue, and T7 promoter in green. (**b**) Violin plot shows the number of genes detected per cell with or without RNA enrichment. Dots represent individual cells. (**c**,**d**) Averaged single-cell DNA accessibility (blue) and DNA methylation (red) profiles at DNase I hypersensitivity sites, split by previously reported signal strength (c), or CpG islands (d)[173]. Shaded area indicates the standard deviation across single cells. (**e**) Averaged single-cell DNA accessibility (blue) and DNA methylation (red) profiles at TSS, segregated by gene expression levels: High (solid), medium (dashed), low (dotted), or undetected (dash dot). (**f**) Boxplot of Pearson's correlation between gene expression and indicated epigenetic features for individual cells (black) at *cis* regulatory elements. Blue dot indicates averaged single-cell correlation, and red square indicates pseudo-bulk correlation. Data in panels (b)-(f) are obtained by applying scMAT-seq to individual H9 hESCs. (**g**, **h**, **i**) H9 (purple) and HEK293T (orange) cells can be separated based on their transcriptome (g), DNA accessibility (h), or DNA methylation (i).

2. The epigenetic landscape of major cell types corresponding to the germ layers and primordial germ cell-like cells

Next, we applied scMAT-seq to study early events in post-implantation human development. While epigenetic regulation of gene expression during post-implantation development and gastrulation have recently been characterized using a similar multiomics single-cell technique in mice, similar studies in human embryos have not been possible despite their fundamental importance to human health and disorders[123]. Further, morphological divergence from mice makes it challenging to directly extrapolate the emergence and maturation of different cell types during this period to human development. Therefore, to study human embryogenesis, we used an organoid model that mimics post-implantation amniotic sacs that can induce early germ layer lineages, similar to previous systems (Fig. 3.2a)[126–132]. As expected, analysis of the transcriptome from scMAT-seq identified epiblast-like cells (EPILC) expressing pluripotency markers POU5F1, NANOG, SOX2 and DPPA4, as well as amniotic ectoderm-like cells (AMLC) expressing TFAP2A, GATA3, HAND1, BMP4,

and CDX2 (Fig. 3.2b,c). Further, EPILCs in these asymmetric cysts have previously been shown to undergo epithelial to mesenchymal transition, mimicking gastrulation and presenting features of posterior primitive streak-/mesoderm-like cells (PPSLC)[126,130]. In agreement with those observations, we found cells expressing markers of the posterior primitive steak, including T, EOMES, LHX1, MESP1, MESP2, GATA6, LEF1, SNAI1 and SNAI2, suggesting that PPSLCs are also derived in our system (Fig. 3.2b,c). Unlike the other organoid models previously developed, we unexpectedly found a small number of anterior primitive streak-/endoderm-like cells (APSLC) within the same organoid system, expressing high levels of NODAL, FOXA2, SOX17, OTX2, CXCR4, and even expressing genes related to the organizer cell fate including GSC and signaling inhibitors DKK1 and CER1 (Fig. 3.2b,c). The emergence of these APSLCs indicates an area of relatively high Activin/Nodal signaling and low BMP signaling within the organoid[176]. In human embryos, in addition to its role at the anterior primitive streak, SOX17 is a critical regulator of primordial germ cells (PGC) specification and unlike in mice, it is upregulated before the commonly used PGC marker PRDM1 (also known as BLIMP1)[125]. Interestingly, a subset of cells were found to express both SOX17 and PRDM1, together with other PGC markers, including NANOG, POU5F1, NANOS3, and TFAP2C (also known as AP2-gamma), indicating the emergence of human primordial germ cell-like cells (hPGCLC) in our system (Fig. 3.2b,c).

In addition to discovering cell types corresponding to the early germ layer lineages and hPGCLCs, we discovered an unexpected population of cells with high expression of SOX2, PAX6 and PAX3, reminiscent of neuroectoderm cells, which typically arise later in development (Supplementary Fig. 3.5a,b)[177]. To investigate

the origin of this population of cells, we performed a simpler experiment to test the differentiation potential of the starting induced pluripotent stem cells (iPSCs) by treating them with the GSK-3 inhibitor CHIR99021 for 24 hours (+CHIR), which is known to activate the WNT pathway, causing differentiation towards the mesodermal lineage in iPSCs[178]. iPSCs with or without CHIR treatment were probed with scMAT-seq, and clustering based on the transcriptome revealed 3 groups of cells – a pluripotent population within untreated cells expressing SOX2, POU5F1, NANOG, and DNMT3B, a mesodermal CHIR-treated population expressing T, EOMES, and NODAL, and a neuroectoderm-like cell (NELC) population present in both CHIR treated and untreated cells, expressing SOX2, PAX6, and PAX3, similar to that observed in the organoid (Supplementary Fig. 3.5c,d). These results indicate heterogeneity within the iPSC population, with a subset of cells that are biased towards the neuroectoderm lineage. Further, as this pre-existing population of NELCs was not responsive to CHIR treatment and displayed limited differentiation potential, it was removed from downstream analysis, highlighting the benefits of performing single-cell measurements on complex biological systems.

We next focused attention on the epigenomes of the different cell types identified in the gastruloid. Together with the rapid emergence of different cell types within a 48-hour window in the organoid, we found that the DNA accessibility and 5mC landscapes were reprogrammed, with distinct profiles for each cell type. Further, both genome-wide epigenetic features could be independently used to accurately cluster cells by cell type (Fig. 3.2d,e). Further, when comparing differentially expressed genes (DEGs) between different cell types, higher gene expression was generally associated with higher DNA accessibility (Fig. 3.2f and Supplementary Fig.

3.6a,b). In contrast, changes in gene expression were significantly correlated to changes in gene body DNA methylation in only a subset of cell types (Fig. 3.2f and Supplementary Fig. 3.6c). Together, these results indicate that reprogramming of DNA accessibility is a major driver of changes in gene expression and the emergence of distinct cell types during *in vitro* gastrulation, whereas gene body methylation plays a more limited role in tuning gene expression by regulating select genes and cell types. To gain better understanding of the dynamics of DNA methylation turnover during gastrulation, we observed that the expression level of the *de novo* DNA methyltransferase DNMT3B was lower in more differentiated cell types, especially PPSLC, PGCLC and AMLC, compared to EPILC, consistent with previous observations that DNMT3B is associated with pluripotency and is responsible for the high levels of non-CpG methylation in hESCs that drops substantially during differentiation (Fig. 3.2c and Supplementary Fig. 3.7a)[164]. In agreement with this, we observed that the levels of 5mCpA relative to 5mCpG dropped for PPSLC, PGCLC and AMLC compared to EPILC (Supplementary Fig. 3.7b). Further, similar to our recent work, scMAT-seq enables strand-specific quantification of DNA methylation, which provides insights into DNA methylation dynamics as increasing levels of asymmetric 5mCpA between the two strands of DNA implies a reduction in the rates of *de novo* methylation[65]. We found that PPSLC, PGCLC and AMLC showed greater strand-specific asymmetry in 5mCpA compared to EPILC, consistent with the expression levels of DNMT3B, suggesting that *in vitro* gastrulation is marked by a drop in genome-wide non-CpG methylation and pluripotency, arising from a global reduction in the rates of *de novo* methylation (Fig. 3.2g and Supplementary Fig. 3.7c).

While changes in DNA accessibility and gene expression were generally correlated across all cell types, fold changes in the expression of individual DEGs did not always result in proportional fold changes in DNA accessibility (Fig. 3.2f and Supplementary Fig. 3.6a,b). To investigate this further, changes in DNA accessibility of DEGs were split into three groups, and as expected, larger increases in DNA accessibility were on average associated with larger increases in gene expression (Fig. 3.2h). However, interestingly we observed outlier genes in all groups that showed fold changes in expression that were significantly higher than the mean, with the group containing genes with the most open chromatin displaying a longer tail of highly DEGs compared to the other two groups. These results suggest that for a subset of genes, other epigenetic features potentially drive changes in gene expression that do not directly alter DNA accessibility within the promoter and gene body, highlighting that the combinatorial action of epigenetic regulators can non-additively tune gene expression.

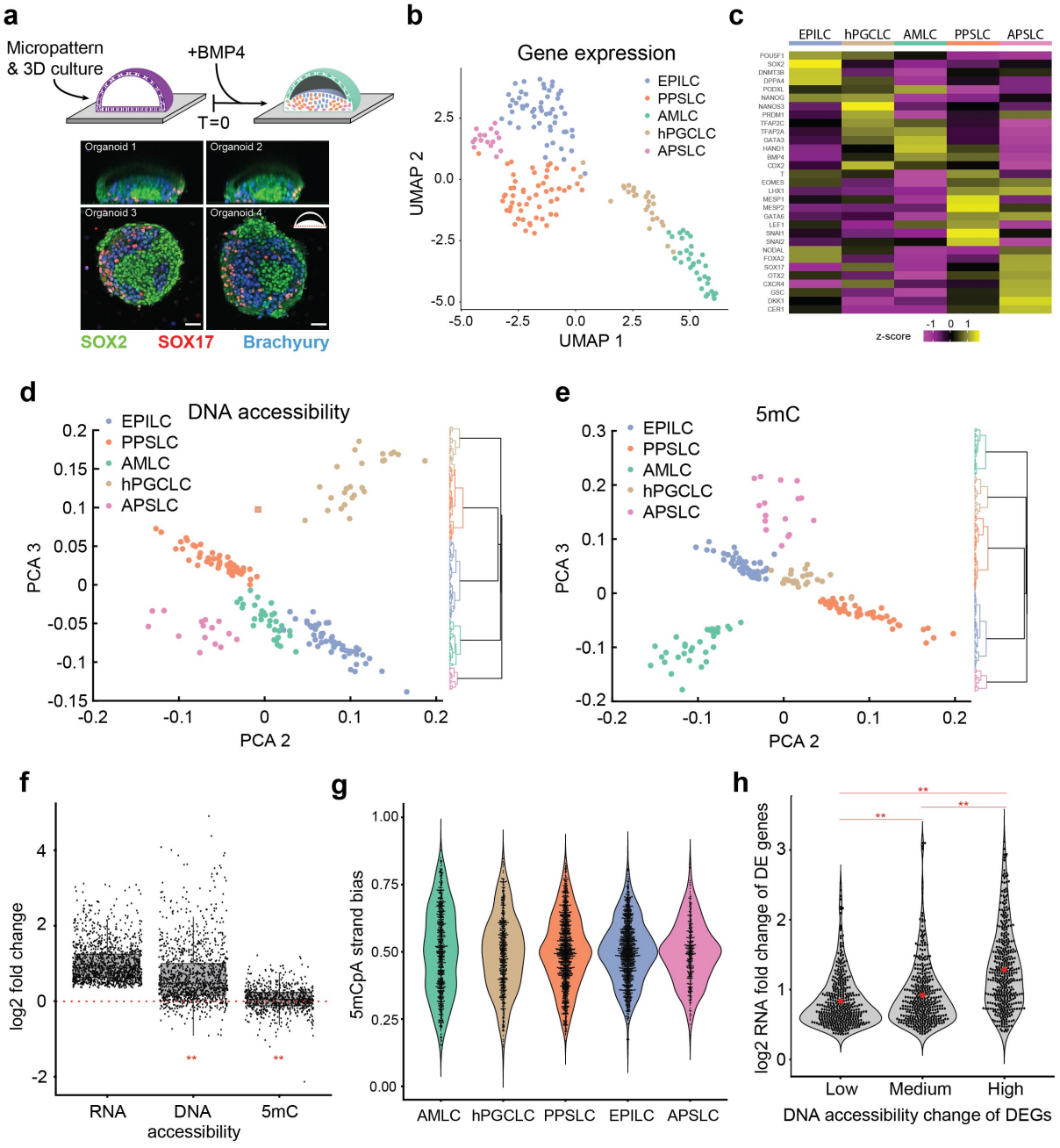


Figure 3.2 | scMAT-seq maps the epigenome and transcriptome of cell types during human gastrulation using a 3D organoid model.

(**a**) Schematic for the *in vitro* generation of 3D post-implantation amniotic sac organoids. Representative immunofluorescence staining of cross sections of the organoids 48 hours post BMP4 addition is shown for SOX2 (green), SOX17 (red), and Brachyury (blue). (**b**) UMAP visualization of the gastruloid based on the single-cell transcriptomes obtained from scMAT-seq. Cell types are assigned to clusters based on established marker genes. (**c**) Heatmap of z-scores for expression of marker genes for different cell types identified in the gastruloid. (**d**,**e**) Principal component projections show that cell types identified in the human gastruloids can also be distinguished using their DNA accessibility landscapes (d) or DNA methylation landscapes (e). Panels also show the corresponding dendrograms. A square border around dots indicate cell type identification by the transcriptome that

differs from the epigenetic feature. (**f**) Boxplot of single-cell averaged log2 fold change in gene expression levels, promoter and gene body DNA accessibility levels, and gene body DNA methylation levels, for DEGs between two cell types. Dots indicate individual DEGs. Note that the log2 fold change in RNA is computed by taking the ratio of expression in a cell type where the gene is expressed at a higher level than in the other cell type. The log2 fold change in DNA accessibility and DNA methylation is shown for the same corresponding pair of cell types. ** indicates a statistically significant change ($p < 0.01$, based on bootstrapped distributions from non-differentially expressed genes) in the log2 fold change of the epigenetic features for DEGs relative to non-differentially expressed genes. (**g**) Violin plot of non-CpG methylation strand bias for each cell type. Strand bias is defined as the ratio of non-CpG methylated sites detected on the plus strand of a chromosome over all non-CpG methylated sites detected on a chromosome. Dots represent individual chromosomes in single cells. (**h**) Violin plot of log2 fold change in gene expression for DEGs partitioned by the relative change in DNA accessibility. Black dots indicate individual DEGs. Red dot indicates the mean of the distribution. ** indicates a statistically significant change ($p < 0.01$, two-sided Mann-Whitney U test) between the distributions.

3. Time resolved epigenetic reprograming during gastrulation and PGCLC development

Finally, we focused attention on the population of hPGCLCs that we identified 48 hours post BMP4 addition. Due to a lack of availability of human embryos at these stages of development, the precursors that specify human primordial germ cells (hPGCs) *in vivo* remain unclear[108,179]. In mice, it is well-established that mPGCs emerge from epiblast cells, and similar results have been found in the developing porcine embryo; however, in the nonhuman primate cynomolgus macaques, PGCs have been found to arise from the extra-embryonic dorsal amnion[180,181]. Further, there is significant divergence between human and mouse in the transcription factor network that is responsible for commitment towards the germ cell lineage, and therefore, the identity of the progenitors that specify hPGCs in humans remain unknown[108,179]. The hPGCLCs that we identified 48 hours post BMP4 addition expressed well-established markers of hPGCs, including SOX17, TFAP2C, PRDM1, NANOG, POU5F1, and NANOS3; however, they were found to not display any genome-wide erasure of DNA methylation, a characteristic feature of PGC

maturation, suggesting these cells were pre-migratory PGCs that had recently undergone specification (Fig. 3.2c and Supplementary Fig. 3.7c)[108,125]. Therefore, to investigate the early events involved in germ cell specification, we systematically characterized younger organoids 20 and 36 hours post BMP4 addition. Using RNA expression data and Monocle 3, individual cells sequenced from the organoid were assigned a pseudotime to indicate the position of a cell along a differentiation trajectory (Fig. 3.3a)[182]. In general, we found the pseudotime assignment to correspond well with organoid age, providing confidence that it accurately described progress along a differentiation lineage (Fig. 3.3b). Similarly, as expected, one of the trajectories from EPILCs was found to bifurcate into two paths, developing into either PPSLCs or APSLCs (Fig. 3.3a). Strikingly, we found that the hPGCLCs and AMLCs bifurcate from a common progenitor population, which suggested the emergence of a precursor within 20 hours after treatment with BMP4. Compared to the EPILCs, we discovered that this progenitor population expressed genes related to the amnion (TFAP2A, GATA3 and CDX2) as well as gastrulating cells (EOMES and T), and downstream targets of BMP4 signaling (BAMBI, ID1-3 and MSX2) (Fig. 3.3c, Supplementary Fig. 3.8a and Supplementary Fig. 3.9). Notably, this was consistent with another recent observation in a disorganized 3D aggregates based system where hPGCLCs were found to emerge from a TFAP2A+ progenitor population[183]. Comparison with this system showed that the progenitors in our organoids are closest to day 1 cells post BMP4 addition in the disorganized aggregates, while hPGCLCs in our system are closest to day 2 cells in the disorganized aggregates (Fig. 3.3d). Further, we investigated DEGs characterizing the mesoderm, amnion and germ cells to find that these genes were more accessible and contained higher

levels of gene body methylation in the progenitor population compared to EPILCs, suggesting that the progenitors are primed towards conversion to AMLCs and PGCLCs (Supplementary Fig. 3.8b-d). Overall, these results suggest that the progenitor population first emerges from EPILCs within 20 hours of BMP4 addition, with transient characteristics of both amniotic- and mesoderm-like cells, before getting specified towards hPGCLCs.

Finally, to systematically map changes in the epigenome along all differentiation trajectories during gastrulation, we grouped genes that varied throughout pseudotime into gene modules based on their expression similarity. In all trajectories, we found that changes in promoter and gene body DNA accessibility closely varied with changes in RNA trajectory (Fig. 3.3e and Supplementary Fig. 3.10). In contrast, gene body DNA methylation was associated to a lesser extent with the transcriptome (Fig. 3.3e and Supplementary Fig. 3.10). To quantify this relationship across all trajectories, gene modules were split into two groups – one that is downregulated (Off) and the other that is upregulated (On) through differentiation. We found that reprogramming of DNA accessibility was highly correlated to changes in gene expression for both genes turning Off and On (Fig. 3.3f). Surprisingly however, while gene body DNA methylation changes were generally well correlated to expression changes for genes that turn Off across all differentiation trajectories, we found a wide range of correlations, depending on the trajectory, for genes that turn On (Fig. 3.3f). Similarly, direct comparison of DNA accessibility and DNA methylation showed high correlation for genes that are downregulated while the correlation was lower, and trajectory dependent, for genes that are upregulated during differentiation (Fig. 3.3f). Previous work has shown that

more open chromatin in highly expressed genes is linked with greater access for the *de novo* methylation machinery, resulting in higher gene body methylation[184]. We hypothesize that the high correlation between DNA accessibility and gene body DNA methylation for genes that are turning Off is possibly due to these regions being inaccessible and therefore, independent of the expression level of the *de novo* DNA methyltransferases. In contrast, for genes that are turning On, we observe a wide range of trajectory-dependent correlations, with differentiation towards AMLCs displaying lower correlation than APSLC and PPSLC lineages, possibly due to the lower expression of DNMT3B resulting in reduced *de novo* methylation activity within gene bodies of AMLCs (Fig. 3.3f, 3.2c and Supplementary Fig. 3.7a). Together, this suggests that increasing DNA accessibility and not gene body methylation is required for gene activation, and that while high DNA accessibility in gene bodies could lead to high DNA methylation, this correlation is decoupled by low *de novo* methyltransferase activity.

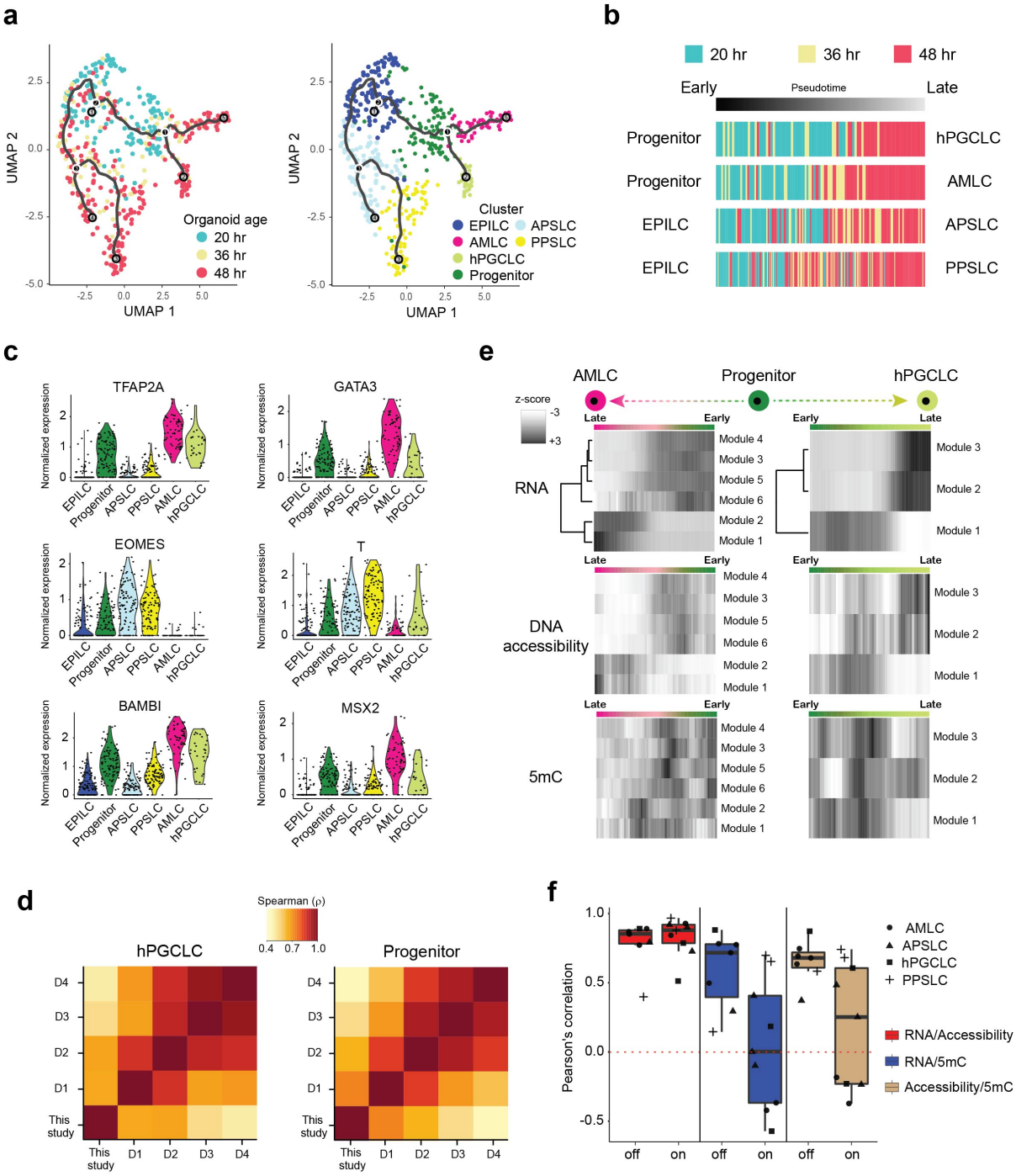


Figure 3.3 | scMAT-seq characterization of time course experiment on gastruloids reveals the identity of hPGCLC progenitors.

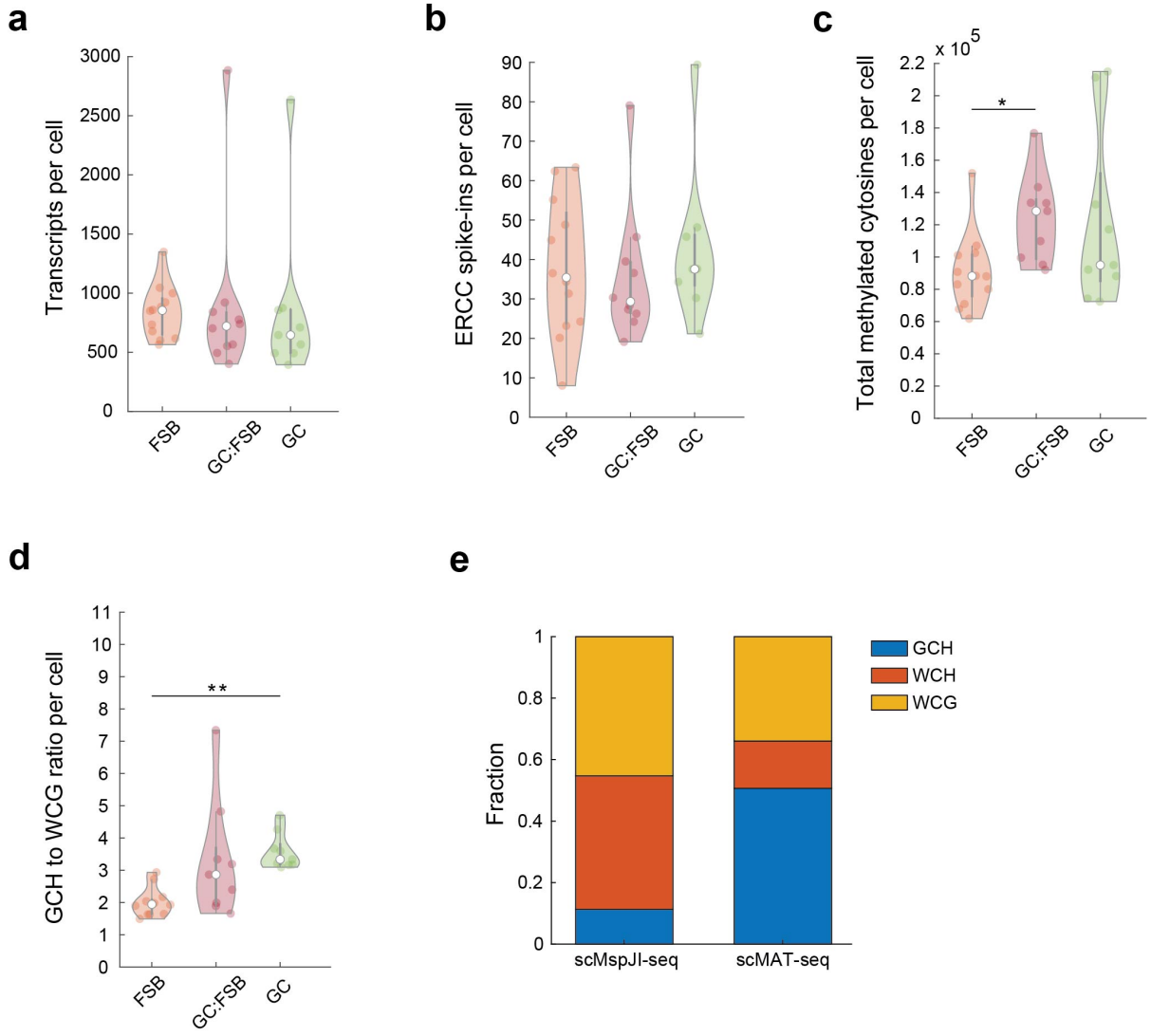
(**a**) (*Left*) UMAP visualization of the single-cell transcriptomes of the gastruloid at different timepoints, 20 hours (blue), 36 hours (tan), and 48 hours post BMP4 addition (ruby). The solid black line shows the inferred differentiation trajectory of cells from Monocle 3. (Right) Assignment of cell types to clusters identified in the time course experiment. (**b**) Relationship between predicted pseudotime trajectory and gastruloid age for different differentiation lineages. (**c**) Violin plots for select genes, highlighting that the progenitor population expresses markers of the amnion (TFAP2A and

GATA3), gastrulating cell fate (EOMES and T), and downstream targets of BMP4 signaling (BAMBI and MSX2). (**d**) Heatmap of Spearman's correlation coefficient comparing gene expression of hPGCLCs and progenitors identified in this study to that found in disorganized aggregates previously described by *Chen et al.*[183]. D1 to D4 indicates day 1 to day 4 post BMP4 addition in the disorganized aggregates. (**e**) Heatmap of z-scores for DEG-derived gene modules along the AMLC and hPGCLC pseudotime trajectory, together with the corresponding changes in DNA accessibility and DNA methylation. The color bar indicates position along the pseudotime, with the progenitor in green, AMLC in pink, and hPGCLC in light green. (**f**) Boxplot of Pearson's correlation between the indicated variables along pseudotime with gene modules being divided into two groups – those that are upregulated (On) or those that are downregulated (Off) with pseudotime.

C. Conclusion

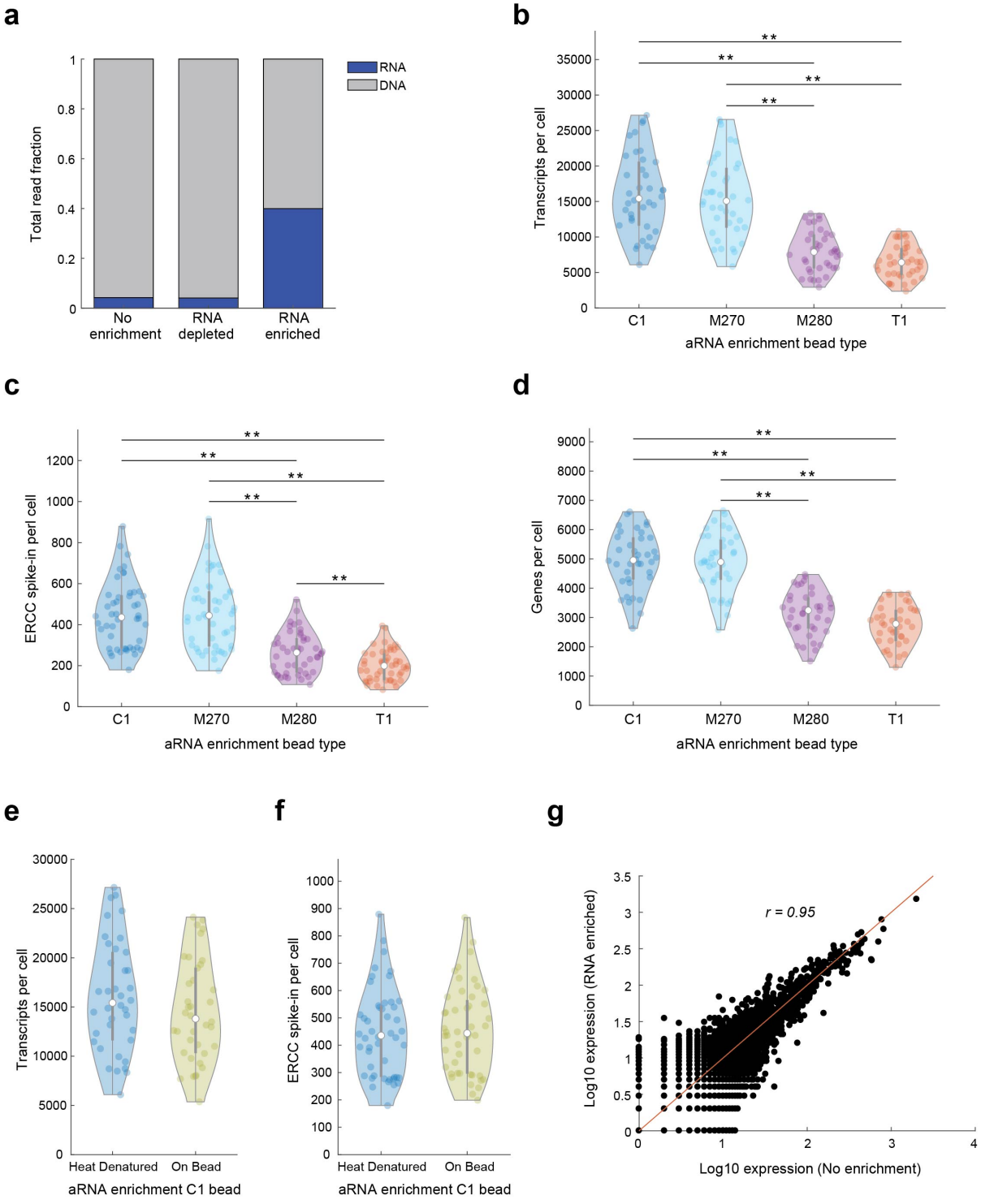
In this report, we developed a multiomics single-cell method scMAT-seq to simultaneously quantify DNA accessibility, DNA methylation and mRNA from the same cell without requiring physical separation of the nucleic acids prior to amplification, enabling efficient and high-throughput mapping of cell types and their corresponding epigenetic profiles *in silico*. When applied to human gastruloids, we show that both the transcriptome and the epigenetic features can be used to identify different cell types. Notably, we discovered that the progenitor population that gives rise to hPGCLCs emerge from EPILCs with transient epigenetic and transcriptional signatures of both amnion- and mesoderm-like cells. In summary, scMAT-seq provides an integrative approach to investigate the role of epigenetic features in regulating gene expression and cell fate decisions in complex and dynamic biological systems.

D. Supplementary figures



Supplementary Figure 3.1 | Optimization of buffer conditions in scMAT-seq.

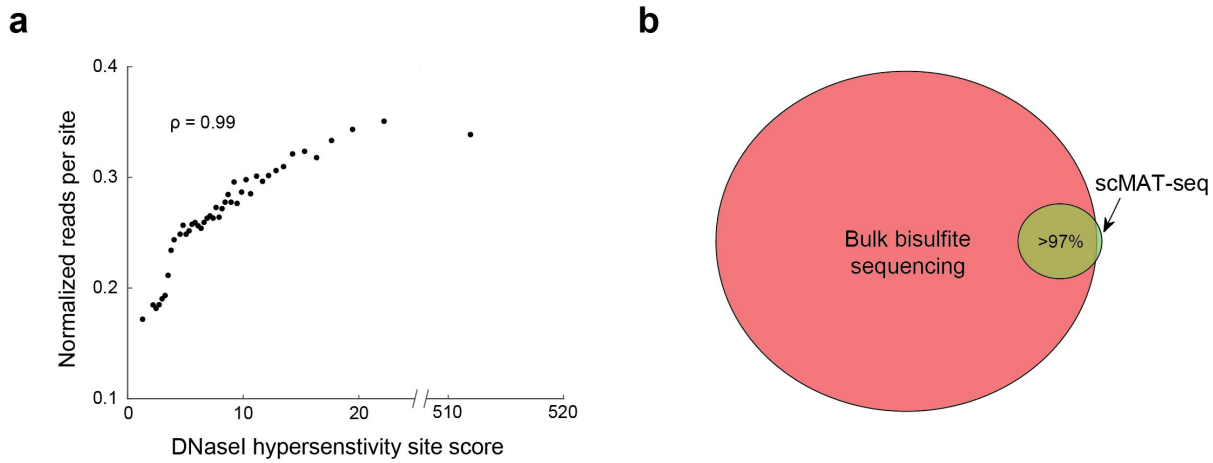
(**a**) Violin plot of the number of transcripts detected per cell in each buffer condition. (**b**) Violin plot of the number of synthetic ERCC RNA spike-in molecules detected in individual cells for different buffer conditions. (**c**) Violin plot of the total number of methylated cytosines detected, including both endogenous and exogenously introduced DNA methylation, in individual cells for different buffer conditions. (**d**) Violin plot of the ratio of methylated cytosines detected in different sequence contexts, comparing DNA accessibility (GCH) to endogenous CpG methylation marks (WCG), in individual cells for different buffer conditions. (**e**) Bar plot comparing detection of methylated cytosines in scMspJI-seq and scMAT-seq using optimized buffer conditions. The relative fractions of DNA accessibility (GCH), endogenous CpG methylation (WCG), and endogenous non-CpG methylation (WCH) are shown in blue, yellow, and orange, respectively. * and ** indicate $p < 0.05$ and $p < 0.01$, respectively (two-sided Mann-Whitney U test, Bonferroni-corrected p-values).



Supplementary Figure 3.2 | Optimization of mRNA enrichment in scMAT-seq.

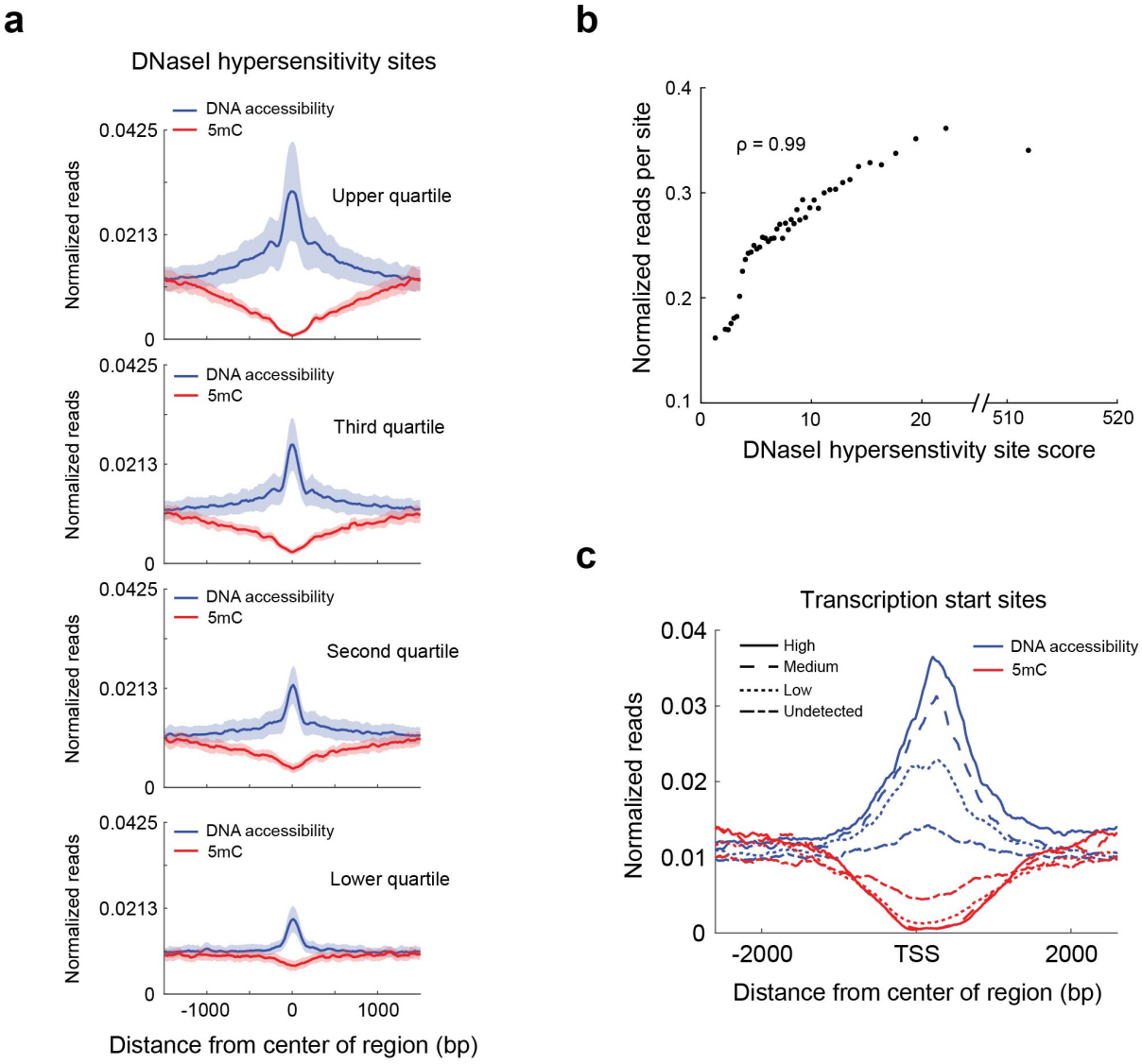
(**a**) Bar plot of the relative fraction of mRNA- and gDNA-derived reads in scMAT-seq without mRNA enrichment (No enrichment), with mRNA enrichment (RNA enriched), or the remaining flowthrough after mRNA enrichment (RNA depleted). (**b-d**) Violin plots of the number of transcripts (b), ERCC spike-in molecules (c), and genes (d) detected per cell using different magnetic streptavidin coated beads for mRNA enrichment. (**e**,**f**) Violin plots of the number of transcripts (e) and

ERCC spike-in molecules (f) detected per cell after dissociating (heat denatured) or retaining (on bead) the aRNA on the magnetic streptavidin C1 beads. (**g**) Scatterplot comparing single-cell averaged gene expression with and without RNA enrichment from the same scMAT-seq sample. * and ** indicate $p < 0.05$ and $p < 0.01$, respectively (two-sided Mann-Whitney U test, Bonferroni-corrected p-values).



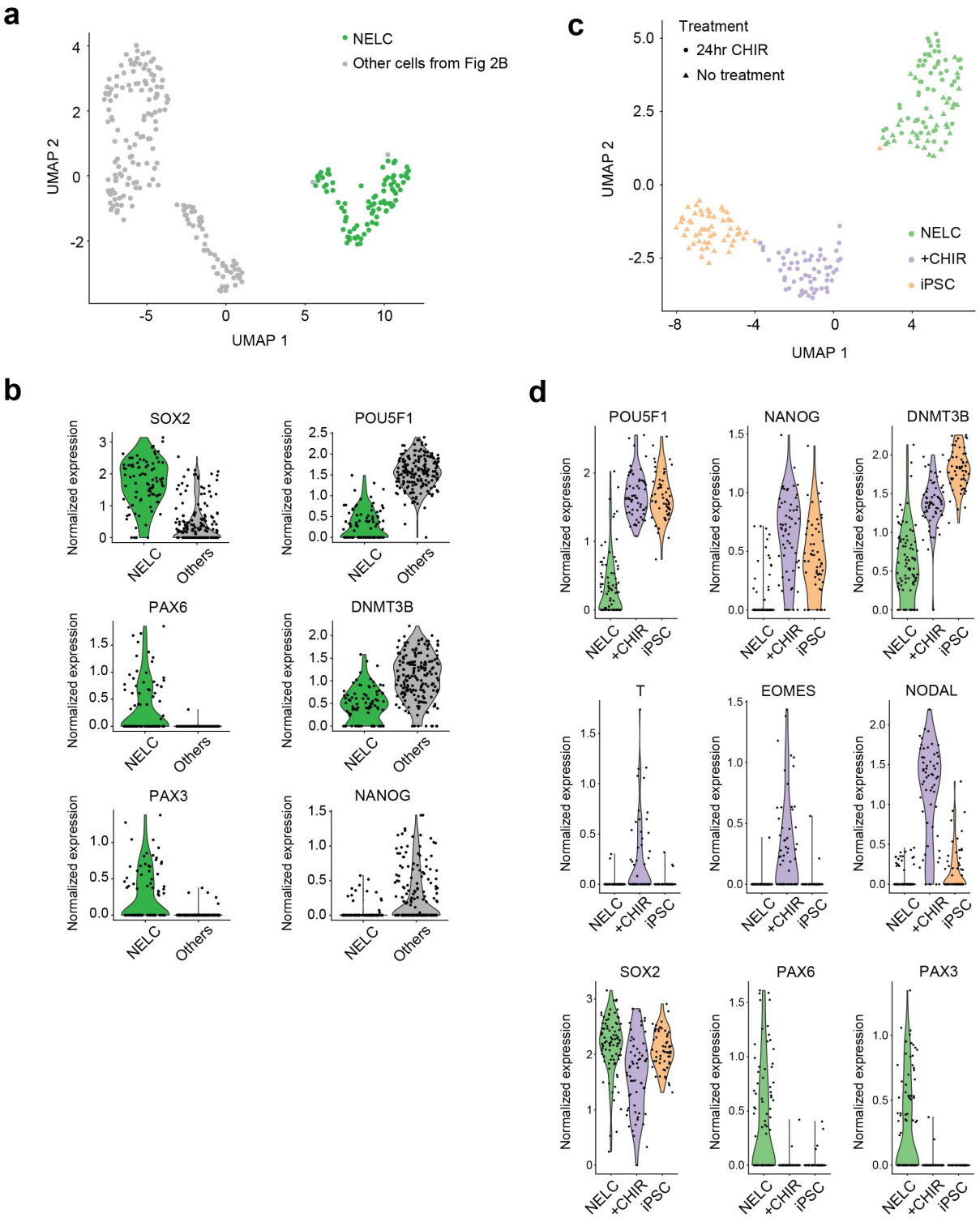
Supplementary Figure 3.3 | scMAT-seq successfully reproduces DNA accessibility and DNA methylation profiles in hESCs.

(**a**) Scatterplot shows that single-cell averaged DNA accessibility in scMAT-seq is highly correlated to DNase I hypersensitivity scores (Spearman's $p = 0.99$)[173]. (**b**) Pie chart shows that greater than 97% of DNA methylation sites detected in single cells using scMAT-seq (green) are also detected in bulk bisulfite sequencing (pink)[174].



Supplementary Figure 3.4 | scMAT-seq can accurately capture the genome-wide DNA accessibility and DNA methylation landscapes after cryopreservation of sorted samples.

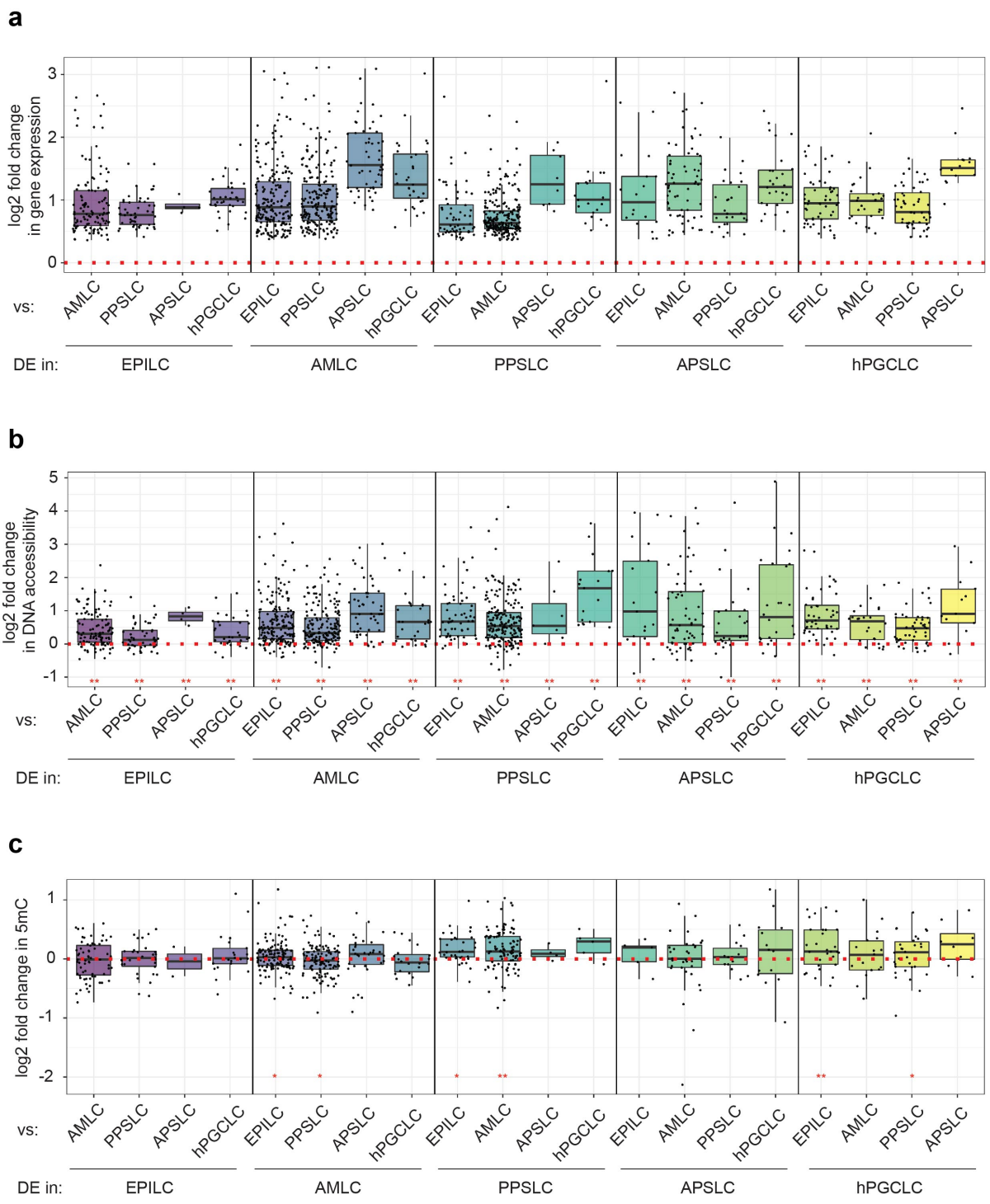
(**a**) Averaged single-cell DNA accessibility (blue) and DNA methylation (red) profiles from cryopreserved hESCs at DNase I hypersensitivity sites, split by previously reported signal strength. Shaded areas indicate standard deviation across single cells[173]. (**b**) Scatterplot shows that averaged single-cell DNA accessibility in scMAT-seq from cryopreserved hESCs is highly correlated to DNase I hypersensitivity scores (Spearman's p = 0.99)[173]. (**c**) Averaged single-cell DNA accessibility (blue) and DNA methylation (red) profiles in cryopreserved hESCs at TSS, segregated by gene expression levels: High (solid), medium (dashed), low (dotted), or undetected (dash dot).



Supplementary Figure 3.5 | Characterizing neuroectoderm-like cells during differentiation of iPSCs.

(**a**) UMAP visualization of single cells in the gastruloid 48-hour post BMP4 addition. Based on established marker genes, a cluster resembling NELCs is detected. (**b**) For the gastruloid characterized in (a), violin plots for select genes are shown, highlighting high expression of

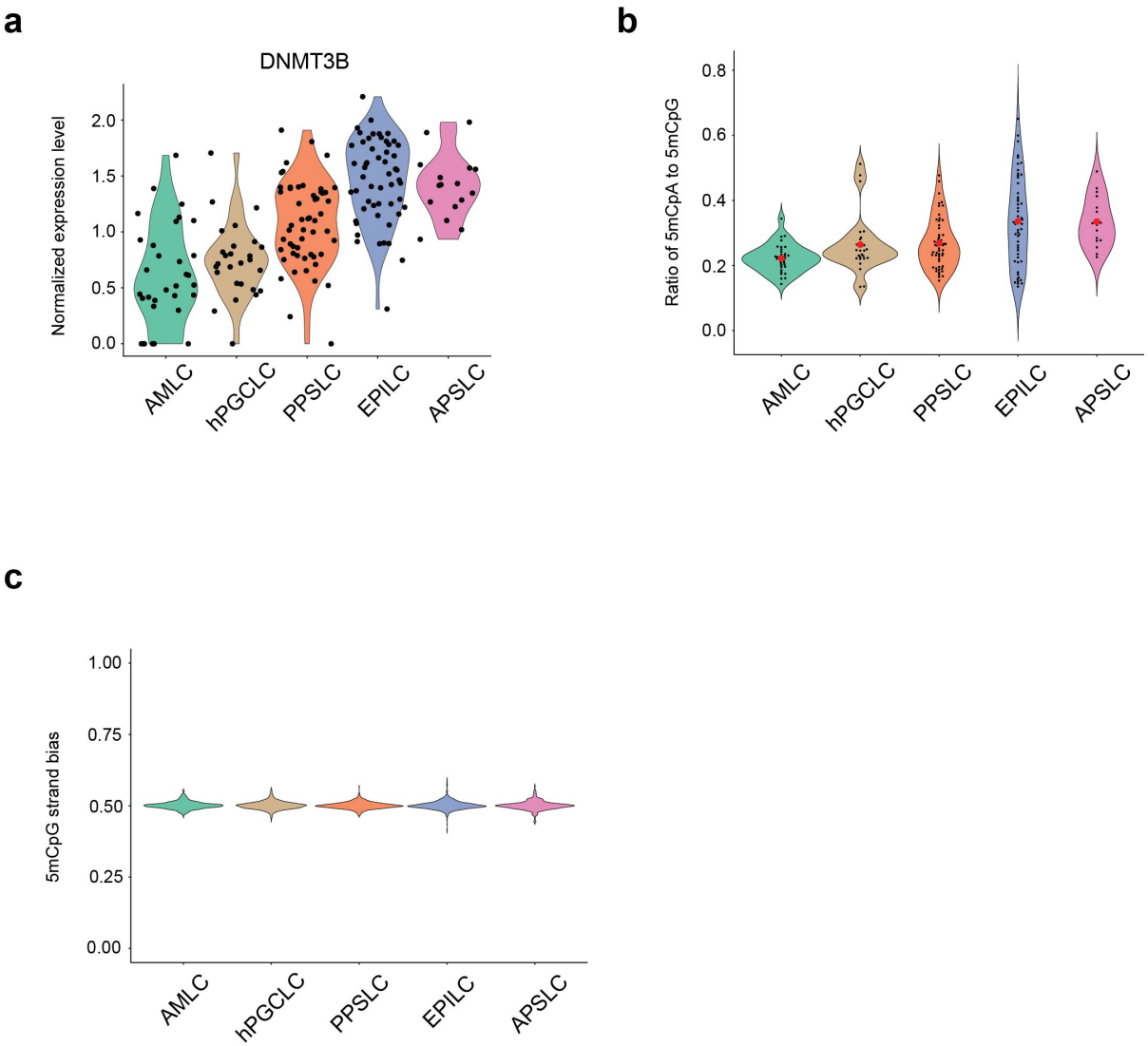
neuroectoderm genes (SOX2, PAX6, and PAX3) and low expression of pluripotency genes (POU5F1, DNMT3B, and NANOG) within the NELC population when compared to all other cells in the organoid. (**c**) UMAP visualization of untreated iPSCs (no treatment) and cells after treatment of iPSCs with CHIR99021 for 24-hours (24hr CHIR). Three populations of cells were detected, NELCs, iPSCs, and +CHIR cells. (**d**) For the system depicted in (c), violin plots for select genes are shown, highlighting high expression of pluripotency genes (POU5F1, DNMT3B, and NANOG) in iPSCs, high expression of mesodermal genes (T, EOMES, and NODAL) in the +CHIR cells, and high expression of neuroectoderm genes (SOX2, PAX6, and PAX3) in NELCs.



Supplementary Figure 3.6 | The epigenetic landscape of cell types identified in the human gastruloids 48 hours post BMP4 addition.

(**a**) Log2 fold change in gene expression for DEGs in one cell type (DE in) compared to another cell type (vs). Note that the log2 fold change in mRNA is computed by taking the ratio of expression in a cell type where the gene is differentially expressed at a higher level compared to the other cell type.

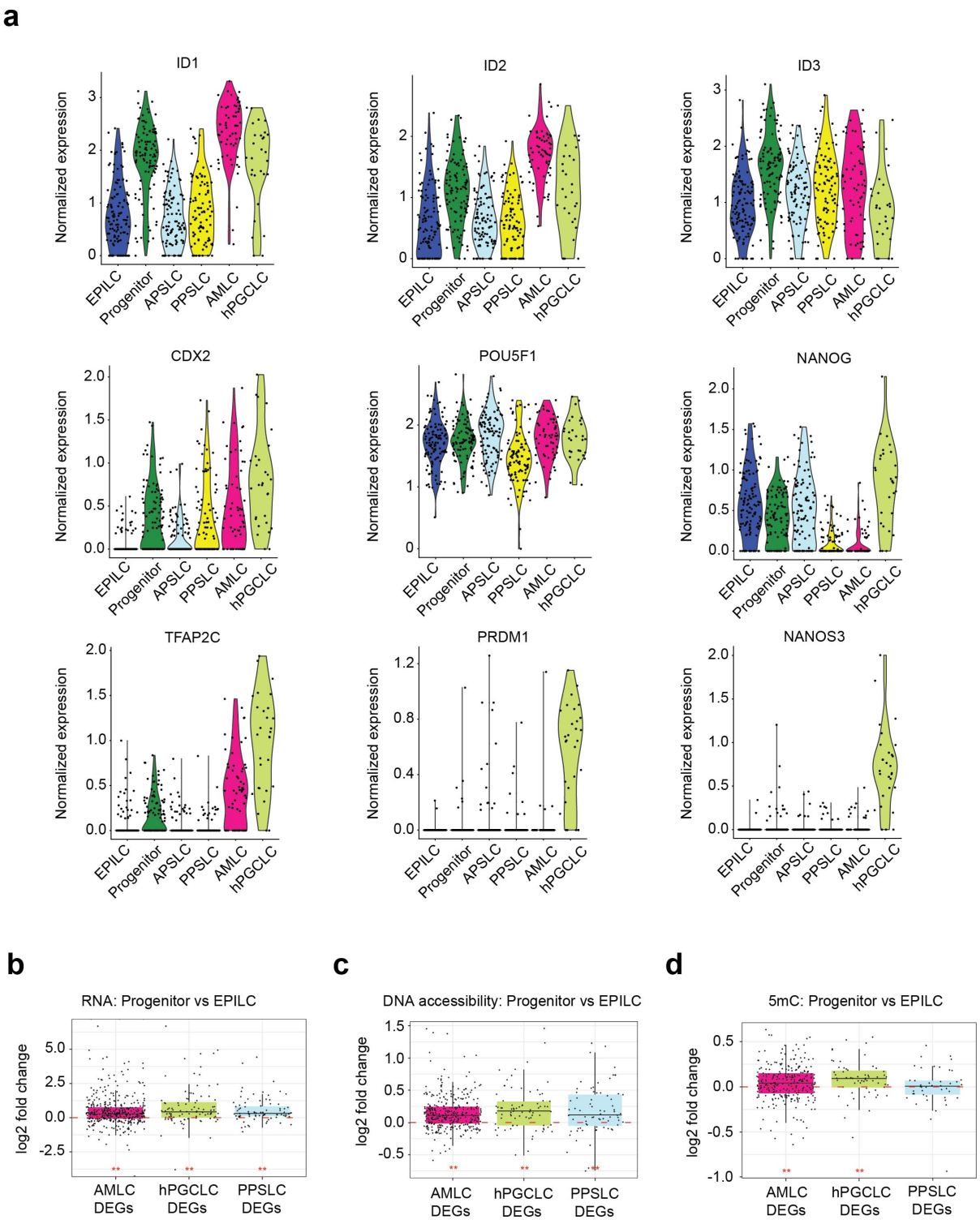
(**b**) Log2 fold change in DNA accessibility (promoter and gene body combined) for the genes depicted in (a). (**c**) Log2 fold change in gene body DNA methylation for the genes depicted in (a). Note that some of the genes shown in (a) are not depicted in (b) or (c) due to low detection across all cells. * and ** indicate a statistically significant change ($p < 0.05$ and $p < 0.01$, respectively, based on bootstrapped distributions from non-differentially expressed genes) in the log2 fold change of the epigenetic features for DEGs relative to non-differentially expressed genes.



Supplementary Figure 3.7 | DNA methylation dynamics in human gastruloids.

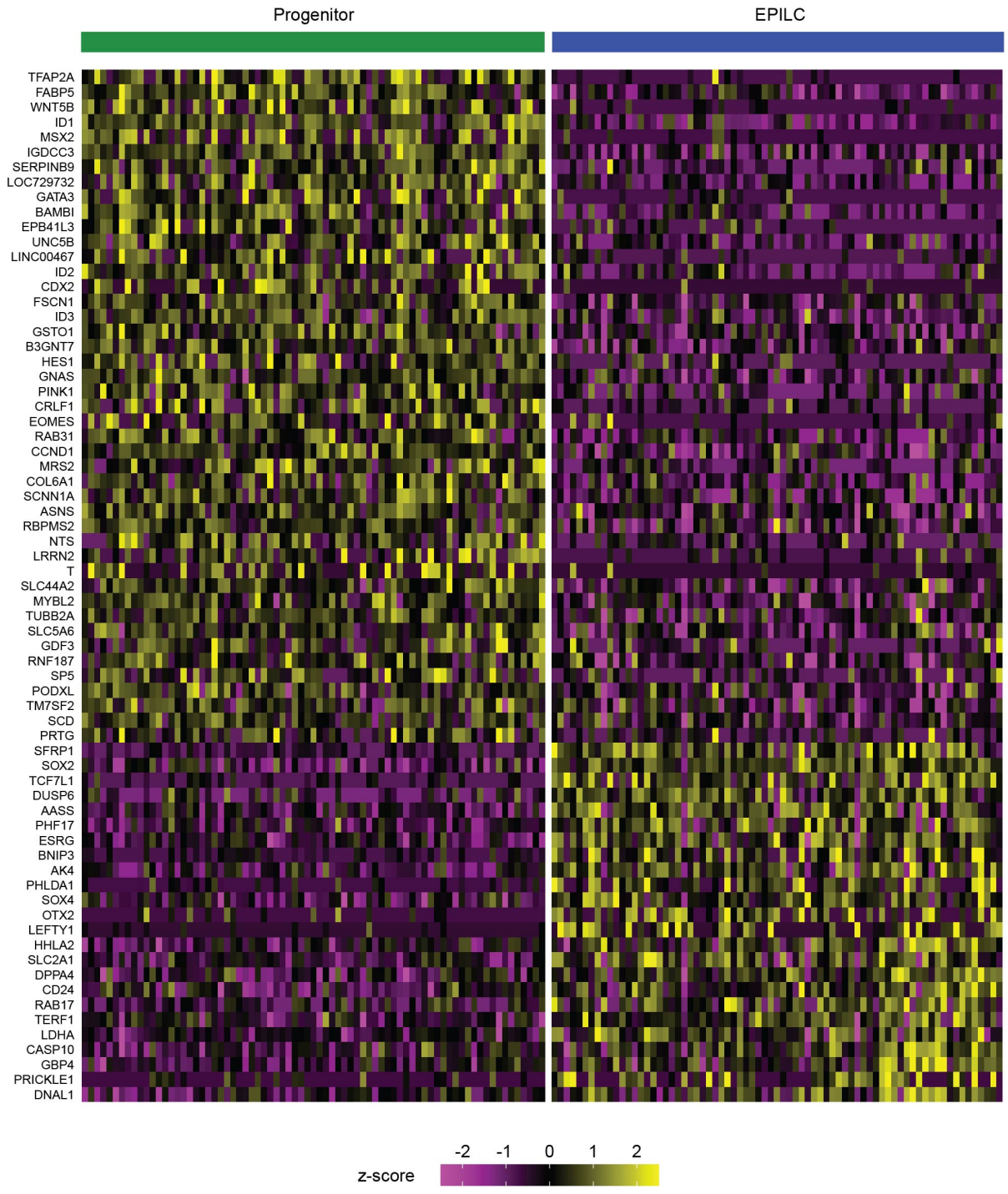
(**a**) Violin plot of gene expression levels for the *de novo* DNA methyltransferase DNMT3B for different cell types detected in the human gastruloid 48 hours post treatment with BMP4. Dots represent individual cells. (**b**) Violin plot of the ratio of 5mCpA to 5mCpG for different cell types. Black points represent individual cells, and the red point indicates the mean of the distribution. (**c**) Violin plot showing the distribution of 5mCpG strand bias in different cell types. 5mCpG strand bias is defined as the ratio of the number of 5mCpG sites detected on the plus strand of a chromosome over all 5mCpG

sites detected on a chromosome. The panel shows that no 5mCpG strand bias is detected in any of the cell types.



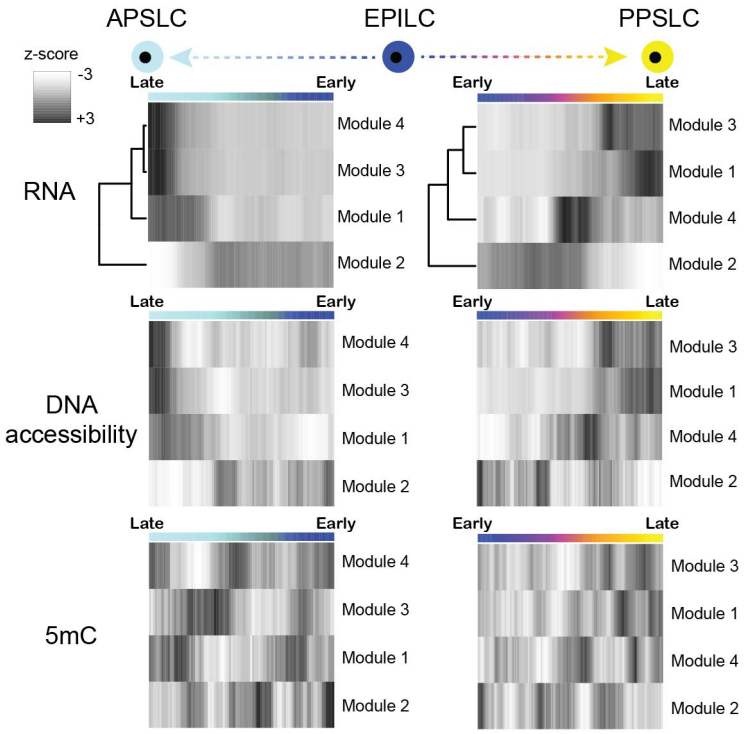
Supplementary Figure 3.8 | hPGCLC and AMLCs bifurcate from a common progenitor population.

(**a**) Violin plots of gene expression levels of select genes for cell types identified in the human gastruloids at different timepoints after BMP4 addition. Genes shown here include targets of BMP4 signaling (ID1-3), an amnion related gene CDX2, genes related to pluripotency (POU5F1 and NANOG), and genes related to hPGCLC fate (TFAP2C, PRDM1, and NANOS3). (**b**) log2 fold change in gene expression for Progenitor cells compared to EPILCs for genes that are differentially expressed in the amnion (AMLC DEGs), hPGCLCs (hPGCLC DEGs) or PPSLCs (PPSLC DEGS) when compared to all other cell types. (**c**) log2 fold change in DNA accessibility (promoter and gene body combined) for the genes depicted in (b). (**d**) log2 fold change in gene body DNA methylation for the genes depicted in (b). ** indicates $p < 0.01$ (one-sample Wilcoxon signed rank test).



Supplementary Figure 3.9 | DEGs between EPILCs and Progenitor cells.

Gene expression heatmap of z-scores for DEGs found between EPILC and Progenitor cells in human gastruloids 20 hours after treatment with BMP4.



Supplementary Figure 3.10 | APSLCs and PPSLCs arise from the EPILCs.

Heatmap of z-scores for DEG-derived gene modules along the APSLC and PPSLC pseudotime trajectory, together with the corresponding changes in DNA accessibility and DNA methylation. The color bar indicates position along the pseudotime, with EPILC in dark blue, APSLC in light blue, and PPSLC in yellow.

4. scMTH-seq: Connecting 5-methylcytosine, the transcriptome, and 5-hydroxymethylcytosine from the same single cell reveals processes responsible for human primordial germ cell maturation and DNA methylation erasure

A. Introduction

There are numerous methods to assess the status of cytosines in the genome[185]. One of the most common method involves bisulfite treatment of DNA to convert unmethylated cytosines to uracils[40]. This technique is very robust and can be adapted to obtain whole genome methylation information or a reduced methylome can be obtained through enzymatic digestion of CG rich sites[41,186]. A drawback of bisulfite sequencing is that it cannot discern between 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), making it less applicable to systems where 5mC dynamics are rapid and transient, like early embryogenesis and primordial germ cell (PGC) development[153]. Traditional bisulfite sequencing has been modified to assess only 5hmC through TAB-seq. In this method, 5hmC is glycosylated, protecting it from TET assisted oxidation of other cytosine modifications to 5caC which is followed by bisulfite sequencing[46]. While powerful, these types of techniques inherently cannot be used to observe both 5mC and 5hmC from the same sample in a single measurement. During PGC development where only a few hundred cells exist and the genome is at least in part actively demethylated through the conversion of 5mC to 5hmC, new technologies are needed[119]. To understand the dynamics between these two related epigenetic features in PGC development, here we have developed

a technique to detect 5mC and 5hmC simultaneously from the same cell (scMH-seq). Additionally, to further connect these epigenetic features to transcription and cell identity, we have incorporated scRNA-seq into this methodology (scMTH-seq) and apply this technique to *in vitro* derived human PGCs.

B. Results

1. 5hmC and non-CpG 5mC are inherited from parental DNA strands at similar rates in hESCs

In our group and others, restriction enzymes that specifically recognize cytosine modifications have been used to create sequencing libraries. To interrogate 5mC, a restriction enzyme MspJI, which specifically recognizes methylated cytosines can be used when preparing a DNA library for sequencing[65,187]. Similarly, AbaSI specifically recognizes 5hmC and can be used for DNA library preparation to quantify 5hmC[34,188]. Digestion with MspJI generates a 4 nucleotide 5' overhang 16 nucleotides downstream from a methylated cytosine, while digestion with AbaSI creates a 2 nucleotide 3' overhang 11-13 base pairs downstream from a 5hmC site. Due to the differences in cutting modality, highly specific ligation of barcoded adapters to distinguish between the two enzymatic digestions is possible even at the single-cell level[35]. In scMH-seq we use both MspJI and AbaSI to specifically digest and capture DNA containing 5mC and 5hmC respectively. While in scMTH-seq, a reverse transcription and second strand synthesis step is performed prior to enzymatic digestion to capture RNA transcripts (Fig. 4.1a). After digestion, corresponding sticky end adaptors containing a T7 promoter, part of the Illumina 5' adaptor, a cell barcode, and for 5mC a unique molecule identifier (UMI) are ligated on. These ligated molecules are amplified, and Illumina libraries are prepared as

described previously[34,65]. Because each molecule contains a barcode indicating its epigenetic feature and cell of origin, reads from each type can be segregated *in silico* and putative 5mC or 5hmC sites are called for each cell by identification of the enzyme cut site from sequences mapped to a reference genome.

To investigate the dynamics of 5hmC conversion from 5mC in pluripotent cells, we successfully applied scMH-seq to 79 human embryonic stem cells and scMTH-seq to an additional 197 human embryonic stem cells and find that multiomics detection levels are similar to those detected in scRNA-seq, scMspJI-seq and scAbaSI-seq alone (Fig. 4.1.b,c)[34,65,189]. 5hmC is known to be a rare epigenetic feature occurring at much lower rates than 5mC in most systems, here we find on average 5hmC is at levels 3.1% that of 5mC in a CpG context (Fig. 4.1.d). Interestingly, the coefficient of variation observed for the level of 5hmC is greater than that of 5mCpG (0.61 vs 0.52 respectively) (Fig. 4.1b,c). 5mC in a CpG context is known to be highly maintained and is likely found a similar levels for these cells, although the measurement will be noisy due to differences in amplification between cells[190]. Both measurements are made in a similar fashion and so both measurements should have similar levels of noise, thus the higher coefficient of variation observed with 5hmC is likely due to high variability in 5hmC levels due to inheritance of varying levels of 5hmC. Because scMH-seq and scMTH-seq are strands specific, this hypothesis can be directly tested by assessing the strand bias of each epigenetic feature. Strand bias of a region for either epigenetic mark is defined as the number of that mark on the plus strand divided by the total detected in the region. A strand bias of 0.5 indicates that both strands of DNA have equal levels of the epigenetic mark of interest. Strand biases deviating from 0.5 indicates

differences between the old strand of DNA and the more recently synthesized DNA strand. This observable difference between complimentary DNA strands is the framework for lineage reconstruction by single-cell 5hmC sequencing and can provide insight into the rate at which an epigenetic mark is accumulating[34,35]. Here we find that 5hmC has high levels of strand bias while the maintained 5mCpG mark deviates very little from 0.5 (Fig. 4.1e). This observed pattern is similar to previous observations using individual 5hmC and 5mC measurements in mouse embryonic stem cells (Supplementary Fig. 4.1).

In human embryonic stem cells, non-CpG methylation was found to have high strand bias, a feature not robustly observed in mouse embryonic stem cells (Fig. 4.1e and Supplementary Fig. 4.1). Non-CpG methylation can potentially show strand bias because DNA methylation maintaince machinery is known to only have a strong preference for hemi-methylated CpG[191]. Like what is seen with 5hmC, the newly synthesized DNA strand has very low levels of non-CpG methylation and if the rates of accumulation of non-CpG methylation are slow relative to the cell cycle, differences between the two strands will be observed. In support of this mechanism, coinciding high levels of strand bias within a cell for both 5hmC and non-CpG 5mC was found (Fig. 4.1.f). Because the inherited DNA strand for any cell will be decorated by high levels of 5hmC and non-CpG 5mC relative to the new DNA strand, this implies that the strand bias of 5hmC and that of non-CpG 5mC would deviate in the same direction about 0.5. Our experimental results confirm this is the case, with a high correlation between 5hmC strand bias and non-CpG strand bias (Pearson's correlation (r) of 0.91, 0.78, and 0.42 for 5mCpA, 5mCpT, and 5mCpC respectively), but a correlation near zero between 5hmC strand bias and 5mCpG

strand bias (r of 0.05) (Fig. 4.1.g and Supplementary Fig. 4.2a-c). Interestingly we also see that the slope of the best fit line between the strand bias of 5hmC and non-CpG methylation differs with a slope of 0.59, 0.28, and 0.12 for 5mCpA, 5mCpT, and 5mCpC respectively. We reasoned that if the rates of accumulation of each epigenetic feature was the same then the slope of this best fit line would be equal to 1. It is well established in embryonic stem cells that 5mCpA is the most abundantly methylated non-CpG dinucleotide, a feature also observed in our data (Fig. 4.1c.)[192]. Since the slope between the strand bias of 5hmC and 5mCpA is below 1 and reduces for 5mCpT and 5mCpC, we conclude that 5mCpA accumulates in human embryonic stem cells more slowly than 5hmC, but more quickly than other non-CpG methylation. To further test this conclusion, we performed the same analysis on each cell individually, and compared the results to *in silico* cells with each epigenetic feature randomly derived from other cells (Fig. 4.1h,i). A very high slope and Pearson's correlation is seen between 5hmC strand bias and 5mCpA for all cells, although a slope greater than 1 is never observed, confirming that 5hmC always accumulates at a faster rate than 5mCpA in this system (Fig. 4.1h,i). For 5mCpT most cells have strong correlations, while for 5mCpC the correlation is lower (Fig. 4.1.i). The low Pearson's correlation for 5mCpC, the most slowly accumulating feature measured here maybe due to high levels of noise when measuring this lowly abundant mark. To further confirm the characteristics observed here are related to the rates of accumulation, we performed stochastic modeling on 5hmC and non-CpG methylation to estimate the turnover rate of each epigenetic mark. Turnover rate is the summation of the forward and reverse reaction rate creating and removing an epigenetic mark and has previously been used to evaluate reaction

rates for 5hmC[34]. In agreement with our previous analysis, we find the turnover rate to be fairly low for both 5hmC and 5mCpA, but much higher for 5mCpT and 5mCpC (average value of 0.98, 2.70, 5.72, and 6.22 per cell division respectively) (Supplementary Fig. 4.3).

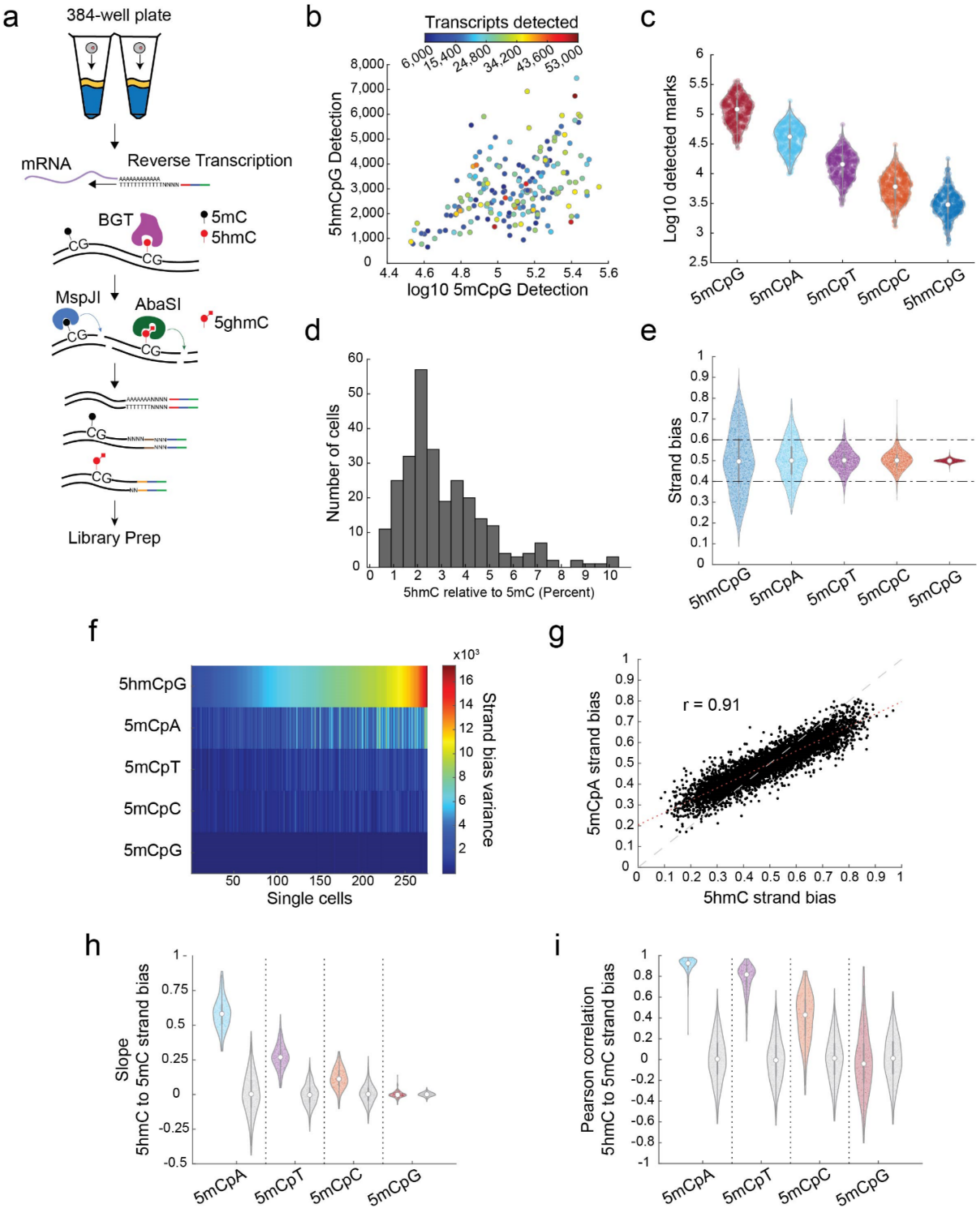


Figure 4.1 | Dynamics between 5hmC and non-CpG methylation in H9 human embryonic stem cells detected by scMH-seq and scMTH-seq.

(**a**) Schematic of scMTH-seq. Cell and detected species specific barcodes are shown in red, brown, and gold for mRNA, 5mC, and 5hmC respectively. The Illumina read 1 sequencing primer is in

blue, and the T7 promoter is in green. (**b**) Detection levels for 5mCpG, 5hmCpG and mRNA transcripts from H9 cells. (**c**) Detection levels for 5mCpG, non-CpG 5mC and 5hmCpG in individual cells. (**d**) The relative ratio of 5hmCpG detection to 5mCpG detected in individual cells. (**e**) Strand bias for 5mCpG, non-CpG 5mC and 5hmCpG. Each dot represents a single chromosome. (**f**) Heatmap of the variance in the strand bias distribution from all autosomal chromosomes of a single cell for 5mCpG, non-CpG 5mC and 5hmCpG. (**g**) A comparison of 5hmCpG and 5mCpA strand bias for the same chromosome of all single cells assessed. (**h,i**) A comparison of 5hmCpG and all other 5mC strand bias for the same chromosome. Each point indicates the slope (**h**) or correlation (**i**) from one cell. Grey dots indicate a *in silico* cell, where strand bias of each feature was randomly assigned from two different cells in the data set.

2. *DND1 and SOX15 expression are promising triggers for passive demethylation and cell cycle arrest in maturing hPGCLCs*

We find that in human embryonic stem cells, because 5mCpG is maintained, there is no connection between its strand bias and that of 5hmC (Fig. 4.1h,i). DNA methylation maintenance is a characteristic of most cell types, but notably the mechanisms involved are inhibited during early preimplantation embryogenesis and PGC development, leading to global methylation loss[193]. Until recently, the ability to create human PGC like cells (hPGCLCs) has been limited to PGCLCs with features akin to recently specified PGCs which have yet to undergo this global demethylation phenomena[194]. The limited ability to achieve mature PGCLCs is in part due to inadequate in vitro culture conditions, which only allow for a few days of PGCLC growth. Recently we have derived a system to support long-term growth of PGCLCs and have shown some cells in this culture under passive demethylation[134]. The role of active demethylation and the effect of 5mC erasure on cell identity in this system are currently unknown. Due to the mixed population present, this understanding cannot be gained using current techniques to evaluate each modality separately. To investigate this, we applied scMTH-seq to PGCLCs 4 days after induction (D4) and PGCLCs cultured for an additional 10 or 21 days after D4 in long term culture

conditions (D4C10 and D4C21 respectively). Unlike undifferentiated stem cells, in all PGCLC conditions we find a subset of cells where high variance in 5hmC strand bias co-occurs with high variance in 5mCpG strand bias, an indicator that passive demethylation is occurring in at least a subset of cells (Fig. 4.2.a). Interestingly only D4 PGCLCs experienced high variance in non-CpG methylation (Fig. 4.2.a). The slope and Pearson's correlation between 5hmC and non-CpG strand bias was accordingly very low for all but D4 PGCLC 5mCpA methylation, indicating very slow non-CpG methylation dynamics in PGCLCs (Supplementary Fig. 4.4a,b). While a limited connection was seen between 5hmC and non-CpG methylation, high correlations were seen between 5hmC and 5mCpG strand bias, a clear indication that 5mC is not being maintained in a subset of cells (Supplementary Fig. 4.4b). We find that our measurements of 5mCpG strand bias using scMTH-seq correspond well to previous measurements using scMspJI alone (Fig. 4.2b)[134]. Using all data sets, we then grouped cells as being maintained (MnT) or unmaintained (UmnT) by using the observed level of 5mCpG strand bias variance in human embryonic stem cells which do not experience passive demethylation as a cutoff. Doing this, we find a large portion of PGCLCs are passively demethylating, with a trend for more passively demethylating cells with longer culturing times (Fig. 4.2b).

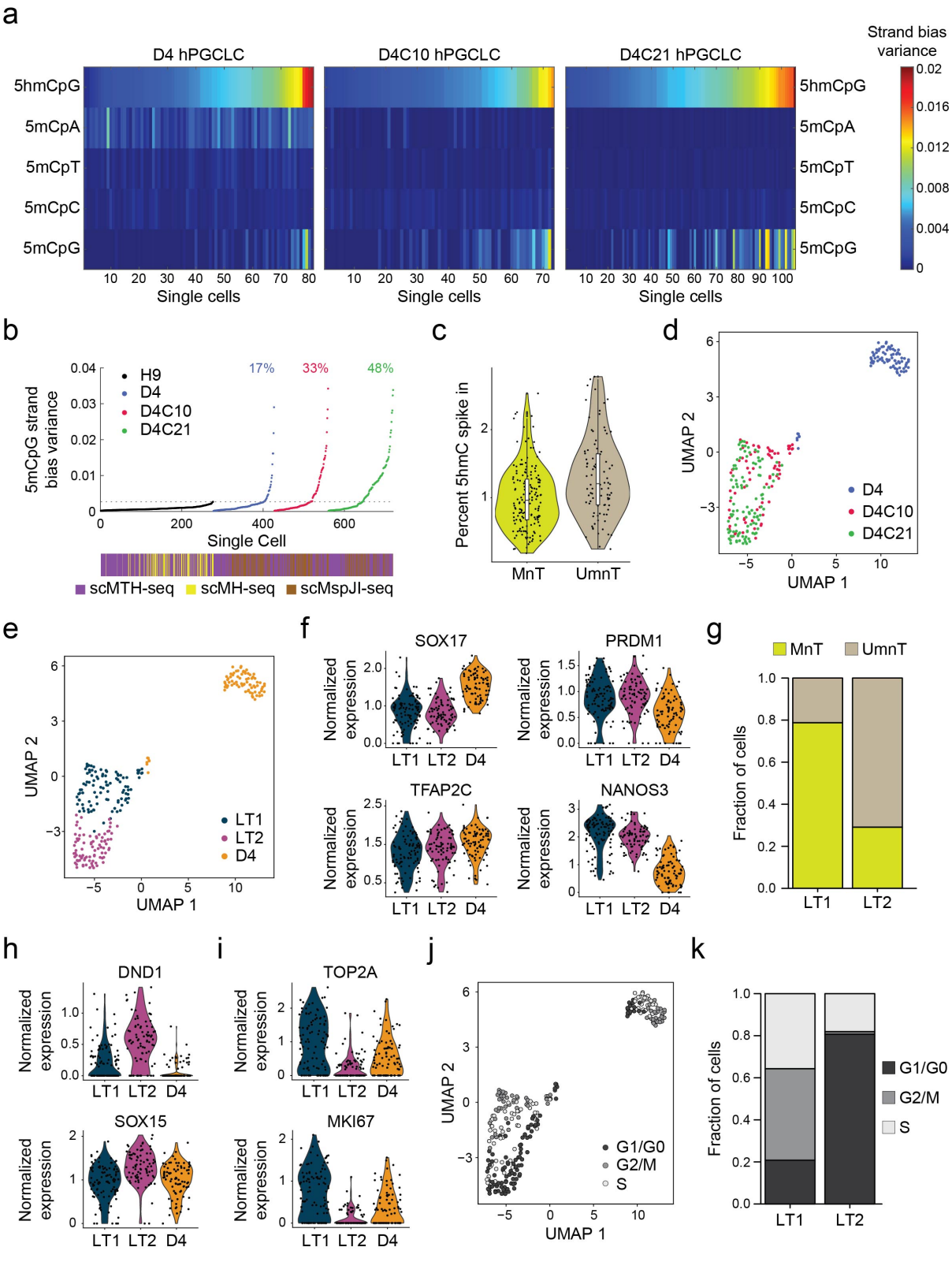


Figure 4.2 | Long term culturing of PGCLCs results in a heterogenous population containing a non-cycling, transcriptionally distinct population that has passively demethylated.

(**a**) Heatmap of the variance in the strand bias distribution from all autosomal chromosomes of a single cell for 5mCpG, non-CpG 5mC and 5hmCpG, discretized by PGCLC culture conditions. (**b**) Comparison of the variance in the strand bias distribution from all autosomal chromosomes of a single cell for 5mCpG. Grey line indicates the maximal level seen in H9 human embryonic stem cells, which were found to have maintaince of DNA methylation. Cells above this line were considered as UmnT. Cells below this line were considered as MnT. Percent values indicate percent of cultured cells in the UmnT group for a given condition. (**c**) Detection of 5hmC spike-in molecules in each cell for the MnT and UmnT groups. (**d,e**) UMAP projection of PGCLC transcriptome discretized by culture condition (**d**) or gene expression based clustering (**e**). (**f**) Normalized expression in each condition of key PGC genes, SOX17, PRDM1 (aka BLIMP1), TFAP2C, and NANOS3. (**g**) Methylation maintenance condition of cells in long-term culture gene expression groups 1 and 2 (LT1 and LT2 respectively). (**h,i**) Normalized expression of genes found to be differentially expressed between LT1 and LT2. (**h**) PGC related genes DND1 and SOX15. (**i**) Cell cycle related genes TOP2A and MKI67. (**j**) UMAP projection of PGCLC transcriptome discretized by predicted cell cycle stage (G1/G0, G2/M, or S phase) (**k**) Predicted cell cycle phase of cells in long-term culture gene expression groups LT1 and LT2.

A temporally regulated wave of active demethylation is known to occur after a period of passive demethylation in PGCs. Using the detection of endogenous 5hmC compared to spike in molecules we find that the PGCs created in this system are relatively immature and do not exhibit robust active demethylation as is seen in vivo (Fig 4.1.c). We then turned our attention to the transcriptome of these cells and found that D4 PGCLCs distinctly separate from those in long term culture conditions (D4), but within long-term culture conditions, two transcriptional states are present, LT1 and LT2 (Fig. 4.2d,e). While levels of key PGC genes like SOX17, PRDM1 (aka BLIMP1), TFAP2C and NANOS3 are expressed in all PGCLCS, some differences are seen between the three transcriptional groups, with the largest changes between the D4 and long-term culture groups (Fig. 4.2f). This transcriptional analysis confirms prior bulk RNA-seq and immunofluorescence data describing the ability of long term culture conditions to maintain a PGCLC state[134].

While these long-term culture conditions maintain a PGCLC state, there is clear heterogeneity within epigenetic state, MnT vs UmnT, and transcriptional state, LT1 and LT2. We hypothesized that these two states may be linked, and indeed we find

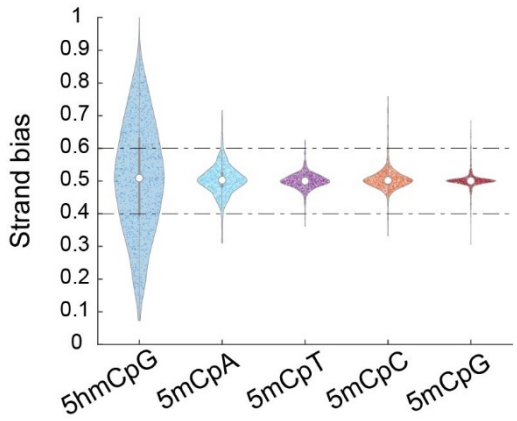
the LT2 transcriptional state is highly enriched for UmnT cells (Fig. 4.2g). Because PGCs demethylate a few days after being specified, it is likely that cells in the LT2 transcriptional group are more mature PGCLCs. In agreeing with this, differentially expressed genes between the two transcriptional states within long term culture, LT1 and LT2, are highly similar between LT1 and D4 PGCLC groups, suggesting that the LT1 group is more transcriptionally like the D4 PGCLCs which have not experienced extended culture (Supplementary Fig. 4.5). As such, these differentially expressed genes are also putative genes for PGC maturation and the initiation of passive demethylation. Notably, DND1 and SOX15 are highly expressed in LT2 when compared to LT1. Both DND1 and SOX15 have been shown to be crucial for proper PGC development in mice, with major losses in PGC numbers occurring after specification in mutants lacking wild type protein expression[195,196]. Additionally, DND1 expression in *Xenopus* directly regulates NANOS1, a key regulator of PGC fate in this organism[197,198]. Interestingly, DND1 has also been strongly implemented in the downregulation of active cell cycle genes, with a mutant form of DND1 causing gonadal teratoma formation in mice[199]. Consistent with this role of DND1, we find active cell cycle genes TOP2A and MKI67 down regulated in the LT2 population, which also expresses DND1 (Fig. 4.2h,i). Cell cycle analysis reveals that most cells in LT2 are non-cycling cells in either G1 or G0 phase, likely at least in part due to high levels of DND1 (Fig. 4.2j,k). The LT2 group occurring mainly in a non-dividing phase, while simultaneously being passively demethylated, suggests that the cells were dividing previously (Fig. 4.2g,k). Because the LT1 group is cycling, they likely give rise to the LT2 population after a cell division where 5mCpG maintaince is impaired. While more study is needed to fully understand how this more mature

population of PGCLCs arises, it is likely that some of the genes identified here are key regulators of this process (Supplementary Fig. 4.5).

C. Conclusion

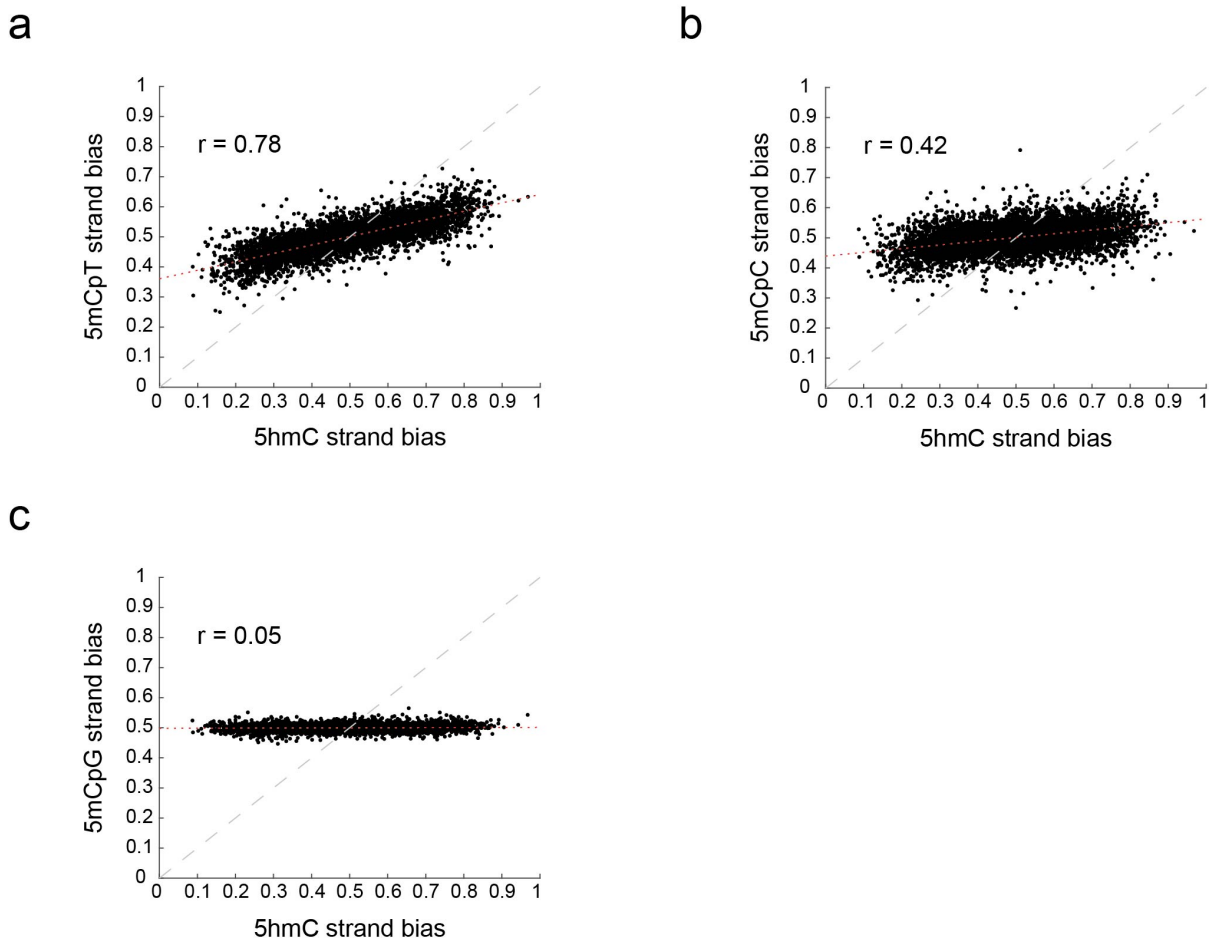
Here we have combined two powerful single cell techniques, scAba-seq and scMspJI-seq into scMH-seq, which is able to identify both 5mC and 5hmC simultaneously from the same cell. Using scMH-seq we identified passive dilution of non-maintained epigenetic marks is a major regulator of the epigenome, with the potential for inherited DNA strands to be highly decorated with 5hmC and non-CpG methylation. Furthering this work, by creating scMTH-seq, we added the ability to obtain the transcriptome from the same single cell as well. We then applied scMTH-seq to a heterogenous PGCLC population under long-term culture conditions where passive demethylation of CpG methylation had been found previously[134]. Under these conditions we observed two transcriptionally distinct sets of cells, one of which was passively demethylating. This same group of cells also exhibited high levels of DND1 and SOX15, two genes critical in PGC development. The expression of DND1 may have induced their more mature state but may have also force these cells out of the active cell cycle. The further investigation of the role of these genes in PGC develop and epigenome remodeling in this system and other PGC model systems will be critical for understanding the maturation process in human primordial germ cells and the formation of germ line tumors.

D. Supplementary figures



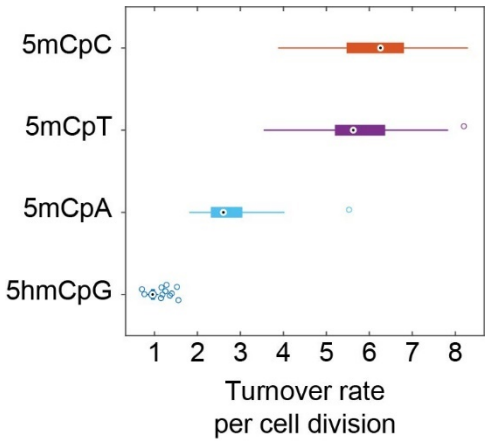
Supplementary Figure 4.1 | 5hmC and 5mC strand bias in mouse embryonic stem cells.

Detection of 5hmC and 5mC were performed in separate cells. Strand bias of 5hmC was detected by scAba-seq. Strand bias of 5mC marks in different dinucleotide context was detected by scMspJI-seq. Data from Sen *et al.* and Mooijman *et al.*[34,65].



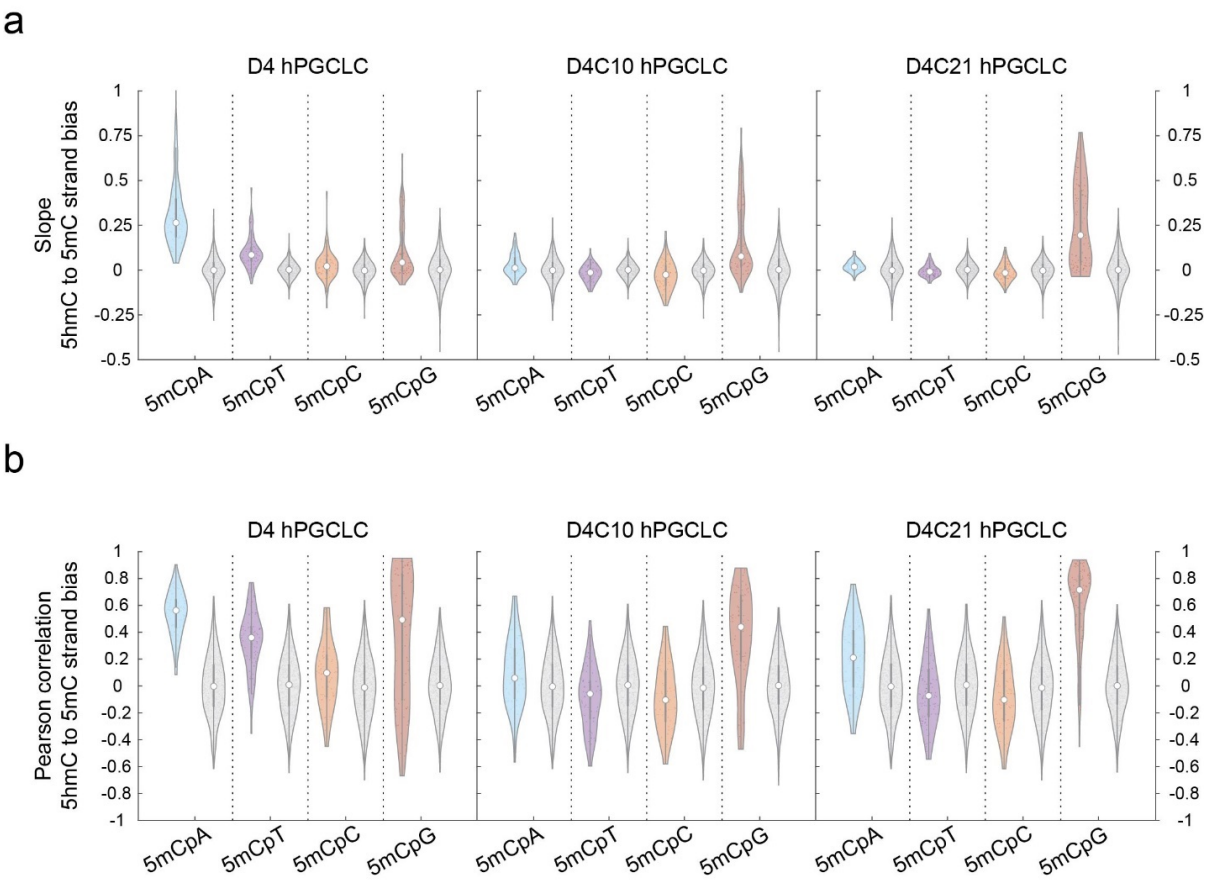
Supplementary Figure 4.2 | 5hmC and non-CpG methylation accumulate together on parental DNA strands.

(**a-c**) A comparison of 5hmCpG and 5mC strand bias for the same chromosome of all single cells assessed. The comparison to 5hmCpG strand bias was made for 5mCpT (**a**), 5mCpC (**b**), and 5mCpG (**c**) strand bias.



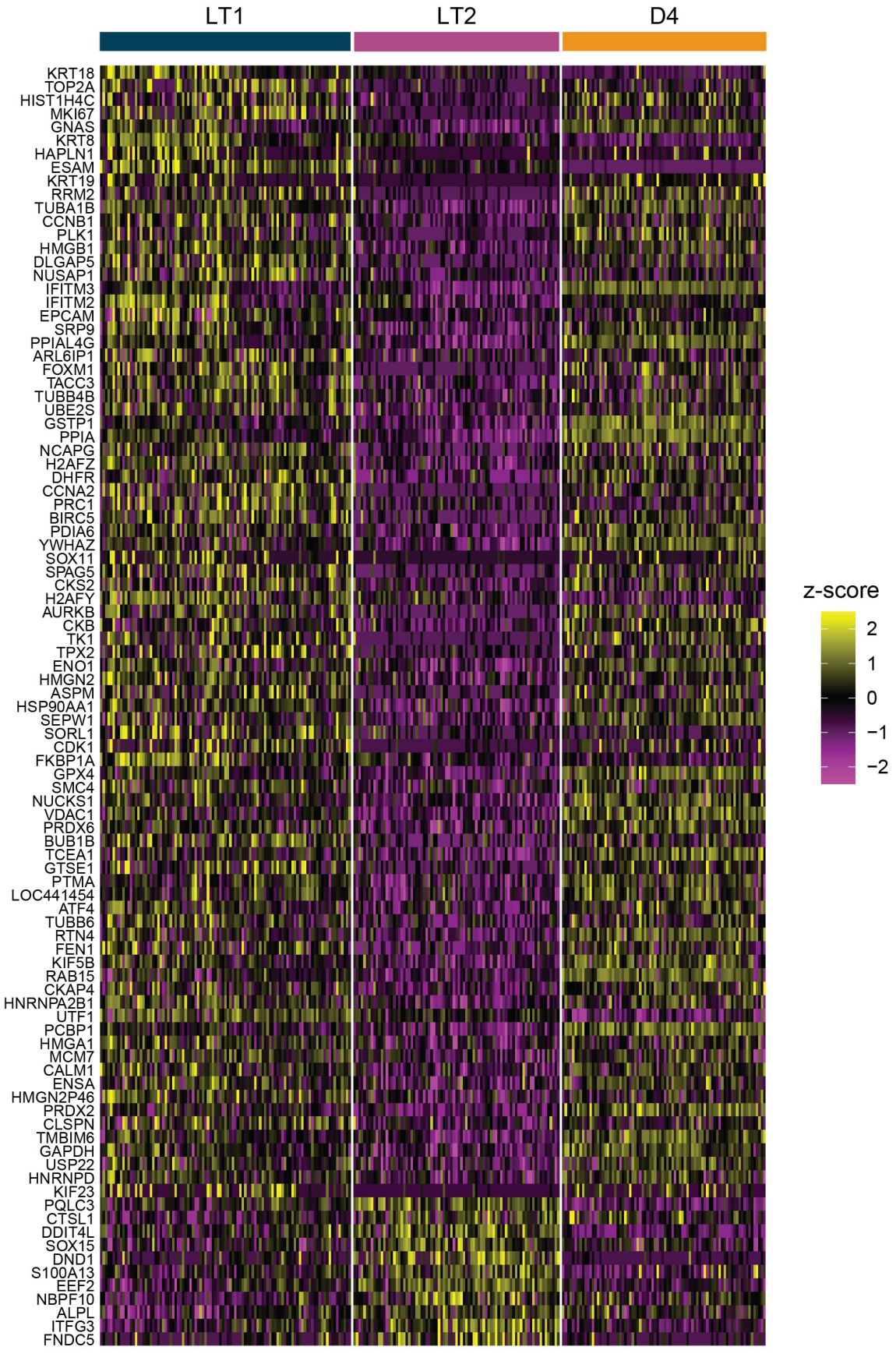
Supplementary Figure 4.3 | Turnover rate for 5hmCpG and non-CpG methylation in H9 human embryonic stem cells.

The turnover rate estimated using model IV described previously by Mooijman *et al.*[34]. Boxplots indicate the results from 100 simulations.



Supplementary Figure 4.4 | Comparison of 5hmCpG and other 5mC strand bias in hPGCLCs.

(**a,b**) A comparison of 5hmCpG and all other 5mC strand bias for the same chromosome. Each point indicates the slope (**a**) or correlation (**b**) from one cell. Grey dots indicate a *in silico* cell, where strand bias of each feature was randomly assigned from two different cells in the data set. Cells were split and plotted separately based on their culture condition as indicated.


LT1
LT2
D4
KRT18
TOP2A
HIST1H4C
MKI67
GNAS
KRT8
HAPLN1
ESAM
KRT19
RRM2
TUBA1B
CCNB1
PLK1
HMGB1
DLGAP5
NUSAP1
IFITM3
IFITM2
EPCAM
SRP9
PPIAL4G
ARL6IP1
FOXM1
TACC3
TUBB4B
UBE2S
GSTP1
PPIA
NCAPG
H2AFZ
DHFR
CCNA2
PRC1
BIRC5
PDIA6
YWHAZ
SOX11
SPAG5
CKS2
H2AFY
AURKB
CKB
TK1
TPX2
ENO1
HMGN2
ASPM
HSP90AA1
SEPW1
SORL1
CDK1
FKBP1A
GPX4
SMC4
NUCKS1
VDAC1
PRDX6
BUB1B
TCEA1
GTSE1
PTMA
LOC441454
ATF4
TUBB6
RTN4
FEN1
KIF5B
RAB15
CKAP4
HNRNPA2B1
UTF1
PCBP1
HMGA1
MCM7
CALM1
ENSA
HMGN2P46
PRDX2
CLSPN
TMBIM6
GAPDH
USP22
HNRNPD
KIF23
PQLC3
CTSL1
DDIT4L
SOX15
DND1
S100A13
EEF2
NBPF10
ALPL
ITFG3
FNDC5
z-score
2
1
0
−1
−2

Supplementary Figure 4.5 | Gene expression differences between hPGCLCs populations in long term culture.

Heatmap of differentially expressed genes found between the two cell populations in long term culture, LT1 and LT2. Color indicates z-score for normalized transcript expression across all PGCLCs sequenced.

5. scDyad&T-seq: heterogenous global demethylation in naïve embryonic stem cells results from differences in DNA methylation maintenance in single cells

A. Introduction

Epigenetic remodeling, including genome wide erasure of 5-methylcytosine (5mC) is associated with the acquisition of pluripotency in mammalian primordial germ cells development and during early preimplantation embryogenesis. In addition to low de novo methylation rates and high 5mC oxidation by TET proteins, during these developmental stages cells exhibit passive demethylation, where 5mC is not faithfully copied to newly synthesized DNA during replication. Recently, we have shown that the genome wide erasure of 5mC in early mouse and human embryogenesis is heterogenous, with a subset of cells experiencing global passive demethylation[65]. Additionally, while most DNA methylation accumulates on the newly synthesized DNA strand very quickly, some fraction of sites remain demethylated for hours after replication and local differences in 5mC levels can affect this 5mC maintenance process[200,201]. These findings indicate that tools to investigate passive demethylation of 5mC are critical in understanding pluripotency. Currently, passive demethylation can be identified through hairpin-bisulfite sequencing, where complimentary DNA strands are physically connected[44,165]. In fact, this concept has even been extended to alternative forms of bisulfite conversion[202]. But, due to the nature of physically connecting the two opposing strands, extensions of these techniques could not be employed to directly investigate the presence of 5hmC on one strand and 5mC on the other strand of a single DNA molecule. Here we

describe CpG Dyad-sequencing (Dyad-seq) which combines enzymatic detection of modified cytosines and traditional nucleobase conversion techniques to identify the presence of hemimethylation or hemihydroxymethylation at the resolution of a single CpG dyad site. We then scale this technique down and allow for the simultaneous detection of RNA transcripts all from the same single cell.

B. Results

1. Detecting 5mC and 5hmC on both strands of the same piece of DNA using Dyad-seq

Dyad-seq describes a generalized method for detecting modified or unmodified cytosines on both strands of a single piece of DNA. We present 4 versions of Dyad-seq, two where the presence of 5mC is known on one strand through the digestion of DNA with MspJI (M-M-Dyad-seq and M-H-Dyad-seq), and two where the presence of 5hmC is known on one strand through the digestion of DNA with AbaSI (H-M-Dyad-seq and H-H-Dyad-seq)[34,65] (Fig. 5.1a). Digested molecules are captured by ligation of the bottom strand to a double stranded adapter containing a sample barcode, UMI, PCR amplification sequence, and corresponding overhang. Next unmodified cytosines are converted enzymatically using APOBEC3A or by using sodium bisulfite to uracil (M-M-Dyad-seq and H-M-Dyad-seq)[50]. These methods measure a combined signal from 5mC and 5hmC, slight modification in the conversion step can result in only the detection of only 5hmC (M-H-Dyad-seq and H-H-Dyad-seq)[46–49,51]. The bottom strand of the adapter is devoid of cytosine and thus it is unaffected by cytosine conversion[56]. Next, random primer extension is used to incorporate part of the Illumina read 2 adapter sequence. The resulting molecules are then PCR amplified and subjected to next generation sequencing. Due to the

use of a 5mC or 5hmC specific endonuclease, the presence of the epigenetic feature can be inferred on the non-amplified strand, while the methylation status of the opposing CpG site can be determine directly from the sequencing results (Supplementary Fig. 5.1a,b). These techniques provide a measurement of 5mC or 5hmC maintenance at single CpG dyad resolution, as well as 5mC or 5hmC percent at single base resolution, as is typically obtained in bisulfite sequencing. Additionally, M-H-Dyad-seq and H-M-Dyad-seq allow for the direct detection of differing epigenetic marks on opposing DNA strands, which is not possible with hairpin bisulfite-based techniques.

To validate M-M-Dyad-seq, we compared mESCs grown with or without Decitabine. Decitabine is a cytosine analog known to directly interact with DNMT1, the 5mC maintenance protein, causing its depletion[203]. Treatment with Decitabine for 24 hours caused global loss of 5mC as well as global loss of 5mCpG maintenance, demonstrating that M-M-Dyad-seq can be used to measure changes in 5mC maintenance as well as global 5mC levels (Supplementary Fig. 5.2a,b). Additionally, in both conditions CpHpG maintenance was very low, consistent with DNMT1s preference for CpG sites and the known phenomena that only CpG sites are maintained in mammalian cells (Supplementary Fig. 5.2c).

After validating technique, we applied Dyad-seq to Mouse embryonic stem cells (mESCs) in an *in vitro* model of epigenetic reprograming during changes in pluripotency. mESCs can exist in a primed or naïve state of pluripotency depending on the culture conditions. Primed mESCs exist when cultured in serum containing conditions supplemented with leukaemia inhibitory factor (LIF) (SL), while naïve mESCs exist when cultured in serum free media containing two inhibitors, GSK3i

(CHIR99021) and MEKi (PD0325901), in addition to LIF (2i)[204]. These two states of mESCs are interconvertible, and mESCs in SL are highly methylated and become hypomethylated when transitioned to 2i[155]. There are three potential causes of this demethylation, 1) active demethylation of 5mC to 5hmC induced by ten-eleven translocation (TET) methylcytosine dioxygenases, 2) reduced *de novo* DNA methylation by DNMT3a and DNMT3b, or 3) passive demethylation from lack of 5mC maintenance upon cell division. Indeed, in this system it has been shown that all three play a role, with passive demethylation likely being the main contributor[205–207]. To further investigate the role of passive demethylation in the transition from SL to 2i, we transitioned SL mESCs to media containing each component of the 2i media for 48 hours and performed all sub-types of Dyad-seq.

The change from SL to the basal media of 2i (No) induced spontaneous differentiation and rapid increase in 5mCpG maintenance (Fig. 5.1b). The addition of LIF into this basal media (BL), as well as the addition of both LIF and GSK3i (G) resulted in limited changes to 5mCpG maintenance (Fig. 5.1b). Basal media containing LIF and MEKi (M) induced an even larger decrease in maintenance than 2i media induced (Fig. 5.1b). Interestingly even as the rate of 5mCpG maintenance decreased in M and 2i, it was rare for dyads containing 5mC to be pair with 5hmC (Fig. 5.1c). In fact, 5hmC was found to rarely occur opposing other hydroxymethylated sites (Fig. 5.1d). In contrast to 5hmC/5hmC levels, 5hmC sites had high levels of 5mC on the opposing DNA strand, which varied similarly to the global levels of 5mC among conditions (Fig. 5.1e,f and Supplementary Fig. 5.3a). These results match well with single-molecule fluorescence resonance energy transfer experiments, which while lacking loci specific information, globally identified

roughly 60% of 5hmCs exist in a 5hmC/5mC state in mESC[208]. Some experiments have demonstrated that TET has a high preference for fully methylated sites over hemimethylated ones, yet, contradictory to this, the crystal structure of TET indicates the non-reactive cytosine is not involved in protein DNA contacts[209,210]. Similar levels of 5hmC/5mC sites compared to genome wide 5mC detection implies that in serum grown mESCs, TET has no preference for fully methylated sites over hemimethylated sites (Fig. 5.1e,f and Supplementary Fig. 5.3a).

In greater than 95% of regions, reduced 5mCpG maintenance and reduced 5mCpG percent were co-observed between SL and 2i, indicating that passive demethylation plays a major role in the hypomethylation observed in 2i mESCs (Supplementary Fig. 5.3b). Global levels of 5hmC were similar across conditions, and no increase was seen for M and 2i, providing further evidence that passive demethylation is the key factor regulating the methylome during this transition (Fig. 5.1g and Supplementary Fig. 5.3c). 2i and M involve the inhibition of the MAPK/ERK pathway, which has been previously shown to induce loss of the 5mCpG maintenance protein (DNMT1) in multiple systems, including in mESCs transitioned from SL to 2i[207,211,212]. To investigate other potential causes of this passive demethylation, we performed RNA-seq on all conditions, where we found each condition to be transcriptionally unique (Supplementary Fig. 5.4a,b). We reasoned that since passive demethylation was observed in the M and 2i conditions, but an increase in 5mCpG maintenance was seen in the No condition, putative causal genes could be identified as those which are highly or lowly expressed in M and 2i when compared to No, but are expressed at intermediate levels in G, BL, and SL. Using this criterion, we observe 61 putative genes. Of these, 39 were highly

expressed in the 2i and M condition with enrichment in pathways associated with pluripotency, negative cell cycle regulation, blastocyst development and the regulation viral life cycle (Fig. 5.1h & Supplementary Fig. 5.4c). The remaining 22 genes were highly expressed in the No Condition and were enriched in pathways such as those associated with negative regulation of ERK1 and ERK2 cascade and mesenchymal cell differentiation (Fig. 5.1h & Supplementary Fig 5.4d). Notably, Dppa3 (developmental pluripotency associated 3 gene) was found to be highly expressed in the M and 2i condition (Fig. 5.1h and Supplementary Fig. 5.4b). Previous studies have found that ectopic expression of Dppa3 lead to global hypomethylation, while Dppa3 knockout leads to global hypermethylation[206,213]. Dppa3 has even been shown to directly bind the PHD domain of UHRF1 (Ubiquitin like with PHD and ring finger domains 1), a critical partner of DNMT1 for 5mCpG maintenance, and displaces it from chromatin, thus inhibiting 5mCpG maintenance[206]. These results show that by combining Dyad-seq with RNA-seq it is possible to identify critical factors driving DNA demethylation and thus changes in pluripotency.

While DNA hypomethylation is a global phenomenon in the naïve state, we and others find intracisternal A particles (IAPs) are protected from 5mC erasure in 2i (Supplementary Fig. 5.5)[205,214]. Interestingly, the elevated 5mC levels of IAPs relative to other genomic locations is also true in M and SL, indicating IAPs are associated with high methylation regardless of cell state (Supplementary Fig. 5.5). High levels of methylation are likely due to higher levels of 5mCpG dyad maintenance found in IAPs relative to other regions (Fig. 5.1i). Recent works have found that DNMT1 maintenance is positively impacted by high levels of methylation

surrounding an individual CpG dyad, suggesting a positive feedback loop between 5mC levels and 5mC maintenance[201]. Further investigating the SL condition, we also found that high methylation level led to high 5mCpG maintenance, with a more pronounced affect in regions with higher CpG density such as CpG Islands (Supplementary Fig. 5.6a-d).

Having discovered 5mCpG maintenance can be impacted by regional differences, we hypothesized that histone modifications could play a role in this process. We find that regardless of the histone modification, high methylation rates are associated with high 5mC maintenance (Fig. 5.1j). At low methylation rates we find regions enriched with H3K9me2/3, enhancers marked by H3K4me1 or H3K27ac, and active promoters marked by H3K9ac have increased 5mC maintenance (Fig. 5.1j). Interestingly, Uhrf1 is critical to 5mC maintenance and its ability to bind H3K9me2/3 with high affinity is well established, providing rational for higher maintenance seen in these regions[215–217]. Other interactions with 5mC maintenance machinery may explain elevated maintenance in these other regions, but surprisingly, while H3K4me3 also marks active promoters and is well correlated with H3K9ac, we find that H3K4me3 as well as H3K36me3 and H3K27me3, had similar levels of 5mC maintenance to genome wide levels *(*Fig. 5.1j*)*[218]*.*

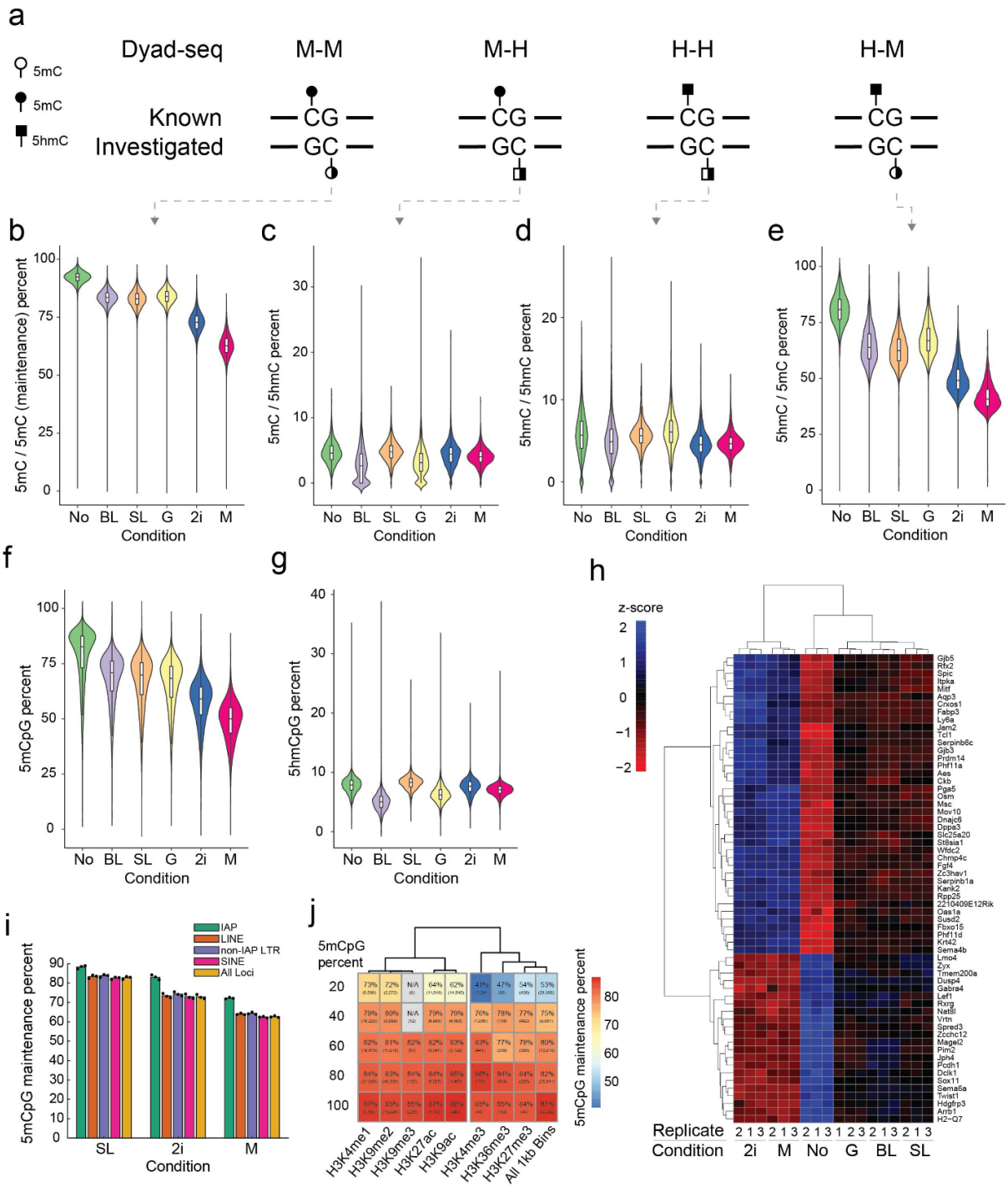


Figure 5.1 | Dyad-seq enables detection of 5mC or 5hmC on both strands of the same piece of DNA.

(**a**) Schematic describing 4 versions of Dyad-seq. M-M-Dyad-seq profiles 5mC on one strand and C or 5mC on the opposing strand. M-H-Dyad-seq profiles 5mC on one strand and C or 5hmC on the opposing strand. H-H-Dyad-seq profiles 5hmC on one strand and C or 5hmC on the opposing strand. H-M-Dyad-seq profiles 5hmC on one strand and C or 5mC on the opposing strand. (**b**) 5mCpG

maintenance as detected by M-M-Dyad-seq. (**c**) The percent of 5mCpGs Dyads with 5hmC on the opposing DNA strand detected by M-H-Dyad-seq. (**d**) The percent of 5hmCpGs Dyads with 5hmC on the opposing DNA strand detected by H-H-Dyad-seq. (**e**) The percent of 5hmCpGs Dyads with 5mC on the opposing DNA strand detected by H-M-Dyad-seq. (**f**) M-M-Dyad-seq detected 5mC in a CpG context from non-Dyad sites. (**g**) M-H-Dyad-seq detected 5hmC in a CpG context from non-Dyad sites. (**b,f**) 100 Kb bin size, (**c-e,g**) 1 Mb bin size. (**h**) Heatmap of differentially expressed genes with a putative role in passive 5mCpG demethylation. (**i**) Detection of maintenance at repetitive elements as detected by M-M-Dyad-seq. (**j**) Detection of 5mCpG maintenance for levels of non-dyad 5mCpG percent at regions enriched for various histone marks. Bracketed numbers indicate total number of regions analyzed.

2. *mESC display heterogenous 5mC maintaince based on their transcriptional state*

The development of Dyad-seq allowed us to characterize the heterogeneity in 5mC maintenance across the genome with the resolution of a single CpG site, but the heterogeneity of 5mCpG maintenance within a sample is still difficult to quantify. M-M-Dyad-seq has some similarities to our previous sequencing technique, scMspJI-seq, thus we hypothesized that unlike currently available hairpin bisulfite techniques, M-M-Dyad-seq could be scaled down to the single cell level[65]. Indeed, single cell 5mC/5mC Dyad-seq (scDyad-seq) is possible and to validate its accuracy, cells treated with Decitabine for 24 hours were analyzed. Most cells experiencing global loss of 5mC and extremely low 5mCpG maintenance but limited changes in 5mCpHpG, validating the method (Supplementary Fig. 5.7a,b).

Other single-cell workflows have benefited greatly from capturing the transcriptome in addition to the epigenome[85,92]. Here in bulk, we also have shown the power of transcriptome analysis when paired with Dyad-seq, so after validating scDyad-seq works in single-cells, we further enhanced the technique by simultaneously identifying the transcriptome (scDyad&T-seq) using a post *in vitro* transcription amplified mRNA enrichment strategy we previously developed in

scMAT-seq (chapter 3). We applied scDyad&T-seq to 106 serum grown mESCs cells, detecting on average 24,384 transcripts, covering 49,626 5mCpG dyads and covering an additional 317,088 CpGs where there is no associated dyad information (Fig. 5.2a). Surprisingly, even when considering regions of similar 5mC content, very high heterogeneity is observed in both 5mCpG maintenance and non-dyad 5mC levels, indicating single-cell resolution can be highly beneficial (Fig. 5.2b). Not surprisingly, since 5mC plays a key role in cell identity and gene expression, we also identify heterogeneity in gene expression.

Using the transcriptome, we identify two subpopulations in our serum grown mESCs, one population that is high in Nanog, Rex1, and Esrrb (Nanog High) and one that is low in these genes (Nanog Low) (Fig. 5.2c, Supplementary Fig. 5.8 and 5.9). It is well established that mESCs grown in serum are transcriptionally heterogenous with bimodal expression of key pluripotency genes, but how their epigenomes regulate these differences is less well studied[85,219]. We find that the Nanog High population is globally hypomethylated and has lower 5mCpG maintenance compared to the Nanog Low group, p-value $8x10^{-6}$ and $3.6x10^{-4}$ respectively (Fig. 5.2.d,e). mESCs can switch from a high to low Nanog state or vice versa stochastically with low frequency[219]. Our results suggest that this stochastic event is connected with changes in DNA methylation, which in turn are established in part by underlying rates of 5mCpG maintenance.

In addition to cell identity, here we have found that histone modifications can impact 5mCpG maintenance. Consistent with bulk findings, H3K9me2/3, H3K4me1, H3K27ac, and H3K9ac tend to have high maintenance related to their 5mC level (Fig 5.2.d,e). Surprisingly, regardless of histone modification, Nanog High cells have

lower 5mCpG maintenance, revealing that cell state is extremely important in governing passive demethylation and can at least in part overcome potential positive interactions between histone marks and DNA maintenance machinery (Fig 5.2.d,e).

While here we have used scDyad&T-seq to understand the heterogeneity of 5mCpG Maintaince in serum grown mESCs cells, this heterogeneity has been previously explored using scMspJI-seq and single cell bisulfite sequencing[65]. These techniques use strand specific 5mC information to calculate strand bias, which is the number of methylated cytosines on the plus strand divided by the total detected in a region[65]. As such, deviations from a strand bias of 0.5 indicate passive demethylation is occurring. Because strand bias is a regional technique, it is inherently limited in its detection of demethylation. Additionally, in most cells it is difficult to discriminate between maternal and paternal DNA which can result in strand bias being obscured and results in an inability to detect passive demethylation in bulk. Strand bias in scDyad&T-seq is equivalent to scMspJI-seq and so we compared the resolution afforded by strand bias vs 5mCpG maintenance. Generally, strand bias does a good job in detecting cells that are experiencing a high level of passive demethylation but struggles to detect cells with more modest passive demethylation (Fig. 5.2f). For one cell (P7L4.78) experiencing high levels of demethylation (average 5mCpG maintenance of 30.7%), very little strand bias was found, while another cell (P7L3.69) experiencing similar levels of demethylation (average 5mCpG maintenance = 25.4%) the cell was found to be highly biased (Fig. 5.2g).

In good agreeance between both measurements, in most serum grown mESCs with high 5mCpG maintenance, no strand bias was observed (Fig. 5.2f and

Supplementary Fig. 5.10a,b). Additionally, a computational method for estimating passive demethylation from nucleobase conversion based 5mC sequencing reads also trended well with the measured 5mCpG maintenance (Supplementary Fig. 5.10c). Together these results indicate that scDyad&T-seq is a highly sensitive method for assessing the maintenance status of individual 5mCpG dyads, the methylation level of non-dyad interrogated CpGs and the transcriptome from the same single cell, allowing for the connection between epigenetic dynamics and cell state.

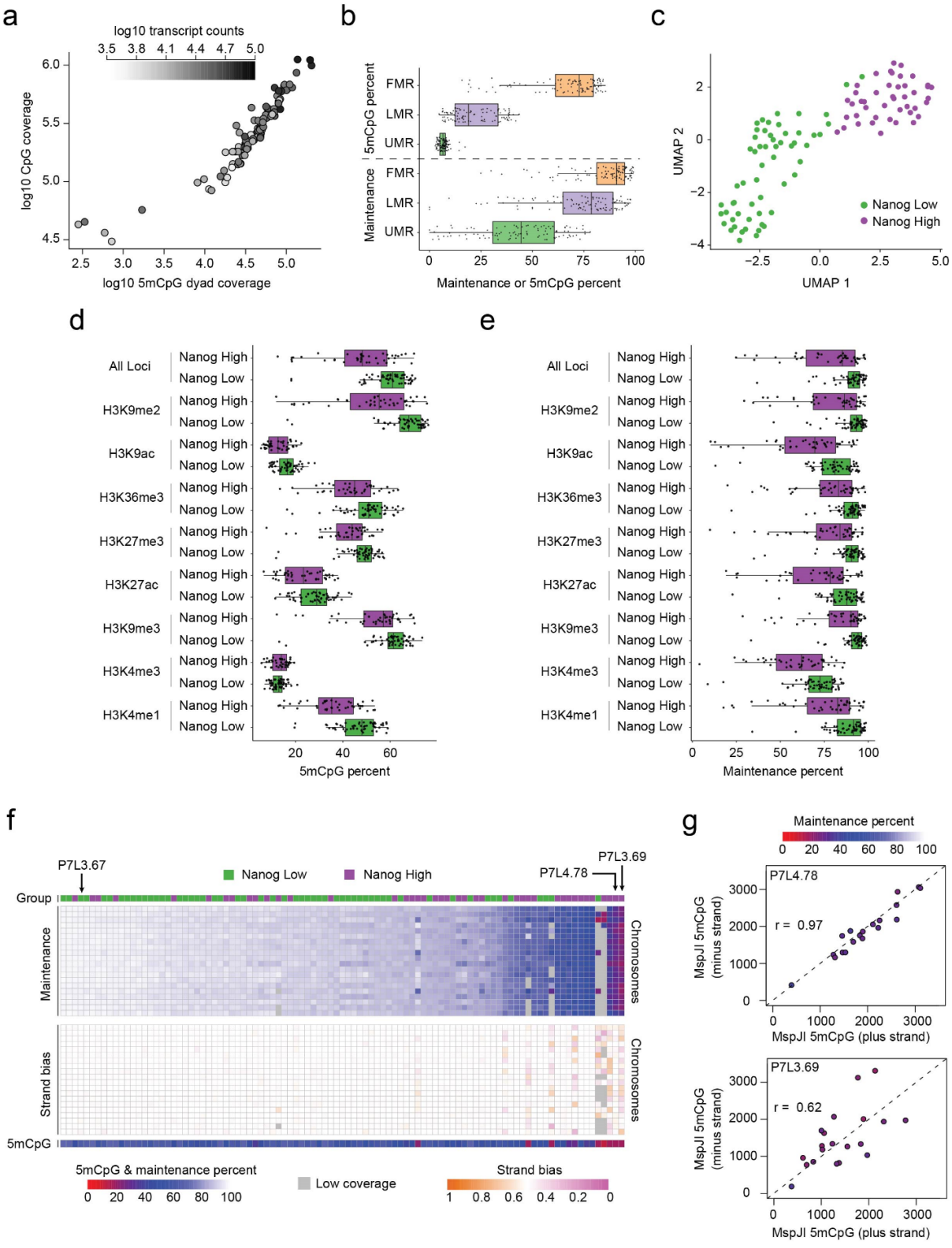


Figure 5.2 | scDyad&T-seq connects cell identity to demethylation dynamics in single cells.

(**a**) Coverage of 5mCpG dyads, non-dyad CpGs, and total transcripts in single cells. (**b**) Methylation and 5mCpG maintenance levels at fully methylated regions (FMR), lowly methylated regions (LMR), and unmethylated regions (UMR) as observed by Stadler *et al*., each dot represents detection in one cell[220]. (**c**) UMAP visualization of serum grown mESCs based on the single-cell transcriptomes obtained from scDyad&T-seq. (**d**) Non-dyad 5mCpG levels in regions of enriched histone marks. (**e**) 5mCpG maintenance levels in regions of enriched histone marks. (d,e) Cells split by groupings in (b). (**f**) Heatmap of 5mCpG maintenance by chromosome indicates an increased level of sensitivity in detecting demethylation when compared to strand bias for the same cell. Transcriptional group and genome wide 5mCpG methylation levels are also reported for the same cells. (**g**) MspJI detection on each strand of a chromosome for two passively demethylating cells from (h), cell P7L4.78 and P7L3.69. A low Persons correlation indicates deviations from a strand bias of 0.5. Point color describes the detected 5mCpG maintenance percent.

3. Heterogenous loss of 5mC maintaince is observed when mESC transition to the naïve state

To further investigate the effect of cell state on DNA maintenance, we applied scDyad&T-seq to mESCs 3, 6, and 10 days after transitioning from serum containing media (Serum) to 2i media (2iD3, 2iD6, and 2iD10 respectively). While 2i cells are remarked for their homogeneity relative to serum grown mESCs, surprisingly we found that the epigenetic reprograming to the naïve state is highly heterogenous with many cells retaining high levels of methylation and 5mCpG maintenance even after 10 days in 2i (Fig 5.3a,b). Using hierarchical clustering, cells were classified as highly or lowly methylated (mC^{Hi} and mC^{Lo}), and highly or lowly maintained (Mnt^{Hi} and Mnt^{Lo}), leading to 4 distinct categories of methylation state (Supplementary Fig. 5.11 and 5.12). We find that during the primed to naïve transition, cells begin generally highly methylated and highly maintained. Passive demethylation then begins, and cells lose 5mC until they reach a lowly methylated and lowly maintained phase, after which, cells remain lowly methylated but the 5mC that still remains is highly maintained (Fig. 5.3c,d). Consistent with this result, regions previously identified as retaining high methylation in the 2i state consistently have higher methylation and maintenance in all timepoints even in serum grown cells[155] (Fig

5.3e,f). These regions were shown to correlate with the presence of H3K9me3, which is located in highly similar regions for serum and 2i grown mESC[155,221]. Together this, along with our previously identified correlation between H3K9me3 and high 5mCpG maintenance, these results indicate that the global impairment to 5mCpG maintenance machinery in 2i grown mESC is partially restored in regions of H3K9me3 likely due to interactions with UHRF1 and enhanced activity of DNMT1 at regions containing high 5mC.

While large epigenetic heterogeneity was observed, we find most cells in 2i are transcriptionally similar and distinct from serum grown cells (Supplementary Fig. 5.13a). Cells grown in 2i did not strongly separate by timepoint, indicating the naïve transcriptional programing is quicky activated once the media is changed (Fig. 5.3g & Supplementary Fig. 5.13a-e). While two broad transcriptional groups were observed, further clustering revealed 4 populations, with two groups related to serum grown cells as discussed previously, Nanog low serum (Nan^{Lo}) and Nanog high (Nan^{Hi}), and two groups related to 2i cells (2i-1 and 2i-2) (Fig. 5.3h).

Interestingly, a handful of cells in the 2iD3 condition cluster with the Nanog low SL cell population and are likely derived from these cells. Nanog low serum grown cells are less likely to survive the transition to 2i when compared to Nanog high cells[222]. This low survival rate is likely due to an epigenetic barrier, as this 2iD3 group contains high 5mCpG maintenance and thus retains high 5mC relative to successfully transitioned 2iD3 cells (Supplementary Fig. 5.14a,b). These non-transitioning cells express low levels of Pou5f1, a similar phenomenon found in other recent work describing these non-transition cells (Supplementary Fig. 5.14c)[222]. In addition, they appear to have undergone spontaneous differentiation to

neuroectoderm lineage and are expression Sox1, but do not survive long term culture and are no longer present by day 10 (Fig. 5.3g,h and Supplementary Fig. 5.14d).

Those cells which did successful transfer into 2i culture could be split into two transcriptionally distinct populations, 2i-1 and 2i-2. The 2i-2 population highly expressed endogenous retrovirus RLTR45 and RLTR45-int, as well as Khdc3 (also known as Filia) known to be involved in safeguarding genomic integrity of mESCs and preimplantation embryos[223,224] (Fig. 5.3i). Additionally, in this same population, we found high expression of Dppa3, which as previously discussed has been implemented in DNA demethylation (Fig. 5.3i). Together this suggests the 2i-2 transcriptional population is experiencing high levels of DNA demethylation resulting in the loss of retroviral silencing and genomic instability, likely due to loss of DNA maintenance. Indeed, 2i-2 cells are highly demethylated while the 2i-1 population of cells, which expresses these transcripts at lower levels retains mainly highly methylated (Fig. 5.3j). The two populations also drastically differ in their 5mCpG maintenance with the 2i-2 having lower global 5mCpG maintenance (Fig. 5.3k). Surprisingly, the two transcriptional populations within the 2i cells appear to be only weakly related to the culturing duration in 2i, as we observe a slight bias for later timepoints to be in the 2i-2 transcriptional group (Supplementary Fig. 5.15a). Instead, as the global levels suggest, we find that the 2i-2 transcriptional group is largely comprised of cells in a methylation state identified as mC^{Lo} and Mnt^{Lo}, while the 2i-1 group still consists of many cells in a mC^{Hi} and Mnt^{Hi} state (Fig. 5.3l). Because we see that the methylation state of mC^{Hi} and Mnt^{Hi} is reduced over time, this suggests that the transcripts expressed by the 2i-2 transcriptional group maybe

important in the long term reprograming to the naïve pluripotency state (Fig. 5.3d). While it is well known that DNA methylation and gene expression become highly decoupled during the transition to the naïve state, we find that this is not true for at least 22 genes and 2 transposable elements (Supplementary Fig. 5.15b). While the transcriptional differences between the two populations are fairly limited, further study is needed to investigate if the epigenetic differences observed in 2i cells lead to differences in differentiation potential and cell fate outcome, as was seen here for the Nanog low serum cells when transitioned to 2i.

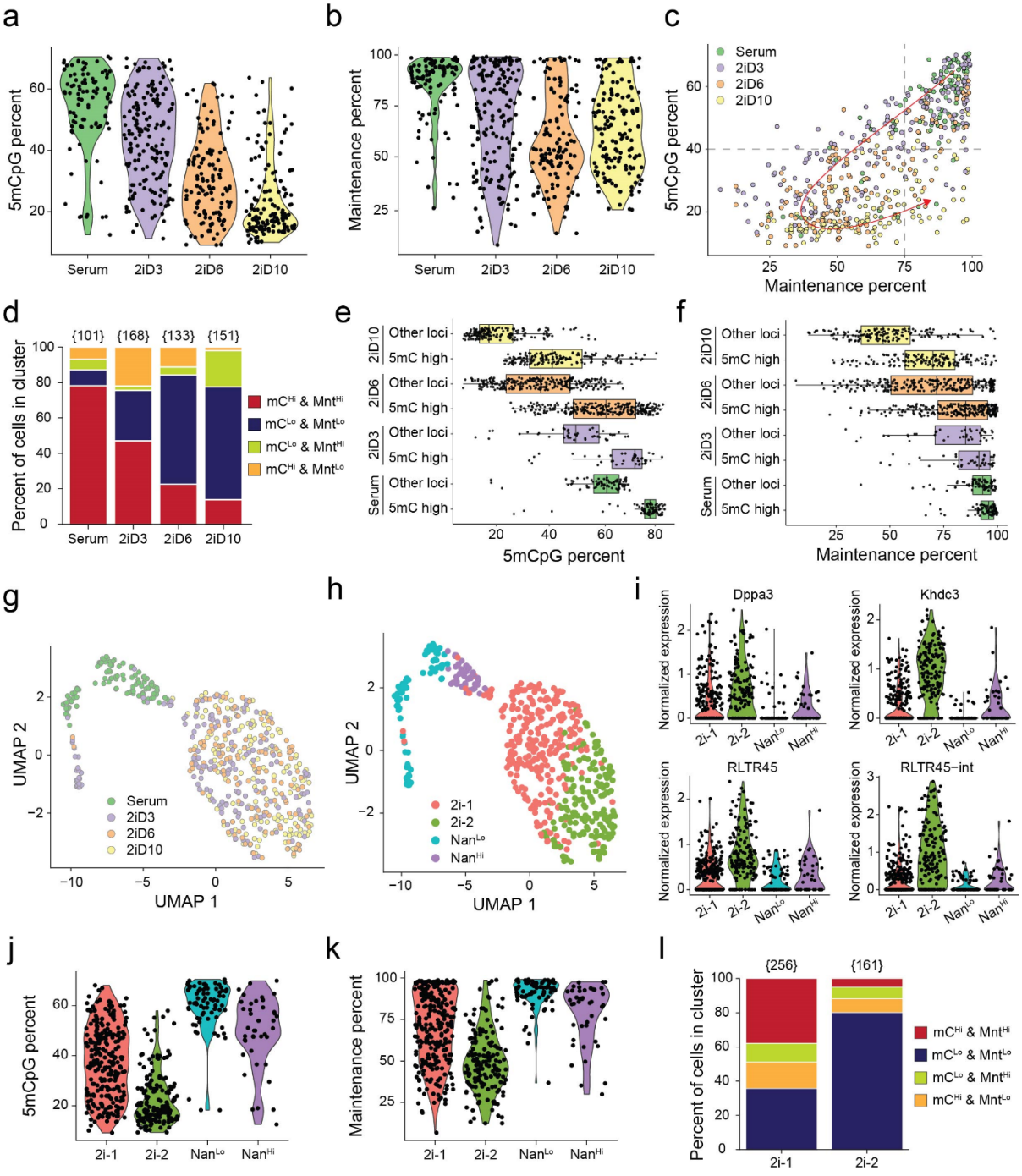


Figure 5.3 | The transition to naïve pluripotency involves transient loss of 5mCpG maintenance, leading to expression of select genes and transposable elements.

(**a,b**) 5mCpG methylation (**a**) and 5mCpG maintenance (**b**) levels for each cell while transitioning to 2i conditions. (**c,d**) Cells transition from highly methylated and highly maintained to a lowly methylated and highly maintained state. (**c**) Overall 5mCpG methylation and 5mCpG maintenance levels for each cell while transitioning to 2i conditions. (**d**) Cells discriminated by measured 5mC dynamics while transitioning to 2i conditions, as described by Supplemental Fig. 5.11 and 5.12. (**e,f**)

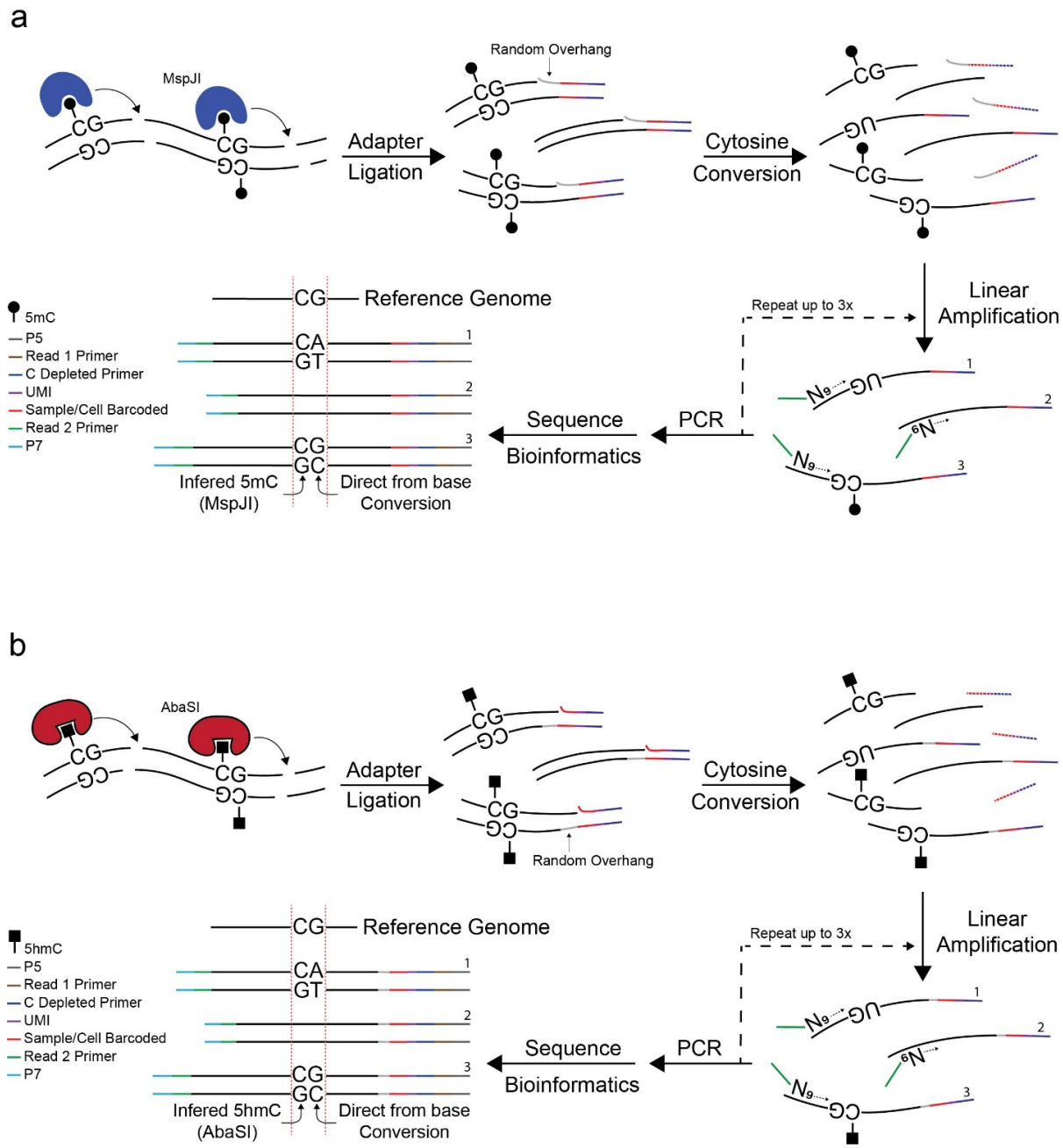
Methylation and 5mCpG maintenance levels at region of high DNA methylation in mESCs grown in long term in 2i (5mC high) as observed by Habibi *et al*. and at all other regions (Other loci), each dot represents detection in one cell[155]. (**g,h**) UMAP visualization of cells transiting to 2i based on the single-cell transcriptomes obtained from scDyad&T-seq, discriminated by time in 2i (**g**) or by transcriptome-based clustering (**h**). (**i**) Gene expression of select genes and transposable elements found highly expressed in the 2i-2 population, Dppa3, Khdc3, RLTR45, and RLTR45-int. (**j,k**) 5mCpG methylation (**j**) and 5mCpG maintenance (**k**) levels for each cell, discriminated by transcriptional cluster shown in (h). (**i**) Measured 5mC dynamics, as described by Supplemental Fig. 5.11 and 5.12, for cells in the 2i transcriptional populations 2i-1 and 2i-2 described in (h). Bracketed numbers indicate total number of cells in that group (d,i).

C. Conclusion

In summary, we developed Dyad-seq which is a general technique for profiling epigenetic marks on opposing DNA strands. One sub type of Dyad-seq, H-M-Dyad-seq was used to observed that 5hmCs are commonly found duplexed with 5mC, an observation that could not be detected with conventional hairpin bisulfite techniques. We then focused our attention to M-M-Dyad-seq and the transition of mESCs from serum to 2i. Here together with RNA-seq, we identified putative genes likely responsible for the passive loss of 5mC. One such gene was Dppa3, which has previously been implemented in this transition and is well known to directly interact with UHRF1, a critical piece of the 5mC maintenance machinery[206]. We then developed scDyad&T-seq by scaling down M-M-Dyad-seq and simultaneously implementing RNA seq on the same single-cells. Using scDyad&T-seq we identified two populations in serum grown mESCs differing in DNA maintenance, 5mC levels, as well as gene expression. Furthermore, when using scDyad&T-seq on mESCs transitioning from serum to 2i, we found the epigenetic reprograming process to be highly heterogenous, with transcription being decoupled from DNA methylation except for a select handful of genes including Dppa3. Further, we show that in addition to cell identity, DNA methylation levels, and histone modifications can

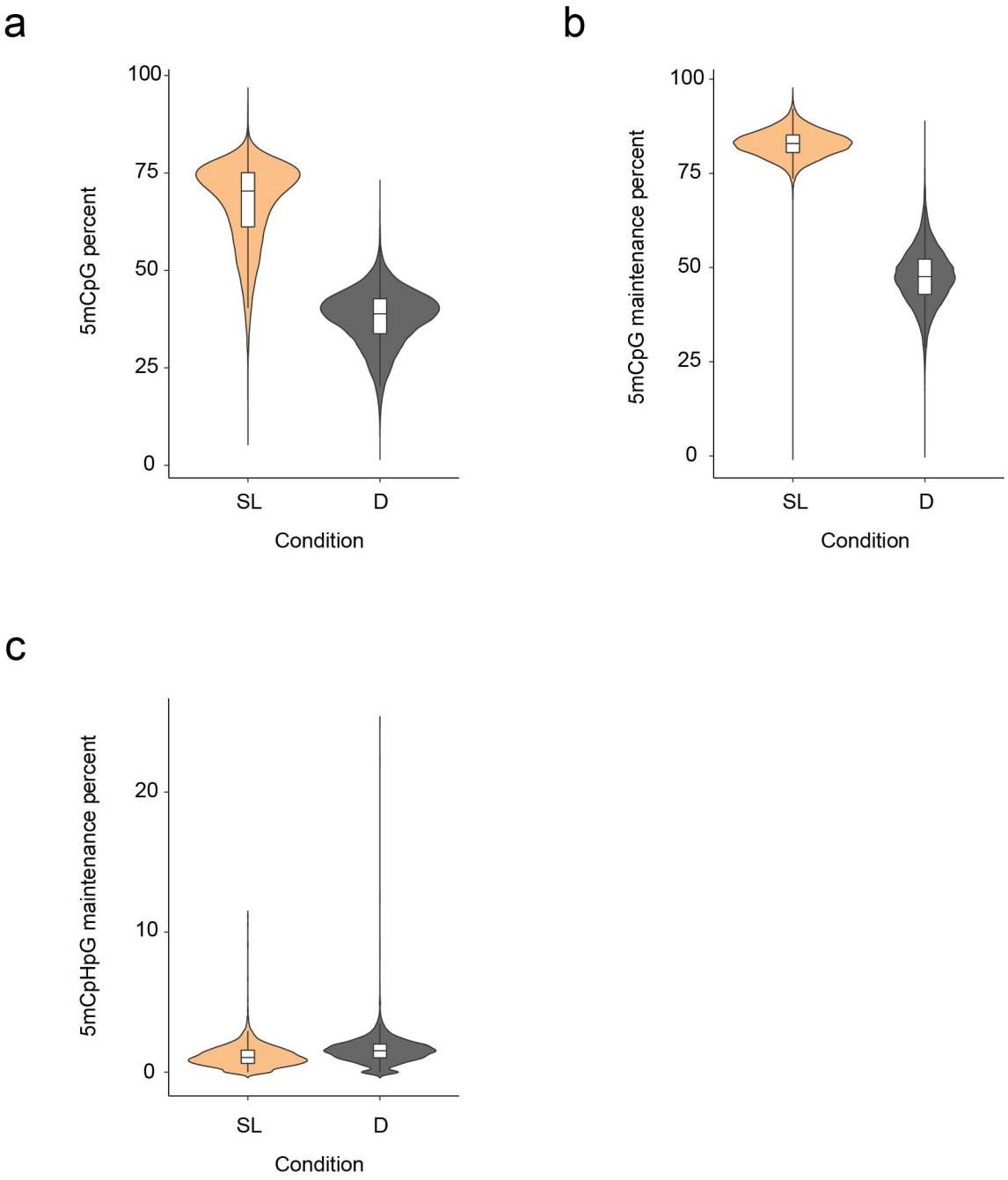
influence the accuracy of the 5mCpG maintaince machinery. Specifically, during the transition to 2i, we find that mESCs retain high 5mC in some areas because of high levels of DNA maintenance which are in part elevated due to H3K9me3 enrichment in these areas[155,221]. Overall, scDyad-seq is an enhancement to both scMspJI-seq and scBS-seq, enabling high resolution detection of passive demethylation and global 5mC levels in thousands of single-cells and optionally with scDyad&T-seq, the method can be easily adapter to obtain transcriptional data from the same cells simultaneously (Supplementary Fig. 5.16a,b).

D. Supplementary figures



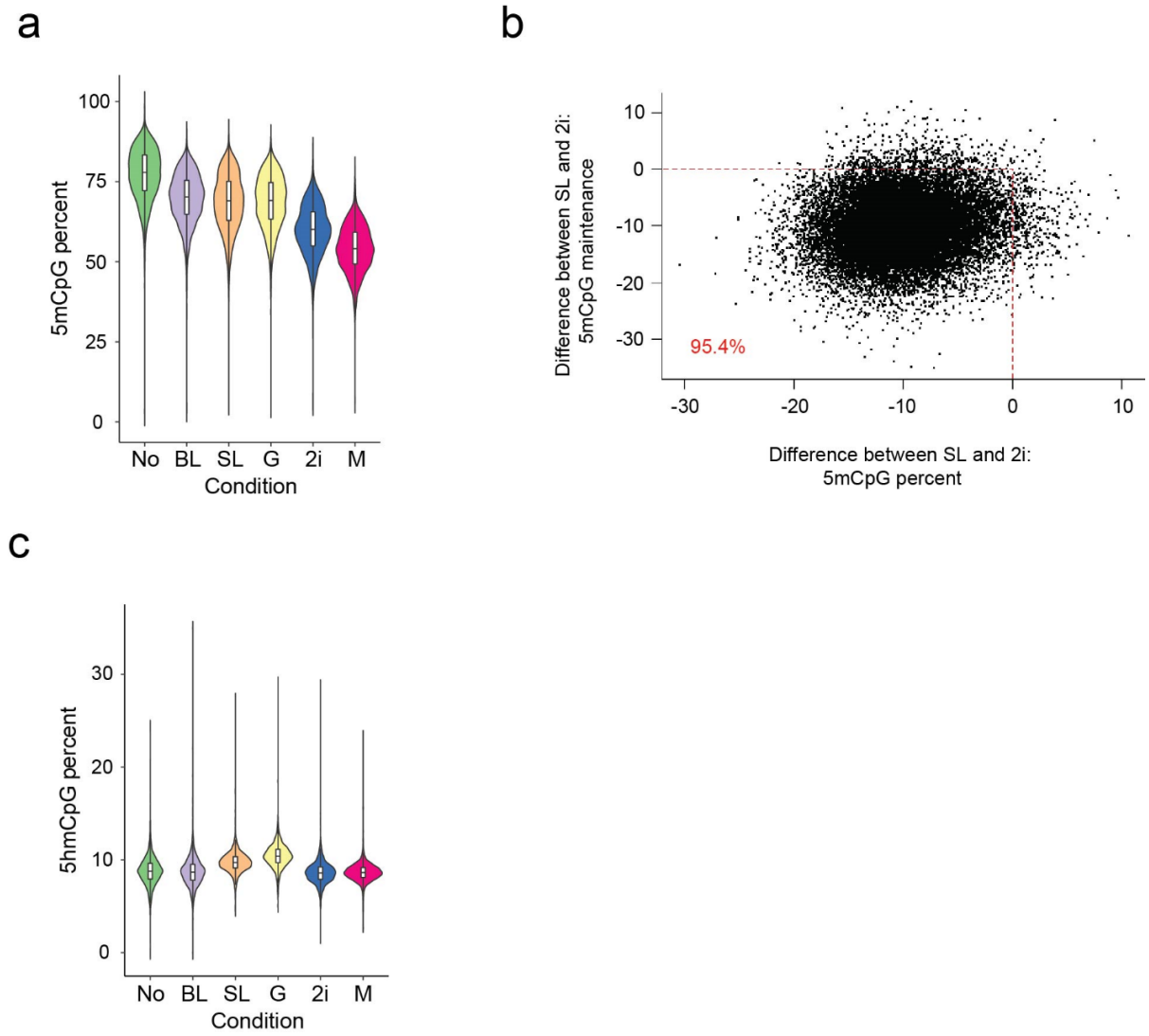
Supplementary Figure 5.1 | Schematic of Dyad-seq methodologies.

(**a**) Describes detection of CpG dyads through MspJI digestion and subsequent nucleobase conversion as used in M-M-Dyad-seq, M-H-Dyad-seq and scDyad&T-seq. (**b**) Describes detection of CpG dyads through AbaSI digestion and subsequent nucleobase conversion as used in H-M-Dyad-seq and H-H-Dyad-seq.



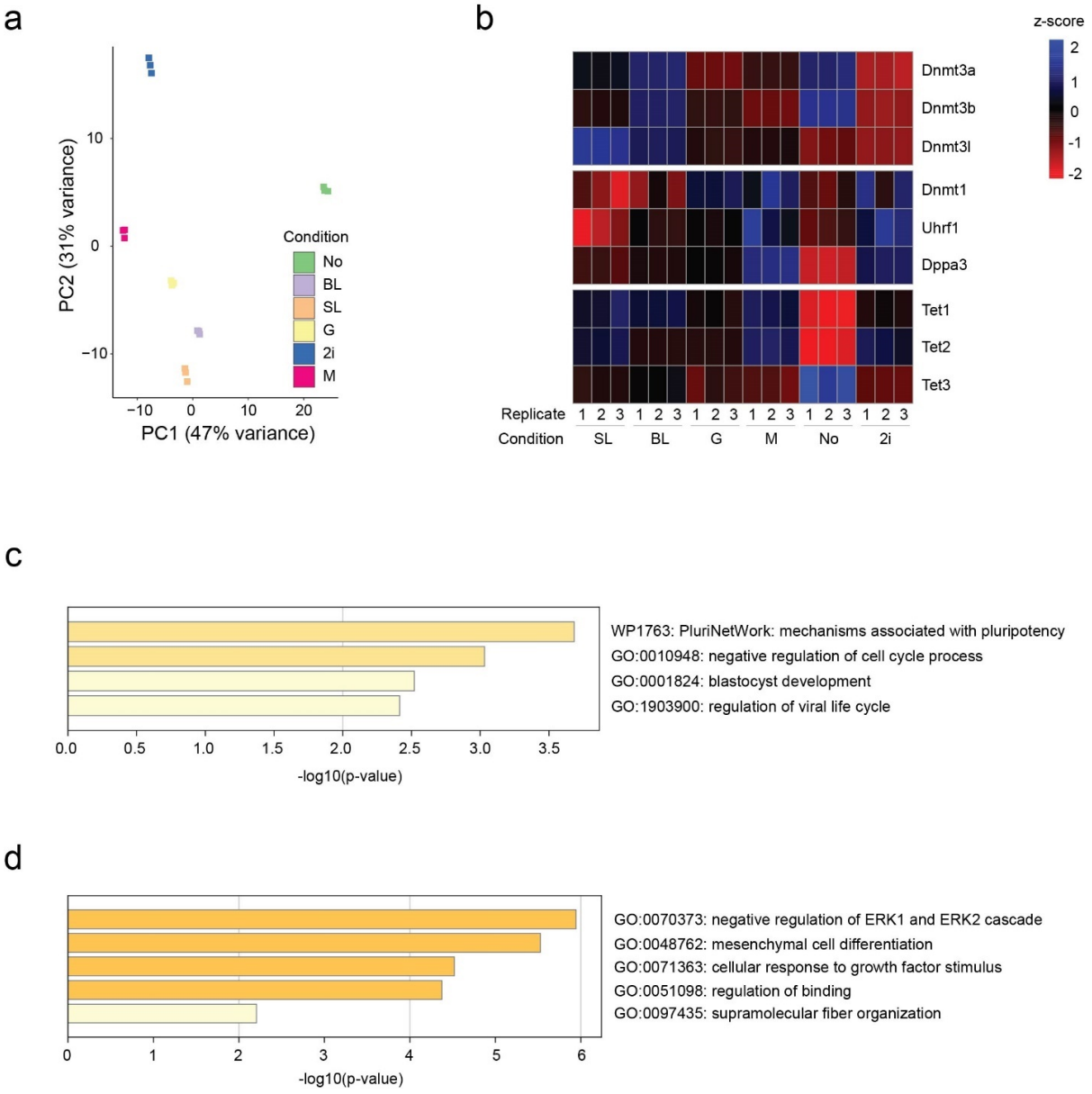
Supplementary Figure 5.2 | M-M-Dyad-seq accurately captures demethylation induced by 24-hour Decitabine treatment.

(**a**) M-M-Dyad-seq detected 5mC in a CpG context from non-Dyad sites for SL and Decitabine treated (D) mESCs. (**b**) 5mCpG maintenance as detected by M-M-Dyad-seq. (**c**) 5mCpHpG maintenance as detected by M-M-Dyad-seq.



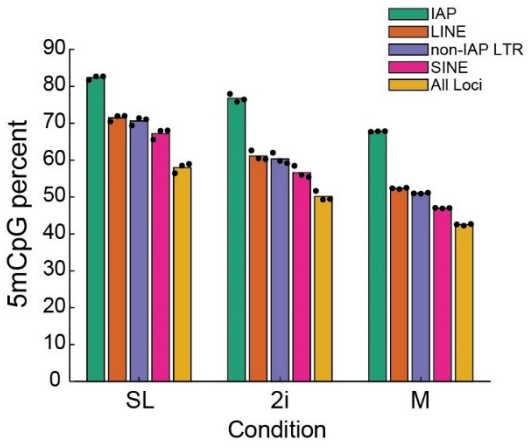
Supplementary Figure 5.3 | Global epigenetic response to 48-hour mESC growth in 2i media components.

(**a**) H-M-Dyad-seq detected 5mC in a CpG context from non-Dyad sites. (**b**) Reduction of 5mCpG maintenance and non-Dyad 5mCpG percent is co-observed when transition from SL to 2i, each dot represents genomic tilling of 100 kb. (**c**) H-H-Dyad-seq detected 5hmC in a CpG context from non-Dyad sites.



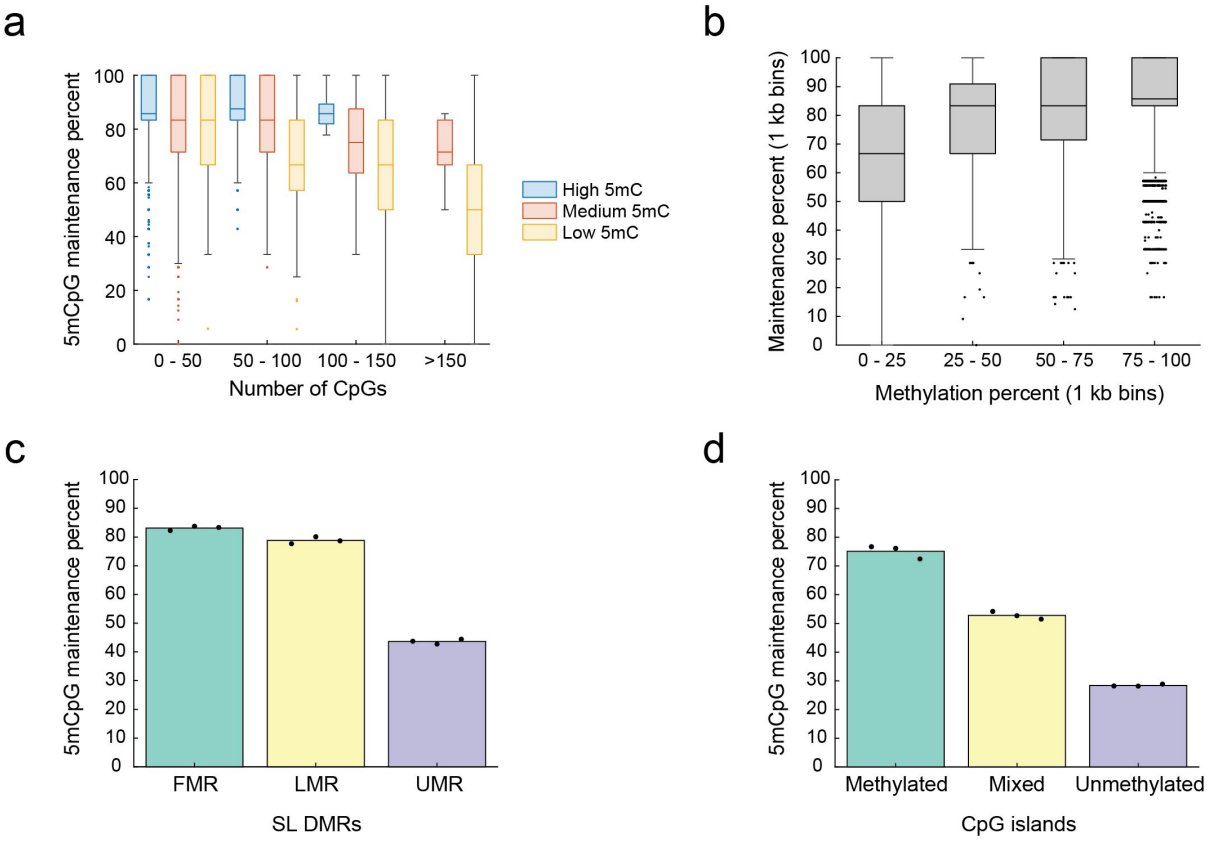
Supplementary Figure 5.4 | Transcriptome response to 48-hour mESC growth in 2i media components.

(**a**) RNA-seq principal component plot. (**b**) Heatmap of genes related to active and passive demethylation processes. (**c,d**) Gene enrichment analysis for differentially expressed genes performed using Metascape[225]. (**c**) Gene set that was highly expressed in the 2i and M condition, lowly expressed in No, and not differentially expressed across SL, BL, and G. (**d**) Gene set that was highly expressed in the No condition, lowly expressed in 2i and M, and not differentially expressed across SL, BL, and G.



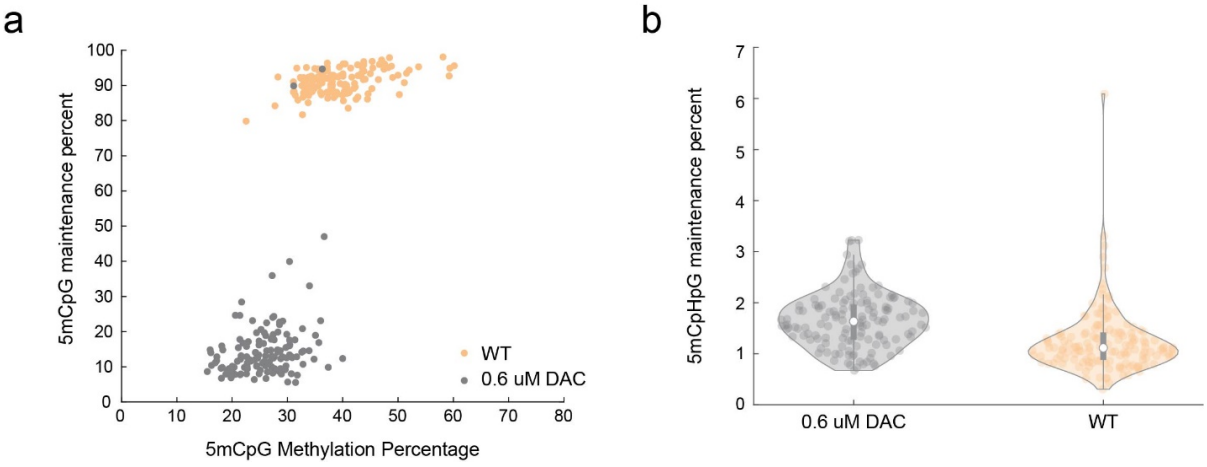
Supplementary Figure 5.5 | IAPs are resistant to demethylation in the primed to naïve pluripotency transition.

Detection of non-dyad 5mCpG percent at repetitive elements after 48-hours in the indicated condition as detected by M-M-Dyad-seq.



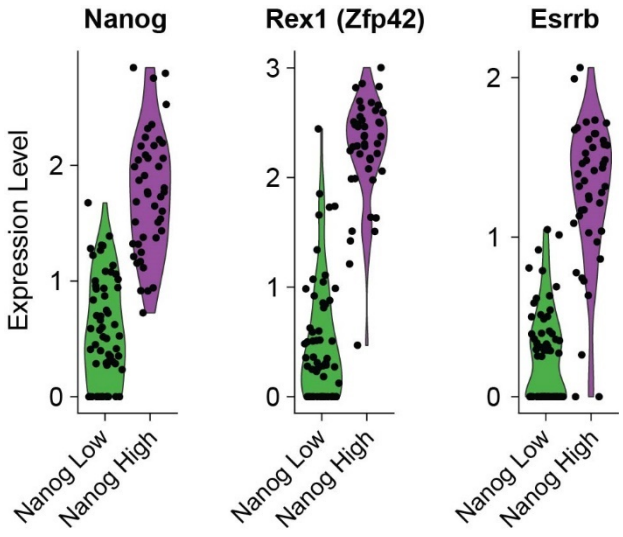
Supplementary Figure 5.6 | High 5mCpG maintenance occurs in regions with high 5mC levels.

(**a**) 5mCpG maintenance in 1 kb genomic bins split by the number of CpGs in the bin as well as the detected non-dyad 5mC level. Low 5mC indicates methylation levels lower than 20%, medium 5mC indicates levels between 20% and 80%, and high 5mC indicates levels greater than 80%. (**b**) 5mCpG maintenance in 1 kb genomic bins split by non-dyad 5mC level. (**c**) 5mCpG maintenance at fully methylated regions (FMR), lowly methylated regions (LMR), and unmethylated regions (UMR) as observed by Stadler *et al.*[220]. (**d**) 5mCpG maintenance at CpG islands split by non-dyad 5mC levels. Methylated indicates CpG islands with greater than 20% methylation, mixed indicates methylation levels between 10 and 20%, and unmethylated indicates CpG islands with less than 10% methylation.



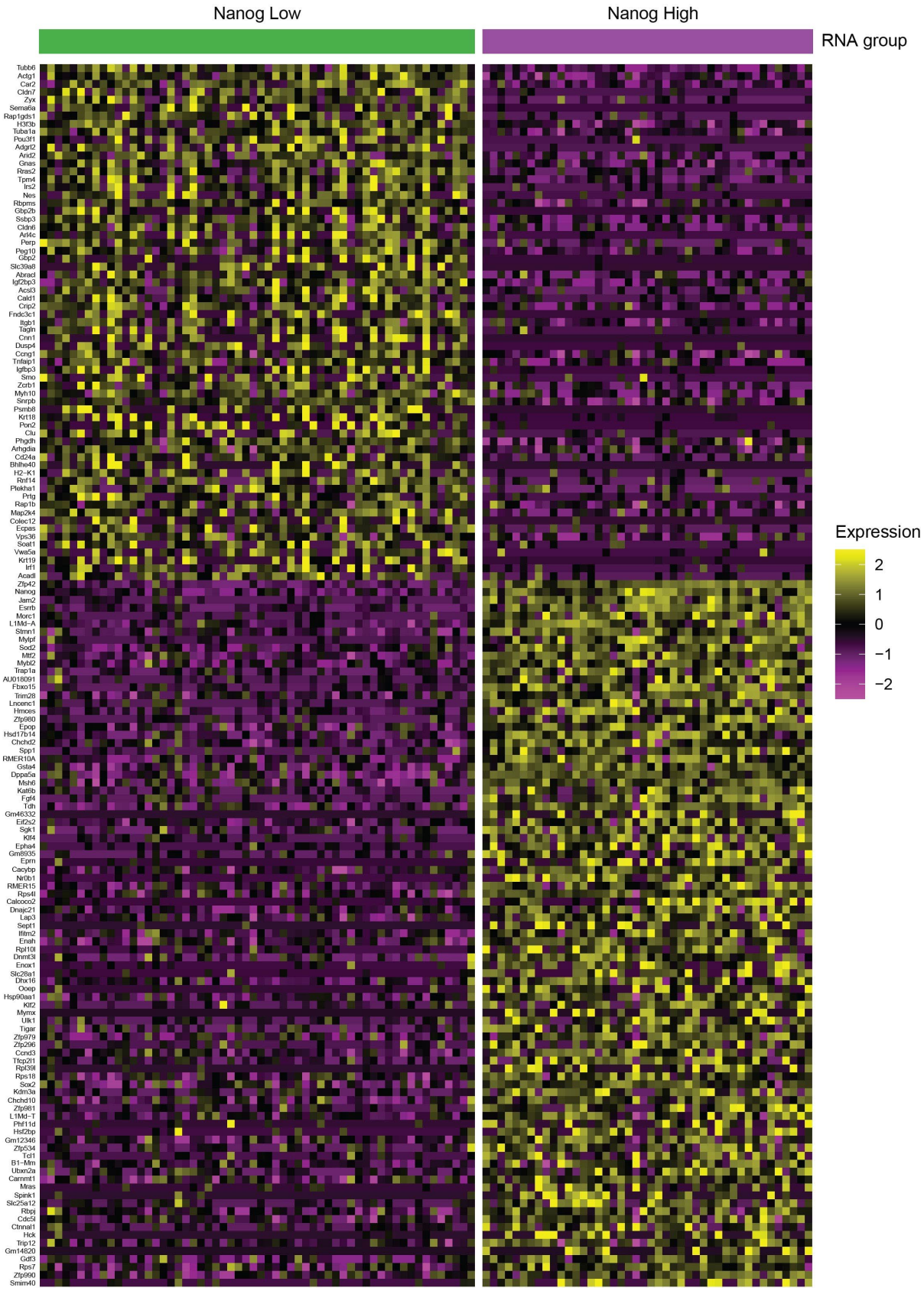
Supplementary Figure 5.7 | scDyad-seq accurately captures demethylation induced by 24-hour Decitabine treatment in single K562 cells.

(**a**) Methylation and maintenance levels of single-cells with or without 24-hour 0.6 μM Decitabine treatment. (**b**) 5mCpHpG maintenance levels of single-cells with or without 24-hour 0.6 μM Decitabine treatment.



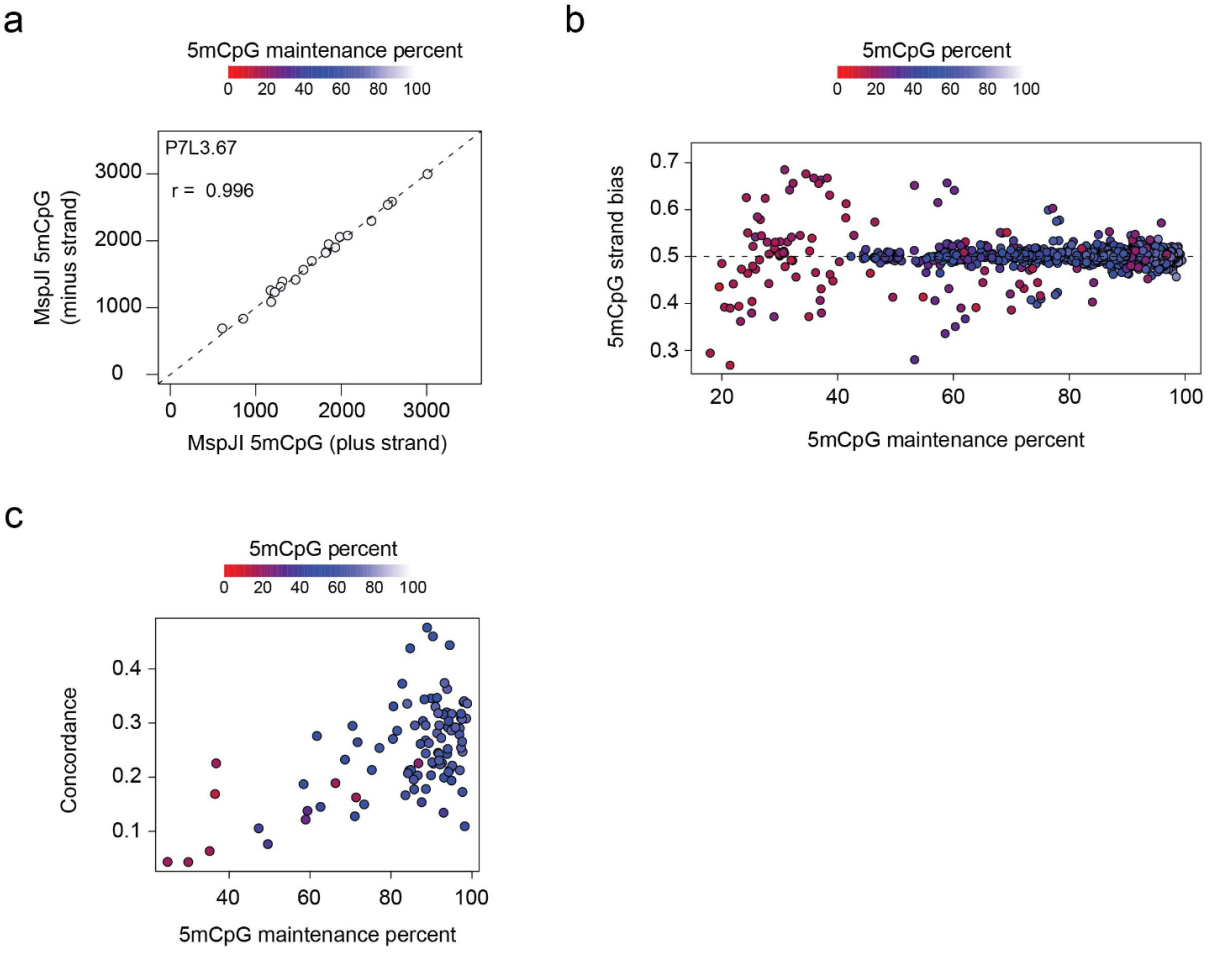
Supplementary Figure 5.8 | Serum grown mESCs are heterogenous, containing a Nanog high and a Nanog low population of cells.

Detection of Nanog, Rex1, and Esrrb expression in the two classified groups based on gene expression (Fig. 5.2c).


Nanog Low
Nanog High
RNA group
Expression
2
1
0
−1
−2

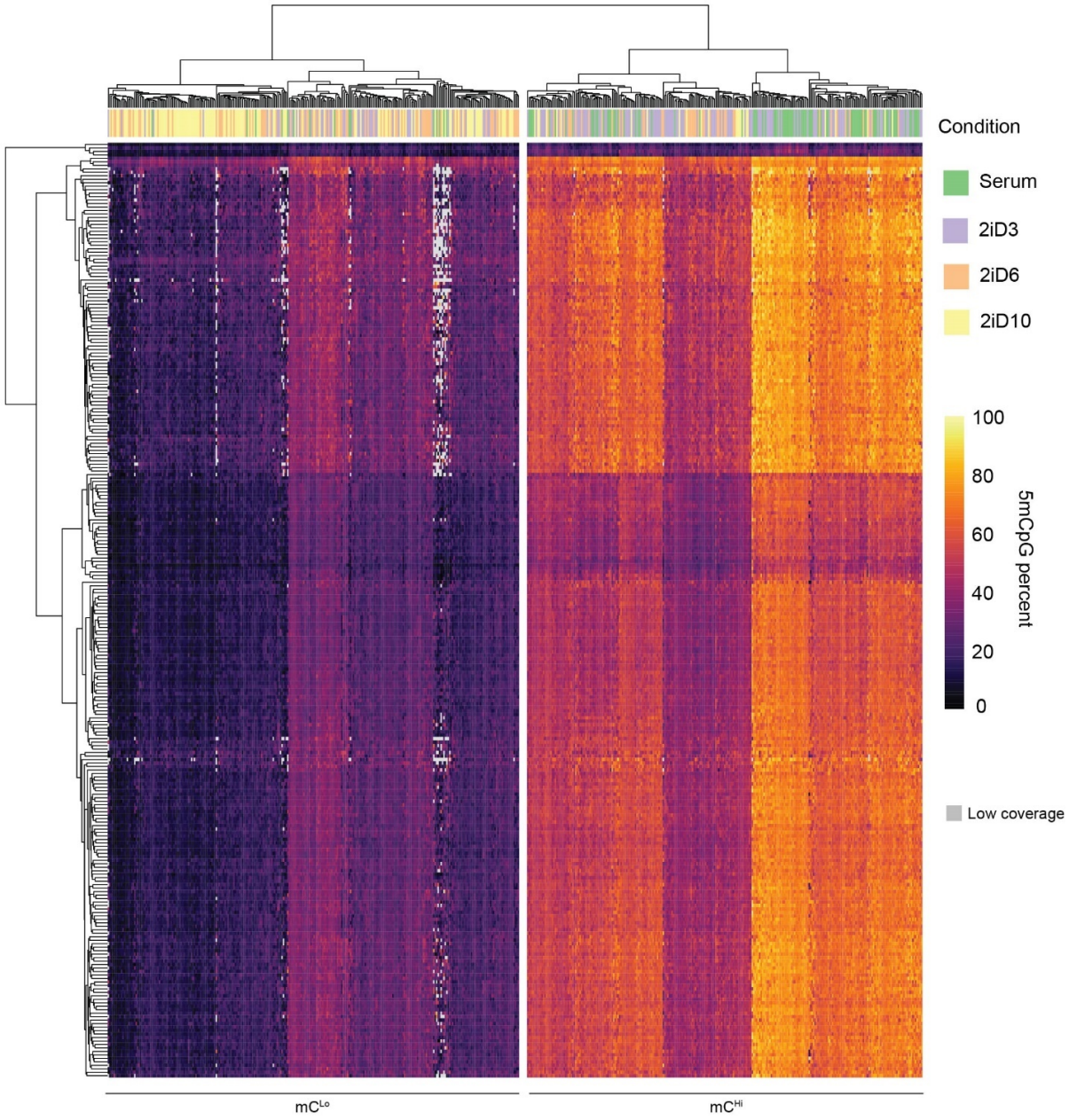
Supplementary Figure 5.9 | Serum grown mESCs contain two subpopulations.

Differentially expressed genes in the two classified groups detected in serum grown mESCs (Fig. 5.2c).



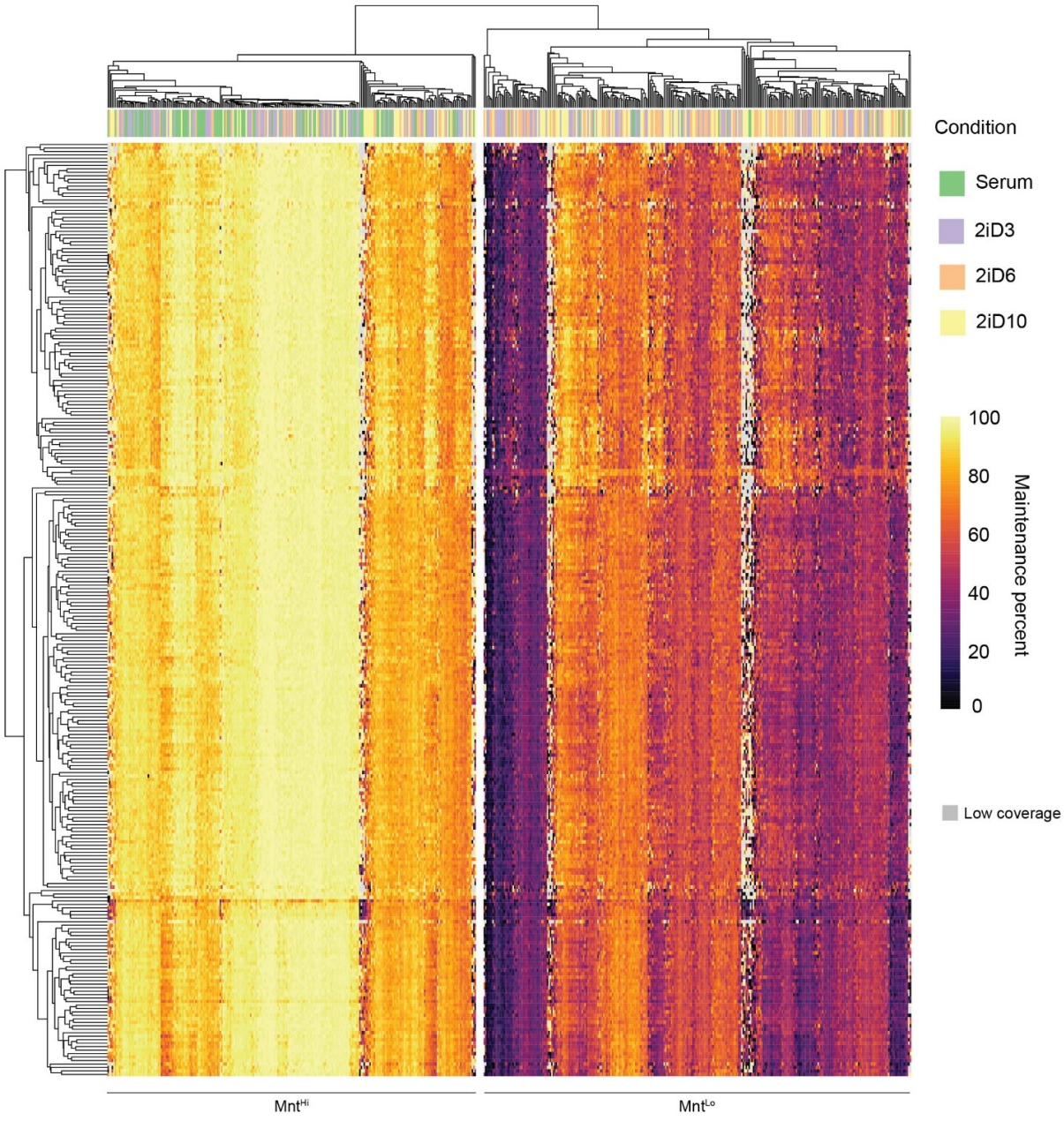
Supplementary Figure 5.10 | Comparison of scDyad&T-seq to scMspJI strand bias measurement in serum grown mESCs.

(**a**) 5mCpG detection by MspJI on the plus and minus strand for each chromosome of a representative highly maintained cell (P7L3.67, see Fig. 5.2f). Point color shows the detected 5mCpG maintenance percent of the chromosome. (**b**) Chromosomal 5mCpG strand bias observed in the detection by MspJI compared to observed 5mCpG maintenance percent. (**c**) Cell wide concordance of methylation calls compared to observed 5mCpG maintenance percent. Concordance was defined as the fraction of reads with at least 5 CpG's covered where 90% or greater of the sites were methylated (b,c) Point color describes the detected 5mCpG percent.



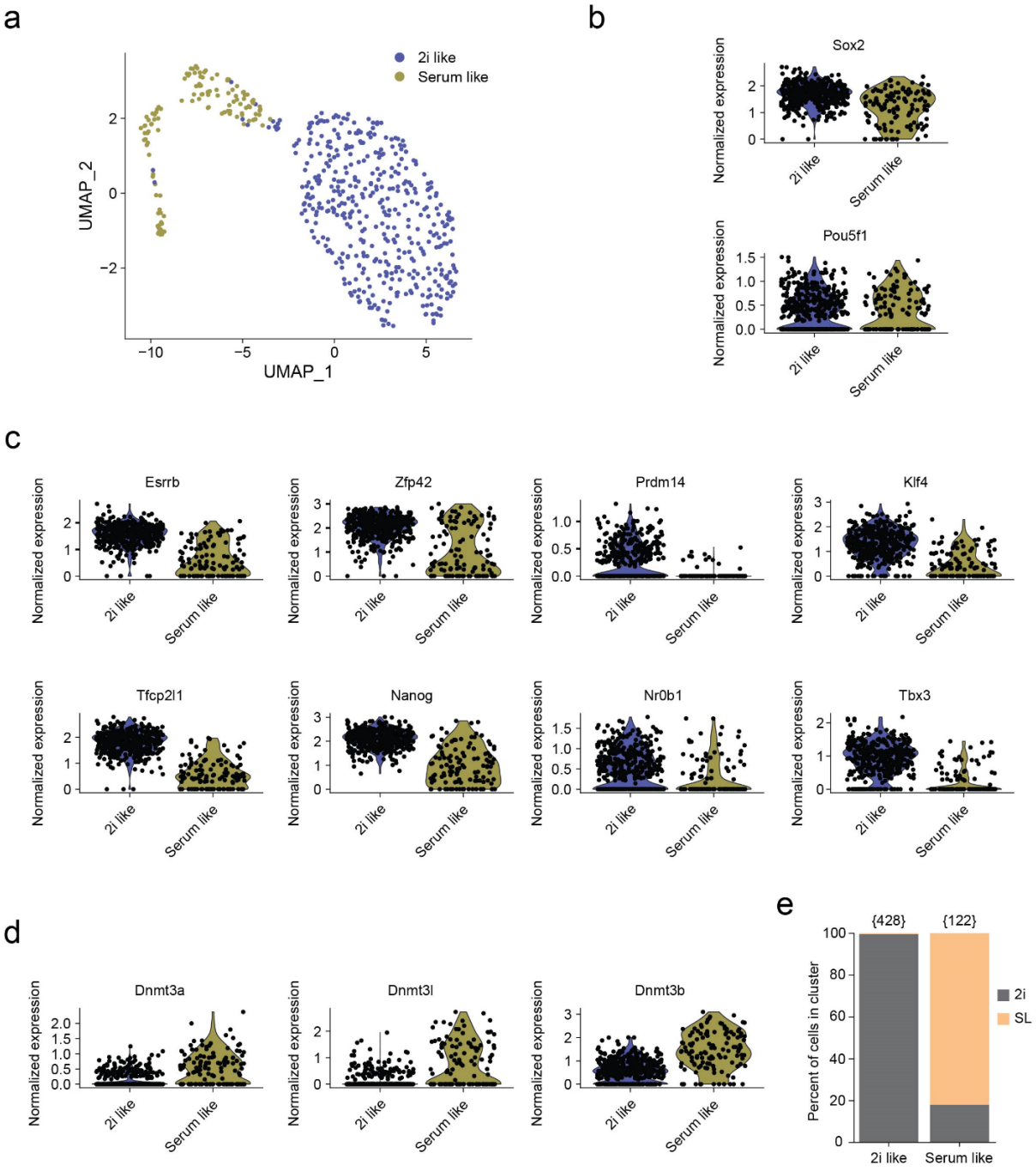
Supplementary Figure 5.11 | Hierarchical cluster of 5mCpG levels in cells transition to 2i conditions.

Based on genome wide 5mCpG levels, cells are clustered into either a 5mC low or 5mC high group, mC^{Lo} or mC^{Hi} respectively.



Supplementary Figure 5.12 | Hierarchical cluster of 5mCpG maintenance levels in cells transition to 2i conditions.

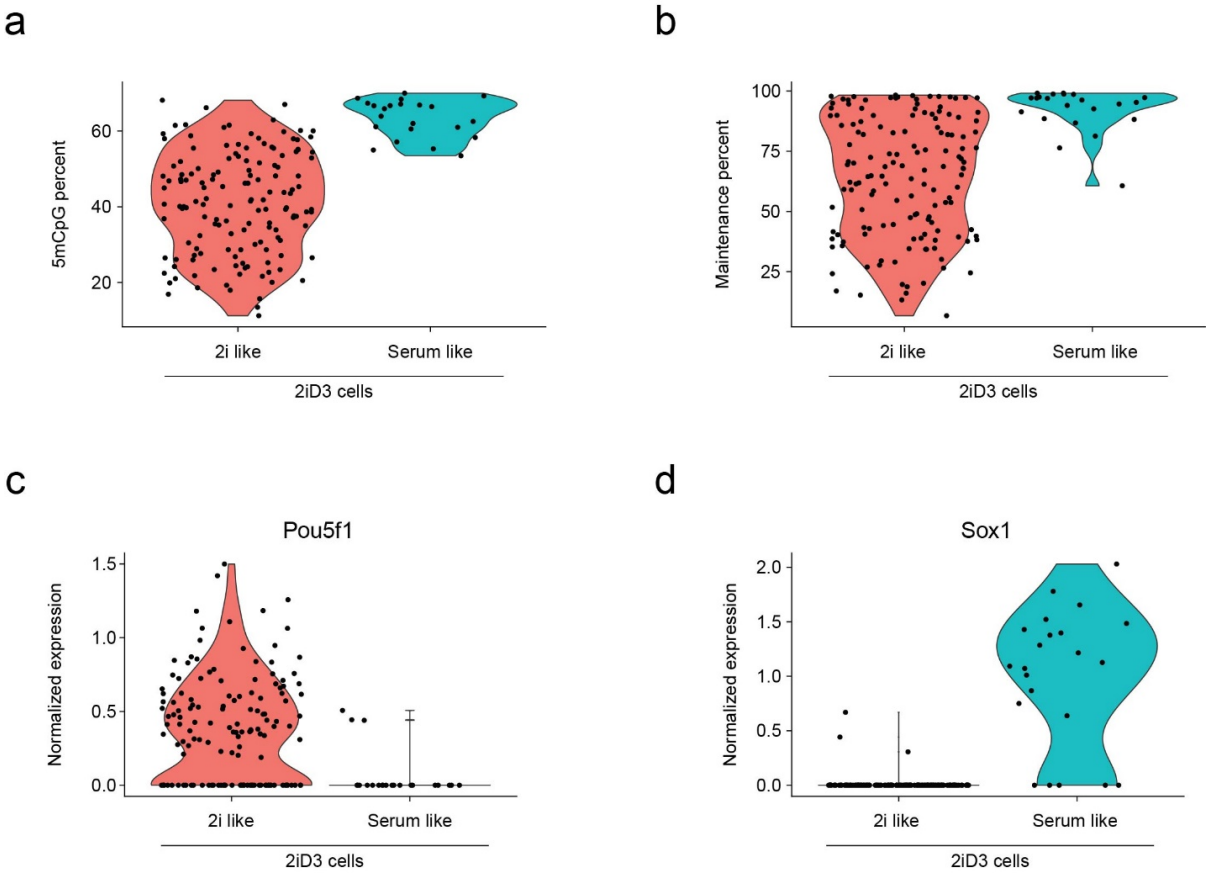
Based on genome wide 5mCpG maintenance levels, cells are clustered into either a low maintenance or high maintenance group, Mnt^{Lo} or Mnt^{Hi} respectively.



Supplementary Figure 5.13 | Broad transcriptional reprograming occurs quickly once mESCs are transitioned to 2i conditions.

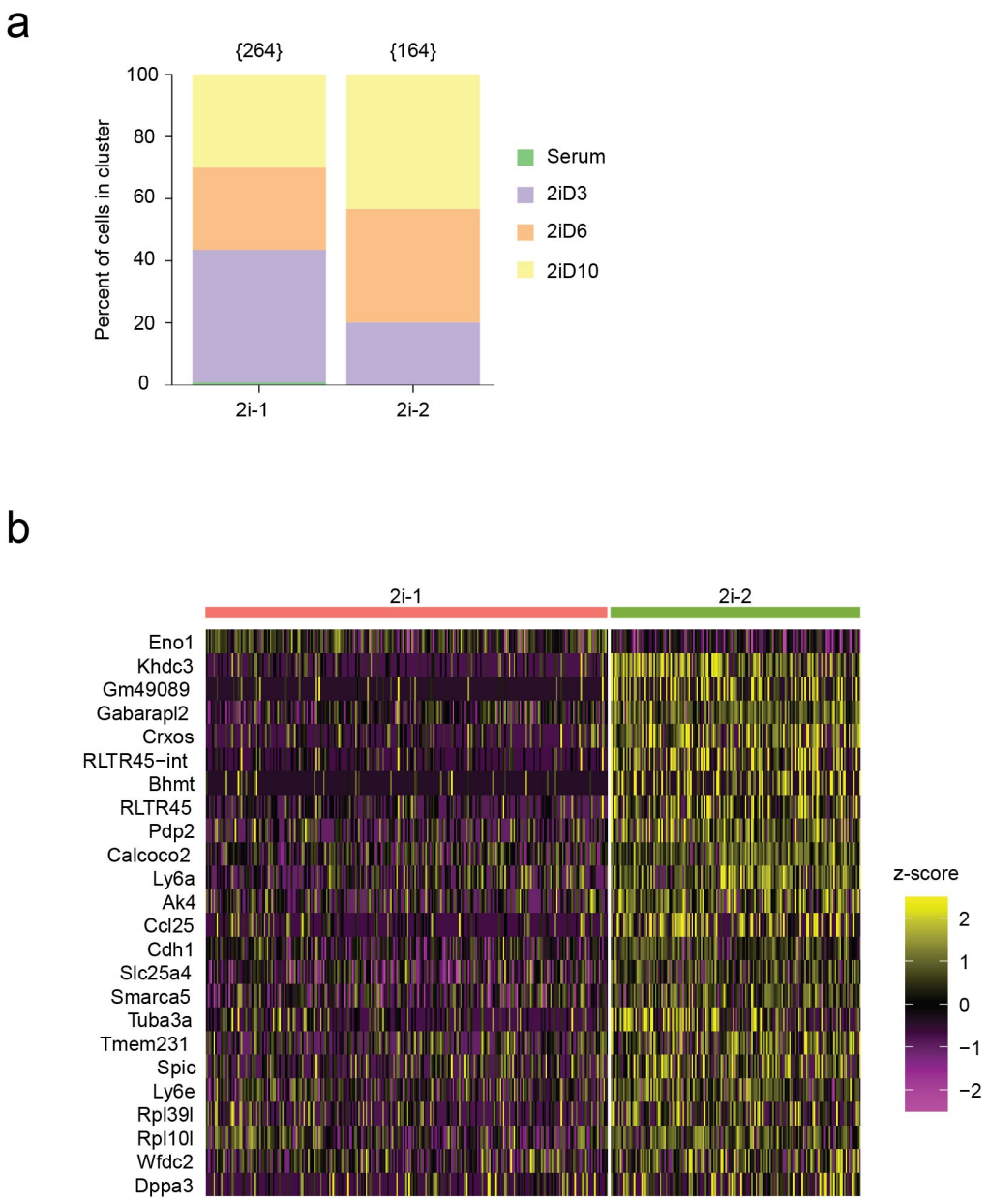
(**a**) UMAP visualization of cells transiting to 2i based on the single-cell transcriptomes obtained from scDyad&T-seq, discriminated by broad transcriptome-based clustering. The cluster names, 2i like and Serum like were assigned based on expression of key marker genes of mESCs in 2i or SL conditions respectively. (**b**) Expression of key pluripotency genes known to be similar between SL and 2i culture[226]. (**c**) Expression of genes known to be highly expressed in 2i mESCs when compared to those grown in SL conditions[226]. (**d**) Expression of genes known to be highly expressed in SL

mESCs when compared to those grown in 2i culture[226]. (**e**) Evaluation of transcriptome assignment accuracy based on known cellular grown conditions, SL or 2i. Bracketed numbers indicate total number of cells in that group.



Supplementary Figure 5.14 | Epigenetic barriers prevent 2i transition in some cells.

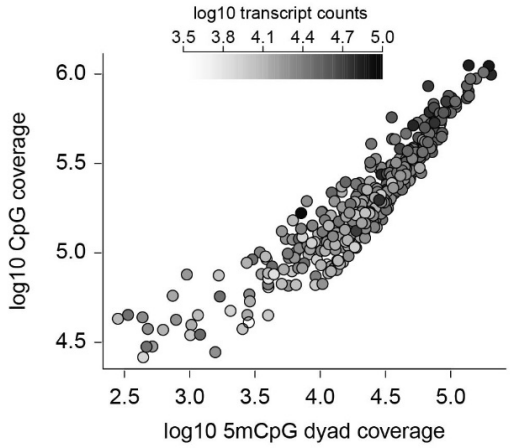
(**a,b**) 5mCpG methylation (**a**) and 5mCpG maintenance (**b**) levels for 2iD3 cells categorized by broad transcriptional group described in Supplementary Fig. 5.13a. (**c-d**) Gene expression of select genes, pluripotency marker Pou5f1 (aka Oct4) (**c**) and early neuroectoderm lineage marker Sox1 (**d**).



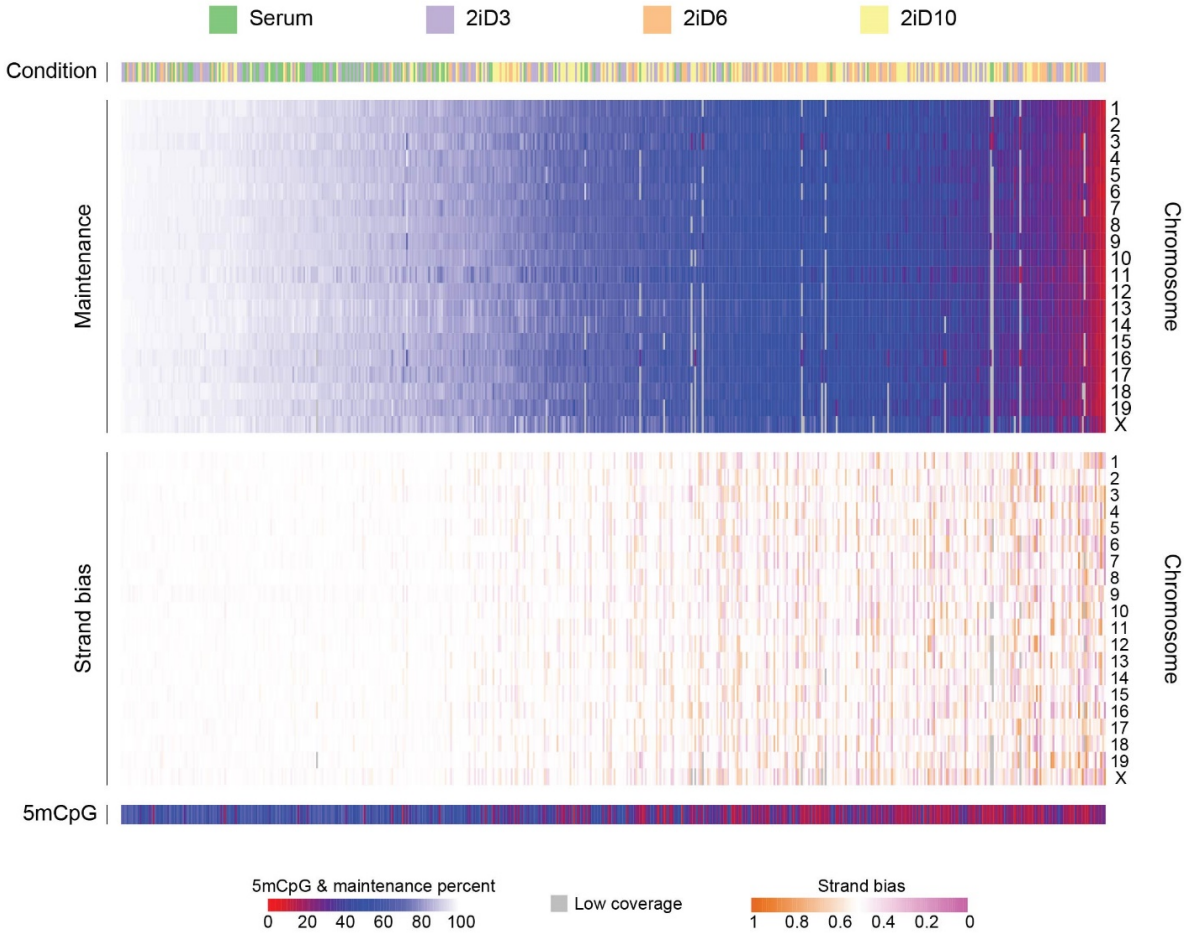
Supplementary Figure 5.15 | mESCs in 2i exhibit two transcriptionally distinct groups.

(**a**) 2i grown mESC transcriptional groups described in Fig. 5.3h, discriminated by time in 2i. Bracketed numbers indicate total number of cells in that group. (**b**) Gene expression of differentially expressed genes between the 2i-1 and 2i-2 population.

a



b



Supplementary Figure 5.16 | scDyad&T-seq connects cell identity to demethylation dynamics in single cells transitioning from SL to 2i.

(a) Coverage of 5mCpG dyads, non-dyad CpGs, and total transcripts in single cells transition from SL to 2i. (b) Heatmap of 5mCpG maintenance by chromosome indicates an increased level of sensitivity in detecting demethylation when compared to strand bias for the same cell. Culture conditions and genome wide 5mCpG methylation levels are also reported for the same cells.

6. scMATH-seq: Detecting 5-methylcytosine, DNA accessibility, RNA transcripts, and 5-hydroxymethylcytosine from the same single cell

A. scMATH-seq in human embryonic stem cells

It is well understood that epigenetic features can drive or repress gene expression. Exactly how these epigenetic features can alter gene expression and cell identity is not fully understood. Even more, the interaction between different epigenetic features in many cases is unknown and needs further study. While many novel techniques are emerging, the lack of modularity and limited number of features detected in one assay severely limits the ability to study epigenetic features and RNA simultaneously from the same single cell[227]. To bridge this gap, we present scMATH-seq a sequencing method to detect the methylome, DNA accessibility, the transcriptome, and the hydroxymethylome simultaneously from the same single-cell. This methodology builds upon the previously described scMAT-seq and scMTH-seq. Apart from the DNA accessibility measurement requiring the methylome to be read simultaneously, the methodology is highly flexible and allows any combination of detection of these four features from the same single cell (Fig. 1a). To do this, individual cells are isolated into reaction wells. Reverse transcription and GpC methylation of open DNA is then performed simultaneously. Next, second strand synthesis of the mRNA into cDNA is performed and then 5hmC sites are glucosylated. Following these steps, genomic DNA is digested sequentially with AbaSI and then MspJI. Digested DNA will be ligated to complimentary adaptors, the reaction wells are then pooled followed by *in vitro* transcription and PCR

amplification for Illumina library preparation. After sequencing, the cell and feature specific barcodes are used to distinguish between mRNA derived reads, 5-methylcytosine (5mC) derived reads, and 5-hydroxymethylcytosine (5hmC) derived reads. Finally, endogenous methylation will be deconvoluted from sequence-specific exogenously induced methylation, enabling quantification of the methylome and DNA accessibility.

B. Current limitations and the future of the single cell multiomics technology

scMATH-seq can detect each feature with high quality, rivaling the detection observed for each mark individually (Fig. 1b)[34,65]. While this method is highly promising, it has yet to be tested in a dynamic system and incurs a drawback of needing high sequencing depths per cell, a feature common to multiomics sequencing techniques. The high sequencing depth needed is due to differing initial levels and potentially differences in capture efficiency between the various detectable features. Notably low detection of RNA transcripts occurs in scMATH-seq and can be problematic in all members of the scMATH-seq methodology family (Fig. 1c). In scMAT-seq, an mRNA derived molecule enrichment strategy was created that is performed after sample amplification, leading to less dropouts and lower sequencing depths required to achieve high quality data. This same approach is applied to scMATH-seq, enriching the library for mRNA derived molecules, resolving the discussed issue (Fig. 1c). Like the mRNA derived molecules, in scMATH-seq, 5hmC derived molecules make up a very low fraction of the sequenced molecules even when compared to scMH-seq, necessitating the derivation of an enrichment strategy to significantly lower the sequencing costs to achieve high quality 5hmC

data (Fig. 1c). Such enrichment strategies are possible but would likely entail a redesign of the double stranded adapters currently used. Some promising methodologies include adding stretches of known sequence, or using alternative promoter sequences, for instance a T3 promoter sequence. If a different stretch of known sequence was added to each double stranded adapter type, it could be designed in a way for optimal biotin-based pulldown (as is done currently in the enrichment of mRNA derived molecules). Alternatively, each could also be designed with a restriction enzyme site for depletion with an endonuclease or with a protospacer adjacent motif (PAM) site for depletion with a CRISPR Cas9 based system[228,229]. Likely other promising enrichment strategies will be developed specifically for low input applications. As single cell multiomics as a field grows, foresight into enrichment strategies will be key for the practicality of the methodologies proposed.

In addition to this limitation, the scMATH-seq family of methods is currently cumbersome and requires multiple days of processing for detection in a few thousand cells. Many new multiomics methodologies have leveraged techniques first developed for scRNA-seq, including droplet based single-cell barcoding, and combinatorial barcoding techniques including split-and-pool[227]. The scaling of these technologies vastly outnumbers that of scMATH-seq but integration of so many epigenetic features has proven difficult with these techniques, specifically for detection of 5mC and 5hmC. It is likely that methods like scMATH-seq will be part of the first wave of techniques to investigate epigenetic marks like 5mC and 5hmC without traditional bisulfite sequencing. I am hopeful that scMATH-seq and the other techniques derived in this manuscript will be a starting point for future researchers to

improve detection in single cell multiomics techniques. I believe that researchers will soon have the ability to profile multiple epigenetic marks and the transcriptome from hundreds of thousands of cells with relative ease, and that this innovation will drive the development of computational tools which can take advantage of the newly developed scale. It is unimaginable the insights into human development, health, and disease that will be uncovered using these techniques, but it is almost certain that our current understanding of the epigenome will be transformed.

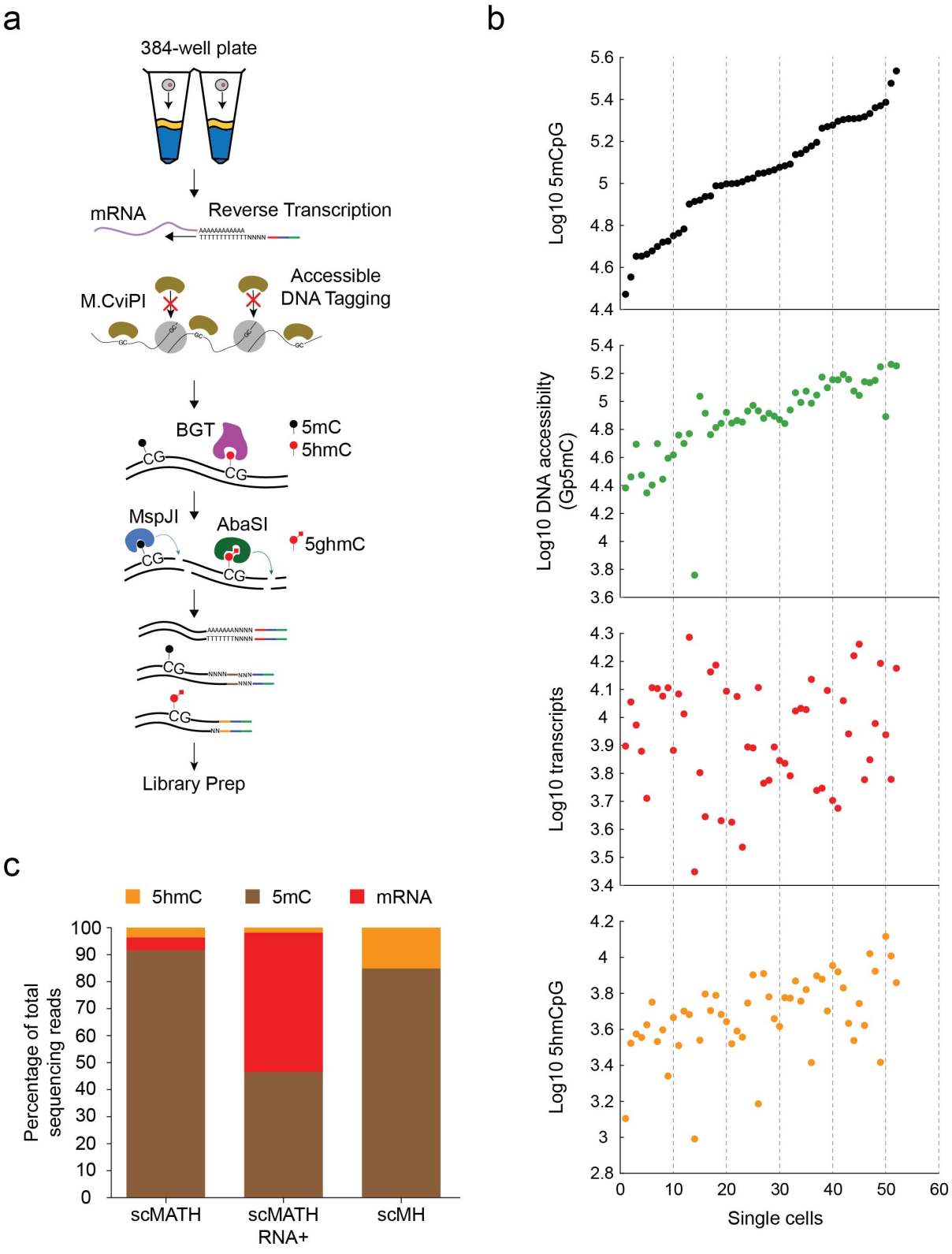


Figure 1 | Detection of 5mC, DNA accessibility, RNA transcripts, and 5hmC simultaneously from the same single cell using scMATH-seq.

(**a**) Schematic of scMATH-seq. Cell and detected species specific barcodes are shown in red, brown, and gold for mRNA, 5mC, and 5hmC respectively. The Illumina read 1 sequencing primer is in blue, and the T7 promoter is in green. (**b**) Detection level for each epigenetic mark in H9 human embryonic stem cells. Each cell is represented in the same x-axis position in all plots, from top to bottom, 5mCpG, DNA accessibility (Gp5mC), gene transcripts, and 5hmCpG. (**c**) Molecule type of origin detected in raw sequencing data for scMATH-seq (scMATH), the RNA enriched scMATH-seq library (scMATH RNA+), and for scMH-seq (scMH).

References

1. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
2. Tsompana, M. & Buck, M. J. Chromatin accessibility : a window into the genome. 1–16 (2014).
3. Miller, J. L. & Grant, P. A. The Role of DNA Methylation and Histone Modifications in Transcriptional Regulation in Humans. in *Epigenetics: Development and Disease,* vol. 61 289–317 (2013).
4. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
5. McBryant, S. J., Adams, V. H. & Hansen, J. C. Chromatin architectural proteins. *Chromosom. Res.* **14**, 39–51 (2006).
6. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311–322 (2008).
7. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).
8. Schones, D. E. *et al.* Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* **132**, 887–898 (2008).
9. Xu, M., Kladde, M. P., Van Etten, J. L. & Simpson, R. T. Cloning, characterization and expression of the gene coding for a cytosine-5-DNA methyltransferase recognizing GpC. *Nucleic Acids Res.* **26**, 3961–6 (1998).
10. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA

methylation within individual DNA molecules Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* 2497–2506 (2012) doi:10.1101/gr.143008.112.

11. Auerbach, R. K. *et al.* Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl. Acad. Sci.* **106**, 14926–14931 (2009).
12. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17**, 877–885 (2007).
13. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. in *Current Protocols in Molecular Biology* 21.29.1-21.29.9 (John Wiley & Sons, Inc., 2015). doi:10.1002/0471142727.mb2129s109.
14. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–90 (2015).
15. Mezger, A. *et al.* High-throughput chromatin accessibility profiling at single-cell resolution. *Nat. Commun.* **9**, 3647 (2018).
16. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, (2019).
17. Cusanovich, D. a *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (80-.).* **348**, 910–914 (2015).
18. Gao, W., Lai, B., Ni, B. & Zhao, K. Genome-wide profiling of nucleosome position and chromatin accessibility in single cells using scMNase-seq. *Nat. Protoc.* **15**, 68–85 (2020).

19. Lai, B. *et al.* Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature* 1 (2018) doi:10.1038/s41586-018-0567-3.

20. Jin, W. *et al.* Genome-wide detection of DNase i hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, (2015).

21. Pott, S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *Elife* **6**, 1–19 (2017).

22. Bird, A. DNA methylation patterns and epigenetic memory DNA methylation patterns and epigenetic memory. *Genes Dev.* 6–21 (2002) doi:10.1101/gad.947102.

23. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33**, 245–254 (2003).

24. Ko, M. *et al.* homeostasis and differentiation of hematopoietic stem cells in mice. **2**, (2011).

25. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472–479 (2013).

26. Gujar, H., Weisenberger, D. J. & Liang, G. The roles of human DNA methyltransferases and their isoforms in shaping the epigenome. *Genes (Basel).* **10**, (2019).

27. Guibert, S. & Weber, M. *Functions of DNA Methylation and Hydroxymethylation in Mammalian Development*. *Current Topics in Developmental Biology* vol. 104 (2013).

28. Suetake, I., Miyazaki, J., Murakami, C., Takeshima, H. & Tajima, S. Distinct

enzymatic properties of recombinant mouse DNA methyltransferases Dnmt3a and Dnmt3b. *J. Biochem.* **133**, 737–744 (2003).

29. Vilkaitis, G., Suetake, I., Klimašauskas, S. & Tajima, S. Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNA methyltransferase. *J. Biol. Chem.* **280**, 64–72 (2005).
30. Tahiliani, M. *et al.* Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science (80-.).* **324**, 930–935 (2009).
31. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (80-.).* **333**, 1300–3 (2011).
32. He, Y.-F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–7 (2011).
33. 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).
34. 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).
35. Wangsanuwat, C., Chialastri, A., Aldeguer, J. F., Rivron, N. C. & Dey, S. S. A probabilistic framework for cellular lineage reconstruction using integrated single-cell 5-hydroxymethylcytosine and genomic DNA sequencing. *Cell Reports Methods* **1**, 100060 (2021).
36. Brinkman, A. B. *et al.* Whole-genome DNA methylation profiling using MethylCap-seq. *Methods* **52**, 232–236 (2010).

37. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **37**, 853–862 (2005).

38. Song, C. X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.* **29**, 68–75 (2011).

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

40. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5- methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 1827–1831 (1992).

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

42. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).

43. Shareef, S. J. *et al.* Extended-representation bisulfite sequencing of gene regulatory elements in multiplexed samples and single cells. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-00910-x.

44. Laird, C. D. *et al.* Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 204–9 (2004).

45. Booth, M. J., Branco, M. R., Ficz, G., Oxley, D. & Krueger, F. Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. **8333**, (2010).

46. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).

47. Liu, Y. *et al.* Subtraction-free and bisulfite-free specific sequencing of 5-methylcytosine and its oxidized derivatives at base resolution. *Nat. Commun.* **12**, 618 (2021).

48. Liu, Y. *et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution*. *Nature Biotechnology* (2019). doi:10.1038/s41587-019-0041-2.

49. Sun, Z. *et al.* Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome Res.* **31**, 291–300 (2021).

50. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* **31**, 1280–1289 (2021).

51. Schutsky, E. K. *et al.* Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4204.

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

53. Guo, H. *et al.* Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* 2126–2135 (2013) doi:10.1101/gr.161679.113.

54. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory

elements in mammalian cortex. *Science (80-.).* **357**, (2017).

55. Luo, C. *et al.* Robust single-cell DNA methylome profiling with snmC-seq2. *Nat. Commun.* **9**, 7–12 (2018).

56. Mulqueen, R. M. *et al.* Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, (2018).

57. Farlik, M. *et al.* Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Rep.* **10**, 1386–1397 (2015).

58. Fabyanic, E. B. *et al.* Quantitative single cell 5hmC sequencing reveals non-canonical gene regulation by non-CG hydroxymethylation. *bioRxiv* 2021.03.23.434325 (2021) doi:10.1101/2021.03.23.434325.

59. Oda, M. *et al.* High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res.* **37**, 3829–3839 (2009).

60. Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* **27**, 361–368 (2009).

61. Suzuki, M. *et al.* Optimized design and data analysis of tag-based cytosine methylation assays. *Genome Biol.* **11**, R36 (2010).

62. Viswanathan, R., Cheruba, E. & Cheow, L. F. DNA Analysis by Restriction Enzyme (DARE) enables concurrent genomic and epigenomic characterization of single cells. *Nucleic Acids Res.* **47**, e122 (2019).

63. Borgaro, J. G. & Zhu, Z. Characterization of the 5-hydroxymethylcytosine-specific DNA restriction endonucleases. *Nucleic Acids Res.* **41**, 4198–4206 (2013).

64. Cohen-karni, D. *et al.* The MspJI family of modification-dependent restriction endonucleases for epigenetic studies. (2011) doi:10.1073/pnas.1018448108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1018448108.

65. Sen, M. *et al.* Strand-specific single-cell methylomics reveals distinct modes of DNA demethylation dynamics during early mammalian development. *Nat. Commun.* **12**, 1286 (2021).

66. Kanter, I. & Kalisky, T. Single Cell Transcriptomics: Methods and Applications. *Front. Oncol.* **5**, 1–8 (2015).

67. Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of single RNA transcripts in situ. *Science* **280**, 585–90 (1998).

68. Warren, L., Bryder, D., Weissman, I. L. & Quake, S. R. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc. Natl. Acad. Sci.* **103**, 17807–17812 (2006).

69. Chen, Y. *et al.* Single-Cell Sequencing Methodologies: From Transcriptome to Multi-Dimensional Measurement. *Small Methods* **2100111**, 2100111 (2021).

70. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

71. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

72. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, (2015).

73. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (80-.).* **357**, 661–667 (2017).

74. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and

spinal cord with split-pool barcoding. *Science (80-.).* **360**, 176–182 (2018).

75. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).

76. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* **33**, 285–289 (2015).

77. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (80-.).* **0730**, eaau0730 (2018).

78. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, (2019).

79. Liu, L. *et al.* Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat. Commun.* **10**, (2019).

80. Xing, Q. R. *et al.* Parallel bimodal single-cell sequencing of transcriptome and chromatin accessibility. *Genome Res.* **30**, 1027–1039 (2020).

81. Zhu, C. *et al.* An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).

82. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).

83. Guo, F. *et al.* Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* **27**, 967–988 (2017).

84. Gu, C., Liu, S., Wu, Q., Zhang, L. & Guo, F. Integrative single-cell analysis of transcriptome, DNA methylome and chromatin accessibility in mouse oocytes.

Cell Res. **29**, 110–123 (2019).

85. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).

86. Hu, Y. *et al.* Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* **17**, 1–11 (2016).

87. Hou, Y. *et al.* Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26**, (2016).

88. Hunt, K. V *et al.* SINEultaneous profiling of epigenetic heterogeneity and transcriptome in single cells. *bioRxiv* 2021.03.25.436351 (2021).

89. Gu, H. *et al.* Smart-RRBS for single-cell methylome and transcriptome analysis. *Nat. Protoc.* doi:10.1038/s41596-021-00571-9.

90. Luo, C. *et al.* Multi-omic profiling of transcriptome and DNA methylome in single nuclei with molecular partitioning. *bioRxiv* 1–18 (2018) doi:doi.org/10.1101/434845.

91. Yan, R. *et al.* Decoding dynamic epigenetic landscapes in human oocytes using single-cell multi-omics sequencing. *Cell Stem Cell* 1–16 (2021) doi:10.1016/j.stem.2021.04.012.

92. Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).

93. Luo, C. *et al.* Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. *bioRxiv* 2019.12.11.873398 (2019) doi:10.1101/2019.12.11.873398.

94. Wang, Y. *et al.* Single-cell multiomics sequencing reveals the functional regulatory landscape of early embryos. *Nat. Commun.* **12**, 1247 (2021).

95. Rossi, D. J., Jamieson, C. H. M. & Weissman, I. L. Stems Cells and the Pathways to Aging and Cancer. *Cell* **132**, 681–696 (2008).

96. Mohn, F. & Schübeler, D. Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends Genet.* **25**, 129–136 (2009).

97. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Publ. Gr.* **17**, 487–500 (2016).

98. Weissman, T. A. & Pan, Y. A. Brainbow: New resources and emerging biological applications for multicolor genetic labeling and analysis. *Genetics* **199**, 293–306 (2015).

99. Spanjaard, B. & Junker, J. P. Methods for lineage tracing on the organism-wide level. *Curr. Opin. Cell Biol.* **49**, 16–21 (2017).

100. Wu, J. & Belmonte, J. C. I. The Molecular Harbingers of Early Mammalian Embryo Patterning. *Cell* **165**, 13–15 (2016).

101. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129–1133 (2010).

102. Vincent, J. J. *et al.* Stage-Specific Roles for Tet1 and Tet2 in DNA Demethylation in Primordial Germ Cells. *Cell Stem Cell* **12**, 470–478 (2013).

103. Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339–344 (2012).

104. Liang, G. & Zhang, Y. Embryonic stem cell and induced pluripotent stem cell: an epigenetic perspective. *Cell Res.* **23**, 49–69 (2013).

105. Olson, C. K. *et al.* In vitro fertilization is associated with an increase in major

birth defects. *Fertil. Steril.* **84**, 1308–1315 (2005).

106. Pera, M. F. Human embryo research and the 14-day rule. *Dev.* **144**, 1923–1925 (2017).
107. Tam, P. P. L. & Behringer, R. R. Mouse gastrulation : the formation of a mammalian body plan. **68**, 3–25 (1997).
108. Hancock, G. V, Wamaitha, S. E., Peretz, L. & Clark, A. T. Mammalian primordial germ cell specification. 1–12 (2021) doi:10.1242/dev.189217.
109. Lawson, K. A. *et al.* Bmp4 is required for the generation of primordial germ cells in the mouse embryo. 424–436 (1999).
110. Carroll, O. *et al.* Blimp1 is a critical determinant of the germ cell lineage in mice. **436**, (2005).
111. Ohinata, Y. *et al.* A Signaling Principle for the Specification of the Germ Cell Lineage in Mice. *Cell* **137**, 571–584 (2009).
112. Richardson, B. E. & Lehmann, R. Mechanisms guiding primordial germ cell migration: Strategies from different organisms. *Nature Reviews Molecular Cell Biology* vol. 11 37–49 (2010).
113. Saitou, M. & Yamaji, M. Primordial Germ Cells in Mice. 1–20 (2016).
114. Hayashi, K. Germ Cell Specification in Mice. **394**, 394–397 (2010).
115. Cusanovich, D. A. *et al.* The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nat. Publ. Gr.* **555**, 538–542 (2018).
116. Bogdanovic, O. *et al.* Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis. 1313–1327 (2011) doi:10.1101/gr.114843.110.

117. Wang, L. *et al.* Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**, 979–991 (2014).

118. Messerschmidt, D. M., Knowles, B. B. & Solter, D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev.* **28**, 812–828 (2014).

119. Hill, P. W. S., Amouroux, R. & Hajkova, P. DNA demethylation, Tet proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: An emerging complex story. *Genomics* vol. 104 324–333 (2014).

120. Yamaguchi, S. *et al.* Dynamics of 5-methylcytosine and 5-hydroxymethylcytosine during germ cell reprogramming. *Cell Res.* **23**, 329–339 (2013).

121. Arand, J. *et al.* In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet.* **8**, e1002750 (2012).

122. Xu, C. & Corces, V. G. Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites. **1170**, 1166–1170 (2018).

123. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, (2019).

124. Qu, X. Bin, Pan, J., Zhang, C. & Huang, S. Y. Sox17 facilitates the differentiation of mouse embryonic stem cells into primitive and definitive endoderm in vitro. *Dev. Growth Differ.* **50**, 585–593 (2008).

125. Irie, N. *et al.* SOX17 is a critical specifier of human primordial germ cell fate. *Cell* **160**, 253–268 (2015).

126. Shao, Y. *et al.* A pluripotent stem cell-based model for post-implantation human amniotic sac development. *Nat. Commun.* **8**, 1–15 (2017).

127. Shao, Y. *et al.* Self-organized amniogenesis by human pluripotent stem cells in a biomimetic implantation-like niche. *Nat. Mater.* **16**, 419–427 (2017).

128. Simunovic, M. *et al.* A 3D model of a human epiblast reveals BMP4-driven symmetry breaking. *Nat. Cell Biol.* **21**, 900–910 (2019).

129. Martyn, I., Kanno, T. Y., Ruzo, A., Siggia, E. D. & Brivanlou, A. H. Self-organization of a human organizer by combined Wnt and Nodal signaling. *Nature* **558**, 132–135 (2018).

130. Zheng, Y. *et al.* Controlled modelling of human epiblast and amnion development using stem cells. *Nature* (2019) doi:10.1038/s41586-019-1535-2.

131. Etoc, F. *et al.* A Balance between Secreted Inhibitors and Edge Sensing Controls Gastruloid Self-Organization. *Dev. Cell* **39**, 302–315 (2016).

132. Karzbrun, E., Khankhel, A. H., Megale, H. C., Glasauer, S. M. K. & Streichan, S. J. Self-organized morphogenesis of a human neural tube in vitro by geometric constraints. *BioRxiv* 1–26 (2021) doi:10.1101/2021.07.24.453659.

133. Sasaki, K. *et al.* Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. *Cell Stem Cell* **17**, 178–194 (2015).

134. Gell, J. J. *et al.* An Extended Culture System that Supports Human Primordial Germ Cell-like Cell Survival and Initiation of DNA Methylation Erasure. *Stem Cell Reports* **14**, 433–446 (2020).

135. Tomlinson, M. J., Tomlinson, S., Yang, X. B. & Kirkham, J. Cell separation: Terminology and practical considerations. *J. Tissue Eng.* **4**, 1–14 (2013).

136. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

137. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet.* **34**, 666–681 (2018).

138. Cedar, H. & Bergman, Y. Programming of DNA methylation patterns. *Annu. Rev. Biochem.* **81**, 97–117 (2012).

139. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–6 (2015).

140. Hackett, J. A. & Surani, M. A. DNA methylation dynamics during the mammalian life cycle. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368**, 20110328 (2013).

141. Leonhardt, H., Page, A. W., Weier, H. U. & Bestor, T. H. A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell* **71**, 865–873 (1992).

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

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

144. Mayer, W., Niveleau, A., Walter, J., Fundele, R. & Haaf, T. Demethylation of the zygotic paternal genome. *Nature* **403**, 501–502 (2000).

145. Inoue, A. & Zhang, Y. Replication-Dependent Loss of 5-Hydroxymethylcytosine in Mouse Preimplantation Embryos. *Science (80-.).* **334**, 194–194 (2011).

146. 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.* **108**, 3642–3647 (2011).

147. Arand, J. *et al.* Selective impairment of methylation maintenance is the major cause of DNA methylation reprogramming in the early embryo. 1–14 (2015).

148. Okamoto, Y. *et al.* DNA methylation dynamics in mouse preimplantation embryos revealed by mass spectrometry. *Sci. Rep.* **6**, 19134 (2016).

149. Guo, H. *et al.* The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014).

150. Smith, Z. D. *et al.* DNA methylation dynamics of the human preimplantation embryo. *Nature* **511**, 611–615 (2014).

151. Okae, H. *et al.* Genome-Wide Analysis of DNA Methylation Dynamics during Early Human Development. *PLoS Genet.* **10**, 1–12 (2014).

152. Maenohara, S. *et al.* Role of UHRF1 in de novo DNA methylation in oocytes and maintenance methylation in preimplantation embryos. *PLoS Genet.* **13**, 1–20 (2017).

153. Huang, Y. *et al.* The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One* **5**, 1–9 (2010).

154. Rooijers, K. *et al.* Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells. *Nature Biotechnology* vol. 37 766–772 (2019).

155. Habibi, E. *et al.* Whole-Genome Bisulfite Sequencing of Two Distinct Interconvertible DNA Methylomes of Mouse Embryonic Stem Cells. *Cell Stem Cell* **13**, 360–369 (2013).

156. Carlson, L. L., Page, A. W. & Bestor, T. H. Properties and localization of DNA methyltransferase in preimplantation mouse embryos: implications for genomic imprinting. *Genes Dev.* **6**, 2536–2541 (1992).

157. Cardoso, M. C. & Leonhardt, H. DNA methyltransferase is actively retained in

the cytoplasm during early development. *J. Cell Biol.* **147**, 25–32 (1999).

158. Howell, C. Y. *et al.* Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. *Cell* **104**, 829–838 (2001).
159. Ratnam, S. *et al.* Dynamics of Dnmt1 Methyltransferase Expression and Intracellular Localization during Oogenesis and Preimplantation Development. *Dev. Biol.* **245**, 304–314 (2002).
160. Cirio, M. C. *et al.* Preimplantation expression of the somatic form of Dnmt1 suggests a role in the inheritance of genomic imprints. *BMC Dev. Biol.* **8**, 9 (2008).
161. Kurihara, Y. *et al.* Maintenance of genomic methylation patterns during preimplantation development requires the somatic form of DNA methyltransferase 1. *Dev. Biol.* **313**, 335–346 (2008).
162. Hirasawa, R. *et al.* Maternal and zygotic Dnmt1 are necessary and sufficient for the maintenance of DNA methylation imprints during preimplantation development. *Genes Dev.* **22**, 1607–1616 (2008).
163. Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5237–42 (2000).
164. Ziller, M. J. *et al.* Genomic distribution and Inter-Sample variation of Non-CpG methylation across human cell types. *PLoS Genet.* **7**, (2011).
165. Zhao, L. *et al.* The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome Res.* **24**, 1296–1307 (2014).
166. Iurlaro, M., von Meyenn, F. & Reik, W. DNA methylation homeostasis in

human and mouse development. *Curr. Opin. Genet. Dev.* **43**, 101–109 (2017).

167. Petrussa, L., Van de Velde, H. & De Rycke, M. Dynamic regulation of DNA methyltransferases in human oocytes and preimplantation embryos after assisted reproductive technologies. *Mol. Hum. Reprod.* **20**, 861–874 (2014).
168. Petrussa, L., Van de Velde, H. & De Rycke, M. Similar kinetics for 5-methylcytosine and 5-hydroxymethylcytosine during human preimplantation development in vitro. *Mol. Reprod. Dev.* **83**, 594–605 (2016).
169. Morgani, S., Nichols, J. & Hadjantonakis, A. K. The many faces of Pluripotency: In vitro adaptations of a continuum of in vivo states. *BMC Dev. Biol.* **17**, 10–12 (2017).
170. Yan, R. *et al.* Decoding dynamic epigenetic landscapes in human oocytes using single-cell multi-omics sequencing. *Cell Stem Cell* 1–16 (2021) doi:10.1016/j.stem.2021.04.012.
171. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, (2016).
172. Pott, S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *Elife* **6**, 1–19 (2017).
173. Encode Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
174. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
175. Jones, P. A. Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
176. Sumi, T., Tsuneyoshi, N., Nakatsuji, N. & Suemori, H. Defining early lineage

specification of human embryonic stem cells by the orchestrated balance canonical Wnt/β-catenin, activin/Nodal and BMP signaling. *Development* **135**, 2969–2979 (2008).

177. Hackland, J. O. S. *et al.* Top-Down Inhibition of BMP Signaling Enables Robust Induction of hPSCs Into Neural Crest in Fully Defined, Xeno-free Conditions. *Stem Cell Reports* **9**, 1043–1052 (2017).
178. Lam, A. Q. *et al.* Rapid and efficient differentiation of human pluripotent stem cells into intermediate mesoderm that forms tubules expressing kidney proximal tubular markers. *J. Am. Soc. Nephrol.* **25**, 1211–1225 (2014).
179. Tang, W. W. C., Kobayashi, T., Irie, N., Dietmann, S. & Surani, M. A. Specification and epigenetic programming of the human germ line. *Nat. Rev. Genet.* **17**, 585–600 (2016).
180. Sasaki, K. *et al.* The Germ Cell Fate of Cynomolgus Monkeys Is Specified in the Nascent Amnion. *Dev. Cell* **39**, 169–185 (2016).
181. Kobayashi, T. *et al.* Principles of early human development and germ cell program from conserved model systems. *Nature* **546**, 416–420 (2017).
182. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
183. Chen, D. *et al.* Human Primordial Germ Cells Are Specified from Lineage-Primed Progenitors. *Cell Rep.* **29**, 4568-4582.e5 (2019).
184. Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V. & King Jordan, I. On the presence and role of human gene-body DNA methylation. *Oncotarget* **3**, 462–474 (2012).
185. Harrison, A. & Parle-McDermott, A. DNA Methylation: A Timeline of Methods

and Applications. *Front. Genet.* **2**, 632–649 (2011).

186. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
187. Huang, X. *et al.* High-throughput sequencing of methylated cytosine enriched by modification-dependent restriction endonuclease MspJI. (2013) doi:10.1186/1471-2156-14-56.
188. Sun, Z. *et al.* High-Resolution Enzymatic Mapping of Genomic 5-Hydroxymethylcytosine in Mouse Embryonic Stem Cells. *Cell Rep.* **3**, 567–576 (2013).
189. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
190. Gowher, H. & Jeltsch, A. Mammalian DNA methyltransferases: New discoveries and open questions. *Biochem. Soc. Trans.* **46**, 1191–1202 (2018).
191. Bashtrykov, P. *et al.* Specificity of dnmt1 for methylation of hemimethylated CpG sites resides in its catalytic domain. *Chem. Biol.* **19**, 572–578 (2012).
192. Pinney, S. Mammalian Non-CpG Methylation: Stem Cells and Beyond. *Biology (Basel).* **3**, 739–751 (2014).
193. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
194. Hancock, G. V, Wamaitha, S. E., Peretz, L. & Clark, A. T. Mammalian primordial germ cell specification. *Development* **148**, 1–12 (2021).
195. Smela, M. P., Sybirna, A., Wong, F. C. K. & Azim Surani, M. Testing the role of sox15 in human primordial germ cell fate [version 2; peer review: 2 approved]. *Wellcome Open Res.* **4**, 1–19 (2019).

196. Youngren, K. K. *et al.* The Ter mutation in the dead end gene causes germ cell loss and testicular germ cell tumours. *Nature* **435**, 360–364 (2005).

197. Lai, F., Singh, A. & King, M. Lou. Xenopus nanos1 is required to prevent endoderm gene expression and apoptosis in primordial germ cells. *Development* vol. 139 1476–1486 (2012).

198. Aguero, T. *et al.* Maternal dead-end 1 promotes translation of nanos1 through binding the eIF3 complex. *Development* **144**, 3755–3765 (2017).

199. Ruthig, V. A. *et al.* The RNA-binding protein DND1 acts sequentially as a negative regulator of pluripotency and a positive regulator of epigenetic modifiers required for germ cell reprogramming. *Dev.* **146**, (2019).

200. Charlton, J. *et al.* Global delay in nascent strand DNA methylation. *Nat. Struct. Mol. Biol.* **25**, 327–332 (2018).

201. Wang, Q. *et al.* Imprecise DNMT1 activity coupled with neighbor-guided correction enables robust yet flexible epigenetic inheritance. *Nat. Genet.* **52**, 828–839 (2020).

202. Lepikhov, K. *et al.* Two are better than one: HPoxBS - hairpin oxidative bisulfite sequencing. *Nucleic Acids Res.* **46**, e88–e88 (2018).

203. Stresemann, C. & Lyko, F. Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine. *Int. J. Cancer* **123**, 8–13 (2008).

204. Ying, Q. L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).

205. Leitch, H. G. *et al.* Naive pluripotency is associated with global DNA hypomethylation. *Nat. Struct. Mol. Biol.* **20**, 311–316 (2013).

206. Mulholland, C. B. *et al.* Recent evolution of a TET-controlled and

DPPA3/STELLA-driven pathway of passive DNA demethylation in mammals. *Nat. Commun.* **11**, 5972 (2020).

207. von Meyenn, F. *et al.* Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells. *Mol. Cell* **62**, 848–861 (2016).

208. Song, C. X., Diao, J., Brunger, A. T. & Quake, S. R. Simultaneous single-molecule epigenetic imaging of DNA methylation and hydroxymethylation. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4338–4343 (2016).

209. Hu, L. *et al.* Crystal Structure of TET2-DNA Complex : Insight into TET-Mediated 5mC Oxidation. 1545–1555 (2013) doi:10.1016/j.cell.2013.11.020.

210. Xing, X. *et al.* Direct observation and analysis of TET-mediated oxidation processes in a DNA origami nanochip. *Nucleic Acids Res.* **48**, 4041–4051 (2020).

211. Lu, R. *et al.* Inhibition of the Extracellular Signal-regulated Kinase/Mitogen-activated Protein Kinase Pathway Decreases DNA Methylation in Colon Cancer Cells. *J. Biol. Chem.* **282**, 12249–12259 (2007).

212. Sarkar, S. *et al.* Histone deacetylase inhibitors reverse CpG methylation by regulating DNMT1 through ERK signaling. *Anticancer Res.* **31**, 2723–2732 (2011).

213. Funaki, S. *et al.* Inhibition of maintenance DNA methylation by Stella. *Biochem. Biophys. Res. Commun.* **453**, 455–460 (2014).

214. Ficz, G. *et al.* FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* **13**, 351–359 (2013).

215. Estève, P. O. *et al.* Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. *Genes Dev.* **20**, 3089–3103 (2006).

216. Harrison, J. S. *et al.* Hemi-methylated DNA regulates DNA methylation inheritance through allosteric activation of H3 ubiquitylation by UHRF1. *Elife* **5**, 1–24 (2016).

217. Zhao, Q. *et al.* Dissecting the precise role of H3K9 methylation in crosstalk with DNA maintenance methylation in mammals. *Nat. Commun.* **7**, (2016).

218. Karmodiya, K., Krebs, A. R., Oulad-Abdelghani, M., Kimura, H. & Tora, L. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics* **13**, (2012).

219. Singer, Z. S. *et al.* Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells. *Mol. Cell* **55**, 319–331 (2014).

220. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).

221. Marks, H. *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**, 590–604 (2012).

222. Hastreiter, S. *et al.* Inductive and Selective Effects of GSK3 and MEK Inhibition on Nanog Heterogeneity in Embryonic Stem Cells. *Stem Cell Reports* **11**, 58–69 (2018).

223. Zheng, P. & Dean, J. Role of Filia , a maternal effect gene , in maintaining euploidy during cleavage-stage mouse embryogenesis. **106**, 7473–7478 (2009).

224. Zhao, B. *et al.* Filia is an ESC-Specific Regulator of DNA Damage Response and Safeguards Genomic Stability. *Cell Stem Cell* **16**, 684–698 (2015).

225. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).

226. Sang, H. *et al.* Dppa3 is critical for Lin28a-regulated ES cells naïve-primed state conversion. *J. Mol. Cell Biol.* **11**, 474–488 (2019).

227. Macaulay, I. C., Ponting, C. P. & Voet, T. Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends Genet.* **33**, 155–168 (2017).

228. Gu, W. *et al.* Depletion of Abundant Sequences by Hybridization (DASH): Using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* **17**, 1–13 (2016).

229. Loi, D. S. C., Yu, L. & Wu, A. R. Effective ribosomal RNA depletion for single-cell total RNA-seq by scDASH. *PeerJ* **9**, 1–18 (2021).

230. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).

231. Van Landuyt, L. *et al.* Closed blastocyst vitrification of biopsied embryos: Evaluation of 100 consecutive warming cycles. *Hum. Reprod.* **26**, 316–322 (2011).

232. Clark, S. J. *et al.* Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat. Protoc.* **12**, 534–547 (2017).

233. Sun, M., Velmurugan, K. R., Keimig, D. & Xie, H. HBS-Tools for Hairpin Bisulfite Sequencing Data Processing and Analysis. *Adv. Bioinformatics* **2015**, 1–4 (2015).

234. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).

235. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

236. Gell, J. J., Zhao, J., Chen, D., Hunt, T. J. & Clark, A. T. PRDM14 is expressed in germ cell tumors with constitutive overexpression altering human germline differentiation and proliferation. *Stem Cell Res.* **27**, 46–56 (2018).

237. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).

238. Wangsanuwat, C., Heom, K. A., Liu, E. & Malley, M. A. O. Efficient and cost-effective bacterial mRNA sequencing from low input samples through ribosomal RNA depletion. 1–21 (2020).

239. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

240. Stephens, M. False discovery rates: a new deal. *Biostatistics* **18**, kxw041 (2016).

241. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

242. Krueger, F. & Andrews, S. R. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

243. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).

Appendix

A. Chapter 2 Methods

1. Cell culture

E14tg2a mouse embryonic stem cells were obtained from American Type Culture Collection (ATCC CRL-182) and the hybrid 129/Sv:CAST/EiJ mouse embryonic stem cells were obtained from Jop Kind's group (Hubrecht Institute). Both lines were tested for mycoplasma contamination. Cells were grown on 0.1% gelatin in ES cell culture media; DMEM (1x) high glucose + glutamax (Gibco), supplemented with 10% FCS (Greiner) 100 µM β-mercaptoethanol (Sigma), 100 µM non-essential amino acids (Gibco), 50 µg/mL Pen/Strep (Gibco) and 1000 U/mL ESGRO mLIF (Millipore). Cells were split every 2 days and media changed every day. Cells were harvested before FACS by washing 3 times with 1x PBS with calcium and magnesium and incubated with 0.05% Trypsin (Life Technologies). Cell were resuspended in ES culture media and cell clumps were removed by passing the cells through a BD Falcon 5 mL polystyrene tube with a filter top.

2. Crispr-Cas9 Dnmt1 knockout

Six gRNA sequences targeting three exons of mouse Dnmt1 were used as described previously[230]. Phosphorylated BbsI compatible restriction overhangs were added to gRNA top and bottom oligos and resuspended at 100 µM in nuclease-free water. Annealing of the oligos was performed in 1x ligation buffer (NEB) using the following program: 97°C for 5 minutes, ramp down by 1°C per 1 minute to 20°C. The pX330 CRISPR-Cas9-GFP gRNA plasmid was a kind gift from Eva van Rooij and mixed with 0.1 µM gRNA oligo. The reaction was simultaneously digested with BbsI

(NEB) and ligated with T4 DNA ligase (NEB) overnight at 16°C. Ligation reactions were transformed into DH5α competent cells and subsequently sequenced using Sanger dideoxy sequencing to confirm the correct insert. All six pX300-gRNA plasmids were pooled and 1 µg was transfected into 2 million E14tg2a cells using Lipofectamine (Life Technologies). A separate pX300 empty vector was also transfected into E14tg2a to serve as a negative control. Two days later, single GFP positive cells were sorted into 384-well plates (BioRad) and subjected to scMspJI-seq.

3. Preimplantation mouse embryo isolation

CAST/EiJ x C57BL/6 hybrid mouse embryos were obtained from four 3-month-old superovulated B6 mothers (injected with pregnant mare serum gonadotropin (PMSG) and human chorionic gonadotropin (HCG) 22 h later), isolated using hyaluronic acid (Sigma), and incubated in M16 medium at 37°C and 5% CO2. The mice were housed at temperatures of 20-24°C, humidity of 45-65%, and a light/dark cycle of 14/10 hours. Individual cells were isolated using Tyrode's solution (Sigma) and trypsin (Life Technologies) and manually deposited into 384-well plates containing lysis buffer and Vapor-lock. Plates were subsequently centrifuged at 1,000 rpm for 1 minute to ensure that cells reach the aqueous phase and then subjected to scMspJI-seq. All animal experiments were approved by the Royal Netherlands Academy of Arts and Sciences and were performed according to the animal experimentation guidelines of the KNAW.

4. Preimplantation human embryo isolation

Supernumerary cryopreserved human embryos were obtained for research from patients undergoing in vitro fertilization (IVF) using standard clinical protocols, at the

Department for Reproductive Medicine, Ghent University Hospital. Cleavage stage embryos, cryopreserved on day 2 or 3 of development, were warmed using EmbryoThaw™ media (Fertipro, Belgium), as outlined by the manufacturer. Blastocyst stage embryos, vitrified on day 5 or 6 of development, were warmed using the Vitrification Thaw kit (Irvine Scientific, Netherlands), as described[231]. Embryos were transferred to either Cook Cleavage or Cook Blastocyst Medium (COOK, Ireland) depending on their developmental stage, and cultured in 20 µL medium droplets under mineral oil (Irvine Scientific, Netherlands) at 37°C, 6% CO_2 and 5% O_2. When required, embryos were briefly treated with Acidic Tyrode's Solution (Sigma-Aldrich, Belgium) for removal of the zona pellucida. All embryos were washed and subsequently dissociated by gentle mechanical dissociation in TrypLE Express Enzyme (Life Technologies, Belgium) using glass capillaries. Single blastomeres were washed and manually deposited into 384-well plates containing lysis buffer and Vapor-Lock. Plates were subsequently centrifuged at 1,000 rpm for 1 minute and stored at -80 °C until further processing. This study was approved by the Ghent University Institutional Review Board (EC2015/1114) and the Belgian Federal Commission for medical and scientific research on embryos in vitro (ADV_060_UZGent). All embryos were donated following patients' written informed consent.

5. scMspJI-seq

Prior to FACS or manual isolation of single cells, 384-well plates (BioRad) are prepared as follows: 4 µL of Vapor-Lock (Qiagen) is manually added to each well using a multichannel pipette followed by 2 µL of lysis buffer (0.2 µL of 25 µg/µL Qiagen Protease, 0.2 µL of 10x NEB Buffer 4 and 1.6 µL of nuclease-free water)

using the Nanodrop II liquid-handling robot (BioNex Solutions). All downstream dispensing steps are performed using the liquid-handling robot. After spinning down the 384-well plates, single cells are deposited into each well of the plate and incubated at 50°C for 15 hours, 75°C for 20 minutes and 80°C for 5 minutes. 5hmC sites in the genome are then glucosylated to block downstream recognition by MspJI by dispensing 0.5 µL of the following reaction mixture: 0.1 µL of T4-BGT (NEB), 0.1 µL of UDP-Glucose (NEB), 0.05 µL of 10x NEB Buffer 4 and 0.25 µL of nuclease-free water. After incubation at 37°C for 16 hours, 0.5 µL the following reaction mixture is added: 0.1 µL of 25 µg/µL Qiagen Protease, 0.05 µL of 10x NEB Buffer 4 and 0.35 µL of nuclease-free water. The plate is then incubated at 50°C for 5 hours, 75°C for 20 minutes and 80°C for 5 minutes. Thereafter, gDNA is digested by the restriction enzyme MspJI by the addition of 0.5 µL of the following reaction mixture: 0.02 µL of MspJI (NEB), 0.12 µL of 30x enzyme activator solution (NEB), 0.05 µL of 10x NEB Buffer 4 and 0.31 µL of nuclease-free water. The digestion is performed at 37°C for 5 hours followed by heat inactivation of MspJI at 65°C for 20 minutes. Next, 0.2 µL of cell-specific double-stranded adaptors are added to individual wells and these adapters are ligated to the fragmented gDNA molecules by adding 0.8 µL of the following reaction mixture: 0.07 µL of T4 DNA ligase (NEB), 0.1 µL of T4 DNA ligase buffer (NEB), 0.3 µL of 10 mM ATP (NEB) and 0.33 µL of nuclease-free water. The ligation is performed at 16°C for 16 hours. Next, wells containing unique cell-specific adapters are pooled using a multichannel pipette and incubated with 0.8x Agencourt Ampure (Beckman Coulter) beads for 30 minutes, washed twice with 80% ethanol and resuspended in 6.4 µL of nuclease-free water. Thereafter, *in vitro*

transcription and Illumina library preparation is performed as described previously in the scAba-seq protocol[34].

6. scMspJI-seq adapters

The double-stranded scMspJI-seq adapters are designed to contain a T7 promoter, 5' Illumina adapter, 3 bp UMI, 8 bp cell-specific barcode, and a random 4-nucleotide 5' overhang. The general design of the top and bottom strand is shown below:

Top oligo:

5' – CGATTGAGGCCGGTAATACGACTCACTATAGGGGTTCAGAGTTCTACA GTCCGACGATCNNN [8 bp cell-barcode] – 3'

Bottom oligo:

5' – NNNN [8 bp cell-barcode] NNNGATCGTCGGACTGTAGAACTCTGAACC CCTATAGTGAGTCGTATTACCGGCCTCAATCG – 3'

The sequence of the 8 bp cell-specific barcode is provided in Supplementary Table 2.1. The protocol for phosphorylating the bottom strand and for annealing the top and bottom strands to generate the double-stranded adapters is described previously in the scAba-seq protocol[34].

7. scMspJI-seq analysis pipeline

scMSpJI-seq libraries were sequenced on an Illumina NextSeq 500 platform. Reads containing the correct cell-specific barcode were mapped to the mouse (mm10) or human (hg19) genome using the Burrows-Wheeler Aligner (BWA) and

filtered for uniquely mapping reads to the genome. Custom scripts written in Perl were then used to demultiplex the data, identify 5mC position, strand information, and remove PCR duplicates. Custom code for analyzing scMspJI-seq data and the accompanying documentation is provided with this work[65].

8. Strand-specific scNMT-seq analysis pipeline

Bisulfite sequencing data from published scNMT libraries (GSE109262) were processed as described previously[52,92]. The first nine bases of the raw reads were trimmed using Trim Galore (v0.5.0) and mapped using Bismark (v20) to the mouse genome (mm10) with the 129/CAST background. SNPs specific to 129/CAST mouse genome were prepared using SNPsplit (v0.3.2) and a list of known variant call files from the Mouse Genomes Project (www.sanger.ac.uk/resources/mouse/genomes/). After mapping with Bismark, duplicate sequences were removed and CpG methylation calls were extracted with strand-specific information. Further data analysis and visualization of the methylation calls used custom scripts that will be made available upon request.

9. Hairpin Bisulfite Sequencing

Hairpin bisulfite sequencing was performed on bulk mouse embryos samples (2- to 64-cell stage mouse embryos). The embryos were treated with protease (1 µL of 25 µg/µL Qiagen Protease, 1 µL of 10x NEB Buffer 4, and 8 µL of nuclease-free water). Then, 0.5 ng of genomic DNA was digested with 20 µL of MspI master mix (1 µL of MspI (NEB), 2 µL 10x NEB CutSmart Buffer in a total volume of 20 µL) and incubated at 37°C for 1 hour. After digestion, the fragmented genomic DNA was ligated with 1 µL of 10 µM phosphorylated hairpin oligo mix (1 µL of NEB T4 ligase, 1 µL of 10x NEB T4 Ligase buffer, 2 µL of 10mM ATP, 5 µL of nuclease-free water)

and incubated overnight at 16°C. The hairpin oligo was prepared as follows: The oligo (5' - G/iMe-dC/iMe-dC/G/iMe-dC/iMe-dC/GG/iMe-dC/GG/iMe-dC/AAG/iBiodT/GAAG/iMe-dC/iMe-dC/G/iMe-dC/iMe-dC/GG/iMe-dC/G – 3') was resuspended in 100 µM of Low-TE. The hairpin oligo was then phosphorylated (1 µL of 100 µM hairpin oligo, 3 µL of 10x T4 Ligase Buffer, 1 µL T4 PNK and 5 µL of nuclease free water) and incubated at 37°C for an hour. Subsequently, the phosphorylated oligo was heated at 94°C and placed in ice water to generate the loop. For purification of the ligation mixture, Dynabeads™ M-280 Streptavidin beads were used following the recommended manufacturer's protocol with the following changes: the bead-ligation mixture was incubated for 1 hour at RT on a rotator and a cold 10 mM Tris-HCl wash step was included. Subsequently, we performed bisulfite sequencing on the sample using the protocol described previously[232]. After sequencing the libraries on a Miseq 300 bp or NextSeq 500 75 bp pair-end run, we used HBS-tools and custom Perl scripts to analyze the methylated CpG dyads[233].

10. Data Availability

Accession code GEO: GSE139984.

Figures associated with raw data: Figure 2.2a-e; Figure 2.3a-d; Figure 2.4a-f; Figure 2.5a,b; Supplementary Figure 2.1a,b; Supplementary Figure 2.2; Supplementary Figure 2.3a,b; Supplementary Figure 2.4a,b; Supplementary Figure 2.5; Supplementary Figure 2.6a-f; Supplementary Figure 2.7; Supplementary Figure 2.8a,b; Supplementary Figure 2.9a-e; Supplementary Figure 2.10a-c.

There are no restrictions on data availability.

11. Code Availability

Custom code for analyzing scMspJI-seq data and the accompanying documentation is provided online with this work[65].

B. Chapter 3 Methods

1. Mammalian cell culture

All mammalian cells were maintained in incubators at 37°C and 5% CO_2. HEK293T cells were cultured on tissue culture treated plastic in high glucose DMEM (Gibco, 10569044) containing L-Glutamine and sodium pyruvate, supplemented with 10% FBS (Gibco, 10437028) and 1x Penicillin-Streptomycin (Gibco, 15140122). H9 cells and hiPSCs (Allen Cell Collection, line AICS-0024) were grown feeder-free on Matrigel (Fisher Scientific, 08-774-552) coated plates in mTeSR1 medium (STEMCELL Tech., 85850). Cells were routinely passaged 1:6 once they reached 75% confluency using 0.25% trypsin-EDTA (Gibco, 25200056) for HEK293T cells, Versene solution (Gibco, 15040066) for H9 cells, and ReLeSR (STEMCELL Tech., 100-0484) for hiPSCs. For FACS sorting, a single-cell suspension was made using 0.25% trypsin-EDTA. The trypsin was then inactivated using serum containing medium. Afterwards, the cells were washed with 1x PBS before being passed through a cell strainer and sorted for single cells into 384-well plates.

2. hiPSC derived mesoderm cell culture

Mesoderm differentiation was performed on hiPSCs (Allen Cell Collection, line AICS-0024). For mesodermal differentiation, the media was replaced with mTeSR1 supplemented with 5 uM of CHIR99021 (STEMCELL Tech., 72052) for 24 hours. Afterwards, the cells were dissociated into a single cell suspension using TrypLE reagent (Gibco, 12563011), and gentle pipetting. TrypLE was then inactivated using

a serum containing medium. Next, the cells were washed with 1x PBS before being passed through a cell strainer and sorted for single cells into 384-well plates.

3. Post-implantation amniotic sac organoid culture

The development of micropatterned 3D stem-cell cultures with a single lumen is described in Karzbrun *et al.*[132]. We applied the protocol here as follows:

Microfabrication of PDMS stamps: PDMS stamps were created with circular features 250 μm in diameter. Stamps were prepared using standard soft-lithography techniques on a four-inch wafer. One layer of photoresist (Microchem, SU-8 2075) is spun onto a silicon wafer at a thickness of 100 μm. Photoresist is exposed to ultraviolet light using a mask aligner (Suss MicroTec, MA6) and unexposed photoresist is developed away to yield multiple arrays of posts. A trimethylchlorosilane layer is vapor deposited on the developed wafer to prevent adhesion. A 10:1 ratio of PDMS and its curing agent (Dow Corning, SYLGARD 184 A/B) is poured onto the wafers and cured at 65°C overnight. The PDMS layer is then peeled off the silicon mold and individual stamps are cut out using a razor blade for future use.

Micro-contact printing: Sterile PDMS stamps and 35 mm diameter custom-made glass-bottomed culture dishes are plasma treated for 1 minute on high setting (Harrick Plasma, PDC-32G) to activate both surfaces. Stamps are pressed features-side to the glass surface and held in place. To passivate the glass surface in nonpatterned regions, 0.1 mg/mL PLL-g-PEG solution (SuSoS AG, Switzerland) is added to the petri dish immediately after securing stamps to the glass surface and incubated for 30 minutes. Stamps are then carefully removed and stamped glass dishes are rinsed several times with PBS containing calcium and magnesium

(PBS++). Laminin-521 (STEMCELL Tech., 77003) is added at a dilution of 5 µg/mL in PBS++ to incubate overnight at 4°C. The following day, stamped glass dishes are rinsed with PBS++ to remove excess unbound, laminin and used within 1-2 weeks.

Stem-cell seeding, lumen formation and differentiation: On Day 1: hiPSCs (Allen Cell Collection, line AICS-0024) are released from well-plate surfaces using non-enzymatic agitation following manufacturer's instructions (ReleSR, STEMCELL Tech.). Cells are resuspended as a single-cell suspension at densities of 750K-1M cells/mL in mTeSR1 containing 10 µM ROCK inhibitor Y27632 (Abcam, ab120129). 200 µL of cell suspension is then pipetted onto prepatterned dishes and allowed to settle for 15 minutes before adding 1 mL of mTeSR1 and allowing cells to settle for 10 additional minutes. Excess media is aspirated, leaving enough liquid to cover patterns and is replaced with fresh 2 mL of mTeSR1. On day 2: mTESR1 media is exchanged with mTESR1 media containing Matrigel (4%, v/v). This triggers lumen formation over 24 hours. On Day 3: Micropatterned colonies have formed a lumen. mTESR1 is exchanged with the addition of 5 ng/ml Recombinant Human BMP4 (Fisher Scientific, 314BP010). Exposure to BMP4 triggers differentiation of cells and is considered 0 hours for experimental purposes. On Day 4 or 5: Samples were collected for single cell sequencing at either 20, 36, or 48 hours post BMP4 supplementation. Samples were dissociated into a single-cell suspension using TrypLE, and gentle pipetting. TrypLE was then neutralized using a serum containing medium. Next, the cells were washed with 1x PBS before being passed through a cell strainer. Finally, single cells were sorted into 384-well plates using FACS.

4. scMAT-seq

4 µL of Vapor-Lock (QIAGEN, 981611) was manually dispensed into each well of a 384-well plate using a 12-channel pipette. All downstream dispensing into 384-well plates were performed using the Nanodrop II liquid handling robot (BioNex Solutions). To each well, 100 nL of uniquely barcoded 7.5 ng/µL reverse transcription primers containing 4 nucleotide unique molecule identifiers (UMI) was added. The reverse transcription primers used here were previously described in Grun *et al.*[234]. Next, 100 nL of lysis buffer (0.175% IGEPAL CA-630, 1.75 mM dNTPs, 1:1,250,000 ERCC RNA spike-in mix (Ambion, 4456740), and 0.19 U RNase inhibitor (Clontech, 2313A)) was added to each well. Single cells were sorted into individual wells of a 384-well plate using FACS. Unless cryopreserved, after sorting, plates were heated to 65°C for 3 minutes and returned to ice. Next, 150 nL of reverse transcription GC tagging mix (0.7 U RNAseOUT (Invitrogen, 10777-019), 1.17x first strand buffer, 11.67 mM DTT, 3.5 U Superscript II (Invitrogen, 18064-071), 0.19 mM SAM, 1.17x GC reaction buffer, 0.1 U M.CviPI (NEB, M0227S)) was added to each well and the plates were incubated at 37°C for 1 hour, 4°C for 5 min, 65°C for 10 min, and 70°C for 10 min. Thereafter, 1.75 µL of second strand synthesis mix (1.74x second strand buffer (Invitrogen, 10812-014), 0.35 mM dNTP, 0.14 U E.coli DNA Ligase (Invitrogen, 18052-019), 0.56 U E.coli DNA Polymerase I (Invitrogen, 18010-025), 0.03 U RNase H (Invitrogen, 18021-071)) was added to each well and the plates were incubated at 16°C for 2 hours. Following this step, 400 nL of protease mix (6 µg protease (Qiagen, 19155), 6.25x NEBuffer 4 (NEB, B7004S)) was added to each well, and the plates were heated to 50°C for 15 hours, 75°C for 20 minutes, and 80°C for 5 minutes. Next, 500 nL of 5hmC-blocking mix (1 U T4-BGT (NEB, M0357L), 6x UDP-glucose, 1x NEBuffer 4) was added to each well

and the plates were incubated at 37°C for 16 hours. Afterwards, 500 nL of protease mix (2 µg protease, 1x NEBuffer 4) was added to each well, and the plates were heated to 50°C for 3 hours, 75°C for 20 minutes, and 80°C for 5 minutes. Next, 500 nL of MspJI digestion mix (1x NEBuffer 4, 8x enzyme activator solution, 0.1 U MspJI (NEB, R0661L)) was added to each well and the plates were incubated at 37°C for 4.5 hours, and 65°C for 25 minutes. Unless otherwise noted, to each well, 320 nL of uniquely barcoded 125 nM double-stranded adapters were added. The double-stranded adapters have previously been described in Sen *et al.*[65]. Next, 680 nL of ligation mix (1.47x T4 ligase reaction buffer, 5.88 mM ATP (NEB, P0756L), 140 U T4 DNA ligase (NEB, M0202M)) was added to each well, and the plates were incubated at 16°C for 16 hours. After ligation, reaction wells receiving different barcodes were pooled using a multichannel pipette, and the oil phase was discarded. The aqueous phase was then incubated for 30 minutes with 1x AMPure XP beads (Beckman Coulter, A63881), placed on a magnetic stand and washed twice with 80% ethanol before eluting the DNA in 30 µL of nuclease-free water. After vacuum concentrating the elute to 6.4 µL, library preparation was performed as previously described in the scAba-seq and scMspJI-seq protocols[34,65]. Libraries were sequenced on an Illumina NextSeq 500 or an Illumina Hiseq 4000, sequencing a minimum of 75 bp on read 1 to detect methylated cytosines. A minimum of 25 bp on read 1 and 50 bp on read 2 was used to detect mRNA. Additionally, unless otherwise stated, detection of mRNA was only performed on RNA enriched samples and detection of methylated cytosines was performed only on unenriched samples.

5. *Optimizing buffer for simultaneous reverse transcription and GpC methylation tagging*

In all experiments, 200 nL of the reverse transcription GC tagging mix was used and the 384-well plate was incubated at 37°C for 1 hour, 4°C for 5 minutes, 65°C for 10 minutes, and 70°C for 10 minutes. For the first strand buffer experiment, this reverse transcription GC tagging mix consisted of 0.7 U RNAseOUT, 2x first strand buffer, 20 mM DTT, 3.5 U Superscript II, 0.16 mM SAM, and 0.1 U M.CviPI. For the GC buffer experiment, this mix consisted of 0.7 U RNAseOUT, 3.5 U Superscript II, 0.16 mM SAM, 2x GC reaction buffer, 0.1 U M.CviPI, and 6 mM of $MgCl_2$. For the 50:50 experiment, this reverse transcription GC tagging mix consisted of 0.7 U RNAseOUT, 1x first strand buffer, 1x GC reaction buffer, 10 mM DTT, 3.5 U Superscript II, 0.16 mM SAM, and 0.1 U M.CviPI. Following this, all other library construction steps were similar to the optimized scMAT-seq procedure. The volume of some reactions were altered as follows. Second strand synthesis was performed by adding 1.3 µL of second strand synthesis mix. The initial protease step was performed by adding 300 nL of protease mix. No 5hmC-blocking mix or secondary protease mix was added. After MspJI digestion, 200 nL of uniquely barcoded 1 nM or 200 nM double-stranded adapters were added. Afterwards ligation was performed by adding 800 nL of ligation mix. All other steps of library construction were unchanged.

6. RNA enrichment

After *in vitro* transcription, 6 µL of amplified RNA product (aRNA) was combined with 2 µL of 1 µM biotinylated polyA primer (Integrated DNA Technologies, standard desalting, 5'- AAAAAAAAAAAAAAAAAAAAAA/3BioTEG/ -3') and incubated for 10 minutes at room temperature. During this incubation, Dynabeads MyOne Streptavidin C1 beads (Invitrogen, 65001) were made RNase-free following the

directions of the manufacturer. In addition, 2x and 1x B&W solution was made according to the manufacturer's directions. After establishing RNase-free conditions, the beads were resuspended in 8 µL of 2x B&W solution. After the 10-minute incubation of aRNA with the biotinylated polyA primer, the beads were mixed in with the solution and incubated for 15 minutes at room temperature with constant shaking at 300 rpm using a thermomixer. Using a magnetic stand, the beads were separated from the supernatant, and the supernatant was discarded. The beads were washed twice with 1x B&W solution. After washing, the beads were resuspended in 10 µL of nuclease-free water. For on-bead processing, this product was taken to reverse transcription. For heat denatured processing, this product was heated to 70°C for 2 minutes, then a strong magnet was used to quickly separate the beads from the supernatant, and the supernatant was transferred to a new tube. In both cases, 5 µL of RNA enriched product was used for reverse transcription, and following this step, library preparation was performed as previously described in the scAba-seq and scMspJI-seq protocols[34,65]. For experiments using other bead types, the C1 beads were exchanged for other streptavidin beads, M270, M280, or T1 (Invitrogen, 65801D), with no other changes.

7. scMAT-seq analysis pipeline

The scMAT-seq analysis pipeline was performed as described previously in Sen *et al.* with minor adjustments[65]. The custom Perl script used to identify 5mC positions in the genome was modified to interrogate the base prior to the called 5mC mark. Called 5mC marks preceded by a G were assigned to the DNA accessibility dataset, and due to the potential off target activity of M.CviPI, only those preceded by an A or T were assigned as endogenous 5mC marks, that were further split by

their context (5mCpG, 5mCpA, 5mCpC, or 5mCpT). Custom codes for analyzing scMAT-seq data and the accompanying documentation is available upon request.

The transcriptome analysis pipeline was similar to that described previously in Grün *et al.* with the following minor adjustments[234]. The right mate of paired-end reads was mapped in the sense direction using BWA (version 0.7.15-r1140) to the RefSeq gene model based on the human genome release hg19, with the addition of the set of 92 ERCC spike-in molecules. Any read mapping to multiple loci were distributed uniformly across those loci. Gene isoforms were consolidated into a gene count, and the UMIs were used to deduplicate reads and provide single-molecule transcript counts for each gene in individual cells[234]. Genes that were not detected in at least one cell were removed from any downstream analysis.

8. Comparison of scMAT-seq to established techniques

DNase I hypersensitivity sites for H9 and HEK293T cells were downloaded from UCSC table browser and sites were grouped based on detection scores[173]. Additionally, the genomic coordinates of CpG islands were downloaded from the UCSC table browser. To compare across datasets, the data was normalized to counts per million and a 75 base pair moving average was plotted for each region of interest. When comparing differing genomic regions within a sample, each region was further normalized by methylated cytosines that were detected when MspJI-seq was performed on bulk H9 gDNA that had been GpC methylated after stripping off chromatin. To do this, H9 gDNA was isolated using the DNeasy Blood & Tissue Kit (Qiagen, 69504). To 1.2 µg of purified H9 gDNA, 10 µL of protease mix (100 µg protease (Qiagen, 19155), 1x GC reaction buffer) was added, and the sample was heated to 50°C for 15 hours, 75°C for 20 minutes, and 80°C for 5 minutes. Next, 10

µL of GC tagging mix (8 U M.CviPI, 640 µM SAM, 1x GC reaction buffer) was added, and the sample was incubated at 37°C for 4 hours. Immediately after, 10 µL of additional GC tagging mix (4 U M.CviPI, 960 µM SAM, 1x GC reaction buffer) was added, and the sample was incubated at 37°C for 4 hours and 65°C for 20 minutes. Afterwards, 100 ng of this DNA was directly used as input to a scaled-up version of the scMspJI-seq protocol[65].

5mC sites detected by scMAT-seq in H9 cells were compared to bulk bisulfite sequencing (GSM706061)[174]. A site was considered methylated if any level of 5mC was detected in the bulk bisulfite or pseudo-bulked scMAT-seq sequencing data. Overlapping 5mCpG sites were counted and compared to the number of non-overlapping sites.

9. *Cluster calling for genome-wide detection of DNA accessibility and 5mC in scMAT-seq*

DNA accessibility and 5mC were quantified within 5 kb bins and then converted to binary scores. Further, pseudobulk profiles were generated using the assigned cell type from the transcriptome. For comparison between H9 and HEK293T cell lines, the top 2% most variable bins between groups were retained. For comparison in the post-implantation amniotic sac organoid, the top 1% most variable bins between groups were retained. After removal of bins with low variance, principal component analysis was performed on the remaining bins for individual cells, and hierarchical clustering was used to assign clusters. Cluster identification was performed through comparison to transcriptome derived cell types, where high similarity was observed between cluster calling for all 3 measurements.

10. Promoter and gene body DNA accessibility and gene body 5mC analysis in scMAT-seq

For the quantification of DNA accessibility and 5mC in individual cells, a small pseudo-count was added prior to estimating the number of UMIs per million counts for each gene. The promoter of a gene was considered as 2,000 base pairs upstream of the transcription start site. Genes with low detection in all cells were removed from downstream analysis for that epigenetic feature.

11. Gene expression analysis

The standard analysis pipeline in Seurat (version 3.1.5) was used for single-cell RNA expression normalization and analysis[235]. For H9 and HEK293T cell lines, cells containing more than 1,000 genes and more than 1,000 unique transcripts, as well as less than 20% ERCC spike-ins, were used for downstream analysis. For the post-implantation amniotic sac organoid, cells containing more than 1,000 genes and more than 4,000 unique transcripts, as well as less than 20% ERCC spike-ins, were used for downstream analysis. The default NormalizeData function was used to log normalize the data. In post-implantation amniotic sac organoids, the cluster identified as NELCs was removed from downstream analysis except where otherwise stated. When analyzing the time course data from the organoid, the FindIntegrationAnchors and IntegrateData functions were used to remove batch- and technique-specific effects. Thereafter, principal components were obtained from the 2,000 most variable genes and the elbow method was used to determine the optimal number of principal components used in clustering. UMAP based clustering was performed by running the following functions, FindNeighbors, FindClusters, and RunUMAP. After clustering, cell types were assigned to groups using known expression markers. To

identify DEGs, the FindAllMarkers or FindMarkers function was used. The Wilcoxon rank sum test was used to classify a gene as differentially expressed, requiring a natural log fold change of at least 0.25 and an adjusted p-value of less than 0.05.

12. Pseudotime analysis

UMAP coordinates and normalized gene expression data for highly variable genes was imported from Seurat to Monocle3 (version 0.2.1.5)[182]. Trajectories were built using the learn_graph function. Cells involved in the four observed trajectories were isolated separately using the choose_graph_segments function. Due to the bifurcation seen in the trajectories, some cells appeared in more than one of the trajectories. For each trajectory, the roots of the trajectories were chosen using the order_cells function to best correspond with cells from the 20-hour post-implantation amniotic sac organoids. After assigning a pseudotime to each cell, the genes varying over each pseudotime were determined using the graph_test function, with genes with a q-value under 0.01 and a Moran's I value above 0.15 considered significant. Significantly varying genes for each trajectory were grouped into gene modules using the find_gene_modules function using a resolution value of 0.05. z-scores for gene expression of each gene module was calculated using the aggregate_gene_expression function. The corresponding z-score for DNA accessibility and 5mC was found and a 10-cell moving average was computed based on the pseudotime. For averaging and plotting, only cells passing quality controls for RNA expression, DNA accessibility and 5mC were considered.

13. Data availability

Sequencing data have been deposited in the Gene Expression Omnibus (GEO) database accession code GEO: GSE181724.

C. Chapter 4 Methods

1. Mammalian cell culture

H9 human embryonic stem cells were grown as described previously in scMAT-seq (chapter 3).

2. PGCLC formation and long-term culture

UCLA2 human embryonic stem cells were cultured and induced into PGCLCs using an incipient mesoderm-like cell intermediate and the creation of disorganized 3d aggregates, as previously described[236]. After 4 days in 3d culture, PGCLCs were sorted and cultured in extended culture conditions containing FR10 medium, as described previously[134]. TRA-1-85 positive single hPGCLCs were isolated into each reaction well of a 384-well plate for scMTH-seq, as described previously[134].

3. scMTH-seq

scMTH-seq processing is like that described in scDyad&T-seq up until the second protease step (chapter 5). One minor difference is that the 5hmC-blocking mix also contained 200 fg of mouse brain DNA (VWR, 76020-078) as a spike-in for estimating 5mC and 5hmC levels between cells. After the second protease step, 500 nL of glucosylated 5hmC digestion mix was added (1x NEBuffer 4 (NEB, B7004S), 1 U AbaSI (NEB, R0665S)) to each well and the plates were incubated at 25°C for 90 minutes, and 65°C for 25 minutes. Next 250 nL of a third protease mix was added (2 µg protease (Qiagen, 19155), and 1x NEBuffer 4) was added to each well, and the plates were heated to 50°C for 3 hours, 75°C for 20 minutes, and 80°C for 5 minutes. Next, 500 nL of MspJI digestion mix (1x NEBuffer 4, 9.5x enzyme activator solution, and 0.1 U MspJI (NEB, R0661L)) was added to each well and the plates were incubated at 37°C for 4.5 hours, and 65°C for 25 minutes.

To each well, 200 nL of uniquely barcoded 20 nM phosphorylated scAba-seq compatible double-stranded adapters were added. Then to each well, 120 nL of uniquely barcoded 125 nM phosphorylated scMspJI-seq compatible double-stranded adapters were added. The sequences of scAba-seq and scMspJI-seq compatible double-stranded adapters have previously been reported[34,65]. Next, 680 nL of ligation mix (1.47x T4 ligase reaction buffer, 6.99 mM ATP (NEB, P0756L), and 140 U T4 DNA ligase (NEB, M0202M)) was added to each well, and the plates were incubated at 16°C for 16 hours. After ligation, reaction wells receiving different barcodes were pooled using a multichannel pipette, and the oil phase was discarded. Library preparation and DNA sequencing for mRNA enriched and non mRNA enriched samples was performed as in scMAT-seq (chapter 3).

4. scMTH-seq analysis pipeline

All sequencing reads were trimmed to 76 bases. Then 5mC, 5hmC, and transcriptome-based reads were separated based on feature specific barcodes. After this, the scMTH-seq analysis pipeline was performed as described previously in Sen *et al.* and Mooijman *et al.*[34,65]. Hg19 and mm10 were used for mapping, with 5mC marks attributed to the mouse genome considered spike-in detections. The transcriptome analysis pipeline was previously described in scMAT-seq. Each feature, 5mC, 5hmC and gene expression were separately analyzed for data quality. If a cell contained at least 30,000 5mC, 300 5hmC, 4,000 transcripts and 1,000 detected genes it was considered successfully amplified in all features. In some cases, a cell only contained high quality information from one or two of these features and so it was used in the analysis only when the cell had high quality data for the feature being analyzed.

5. Gene expression analysis

The standard analysis pipeline in Seurat (version 3.1.5) was used for single-cell RNA expression normalization and analysis[235]. Cells containing more than 1,000 genes and more than 4,000 unique transcripts, as well as less than 20% ERCC spike-ins, were used for downstream analysis. The default NormalizeData function was used to log normalize the data. Thereafter, principal components were obtained from the 2,000 most variable genes and the elbow method was used to determine the optimal number of principal components used in clustering. UMAP based clustering was performed by running the following functions, FindNeighbors, FindClusters, and RunUMAP. After clustering, cell types were assigned to groups using known expression markers. To identify DEGs, the FindAllMarkers or FindMarkers function was used. The Wilcoxon rank sum test was used to classify a gene as differentially expressed, requiring a natural log fold change of at least 0.3 and an adjusted p-value of less than 0.01. Cell cycle analysis was performed as described in the Seurat cell cycle vignette using cell cycle genes derived previously[237].

6. Turnover rate modeling

The turnover of each feature (5hmC, 5mCpA, 5mCpT, 5mCpC) was modeled separately using autosomal chromosome detection data. Features were modeled using model IV described previously by Mooijman *et al*.[34]. For each feature, 100 simulations were performed.

7. Code availability

Codes for analyzing scMTH-seq data and the accompanying documentation has been described previously by Sen et al. and Mooijman et al.[34,65].

D. Chapter 5 Methods

1. Mammalian cell culture

All mammalian cells were maintained in incubators at 37°C and 5% CO_2. Mouse embryonic cell line ES-E14TG2a (E14) were grown on gelatin (Millipore Sigma, ES-006-B) coated tissue culture plates. Media was created using high glucose DMEM (Gibco, 10569044), 1% non-essential amino acid (Gibco, 11140050), 1% Glutamax (Gibco, 35050061), 1x Penicillin-Streptomycin (Gibco, 15140122), and 15% stem cell qualified serum (Millipore Sigma, ES-009-B). The media was frozen in aliquots and used for a maximum of 2 weeks after thawing, kept constantly at 4°C. Once thawed, 1 µL of beta-mercaptoethanol (Gibco, 21985023), and 1 µL of LIF (Millipore, ESG1106) were added for every 1 mL of thawed media. Daily, the cells were washed with 1x DPBS (Gibco, 14190250) and the media was exchanged. Cells were routinely passaged 1:6 once they reached 75% confluency using 0.25% trypsin-EDTA (Gibco, 25200056). E14 cells grown under these conditions also describe the SL experimental group. For FACS sorting, a single-cell suspension was made using 0.25% trypsin-EDTA. The trypsin was then inactivated using serum containing medium. Afterwards, the cells were washed with 1x DPBS before being passed through a cell strainer and sorted for single cells into 384-well plates.

K562 cells were grown in RPMI (Gibco, 61870036) with 10% serum (Gibco, 10437028) and 1x Penicillin-Streptomycin. When cells reached a density of approximately 1 million cells per mL, they were split and resuspended at a density of 200,000 cells per mL. Cells were washed and FACS sorted as described for E14.

2. 24-hour Decitabine culture

E14 mouse embryonic stem cells were cultured as described. Upon passage of the E14 cells, SL media was supplemented with 0.05 μM of Decitabine. After 24 hours, cells were lifted using 0.25% trypsin-EDTA. The trypsin was then inactivated using serum containing medium. The cells were washed with 1x DPBS and then resuspended in 200 μL of DPBS. Genomic DNA was extracted using the DNeasy kit (Qiagen, 69504) according to the manufacturer's recommendations.

K562 cells were cultured as described. Upon passage the media was supplemented with 0.6 μM of Decitabine. After 24 hours the cells were washed and FACS sorted as described previously here.

3. 48-hour 2i media component experiment

E14 mouse embryonic stem cells were cultured as described. Upon passage, cells were resuspended in the following media according to their condition. Commercial 2i media containing LIF (Millipore, SF016-200) was used for BL, G, 2i, and M experiments. For 2i, all components were used according to the manufacturer's recommendations. For G and M, only the GSK3B inhibitor and MEK1/2 inhibitor was added respectively. For BL, no inhibitors were introduced. For No, commercial 2i media without LIF (Millipore, SF002-100) was used with no inhibitors introduced. After 24 hours, the cells were washed with 1x DPBS and the media was exchanged. 48 hours after the initial media switch, the cells were collected using 0.25% trypsin EDTA, quenched using serum containing media, washed in 1x DPBS and finally resuspended in 1x DPBS. The sample was then split in half. One half was resuspended in 200 μL of DPBS and had its genomic DNA extracted as described previously. The other half was resuspended in 500 μL of

TRIzol reagent (Invitrogen, 15596018) and total RNA was extracted according to the manufacturer's recommendations. Each condition was performed in triplicate.

4. Chip-seq data processing

The following published ChIP datasets were used in this study (GEO accessions): GSM1000123 (H3K9ac), GSE74055 (H3K9me1 and H3K27ac), GSE23943 (H3K4me3, H3K9me3, H3K27me3, and H3K36me3), and GSE77420 (H3K9me2). For all, the processed data file was downloaded from GEO and further processed if needed. For GSE74055, in 1kb bins the bigwigCompare tool on Galaxy (version 2.1.1.20160309.6) was used to compare enriched datasets to the input data, bins with a log2 enrichment score greater than 2 were considered enriched regions. For GSE23943, peak calling was performed using MACS2 on Galaxy, the resulting narrow peaks file was used as enriched regions. For GSE77420, in 2kb bins the enrichment score for serum grown H3K9me2 was compared to the input serum score. Regions were considered enriched if the H3K9me2 score was greater than the input score for both replicates. When applicable, enriched regions were converted from mm9 to mm10 using the UCSC genome browser LiftOver tool.

5. Dyad-seq Adapters

The double stranded Dyad-seq adapters are designed to be devoid of cytosine on the bottom strand. They contain a PCR sequence, a 4-base pair UMI, and a 10-base pair cell-specific barcode. For Dyad-seq using MspJI as a restriction enzyme (M-M-Dyad-seq and M-H-Dyad-seq), the adapters contain a 4 base pair 5' overhang.

For bulk Dyad-seq, cell-specific barcodes were used as replicate specific barcodes.

Top oligo: 5’- NNNN [10 base pair barcode] HHWHCCAAACCCACTACACC -3’

Bottom oligo: 5’- GGTGTAGTGGGTTTGGDWDD [10 base pair barcode] -3’

The sequence of the 10 base pair cell-specific barcode for scDyad&T-seq is provided in Supplementary Table 5.1. This design and the first three barcodes were also used for M-H-Dyad-seq.

For M-M-Dyad-seq, a prototype of this design was used consisting of a 3-base pair UMI and an 8 base pair sample-specific barcode (Supplementary Table 5.2).

Top oligo: 5’- NNNN [8 base pair barcode] HHHCCAAACCCACTACACC -3’

Bottom oligo: 5’- GGTGTAGTGGGTTTGGDDD [8 base pair barcode] -3’

The sequence of the 8 base pair cell-specific barcode for M-MDyad-seq is provided in Supplementary Table 5.2.

For Dyad-seq using AbaSI as a restriction enzyme (H-H-Dyad-seq and H-M-Dyad-seq), the adapters instead contain a 2 base pair 3’ overhang.

Top oligo: 5’- [10 base pair barcode] HHWHCCAAACCCACTACACC -3’

Bottom oligo: 5’- GGTGTAGTGGGTTTGGDWDD [10 base pair barcode] NN -3’

The sequence of the 10 base pair cell-specific barcode for H_M-CpG-Dyad-seq and H_H-CpG-Dyad-seq is provided in supplementary Table 5.3.

All adapters were left unphosphorylated. The protocol for annealing the top and bottom strands to create double-stranded adapters is described previously in scAba-seq[34].

6. Bulk CpG-Dyad-seq

For all bulk CpG-Dyad-seq 100 ng of purified genomic DNA was resuspended in 20 µL of 5hmC blocking mix (1x CutSmart buffer (NEB, B7204S), 2.5x UDP-glucose, 10 U T4-BGT (NEB, M0357L)) and the samples were incubated at 37°C for 16 hours. Afterwards, 10 µL of protease mix (100 µg protease (Qiagen, 19155), 1x CutSmart buffer) was added to each sample, and the samples were heated to 50°C for 5 hours, 75°C for 20 minutes, and 80°C for 5 minutes. After this point, samples differed based on sub-type of CpG-Dyad-seq used.

For M-M-Dyad-seq, 10 µL of MspJI digestion mix (2 U MspJI, 1x enzyme activator solution, 1x CutSmart buffer) was added to each sample and the samples were heated to 37°C for 5 hours, and 65°C for 20 minutes. Next 1 µL of barcoded 1 µM M-M-Dyad-seq adapter was added. Then 9 µL of ligation mix (1.11x T4 ligase reaction buffer, 4.44 mM ATP (NEB, P0756L), 2000 U T4 DNA ligase (NEB, M0202M)) was added to each sample, and the samples were incubated at 16°C for 16 hours.

For M-H-Dyad-seq, all steps were the same as M-M-Dyad-seq with the exception that 1 µL of barcoded 1 µM M-H-Dyad-seq adapter was added instead of the M-M-Dyad-seq adapter.

For H-M-Dyad-seq and H-H-Dyad-seq, 10 µL of AbaSI digestion mix (10 U AbaSI (NEB, R0665S), 1x CutSmart buffer) was added to each sample and the samples were heated to 25°C for 2 hours, and 65°C for 20 minutes. Next 1 µL of barcoded 1

μM AbaSI-Dyad adapter was added. Then 9 μL of ligation mix (1.11x T4 ligase reaction buffer, 4.44 mM ATP (NEB, P0756L), 2000 U T4 DNA ligase (NEB, M0202M)) was added to each sample, and the samples were incubated at 16°C for 16 hours.

After ligation, up to three barcoded libraries of the same type were pooled and all Dyad-seq types were subjected to a 1x AMPure XP bead cleanup (Beckman Coulter, A63881), and eluted in 40 μL of water.

For M-M-Dyad-seq and H-M-Dyad-seq were then evaporated to a total of 28 μL and subjected to nucleobase conversion using the NEBNext enzymatic methyl-seq conversion module (NEB, E7125S) according to the manufacturer's recommendations apart from doing the final elution step in 40 μL of water. Bisulfite conversion can also be used but results in less complex sequencing libraries (data not shown).

For M-H-Dyad-seq and H-H-Dyad-seq nucleobase conversion was performed using the NEBNext enzymatic methyl-seq conversion module similarly to what has been described previously[49]. Briefly, samples were evaporated to a total of 17 μL. Then 4 μL of formamide (Sigma-Aldrich, F9037-100ML) was added and the samples were heated to 85°C for 10 minutes before being quenched on ice. APOBEC nucleobase conversion was performed as described by the manufacturer except the incubation was held at 37°C for 16 hours. After which the manufacturer's recommendations were followed apart from doing the final elution step in 40 μL of water.

To the nucleobase converted libraries all Dyad-seq types were subjected to one round of linear amplification. To do this, 9 μL of amplification mix was added (5.56x

NEBuffer 2.1 (NEB, B7202S), 2.22 mM dNTPs (NEB, N0447L), and 2.22 uM Linear amplification 9-mer: 5’- GCCTTGGCACCCGAGAATTCCANNNNNNNNN -3’) and the samples were heated to 95°C for 45 seconds before being quenched on ice. Once cold, 100 U of high concentration Klenow DNA polymerase (3’-5’ Exo-) (fisher scientific, 50-305-912) was added. Then samples were quickly vortexed, spun down and then incubated at 4°C for 5 minutes, followed by an increase of 1°C every 15 seconds at a ramp rate of 0.1°C per second until the samples reach 37°C which was then held for an additional 1.5 hours. Afterwards a 1.1x AMPure XP bead cleanup was performed, and the samplers were eluted in 40 µL of water before being evaporated down to 10 µL. The entire sample was then used in a linear PCR reaction by adding 15 µL of solution (1.67x high-fidelity PCR mix (NEB, M0541L) and 0.67 µM Extended RPI primer: 5’-AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGACGA TCGGTGTAGTGGGTTTGG-3’) and performing PCR as follows, an initial denaturing of 98°C for 30 seconds, followed by 16 cycles of 98°C for 10 seconds, 59°C for 30 seconds, and 72°C for 30 seconds, after these cycles the samples were held at 72°C for 1 minute. Next 5 µL of linear PCR product was amplified further in a standard Illumina library PCR reaction, incorporating a unique indexed i7 primer. The remaining linear PCR product was stored at -20°C. Two 0.825x AMPure XP bead cleanups were performed in the sequencing libraries with a final elution of 15 µL of water. The libraries were then quantified on an Agilent Bioanalyzer and Qubit fluorometer. Finally, libraries were subjected to Illumina sequencing on the HiSeq plateform obtaining 150 bp reads from both ends.

7. Bulk RNA-seq

Total RNA was extracted using TRIzol (Ambion, 15596018). 50 ng of total RNA was heated to 65°C for 5 minutes and returned to ice. Afterwards it was combined with 9 uL of reverse transcription mix (20 U RNAseOUT (Invitrogen, 10777-019), 1.11x first strand buffer, 11.11 mM DTT, 0.56 mM dNTPs (NEB, N0447S), 100 U Superscript II (Invitrogen, 18064-071), and 25 ng of barcoded reverse transcription primer) and the sample was incubated at 42°C for 75 minutes, 4°C for 5 minutes, and 70°C for 10 minutes. Each replicate received a different barcoded reverse transcription primer. The reverse transcription primers used here were described previously[238]. Afterwards, 50 µL of second strand synthesis mix (1.2x second strand buffer (Invitrogen, 10812-014), 0.24 mM dNTPs (NEB, N0447S), 4 U E.coli DNA Ligase (Invitrogen, 18052-019), 15 U E.coli DNA Polymerase I (Invitrogen, 18010-025), 0.8 U RNase H (Invitrogen, 18021-071)) was added to each sample and the samples were incubated at 16°C for 2 hours. The barcoded replicates were then pooled, and a 1x AMPure XP bead (Beckman Coulter, A63881) cleanup was performed, eluting in 30 µL of water, which was subsequently evaporated to 6.4 µL. The molecules were amplified with IVT and an Illumina sequencing library was prepared as described in CEL-Seq2[171]. Libraries were subjected to Illumina sequencing on the HiSeq plateform obtaining 150 bp reads from both ends.

8. Bulk RNA-seq analysis

Bulk RNA-seq data was processed as described previously (scMAT-seq, chapter 3) with the following modification, reads were mapped to the RefSeq gene model based on the mouse genome release mm10, with the addition of the set of 92 ERCC spike-in molecules.

DESeq2 was used for normalization and differential gene expression calling[239]. Each condition was compared pairwise and adaptive shrinkage was used to adjust the log fold change observed[240]. For differential gene calling an adjusted p-value cutoff of 0.01 and a shrunken log fold change cutoff of 0.75 was used. For visualization and clustering, variance stabilizing transformation was performed and batch effects from differing reverse transcription primer barcodes was removed using the removeBatchEffect function in the limma package[241].

9. scDyad&T-seq

4 µL of Vapor-Lock (QIAGEN, 981611) was manually dispensed into each well of a 384-well plate using a 12-channel pipette. All downstream dispensing into 384-well plates were performed using the Nanodrop II liquid handling robot (BioNex Solutions). To each well, 100 nL of uniquely barcoded 7.5 ng/µL reverse transcription primers containing 6 nucleotide unique molecule identifiers (UMI) was added. The reverse transcription primers used here were previously described in Grun *et al.* with the exception that a UMI length of 6 was used[234]. Next, 100 nL of lysis buffer (0.175% IGEPAL CA-630, 1.75 mM dNTPs (NEB, N0447S), 1:1,250,000 ERCC RNA spike-in mix (Ambion, 4456740), and 0.19 U RNase inhibitor (Clontech, 2313A)) was added to each well. Single cells were sorted into individual wells of a 384-well plate using FACS and stored at -80°C until used. To begin processing, plates were heated to 65°C for 3 minutes and returned to ice. Next, 150 nL of reverse transcription mix (0.7 U RNAseOUT (Invitrogen, 10777-019), 2.33x first strand buffer, 23.33 mM DTT, and 3.5 U Superscript II (Invitrogen, 18064-071)) was added to each well and the plates were incubated at 42°C for 75 minutes, 4°C for 5 minutes, and 70°C for 10 minutes. Thereafter, 1.5 µL of second strand synthesis mix

(1.23x second strand buffer (Invitrogen, 10812-014), 0.25 mM dNTPs (NEB, N0447S), 0.14 U E.coli DNA Ligase (Invitrogen, 18052-019), 0.56 U E.coli DNA Polymerase I (Invitrogen, 18010-025), 0.03 U RNase H (Invitrogen, 18021-071)) was added to each well and the plates were incubated at 16°C for 2 hours. Following this step, 650 nL of protease mix (6 µg protease (Qiagen, 19155), 3.85x NEBuffer 4 (NEB, B7004S)) was added to each well, and the plates were heated to 50°C for 15 hours, 75°C for 20 minutes, and 80°C for 5 minutes. Next, 500 nL of 5hmC-blocking mix (1 U T4-BGT (NEB, M0357L), 6x UDP-glucose, 1x NEBuffer 4) was added to each well and the plates were incubated at 37°C for 16 hours. Afterwards, 500 nL of protease mix (2 µg protease, 1x NEBuffer 4) was added to each well, and the plates were heated to 50°C for 3 hours, 75°C for 20 minutes, and 80°C for 5 minutes. Next, 500 nL of MspJI digestion mix (1x NEBuffer 4, 8x enzyme activator solution, and 0.1 U MspJI (NEB, R0661L)) was added to each well and the plates were incubated at 37°C for 4.5 hours, and 65°C for 25 minutes. To each well, 280 nL of uniquely barcoded 250 nM unphosphorylated double-stranded adapters were added. Next, 720 nL of ligation mix (1.39x T4 ligase reaction buffer, 5.56 mM ATP (NEB, P0756L), 140 U T4 DNA ligase (NEB, M0202M)) was added to each well, and the plates were incubated at 16°C for 16 hours. After ligation, reaction wells receiving different barcodes were pooled using a multichannel pipette, and the oil phase was discarded. The aqueous phase was then incubated for 30 minutes with 1x AMPure XP beads (Beckman Coulter, A63881), placed on a magnetic stand and washed twice with 80% ethanol before eluting the DNA in 30 µL of nuclease-free water. After vacuum concentrating the elute to 6.4 µL, *in vitro* transcription (IVT) was performed as previously described in the scAba-seq and scMspJI-seq protocols[34,65]. RNA

enrichment from IVT product was performed as described in scMAT-seq (chapter 3) with the following modifications. The entire IVT product was used for enrichment, 4 µL of biotinylated polyA primer, and 8 µL of Dynabeads MyOne Streptavidin C1 beads (Invitrogen, 65001) were used and resuspended in 24 µL of 2x B&W solution after establishing RNase-free conditions. Additionally, the supernatant of material after combining all three components was saved for additional processing.

The supernatant of the RNA enrichment contains unamplified barcoded scDyad-seq DNA molecules. A 1x AMPure XP bead cleanup was performed using a 30 minute incubation and was eluted in 40 µL of water. Samples were then evaporated to 28 µL and nucleobase conversion was performed as described for bulk M-M-Dyad-seq. Samples were then subjected to four rounds of linear amplification. The first round was the same as described for bulk Dyad-seq. In subsequent rounds samples were first heated to 95°C for 45 seconds before being quenched on ice. Once cold, 5 µL of amplification mix was added (1x NEBuffer 2.1 (NEB, B7202S), 2 mM dNTPs (NEB, N0447L), 2 uM Linear amplification 9-mer, and 10 U of high concentration Klenow DNA polymerase (3'-5' Exo-) (fisher scientific, 50-305-912)). Then samples were quickly vortexed, spun down and then incubated the same as performed in the first round of linear amplification. After 4 rounds of linear amplification, sequencing libraries were prepared the same way as described for bulk Dyad-seq. Libraries were subjected to Illumina sequencing on the HiSeq plateform obtaining 150 bp reads from both ends.

scDyad-seq is performed similarly to scDyad&T-seq, except the initial reverse transcription and second strand synthesis steps are replaced with dispenses of 1x

NEBuffer 4. Additionally, because the transcriptome is not captured, IVT is not performed and steps involving aRNA enrichment and processing are omitted.

10. Dyad-seq analysis pipeline

CpG dyads containing potential information from both DNA strands was analyzed separately from non-dyad methylation or hydroxymethylation levels. To analyze CpG dyads, read 1 was trimmed to 86 base pairs, and then exact duplicates were removed using Clumpify from BBTools. Next, reads containing the correct PCR amplification sequence and correct barcode were extracted. These reads were then trimmed using the default settings of TrimGalore. For mapping, Bismark was used in conjunction with Bowtie2 v2.3.5 to map to MM10[242]. For experiments using K562 cells, hg19 was used. After mapping, Bismark was used to further deduplicate samples based on UMI, cell barcode and mapping location. For library preparation using MspJI, a custom Perl script was used to identify 5mC positions in the genome as detected by MspJI and interrogate the methylation status of the opposing cytosine in a CpG or CpHpG dyad context from the nucleobase conversion. For library preparation using AbaSI, a custom Perl script was used to identify 5hmC positions in the genome as detected by AbaSI and interrogate the methylation status of the opposing cytosine in a CpG dyad context from the nucleobase conversion. To analyze non-dyad methylation or hydroxymethylation, the cell barcode and UMI were transferred from read 1 to read 2. Read 1 was trimmed using TrimGalore in paired end mode. The 5’ end of read 1 was clipped by 20 bases to remove potential bias from enzymatic digestion. The 5’ end of read 2 was clipped by 9 bases to remove potential bias from the linear amplification 9-mer. The 3’ end of read 1 was also hard clipped 9 bases after detection of the Illumina adapter was performed. The 3’ end of

read 2 was hard clipped 34 bases after detection of the PCR amplification sequence CCACATCACCCAAACC, removing any potential bias from the enzymatic digestion as well as removing bases corresponding to barcodes and UMIs. Each read was mapped separately to MM10 using Bismark. Using Bismark, both resulting sam files were deduplicated further using UMI, cell barcode and mapping location. The bismark_methylation_extractor tool was then used to extract detected cytosines. After which custom Perl code was used to connect detected cytosines to their respective cell. CpG detection files from the same DNA strand for read 1 and read 2 were then combined and any duplicate detections where the same cytosine was read in both read 1 and read 2 were removed using the UMI. Custom codes for analyzing Dyad-seq data and the accompanying documentation is provided with this work (Supplementary Software). To threshold between successful and unsuccessfully sequenced cells, a minimum detection threshold of at least 25,000 non-dyad covered CpG sites was used. Cells that passed the threshold of detection for 5mC or the transcriptome were considered in the analysis of that feature. For downstream anlysis, genomic regions with extremely low or high coverage were excluded as they likely result from mapping artifacts. Afterwards, for clustering binned dyad and non-dyad methylation levels were subjected to hierarchical clustering and the optimal number of clusters was assigned using silhouette scores.

11. scDyad&T-seq gene expression analysis

Read 2 was trimmed using the default settings of TrimGalore. After trimming, STARsolo (STAR aligner version 2.7.8a) was used to map the reads to MM10 using the gene annotation file from Ensembl. The reads were again mapped to MM10 using the transposable elements annotation file described in TEtranscripts[243].

Transcripts with the same UMI were deduplicated, transcripts detected from genes or transposable elements were analyzed together and any that were not detected in at least one cell were removed from any downstream analysis.

The standard analysis pipeline in Seurat (version 3.1.5) was used for single-cell RNA expression normalization and analysis[235]. Cells containing more than 500 genes and more than 2,000 unique transcripts, were used for downstream analysis. The default NormalizeData function was used to log normalize the data. The top 2,000 most variable genes were used for making principal components and the elbow method was used to determine the optimal number of principle components to use in clustering. UMAP based clustering was performed by running the following functions, FindNeighbors, FindClusters, and RunUMAP. To identify DEGs, the FindAllMarkers or FindMarkers function was used. The Wilcoxon rank sum test was used to classify a gene as differentially expressed, requiring a natural log fold change of at least 0.1 and an adjusted p-value of less than 0.05.

E. Chapter 6 Methods

1. Mammalian cell culture

H9 human embryonic stem cells were grown as described previously in scMAT-seq (chapter 3).

2. scMATH-seq

scMATH-seq processing is like that described in scMTH-seq with minor differences. No spike-in molecules were added, and the reverse transcription step is replaced by a simultaneous reverse transcription and GC tagging step. To do this step, 150 nL of reverse transcription GC tagging mix (0.7 U RNAseOUT (Invitrogen, 10777-019), 1.17x first strand buffer, 11.67 mM DTT, 3.5 U Superscript II

(Invitrogen, 18064-071), 0.19 mM SAM, 1.17x GC reaction buffer, and 0.1 U M.CviPI (NEB, M0227S)) was added to each well and the plate was incubated at 37°C for 1 hour, 4°C for 5 min, 65°C for 10 min, and 70°C for 10 min. After this point, second strand synthesis and all further library preparation steps are as described in scMTH-seq. DNA sequencing for mRNA enriched and non mRNA enriched samples were performed as in scMAT-seq (chapter 3).

3. scMATH-seq analysis pipeline

All sequencing reads were processed as described in scMTH-seq (chapter 4) with the minor change that 5mC based reads were further processed to investigate the presence of GpC methylation as described in scMAT-seq. Each feature, 5mC, 5hmC and gene expression were separately analyzed for data quality. If a cell contained at least 10,000 5mCpG, 10,000 Gp5mC, 300 5hmC, and 1,000 transcripts was considered successfully amplified in all features.

Supplementary Tables

1. Chapter 2

Supplementary Table 2.1 | Table listing 8 bp cell-specific barcodes used in scMspJI-seq

Cell barcode	Sequence
1	GCGAGATT
2	CATTCCAC
3	CCGATGAT
4	AGCTTAGC
5	CGTTACTG
6	TTCGCTTG
7	CTACTGCT
8	AAGAGAGC
9	AACTGTGG
10	CACATCAG
11	AGTCAGTC
12	CGTTGTCA
13	TAGGAACG
14	CCTGATCT
15	AACGAGCA
16	GCAGTAAC
17	TTCTCGAC
18	GTCCAATC
19	AACTCACC
20	CTGCGAAT
21	ACGTTACC
22	AGTTGCAC
23	AATAGCCG
24	ACCTCTAC
25	TTCGACGT
26	TTGATCCG
27	GTACAGGT
28	ACCACCTT
29	GGATTCGA
30	CCGTTAAG

31	GTTCGGAA
32	CCACCATT
33	CGATCGAT
34	GTCTGTAC
35	ACGCCTTA
36	CAGGATTC
37	GGAAGATC
38	TCAGACGA
39	TGCGCTAA
40	GAGAATGC
41	TTAGCGTG
42	TGAAGGCT
43	CTTAGCAG
44	AAGCTACC
45	ACATCTGC
46	CGCATTAC
47	CCTAGATC
48	CATCCAGA
49	GGTCTTGA
50	GACAGATG
51	GAACAGCT
52	ACGAGCAA
53	TCCTTCTC
54	GCGTGTAA
55	CAGCCATA
56	GAATTGCC
57	AATCAGCC
58	CTCAACAC
59	GCAGATAC
60	TCGCTTGT
61	AGTCTTCG
62	TAGAGGCA
63	CCTTGGTT
64	AGAACGCA
65	GTATACGC
66	ACTGCTAG
67	ATCGGTGA
68	GACCATGA
69	TCCAAGGT

70	GCCAACAT
71	GCGTCAAT
72	AGCCAAGT
73	ACGTCAGA
74	TCACCTGA
75	GCAATCCT
76	AATTCGCC
77	TGAAGCTC
78	GTCCGATA
79	CCTGTAGT
80	CAGACTGT
81	TGTAGCCT
82	GATGCCAT
83	AACGGCAT
84	GATAGCAC
85	TACGGTTC
86	TGGTTGGA
87	TCGTGTAC
88	TAGCGGAA
89	CTAGGCTA
90	GCTGTGTA
91	CAGGTCTT
92	AAGAGCCA
93	GCATGACT
94	TTACGGTC
95	ACGCATAC
96	GATGCAAC

2. Chapter 5

Supplementary Table 5.1 | Double stranded adapter barcodes used in scDyad&T-seq.

Barcodes described 5’ to 3’ for the bottom adapter. Top adapter barcode sequences are the reverse compliment of those listed. Barcodes 1, 2 and 3 are also used in M-H-Dyad-seq.

Cell barcode	Cell barcode
1	AGAGATGGAA
2	TTGGATGGTA
3	TGTTTGTAGG

4	TGAAGAGAAG
5	AGTGTGAAGT
6	AAGTTGATGG
7	GAATGGTGAT
8	AGGTTGAGTT
9	GATTAGGTGA
10	GAAGATTGGT
11	AGAGGAAGAA
12	TGGTGTTATG
13	GAGAAGAAGA
14	GATATGGAAG
15	GTTAGGTAAG
16	GTTAGTAGAG
17	TGGTGTATGA
18	AGGAAAGTAG
19	GAGATAAAGG
20	TTGAAGGAGT
21	AGTGTGAGAA
22	GTAGATAGAG
23	AAGTGTTGAG
24	GTGAGTAGTT
25	AGAGGTTAGA
26	TGGTTGAAGA
27	ATGATGAAGG
28	AGTTGAGGAA
29	GAAAGTGATG
30	AAGGTTGGAA
31	GAATGGTATG
32	GAAGAGAAAG
33	AGGTTGAAAG
34	TAGGATGGAT
35	TAGGTGAAGT
36	ATGAGTGGAA
37	TAGATGTAGG
38	GATATGAGGT
39	GAAGAAAGTG
40	ATGAGAGTGA
41	AGGAGTATAG
42	GTGATAGATG

43	TGAGGTAGTA
44	GTGTGTAGAA
45	GAGAGTTGTA
46	GAGAAAGGTA
47	TTGATGGAAG
48	TGTATGGATG
49	AAGTAGGAAG
50	TGTGATGAAG
51	GATTGTGGTA
52	GATAGATAGG
53	GTATGGAAAG
54	GAGAAAGAAG
55	AGTGAAAGGA
56	TGATTGTGTG
57	TGAGATATGG
58	TAGTTTGAGG
59	AAGGTAGAAG
60	AGAGAGAAGA
61	GTTGGAAGAA
62	GAAGGATGTA
63	GTAAAGGAAG
64	ATGGAGAGTA
65	AGGAAGTTGT
66	TGATGGAGAT
67	TGAATGGTAG
68	GTTTGAAGGT
69	AGTTTGTGGT
70	AGATGTGAAG
71	AGATATGTGG
72	GAGAATAGTG
73	TTGGAAGAAG
74	TGGTTGTTAG
75	TGTTGAGATG
76	TAGGTTGTAG
77	GTGGTAAAGT
78	GAAGGAAAGA
79	GATAATGAGG
80	ATGTTGGTAG
81	GAGTGATAAG

82	AAGGTGAATG
83	TAGGAGTAAG
84	GTGAGATGAT
85	TGAGGTTAAG
86	TGAGAAAGGT
87	AGAGTTGATG
88	TTGGAAAGTG
89	GTTAGGAAGA
90	AGGATTGAAG
91	TAGAAGAGGT
92	GTGAAAGAGA
93	GTATGAGGAA
94	TGAAAGAGTG
95	GTTGTGAAAG
96	GAAAGAAGGT

Supplementary Table 5.2 | Double stranded adapter barcodes used in M-M-Dyad-seq.

Barcodes described 5’ to 3’ for the bottom adapter. Top adapter barcode sequences are the reverse compliment of those listed.

Barcode #	Cell barcode
1	ATATGGAG
2	AGGGATTG
86	TGGTTGGA

Supplementary Table 5.3 | Double stranded adapter barcodes used in H-M-Dyad-seq and H-H-Dyad-seq.

Barcodes described 5’ to 3’ for the bottom adapter. Top adapter barcode sequences are the reverse compliment of those listed.

Barcode #	Cell barcode
1	GAGAATGTGT
2	AGGTAAGATG
3	TGTAAGTGAG