UNIVERSITY OF CALIFORNIA

Santa Barbara

Simultaneous Measurement of DNA Methylation and Genome-Nuclear Lamina Interactions in Single Cells

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Chemical Engineering

by

David Jack Podorefsky

Committee in charge:

Professor Siddharth Dey, Chair

Professor Arnab Mukherjee

Professor Michelle O'Malley

Professor Adele Doyle

December 2023

The dissertation of David Jack Podorefsky is approved.

Arnab Mukherjee

Michelle O’Malley

Adele Doyle

Siddharth Dey, Committee Chair

December 2023

VITA OF DAVID PODOREFSKY
December 2023

EDUCATION

PhD in Chemical Engineering, Bioengineering Emphasis
University of California, Santa Barbara — Fall 2018 - Present
Santa Barbara, CA — *Expected degree conferral:* December 2023

Bachelor of Science in Chemical Engineering
University of Massachusetts Amherst — Fall 2014 - Spring 2018
Amherst, MA

Summa Cum Laude
GPA: 3.92

SKILLS

Molecular Biology: Cell culture, PCR, gel electrophoresis, western blotting
Genomics: Next-generation sequencing library preparation, genome mapping (Bash, Perl, MATLAB)
Computational Biology: PyMOL, image processing (ImageJ, Imaris), Simulink
Instruments: Fluorescence activated cell sorting, fluorescence microscopy, high-performance liquid chromatography
Engineering software: Aspen Plus, LabView, COMSOL, Mathematica
Other: Microsoft Office, HTML, CSS, Illustrator, Premiere, After Effects

RESEARCH

University of California, Santa Barbara — Santa Barbara, CA
Graduate Researcher; Advisor: Siddharth S. Dey — January 2019 - Present
Simultaneous measurement of DNA methylation and genome-nuclear lamina interactions in single cells

- Profiled epigenome of human myeloid leukemia cells to study the link between the 5-methylcytosine DNA modification and the 3-dimensional organization of genetic information within the same cell
- Sorted single cells with FACS, prepared DNA libraries with a liquid handling robot for Illumina sequencing, aligned reads to human genome employing Perl, and measured epigenetic correlations with MATLAB

University of Massachusetts Amherst — Amherst, MA
Undergraduate Researcher; Advisor: Shelly R. Peyton — Summer 2016 - May 2018
Mesenchymal Stem Cell Traction Forces in 3D, Degradable PEG Hydrogels

- Measured human stem cell traction on the extracellular matrix *in vitro* with peptide functionalized poly(ethylene glycol) based tissue scaffolds, tuning stiffness, adhesivity, and degradability

- Calculated displacement and cell traction utilizing confocal microscopy, iterative digital volume correlation and stress computation in MATLAB, and reconstructed results with 3D renderings

PUBLICATIONS

David J. Podorefsky, Alex J. Chialastri, Siddharth S. Dey. The simultaneous measurement of 5-methylcytosine and genome-nuclear lamina contacts in single leukemia cells (in preparation)

Alex Chialastri, Eyal Karzbrun, Aimal H. Khankhel, Neha Saxena, David J. Podorefsky, Monte J. Radeke, Sebastian J. Streichan, Siddharth S. Dey. Integrated single-cell sequencing reveals principles of epigenetic regulation of human gastrulation and germ cell development in a 3D organoid model (in revision, bioRxiv doi: 10.1101/2022.02.10.479957, Nature Genetics)

David J. Podorefsky, Lauren E. Jansen, Jacob A. Ong, Carey E. Dougan, Christian Franck, and Shelly R. Peyton. Mesenchymal Stem Cell Traction Forces in 3D, Degradable PEG Hydrogels (in preparation)

TALKS

22nd UC Symposium on Bioengineering & Biotechnology Industry Showcase, Poster, August 9, 2022

5th International Conference on Epigenetics and Bioengineering (EpiBio), Talk, November 5, 2021

12th Amgen-Clorox Graduate Student Symposium, Poster, October 4, 2019

1st Year ChE Graduate Student Research Symposium, Talk, September 24, 2019

24th & 23rd Massachusetts Statewide Undergraduate Research Conferences, Poster, Spring 2018 & 2017

Massachusetts State Science & Engineering conference at MIT, Poster, Spring 2013

TEACHING

Separation Processes Teaching Assistant
Department of Chemical Engineering, UC Santa Barbara Fall 2020

- Held weekly office hours, graded weekly homework assignments, and put together answer keys

Chemical Reaction Engineering Teaching Assistant
Department of Chemical Engineering, UC Santa Barbara Spring 2022

- Held weekly office hours, graded homework and wrote Aspen questions, and gave review session for midterm

MENTORING

Mentor for Undergraduate Student (An)
Department of Chemical Engineering, UC Santa Barbara Spring 2022

- Shadowed DNA library preparation and performed PCR and solid-phase reversible immobilization purification

Mentor for Undergraduate Student (Sean)
Department of Chemical Engineering, UMass Amherst Summer - Fall 2020

- Remotely taught undergraduate student 3D traction force microscopy image processing pipeline

Mentor for Undergraduate Student (Jacob)
Department of Chemical Engineering, UMass Amherst Fall - Spring 2018

- Taught 3D PEG traction hydrogel synthesis, high-performance liquid chromatography, & peptide resin cleavage

Engineering the Cell - Mentor for High School Students (Amma and Princess)
Department of Chemical Engineering, UMass Amherst Summer 2017

- Taught high-performance liquid chromatography for analysis and purification of in-house made oligopeptides

SERVICE

Organizing Committee:
University of California, Santa Barbara Spring - Summer 2022
Bioengineering Research Symposium and Biotechnology Industry Showcase

- Attended weekly organizing meetings, liaison for speakers, organized industry panel event, pickup van logistics

Organizing Committee:
University of California, Santa Barbara Spring - Fall 2022
BioE Tech Talks (Podcasts)

- Attended weekly organizing meetings with BMES, arranged presentation and Q&A session with Eppendorf

Outreach Volunteer:
It’s A Material World at Hollister Elementary School Spring 2019
Materials Research Outreach Program at El Camino Elementary School Winter 2019

- Taught kids and parents basic Science with the intention of inspiring first-generation college students in STEM

BioE Holiday Party & Lab Tours for High School Students:

University of California, Santa Barbara December 2019

- BMES organized holiday party and lab tour for high school students, gave single-cell library pooling demo

AWARDS & HONORS

5th International Conference on Epigenetics and Bioengineering NSF Grant, October 2021

National Science Foundation *Graduate Research Fellowship Program* Honorable Mention, Spring 2018

Dean's List, University of Massachusetts Amherst, Fall 2014 - Spring 2018

John and Abigail Adams Scholarship, Fall 2018

1st place at Massachusetts State Science & Engineering conference at MIT, Spring 2013

AFFILIATIONS

Society for Biological Engineering	Summer 2021 - Present
UCSB Chapter of Biomedical Engineering Society (BMES), VP	Fall 2019 - Present
UCSB & UMass Chapters of American Institute of Chemical Engineers	Fall 2014 - Present
International Society for Pharmaceutical Engineering (ISPE)	Fall 2014 - Present

REFERENCES

Prof. Siddharth Dey (thesis supervisor)

Prof. Shelly Peyton (undergraduate supervisor)

ABSTRACT

Simultaneous Measurement of DNA Methylation and Genome-Nuclear Lamina Interactions in Single Cells

by

David Jack Podorefsky

DNA methylation (5-methylcytosine or 5mC) and the 3-dimensional organization of the genome within the cell nucleus are two critical epigenetic features that regulate gene expression and cellular behavior, and aberrant patterns of these epigenetic features are associated with cancer and disease[1,2]. In cancer, widespread regions of DNA methylation loss, termed hypomethylation, have previously been found positioned at the periphery of the nucleus, the nuclear lamina (NL), suggesting DNA methylation and genome organization could be linked epigenetic states[1]. However, the direct genome-wide relationship between 5mC and genome-NL interactions remains obscured in bulk measurements.

To overcome this limitation, we developed a new single-cell sequencing technology (sc5mC+DamID-seq) to simultaneously measure DNA methylation and genome-NL interactions from the same cell. sc5mC+DamID-seq uses mark-specific barcoded adapters and cytosine deamination to identify both epigenetic features and individual cells, enabling us to profile thousands of single cells per day.

By applying sc5mC+DamID-seq to chronic myelogenous leukemia cells (KBM7), we observed hypomethylation at regions contacting the NL, known as lamina associated domains (LADs). Interestingly, we also discovered that genomic regions that contact the NL more frequently in single cells display the greatest loss and variability in 5mC. Further, while LADs appear as continuous stretches of contact with the NL in bulk sequencing, LADs at the single-cell level frequently display segments of noncontact with the NL, which we found influences the mean levels of 5mC in LADs. Finally, to test the connection between these two epigenetic features, we globally demethylated the epigenome, which relocated more variable genome-NL contact regions towards the nuclear interior. Thus, simultaneous single-cell measurements in sc5mC+DamID-seq has enabled us to systematically uncover the relationship between DNA methylation and genome organization in cancer cells. As an addition to the protocol, the transcriptome, the ensemble of mRNA produced by a cell, was measured with 5mC and genome-NL contacts, to reveal how epigenetic features directly correlate with gene expression. The mRNA modification m^6A was also compared to the epigenetic features to find the relationship between the epigenome and epitranscriptome.

TABLE OF CONTENTS

I. Introduction

A. Motivation and Specific Aims

The epigenome is an ensemble of chemical modifications to DNA and proteins that turn genes on and off without altering the sequence itself. This switch allows cells to display different gene expression while retaining the same genetic material and plays a critical role in influencing cellular function. In this project, I aim to better understand how the epigenome regulates gene expression in single cells by simultaneously measuring two epigenetic features, DNA modification by 5-methylcytosine and genome-nuclear lamina contacts (a facet of genome organization), within the same cell.

The overall gap in knowledge is how does 5mC and genome-NL contacts correlate within the same cell, and how do they combine to affect gene expression. Although previous bulk studies in cancer cells have demonstrated that regions of depleted DNA methylation with foci of increased DNA methylation, termed partially methylated domains, are situated at the nuclear lamina, seen in **Fig. 1**, these epigenetic features have not been measured simultaneously within the same cell[1]. Single-cell (sc) sequencing can detect heterogeneity

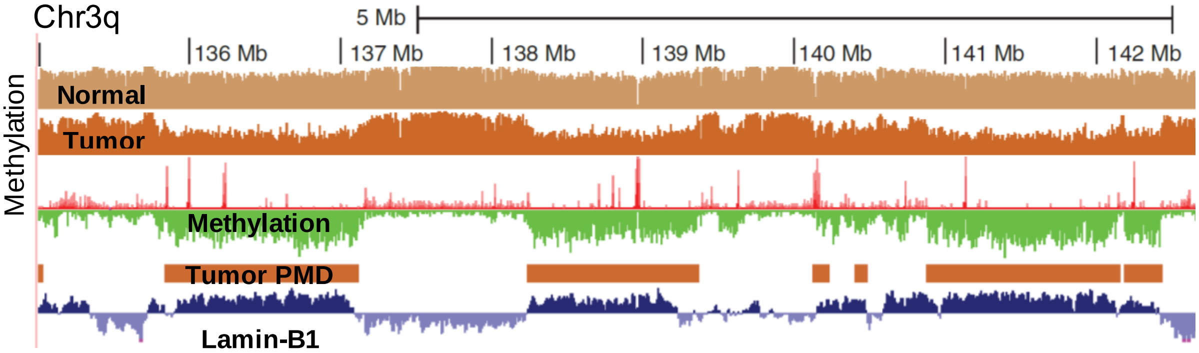


Fig. 1 | DNA methylation and genome-nuclear lamina contacts within bulk populations of normal and cancer cells. Regions of methylation loss are marked as partially methylated domains. Location of NL coincides with Lamin-B1 protein. Figure adapted from B.P. Berman *et al. Nature Genetics* (2012)

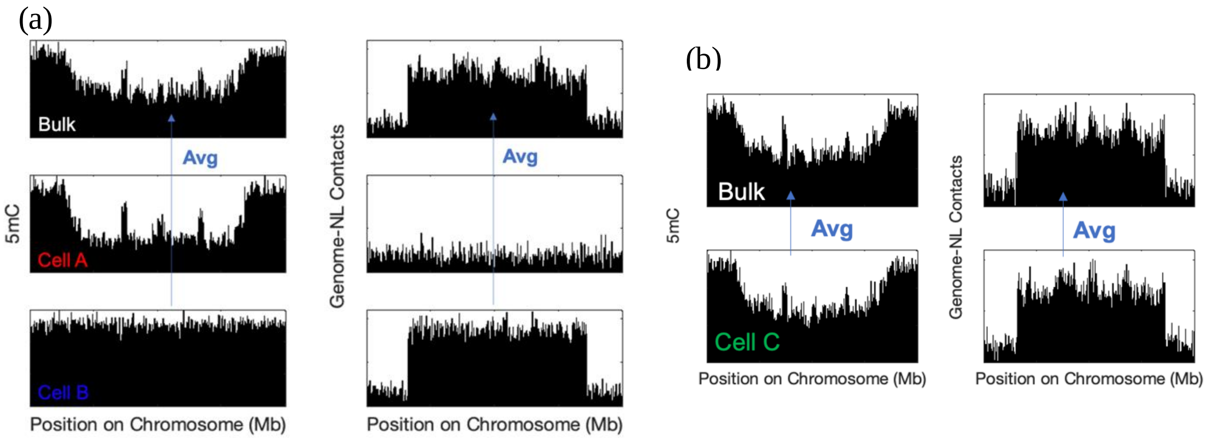


Fig. 2 | Simulated 5mC and genome-NL contact profiles in bulk and in single cells (**a**) Population is comprised of Cell A and Cell B profiles. (**b**) Population is comprised of only Cell C profiles.

within a cell population, which is hidden by bulk measurements. As seen in **Fig. 2**, the relation between DNA methylation and genome-NL contacts within the same cell may not hold at the bulk level due to measuring an ensemble of states, creating two possibilities. The first possibility is the cells in the population may look like Cell A and Cell B **(Fig. 2a)**. Cell A may display a region with a loss of DNA methylation, termed hypomethylation, corresponding to no NL contacts, whereas, in the same region, Cell B may display no changes in DNA methylation and an NL contact. However, when averaged in bulk, there is hypomethylation and a corresponding NL-contact, which is not indicative of what is occurring at the single-cell level. The second possibility is the cells in the bulk population may all look like Cell C, in which 5mC and genome-NL contacts are actually correlated **(Fig. 2b)**. Therefore, bulk measurements of these epigenetic features are not sufficient because it is unknown if they are truly linked within the same cell, or if they only appear to be as a consequence of the measuring an ensemble. Simultaneous measurement of epigenetic features must be made within the same cell for complete characterization of their correlations.

I hypothesize that 5mC and genome-NL interactions are linked within single cells because the activation of gene expression is regulated by changes in 5mC on gene promoters and coincides with the unpacking of condensed, repressed genes situated at the NL that reposition to the nuclear interior for transcription[3]. Therefore, these epigenetic features may be functionally related, in that altering the 5mC of a gene positioned at the NL may reposition the lamina associated domain (LAD). Vice versa, repositioning a LAD may change the 5mC of the LAD. Since high levels of DNA methylation in promoter regions and genome-NL contacts both correlate with low gene expression, discovering how features of the epigenome are related and how they combine to regulate gene expression could improve our understanding of their dysregulation in disease, such as cancer.

1. Aim to Simultaneously Measure 5mC and Genome-NL Contacts in Single Cells:

Although 5mC and genome-NL contacts have both been measured independently in single cells and correlated across bulk studies, they have never been measured simultaneously in single cells. We hypothesize by using a specially engineered cell line and next generation sequencing technology, 5mC and genome-NL contacts can be measured simultaneously in the same cell. Unlike 5mC, which is endogenous to DNA, detection of NL contacts requires the addition of a fusion protein into each cell that indicates when a genomic region makes the contact, however its integration in a manner suitable for single-cell sequencing has remained a challenge. Although transiently adding the construct has been achieved on the order of days in previous bulk DamID studies, it may show incorporation variability between each cell, which would foil the accuracy of our single-cell studies[4]. The alternative is creating stable cell lines with this construct, which is very labor intensive and takes on the order of months[5]. Even with the fusion protein added, NL contact marks are not maintained when the DNA replicates, requiring the addition of the fluorescence ubiquitination cell cycle indicator (FUCCI system) into the cell line to select for cells in the optimal phase[6,7]. Due to these reasons, single-cell DamID studies have been limited. In the Dey lab, we have a stable cell line (human myeloid leukemia (KBM7), courtesy of Jop Kind) expressing the fusion protein for detection of NL contacts, with the FUCCI system, overcoming the limitation of having to make a new cell line. The second challenge in developing this technology is performing reactions on individual cells after they are sorted into wells of a plate with fluorescence-activated cell sorting (FACS). The reagent volumes required for each cell in a well is lower than what can be accurately measured with a pipette and is challenging to process the vast number of cells manually, however the Dey lab has a liquid handling robot to quickly dispense nanoliters of reagents into wells. To measure the epigenetic marks, we will use restriction enzymes known to recognize 5mC and genome-NL contact sites, to which we will ligate adapters for sequencing. Potential outcomes include either the ability to measure both epigenetic features in single cells, only one of the epigenetic features, or both can be detected but the signal of one is lower than the threshold required for analysis. Our strategy is to first measure 5mC and genome-NL contacts in bulk, since it is less technically challenging than in single cells, and then validate they can both be measured independently in single cells. After this validation, we will measure them simultaneously in single cells. With a specially engineered cell line for measuring genome-NL contacts and a robot for dispensing small quantities of reagents to prepare DNA libraries, it is possible for us to measure 5mC and genome-NL contacts simultaneously in single cells.

2. Aim to Quantify the relationship between DNA methylation and genome organization in individual cells:

I hypothesize that 5mC and genome-nuclear lamina contacts are linked features in cells, and partially methylated domains contact the NL. Although bulk studies have shown an anticorrelation between DNA methylation and genome-NL contacts in cancer cells, it is unknown if this relationship holds at the single-cell level. I aim to first elucidate whether DNA methylation and genome organization are anticorrelated at the single-cell level by measuring and comparing 5mC and genome-NL contacts profiles in KBM7, implementing sc5mC+DamID sequencing from Aim 1. Single-cell sequencing of genome-NL contacts has additionally demonstrated that certain LADs display NL contact heterogeneities within the same cell line, referred to as contact frequency[8]. If 5mC and genome-NL contacts are indeed linked in single cells, I will measure the contact frequency of individual LADs by observing what percentage of the sequenced cells display the LAD, and then compare it with the level of methylation in the corresponding location on the chromosome. Potential outcomes include either more variability in LADs produces changes in 5mC levels, or LAD variability does not affect 5mC levels. Through this I can establish if there is a correlation between the extent of hypomethylation in a partially methylated domain and the frequency in which it contacts the NL.

3. Aim to Assess causality between DNA methylation and genome organization:

It is currently known that 5mC and genome-NL contacts are linked features in bulk, and upon completion of Aim 2, it will be known if this relationship holds in single cells. If these epigenetic features are connected in single cells, I ask if they are functionally linked to each other. I hypothesize that they display causation, in that changing global DNA methylation will reposition LADs, and conversely, disrupting the attachment of LADs to the NL will alter the 5mC of the genomic region. I will prove if there is causality by first drugging the cells with a DNA methylation inhibitor to reduce global 5mC, and then performing 5mC+DamID sequencing to compare the LAD (and 5mC) profiles in the drugged and control cells to evaluate if LADs repositioned themselves once 5mC is lost. Potential outcomes include changing global 5mC repositions the majority of LADs to the nuclear interior, suggesting the epigenetic features may be functionally linked, or changing global 5mC does not reposition the LADs, suggesting they may only be correlated and rely on an additional component. To further prove causality, I could conversely reposition LADs to the nuclear interior using dCas9 attached to an enzyme that changes DNA methylation, perturbing the boundary elements of LADs that are involved in maintaining chromatin structure. Potential outcomes are repositioning the LAD changes the methylation of the region, suggesting causality, repositioning the LAD does not change methylation, or changing the methylation of only the boundary elements does not reposition LADs. A backup strategy is using siRNA to knock down attachments of genomic regions to the NL and then observing if there are changes in 5mC. If both changing the 5mC repositions the LADs, and repositioning the LADs changes 5mC, it would support these two epigenetic features display a causal relationship. Causality of these two epigenetic features has yet to be explored due to the lack of simultaneous measurements within the same cell, and thus the method we propose will provide insight into the epigenome as a multifaceted regulator of gene expression.

4. Aim to Combine Transcriptome Measurements with 5mC and genome-NL contacts:

After 5mC and genome-NL contacts have been successfully measured within the same cell, I will further expand the technique to measure the transcriptome, which is the ensemble of all mRNA expressed by a cell. Unlike DNA methylation and spatial organization of the genome, which are seen on the DNA, this measurement pertains to the RNA, making it much easier to measure simultaneously with the other features. This can be achieved by utilizing the polyA tail on mRNA as a handle for attaching a complementary sequencing adapter, prior to ligating adapters to the DNA to measure the 5mC and genome-NL contacts.

B. Background

1. 5mC Modification and Demethylation

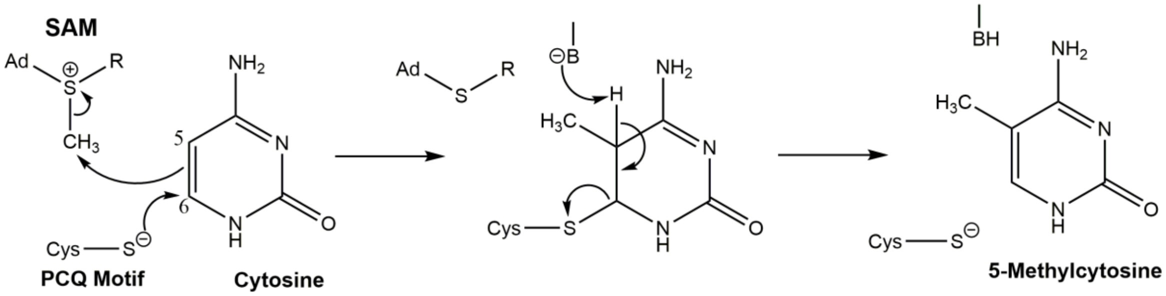


Fig. 3 | Catalytic mechanism of DNA methyltransferase. The PCQ motif of DNMT makes a nucleophilic attack on carbon 6 of cytosine to push electrons from the cytosine ring, forming a covalent bond between carbon 5 and a methyl group originating from SAM. A base provided by the enzyme deprotonates carbon 5 to form 5-methylcytosine. Reaction scheme adapted from F. Lyko, *Nature Reviews* (2018)

The most common DNA modification is 5-methylcytosine (5mC), with 60-80% of mammalian CpG sites containing this modification[9]. High levels of this mark are associated with gene repression when on gene promoters, transcriptional regulation, and stably maintaining genes in a silent state, such as in genomic imprinting and restraining transposable elements[10]. DNA methylation in the gene body is correlated with the expression of transcripts[11].

DNA methyltransferases (DNMTs) both add and maintain 5mC on newly synthesized DNA strands[12]. 5mC is added *de novo* by DNA methyltransferase 3A (DNMT3A) and DNMT3B by a catalytic mechanism involving the transfer of the methyl group from S-adenosylmethionine (SAM) to carbon 5 of the cytosine, to form 5-methylcytosine **(Fig. 3)**. Following DNA replication, the maintenance DNA methyltransferase, DNMT1, maintains 5mC on newly synthesized daughter strands by recognizing hemi-methylated CpG sites through its partner UHRF1, and performing the same reaction mechanism through the conserved PCQ motif[13]. Overexpression of DNMTs produces increased levels of 5mC and disruption of the enzyme has been shown to decrease levels of 5mC[14].

Removal of the methyl group from cytosine occurs by two mechanisms: passive demethylation and active demethylation[13]. Passive demethylation occurs by dilution of the mark on newly synthesized strands during replication due to a lack of maintenance by DNMT1. Active demethylation consists of a series of oxidations that converts 5-methylcytosine to 5-hydroxmethylcytosine (5hmC), 5-formylcytosine (5fC), and then 5-carboxylcytosine (5caC). The latter two DNA modifications are eventually excised by thymine DNA glycosylase (TDG), followed by base excision repair (BER) to restore the cytosine **(Fig. 4)**[13]. Knocking out TDG resulted in a 2-fold increase of 5fC without any significant change to the levels of 5hmC, suggesting 5mC must be oxidized at least twice in active demethylation for its conversion back to cytosine[15]. Ten-eleven

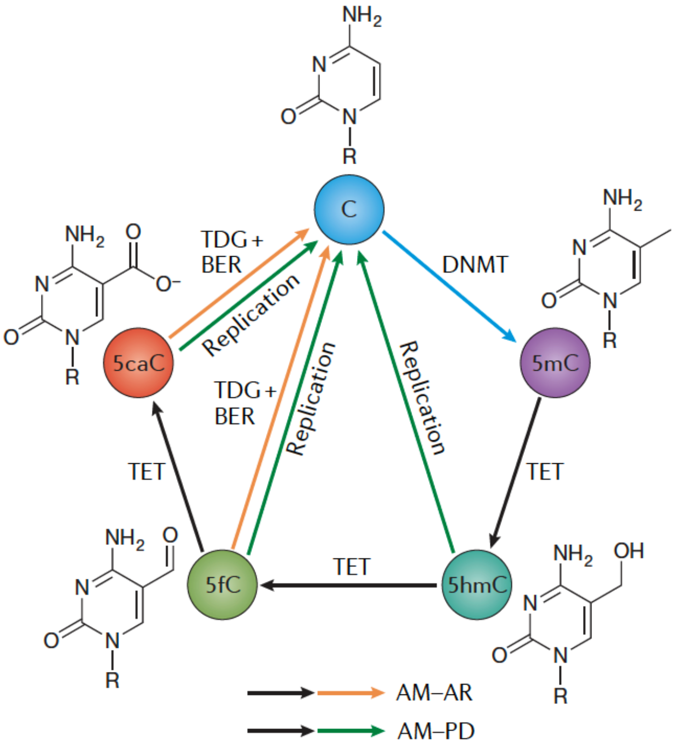


Fig. 4 | DNA methylation cycle. A methyl group is added to cytosine by DNMT and is converted by TET to oxidized states. Demethylation occurs passively, shown by the green arrows in active modification-passive dilution (AM-PD), and it occurs actively shown by the orange arrows in active modification-active removal (AM-AR). Figure adapted from X. Wu *et al. Nature Reviews* (2017) translocation (TET) family dioxygenases are responsible for the oxidation of 5mC, 5hmC, and 5fC, and TET2 has a 4.9-7.6-fold faster initial reaction rate for its substrate of preference, 5mC[16]. In order to perform its function in active demethylation, TET requires Fe(II) as a cofactor, and oxygen and α-ketoglutarate (α-KG) as substrates, the latter of which is derived from isocitrate by isocitrate dehydrogenase (IDH) enzymes[17]. IDH mutants are commonly present in many cancers and generate (R)-2-hydroxyglutarate ((R)-2-HG), an inhibitory substrate analog for α-KG-dependent proteins, hindering TET activity, and the mutations at Arg100 and Arg132 in IDH1 and Arg140 and Arg172 in IDH2 increase binding affinity for NADPH which reduces α-KG to (R)-2HG[17–19]. This mutation in cancer ultimately interferes with conversion of 5mC to 5hmC, affecting the DNA methylation cycle.

2. The Nuclear Lamina, Genome Organization, and Laminopathies

The nuclear lamina is a fibrous protein layer that lines the nucleoplasmic side of the inner nuclear membrane and extends into the nucleoplasm[20]. It is primarily composed of lamins, which are type V intermediate filaments, and lamina associated proteins and play roles in chromatin modification, transcriptional repression, and maintaining structure[20]. The large surface area and architecture of the nuclear lamina harbors and constrains peripheral chromatin by acting as an anchoring platform for physical attachment[21]. Regions of DNA that contact the nuclear lamina are called lamina associated domains (LADs) and are characterized in normal cells by low levels of gene expression **(Fig. 5)**[22]. Human and mouse cells have 1000-1500 LADs distributed across all chromosomes, with a median size of 0.5 Mb, and comprise over one third of the genome, making it a monumental epigenetic feature[23]. Since LADs are condensed (heterochromatic) regions of chromatin, gene expression within these domains is likely controlled by the repressive environment,

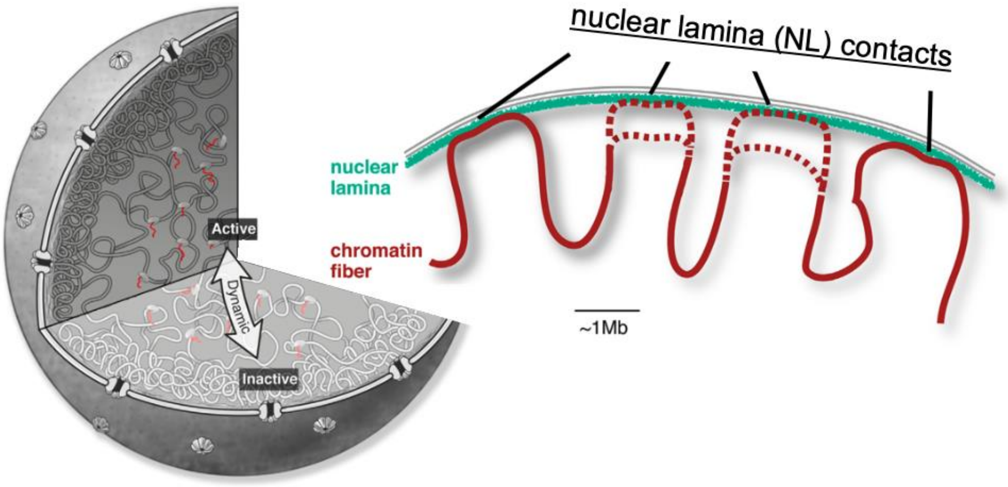


Fig. 5 | 3D organization of the genome. Inactive genes are positioned close to the periphery of the nucleus (the nuclear lamina) in densely packed heterochromatin and display low levels of gene expression. Active genes are located towards the nuclear interior in expressive euchromatin. Gene expression is a function of DNA accessibility in the nucleus governed by genome organization. Figure adapted from Marta Melé *et al. Molecular Cell* (2016) and Jop Kind *et al. Cell* (2015).

suggesting the role of higher-order organization of chromatin for long range gene silencing[24]. Indeed, the spatial organization of chromatin is implicated in gene expression, since euchromatin (uncondensed state) is gene rich, transcriptionally active, and located in the nuclear interior, whereas heterochromatin is gene poor, transcriptionally inhibited, and located at the nuclear periphery[3]. There are three main models that elucidate how nuclear organization and gene expression are interconnected and why nuclear reorganization may occur. First, activation of genes may require repositioning of DNA to transcriptionally active regions, termed transcription factories, which has been observed through the colocalization of transcriptionally active genes at these regions[25]. Second, transcription of long non-coding RNA (lncRNA) can reorganize chromatin by the RNA product marking a location for protein binding to occur, which pulls the DNA into a new confirmation for more permissive expression, termed the Cat’s Cradle Model[3]. Third, transcriptional interference contributes to chromatin remodeling since the act of transcription at one site can produce a negative affect at another site, exemplified by transcription of *Bxd* ncRNAs repressing the expression of *Ubx* in *cis*[26].

A previous study disrupted the heterochromatic state of LADs by targeting them with the acidic-activating domain (AAD) of viral protein VP16, resulting in destabilization in the peripheral positioning of LAD-NL contacts in single cells[27]. Coinciding with this event was epigenetic changes; LADs that remained at the periphery retained H3K9me2, a histone modification enriched in LADs and found at the nuclear periphery, whereas LADs that relocated to the nuclear interior displayed less H3K9me2[27]. Knocking down G9a, a histone lysine methyltransferase (HKMT) that adds dimethylation to H3K9, resulted in less association of LADs with the NL, supporting that epigenetics may be a significant factor in nuclear organization[27]. LADs that were not in contact with the NL displayed higher levels of H3K36me3, a marker of transcriptional elongation, suggesting expression of derepressed genes in LADs is dependent on nuclear organization, which is regulated by the epigenome[27].

Mutations in lamins and their associated proteins in the nuclear lamina are known to cause laminopathic diseases[20]. Lamin A (LMNA) mutations are widespread with over 330 disease causing mutations, making it one of the most mutated genes in the human genome, and is associated with muscle diseases such as Emery Dreifuss Muscular Dystrophy (EDMD), adipocyte diseases such as Dunnigan-type familial partial lipodystrophy, neuronal diseases such as Charcot-Marie-Tooth (CMT) disease, and accelerated aging such as Hutchinson-Gilford progeria syndrome (HGPS)[20]. It is thought the diseases emerge by affecting interactions between the nucleus and cytoskeleton networks, hindering rearrangement of the nuclear lamina, and overall misregulating genes, such as reducing response of stress-dependent gene expression upon application of mechanical stress[20]. In the most frequently occurring laminopathy, EDMD, mutations are present in both LMNA and emerin (a nuclear lamina-associated protein family in the inner nuclear membrane), interfering with heterochromatin formation at the Sox2 locus, resulting in mRNA overexpression that inhibits myogenesis, preventing the formation of muscular tissue[28]. Likewise, lamin mutations are present in many types of cancer, supporting the significance of the nuclear lamina in maintaining proper gene expression. In gastrointestinal cancer and prostate cancer, studies have found reduction or lack of expression of lamins A, B1, and C[29]. In skin cancer, reduction or lack of expression of lamin A in basal cell carcinomas correlated with increased tumor proliferation rate[30]. Thus, abnormalities at the NL influence disease by interfering with the spatial organization of genetic information in concert with epigenetic regulation.

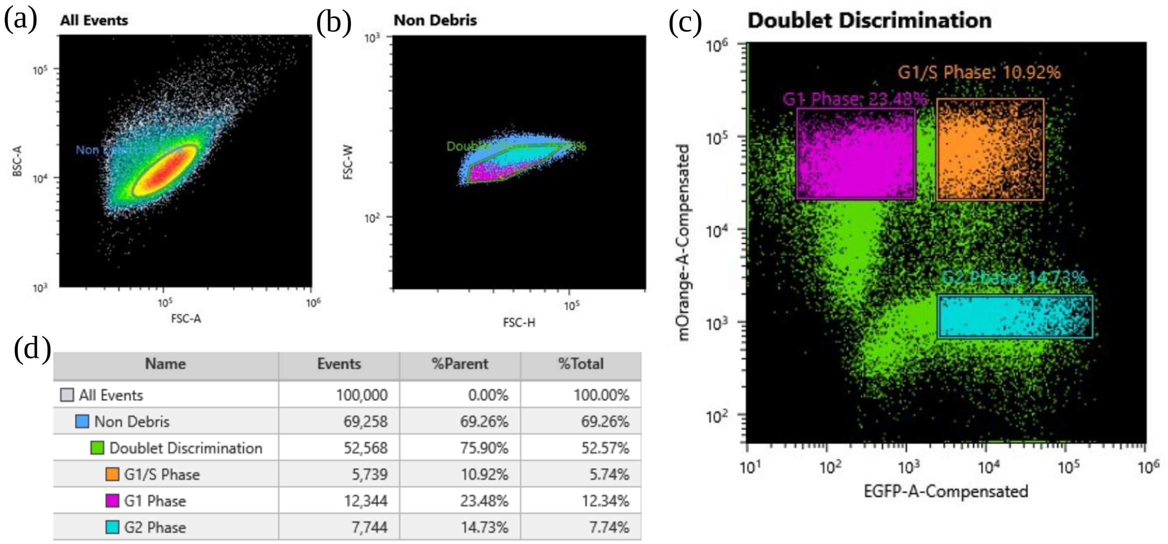
3. Hyper, Hypomethylation, and Partially Methylated Domains in Cancer

The regulatory effects of 5-methylcytosine and nuclear organization combine to dysregulate gene expression in cancer. Hypermethylation, or overly methylated DNA, and hypomethylation, or undermethylation of DNA, together form partially methylated domains (PMDs), and are collectively found in almost every type of cancer[31]. Tumor suppressor genes are inactivated by *de novo* DNA methylation of CpG islands overlapping gene promoters, preventing growth inhibition of tumors[32]. In acute myelogenous leukemia (AML) for example, the tumor suppressor gene $p15^{INA4B}$ is silenced by hypermethylation, helping the cancer to proliferate[33]. Coinciding with the aberrant DNA methylation are elevated levels of DNA methyltransferases DNMT1, DNMT3A, and DNMT3B in cancer tissue compared to normal tissue[14]. Repeated DNA sequences that collectively cover half of the genome, including satellite DNA, a tandemly repeating non-coding DNA that is a main component of heterochromatin, Alu, and long interspersed elements (LINE), are hypomethylated in cancer[31]. Coding DNA is also hypomethylated in cancer, such as the gene promoter for enzyme urokinase (uPA) that catalyzes the activation of plasmin, which is essential for tumor metastasis by degrading extracellular matrix proteins to aid cell proliferation, migration, and angiogenesis[2]. In highly invasive breast cancer, the promoter was demethylated and cells expressed uPA mRNA, whereas in normal and low invasive breast cancer cell lines, uPA was not expressed, and the promoter remained methylated[2]. It is hypothesized that there is a connection between hypermethylation and hypomethylation during tumorigenesis; overactivity of *de novo* DNA methyltransferases that methylate CpG islands overlapping tumor suppressor genes may be counteracted by widespread demethylation, as a form of epigenetic repair[31]. However, removal of the DNA methylation may be inefficient and rather than correcting the foci of hypermethylation, it may result in demethylation of entire regions, producing hypomethylation. Conversely, overexpression of hypomethylated genes may be silenced with genome wide *de novo* DNA methylation, ultimately producing hypermethylation, as previously observed *in vivo* during multistage hepatocarcinogenesis in which preneoplastic nodules progressively displayed hypomethylation and an increase of p53 mRNA during the first 36 weeks of the study, followed by a decrease in p53 mRNA and relative hypermethylation when reaching hepatocellular carcinoma 18 weeks later[34].

If a cell gains a mutation in a gatekeeping pathway, one that maintains checks and balances on cell division and cell death, such as the gatekeeper adenomatous polyposis coli (APC) gene in colon tumors, tumorigenesis is initiated[35,36]. Nuclear lamins are involved in tumor suppression by protecting mutations from occurring in tumor suppressor pathways, and therefore if lamins are mutated, cancer cells can evade apoptosis, resulting in uncontrolled cell proliferation[20]. Indeed, hypermethylation of the lamin A/C CpG island-promoter is frequently found in leukemia and lymphoma resulting in reduction of lamin expression and is associated with decreased survival rates in nodal diffuse large B-cell lymphoma patients[37]. Since tumor partially methylated domains coincide with lamina associated domains, observed by their overlap with the locations of lamin-B1, seen earlier in **Fig. 1**, genome spatial organization, through NL attachment, and DNA methylation are likely linked epigenetic features that contribute to cancer when dysregulated[1].

C. Methods and Techniques

1. Fluorescence-activated cell sorting



Name	Events	%Parent	%Total
☐ All Events	100,000	0.00%	100.00%
🟦 Non Debris	69,258	69.26%	69.26%
🟩 Doublet Discrimination	52,568	75.90%	52.57%
🟧 G1/S Phase	5,739	10.92%	5.74%
🟪 G1 Phase	12,344	23.48%	12.34%
🟦 G2 Phase	7,744	14.73%	7.74%

Fig. 6 | FACS plot of KBM7 FUCCI cells. (a) Remove debris through sorting gate on BSC-A vs FSC-A. **(b)** Remove doublets by selecting for low FSC-W for each FSC-H. **(c)** Gates for sorting G1 Phase (mOrange), G1/S phase (mOrange + EGFP), and G2 Phase (EGFP). **(d)** Fraction of sample corresponding to each gate, e.g. G1/S phase is 5.74% of All Events.

To sequence the epigenome of individual cells, they must first be isolated. The approach implemented to separate cells involves using fluorescent activated cell sorting (FACS) to sort single cells into individual wells of a 384 well plate for library preparation reactions. Cells are suspended in the FACS machine and flowed through a narrow channel to produce a stream of individual cells[38]. These cells then pass through a laser one by one that detects what stage of the cell cycle the incoming cell is at based on fluorescent excitation of the FUCCI reporter system, while additionally confirming individual cells are being sorted, rather than debris, using a Hoechst or DAPI nuclear counterstain[39]. Gates are created based on the desired fluorescent signals above a threshold level **(Fig. 6)**, and droplets containing each single cell are electrically charged and deflected as they pass through charged deflector plates, into either single wells in the 384 well plate or into the discarded cells collection tube if the criteria was not met[38]. In the proposed method, cells are initially gated to remove debris and doublets, and are then sorted for those containing both FUCCI markers, indicative of the G1/S phase transition during which chromosomal order is stable and LADs are in contact with the NL[40].

2. Methods for quantifying genome-nuclear lamina interactions in single cells

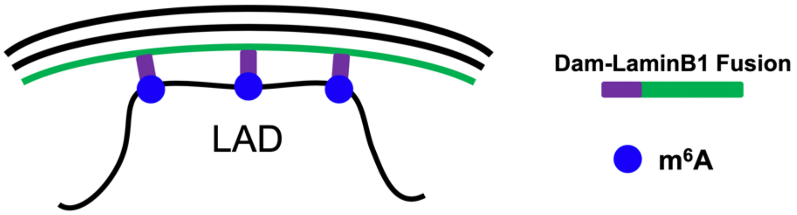


Fig. 7 | Strand of genomic DNA (black wavy line) contacts the nuclear lamina and the Dam fusion adds the m^6A (blue circles), marking the genome-NL interaction.

A wide range of techniques have been implemented for quantifying interactions between lamina associated domains and the nuclear lamina. One of the original techniques to study chromatin dynamics used a GFP-Lac repressor fusion protein (GFP-LacI) that targeted *lac* operator (*lac*O) sequences within various euchromatin sites in live *Drosophila* spermatocyte nuclei, permitting visualization of interphase chromatin motion and organization throughout the progression of the cell cycle, however it was not directly applied to study LADs[41]. Using fluorescence *in situ* hybridization (FISH) to probe regions that were known to either highly or lowly associate with lamin-B1, showed the former tended to be located closer to the nuclear rim, however FISH can only be performed in fixed cells and light microscopy has a limited resolution, making it difficult to determine whether the region directly contacted the NL, or was simply close by[22]. DamID measures contacts between the genome and the nuclear lamina by transfecting cells with a fusion of lamin-B1 and bacterial protein, DNA adenine methylase (Dam), that methylates adenines in a “GATC” sequence context, a modification known as N^6-methyladenosine (m^6A)[27]. Thus, when genomic DNA comes in close proximity to the nuclear periphery, the Dam-LaminB1 fusion protein methylates the A’s in GATC sequences, leaving a chemical modification on the DNA **(Fig. 7)**. To observe genome-NL contacts during specific time frames, the Dam-LaminB1 fusion is regulated by a destabilization domain (DD). Fusion of the DD to the protein of interest, in this case the Dam-LaminB1, results in instability, and consequently degradation of the entire fusion protein, preventing m^6A markers from being added to LADs. However, with the addition of a ligand, known as Shield1 (Shld1), the fusion protein is protected from degradation and will be functional[42]. Therefore, by adding Shld1 to the cell culture before cell sorting, genome-NL contacts can be observed within a chosen time frame, to ensure cells are given the maximum amount of time for the m^6A marks to be placed on the LADs before the mark is diluted from DNA replication[27]. The concentrations of Shld1 used for DamID ranged from 0.5 nM to 500 nM, requiring us to test a range of concentrations to find the optimal concentration for our KBM7 cell line[8,27,43–46]. One method used to assess genome-NL interactions was a Tet-Off system, in which the removal of Doxycycline (Dox) permitted the transcription of an m^6A tracer for visualization of which regions obtained the mark from contacting the nuclear lamina[47]. Although this allowed visualization of m^6A tracer signal colocalization with the nuclear periphery after only 5 hours of Dam-Lamin B1 induction, the method is qualitative and does not provide sequence information of the LAD, as well as requiring constant dosing of cells with Dox prior to induction[27].

To measure genome-NL interactions both in single cells and quantify them in terms of their sequence, a next generation sequencing (NGS) approach may be implemented. Briefly, DNA of single cells is cleaved by DpnI, an enzyme that recognizes m^6A on GATC motifs, making a blunt cut between m^6A and T, and barcoded adapters are ligated to the cut sites for sequencing, such that the NL contacting genomic regions can be identified by mapping the fragments back to the human genome[8]. This method for measuring genome-NL contacts is versatile and could be combined sequencing other epigenetic features, such as 5mC, or the transcriptome, and therefore an in-depth description and schematic of the method will be provided in *Method for simultaneously measuring genome-NL contacts and 5mC in single cells* and *Chapter II: Developing Strategies for Measuring Single Cell 5mC & Genome-NL Contacts*. Recently, DamID was adapted to simultaneously measure protein-DNA contacts and the transcriptome in single cells, confirming it can be implemented in a multi-omics manner[43]. Lastly, this technique was recently adapted into a microfluid approach, called μDamID, that uses a similar workflow and includes both m^6A tracer imaging and library preparation, while eliminating the need for FACS, however it is restricted to only 10 single cells in parallel, and thus the proposed method has a much greater throughput of 384 single cells[44].

3. Methods for quantifying 5mC in single cells

The method for measuring 5mC in single cells is much less involved than measuring genome-NL contacts, since the mark is endogenously present. 5mC can be globally detected using immunofluorescence by tagging 5mC with antibodies, and many have been created and improved upon since their original development in 1982[48,49]. Rather than an antibody approach, an inverse detection method could be implemented to detect which regions in the genome are not methylated, such as in a CpG methyltransferase (M.SssI) assay, in which all non-methylated cytosines in a "CG" sequence context are methylated using radiolabeled methyl groups from S-adenosyl-L-[methyl ^{3}H] methionine (SAM[^{3}H]) and radiometrically measuring the lack of original 5mC[50]. However, these methods will only provide information about the total 5mC level, and not any sequence information, but these methods could still prove useful in measuring the extent of DNA hypomethylation in cancer. Measuring the specific sites of 5mC in sequences of genomic DNA has been achieved using a method called bisulfite sequencing[51]. In this method, sodium bisulfite is reacted with genomic DNA resulting in the conversion of cytosine to uracil, while 5-methylcytosine is protected from the conversion. By then amplifying a region of interest with PCR, all uracil nucleotides in the bisulfite converted strands become thymine nucleotides in the PCR product, and 5-methylcytosine amplifies as cytosine. Thus, by deciphering the unchanged cytosines in the PCR product of bisulfite reacted strands, relative to the unreacted strands, 5mC sites can be inferred. This method has limitations, such as the need to PCR amplify and sequence both reacted and unreacted strands if reference sequences are not available, such as variations in alleles, and cytosines adjacent to methylated CpG sites are partially resistant to the bisulfite treatment, leading to 5mC artifacts[52]. However, there are 5mC recognizing restriction endonucleases that can cleave downstream from the site, allowing for adapter ligation and sequencing of the DNA modification without the need for base pair conversions and comparisons to unreacted sequences, similar to DpnI in DamID for recognizing m^{6}A sites. MspJ1 restriction enzymes recognize ^{me}CNNR in the 5’ to 3’ direction (where R = G or A) and generates a cut 12 nucleotides downstream of the 5mC nucleotide on the plus strand, and 16 nucleotides on the minus strand, generating a four-base 5’ overhang[53]. Adapters may be ligated to these fragments and sequenced to detect the 5mC sites in the epigenome. Although many of these approaches were originally developed to measure 5mC in a bulk population of cells, they may all be done at the single-cell level by first FACS of the cells, then performing reactions on individual cells in wells.

4. Method for simultaneously measuring genome-NL contacts & 5mC in single cells

Despite the wide range of methods for measuring genome-NL contacts and 5mC, no group yet has implemented simultaneous sequencing of these epigenetic features in single cells. One method however, methylation-specific fluorescence *in situ* hybridization (MeFISH), has allowed visualization in single nuclei of hypomethylation at satellite DNA, the main component of heterochromatin, which typically resides at the NL[54]. The method utilizes interstrand complexes formed by osmium and bipyridine-containing nucleic acids (ICON) probes which have differential affinity for cytosine and 5-methylcytosines on the DNA target[54]. By designing fluorescently labeled probes with a specificity for satellite sequences, *in situ* hybridization was achieved, allowing for detection of satellite locations and DNA methylation by observation of the FISH signal[54]. Crosslinking of the ICON probes with osmium creates linkages to only 5mC, and removal of non-crosslinked probes by denaturation produces probes only on the 5mC. Thus, a second measurement, the MeFISH signal, is captured, displaying the extent of 5mC which is proportional to the fluorescent signal. Through this, researchers have observed a remarkable decrease in peripheral fluorescent signal at classical satellites, representing hypomethylation in human lymphoblast cells from patients with ICF syndrome, a disease that is characterized by low levels of DNA methylation at the classical satellites[54]. However, similar to the FISH technique in the section describing methods of measuring genome-NL contacts, it is difficult to determine whether the region directly contacts the NL or was simply close by. Similar to some of the methods for measuring 5mC, this technique does not reveal information about which cytosines are methylated in the sequence, and rather displays a broader picture of the extent of hypomethylation. Finally, it requires *a priori* knowledge of which sites may exhibit hypomethylation in order to design ICON probes to target the regions, and in practicality we are seeking to discover previously unknown aberrantly methylated regions with our method. Although measuring 5mC and genome-NL contacts within the same cell have never been achieved before, it has been measured on the same molecule by using the third generation sequencing Nanopore technology combined with DamID, detecting cytosine methylation state on DpnI cut fragments based on the difference in current compared to unmethylated cytosine[55,56]. An alternative single molecule approach was achieved with antibody-binding protein A fused to deoxyadenosine methyltransferase Hia5 (pA-Hia5) to target lamin B1 and add m^{6}A methylation to the nearby DNA contacting this nuclear lamina protein, allowing for its simultaneous measurement with 5mC when sequenced with Nanopore[57]. Both single-molecule methods are limited by how long of DNA strands they can measure, and therefore fail to capture the 5mC and genome-NL contact profile of entire chromosomes in a single cell. Measurement of very long stretches is necessary to observe the full contact and noncontact runs within LADs, as these alternations of contact show significant correlations with the 5mC of the same cell (discussed in detail in *Chapter III*).

5. Transcriptome and Epitranscriptome and Methods of Measurement

The genetic information stored in the DNA of an organism is expressed through its RNA transcripts, and the collection of all of an organism’s transcripts is known as its transcriptome[58]. By profiling the RNA a cell produces, the cell’s behavior can be understood. It is critical to measure the RNA of individual cells rather than a bulk population because cells display variability in the their gene expression, meaning the average measurement may not represent what is actually occurring in each cell. A variety of techniques have been developed to achieve single-cell measurements, however some are very limited such as real-time PCR measurements or microscopy, which can only measure some of the organism's genes rather than the full set[59]. To address these issues, high-throughput sequencing based approaches, known collectively as "RNA-seq", have been developed to overcome these limitations and measure the full transcriptome of single cells, while also solving the low material input issue by implementing amplifications steps. One commonly used 384 well plate method, called CEL-seq2, uses the polyA tail present on eukaryotic mRNA to attach a barcoded adapter with a unique molecular identifier (UMI) through reverse transcription (RT) to keep track of which RNA originated from each cell and transcript counts[60]. This method also pioneered using *in vitro* transcription (IVT) as a way to amplify the material prior to attaching sequencing adapters with random priming and PCR[60]. An alternative method to this, known as Smart-seq2, still utilizes an Oligo(dT) primer, however uses RT and terminal transferase to attach nucleotides to the 3' end of the synthesized cDNA to act as anchor for a locked nucleic acid containing template switching oligo to reverse transcribe the first strand[61]. Rather than IVT to amplify and random priming to incorporate the sequencing adapter, PCR amplification is performed followed by tagmentation with Tn5 to add the handles for the sequencing adapters[61]. Microfluidic RNA-seq approaches have been developed to increase cell throughput, such as Drop-seq, which encapsulates individual cells in nanoliter droplets alongside microparticles that have 10^8 barcoded RNA-seq adapters attached to them, each with a UMI and a unique barcode per microparticle, allowing for production of a library containing 10,000 single-cell transcriptomes[62]. Extensions of RNA-seq exist such as scSLAM-seq which can measure transcription dynamics through the addition of the nucleoside analogue 4-thiouridine (4sU) to cell culture medium two hours before cell sorting[63]. Since 4sU is incorporated into newly transcribed RNA, adding iodoacetamide (IAA) prior to sequencing converts the 4sU into cytosine, allowing old and new RNA to be identified based on which RNA contains the U-to-C conversions[63]. RNA-seq can be paired with other measurements in the same cell, such as the genomic DNA in DR-seq, which PCR amplifies primarily DNA-derived fragments while separately IVT amplifying mRNA-derived fragments that had been attached to a CEL-Seq-like poly(T) adapter containing a T7 promoter[64].

Similar to the way DNA can be chemically modified, transcribed RNA can be further modified by over 170 chemical modifications, known as the epitranscriptome[65]. Some of these modifications include pseudouridine (Ψ), N^6-methyladenosine (m^6A), N^1-methyladenosine (m^1A), 5-methylcytidine (m^5C), N^4-acetylcytidine (ac^4C), N^7-methylguanosine (m^7G), and ribose methylations (N_m), which have the potential to regulate the fate of transcribed RNA[65]. Of these modifications, m^6A has been most well studied, since it is the most abundant RNA modification, and it is known to regulate chromatin state and transcription as well as influence cancer progression when dysregulated[66,67]. Like how methylation can be written and erased from the DNA by DNMT and TET, respectively, m^6A is written to RNA by N6-adenosine-methyltransferase (METTL3) and erased by alpha-ketoglutarate-dependent dioxygenase FTO and ALKBH5 (AlkB Homolog 5, RNA Demethylase)[65]. The m^6A RNA modification can be measured with an immunoprecipitation technique known as MeRIP-seq, that binds a Dynabead coupled-antibody specific to m^6A on fragmented RNA that had been extracted from a cell, allowing for its separation from the total RNA[68]. After separation, the m^6A containing RNA undergoes RNA-seq and the location of the epitranscriptomic mark can be determined. However, this method does have a limitation in that m^6A antibodies can also recognize the structurally similar m^6Am RNA modification, motivating the development of DART-seq, an antibody-free method for detecting m^6A[69]. This method uses APOBEC1, a cytosine deaminase, fused to the m^6A-binding YTH domain to convert cytidine bases adjacent to m^6A into uracils, such that after RNA-seq is completed, the C-to-U mutations indicate where the m^6A was present. There is also a more general approach to detecting whether or not epitranscriptomic modifications are on the RNA, which has been demonstrated on tRNA in a technique called AQRNA-seq[70]. Since cDNA synthesis by reverse transcription is frequently blocked or mutated by RNA modifications, the location where the RT ends indicates the presence of the mark. The mark that was present can be determined by searching the location in a database of all the RNA modifications in a given cell line if that cell line had been well studied. Additionally, by dividing a bulk sample into two and using AlkB treatment on one fraction to remove all the marks, the full RNA sequence can be mapped in that fraction, while also measuring the truncated sequences in the other fraction, for less documented cell lines.

6. Histone Modifications and Methods for Measuring

In order to package the DNA in the nucleus into a condensed structure, it is wrapped around a histone octamer consisting of two copies of each of the four core histones: H2A, H2B, H3 and H4[71]. The N-terminal tails on the histones are often modified at their lysine and arginine amino acids by methylation and acetylation **(Fig. 8)**, which controls gene expression through gene activation and repression **(Table 1)**[71]. Histone methylation and acetylation are added by histone methyltransferases and acetyltransferases, respectively, such as EZH2 for adding trimethylation (me3) to histone 3 lysine 27 (H3K27) to cause gene repression, and

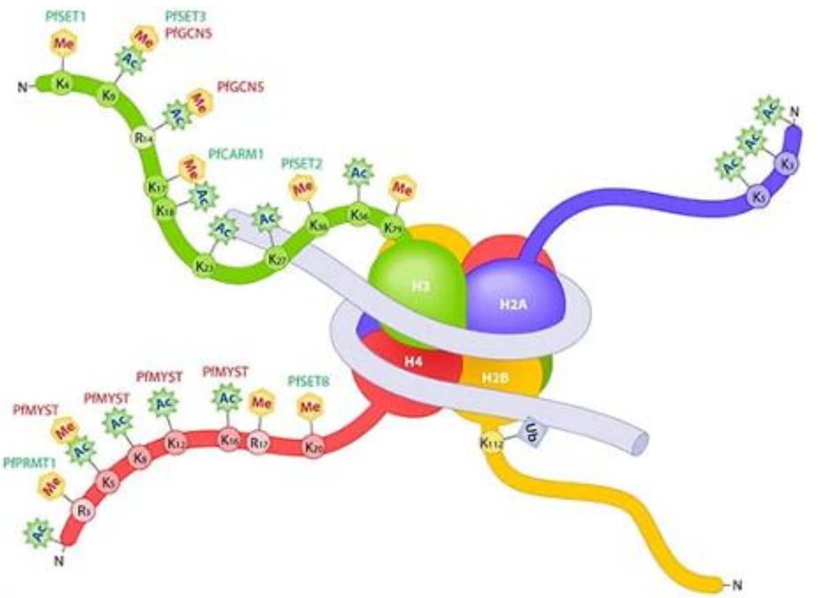


Fig. 8 | Overview of Activating and Repressive Histone Modifications. Histone octamer with modifications on N-terminus tails (Image Source: *AMSBIO, Histones and Nucleosomes*)

removed by histone demethylases and deacetylases, respectively, such as lysine-specific demethylase 1A (KDM1A) to remove the monomethylation from the transcription activating H3K4me1[72]. ChIP-seq (chromatin immunoprecipitation followed by sequencing) is a strategy to identify where histone modifications are located by first fragmenting the sample, and then subsequently using an antibody specific to the histone modification of interest to enrich for the locations in the genome containing the marks[73]. Following the isolation of the DNA-histone complex, the histones are degraded with a protease, adapters are added to the DNA adjacent to the

Modification	Function	Location
H3K4me1	Activation[71]	Enhancer[72] and promoter[71]
H3K4me2	Activation[72]	Promoter[71]
H3K4me3	Activation[72]	Promoter[72] and bivalent domains[71]
H3K9ac	Activation[74]	Enhancer and promoter[71]
H3K9me3	Repression[71]	Gene poor regions (satellite repeats in telomeres and pericentromeres)[77]
H3K27ac	Activation[75]	Enhancer and promoter[71]
H3K27me3	Repression[72]	Promoters, gene body, bivalent domains[71]
H3K36me3	Activation[71]	3’ end of gene body[71]
H3K79me2	Activation[75]	Gene body[75]
H4K16ac	Activation[76]	LINE1 5’ untranslated regions[78]

Table 1 | Histone modification table of function and locations found on the chromosomes.

modified histone, and next-generation sequencing is performed to determine where the mark had originated[73]. ChIP-seq has its limitations though, such as requiring many cells to overcome its low yield that produces high background noise and low signal, so alternative strategies have been developed, such as CUT&Tag (Cleavage Under Targets and Tagmentation), which works at the single-cell level[79]. Primary and secondary antibodies bind the histone modification and allow for the attachment of a hyperactive Tn5 transposase-Protein A (pA-Tn5) fusion protein holding sequencing adapters, which are inserted directly into the DNA near the histone modification, facilitating the next steps of library preparation[79]. Histone modification can be measured alongside other features in the same cell, such as the transcriptome in Paired-Tag (parallel analysis of individual cells for RNA expression and DNA from targeted tagmentation)[80]. This strategy combines pA-Tn5 for detecting the histone modification and RT with CEL-Seq-like adapters for the transcriptome component. Reactions are performed on nuclei in different wells with well specific barcodes, each for the transposase and RT primers, followed by a ligation-based combinatorial barcoding strategy and splitting the chromatin DNA and cDNA into two sequencing libraries corresponding to each profile[80].

II. Developing Strategies for Measuring Single Cell 5mC and Genome-NL Contacts

A. Simultaneous digestion with DpnI and MspJI

1. Nonforked 5mC adapters (v1.0)

I developed a new method that involves incubating genomic DNA acquired from DD-Dam-LaminB1 cells with both DpnI and MspJ1 to fragment the sequences at locations that either contacted the NL or contain 5mC, thereby producing fragments that adapters could be ligated to for sequencing. To measure NL contacts, we use a stable, clonal human cancer myeloid leukemia cell line (KBM7) that has the LmnB1-Dam fusion protein for measuring the NL contacts and the FUCCI system for determining the stage of the cell cycle. The ligand Shld1 activates the fusion protein and is added 15 hours before sorting such that cells have the maximum amount of time for Dam to place m^{6}A NL contact marks at LADs, before there is an accumulation of marks upon genome replication (after S phase). The concentration of Shld1 had been increased from an initial 0.5 nM to 500 nM for this method, producing a greater percentage of Dam reads out of the total 5mC+DamID reads. To allow sequencing of individual epigenomes, the FACS machine sorts the Shld1 treated single cells that are in the G1/S phase (via the FUCCI system) into different wells of a 384-well plate. A nanoliter dispensing liquid handling-robot performs the subsequent reactions on the cells to ensure rapid and accurate pipetting below the limit of manual pipetting. Cells are lysed to extract genomic DNA and protease is added to make the chromatin accessible for enzymatic cleavage **(Fig. 9a)**. Protease needs to be added because closed chromatin has been shown to restrict the ability for enzymes to interact with DNA in those regions, which is the basis for

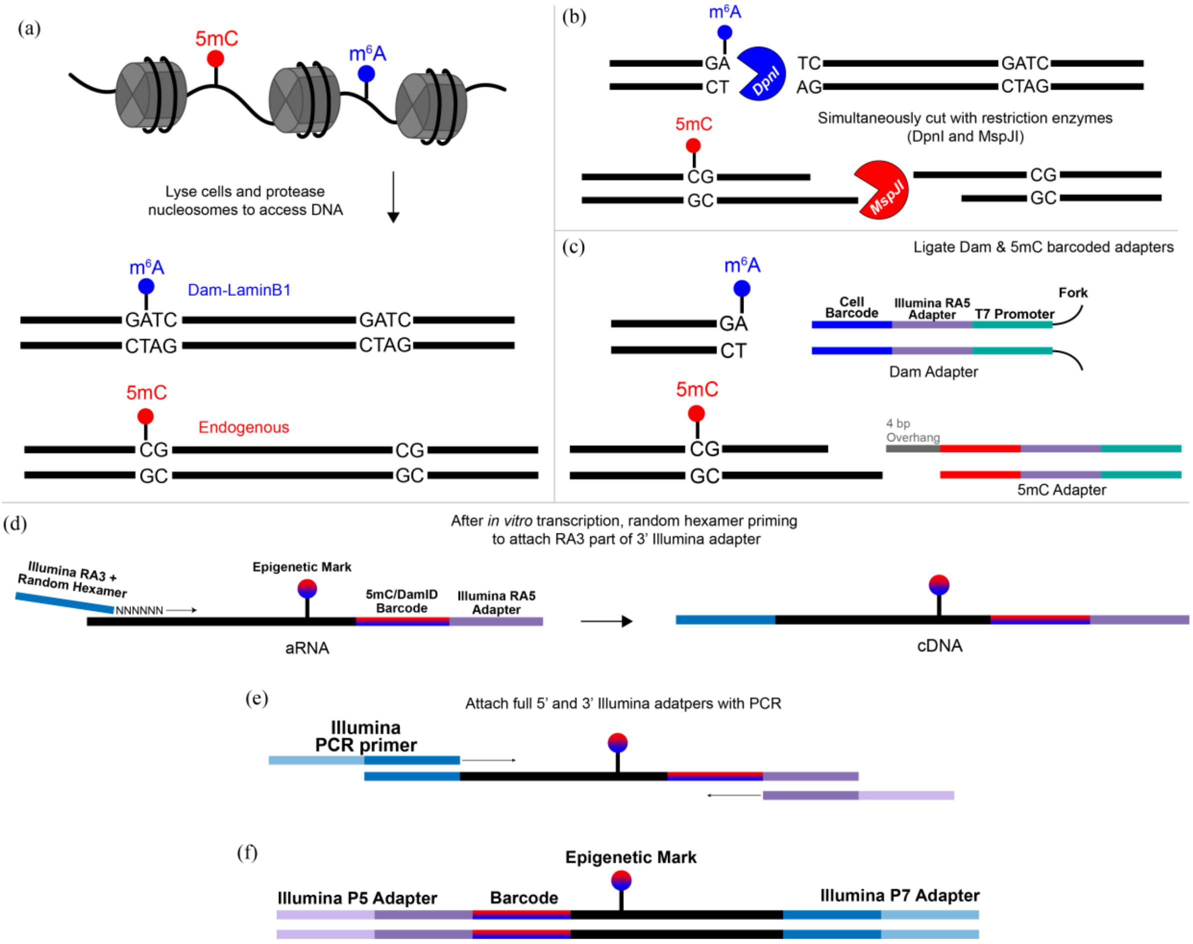


Fig. 9 | Single cell 5mC+DamID (v1.0) Library Preparation Schematic. (**a**) Lysing cells protease nucleosomes to access DNA. (**b**) DNA digestion with DpnI and MspJI restriction enzymes to detect genome-NL contacts and 5mC marks, respectively. (**c**) Attachment of barcoded adapters to DNA fragments by random hexamer priming (**d**) cDNA of amplified, reverse transcribed fragments. (**e**) PCR primers for incorporation of Illumina sequencing adapters (**f**) Final library structure.

many DNA accessibility sequencing techniques such as DNase-seq and ATAC-seq[81]. Restriction enzymes are added to the wells to simultaneously digest the DNA sequences pertaining to each mark. The DpnI enzyme recognizes and cleaves DNA at the adenine methylated GA^{me}TC sites, corresponding to genome-nuclear lamina interactions, and the MspJ1 enzyme cuts downstream from sites containing 5mC **(Fig. 9b)**. Cleavage of mammalian genomic DNA with DpnI is blocked by overlapping CpG methylation, reducing ambiguity of reads in regions containing both marks. Two types of adapters, pertaining to either the Dam or 5mC epigenetic mark (blue/red), are ligated to each respective cut site **(Fig. 9c)**. The first component of each adapter is a cell specific barcode to discern which cell the epigenomic information corresponds to, as outlined in CEL-seq, and a unique barcode is used for each epigenome in the well plate[60]. The second component is the priming site for incorporation of the Illumina P5 adapter, known as RA5, which is one of the adapters necessary for sequencing. The third component of the adapter is a T7 promoter, which allows for *in vitro* transcription (IVT) to linearly amplify the DNA fragments from the mere picograms in the single cell to nanograms. Linear amplification ensures low abundance sequences are not lost due to bias in the amplification, which would normally occur in a typical exponential amplification, such as PCR[82]. Lastly, the 5mC adapter has an overhang for ligating to only MspJ1 cut sites, and the DamID adapter has a forked end to prevent the adapters from forming long chains by blunt end-to-end ligation. Following ligation, the barcoded 96-cell libraries are pooled into single tubes and reactions can then be performed as in bulk, using manual pipetting. Bead clean-up, a form of solid-phase chromatography with reversible immobilization using magnets, is implemented to purify the DNA from the reagents[83]. *In vitro* transcription is performed to amplify the genetic information, the RNA product is fragmented to increase sequencing efficiency, and is followed by an additional bead cleanup[84]. Next a random hexamer reverse transcription (RT) primer is used for reverse transcription of the amplified RNA (aRNA) back to complimentary DNA, adding the RA3 sequence, the priming site for incorporation of the Illumina P7 adapter **(Fig. 9d)**. Lastly, Illumina PCR primers are used to amplify the library into its final form, incorporating the full 5’ Illumina P5 and 3’ Illumina P7 sequencing adapters via the RA5 and RA3 priming sequences, respectively **(Fig. 9e)**. The advantage of Illumina is many samples can be sequenced in parallel, resulting in high efficiency with a relatively low cost per base sequenced, which is likely why Illumina has formed a monopoly in the sequencing market over the past decade[85]. The raw Illumina read output are then mapped back to the human reference genome using Burrows-Wheeler Alignment (BWA) tool, which provides fast and accurate alignment for both short and long reads, which is interpreted by SAMtools to remove PCR duplicates, and exported as a delimited file consisting of cell, chromosome, position the read mapped to, and strand for both the 5mC and Dam (DNA m^6A) reads[86–88]. By implementing this method, the sites of 5mC and genome-NL contacts within single cells can be determined.

To test the efficiency of the method, mapping statistics were performed, including measuring the ratio of Dam:5mC reads, read counts of each mark, fraction of sequenced reads that mapped to the genome, and the redundancy of reads. To pick the best cells for investigation, sequenced cells in each condition were sorted by those with the most Dam reads, and all cells with under 10,000 5mC reads were removed. Initial attempts indicated an abundance of 5mC reads at the expense of few Dam reads, so the 5mC:DamID adapter concentration ratio was tuned with the aim to improve read acquisition, holding the concentration of DamID adapter constant **(Fig. 10a)**. The goal was to increase Dam reads without reducing uniquely mapped reads of both epigenetic marks. Although lowering the 5mC adapter concentration increased the percentage of Dam reads out of the total number of 5mC and Dam reads **(Fig. 10a)**, it decreased the amount of unique 5mC reads per cell to some extent **(Fig. 10b)** and the overall complexity of the library, seen by noisy 5mC profiles containing more empty bins and less dynamics in the read count along the chromosomes

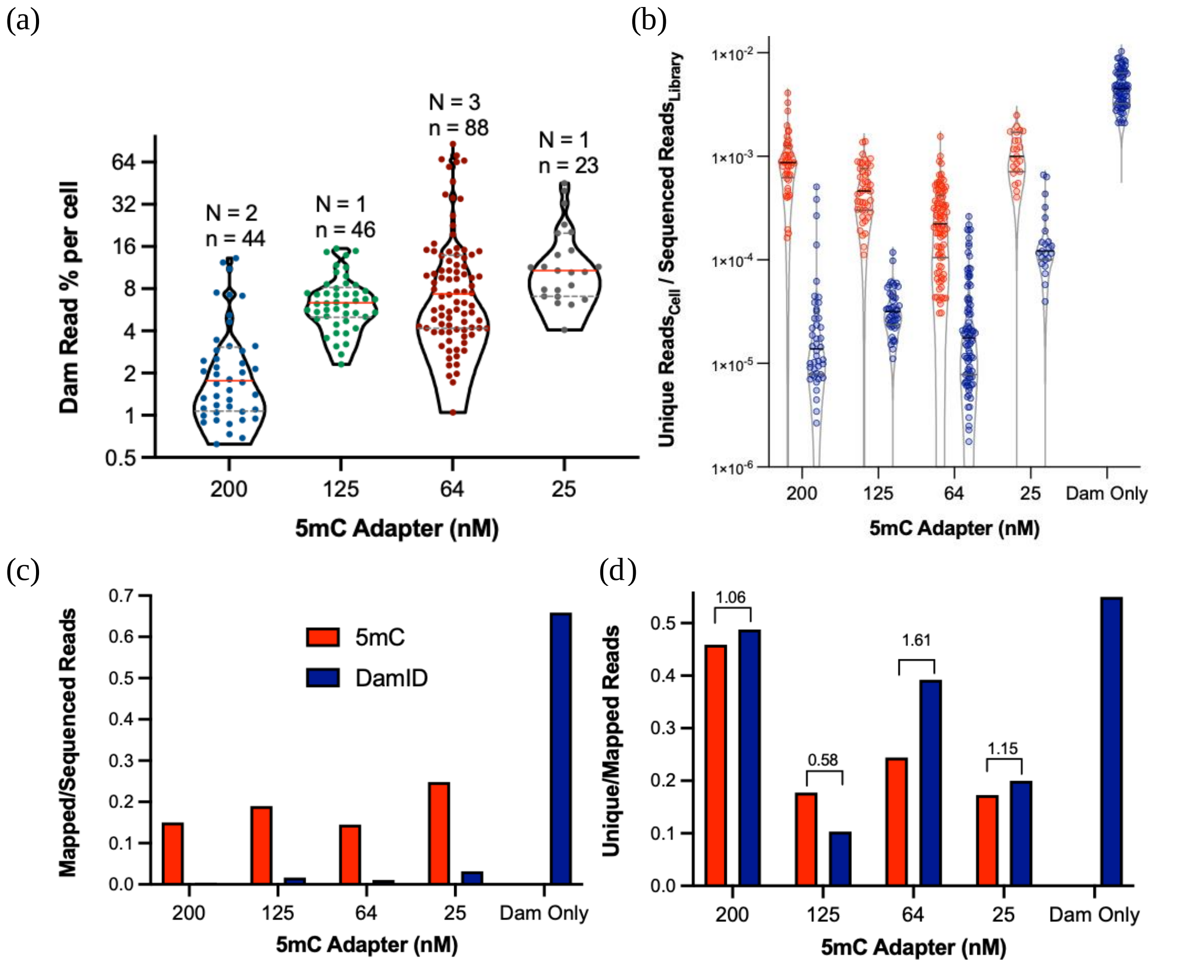


Fig. 10 | sc5mC+DamID v1.0 Mapping Statistics. (a) Percentage of total reads that are Dam (originating from DpnI, the genome-NL contacts). **(b)** Unique reads/total reads (normalizing to sequencing depth) for each mark (5mC in red and Dam in blue) in each cell. Showing only cells with > 10,000 unique 5mC reads. **(c)** Redundant reads, RABA/(FASTQ/4) and **(d)** Unique/redundant reads, FABA/RABA for each library.

(data not shown). The DamID signal was only good quality in a low fraction of the cells, as most had high background noise, making it challenging to call which regions were LADs and iLADs. A histogram of Dam reads in single-cell DamID only displays a bimodal distribution, the left peak consisting of off target m^{6}A, considered background noise, and the right peak consisting of regions that made nuclear lamina contacts, and therefore are LADs **(Fig. 11)**. Dam reads were normalized by observed over expected (OE) (*Appendix B*). The basis for this is that although adjacent or close by GATC sequences exist in the human genome, they cannot necessarily be mapped due to the restriction enzymes creating too short of fragments **(Fig. 24c)**, resulting in inaccurate normalization of the data in regions containing closely spaced GATC sequences. To generate the mappable GATCs to use as a normalization, *in silico* DpnI digestion was performed on the human genome by generating potential DamID read sequences of 65 nt (in both directions) from the reference human genome, which were aligned back to the reference genome and processed like the DamID data[89]. By design, OE scores greater than 1 correspond to DamID reads that are greater than expected by chance under the null hypothesis that the entire genome randomly contacts the NL, and therefore bins that contain an OE score greater than 1 are considered to be in contact with the nuclear lamina[8]. Indeed, the histogram of OE scores, seen in **Fig. 11**, displays a bimodal distribution of values with the right peak beginning at OE = 1, suggesting regions of the genome associated with the right peak are making NL contacts. Comparing the unique reads/sequencing depth to scDamID only with same concentration of Dam adapters (64 nM), there are orders of magnitude more Dam reads verses the combined 5mC+DamID method **(Fig. 10b)**.

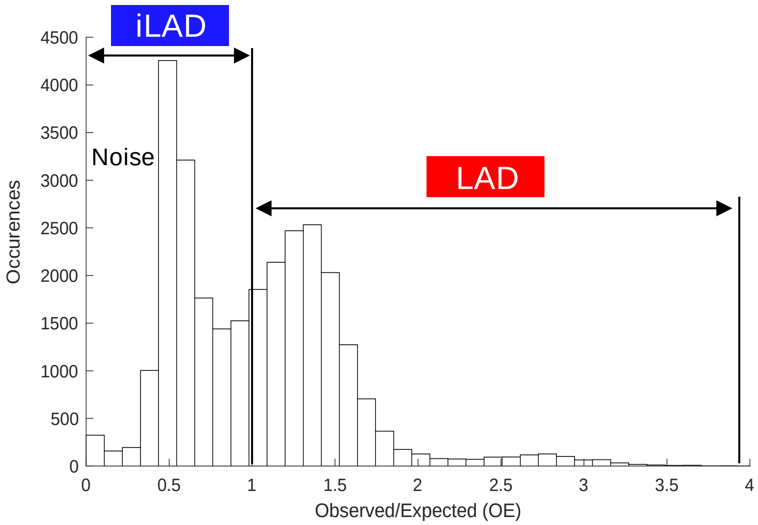


Fig. 11. Observed over expected (OE) Dam reads in 85 single cells. OE scores per cell were calculated in 100 kb bins and occurrences across the single cells were summed.

Due to abundance of 5mC and low read counts of Dam, we believed the Dam and MspJI adapters may be interacting in a way to ultimately decrease the amount of Dam reads. The lowest concentration 25 nM 5mC adapter produced the best mapping/sequencing depth for 5mC and DamID reads, which supports the idea that too high of a 5mC adapter concentration may increase the amount of junk reads in the library, as Dam mapping generally decreased at higher 5mC adapter concentrations **(Fig. 10c)**. Unique read counts were worse when using low 5mC adapter concentrations compared to the 200 nM 5mC adapter condition, indicating the presence of redundant reads, suggesting higher adapter concentrations increase library complexity and sequencing efficiency **(Fig. 10d)**.

2. Forked 5mC Adapters (v1.1)

Even though 5mC and genome-NL contacts could be measured at high resolution in the same cell with the DpnI and MspJI approach **(Fig. 12a**) the efficiency was low because many cells had noisy Dam signal despite good 5mC quality, depicted in **Fig. 12b**. Visually, good quality Dam signal is when the background noise is below OE = 1, seen in **Fig. 12a** where the single-cell LAD peaks on chromosome 17 are all in similar locations and above OE = 1. However, a large fraction of the profiles had bad quality (noisy) Dam signal, like in **Fig. 12b** where the baseline OE score > 1, making it difficult to differentiate between signal and noise, and therefore genome-NL contacts cannot be interpreted. Due to this issue, the next step was to optimize the method to improve the Dam signal. The first optimization of the method was adding a fork to the tail of the 5mC adapter to prevent a blunt end ligation of the Dam adapter to it, which could result in interference of T7 binding and amplification. Forking the 5mC adapter didn’t change the 5mC profiles of the cells when compared to using the original adapter **(Fig. 12c)**. The second method optimization was increasing the Dam:5mC adapter

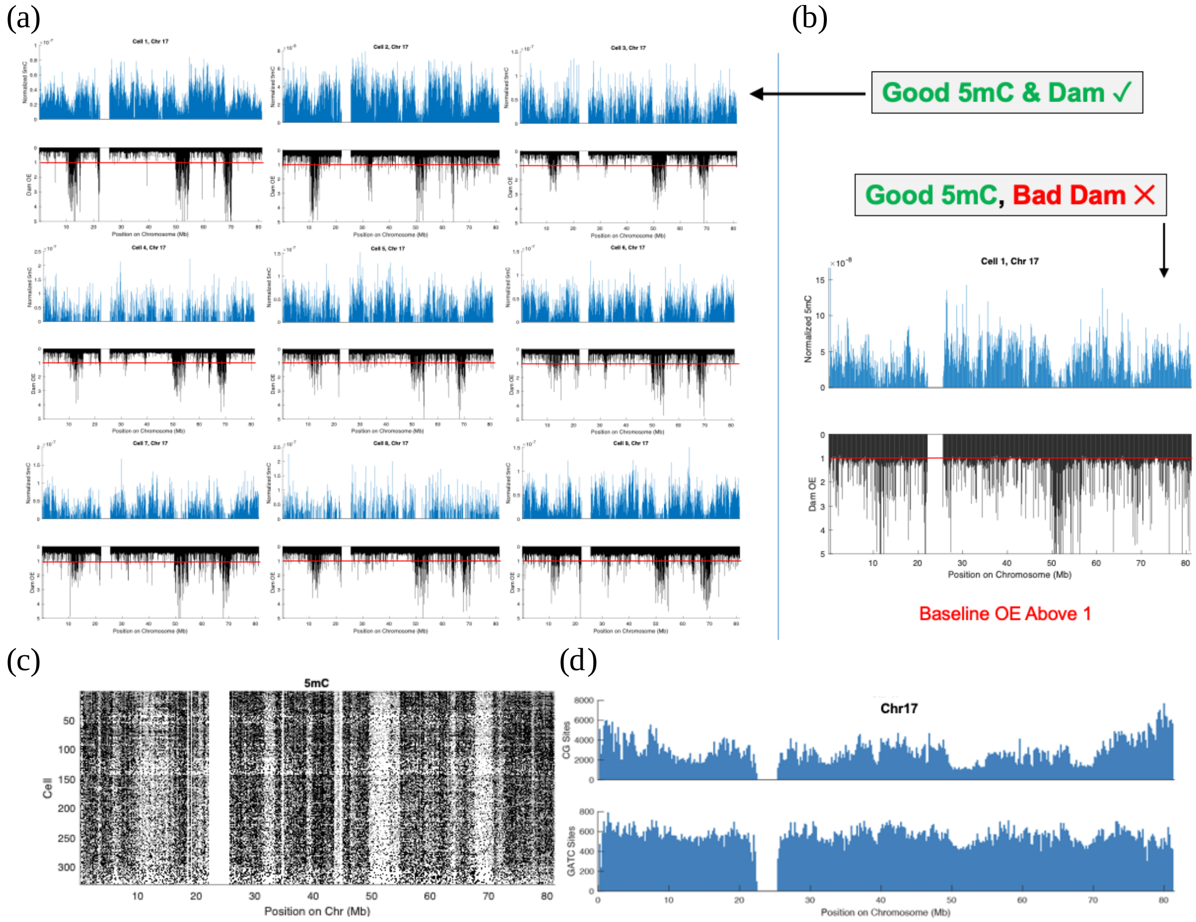


Fig. 12 | Optimization of Method: Improve Dam Signal. (a) Top 5mC and Genome-NL contact profiles from the same cell on Chr 17, with DpnI and MspJI approach. **(b)** Cell with good 5mC, but bad Dam signal. (a,b) Red horizontal lines drawn at OE = 1 **(c)** Single cell 5mC profiles from cells with and without forked 5mC adapters. **(d)** CG (top) and GATC (bottom) site profiles on chromosome 17.

concentration ratio to acquire more Dam reads. The human genome has 28,217,448 CG sites and of these 60-80% are methylated, resulting in 19.8 million 5mC sites that could be digested by MspJI[9]. In terms of possible genome-NL contact sites, there are 7,127,609 GATC for Dam-LaminB1 fusion to add m^{6}A, and considering 40% of the genome is LADs, there are approximately 2.9 million sites that could be digested by DpnI **(Fig. 12d)**[22]. The Dam signal is therefore harder to detect, considering it seven times scarcer than the 5mC signal, so to capture all of its signals more Dam adapter could be used by increasing its concentration in the reaction to improve the Dam adapter-DNA fragment collision frequency that precedes the ligation event. The last change of method was eliminating the double digestion with DpnI and MspJI, and instead cutting with the restriction enzymes at separate times to prevent collision on the DNA. Since DpnI is monomeric in structure and MspJI is tetrameric, it is possible that the latter may hinder the accessibility of the former to the DNA for cutting, possibly blocking a cut[90,91]. The order of restriction enzyme activity may also influence recognition of 5mC and genome-NL contact sites. The potential outcomes are adding the DamID recognizing enzyme first, then 5mC recognizing enzyme second will increase Dam read count, and vice versa. To test this hypothesis, individual experiments would be performed by the enzymes in both orders, deactivating the first before the addition of the second, and compare 5mC and Dam read counts to adding both at the same time. However, I observed in sc5mC only sequencing and scDamID only sequencing, the 5mC and genome-NL contacts were anti-correlated, meaning the enzymes shouldn't collide because they don't cut in the same location. This approach is also disadvantageous because it will take an additional day to preform library preparation to include a second digestion, so it would be more practical to focus on other optimization strategies. Upon testing this optimization, there was little product.

Forking the 5mC adapter could improve the Dam signal by preventing multiple adapter interaction issues from occurring. The first possible adapter interaction issue that would lead to reduction in the acquisition of Dam signal is if the blunt ended DamID adapter ligates to the tail of the 5mC adapter, because it would sequester it and prevent its ligation to the DpnI cut site **(Fig. 13a)**. The second possible issue occurs if the Dam adapter ligates to the tail of the 5mC adapter, and then this unit further ligates to the MspJI cut site. T7 RNA polymerases

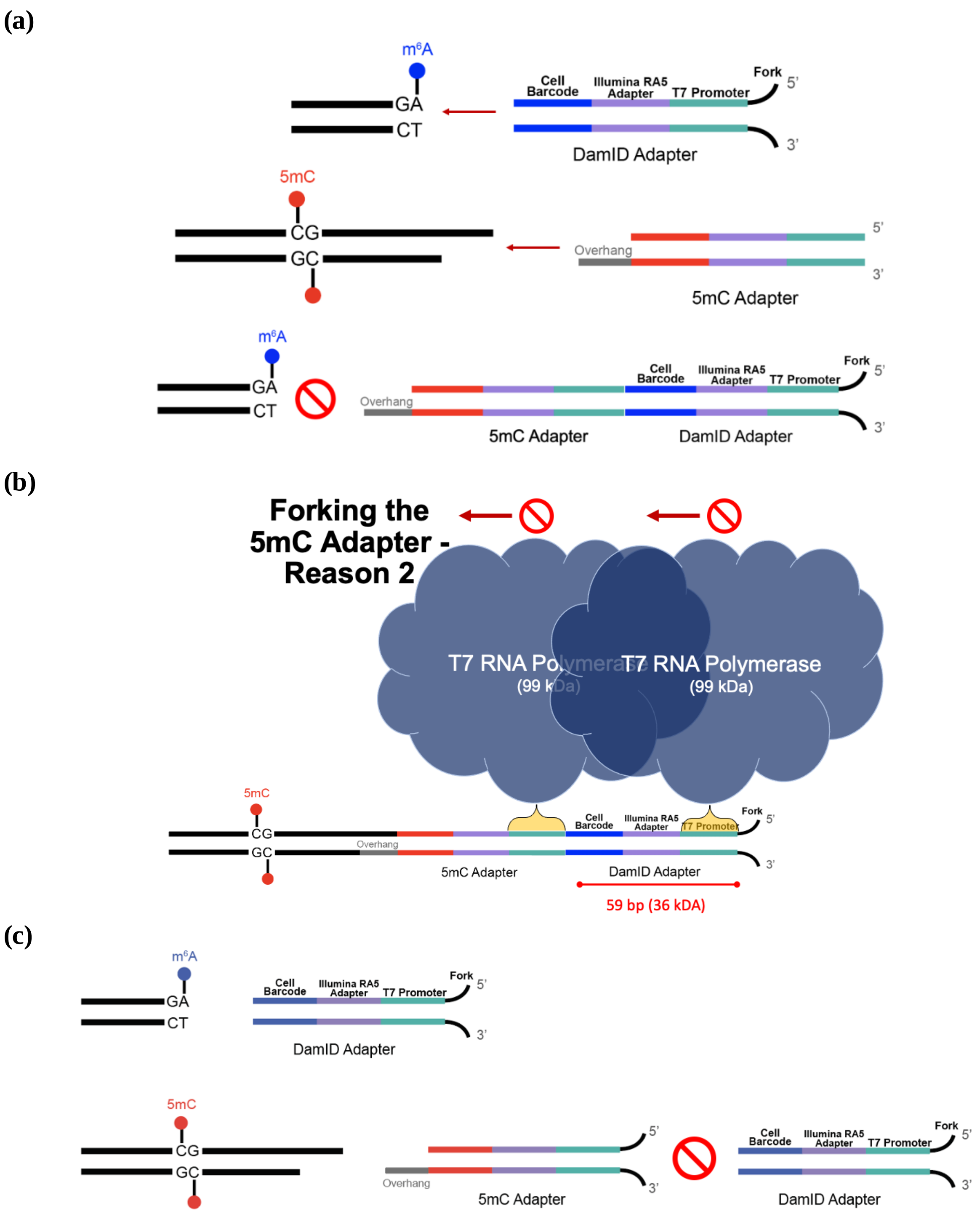
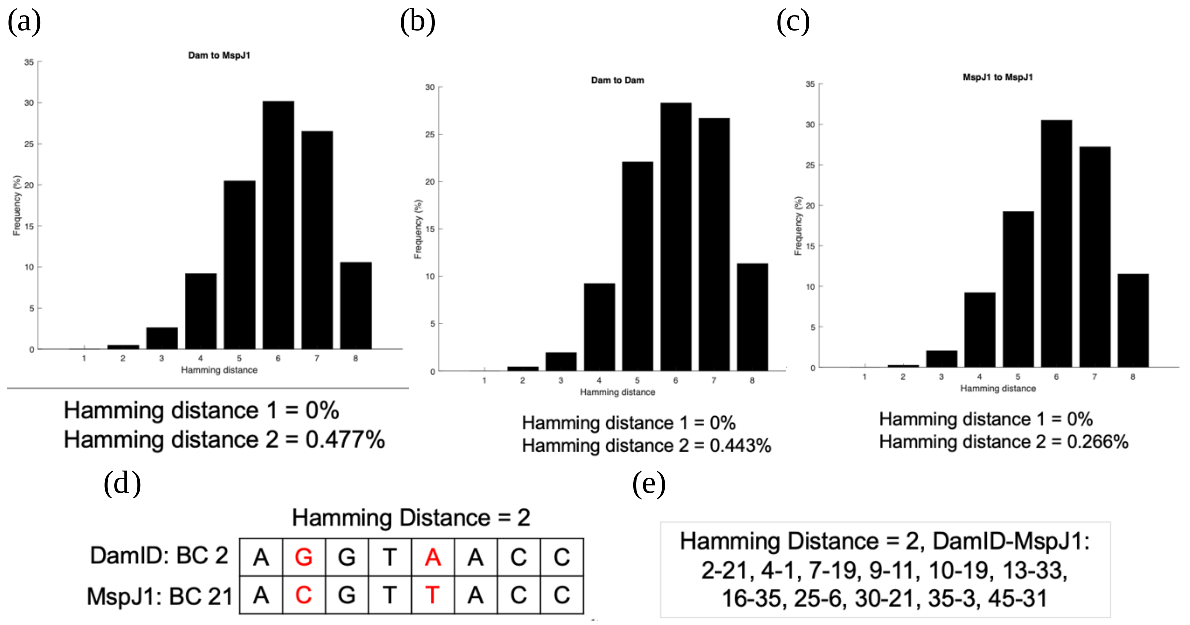


Fig. 13 | Rationale for forking 5mC adapters. (a) Dam and 5mC adapter can both bind to their respective cut sites, but the Dam adapter can also bind to the 5mC adapter. **(b)** 5mC-Dam adapter hybrid hinders T7 binding and could incorrectly bind to Dam adapter for 5mC. T7 Polymerase[92] drawn to approximate scale, as 2x the length of the DamID adapter, assuming molecular weight is proportional to volume, using 1 bp = 618 Da. **(c)** Forking 5mC adapter prevents formation of 5mC-Dam adapter hybrid.

may compete for the T7 promoter binding site, likely hindering each other's activity **(Fig. 13b)** due to the large size of the enzymes compared to the distance between the T7 promoters on the adapters. Thirdly, if the material is amplified starting at the T7 promoter site on the DamID adapter, the resulting sequence will contain two Illumina RA3 priming sites. This will cause problems with the PCR amplification **(Fig. 13b)** and mapping to the genome after Illumina sequencing, because it will try to align the full 5mC adapter + DamID cell barcode sequence to the genome, resulting in an unmappable read. By forking the 5mC adapter, there is no longer the risk of the DamID adapter and 5mC adapter hybridizing, preventing these possible interactions **(Fig. 13c)**.

Since the adapters were being redesigned, the hamming distance between the MspJI and DamID adapters were checked to insure minimal barcode sequence overlap. The hamming distance is a measure of how different two strings are, and in this context, it is the number of



Hamming distance 1 = 0%
Hamming distance 2 = 0.477%

Hamming distance 1 = 0%
Hamming distance 2 = 0.443%

Hamming distance 1 = 0%
Hamming distance 2 = 0.266%

Hamming Distance = 2

DamID: BC 2	A	G	G	T	A	A	C	C
MspJ1: BC 21	A	C	G	T	T	A	C	C

Hamming Distance = 2, DamID-MspJ1:
2-21, 4-1, 7-19, 9-11, 10-19, 13-33,
16-35, 25-6, 30-21, 35-3, 45-31

Fig. 14 | Adapter Hamming Distance between **(a)** Dam to 5mC adapter, **(b)** Dam to Dam adapter, and **(c)** 5mC to 5mC adapters. **(d)** Hamming distance example. Differences in bases highlighted in red **(e)** Dam to 5mC adapters that have hamming distance of 2.

corresponding locations that are different between two single-cell barcodes. An example of this is in **Fig. 14d**, where the hamming distance between the DamID adapter with barcode 2 and the MspJI adapter with barcode 21 is two because the 8 base pair sequences are identical except for the nucleotides in position 2 and 5, highlighted in red. It is best for the barcodes to have a high hamming distance in case a base is misread during sequencing, causing the signal to be interpreted as originating from a different cell. When comparing the DamID adapter barcodes to the MspJI adapter barcodes, a hamming distance of two only occurs 0.477% of the time, and a hamming distance of one doesn't occur, indicating a low probability of misinterpretation **(Fig. 14a)**. Likewise, when comparing the DamID adapters to themselves, a hamming distance of two occurs at 0.443% and hamming distance one occurs 0% of the time **(Fig. 14b).** When comparing the MspJI adapters to themselves, a hamming distance two occurs 0.266% of the time and a hamming distance of one does not occur **(Fig. 14c)**. Since there is a lack of sequence similarity between the barcoded adapters, it is unlikely that a misread will cause the signal to be interpreted as originating from another cell.

Using the forked 5mC adapters produced more consistency between sequencing data sets in terms of the unique read counts/sequencing depth, the unique over mapped reads, and mapped over sequenced reads. To test efficiency of the forked 5mC adapters on signal acquisition, the 5mC and Dam adapters were each varied from 64 nM – 128 nM. Using higher ratios of Dam:5mC adapters, such as 128:64 nM, produced the highest percentage of Dam reads per cell, and when using an equal ratio of adapters, such as 128:128 nM and 64 nM:64 nM, the condition with the higher concentration of Dam adapter had a slightly higher median and upper quartile **(Fig. 15a)**. The forked 5mC adapters produced a more consistent mapping **(Fig. 15b-d)** between conditions with varying adapter concentrations than when

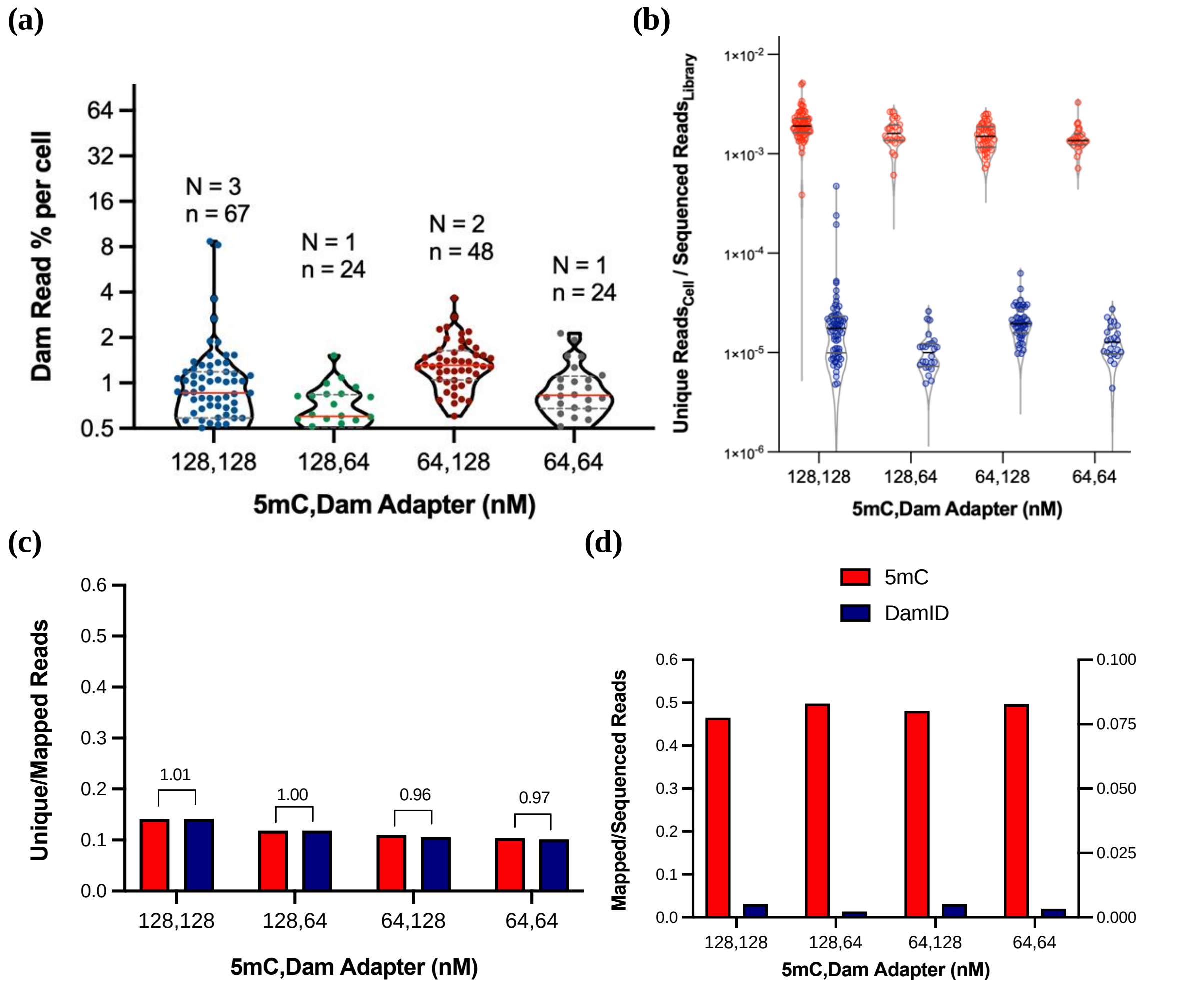


Fig. 15 | sc5mC+DamID (v1.1) Mapping Statistics with forked 5mC adapters. (a) Percentage of total reads that are Dam (originating from DpnI, the genome-NL contacts). **(b)** Unique reads/total reads for each mark (5mC in red and Dam in blue) in each cell. Showing only cells with > 10,000 unique 5mC reads **(c)** Redundant reads, RABA/(FASTQ/4) and **(d)** Unique/redundant reads, FABA/RABA for each library.

using the unforked 5mC adapters **(Fig. 10b-d)**, suggesting forking the adapters reduced the adapter interactions that lead to the inconsistencies when simultaneously measuring 5mC and Dam signal. While substantially improving the 5mC signal, this optimization did not improve the average Dam read count per cell, and in fact it actually decreased the amount of Dam reads compared to the unforked 5mC adapter condition with the same adapter concentration **(Fig. 16b)**. The improvement it made was increasing the mapping efficiency of 5mC signal

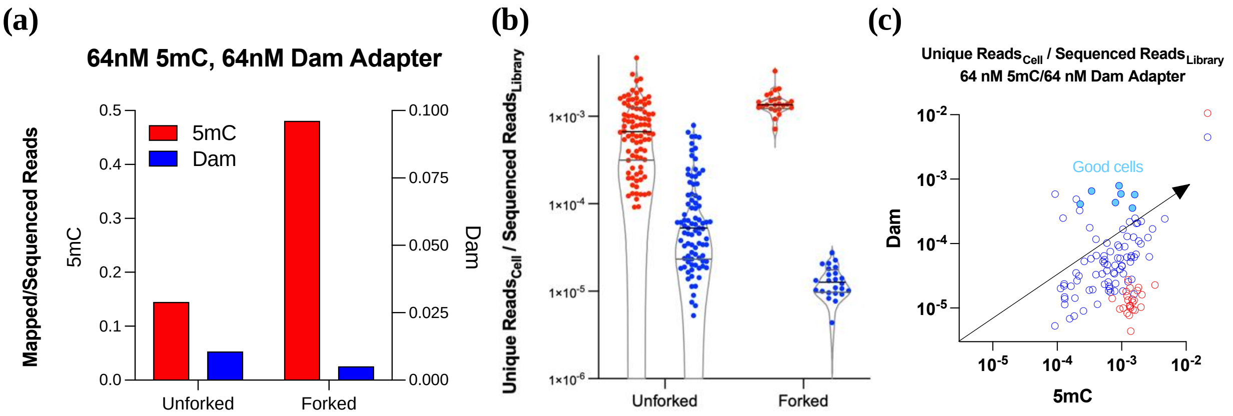


Fig. 16 | Comparisons between data quality with forked and unforked 5mC adapters (a) Mapped/sequenced reads for 5mC and Dam **(b)** Unique/sequenced reads **(c)** Dam vs. 5mC reads of the same cell. Forked 5mC Adapter (red), Unforked 5mC (blue).

adapters do not fall closer to that region on the plot, and rather fall farther away. This suggests that although forking the 5mC adapters improves the reproducibility and consistency of reads between cells and samples, it alone is not the strategy to improving Dam read count.

The next optimization was increasing the Dam:forked 5mC adapter ratio to 8 and 16, given there are seven times fewer possible DpnI cut sites than MspJI cut sites. With the combination of these optimizations, it worked better in one cell than with the unforked 5mC adapters **(Fig. 17a, top left cell)**. Distinct LADs can be seen in other cells from the same conditions **(Fig. 17a,b next 4 cells)**, however their baseline OE score was a bit higher. The efficiency of good quality cells was still low, so the Dam background noise was quantified in each cell by calculating genome-wide mode and median OE scores, which were compared to the total Dam read count in that same cell. The top cells in **Fig. 17a,b** had median/mode OE baseline scores below OE = 1 and many Dam reads, on the order of the previously observed threshold of >60k Dam reads that indicated low baseline OE, however there were few cells with a low background noise baseline (2% efficiency), seen in **Fig. 17c**. The mapping

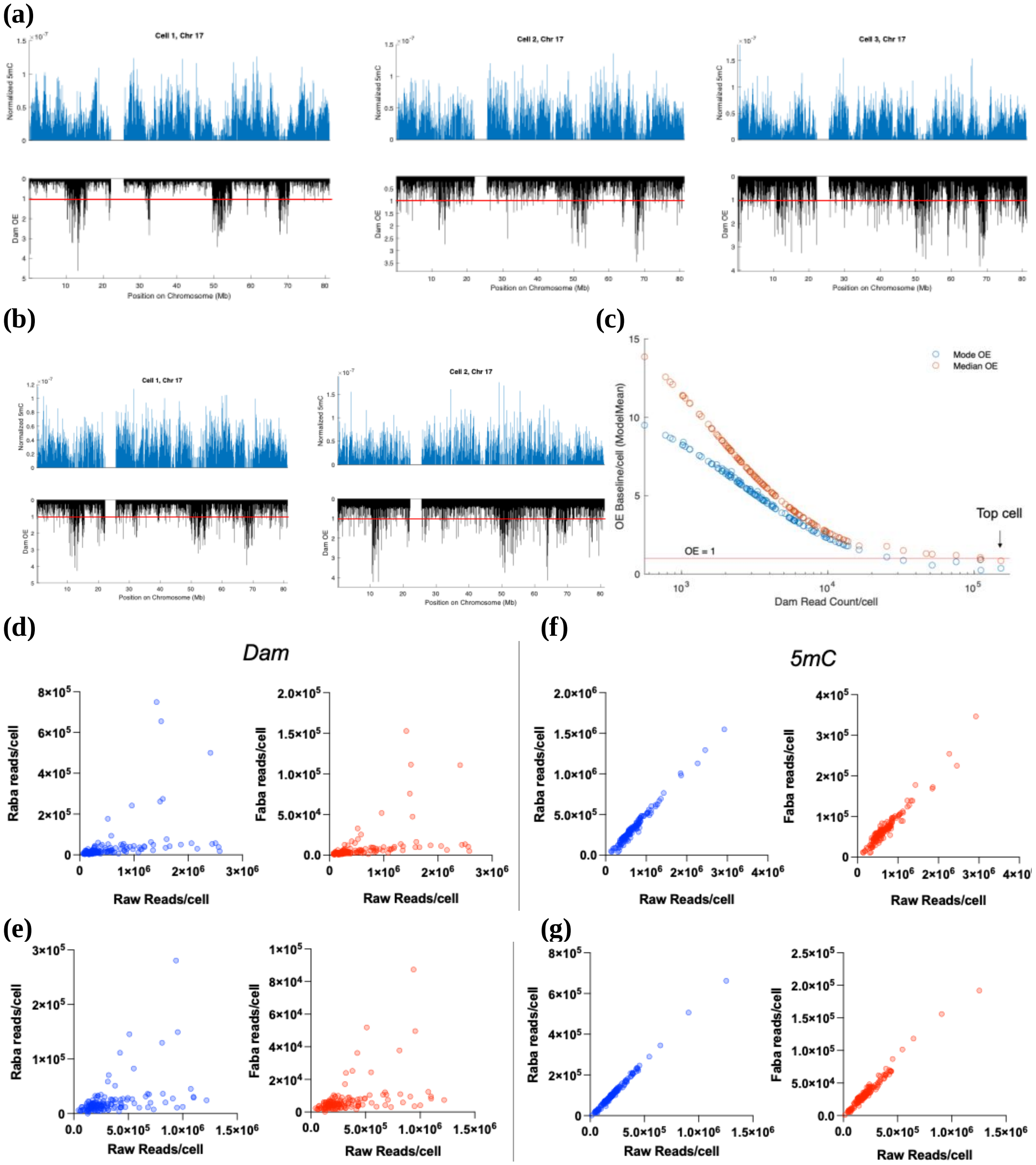


Fig. 17 | (a) Single cell 5mC+DamID profiles with forked adapters. 64nM/512 nM Forked 5mC/Dam adapters. Red line is OE = 1 **(b)** 64nM/1024 nM forked 5mC/Dam adapters. **(c)** OE baseline (mode in blue and medium in red) along the chromosomes vs. Dam read count of the same cell **(d)** Redundant (Raba) or unique (Faba) Dam reads vs. raw reads of the same cell for 64nM/512 nM Forked 5mC/Dam adapters and **(e)** 64nM/1024 nM forked 5mC/Dam **(f)** Redundant (Raba) or unique (Faba) 5mC reads vs. raw reads of the same cell for 64nM/512 nM Forked 5mC/Dam adapters and **(g)** 64nM/1024 nM forked 5mC/Dam.

efficiency and unique read counts were further measured as a function of the raw read count per cell. Typically, as sequencing depth (raw reads) of each cell increases, the redundant (PCR duplicate) and unique reads should increase proportionally. For the 5mC reads, this was the case for both adapter concentrations, producing a positive linear correlation **(Fig. 17f,g)**, however for the Dam reads, the deeper sequenced cells didn't necessarily have more redundant and unique reads, as the correlation line was flatter **(Fig. 17d,e)**. In fact, the top cells with the lowest baseline OE weren't the cells with the highest raw read count, seen by the most vertical data points on the y-axis in the middle of the plots in **Fig. 17d,e**.

To determine what further optimizations could be performed, the analyze index sort data function on the cell sorter software was used to identify the sorting parameters of the top cell from the forked 5mC adapter conditions (64nM/512 nM Forked 5mC/Dam adapters, library 1, barcode 4). By referencing the cell barcode from the sequencing data, the cell's location on the 384 well plate was found **(Fig. 18b)** and traced back to the sorting plots **(Fig. 18a)**. (The adapter plate was designed with 48 unique barcodes on the left or right half, and the barcodes were arranged sequentially on each half from left to right like words on a page, e.g. first row 1-6, then second row 7-12. Since the liquid handling robot dispenses barcodes clockwise into four adjacent wells of the 384 plate, the cell is identified by going to the right on the 384 well plate by 2x2 boxes for each barcode. Since library 1 is in the top left quadrant, and barcode 4 will shift the well to the right from A1 by blocks of 2, four times, the cell was in A7). This cell had high EGFP and mOrange signal from the FUCCI system, indicating it was at the correct stage of the cell cycle. It was possible then that other cells on the plate had noisier Dam signal because they were at the wrong stage of the cell cycle, such as at G1 phase (red), where not all the genome NL-marks were placed, or the cells passed through S phase into G2

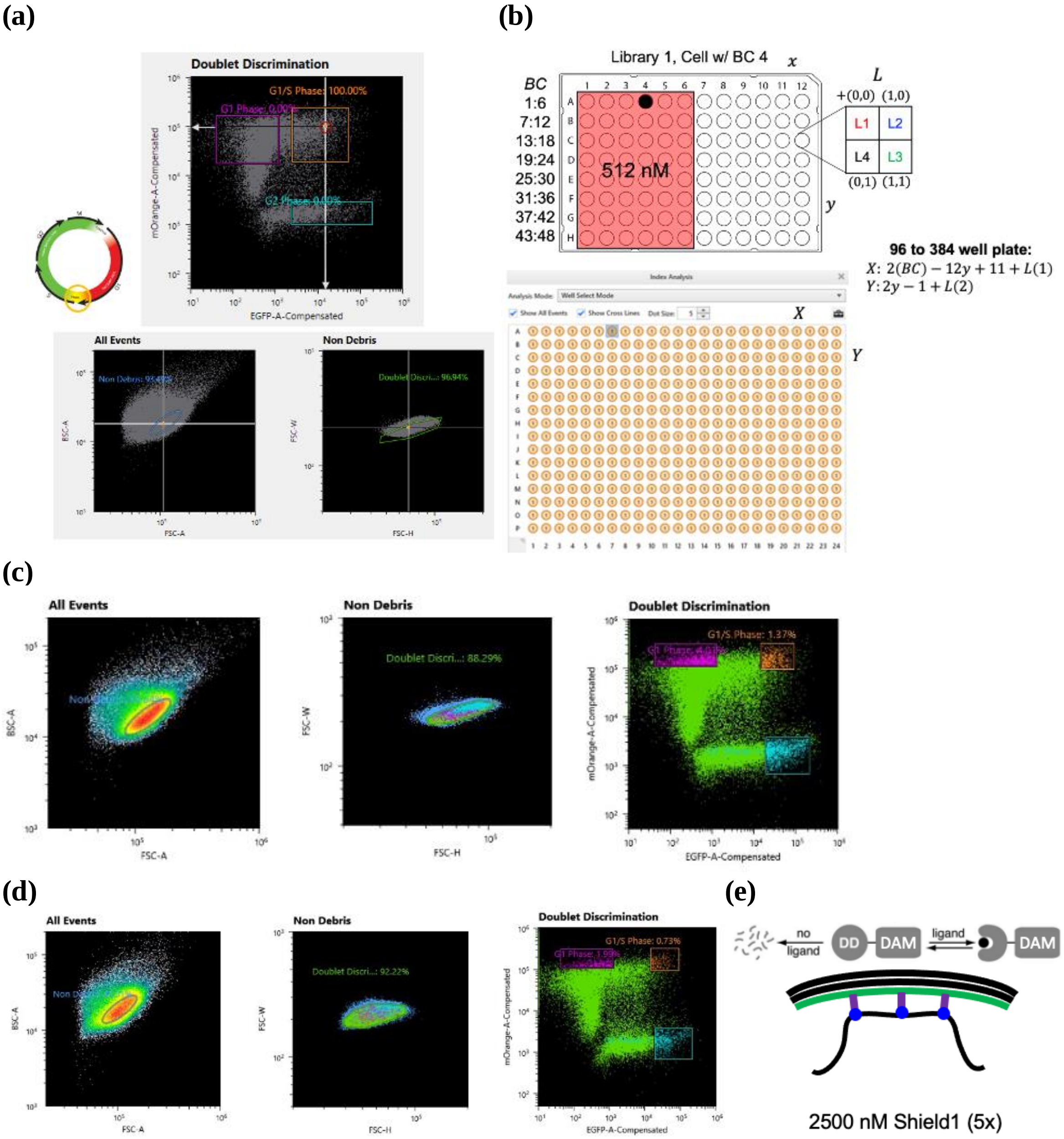


Fig. 18 | (a) Analyzing index sort data of top cell to determine why it had worked. **(b)** Top cell location in 384 well plate. **(c)** Sort the Most FUCCI ++ cells, in orange box. **(d)** Sort the most FUCCI ++ cells and also **(e)** increase [Shield1] to 5x more than the standard. Image adapted from L. A. Banaszynski et al. *Cell* (2006)

phase (green) where there will an accumulation of m^6A marks on the DNA, resulting in noise. The solution to this would be to select finer region by increasing the EGFP and mOrange threshold **(Fig. 18c**, orange box). However, some of the other cells that worked had lower mOrange signal, and one had lower EGFP signal, so it was ambivalent that sorting the highest fluorescence FUCCI++ cells would be the best solution to improve the Dam signal. Another possibility was the DD-LaminB1-Fusion that places the m^{6}A marks on the LADs wasn't being completely activated by the ligand Shield1 at 500 nM, resulting in too much background noise. Contact marks are placed only when destabilization domain of the fusion protein is induced by ligand, suggesting increasing the Shield1 ligand concentration could increase the amount of m^{6}A place at the LADs. Despite sorting for the most FUCCI ++ cells and increasing the Shield1 concentration from 500nM to 5x at 2500nM **(Fig. 18d,e)**, libraries were unable to be generated (not enough material on bioanalyzer).

B. DpnI Digestion with TET2/APOBEC Conversion

1. Random priming (v2.0)

To improve the Dam signal, a new version of the protocol was implemented using a strategy to orthogonally measure the 5mC and genome-NL contact marks, preventing interactions between the restriction enzymes as well as between the adapters. The simultaneous digestion step with DpnI and MspJI was eliminated, since MspJI derived 5mC seemed to outnumber the Dam reads. The advantage of this new protocol is now only one restriction enzyme is used, DpnI, and only one type of adapter, a modified version of the Dam adapter, to prevent competition for DNA during digestion and potential adapter interactions, while also reducing the cost previously associated with needing a set of adapters for each mark. With only one set of adapters, the adapter dispensing step is much faster, at 10 minutes/plate, rather than 20 minutes/plate. This new version of the protocol can measure every modified and unmodified cytosine at single nucleotide resolution, versus the previous method with MspJI that creates cuts 12/16 nucleotides downstream from the methylated CpG site, reducing the resolution of 5mC site detection. The *in vitro* transcription step was replaced with linear PCR, so there is no longer a degradable RNA intermediate, which would likely produce a higher yield. Nearly the same liquid handling robot protocol is used before pooling cells, so very few adjustments are required. The second part of the protocol is different, when the libraries are processed in bulk, and includes an enzymatic version of bisulfite conversion to orthogonally measure the 5mC marks. Besides the many pros to this protocol, there is only one con, which is the protocol is longer by a couple days. However, if it produces much better results, the extra time is then worth it.

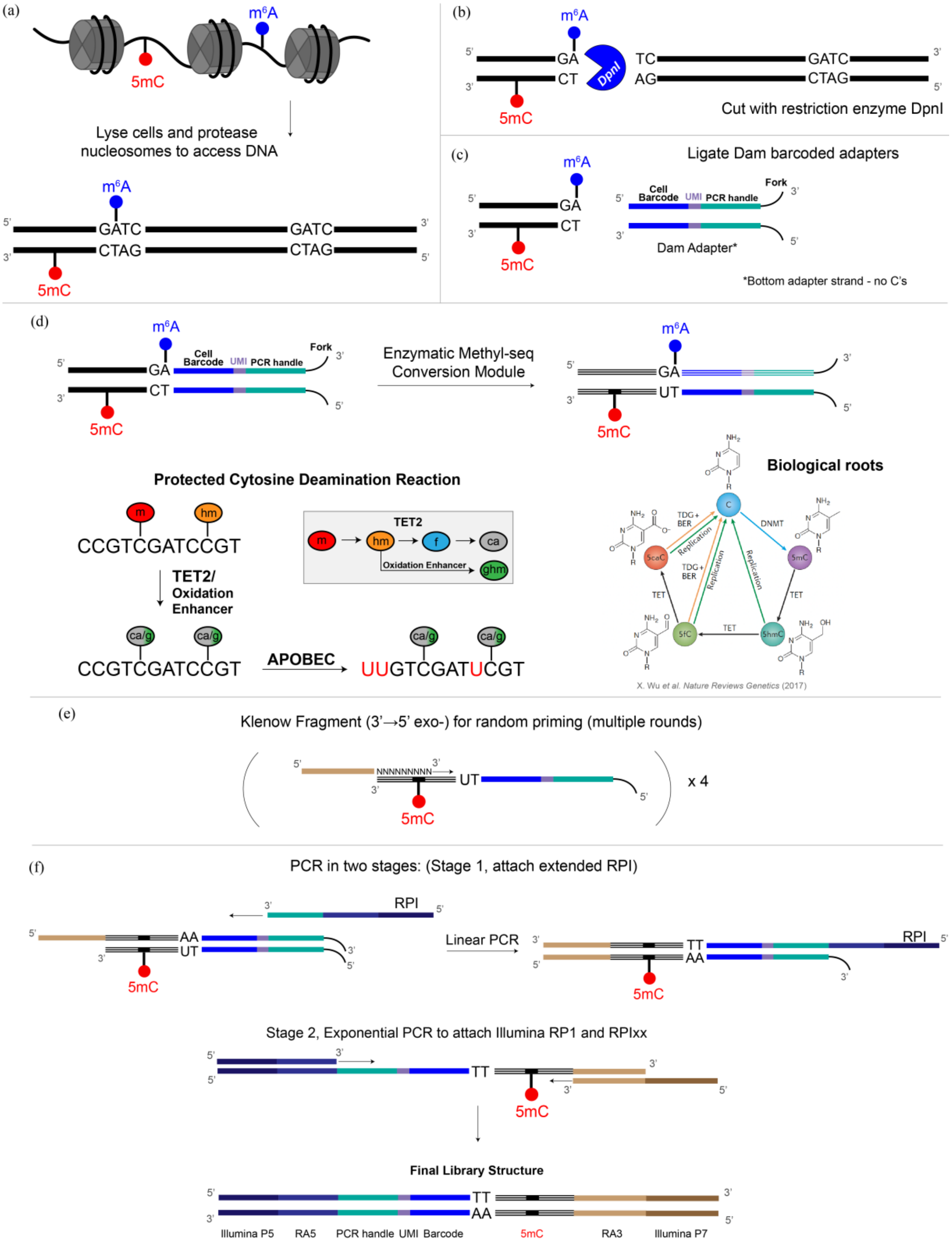


Fig. 19 | sc5mC+DamID (v2) Schematic. (a) Lyse cells and protease nucleosomes to access DNA. **(b)** Cut m^{6}A mark in GATC context with DpnI. **(c)** Ligate Dam barcoded adapters. **(d)** Enzymatic version of bisulfite conversion to identify 5mC. **(e)** Random priming to attach Illumina RA3 adapter. **(f)** PCR amplification to generate final library

The new protocol starts the same with lysing the sorted cells and then adding protease to unwind the DNA wrapped around the histones to make the 5mC and m^{6}A accessible for restriction digestion **(Fig. 19a)**. Now, only DpnI is added to DNA to create cuts at the m^{6}A genome-NL contact sites **(Fig. 19b)**. This is advantageous over the previous method because now the tougher signal is detected first, while there also no longer being a competition with MspJI which typically took all the reads. Next, a modified version of the Dam adapter is ligated to the restriction cut site **(Fig. 19c)**. This adapter contains the cell specific barcode as before to differentiate what material had originated from each cell, as well as the fork to prevent long chains of adapters forming due to blunt end ligations. Now there is a PCR handle, instead of a T7 promoter, for linearly amplifying the material and attaching one of the Illumina adapters. The bottom strand of the adapter was designed to have no cytosines, allowing bisulfite-like conversion to be performed while maintaining the sequence of the cell barcode and PCR handle. The Enzymatic Methyl-seq conversion module is used to orthogonally measure the 5mC downstream from the DpnI cut site by converting all unmodified cytosines into uracils and protecting the 5mC from the conversion, allowing us to detect the sites with and without 5mC, a limitation with original strategy that used MspJI[93]. TET2 is used to convert all the 5mC and 5hmC into 5-carboxycytosine (5caC) through the cascade reaction that occurs natively in the cell during active demethylation **(Fig. 19d,bottom-right)**. To actively remove methylation in cells, an iterative oxidation of 5mC to 5-hydroxymethylcytosine (5hmC) to 5-formylcytosine (5fC) to 5caC with TET occurs, and this is simulated through the addition of the TET2 enzyme to the reaction mixture. The conversion is further enhanced by adding an oxidation enhancer to glucosylate 5hmC to glucosylated 5mC (g5hmC) **(Fig. 19d,bottom-left)**. APOBEC, a cytosine deaminase, is added to convert all unmodified cytosines into uracils, while leaving the 5caC and g5hmC protected from the deamination. This reaction converts all of the DNA fragment (indicated by stripes in **Fig. 19d,top**), except for the 5mC sites and the bottom strand of the adapter that had no cytosines. Next, the Illumina indexing primers are attached, and the signal is amplified from picograms to micrograms, so there is a sufficient amount of material for sequencing. Random priming with Klenow fragment (3’→5’ exo-) is used to attach the RA3 of the Illumina adapter for sequencing, filling in the rest of the DNA with complementary bases **(Fig. 19e)**[94]. PCR is then performed in two stages. Linear PCR is performed with an extended version of the RPI Illumina primer, containing the sequence complementary to the PCR handle on the Dam adapter, to attach the full Illumina P5 adapter for sequencing, while simultaneously amplifying the material **(Fig. 19f,top)**. Only the bottom strand of the deaminated sequence will amplify because its sequence was unchanged by the deamination, and the fork on the strand will also not amplify based on the design of the extended RPI primer. Lastly, the standard RPI and indexing RPI[1-48] Illumina primers are used to exponentially amplify the material and attach the 6 base pair i7 indexing sequence (P7), similar to the final step of the MspJI and DpnI version of the protocol **(Fig. 19f,middle)**. This will produce the final library structure **(Fig. 19f,bottom)** containing the barcoded adapter ligated to a TT dinucleotide. It is adjacent to a TT dinucleotide because initially the Dam-LaminB1 fusion places m^{6}A in a GATC context and then DpnI will cut between the m^{6}A and the T, resulting in a TC dinucleotide on the bottom strand to which the adapter had ligated. This is then converted to TU because of the APOBEC deamination, in all scenarios because cleavage of mammalian genomic DNA by DpnI is blocked by overlapping CpG methylation, so it cannot remain in a TC context. Through random priming, the TU will be converted to its complement, AA, and following the PCR amplification, it will be converted back to TT. On the final library structure after the TT dinucleotide ligation site, is the deaminated sequence where all 5mC sites are unchanged. This more advanced method for simultaneously detecting 5mC and genome-NL contacts was implemented, and since it first detects the genome-NL sites and then the 5mC is read off those fragments at single nucleotide resolution, it may improve the Dam signal.

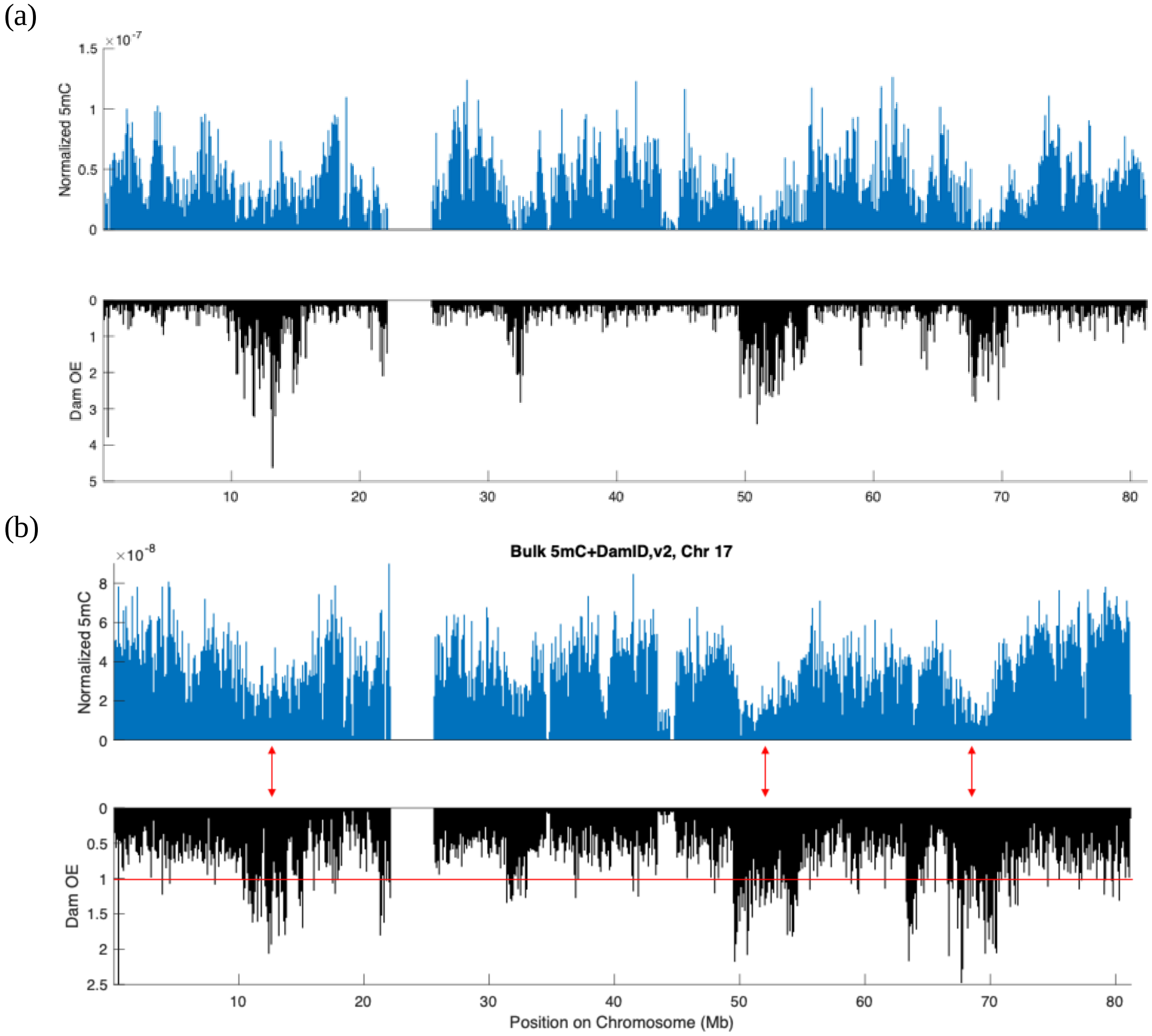


Fig. 20 | 5mC+DamID Versions 1 & 2. (a) 5mC (top) and genome-NL contact (bottom) profiles on chromosome 17 in a single cell with MspJI and DpnI, version 1. **(b)** 5mC (top) and genome-NL contact (bottom) profiles on chromosome 17 in bulk with DpnI and TET2/APOBEC, version 2. The same LADs correspond to same hypomethylated regions as version 1, shown by red arrows. Horizontal line at OE = 1.

5mC and Dam profiles from each of these versions of the method were compared, and they produced similar profiles. **Fig. 20a** shows the 5mC and Dam profiles of a single cell from the first version of the method that used DpnI and MspJI. **Fig. 20b** shows the simultaneous measurement of 5mC and Dam on the same DNA molecule for a bulk population using the new method, DpnI and TET2/APOBEC. The LADs and 5mC hypomethylation are in the same locations suggesting these methods are interchangeable for simultaneously measuring these features of the epigenome in the same cell. However, this protocol was only implemented in bulk, and single cell data is required to prove the features are truly anticorrelated.

Comparing the sequenced 5mC component of sc5mC+DamID v1 (MspJI and DpnI) to the 5mC component of bulk 5mC+DamID v2 (DpnI & TET2/APOBEC, where the 5mC is measured off cytosine deaminated Dam reads), the locations of the 5mC hypomethylation on chromosome 17 are the same between version 1 and 2 **(Fig. 21a)**, validating both methods of detecting 5mC are interchangeable. Now with version two of the protocol, both methylated and unmethylated CpG's can be measured, a limitation with the original MspJI method. On the top track of **Fig. 21b** is the CG normalized 5mC reads, producing the typical 5mC profile on Chr 17, the middle track now contains the unmethylated cytosines in the CpG context, and the bottom track is the 5mC percentage in the genomic bins, where methylated CpGs are normalized by the sum of both methylated and nonmethylated CpGs (the total CpG sites). Measuring both methylated and non-methylated CpG's with the new method is a significant advancement because now we are using half of the number of adapters and making three measurements within the same cell instead of two: 5mC and unmethylated CpGs at single nucleotide resolution and Dam signal. Comparing the Dam signal on Chr 17 between bulk

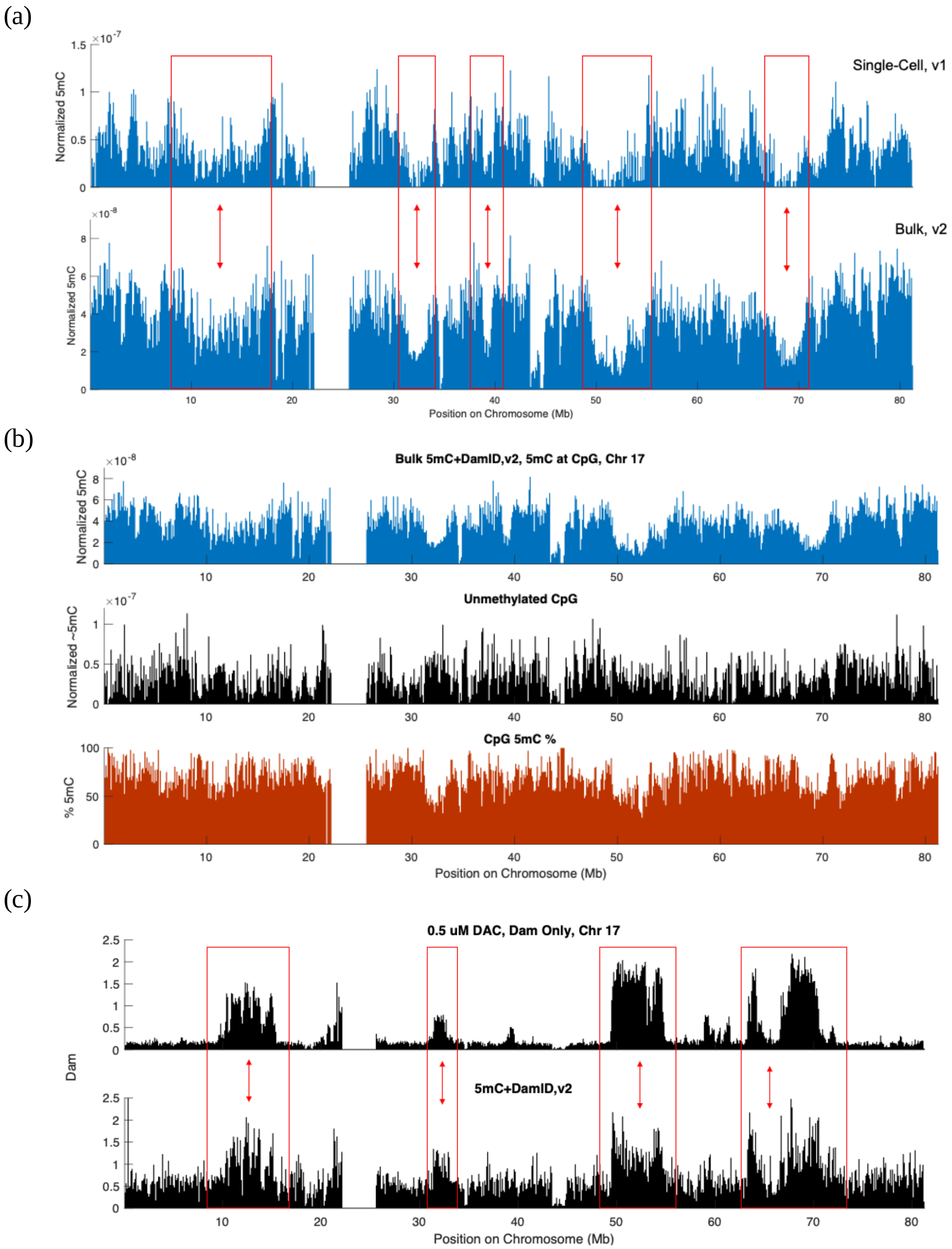


Fig. 21 | (a) 5mC profile from sc5mC+DamID v1 vs. from bulk 5mC+DamID v2 on gDNA treated with DMSO for 3 days **(b)** Bulk 5mC+DamID v2 normalized 5mC profile (top), unmethylated CpGs (middle), percent 5mC on CpGs (bottom). **(c)** Dam profile from bulk 5mC+DamID v1 vs. from bulk 5mC+DamID v2 on gDNA treated with DAC for 3 days.

DamID only (using the original Dam adapters and version one of the library preparation) to the Dam component of 5mC+DamID version two (with the redesigned Dam adapters and APOBEC deamination) on the same gDNA (treated with 0.5 µM DAC for 3 days, more information about this condition in *Chapter IV*), the same LADs appear along the chromosome **(Fig. 21c)**. The Dam signal is distinguishable from noise with a baseline (background) OE score below one, indicating it is possible to measure genome-NL contacts with this new method. Although these results imply the anti-correlation between methylation and genome nuclear lamina contacts, since the locations of hypomethylation correspond with LADs in this new method, it doesn't prove the anti-correlation because the bulk measurement obscures the correlations between epigenetic features, since it is comparing averages, despite the marks being measured on the same molecule. This new method therefore needed to be scaled down to the single-cell level to observe the anticorrelation.

Prior to performing the method with the full set of unique single-cell barcoded adapters, the protocol was implemented at the "pseudo" single-cell level (psc), in which the same barcoded-adapter was dispensed for each cell, still using the same reagent volumes as the single-cell technique. The 5' end of the adapter that ligates to the 3' end of the DpnI cleaved fragment needs to be phosphorylated for the attachment. Two variations were tested, those phosphorylated in-house using T4 PNK, which catalyzes the phosphorylation of the 5'-hydroxyl terminus of polynucleotides using inorganic phosphate derived from ATP, and others that were pre-phosphorylated[95]. Prephosphorylating the adapters generally produces better phosphorylation efficiency, but is more costly than phosphorylating in house. If the method works with the in-house phosphorylated adapters at a similar efficiency to the pre-phosphorylated adapters, it will cost significantly less to perform the single-cell experiment,

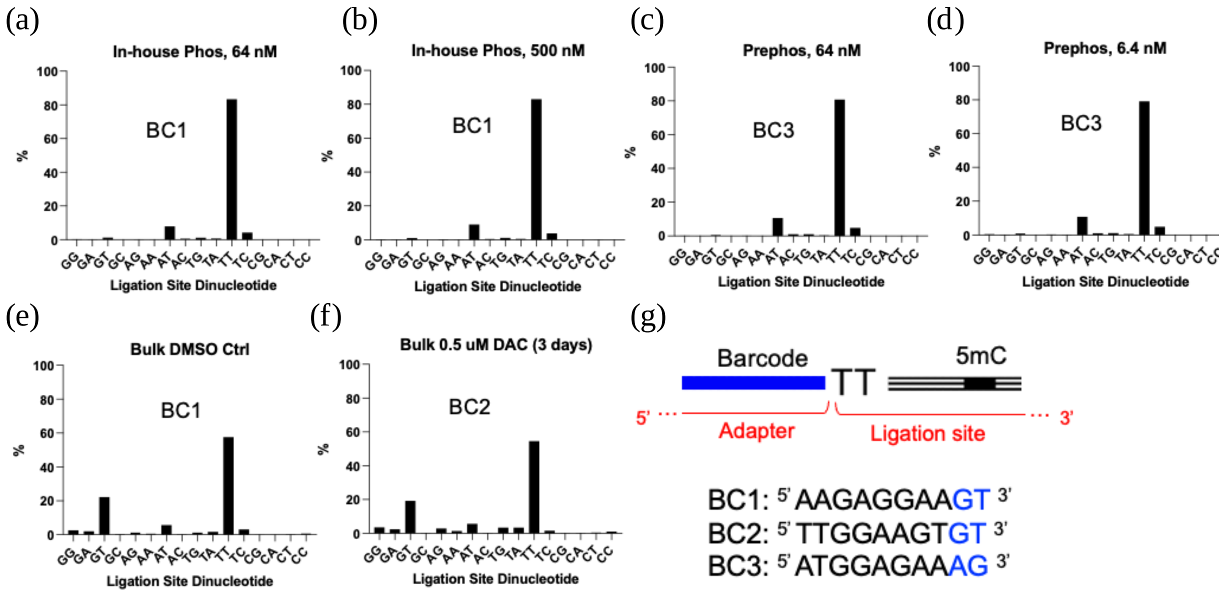


Fig. 22 | psc5mC+DamID (v2.0) adapter ligation site is primarily TT. (a) Adapter ligation site with In-house phosphorylated, 64 nM adapter **(b)** In-house phosphorylated, 500 nM adapter **(c)** Prephosphorylated, 64 nM adapter **(d)** Prephosphorylated, 6.4 nM adapter. **(a-d)** are on pseudo single cell libraries of 96 cells **(e)** Bulk DMSO treat library **(f)** Bulk 3 day 0.5 µM DAC treatment, (a-f) Plus strand reads shown only (minus strands had nearly the same percentage of ligation sites) **(g)** Anticipated binding configuration, with sequences of each barcode.

allowing for the protocol to be more reproducible. Both pre-phosphorylated and in-house phosphorylated adapters were tested at two concentrations, like how the forked adapters were tested at different concentrations, with the goal of improving the Dam signal acquisition. After sequencing the libraries, the same analysis pipeline was run on each psc5mC+DamID v2.0 library and the bulk 5mC+DamID v2.0 DMSO/DAC conditions that were the basis for this method being possible. The ligation site of the redesigned Dam adapter was measured, and in all four of the psc libraries ~80% of the ligation sites were at the dinucleotide TT **(Fig. 22a-d)**. In the bulk version of this protocol on the DMSO/DAC conditions, there was only a 55% ligation to TT **(Fig. 22e,f)**, suggesting there was an improvement when scaling it down to the pseudo single-cell level.

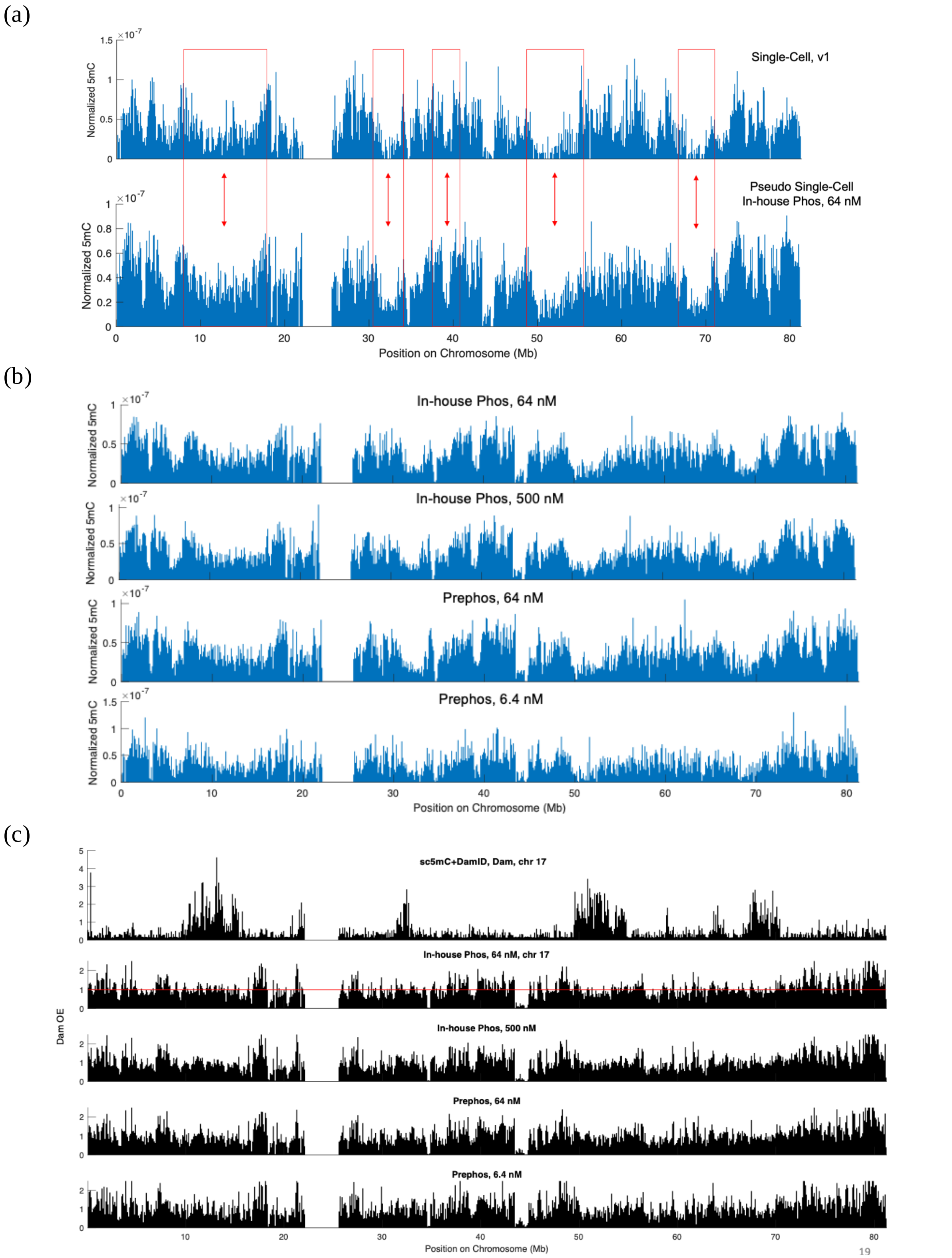


Fig. 23 | (a) 5mC profiles from sc5mC+DamIDv1 (top) and pseudo single-cell 5mC+DamID v2 (bottom) with 64 nM in-house phosphorylated adapters **(b)** 5mC profile comparisons from pseudo single-cell 5mC+DamID v2 with various phosphorylations and concentrations **(c)** Dam profiles with same order and conditions as (b).

The 5mC profile of the pseudo single-cell libraries resembled the single-cell version 1 profile. Comparing the 5mC component of sc5mC+DamID v1 to the in-house phosphorylated (64 nM) psc5mC+DamID v1 **(Fig. 23a)**, the hypomethylation along the chromosome was in the same locations, suggesting MspJI and TET2/APOBEC are interchangeable at likely the sc level for measuring 5mC. Each of the pseudo single-cell conditions, with different adapter concentrations and ways of phosphorylation, had similar 5mC profiles, supporting a robustness in the method **(Fig. 23b)**. A weaker signal was observed with the Prephos, 6.4 nM adapter condition, however this is most likely explained by it having the lowest adapter concentration. The Dam component of the pseudo single-cell libraries did not match the profile of the version 1 single-cell method. The top track of **Fig. 23c** is the expected Dam signal, a profile from sc5mC+DamID version 1, however in the next four tracks corresponding with psc5mC+DamID v2.0, the LADs were not visible. Interestingly, the bad Dam signal profile was different than the noisy profile in version 1: the baseline (background) OE score was not too high, as peaks would still be seen at the LAD regions, the issue now is the signal appears inverted. This is seen best by the LAD regions positioned at 10 Mb and 50 Mb, where a drop in Dam signal, rather than in increase in signal is observed while the surrounding iLAD regions have more signal. The rationale for this phenomenon is the random priming is biased for the iLAD fragments over the LAD fragments, thereby reducing the LAD signal **(Fig. 24a)**. The size of the LAD fragments for priming will be smaller than the iLAD fragments because there is less space between the DpnI cut sites due to the abundance of m^{6}A contact marks added by the Dam-lamin B1 fusion **(Fig. 24b)**. Given the mappable inter-GATC distance is roughly 200 bp **(Fig. 24c)** and the random primer is a 9-mer, if the primer doesn't bind towards the tail of the DNA

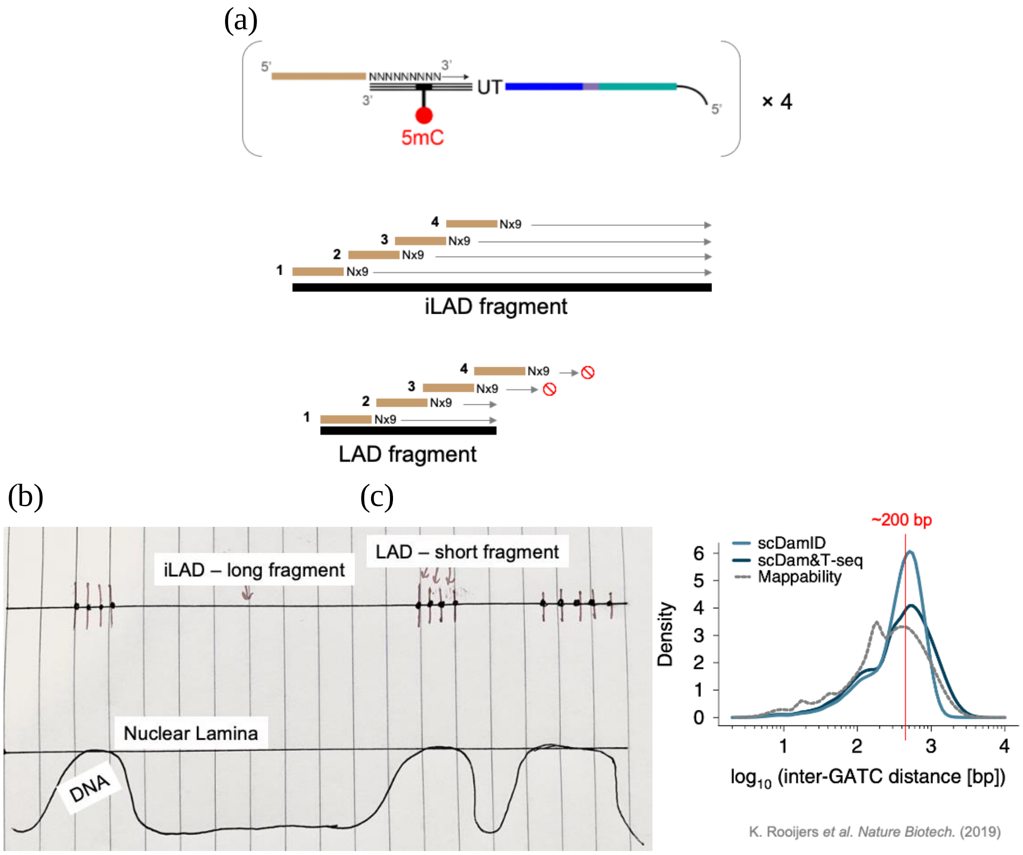


Fig. 24 | (a) Random priming schematic on both iLAD and LAD derived fragments **(b)** Cut sites along DNA for regions contacting nuclear lamina or iLADs. LADs produce shorter fragments than iLADs because there is a higher density of m^{6}A marks **(c)** Inter-GATC distance and mappability of GATC sites using the DpnI restriction enzyme and adapter ligation method, adapted from K. Rooijers *et al. Nature Biotechnology* (2019).

fragment, the complementary synthesized strand may be too short to be compatible with the rest of library preparation. It is important to note iLADs also have regions of contact with the nuclear lamina (referred to as contact segments, discussed in *Chapter III*), however they are spaced far apart resulting in long iLAD fragments. With each round of random priming, the LAD/iLAD fragments will get smaller, ultimately creating a bias for iLAD signal over LAD signal. This random priming strategy for adding the RA3 adapter therefore turns it into a game of “pin the tail on the molecule”, it that if the primer isn’t bound towards the tail region, the molecule may be lost for sequencing. The solution to this issue is less rounds of random priming or an entirely different priming strategy will reduce the priming bias for iLADs, and it will therefore improve the Dam signal.

To test this hypothesis, the protocol was repeated with only one round of random priming. The LADs were still not visible however, and produced the same inversion of iLAD and LAD signal because the LAD fragments were selected against by random priming **(Fig. 25a)**. The dinucleotide at the ligation site however remained as TT at 80%, indicating the adapters were binding to the GA^{m6}↓TC DpnI cut sites **(Fig. 25d)**. Despite the masked LAD signal in this condition that had produced a Dam profile that resembled the 5mC signal, the regions where hypomethylation occurred, coinciding with the LADs, could still be detected by subtracting the two signals. In version 2 of the method, first the Dam read is acquired and

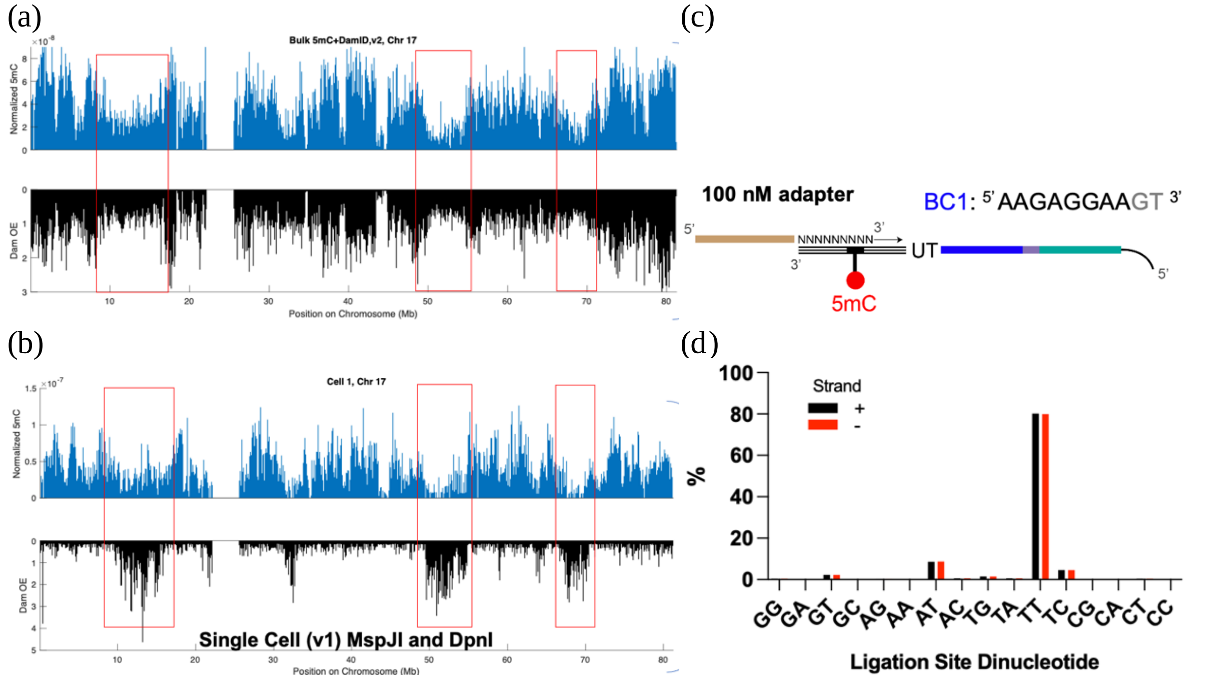


Fig. 25 | (a) 5mC and Dam profiles of psc5mC+DamID v2, with one round of random priming, 100 nM adapter, normalized by CG content and Dam OE **(b)** 5mC and Dam profiles of sc5mC+DamIDv1 as reference. **(c)** Schematic of random priming the deaminated adapter ligated ssDNA fragment. **(d)** Adapter ligation site dinucleotide on plus and minus strands.

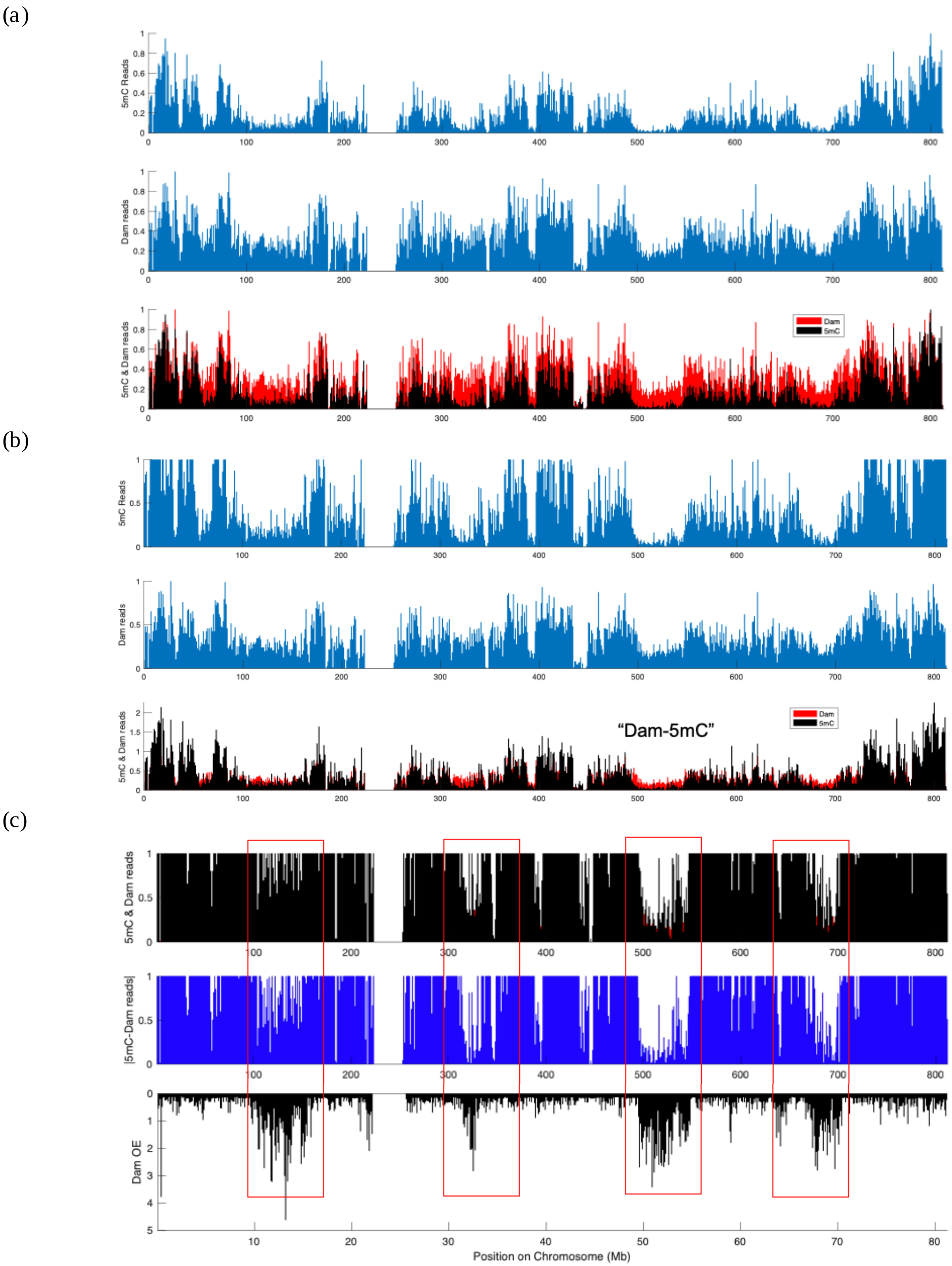


Fig. 26 | (a) (Top) 5mC and (Middle) Dam profile normalized on a scale of 0-1. (Bottom) 5mC and Dam read overlap. **(b)** (Top) 5mC normalized by a skewed scale, $\frac{5mC_{bin,i}\{chr_j\}}{Max(5mC_{bin,i}\{chr_j\}-scaling)}$, and (Middle) Dam profile normalized on a scale of 0-1. (Bottom) Skewed 5mC and Dam read overlap. **(c)** (Top) Skewed 5mC and Dam overlap at minimum scaling constant. (Middle) Absolute value difference between skewed 5mC and Dam reads. (Bottom) LAD profile then the 5mC is read off that fragment, therefore any region containing Dam signal and low 5mC is a hypomethylated region. To observe this and prove the 5mC is correctly being detected at the pseudo single-cell level, the 5mC and Dam reads were normalized on a scale of 0-1 (instead of by CG content or Dam OE), and the profiles were overlaid **(Fig. 26a)**. At regions such as around bin 500 and 700 (corresponding to 50 and 70 Mb), there were a significant amount of Dam reads (red) and few 5mC reads (black), indicating low 5mC at these regions. To better visualize the hypomethylation, the 5mC reads were "stretched" over the Dam reads until overlapping, by subtracting a scaling constant from the denominator of the normalization **(Fig. 26b)**. By scaling the normalization, the regions with low 5mC and high Dam reads are visible (seen in red), which correspond to the hypomethylated LADs, validating this method works but the Dam signal is biased towards iLADs. By applying a range of scaling constants to the 5mC normalization and measuring the absolute difference of 5mC and Dam signals, most regions quickly increase past unity to infinity **(Fig. 26c)**. However, in hypomethylated LAD regions, such as bin 500, the difference initially approaches zero before increasing to infinity much slower than in the iLAD regions. Regions where the absolute difference between 5mC and Dam reads approach infinity slower primarily correspond with the locations of the LADs, when comparing the hypomethylated regions in psc5mC+DamID v2.0 to the Dam profile of a single cell from sc5mC+DamID v1. Therefore, the regions with low 5mC reads relative to the number of Dam reads are the hypomethylated LADs, validating the v2 method can accurately measure 5mC, while suggesting there is an issue with iLAD priming bias based on the Dam signal.

2. TACS Ligation (v2.1)

A random priming alternative, based on a method called TACS (terminal deoxyribonucleotidyl transferase (TdT)-assisted adenylate connector-mediated ssDNA) ligation, was implemented to improve the Dam signal[96]. The novelty to this technique is it doesn't require a template to attach the RA3 component of the Illumina adapter, and instead uses TdT to incorporate nucleotides onto the 3'-OH termini of ssDNA, to which a single stranded ligation of the RA3 adapter occurs. To keep the P7 Illumina adapter as close as possible to the deaminated DNA fragment, a short string of nucleotides are added **(Fig. 27a)** by using ATP instead of conventional dNTPs. ATP is converted into ADP and then AMP by hydrolysis, and since AMP is effectively adenosine, using ATP instead of dNTPs will limit the amount of RNA bases added to the tail of the fragment to between 2 and 4. Since RNA bases are added, T4 RNA ligase I can be used to connect the 5' phosphoryl-terminated donor on an RA3 adapter to the 3' hydroxyl-terminated acceptor, the ssRNA tail of the fragment. By designing the RA3 as dual phosphorylated, end to end ligations of the adapter cannot occur, preventing its concentration from decreasing in the reaction, like the benefit of forking the Dam adapters. An advantage to using T4 RNA Ligase I is it incubates at a low temperature (16°C), preventing the activity of RNAses that could degrade TdT added RNA bases. PEG is additionally used in the reaction to increase the ligation efficiency. The original TACS protocol claimed that single stranded ligation to bisulfite cleaved DNA is not practical because TdT failed to modify the majority of the 3' ends, and instead they performed one round of random priming and then TACS ligation[96]. This approach wouldn't be useful for sc5mC+DamID v2 because any random priming will bias the Dam signal for iLADs and reduce the LAD signal. Although cleavage of the deoxyribose ring from bisulfite conversion

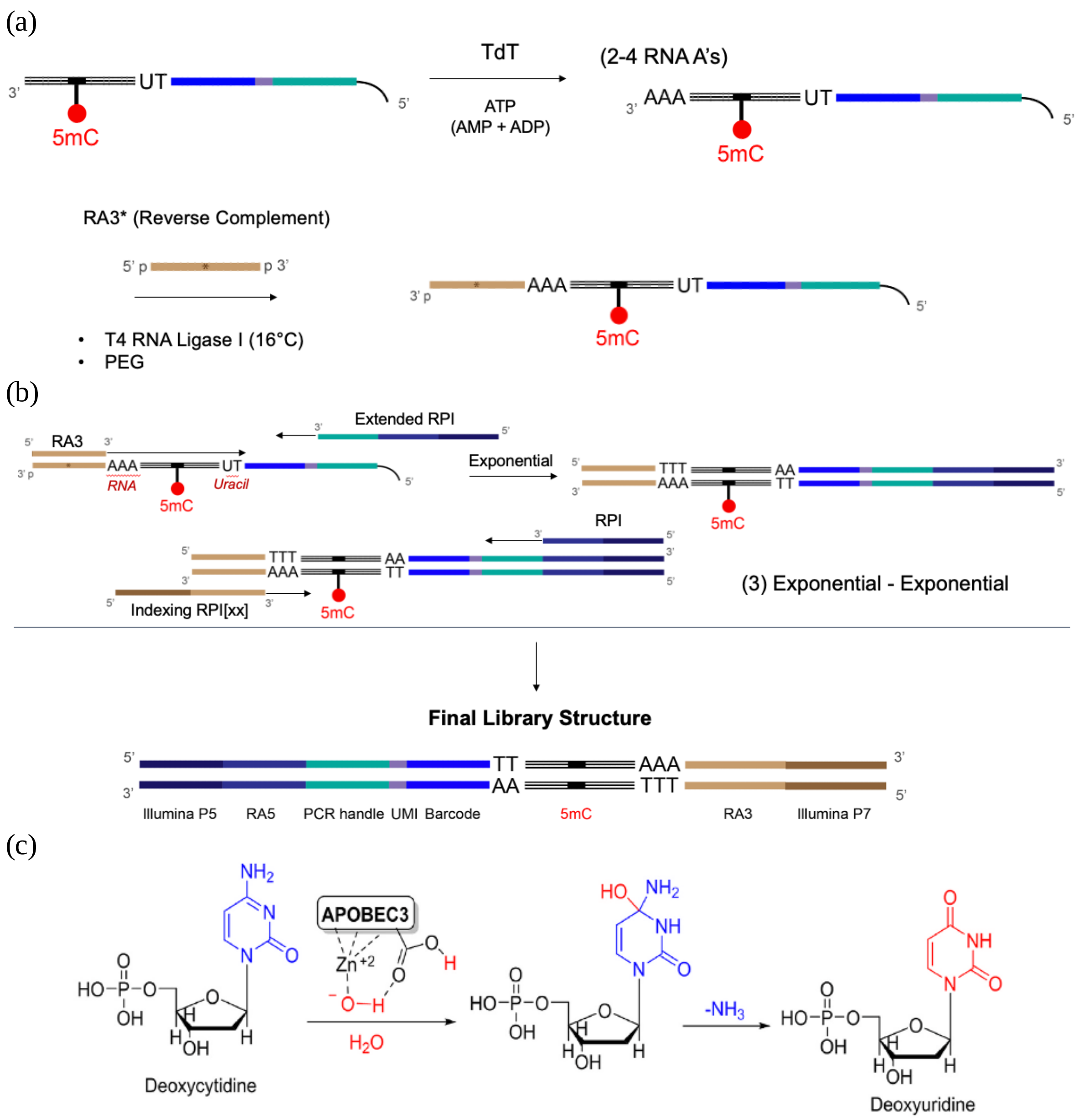
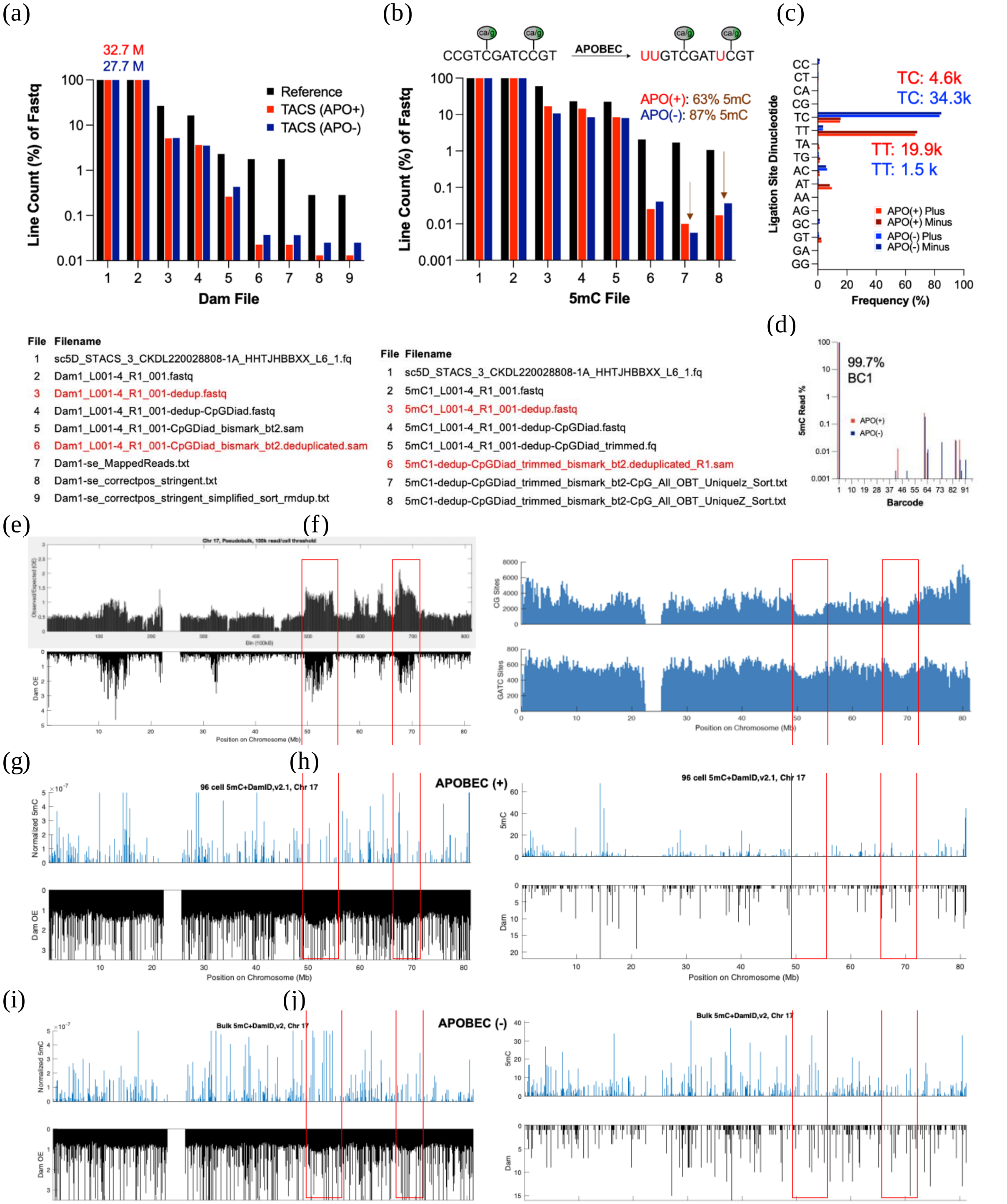


Fig. 27 | sc5mC+DamID v2.1 Random priming alternative: TACS ligation. Schematic of **(a)** A-tailing and single stranded adapter ligation and **(b)** PCR amplification to generate final library. **(c)** APOBEC3A deamination reaction mechanism, image adapted from S. Revathidevi *et al Cancer Lett.* (2021).

could make the 3' terminal structure unrecognizable by TdT, interfering with it's ability to add an A-tail, the NEBNext Enzymatic Methyl-seq conversion module is less harsh on the DNA because it is an enzymatic approach[93,96], meaning this strategy could be possible. The APOBEC3A enzyme shouldn't damage the 3' deoxyribose ring of the DNA during the cytosine deamination based on the reaction mechanism **(Fig. 27c)** since it only modifies the pyrimidine, and therefore it shouldn't prevent TdT activity[97]. After ligation, the single stranded fragment is then amplified by a series of exponential PCR steps, however there are two limitations at this stage. The first issue is there are RNA bases (**Fig. 27b**) that must be recognized by a DNA polymerase, and the second issue is there are uracils (non-typical DNA bases) that are in the fragment needing to be amplified. The first issue is addressed by only adding only a few RNA bases to the tail through the ATP to AMP decomposition, which should not stall a robust DNA polymerase. The second issue is addressed by using a robust uracil tolerant DNA polymerase, Q5U, which is optimized for uracil-containing NGS library amplification and is compatible with NEBNext Enzymatic Methyl-seq kit. The first PCR is performed with Q5U, using an RA3 primer and the extended RPI primer (from v2.0) to produce a DNA molecule absent of uracil and RNA bases, while attaching the full Illumina P5 adapter (**Fig. 27c**). The second PCR uses the same RPI and indexing RPI[xx] primers (from v1.0) and the standard DNA polymerase, NEB Next High Fidelity, now that there are no uracils, to generate the final library seen in **Fig. 27c**, containing a series of A's between the deaminated fragment and RA3. This protocol generated pseudo single-cell libraries which were further sequenced and analyzed for the method efficiency.

Several levels of checks suggest TACS ligation could work at the single-cell level, however it may be low efficiency. The line count/sequencing depth of the 5mC and Dam output files was measured for the psc5mC+DamID v2.1 APOBEC(+), APOBEC(-), and the bulk 5mC+DamID v2.0 DAC reference conditions. Many reads were lost at the deduplication stages (e.g. *dedup.fastq and *deduplicated.sam) compared to the reference **(Fig. 28a,b)**, suggesting a low efficiency of TACS ligation, with many reads being PCR duplicates.

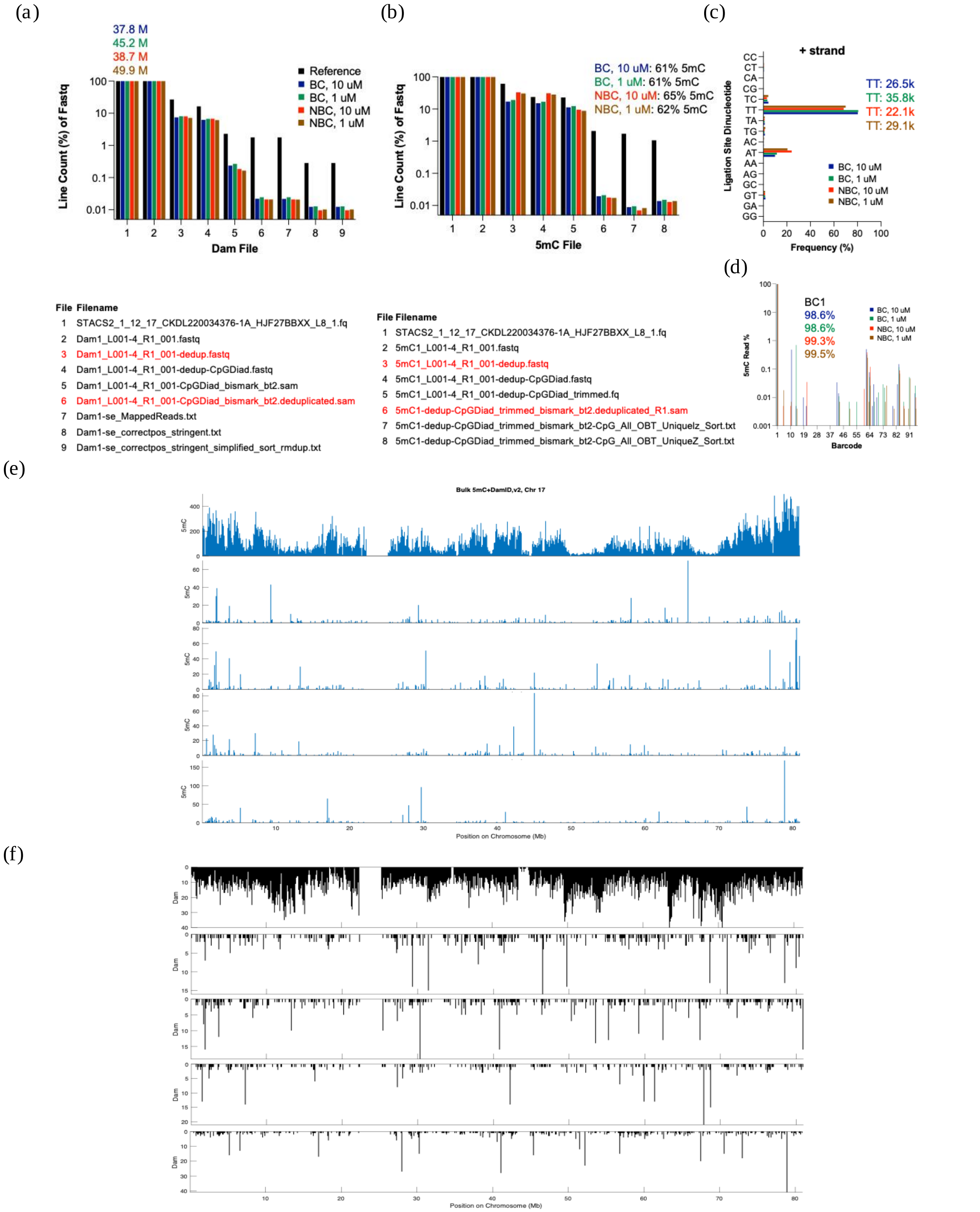


File	Filename
1	sc5D_STACS_3_CKDL220028808-1A_HHTJHBBXX_L6_1.fq
2	Dam1_L001-4_R1_001.fastq
3	Dam1_L001-4_R1_001-dedup.fastq
4	Dam1_L001-4_R1_001-dedup-CpGDiad.fastq
5	Dam1_L001-4_R1_001-CpGDiad_bismark_bt2.sam
6	Dam1_L001-4_R1_001-CpGDiad_bismark_bt2.deduplicated.sam
7	Dam1-se_MappedReads.txt
8	Dam1-se_correctpos_stringent.txt
9	Dam1-se_correctpos_stringent_simplified_sort_rmdup.txt

File	Filename
1	sc5D_STACS_3_CKDL220028808-1A_HHTJHBBXX_L6_1.fq
2	5mC1_L001-4_R1_001.fastq
3	5mC1_L001-4_R1_001-dedup.fastq
4	5mC1_L001-4_R1_001-dedup-CpGDiad.fastq
5	5mC1_L001-4_R1_001-dedup-CpGDiad_trimmed.fq
6	5mC1-dedup-CpGDiad_trimmed_bismark_bt2.deduplicated_R1.sam
7	5mC1-dedup-CpGDiad_trimmed_bismark_bt2-CpG_All_OBT_Uniquelz_Sort.txt
8	5mC1-dedup-CpGDiad_trimmed_bismark_bt2-CpG_All_OBT_UniqueZ_Sort.txt

Fig. 28 | psc5mC+DamID v2.1 (TACS) library statistics and profiles. (a) Line count of Dam and **(b)** 5mC files, normalized to raw read count. **(c)** Ligation site dinucleotide frequency **(d)** Barcode on 5mC reads **(e)** (Top) Pseudobulk DamID and (Bottom) Dam component of sc5mC+Dam v1 **(f)** CG (top) and GATC (bottom) site profiles on chromosome 17 **(g)** TACS (with APOBEC) 5mC and Dam profiles (normalized) and **(h)** nonnormalized **(i)** TACS (without APOBEC) 5mC and Dam profiles (normalized) and **(j)** nonnormalized.

Despite this, the APOBEC conversion was detectable, seen through the line count of the 5mC output files, file 7 (the non-methylated CpG reads) and file 8 (the methylated CpG reads). The APOBEC(-) condition had more 5mC reads than the APOBEC(+) condition, and the APOBEC(+) condition had more non-methylated CpG reads than the APOBEC(-) condition **(Fig. 28a,b)**. This was what was expected if the APOBEC conversion were working correctly, since the code calls a site as 5mC if it were protected from the deamination, and without APOBEC there wouldn't be conversion, which would be seen as more 5mC. The APOBEC(+) condition had 63% 5mC and APOBEC(-) had 87% 5mC, compared to the psc5mC+DamID v2.0 one round of random priming condition with 68% 5mC. To further validate the method was working, the adapter ligation site was measured; in the APOBEC(+) condition, the correct ligation dinucleotide was detected (>67% TT) and likewise in the non-deaminated APOBEC(-) condition, the dinucleotide was mostly (84%) TC **(Fig. 28c)**. However, there were very few reads in the final Dam output file, since it started with 32.7 M raw reads and ended with 19.9 k TT reads. For sc5mC+DamID v1 with DpnI and MspJI, typically 60k Dam reads/cell produced high signal to noise with detectable LADs, however this TACS condition had 96 cells with under 20k reads total. This means there would have to be 289x more signal in the APOBEC(+) condition for LADs to be detectable, assuming a similar trend. A further check of this method indicated nearly all (99.7%) of the 5mC reads, read off the DpnI cut fragments, had adapter barcode 1 **(Fig. 28d)**, the adapter that had been used for this library preparation. The low reads counts are evident in the 5mC and Dam profiles along chromosome 17 **(Fig. 28h,j)**. With APOBEC, there were low reads and little signal, appearing as scattered noise. When normalizing the Dam signal by OE, shallow peaks appear at LAD locations 50 and 70 Mb, referencing the Dam component of sc5mC+DamID



File	Filename
1	STACS2_1_12_17_CKDL220034376-1A_HJF27BBXX_L8_1.fq
2	Dam1_L001-4_R1_001.fastq
3	Dam1_L001-4_R1_001-dedup.fastq
4	Dam1_L001-4_R1_001-dedup-CpGDiad.fastq
5	Dam1_L001-4_R1_001-CpGDiad_bismark_bt2.sam
6	Dam1_L001-4_R1_001-CpGDiad_bismark_bt2.deduplicated.sam
7	Dam1-se_MappedReads.txt
8	Dam1-se_correctpos_stringent.txt
9	Dam1-se_correctpos_stringent_simplified_sort_rmdup.txt

File	Filename
1	STACS2_1_12_17_CKDL220034376-1A_HJF27BBXX_L8_1.fq
2	5mC1_L001-4_R1_001.fastq
3	5mC1_L001-4_R1_001-dedup.fastq
4	5mC1_L001-4_R1_001-dedup-CpGDiad.fastq
5	5mC1_L001-4_R1_001-dedup-CpGDiad_trimmed.fq
6	5mC1-dedup-CpGDiad_trimmed_bismark_bt2.deduplicated_R1.sam
7	5mC1-dedup-CpGDiad_trimmed_bismark_bt2-CpG_All_OBT_Uniquelz_Sort.txt
8	5mC1-dedup-CpGDiad_trimmed_bismark_bt2-CpG_All_OBT_UniqueZ_Sort.txt

Fig. 29 | psc5mC+DamID v2.1 (TACS), experiment two, library statistics and profiles. (a) Line count of Dam & **(b)** 5mC files, normalized to raw read count. **(c)** Ligation site dinucleotide frequency **(d)** Barcode on 5mC reads **(e)** 5mC profile from (Top) Bulk 5mC+DamIDv2, (Next four) Bead clean up (BC) with 10 µM dual phosphorylated RA3 adapter, BC-1 µM, No BC-10 µM, No BC-1 µM **(f)** Dam profiles from same conditions & order as (e).

v1 **(Fig. 28g)**. However, this is likely due to less GATC sites at those locations **(Fig. 28f)**, effectively raising the peak because of the normalization. Similar results were observed in the condition prepared without APOBEC **(Fig. 28i,j)**. These results suggest that TACS may be possible at the single-cell level, but the signal is too low with the current implementation, requiring further optimization to improve the signal.

The experiment was repeated with two optimizations, (1) with or without a bead clean up between the TdT and ligation, and (2) decreasing the concentration of the dual phosphorylated RA3 adapter from 100 μM to 10 μM and 1 μM. A similar analysis was performed on the data, and again many reads were lost during the deduplication **(Fig. 29a,b)** stages, suggesting the presence of PCR duplicates. Despite this, the expected 5mC percentage was measured in all conditions (61-65%), compared to the 68% 5mC observed in the psc5mC+DamID v2.0 one round of random priming condition. Eliminating the bead cleanup step lowered the TT dinucleotide ligation site to nearly 70%, compared to the condition with the bead cleanup that had the correct adapter ligation site at greater than 80% **(Fig. 29c)**. By decreasing the dual phosphorylated RA3 adapter concentration to 1 μM, there were slightly more Dam reads than in the 100 μM condition from the previous TACS experiment, however the reads were still too low (35.8k for 96 cells) compared to the typical 60k Dam reads/cell that is associated with good Dam signal. Despite the low read count, the 5mC reads again had the barcode 1 sequence at 99%, confirming the correct final sequence structure in the libraries **(Fig. 29d)**. The 5mC and Dam signal on chromosome 17 however were still too low and scarce **(Fig. 29e,f)**. For the 5mC, the top track is the reference and the bottom four have either the bead cleanup or no bead cleanup, and difference RA3 adapter concentrations. Low read counts are observed for the 5mC, suggesting the optimizations didn’t improve the signal. The Dam signal is low compared to the reference **(Fig. 29f)**, with few reads observed and the lack of well-defined LAD peaks. These results suggest that although TACS ligation seemed like a good strategy to eliminate the random priming step, the efficiency was too low with our implementation. The frequent PCR duplicates suggest an issue with the single stranded ligation and likely only a fraction of the molecules gained the RA3 adapter that permits PCR amplification.

3. A-Tailing and Poly-T Priming (v2.2)

An alternative strategy was implemented to eliminate the random priming, in which a tail was added to the deaminated fragment, like with TACS, however now the tail is primed with a complementary sequencing containing the RA3 adapter, rather than ligating it directly to the tail. Although any nucleotide could be used for the tail, adenine was chosen again. It would seem best to add a tail of cytosines and prime with an adapter containing a complementary string of guanines because (1) after the deamination all non-methylated cytosines would be converted to uracils, so the primer would be very specific to the tail, given the deficiency of cytosines on the DNA, and (2), the C/G bond is stronger than the A/T bond because of an additional hydrogen bond in the pairing[98]. However, oligonucleotides with four or more consecutive G bases can form strong secondary structures known as guanine tetraplexes, which could interfere with the priming[99]. Therefore, a tail of A’s was added, however DNA bases (dATP) were used, and the length was controlled by varying the concentration of dATP in the reaction, and varying the TdT incubation time **(Fig. 30a)**. The TdT enzyme has a higher initial activation energy for the first nucleotide addition, whereas subsequent nucleotides are added more readily, so 15- and 30-minute incubations were

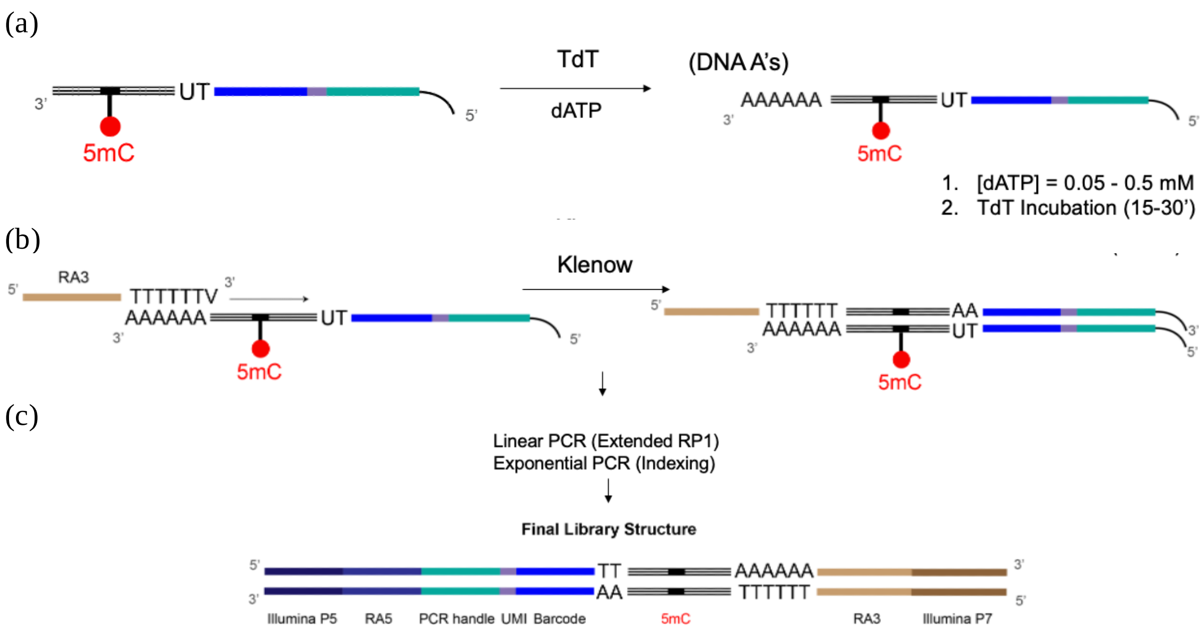
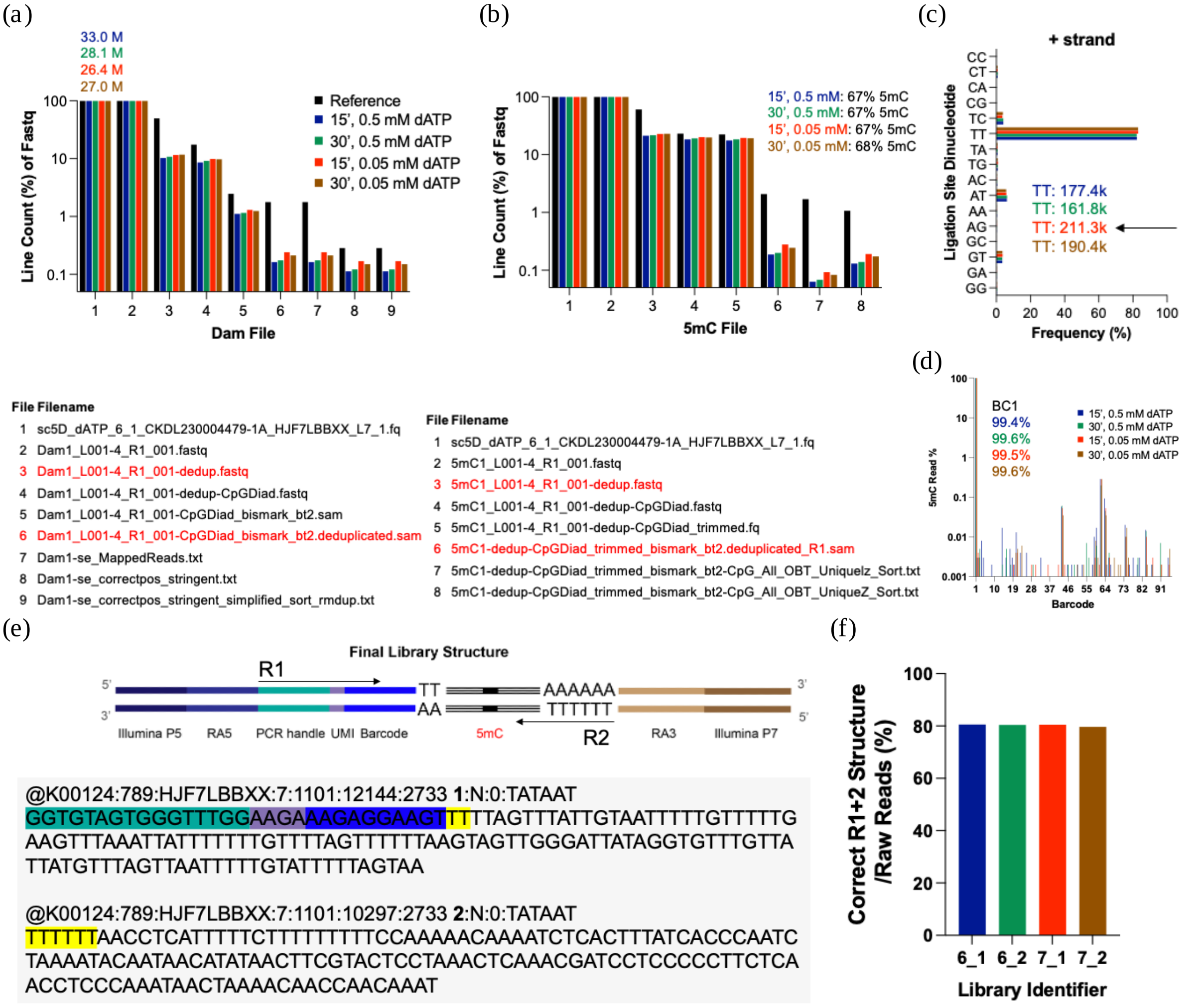


Fig. 30 | sc5mC+DamID (v2.2) Random priming alternative: A-Tailing and Poly-T priming. Schematic **(a)** A-tailing with TdT and dATP then **(b)** synthesizing first strand with RA3(TTTTTT)V primer using Klenow and **(c)** PCR in two stages to generate final library.

chosen to provide enough time for the first addition[100]. The complementary adapter to the A tail was designed as 5'-(RA3)TTTTTT(V)-3', since the "V" nucleotide (guanine/cytosine/adenine) can pair with any nucleotide except A, forcing the adapter to bind to the first six A's on the tail. A bead cleanup was performed after TdT because this increased the TT ligation site dinucleotide frequency **(Fig. 29c)** and decreased low base pair peaks on the bioanalyzer of the final library (data not shown) in the second TACS experiment. Klenow was used to generate the complementary strand downstream the primer because it can interpret uracils **(Fig. 30b),** confirmed by its use in random priming in 5mC+DamID v2.0. After synthesizing the complementary strand, a similar protocol to v2.0 was implemented, where the Illumina P5 adapter is attached with the extended RPI primer via linear PCR, and then exponential PCR is performed with RPI and indexing RPI[xx] to generate the final library **(Fig. 30c)**. The final structure contains a string of six A's between the DNA fragment and RA3 sequence, due to the primer having the TTTTTT(V) sequence.

The psc5mC+DamID v2.2 library preparation was successful, and these four conditions were sequenced. Similar analysis to TACS was performed, and many reads were lost at the deduplication steps **(Fig. 31a,b)** suggesting the presence of PCR duplicates, however the final line count was much closer to the reference than with TACS. The expected 5mC



File	Filename
1	sc5D_dATP_6_1_CKDL230004479-1A_HJF7LBBXX_L7_1.fq
2	Dam1_L001-4_R1_001.fastq
3	Dam1_L001-4_R1_001-dedup.fastq
4	Dam1_L001-4_R1_001-dedup-CpGDiad.fastq
5	Dam1_L001-4_R1_001-CpGDiad_bismark_bt2.sam
6	Dam1_L001-4_R1_001-CpGDiad_bismark_bt2.deduplicated.sam
7	Dam1-se_MappedReads.txt
8	Dam1-se_correctpos_stringent.txt
9	Dam1-se_correctpos_stringent_simplified_sort_rmdup.txt

File	Filename
1	sc5D_dATP_6_1_CKDL230004479-1A_HJF7LBBXX_L7_1.fq
2	5mC1_L001-4_R1_001.fastq
3	5mC1_L001-4_R1_001-dedup.fastq
4	5mC1_L001-4_R1_001-dedup-CpGDiad.fastq
5	5mC1_L001-4_R1_001-dedup-CpGDiad_trimmed.fq
6	5mC1-dedup-CpGDiad_trimmed_bismark_bt2.deduplicated_R1.sam
7	5mC1-dedup-CpGDiad_trimmed_bismark_bt2-CpG_All_OBT_Uniquelz_Sort.txt
8	5mC1-dedup-CpGDiad_trimmed_bismark_bt2-CpG_All_OBT_UniqueZ_Sort.txt

```
@K00124:789:HJF7LBBXX:7:1101:12144:2733 1:N:0:TATAAT
GGTGTAGTGGGTTTGGAAGAAAGAGGAAGTTTTTAGTTTATTGTAATTTTTGTTTTTG
AAGTTTAAATTATTTTTTTGTTTTAGTTTTTTAAGTAGTTGGGATTATAGGTGTTTGTTA
TTATGTTTAGTTAATTTTTGTATTTTTAGTAA

@K00124:789:HJF7LBBXX:7:1101:10297:2733 2:N:0:TATAAT
TTTTTTAACCTCATTTTTCTTTTTTTTTCCAAAAACAAAATCTCACTTTATCACCCAATC
TAAAATACAATAACATATAACTTCGTACTCCTAAACTCAAACGATCCTCCCCCTTCTCA
ACCTCCCAAATAACTAAAACAACCAACAAAT
```

Fig. 31 | psc5mC+DamID v2.2 (A-Tailing and Poly-T priming) library statistics. (a) Line count of Dam and **(b)** 5mC files, normalized to raw read count. **(c)** Ligation site dinucleotide frequency, TT count is sum of +/- strands. **(d)** Barcode on 5mC reads. **(e)** Schematic of final library structure in the raw sequence files, each block corresponding to read 1 then read 2 of the same read (1st lines are sequence identifier with read # bolded, 2nd lines are the DNA sequence). **(f)** Percentage of paired reads with the correct final library structure.

percentage was observed in all conditions (67-68%), which was the closest so far to the 68% 5mC detected in the psc5mC+DamID v2.0 one round of random priming condition, suggesting this strategy was an improvement over TACS. The 15-minute TdT incubation condition with the lower dATP concentration (0.05 mM) produced the highest read efficiency for both 5mC and Dam, seen by the red bar in **Fig. 31a,b** for Dam file 9 and 5mC files 7 & 8. In this top condition, 83% of the ligation sites had the TT dinucleotide and there were 211.3 k TT bound sites with an input of 26.4 M raw reads, an order of magnitude more than with TACS **(Fig. 31c)**. Nearly all (99.5%) of the 5mC reads had the right barcode, further supporting the method was working correctly **(Fig. 31d)**. As additional validation of psc5mC+DamID v2.2, the raw sequence files (.fastq) were checked for the correct final library structure on Read 1 and Read 2. My Perl script goes through Read 1 line by line, and if the PCR handle and correct barcode are detected, the line is flagged as correct. Then it goes through Read 2, and for all the flagged lines, if there is a string of six T's (originating from the correct binding of the (RA3)TTTTTT(V) primer), the whole read is marked as correct **(Fig. 31e)**. 80% of the raw reads had the correct sequence structure **(Fig. 31f)**, supporting the novel priming strategy works.

Similar 5mC profiles were observed in all four of the psc5mC+DamID v2.2 (dATP) libraries compared to the reference condition **(Fig. 32a)**. This was the first time detecting 5mC signal at the pseudo single-cell level without the random priming method, by using the novel A-tailing and polyT priming approach. Although there was much more Dam signal in the four dATP libraries than in the TACS libraries, the profile on chromosome 17 was incorrect when compared to the reference **(Fig. 32b)**. Comparing the reference and the top dATP condition (15 minute incubation and 0.05 mM dATP), there is a loss of 5mC,

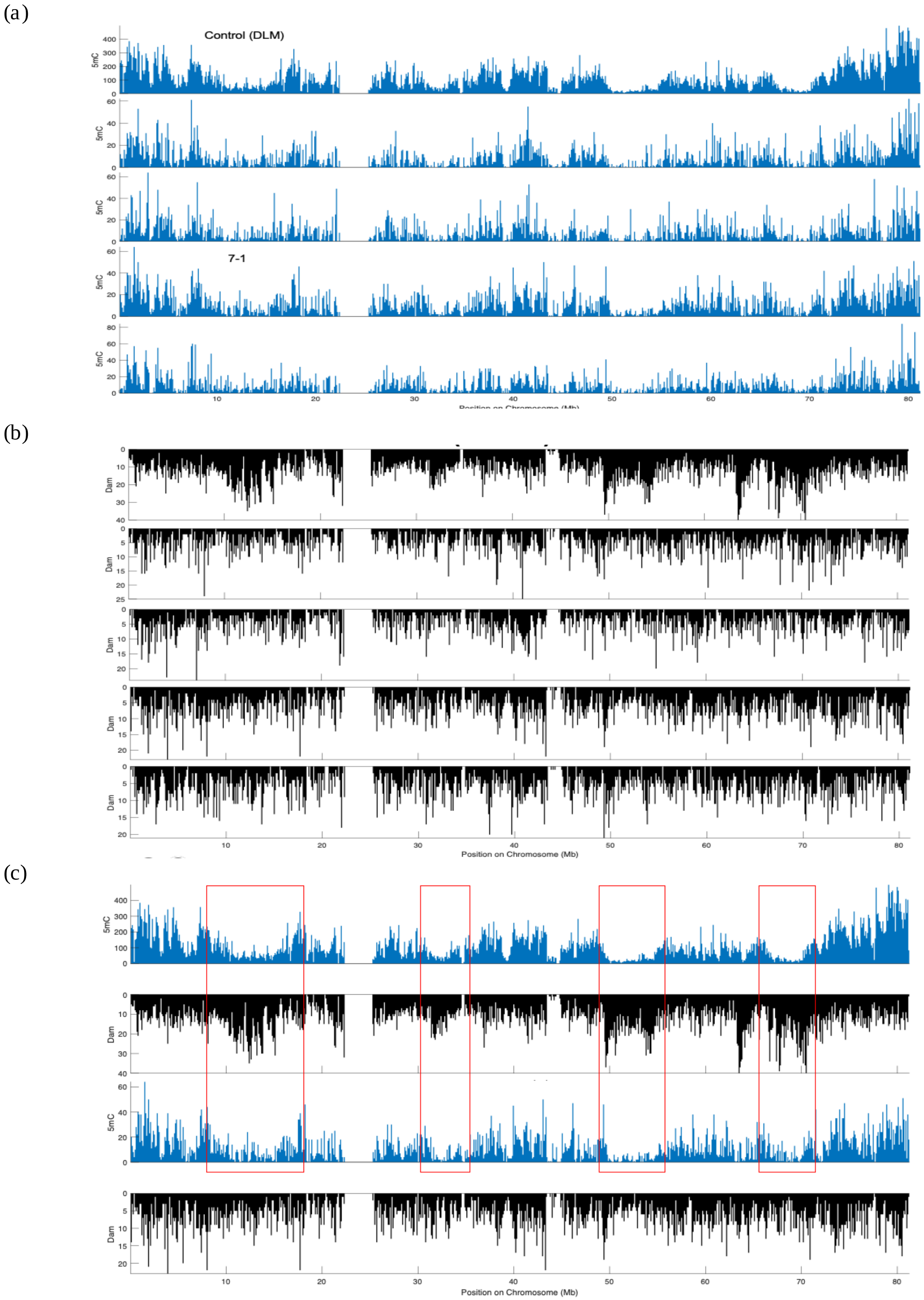


Fig. 32 | psc5mC+DamID v2.2 (A-Tailing and Poly-T priming) libraries vs. references. (a) 5mC profile from (Top) Bulk 5mC+DamID v2.0 (Next four) 15’ TdT with 0.5 mM dATP, 30’ with 0.5 mM dATP, 15’ with 0.05 mM dATP, 30’ with 0.05 mM dATP **(b)** Dam profiles from same conditions and order as (a). **(c)** 5mC and Dam profiles from (Top) bulk 5mC+DamID v2.0 and (Bottom) psc5mC+DamID v2.2, 15’ TdT with 0.05 mM dATP.

hypomethylation, at the LADs **(Fig. 32c)**, however the LAD signal is wrong. Given the success of this method in detecting 5mC, which is read off the DpnI cut fragments, further optimizations were made to improve the Dam signal.

To optimize this (RA3)TTTTTT(V) priming strategy, a three variable, two level factorial design was implemented. The first variable was the TdT concentration, which was increased based on the hypothesis that more enzyme would result in more tails added to the fragments, producing an increase in efficiency. The second optimization was both increasing and decreasing the concentration of the (RA3)TTTTTT(V) primer. Since decreasing the dual phosphorylated RA3 adapter from 100 μM to 10 & 1 uM improved the TACS reads, the (RA3)TTTTTT(V) primer was likewise decreased. Lastly, the polyT priming was increased from one to two rounds, since additional priming would likely increase the amount of material gaining the RA3 adapter, and it was previously observed that read complexity increases with up to four rounds of priming at the single-cell level (data not shown). For all of these conditions, 15 minute TdT incubation was used and 0.05 mM dATP, based on the parameters of the top psc5mC+DamID v2.2 condition. This experiment was performed on

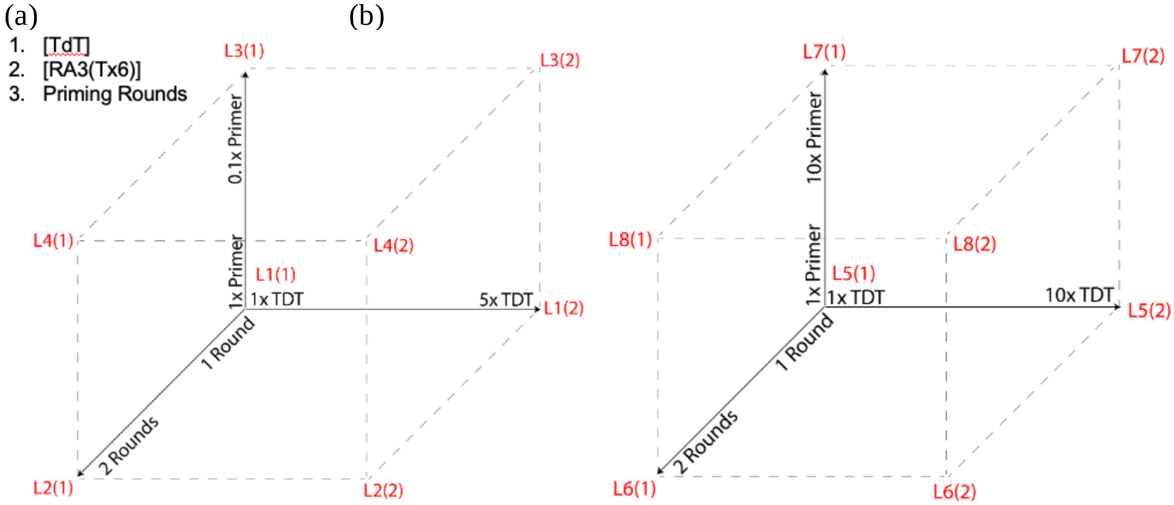
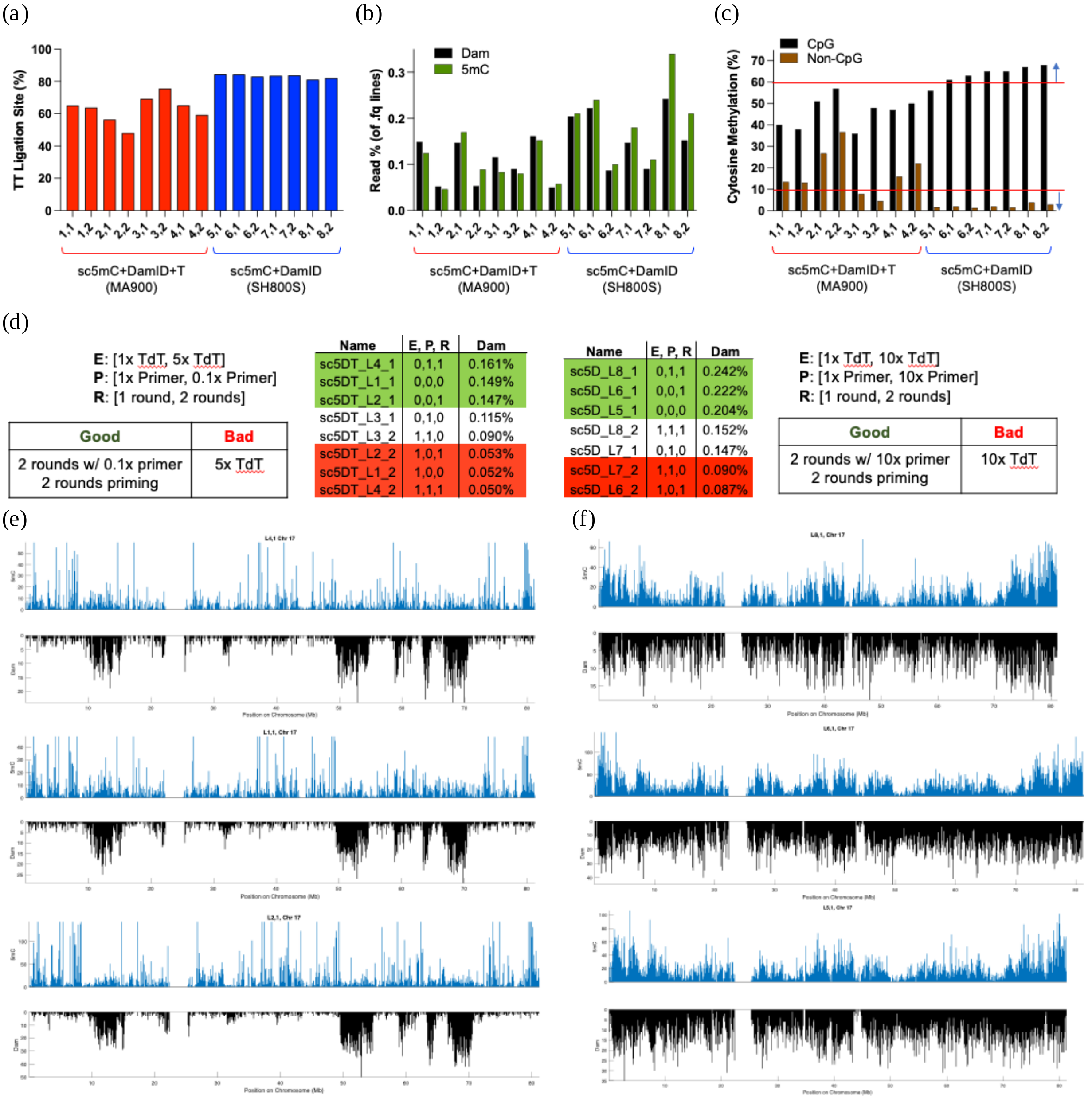


Fig. 33 | 2^3 Factorial Designs for dATP version of sc5mC+DamID+T and sc5mC+DamID. (a) sc5mC+DamID+Transcriptome method **(b)** sc5mC+DamID method two 384 well plates of single cells, one plate with the "standard" protocol (5mC+DamID), and the second with an extended version of the protocol (5mC+DamID+Transcriptome), which will be discussed in more depth in *Chapter V: Adding Transcriptome Measurements and Epitranscriptome*. When including the transcriptome measurement, the volumes and reagents are different than the standard protocol, so a control condition was performed for each of the factorial designs **(Fig. 33)**.

Of the 16 possible conditions, 15 libraries were successfully synthesized and sequenced. Similar to the previous sets of experiments, sequencing efficiencies were measured as part of the analysis. The ligation site dinucleotide was most frequently TT for all the conditions **(Fig. 34a)** however, there was a better efficiency for the non-transcriptome version of the protocol. To measure the 5mC and Dam sequencing efficiencies, the lines in the final 5mC and Dam output files were counted and normalized to the number of raw read **(Fig. 34b)** as before with the TACS and other dATP analysis. Based on the Dam read output efficiency, the top 5mC+DamID+T conditions were (1) decreasing the primer concentration and increasing priming rounds, (2) the control, and (3) increasing the priming rounds **(Fig. 34d)**. The worst conditions all had an increase in the TdT concentration. The single variable that had the greatest positive effect was the priming rounds because the Dam read output efficiency was higher for two rounds of priming (0.147%) than for lowering the primer concentration to 0.1x (0.115%). For the standard, non-transcriptome version of the protocol, a similar efficiency ranking of the variables was observed, where only increasing the rounds of priming produced a 0.222% Dam read output efficiency compared to only increasing the primer concentration (0.147%). The worst conditions all had increased TdT enzyme concentration. In addition to measuring the 5mC percentage in the CpG context, the cytosine methylation in the non-CpG



E: [1x TdT, 5x TdT]
P: [1x Primer, 0.1x Primer]
R: [1 round, 2 rounds]

Good	Bad
2 rounds w/ 0.1x primer 2 rounds priming	5x TdT

Name	E, P, R	Dam
🟩 sc5DT_L4_1	0,1,1	0.161%
🟩 sc5DT_L1_1	0,0,0	0.149%
🟩 sc5DT_L2_1	0,0,1	0.147%
sc5DT_L3_1	0,1,0	0.115%
sc5DT_L3_2	1,1,0	0.090%
🟥 sc5DT_L2_2	1,0,1	0.053%
🟥 sc5DT_L1_2	1,0,0	0.052%
🟥 sc5DT_L4_2	1,1,1	0.050%

Name	E, P, R	Dam
🟩 sc5D_L8_1	0,1,1	0.242%
🟩 sc5D_L6_1	0,0,1	0.222%
🟩 sc5D_L5_1	0,0,0	0.204%
sc5D_L8_2	1,1,1	0.152%
sc5D_L7_1	0,1,0	0.147%
🟥 sc5D_L7_2	1,1,0	0.090%
🟥 sc5D_L6_2	1,0,1	0.087%

E: [1x TdT, 10x TdT]
P: [1x Primer, 10x Primer]
R: [1 round, 2 rounds]

Good	Bad
2 rounds w/ 10x primer 2 rounds priming	10x TdT

Fig. 34 | psc5mC+DamID(+T) v2.2 (A-Tailing and Poly-T priming) library statistics. (a) Ligation site TT dinucleotide frequency **(b)** Final line count of Dam and 5mC files, normalized to raw read count. **(c)** Cytosine methylation percentage in the CpG and Non-CpG context **(d)** Summary of best and worst conditions **(e)** 5mC and Dam profiles from top psc5mC+DamID+T and **(f)** psc5mC+DamID conditions.

context (CHG or CHH, where H is A/C/T) was also measured to assess if the APOBEC was correctly deaminating **(Fig. 34c)**. The 5mC percentage in the CpG context should be between 60-80% and the methylation in the non-CpG context should be low (<10%). Both methylation percentages were correct for the sc5mC+DamID conditions (sorted on the Sony

SH800S), however for the sc5mC+DamID+T conditions (sorted on the Sony MA900) the methylation was low in the CpG context and high in the CHG/CHH context. Examining the non-normalized Dam and 5mC signals of the top three conditions from each plate, the Dam signal was correct for the 5mC+DamID+T plates that were sorted on the MA900, however their 5mC profiles were not correct along most of chromosome 17 **(Fig. 34e)**. This was the first time the Dam signal was correct, displaying high signal to noise, at the pseudo single-cell level with the v2 protocol (without MspJI), making this a milestone. For the standard 5mC+DamID plates that were sorted on the SH800S, the Dam profiles were wrong, however the 5mC profile were correct, with hypomethylation at the LADs **(Fig. 34f)**. This was surprising because a mostly similar protocol was performed on each of the plates and the tubes were handled at the same time. What likely occurred was the variations between the sc5mC+DamID+T and sc5mC+DamID protocols lead to the former favoring the Dam signal and the latter favoring the 5mC signal, since the psc5mC+DamID v2.2 experiment was repeated with a plate sorted on the MA900 and produced similar results (data not shown).

This raised the idea that the 5mC signal may be improved when there are many cuts along the chromosome, since the 5mC is read off the fragments that the Dam adapters bind to. Therefore, the 5mC may have had worse quality in the 5mC+DamID+T experiment because there were only cuts at the LADs. This would suggest the 5mC percentages in these LAD regions should be correct, whereas in the surrounding iLADs the percentages should be incorrect. To test the hypothesis that accurately measuring all 5mC may require frequent digestion, condition L4_1 (the library with the best Dam signal on the 5mC+DamID+T plate, seen in **Fig. 35c**) was used to determine LAD and iLAD coordinates, and the 5mC percentage was calculated at LAD and iLAD regions in each chromosome for the top three

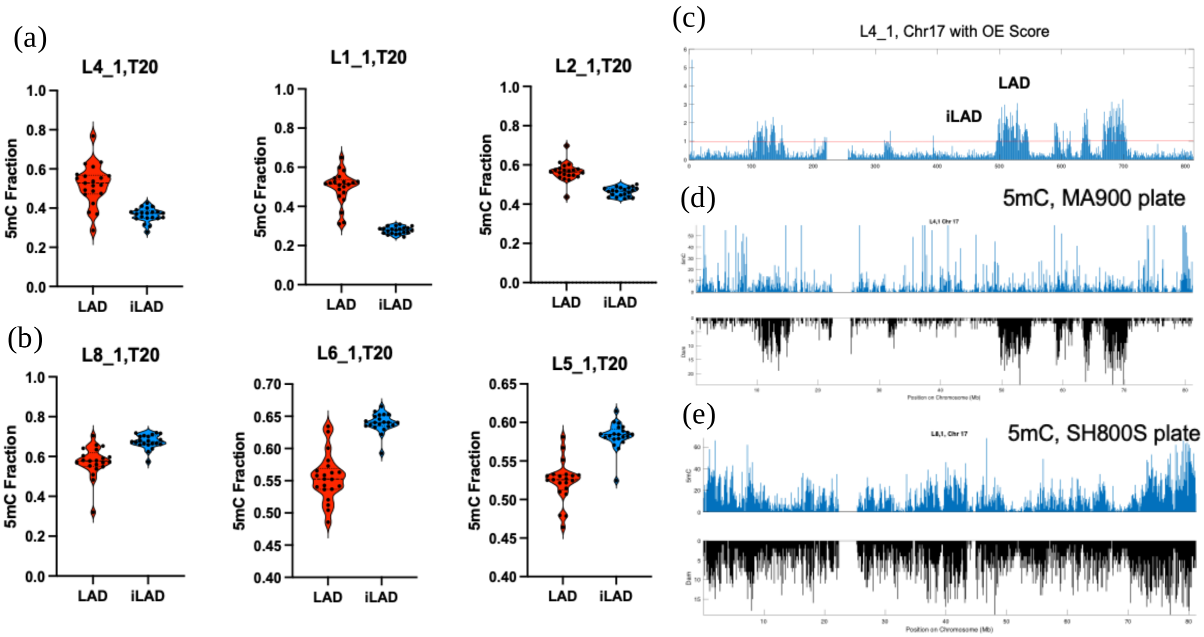
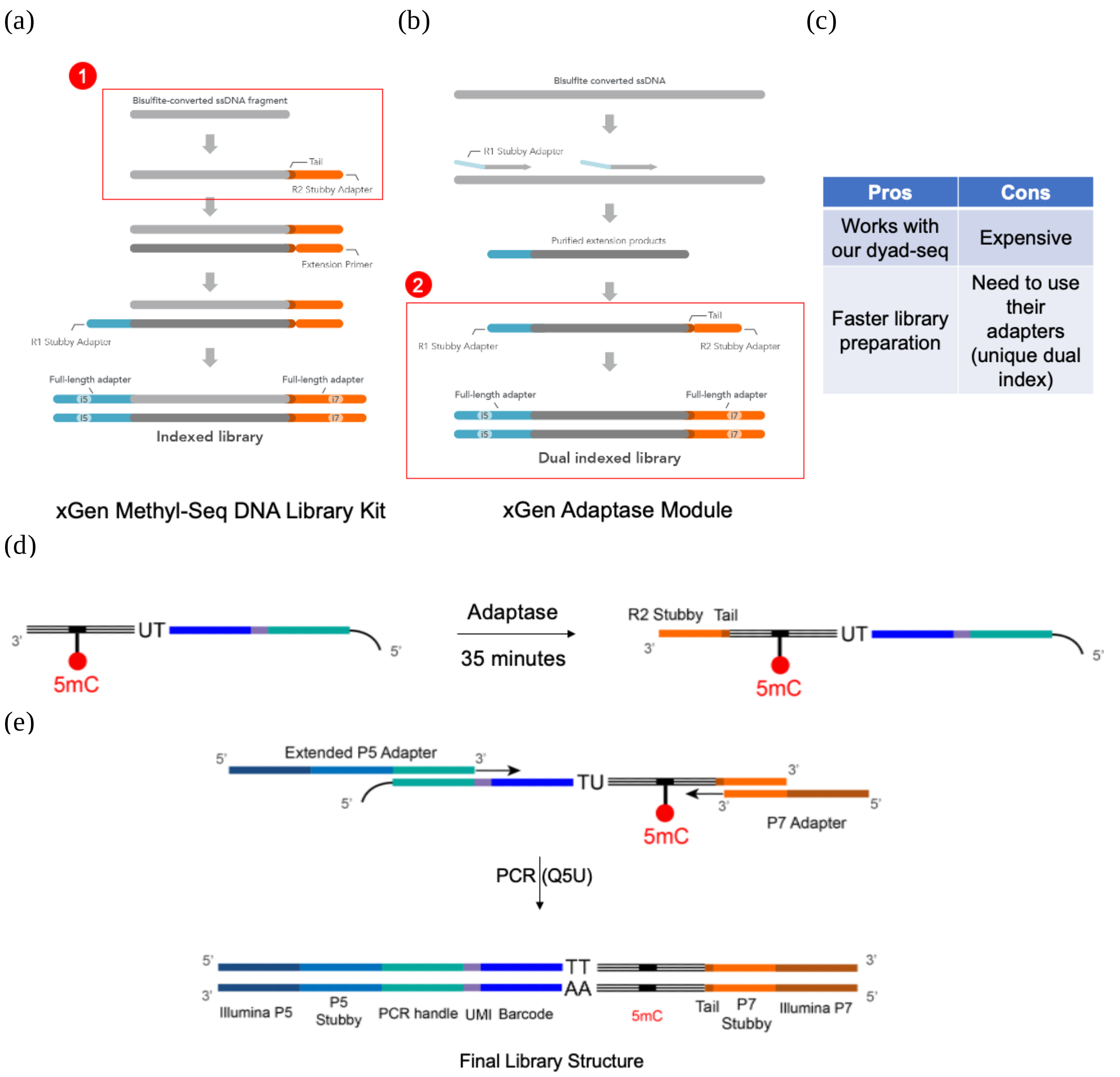


Fig. 35 | Detecting iLAD methylation may require frequent digestion. (a) Chromosomal percentage of 5mC in LAD and iLAD regions in top psc5mC+DamID+T and **(b)** psc5mC+DamID conditions. For (a,b) it was considered a LAD/iLAD if there were 2 bins or greater of contact/noncontact. Thresholded for bins containing a minimum of 20 CpG 5mC plus non-methylated CpG reads. **(c)** Dam signal of psc5mC+DamID+T v2.2 condition L4_1, the library with the best Dam signal in all of version 2, OE normalized, horizontal red line drawn at OE = 1 **(d)** 5mC and Dam Signal of psc5mC+DamID+T v2.2 condition L4_1 and **(e)** psc5mC+DamID v2.2 condition L8_1.

conditions from each plate. The conditions from both plates had similar LAD methylation percentages, with a median in the 50%'s **(Fig. 35a,b, red)**, suggestion the LAD methylation was correct. The MA900 plates however, where most of the Dam reads were only at the LADs, had 5mC percentages higher in the LADs than in the iLADs **(Fig. 35a)**. The trend is supposed to be the other way around, since LADs are hypomethylated compared to iLADs, which was seen with the SH800S plates **(Fig. 35b)** that contained more uniform Dam signal instead of LAD peaks **(Fig. 35d,e)**. This suggests that 5mC may only be good quality when there are many cuts along the chromosomes, a trend observed in all of the v2 libraries so far.

4. Adaptase (v2.3)

One last random priming alternative was attempted to improve the Dam signal, the xGen Adaptase Module. Although this method was designed originally for bisulfite converted DNA from single cells, the manufacturer (IDT) claimed it should work with enzymatic converted DNA fragments, such as in the APOBEC strategy used in all of v2 of the protocol. This strategy is very similar to the TACS method, in that an Adaptase step simultaneously performs end repair, tailing and ligation of an "R2 stubby adapter" to the 3' end of each fragment. The tailing step is similar to TdT, however it adds a low complexity, G-rich polynucleotide tail, with a median length of eight bases[101], and the R2 stubby adapter serves the same purpose as the dual phosphorylated RA3, for later attaching the full P7 Illumina sequence via PCR. IDT released two versions of the adaptase protocol: (1) **(Fig. 36a)** where the R2 stubby adapter was attached directly to bisulfite converted ssDNA fragments[102], tested at the 100 pg – 100 ng input range, and (2) **(Fig. 36b)**, where random priming is first used, however PCR is performed directly on ssDNA that contains priming sites on the 5' and 3' ends[103], and the version was tested on single cells. Parts of each of these protocols were combined **(Fig. 36a,b, red box)**, to create a method for sc5mC+DamID that doesn't include random priming and has just one PCR step. The benefit to this strategy is it integrates well with the 5mC+DamID v2 protocol and allows for faster library preparation, with only two streamlined reactions after the cytosine deamination to reach the final library. The downside is the Adaptase kit is more expensive than the other priming strategies and not compatible with the aforementioned RPI and RPI[xx] primers used for PCR, requiring the need to redesign the PCR primers. The schematic of the method is seen in **Fig. 36d**, where the adaptase step adds a tail and ligates the R2 stubby adapter to the 3' end of the bottom strand



Pros	Cons
Works with our dyad-seq	Expensive
Faster library preparation	Need to use their adapters (unique dual index)

Fig. 36 | sc5mC+DamID (v2.3) Random priming alternative: Adaptase. Schematic of **(a)** xGen Methyl-Seq DNA library prep kit protocol and **(b)** xGen Adaptase Module protocol. **(c)** Pros and cons of adaptase method. **(c)** Schematic of sc5mC+DamID using adaptase to add R2 stubby adapter tail and **(d)** PCR of fragment to generate final library.

of the Dam adapter-ligated fragment. From here, it goes directly into the PCR step, using both the PCR handle on the Dam adapter and the R2 stubby adapter as the PCR priming sites **(Fig. 36e)**. Since the protocol doesn't use the typical RPI and RPI[xx] primers, custom versions of the TruSeq HT combinatorial dual index adapters were designed and used for the adaptase experiments. The 5’ PCR primer is a joined Illumina P5 + P5 stubby + PCR handle sequence and the 3’ PCR primer is the reverse complement of the joined Illumina P7 and P7 stubby sequence that was truncated to lower the ΔT_m between the primers to ensure a universal annealing temperature.

C. Simultaneous digestion with DpnI & a Frequent Cutter (HpyCH4V or HincII) (v3.0)

Motivated by the results of the *A-Tailing and Poly-T Priming (v2.2)* section, signal needs to be acquired everywhere along the chromosome to measure all the 5mC. However, this appears to create a paradox, because cuts are only made by DpnI at the G(m^6A)↓TC motif that primarily occurs in LADs. Since the 5mC is read off only these fragments, 5mC will mostly be measured in the LADs when the Dam signal is correct, characterized by having high signal to noise. To measure both 5mC and genome-NL contacts with 5mC+DamID v2, cuts must be made everywhere along the chromosome while also cutting at the LADs in a way that is differently identifiable. This is best achieved by using a frequent blunt cutting enzyme that cleaves in a different context than DpnI, is incubated at the same temperature, and is able to be heat inactivated. Restriction enzymes HincII and HpyCH4V fit these criteria. HincII cuts in a GTY↓RAC context and is blocked by some combinations of overlapping CpG methylation[104], resulting in a cut frequency of <1/1024 bases, due to the presence of these R/Y variable bases. HpyCH4V creates blunt cuts in a TG↓CA context[105] and is not sensitive to CpG methylation, resulting in a cut frequency of 1/256 bases. DpnI is added simultaneously with HincII creating blunt cuts at both genome-NL contact sites and < 1/1024 bp, respectively **(Fig. 37b)**. Since blunt cuts were made, the blunt-ended Dam adapter

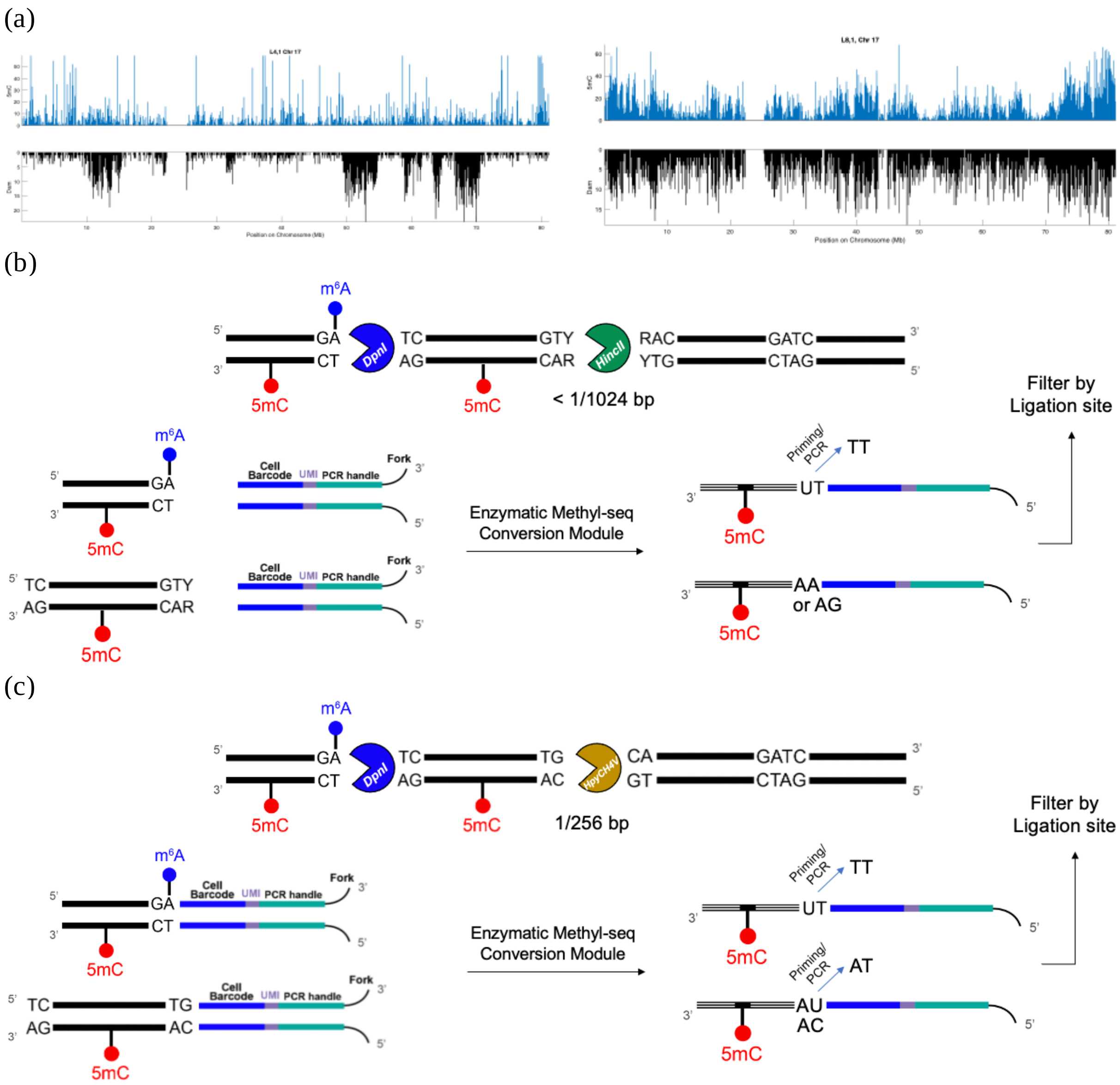
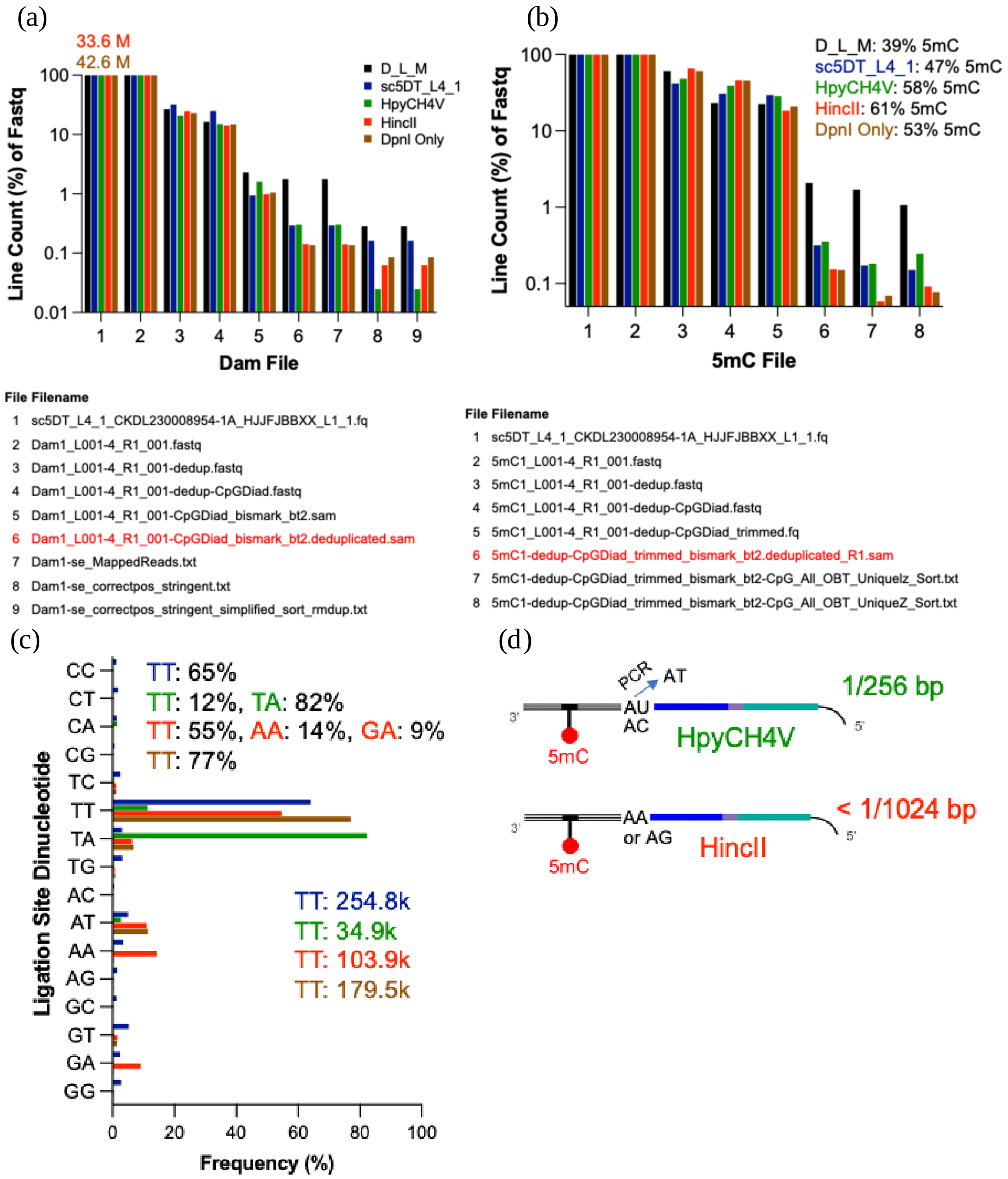


Fig. 37 | Improve 5mC signal acquisition with frequent DNA digestion. (a) 5mC and Dam profiles when detecting Dam signal (Left) only at the LADs and (Right) everywhere along the chromosome. **(b)** Simultaneously cutting the DNA with DpnI and HincI or **(c)** HpyCH4V (top), ligating Dam adapters (left), and fragment sequence after deamination (right).

can ligate to both restriction digestion sites, and different dinucleotides will be adjacent to the ligation sites following the cytosine deamination. 5’ TU is adjacent to the DpnI cut site, which will be converted to the extensively discussed TT dinucleotide following PCR amplification, and either 5’ GA or AA is adjacent to the HincII cut site. By filtering for only

TT ligation sites during the Dam analysis, only the genome-NL contact profile will be visible **(Fig. 37a, left)**, while 5mC will be detected all along the chromosome due to the frequent cutting **(Fig. 37a, right)**. Using HpyCH4V follows a similar scheme to HincII **(Fig. 37c)**, however it cuts more frequently (1/256 bases) and the 5' CA and UA is adjacent to the HpyCH4V cut site (depending on if the cytosine was unmodified/converted), and the UA is converted to TA following PCR amplification. By filtering the DpnI "TT" sites from the HpyCH4V cut sites, Dam signal can be discerned from the frequent cutting used for the 5mC signal.

The psc5mC+DamID v2.3 and v3.0 methods with adaptase were performed on the MA900 plates, from the same round of sorting as the plate with the correct Dam profiles. An efficiency analysis was performed on the adaptase libraries, like what had been done with the TACS and dATP libraries. Most reads were lost at the .sam deduplication file **(Fig. 38a,b)** for the adaptase HpyCH4V, HincII, and DpnI only conditions compared to the reference, suggesting the presence of PCR duplicates. The 5mC percentage was slightly lower than expected with a range of 53-61% 5mC compared to the psc5mC+DamID v2.0 one round of random priming condition that had 68% 5mC. The effect of frequent cutting was observable, as the HpyCH4V condition had the most lines in the non-methylated CpG and 5mC files, 7 and 8, **(Fig. 38b, green bar)**, and the least lines in the final Dam output file **(Fig. 38a, green bar)**. This is due to most of the detected sites pertaining to the HpyCH4V restriction sites, rather than the DpnI cut sites. The ligation site dinucleotide of this condition had 82% of adapters bound to TA sites, the correct restriction site of HpyCH4V, verses 12% binding to the DpnI "TT" site **(Fig. 38c)**, suggesting HpyCH4V is cutting too frequently and outnumbering the genome-NL contact site cuts. HincII however had 55% TT sites and 14%



File	Filename
1	sc5DT_L4_1_CKDL230008954-1A_HJJFJBBXX_L1_1.fq
2	Dam1_L001-4_R1_001.fastq
3	Dam1_L001-4_R1_001-dedup.fastq
4	Dam1_L001-4_R1_001-dedup-CpGDiad.fastq
5	Dam1_L001-4_R1_001-CpGDiad_bismark_bt2.sam
6	Dam1_L001-4_R1_001-CpGDiad_bismark_bt2.deduplicated.sam
7	Dam1-se_MappedReads.txt
8	Dam1-se_correctpos_stringent.txt
9	Dam1-se_correctpos_stringent_simplified_sort_rmdup.txt

File	Filename
1	sc5DT_L4_1_CKDL230008954-1A_HJJFJBBXX_L1_1.fq
2	5mC1_L001-4_R1_001.fastq
3	5mC1_L001-4_R1_001-dedup.fastq
4	5mC1_L001-4_R1_001-dedup-CpGDiad.fastq
5	5mC1_L001-4_R1_001-dedup-CpGDiad_trimmed.fq
6	5mC1-dedup-CpGDiad_trimmed_bismark_bt2.deduplicated_R1.sam
7	5mC1-dedup-CpGDiad_trimmed_bismark_bt2-CpG_All_OBT_Uniquelz_Sort.txt
8	5mC1-dedup-CpGDiad_trimmed_bismark_bt2-CpG_All_OBT_UniqueZ_Sort.txt

Fig. 38 | psc5mC+DamID v2.3 and v3.0 (Adaptase - DpnI only & DpnI+HincII/HpyCHV) library statistics. **(a)** Line count of Dam and **(b)** 5mC files, normalized to raw read count. **(c)** Ligation site dinucleotide frequency. **(d)** Expected ligation site for blunt adapters onto HpyCH4V and HincII cut sites.

AA and 9% GA sites, the correct restriction sites for this enzyme, validating that blunt cutting less frequently than HpyCH4V **(Fig. 38c)** increases the proportion of TT ligation sites. The DpnI only condition had 77% TT ligation sites, however overall, the adaptase method had lower efficiency than the v2.2 priming strategy **(Fig. 38a,b, brown vs. blue bars)**. Comparing the sc5mC+DamID version 1 CG normalized 5mC profile to the 5mC profiles of the adaptase conditions, only HpyCH4V produces the correct profile **(Fig. 39a)**. The HincII and DpnI only profiles are noisy from too few reads, lacking solid peaks and methylation dynamics along the chromosome, such as hypomethylation seen in the

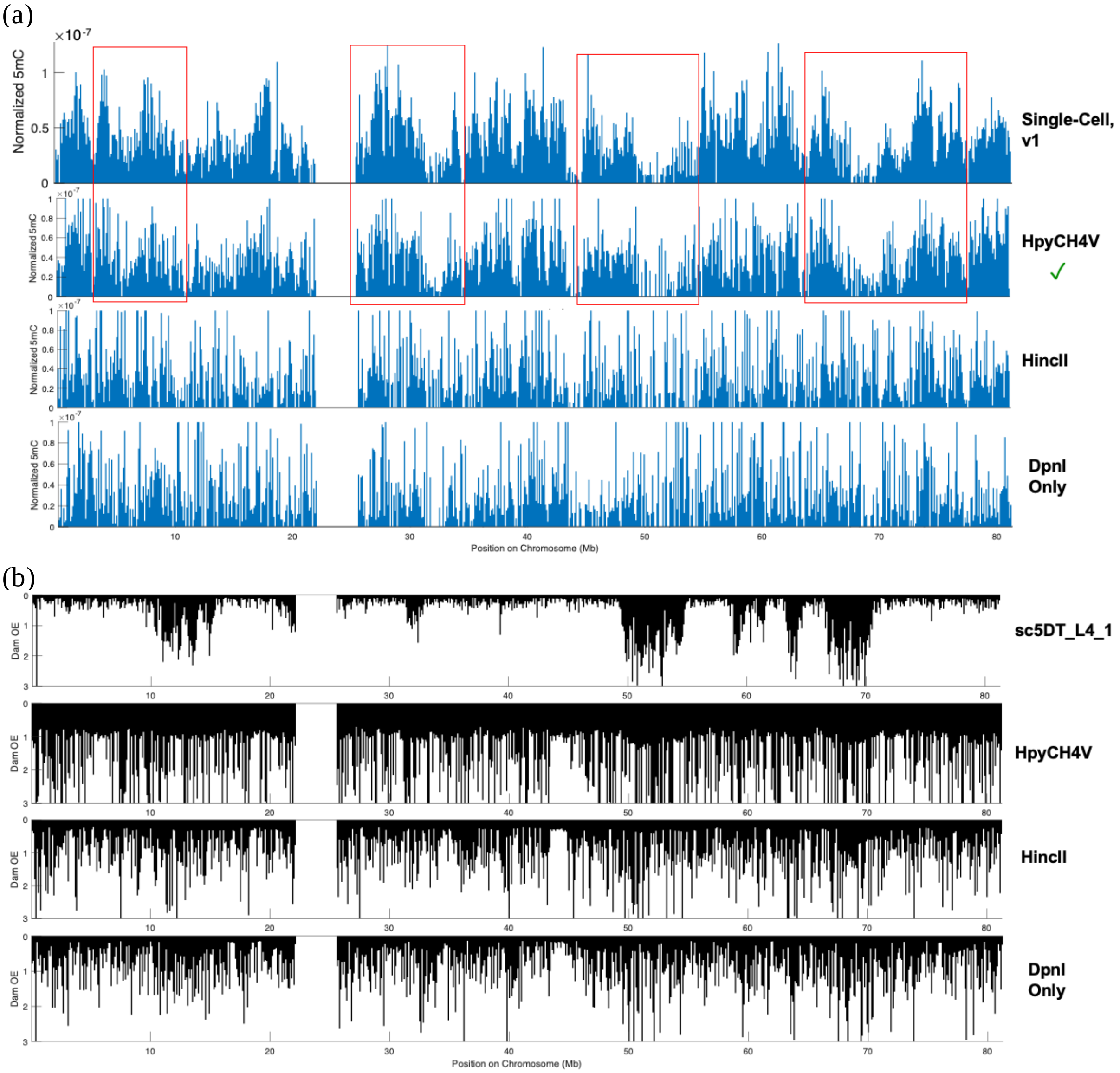


Fig. 39 | Adaptase 5mC and Dam profiles on 96 cells chromosome 17 vs. reference. (a) CG normalized 5mC profiles of sc5mC+DamID v1, psc5mC+DamID v3.0 (HpyCH4V then HincII), psc5mC+DamID v2.3. **(b)** OE normalized Dam profiles of sc5mC+DamID+T (L4_1), psc5mC+DamID v3.0 (HpyCH4V then HincII), psc5mC+DamID v2.3.

HpyCH4V condition. However, this confirms the hypothesis that more frequent digestion improves the 5mC signal, meaning this should fix the issue when the Dam is correct, but 5mC is only correctly measured in the LADs and not the iLADs. Despite the 5mC profile being correct in the adaptase condition with the addition of the HpyCH4V, the adaptase Dam profiles did not match the top dATP Dam profile **(Fig. 39b)**. This could be explained by the low efficiency of the adaptase strategy at the single-cell level, seen by generally more PCR duplicates compared to v2.2, or because of the variation in the reagents and volumes between the sc5mC+DamID+T and sc5mC+DamID methods. This suggests reverting to the A-Tailing and Poly-T priming strategy (v2.2) with the sc5mC+DamID+T method, using additional rounds of priming, the most effective variable in the full factorial design, and adding HincII or HpyCH4V, as in v3.0, to simultaneously digest with DpnI.

III. Single-Cell 5mC and DamID Results

A. 5mC & Genome-NL Contact Anticorrelation in Single Cell

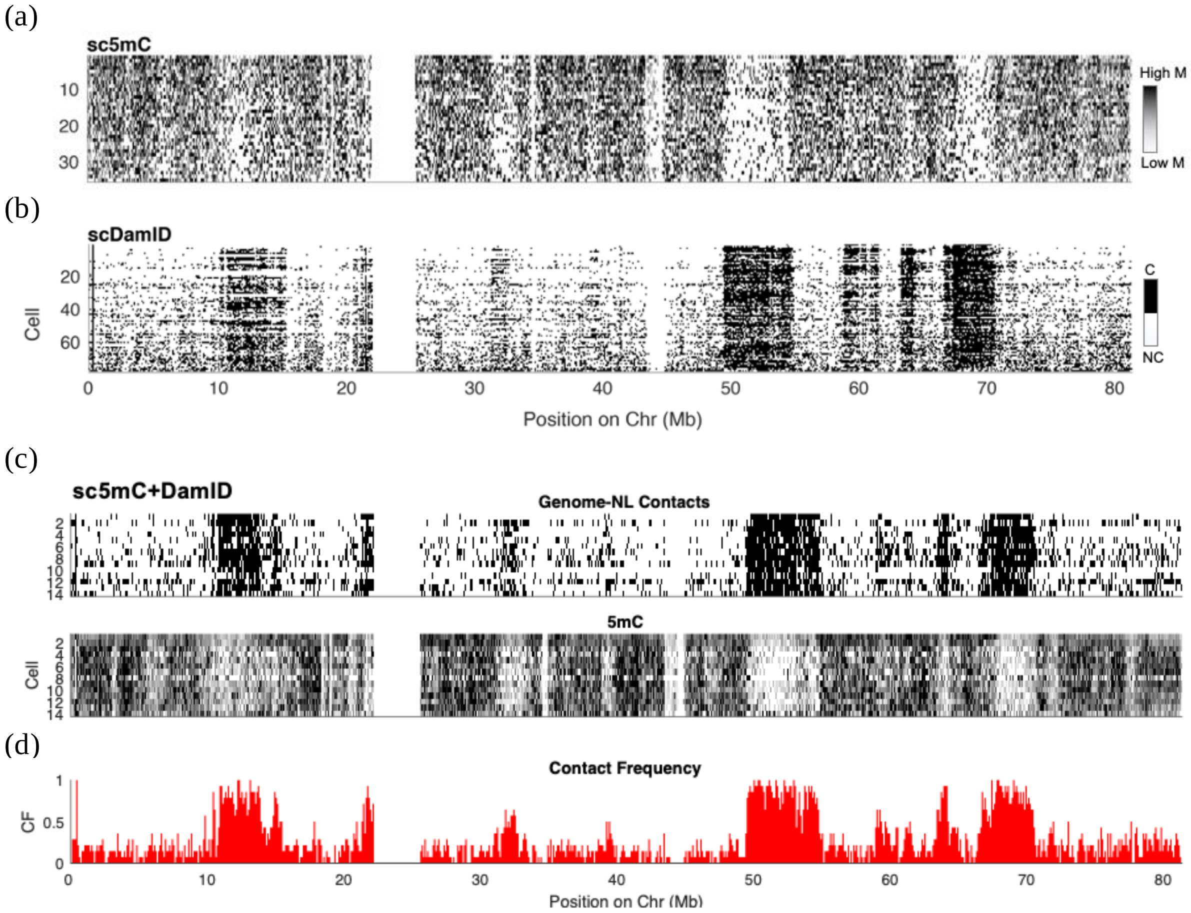


Fig. 40 | 5mC and Genome-NL Contact Profiles on Chr 17 in Single Cells. (a) Measured only 5mC or **(b)** only genome-NL in many cells by using either MspJI or DpnI. Row: cell, column: pos. on chr 17, Pixel intensity: amount of 5mC/genome-NL contact. **(c)** Combined measurements – 5mC + DamID, same rows represent corresponding profiles of the same cell. For (a,c), only cells that contained over a threshold of 10,000 total 5mC reads were considered for analysis. For (a-c), spikes of technical artifacts from mapping repetitive sequences near the centromere were removed. **(d)** Contact frequency of chromosome 17 in single cells from sc5mC+DamID. Values range between 0 and 1 indicating no cells make NL contacts within the bin or all cells make NL contacts within the bin, omitting the centromeres.

This chapter aims to answer if 5mC and genome-NL contacts are anticorrelated at the single-cell level. It has been described previously that they are anticorrelated in bulk, but it has never been observed along a full chromosome at the single-cell level[1]. To lay the foundation for simultaneous sequencing, first single-cell 5mC only sequencing was performed, using a strategy similar to v1, omitting simultaneous digestion and just using MspJI, generating the profile in **Fig. 40a**. The cells display heterogeneity in their 5mC expression and hypomethylation, such as around 50 Mb and 70 Mb. Single-cell DamID was performed similar to v1, but only using DpnI in the digestion step, generating the LAD profile in **Fig. 40b**. Heterogeneity in contact within LADs is observed between the cells, and in some cells there is an absence of certain LADs, such as the LAD at 10 Mb in cell 18. To further validate the LADs were measured correctly, the percentage of the chromosomal bins that were considered LADs out of the total LAD and iLAD bins were assessed for each chromosome. Chromosome 18 displayed the greatest percentage of LADs and chromosome 19 displayed the least percentage of LADs, seen in **Fig. 41a**, which is in agreement with previous studies that indicated chromosome 18 tends to be located at the periphery, and chromosome 19 tends to be located at the interior[106,107]. Even though these 5mC and genome-NL contact marks were observed in different cells, the locations of the LADs are visible and correspond with the locations lacking methylation. However, this does not fully prove methylation loss occurs at genome-NL contacts, requiring making the measurements within the same cell. Using the sc5mC+DamID v1 method, 5mC and genome-NL contacts marks were measured within the same cell **(Fig. 40c)**. The combined and separate epigenetic measurements produce the same profiles, and there is a visual anticorrelation between the 5mC and LADs. Since the same hypomethylation profile is present in sc5mC as in the 5mC measurement of the combined, it confirms the observed hypomethylation in the combined measurement is not a result of adding both enzymes, MspJI and DpnI, at the same time, creating competition for the sites and a loss of reads.

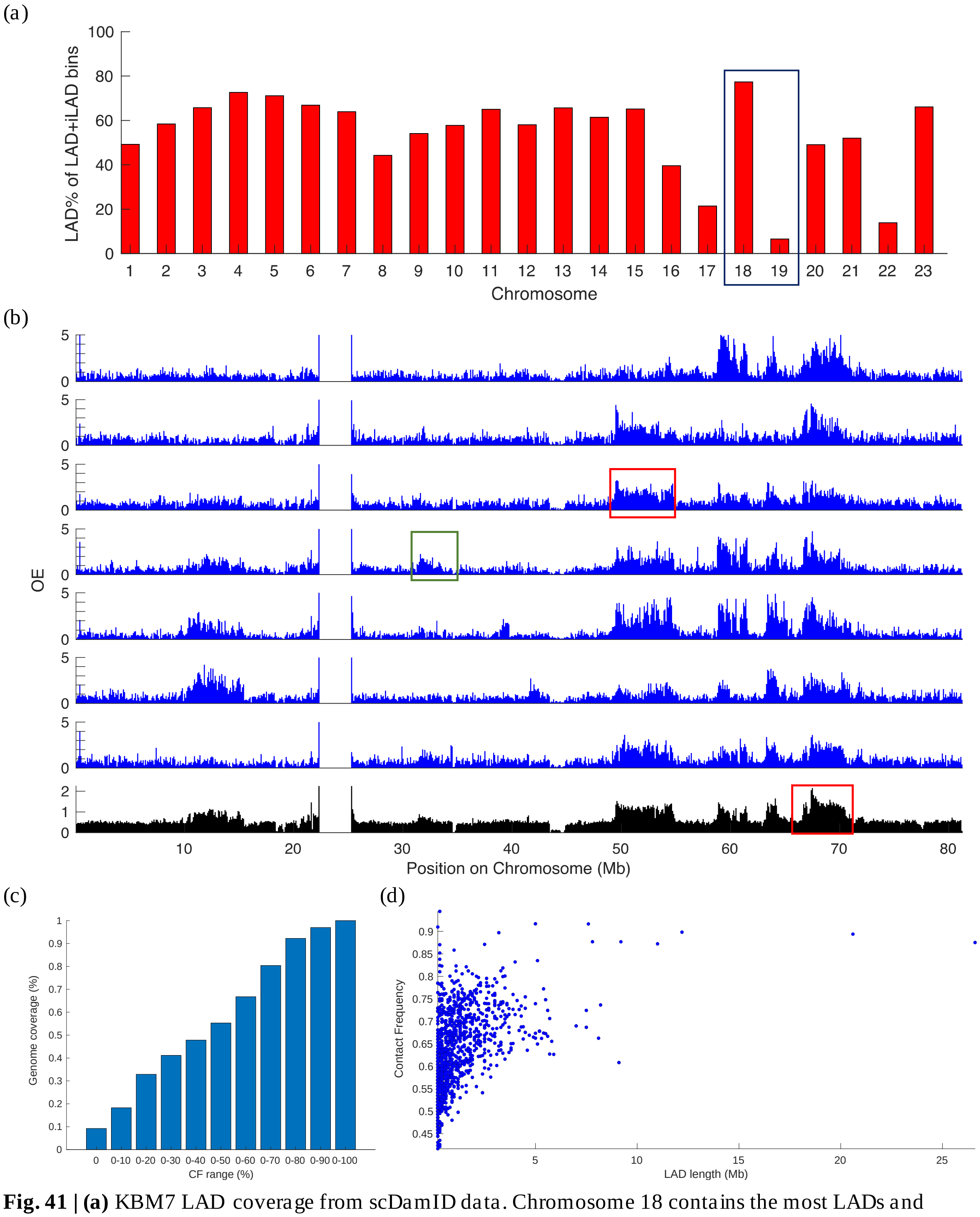


Fig. 41 | (a) KBM7 LAD coverage from scDamID data. Chromosome 18 contains the most LADs and chromosome 19 displays the fewest number of LADs. **(b)** LAD heterogeneity between single cells on Chromosome 17 from scDamID data. First seven lanes are LAD profiles of seven single cells, and the last lane is the Dam read ensemble of all the single cells. Consistent and variable LADs are highlighted by red and green boxes, respectively. **(c)** Cumulative contact frequency distribution of contact frequencies in all chromosomes. LADs within the lower CF range represent more variable LADs, whereas LADs in the higher CF ranges represent consistent LADs. **(d)** Contact frequency of LADs vs. the length of the LAD.

Focusing on chromosome 17, Dam reads from seven single cells, measured with scDamID, were plotted alongside the ensemble of Dam reads **(Fig. 41b)**. They displayed consistent LADs, those making genome-NL contact in most cells (such as the LAD at 50-55 Mb and 65-70 Mb), variable LADs, regions only making contact in some of the cells (such as the LAD at 32-35 Mb), and inter-LADs where OE < 1. Motivated by the heterogeneity in genome-NL contacts in the single cells, the contact frequency (CF), or in what fraction of cells a region contacts the NL was quantified with the single-cell measurements of genome-NL contacts. It is scored as a fraction between 0 and 1, where 0 represents none of the cells making an NL contact in the bin and 1 representing all cells making an NL contact in the bin. A LAD is considered consistent, if the contact frequency is closer to 1, such as the LAD around 50 Mb, that appears in most cells **(Fig. 40d)** or it is considered a variable LAD if the contact frequency is closer to 0.5, such as the LAD around 30 Mb. Contact frequencies were plotted cumulatively in **Fig. 41c** for the entire genome. The contact frequency of each LAD was plotted against the length of the same LAD **(Fig. 41d)**. As the length of the LADs increase, the CFs tend to shift towards higher values, supporting the notion that LADs represent long stretches of DNA that regulate gene expression by peripheral positioning into heterochromatin.

The 5mC levels were measured in LADs and iLADs of a single cell, using the OE normalized genome-NL contacts to determine which regions were at, and away, from the nuclear periphery **(Fig. 42a)**. Lower amounts of 5mC were observed in the LADs compared to in the iLADs, suggesting that 5mC and genome-NL contacts are anticorrelated at the single cell level. Single-cell DamID data was then used to compute the contact frequency and single-cell 5mC data was used to compute the 5mC as a function of its contact frequency.

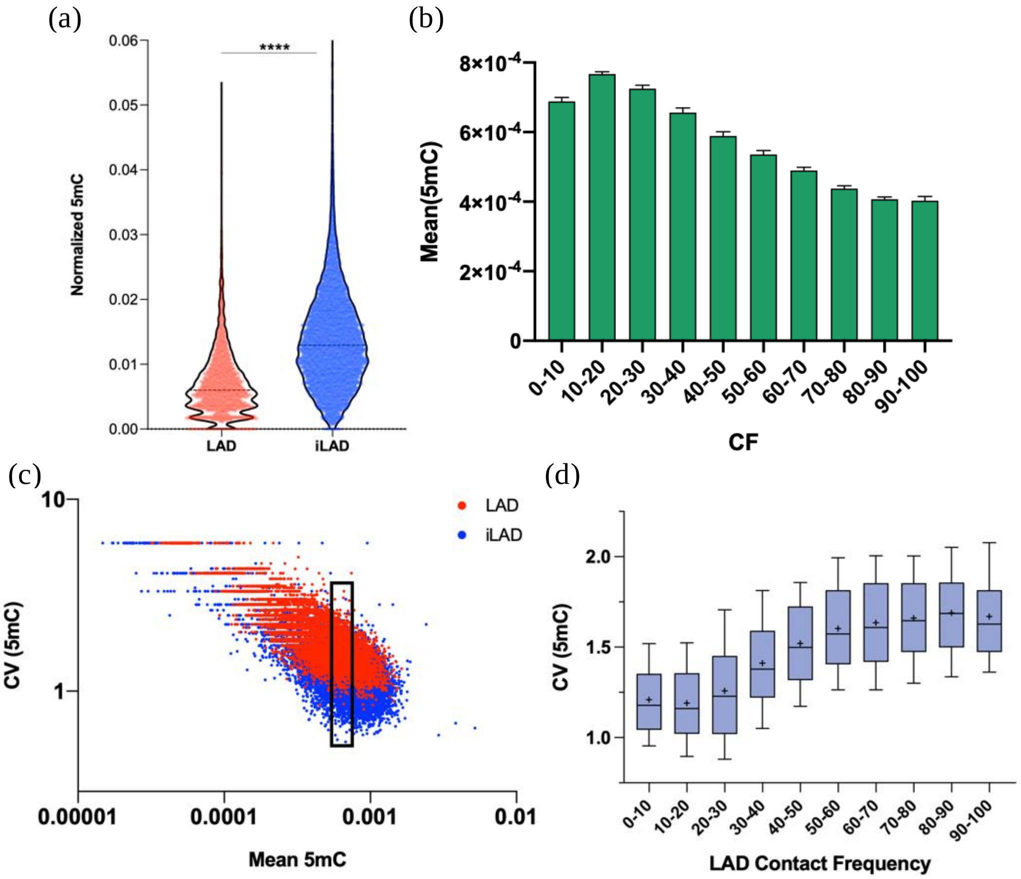


Fig. 42 | (a) 5mC & Genome-NL Contact Anticorrelation in the Single Cell. 5mC levels in 100 kB bins marked as LADs or iLADs. CG normalized. Statistical significance assessed with a Mann-Whitney test, P value < 0.0001. **(b)** Mean 5mC levels vs. contact frequency. **(c)** Coefficient of variation (CV) of 5mC vs. mean 5mC in LAD or iLAD regions. Each point is position along the chromosome. **(d)** CV(5mC) vs LAD contact frequency at fixed mean, defined by window drawn in (c). For b-d 5mC is from sc5mC only cells, and relations are from corresponding bins between single cells. For a-d, LAD/iLAD coordinates and contact frequencies along chromosomes were defined from pseudobulk DamID data.

Regions that frequently contact the nuclear lamina, consistent LADs, tended to have lower 5mC levels than more variable LADs **(Fig. 42b)**. This suggests the extent of methylation loss depends on the amount of contact with the periphery in the cell line. The 5mC levels were analyzed in corresponding bins of single cells along the chromosome and assessed for variability, considering whether they were in LADs or in iLAD regions **(Fig. 42c)**. LADs (red) tended to reach lower maximum amounts of 5mC than iLADs (blue) and have more variability in 5mC at any mean level of 5mC. iLADs however had lower noise in 5mC for any average level of 5mC, suggesting a similar iLAD methylation pattern between cells, rather than the cells experiencing varying degrees of methylation loss in the LADs. After controlling for mean 5mC levels, it was found as contact frequency increases, the variability in 5mC increases, suggesting in this cancer cell line, the frequent NL contacts are subject to a greater heterogeneity in their methylation loss **(Fig. 42d)**. The reason the 5mC level needs to be controlled for while computing the variability in 5mC, is that a low 5mC baseline could be more prone to fluctuations in 5mC, that could be mistaken for a higher CV(5mC). To summarize, there is lower 5mC in LADs compared to iLADs, and the most hypomethylation, or methylation loss, occurs at frequent genome-NL contacts in this cancer cell line.

B. LAD Multivalency and Contact Runs Correlates with 5mC Levels

Motivated by the variations in genome-NL contacts at the single-cell level, the gaps between the contacts were examined for methylation trends. Presented in **Fig. 43a** are three high resolution 5mC+DamID profiles, each corresponding to simultaneous measurements from the same cell. Even though they are profiles of the same chromosome, they display different distributions of contact lengths in similar LAD regions **(Fig. 43b)**. This suggests some cells don’t maintain stretches of genome-NL contact for as long within the same locations on the chromosome, and this phenomenon will be referred to as the “contact run”. By measuring both epigenetic marks within the same cell, this contact run metric can be used to better understand why certain LADs tend to be more hypomethylated. Some cells have more fragmented contact runs, seen by comparing the contacts in the LAD regions

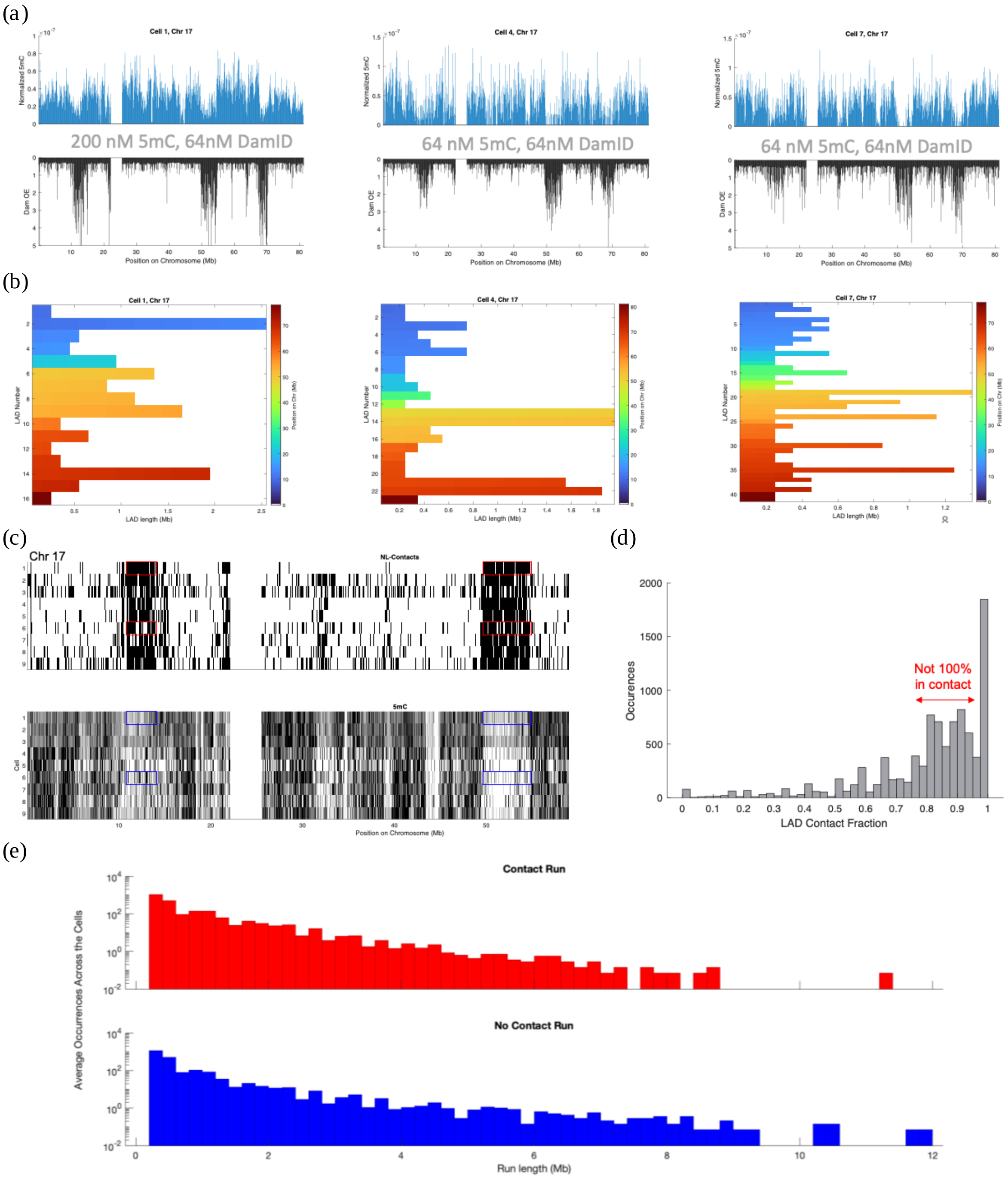


Fig. 43 | LAD Multivalency and Contact Runs. (a) 5mC and Dam profiles from sc5mC+DamID v1.0 **(b)** Contact lengths of LADs in chromosome 17 from cells in (a). LADs are called when 2 consecutive bins have OE > 1 **(c)** 5mC and Dam profiles from the same cell, highlighting the differences in contact runs and 5mC profiles. **(d)** Fraction of pseudobulk called LAD in contact at the single-cell level. **(e)** Average distribution of LAD run lengths for genome-NL contacts (top) and non-contacts (bottom). Normalized by cell count in sample.

highlighted in red in cell 1 and cell 6 in **Fig. 43c**, and in fact a significant number of LADs are not 100% in contact at the single-cell level, ranging from 80-90% in **Fig. 43d**. This suggests a coordination between neighboring segments, meaning an entire region could still be anchored to the nuclear lamina, even though it is less than 100% in contact. This phenomenon has been observed previously[8], and it is analogous to how a dress shirt can still be attached without buttoning every button. These contact run lengths can extend up to around 9 Mb long, based on their average distribution between the cells **(Fig. 43e)**. Given the variability in contact runs within LADs at the single-cell level, it was investigated if the methylation patterns in LADs with more contacts are different than those with less contacts.

By defining the LAD boundary coordinates using a pseudo-bulk interpretation of the scDamID data, the space between genome-NL contact runs becomes visible **(Fig. 44a)**. For example, the contacts of cell 1 and cell 6 are highlighted within the LAD at position 50-55 Mb on chromosome 17, and cell 1 makes more contact (shown in black) along the region, whereas cell 6's contact is broken into more and smaller segments, containing white segments of noncontact. Within the LADs, the segments of contact have less 5mC than the segments of noncontact, suggesting LAD 5mC is partitioned and is concentrated in the noncontact segments **(Fig. 44b)**. The amount of 5mC within each of these contact and noncontact segments is further dependent on the length of the segment **(Fig. 44c)**. Longer contact (and noncontact) segments are more hypomethylated than shorter segments, and the noncontact segments have greater average amounts of 5mC for all segments lengths up to around 1 Mb. The noncontact segment trend past 1 Mb is likely explained by the limited number of noncontact segments of this size, and likewise the behavior of the contact segments approaching 10 Mb (frequencies seen in **Fig. 43e**). However, there appears to be

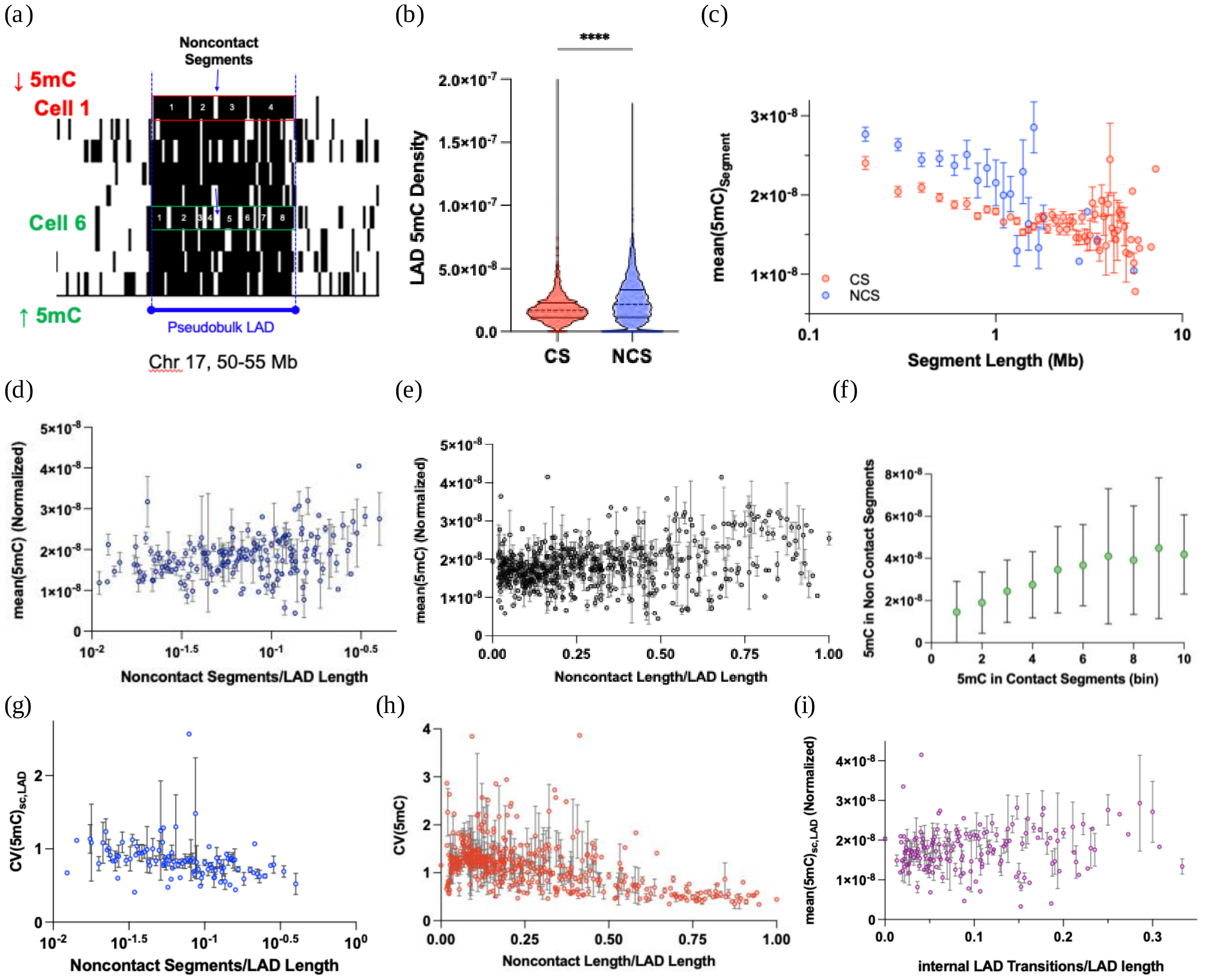


Fig. 44 | Space Between Multivalent Contacts Correlates with 5mC Levels. (a) Contact segments and non-contact segments in LAD at 50-55 Mb on Chr 17. **(b)** Sum of 5mC levels in contact and non-contact segments of LADs normalized by total run lengths. Kolmogorov-Smirnov test (P value <0.0001) **(c)** Average level of 5mC in contact and non-contact segments as a function of segment length. (**d)** Average 5mC levels in each single cell LAD vs. the number of non-contact segments or **(e)** non-contact length, both normalized by LAD length. **(f)** 5mC in non-contact segments vs. in contact segments of the same single-cell LAD. **(g)** CV of 5mC levels in each single cell LAD vs. the number of non-contact segments, at a fixed mean 5mC, or **(h)** non-contact length, both normalized by LAD length. **(i)** Average 5mC of each single cell LAD vs. transitions between contact and non-contact regions (2 or more consecutive bins), normalized to LAD length. (b-i) Includes only LADs spanning over 500 kb (5 bins) and in (b,d,e,g-i) each data point is 1 LAD.

some degree of balance in contact and noncontact segment 5mC, because when there is lower 5mC in contact segments, there is also lower 5mC in the noncontact segments of the same single-cell LAD **(Fig. 44f)**. As LADs fragment, represented by more segments of noncontact, they tend to reach higher average levels of 5mC and less 5mC variation along the single-cell LAD **(Fig. 44d,g)**, however this trend is noisier and weaker than 5mC vs. segment length. Similarly, LADs that have a greater fraction of the region not in contact tend to mostly reach higher average levels of 5mC and lower 5mC variation along each single-cell LAD **(Fig. 44e,h)**, however again the trends are weaker than the 5mC vs. segment length. This may be explained by using small and non-uniform bin sizes in the plot, but it is most likely explained by not factoring in how long the contact and noncontact segments are, which appears to have the stronger correlation with 5mC. For example, LAD "A" could have a noncontact fraction of 0.5 with two long noncontact segments and three long contact segments, and a similarly sized LAD "B" could also have a noncontact fraction of 0.5, but contain five small noncontact segments alternating with six small contact segments. Because smaller segment lengths are well correlated with higher levels of 5mC, LAD "B" would contain more 5mC than LAD "A", despite having the same noncontact length, resulting in a confounded correlation, such as in **Fig. 44e**. Therefore, an alternative analysis could consider how frequent the contact and noncontact alternates (or transitions) within a region, as in **(Fig. 44i)**. With more switching between contact and noncontact, the 5mC tends to increase, most likely due to the noncontact and contact segment sizes shrinking. However, this analysis does not consider the lengths of each contact/noncontact segment, only the number of transitions between them. This could lead to states of parity in transition number but different average 5mC levels in the LAD, similar to the aforementioned LAD "A" / LAD "B" cases, e.g. a LAD transition in the form of $\text{Short}_{\text{Contact}}\rightarrow\text{Long}_{\text{Noncontact}}\rightarrow\text{Short}_{\text{Contact}}\rightarrow\text{Long}_{\text{Noncontact}}$ (3 transitions) or $\text{Long}_{\text{Contact}}\rightarrow\text{Short}_{\text{Noncontact}}\rightarrow\text{Long}_{\text{Contact}}\rightarrow\text{Short}_{\text{Noncontact}}$ (3 transitions). These results elucidate the previous observation that LADs are hypomethylated, but it is not simply the DNA contacts the NL so there is loss of 5mC, but the LAD hypomethylation depends on the length of the segments in contact (and not in contact) with the NL, where longer stretches of contact are correlated with less 5mC. Further, within the LAD region, 5mC density is partitioned between the contact and noncontact segments, where more 5mC is found in noncontact segments than contact segments of the same length.

C. iLAD 5mC Partitions into Contact and Noncontact Segments

With noncontact segments appearing in LADs at the single-cell level and displaying trends with increased methylation, iLADs were studied to determine whether they display the opposite correlation, making some contact with the NL lamina, causing loss in methylation, depicted in **Fig. 45a**. LADs and iLADs display parallel trends, including segmentation based on contact frequency, revealing a similar negative correlation between noncontact segments & LAD contact frequency and contact segments & iLAD noncontact frequency (in what fraction of cells the region is situated away from the nuclear lamina). LADs frequently situated at the NL tend to have fewer noncontact segments than more variable LADs, that have more noncontact segments **(Fig. 45b)**. The same trend holds true for iLADs, where consistent iLADs, domains that rarely contact the NL, contain few contact segments at the single-cell level, whereas more variable iLADs make some contact with the NL **(Fig. 45c)**, however not enough for the entire domain to be situated at the periphery. Like LADs, iLAD 5mC density is partitioned into its noncontact and contact segments, where most 5mC is in the noncontact regions **(Fig. 45d)**. In terms of segment length, longer contacts in iLADs have less average 5mC in the segment and are more hypomethylated than shorter contacts, and there was always more 5mC in noncontact segments of the same length **(Fig. 45e)**.

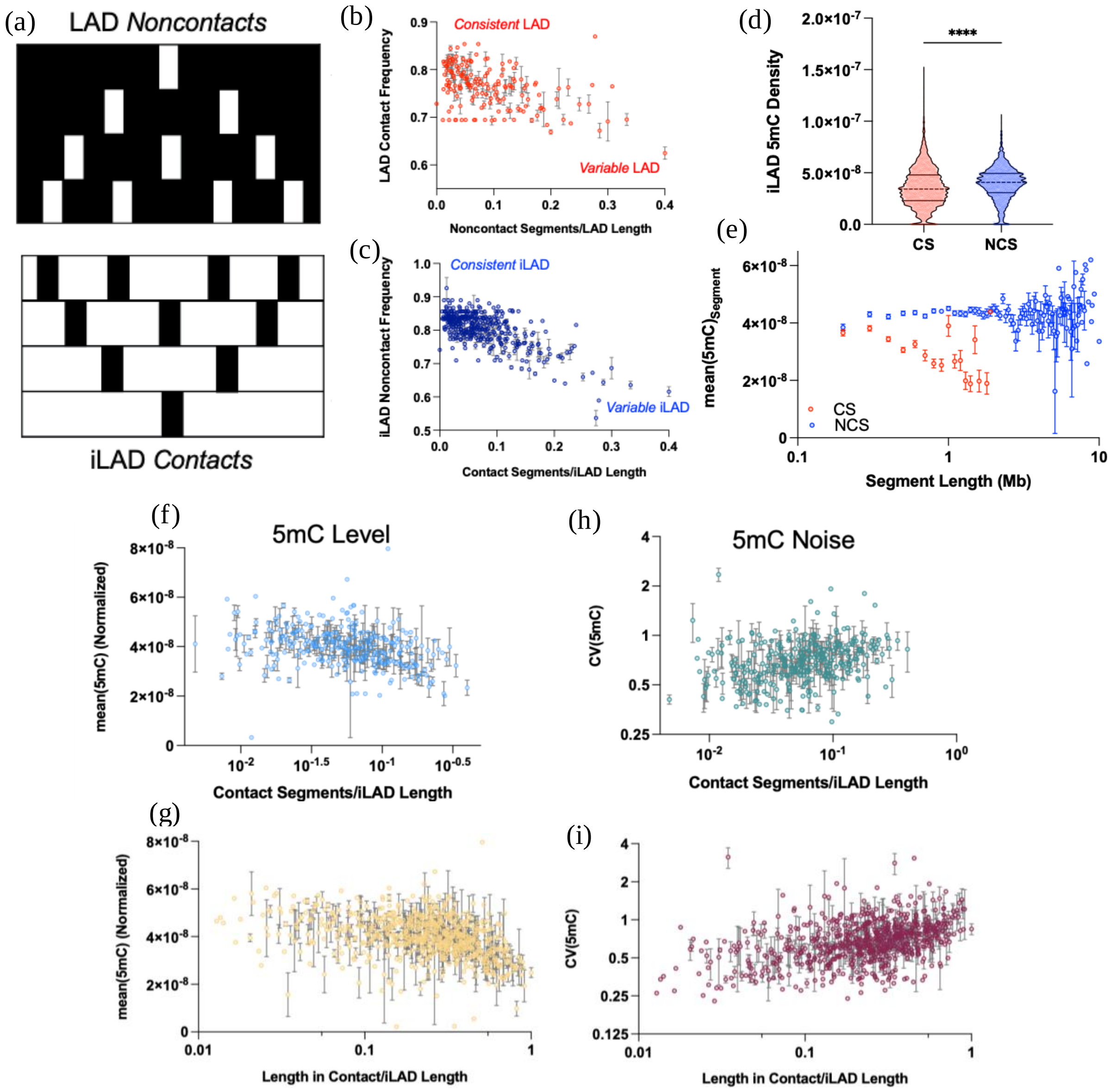


Fig. 45 | iLAD 5mC Partitions into Contact and Noncontact Segments. (a) Depiction of noncontact segments in LADs (top) and contact segments in iLADs (bottom). **(b)** LAD contact frequency as a function of noncontact segments and **(c)** iLAD noncontact frequency as a function of contact segments, each normalized by LAD or iLAD length, respectively. **(d)** Sum of 5mC levels in contact and non-contact segments of iLADs normalized by total run lengths. Kolmogorov-Smirnov test (P value <0.0001) **(e)** Average level of 5mC in iLAD contact and non-contact segments as a function of segment length. (**f)** Average 5mC levels in each single cell iLAD vs. the number of contact segments or **(g)** contact length, both normalized by iLAD length. **(h)** CV of 5mC levels in each single cell iLAD vs. the number of contact segments or **(i)** contact length, both normalized by iLAD length.

Therefore, the segments in iLADs display the same trend with methylation as in LADs **(Fig. 44c)** and more generally this observation supports regions that contact the NL are hypomethylated. However, there is no change in 5mC based on the noncontact segment length, likely because iLADs in cancer cells aren't normally hypomethylated like the LADs. iLADs with more contact segments reached lower average 5mC than those iLADs with fewer contact segments **(Fig. 45f)**, and a similar trend held with contact fraction **(Fig. 45g)**. Likewise, iLADs with more contact segments had more variable 5mC along the region **(Fig. 45h)**, however this was a weak correlation, and a greater contact fraction in iLADs displayed a positive correlation with variability in 5mC **(Fig. 45i)**. These trends were weaker than the segment trends in iLADs and are again explained by the contact segment size in iLADs contributing more to the 5mC levels than the number of contact segments or fraction of contact in the iLAD. The range of mean iLAD 5mC levels for a fixed number of contacts or contact fraction is most likely explained again by parity situations, such as iLAD "A" containing the same number of contact segments or contact fraction as iLAD "B", but the sizes of the segments are different, resulting in different amounts of 5mC. To summarize, contacts within iLADs observed at the single-cell level correlate with losses of 5mC, and further reflect genome-NL contacts are hypomethylated in cancer.

D. Activating and Repressive Histone Modification on Contact Frequency

Publicly available data KBM7 ChIP-Seq sets[108,109] were used to draw more connections between how genome-NL contacts and epigenetic features are related, focusing on the correlation between activating and repressive histone modification and contact frequency. Consistent LADs tend to have fewer activating histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K79me2, H4K16ac) than variable LADs **(Fig. 46a-h)**. Since LADs are typically thought of as transcriptionally silent condensed DNA

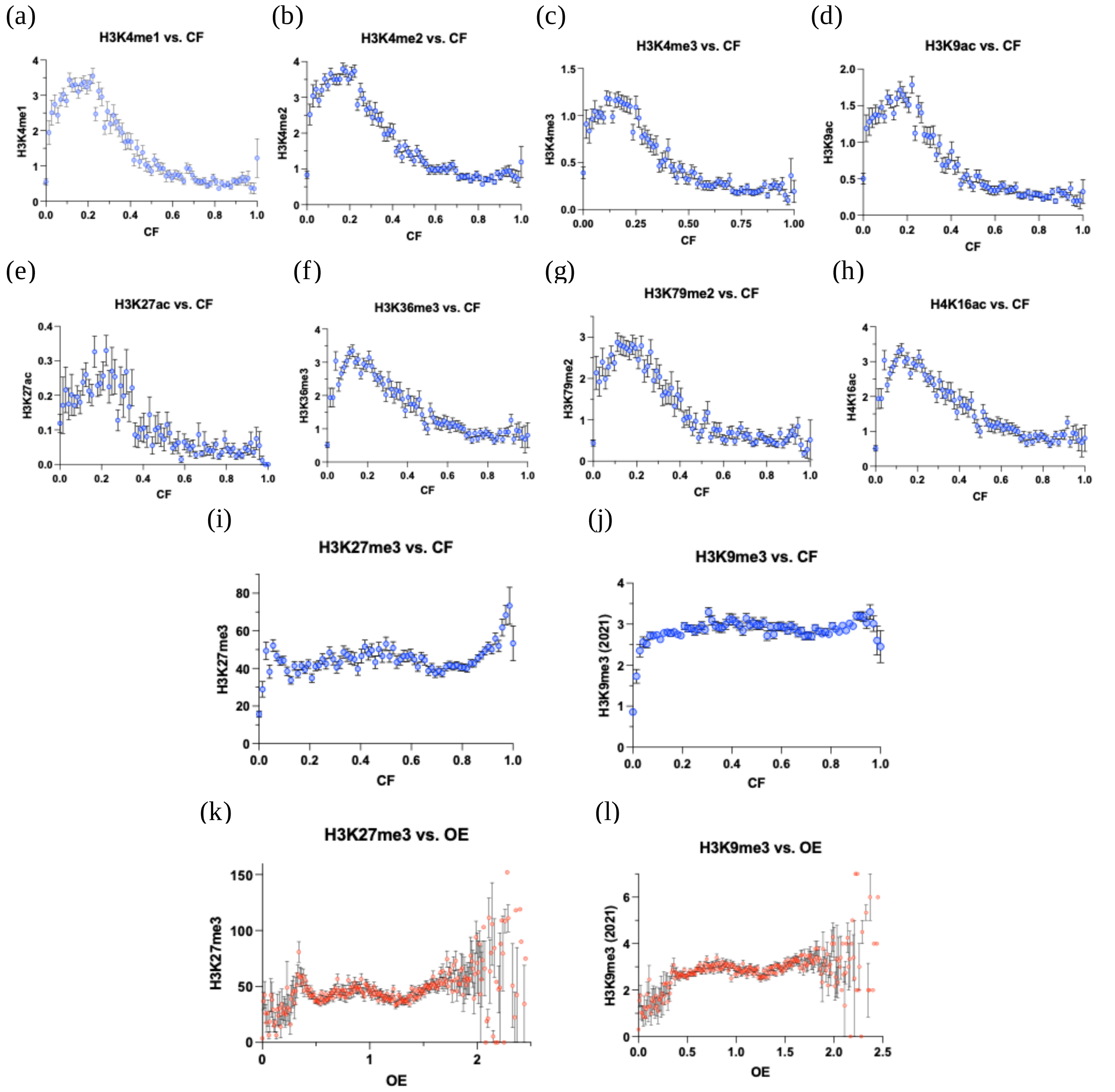


Fig. 46 | Activating and Repressive Histone Modifications on Contact Frequency. (a) Amount of H3K4me1, **(b)** H3K4me2, **(c)** H3K4me3, **(d)** H3K9ac, **(e)** H3K27ac, **(f)** H3K36me3, **(g)** H3K79me2, **(h)** H4K16ac, **(i)** H3K27me3, and **(j)** H3K9me3 as a function the contact frequency of the bin. **(k)** H3K27me3 and **(l)** H3K9me3 marks as a function of genome-NL contact OE score.

at the periphery, it agrees that regions frequently in contact with the NL will have less of these activating histone modifications. DNA that is less frequently in contact with the NL, considered as more uncondensed and transcriptionally active towards the nuclear interior, would contain more activating modifications. Consistent LADs tended to have more repressive histone modifications (H3K27me3, H3K9me3) than variable LADs **(Fig. 46l,j)**. Considering H3K9 methylation is linked with NL tethering[110,111], it is unsurprising this mark was most present at the regions that most frequently contacted the NL.

E. Proposed Schematic of Noncontact Segment Phenomena in LADs

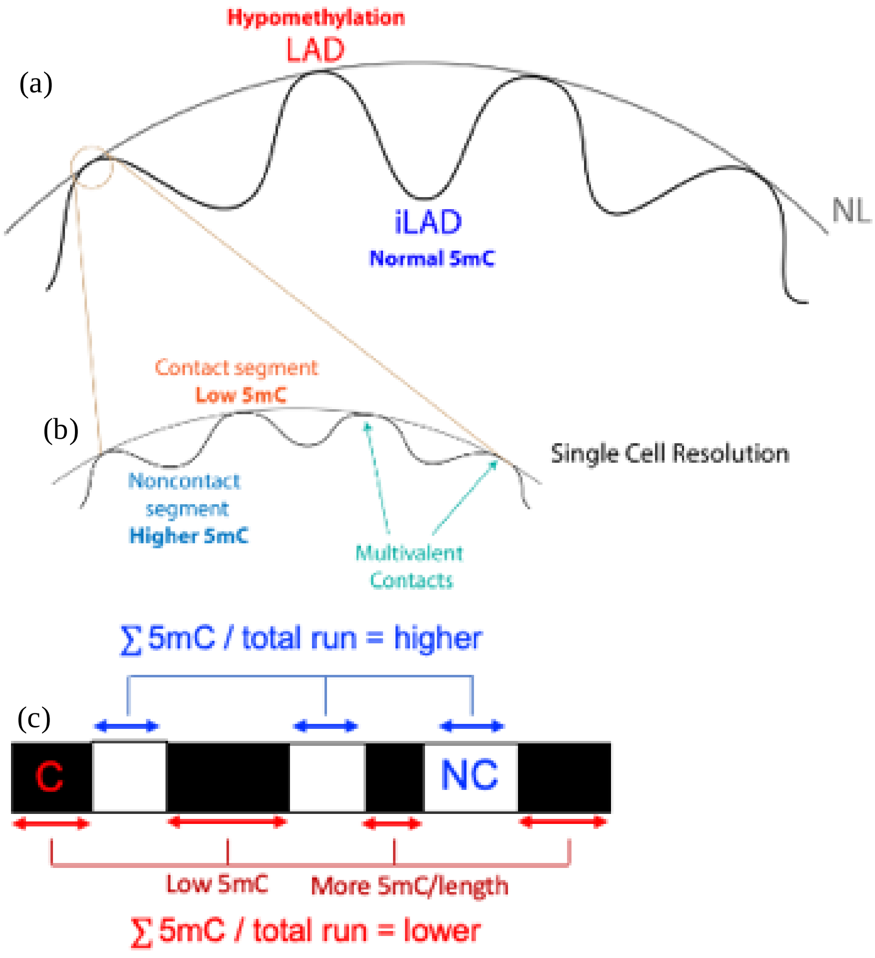


Fig. 47 | Proposed schematic of noncontact segment phenomena in LADs **(a)** Hypomethylation occurring at genome-NL contacts, LADs, and normal methylation at regions away from nuclear lamina, iLADs. **(b)** Within LADs there are segments of contact that contain less 5mC than segments of noncontact of the same size. **(c)** Higher concentration of 5mC is in LAD noncontact segments per Mb, and longer segments have less 5mC per Mb.

Considering noncontact segments in LAD regions appear at the single-cell level, the understanding of the spatial organization of genetic information within the nucleus and its links with methylation is expanded. A conventional image of LADs and iLADs may be thought of as what is depicted in **Fig. 47a**, and with simultaneous measurement of genome-NL contacts and 5mC within the same cell, it is understood LADs have methylation loss compared to iLADs. However within these LAD regions, determined by an OE score greater than one in the bulk context, single cells display variability in contact, each having unique contact runs and segments of noncontact **(Fig. 47b)**, creating multivalent contacts to anchor the region to the NL. With simultaneous epigenetic measurements, greater 5mC density was observed within the noncontact segments than in the contact segments (**Fig. 47c**, summing the 5mC), which can be thought of as smaller "subLADs", since they display the same hypomethylated trend as their larger parent LAD. Long segments of contact have less 5mC per Mb than short segments, representing the extent of methylation loss in a bulk defined LAD region is dependent on the lengths of the continuous stretches of contact observed at the single-cell level.

IV. Decitabine to Reorganize the Genome

A. Schematic and DAC dosage studies

Considering the correlation between 5mC and genome-NL contacts in CML cells at the single-cell level, these marks could be functionally linked, or the observation could be a consequence of other epigenetic components. To determine if there is a direct connection, two scenarios could be tested: (1) changing the 5mC and observing if it produces a change in the genome organization, and (2) disrupting the genome organization to change the 5mC levels. Scenario 1 was implemented using an FDA approved chemotherapy drug, decitabine (Dacogen, DAC), to globally decrease 5mC levels in KBM7, with the hypothesis that this

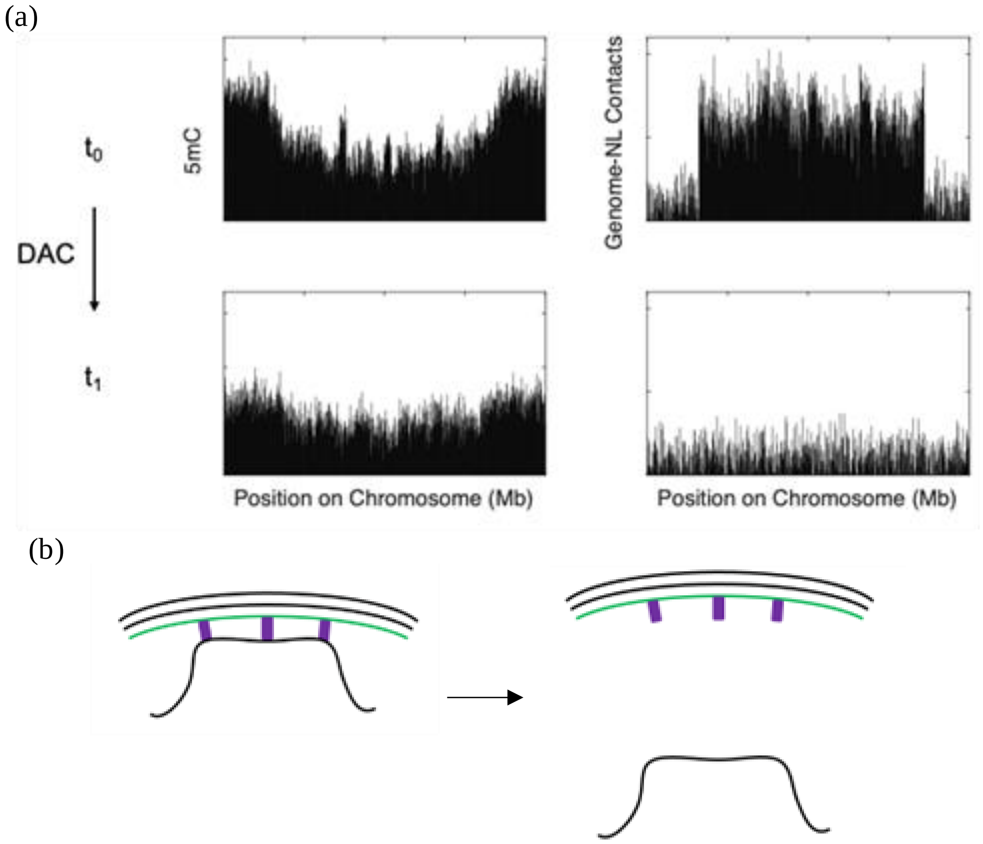


Fig. 48 | Decrease Global 5mC with Decitabine to Reposition LADs (a) Hypothetical 5mC and genome-NL contact profiles initially at t_0, DAC is added, and t_1 is after changes to the epigenome occur. **(b)** Repositing of LAD to nuclear interior.

would impact the genome organization and reposition the LADs if a connection exists. Decitabine is a cytosine analog that incorporates itself into DNA to irreversibly sequester DNA methyltransferases (DNMTs) thereby inhibiting DNA methylation, and is prescribed in chemotherapy to treat myeloid leukemia[112]. Upon addition to cell growth media, decitabine would be expected to globally demethylate the epigenome via passive demethylation, from the usual 60-80%[9] to a significantly lower percentage **(Fig. 48a, left)**. Once the methylation is lost, if hypothesis that the epigenetic features are functionally linked holds true, the domains would reposition **(Fig. 48a, right)**. How they would reposition is the greater question; at first one may imagine all the heterochromatin would decondense to euchromatin and relocate to the nuclear interior, however this idea isn't realistic due to the enormous amount of genetic material that must be compacted in order to fit within the nucleus. Therefore it may be best to envision the response as something similar to searching for a parking spot in a densely packed lot; as soon as spaces become available, they will be promptly filled, and thus it will create a situation more along the lines of a genome reorganization, rather than all of the LADs, covering 40% of the genome[22], becoming detached. This reorganization could create a profound effect on expression and could be thought of as switching from a minor to a major musical chord, in that 2/3 of the tones (contacts) remain unchanged, and the one that is shifted by the smallest interval creates a dramatic change in the expression, from sad to happy.

Initial studies tested the effect of decitabine dosage on KMB7 cell viability and proliferation, in order to determine the optimal dosage to use in DamID sequencing experiments. Cells were dosed with serial dilutions of either DAC in DMSO or DMSO only as a control, for three days, and measured for cell count and viability on day three. Cells

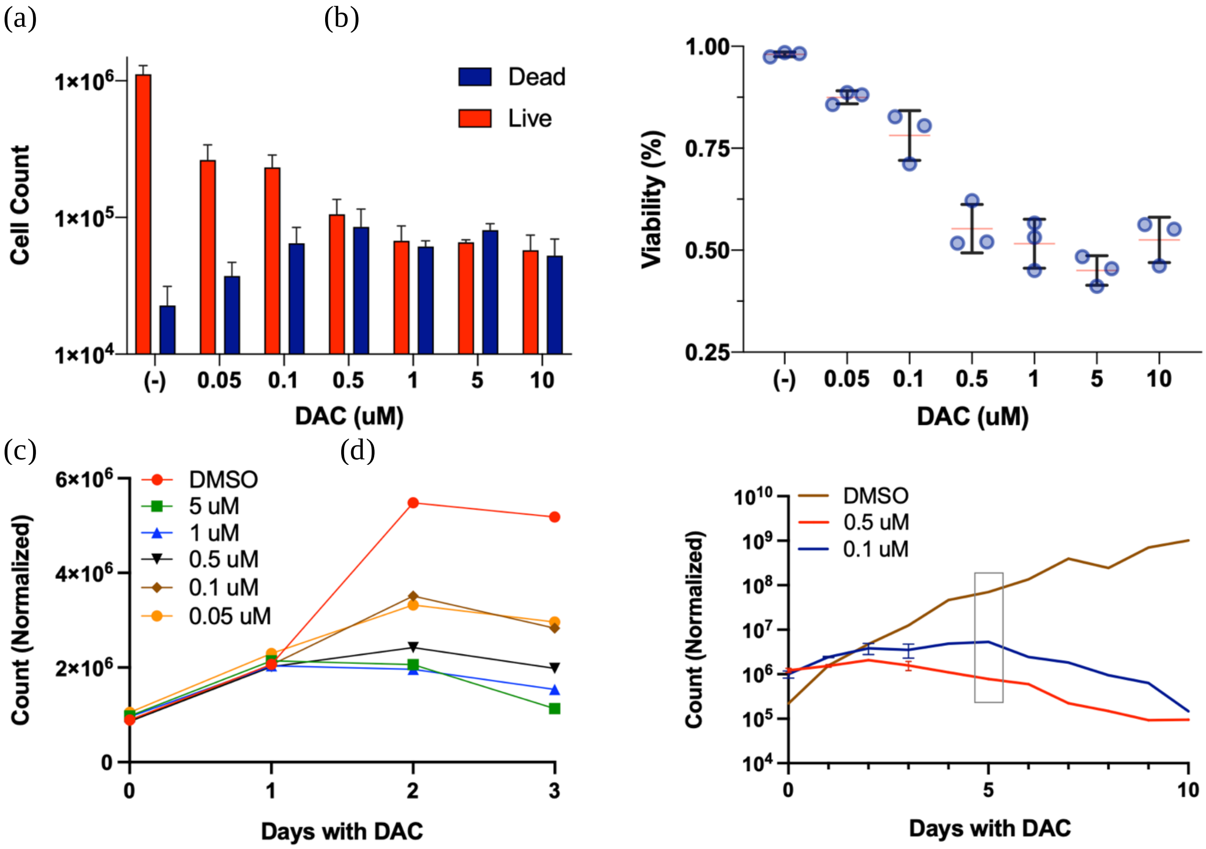


Fig. 49 | Decitabine dosage affects KBM7 viability and proliferation. (a) Cell count and **(b)** viability in response to DAC treatment. Measured at the end of 3 days. **(c)** Cell proliferation in response to DAC treatment, measured daily for 3 days and **(d)** 10 days.

viability decreased, relative to the control, in a dose dependent manner and began to level out when DAC was greater than or equal to 0.5 μM **(Fig. 49a,b)**. This suggested that the optimal dosage, the minimum dosage needed to produce an effect on the epigenome, would likely be 0.1-0.5 μM. Higher than this dosage would likely lead to increased cell death, without increased global demethylation. The experiment was repeated, omitting the unnecessarily high 10 μM dose, and cells were counted each day to determine how soon the drug starts to influence cell viability. Cell viability was affected by the DAC starting on day 2 **(Fig. 49c)**, seen by the DMSO condition cells starting to proliferate at a faster rate, with the cell count each day changing in a dose dependent manner. A longer-term study was performed, utilizing a finer tuned 0.1-0.5 μM DAC range, to determine if cell growth would become arrested, since it is known that this drug can cause cell cycle arrest at the G1 and G2/M phases[113]. Cell count was measured daily for 10 days, and it was found that cells stop growing after 5 days in culture **(Fig. 49d)**, informing us to limit their DAC exposure to 5 days, or to start with a high seeding density to account for cell cycle arrest for there to be sufficient cells for sequencing. Genomic DNA (gDNA) was harvested from cells in each of these conditions and DamID was performed in bulk to determine if this drug, known for global demethylation, was capable of repositioning LADs.

B. Global Demethylation Triggers LAD Repositioning

Since decitabine is known to globally demethylate the epigenome, cells that have been exposed to this drug should have a reduction in their 5mC levels compared to a non-drugged control. 5mC percentage at CpG sites was measured in bulk using 5mC+DamID version 2.0 **(Fig. 50, top)**, displaying a reduction in the 0.5 μM DAC condition. Compared to the control treated with DMSO for three days **(Fig. 50, bottom)** that contained a higher percentage of

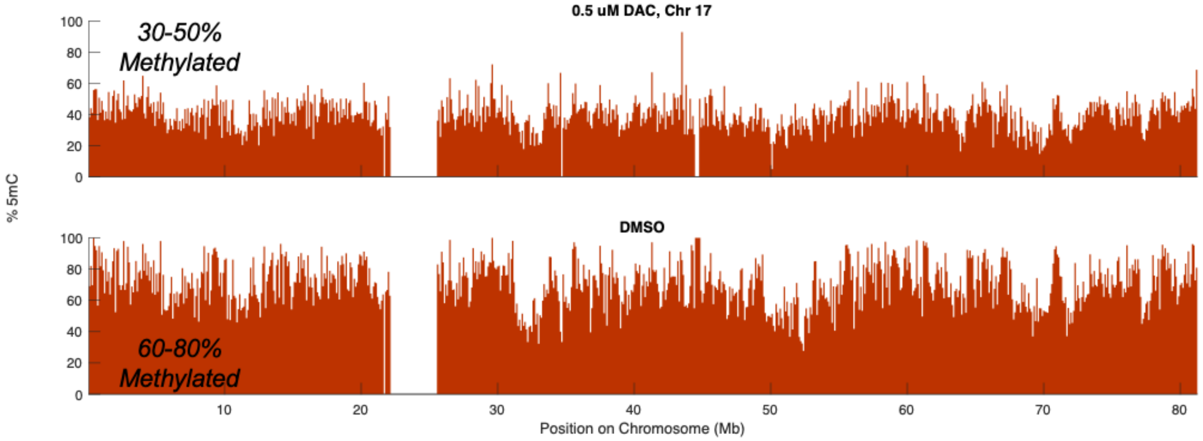


Fig. 50 | Decitabine Reduces Global 5mC Levels. 5mC profiles of three day 0.5 μM DAC (top) or DMSO treated (bottom) KBM7. 5mC measured in bulk with 5mC+DamID v2.0

methylation, the significant 5mC reduction with the DAC suggests the drug incorporated itself into the DNA to irreversibly sequester the DNA methyltransferases.

Given the significant reduction in 5mC levels, the DamID method (similar to the v1.0 method, omitting MspJI) was used to measure the genome-NL contacts in response to the global demethylation to determine if 5mC reduction is capable of triggering genome reorganization. Three characteristic changes in LADs were observed across the chromosomes, which can be described as “sinking”, “eroding”, and “cracking” **(Fig. 51a)**. “Sinking” is when a genome-NL contact decreases in contact score relative to the control, which represents a decrease in the average contact score of the bulk population, suggesting fewer cells are making the contact. This is equivalent to a reduction in contact frequency if it were measured at the single-cell level, and overall, there are less marks being placed on the genomic region by the Dam-LaminB1 fusion protein. “Eroding” is when the side of a LAD moves away from the NL, analogous to when ocean bluffs erode. “Cracking” is when just part of the LAD moves towards the nuclear interior, which could indicate a weak point destabilizing and detaching. Using the DMSO control to determine the LAD coordinates, the same LAD regions in the control condition reached higher average contact scores than in the DAC condition **(Fig. 51b)**, suggesting there were more attachments to the NL at LAD regions in the control. Correspondingly, LADs in the DAC condition reached higher noise in contact **(Fig. 51c)**, representing more variability along the profile, which could occur with loss in contacts, such as “sinks”, “erosion”, and “cracking”. Parallel lines were drawn to encompass 90% of the data in both of these plots **(Fig. 51b,e)**, and the 10% of LADs having the highest contact score in the control condition were in fact LADs in the DAC condition that had high contact noise **(Fig. 51c)**, suggesting certain LADs are breaking contact in

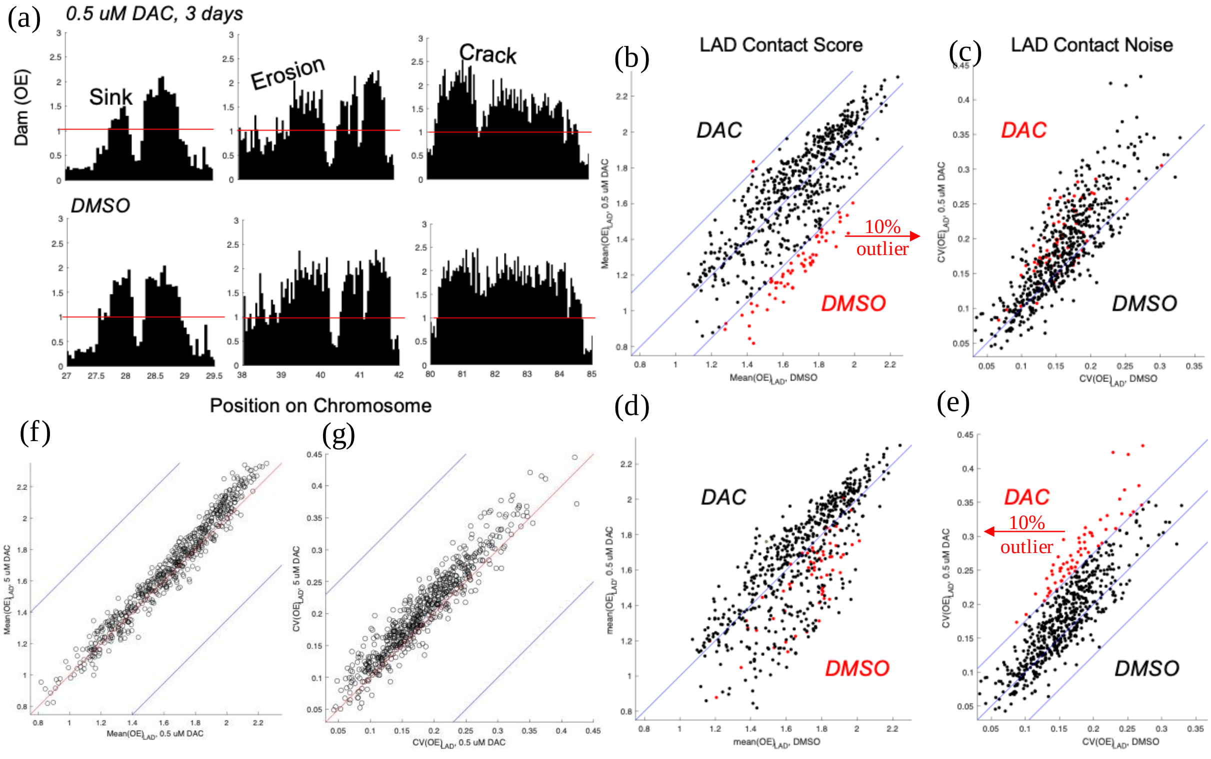


Fig. 51 | Global Demethylation Triggers LAD Repositioning in Bulk. (a) Comparison between OE scores of the same locations in DAC and DSMO conditions. LADs in DAC treated cells appear to "sink", "erode", or "crack". **(b)** Average or **(c)** coefficient of variation of OE score of the same LAD in 0.5 µM DAC and DMSO treated condition, each point is a LAD. Equidistant parallel lines from y = x drawn to encompass 90% of data of (b). Red points correspond to same LADs between (b,c). **(d)** and **(e)** are same data points as (b) and (c), except parallel lines to encompass 90% of data drawn from (e). **(f)** Average or **(g)** coefficient of variation of OE score of the same LAD in 5 µM and 0.5 µM DAC treated condition, each point is a LAD.

response to the DAC treatment. Likewise, the LADs in the DAC condition that had the highest noise along their contact, corresponded with the same LADs in the DMSO control condition that had higher contact scores **(Fig. 51d)**, further validating certain LADs are experiencing repositioning likely due to the phenomena of sinking/eroding/cracking. A similar correlation was additionally observed when comparing 0.5 µM DAC to 5 µM DAC treated cells, where the LADs in the higher DAC concentration condition had more contact noise along their profile **(Fig. 51g)**, however there wasn't a significant deviation from the identity line when comparing OE scores **(Fig. 51f)**. Since low values and baselines are subject to more variability, the average OE score of each LAD was compared to the CV(OE)

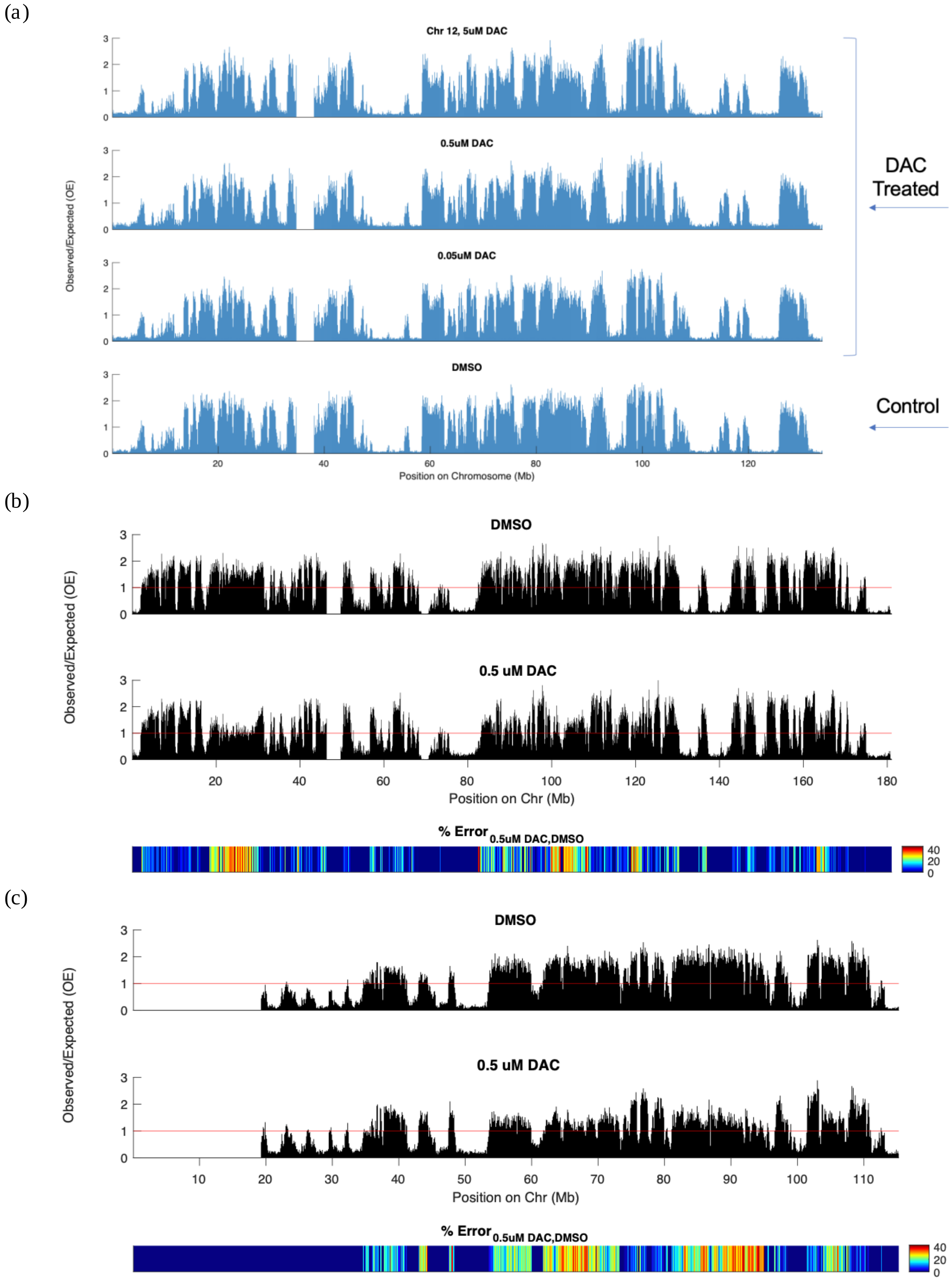


Fig. 52 | DAC induced LAD repositioning in along chromosome. (a) Bulk DamID profiles of KBM7 treated with 5 μM, 0.5 μM, 0.05 μM DAC, or DMSO. Chr 17, normalized as OE. **(b)** Dam profile of DMSO and 0.5 μM DAC treated cells, with heatmap of percent difference between OE scores along chromosome, for chromosome 5 and **(c)** chromosome 13.

of that same LAD, and it was confirmed that these parameters do not display a correlation (data not shown). Since the CV was not a strong function of the mean, it suggests the observed correlations are not due to fluctuations, such as LADs with low levels of signal producing the contact noise along the chromosomes. These results suggest there are changes in contact with the addition of DAC, characterized by LADs frequently having lower OE and higher noise along the chromosome profile relative to the control condition.

Through DamID on DAC treated bulk KBM7 samples, changes to genome-NL contact profiles were observed relative to the DMSO control. The changes reflect reorganization of the genome, rather than all the LADs moving towards the nuclear interior. Plotting the contact score verses the position on the chromosomes revealed some of the changes **(Fig. 52b,c)**, however the repositioning was best seen by computing the percent difference in OE score between corresponding bins of the DAC and DMSO condition, omitting all of the

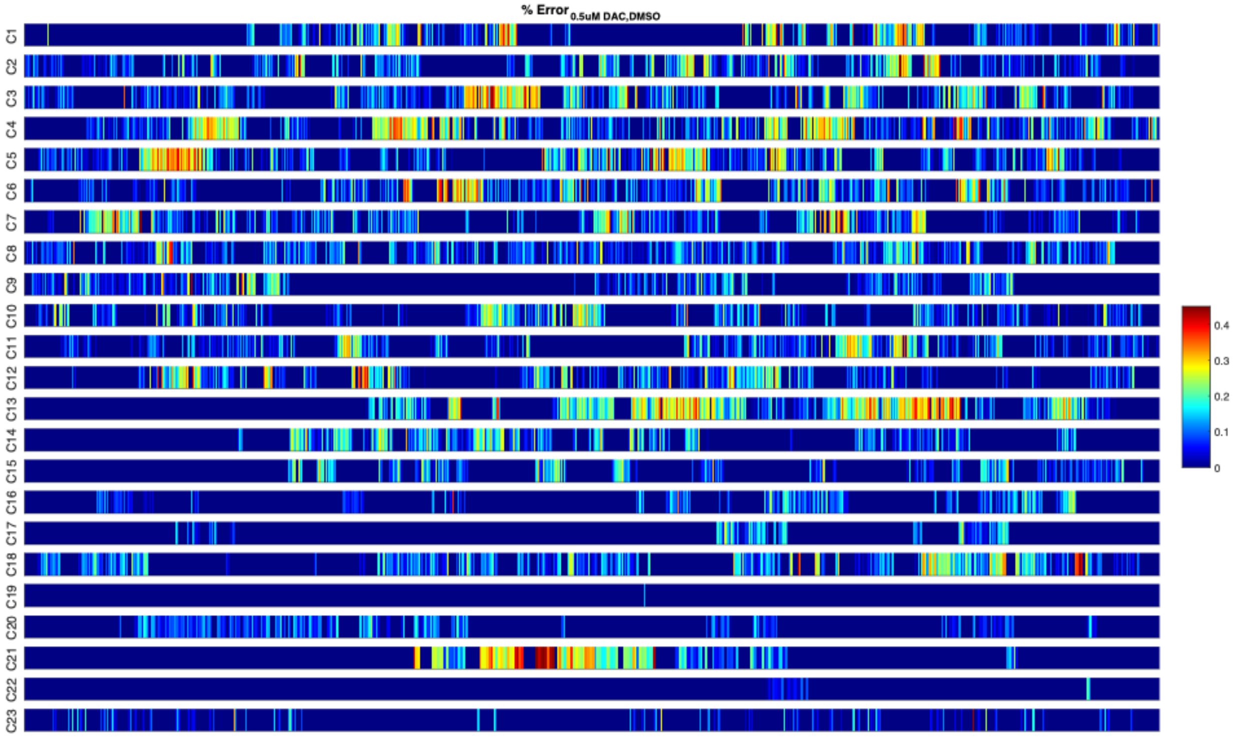


Fig. 53 | LAD repositioning in each chromosome. Heatmap of percent difference between 0.5 μM DAC and DMSO OE scores along each chromosome.

iLAD and centromeric regions in the analysis. On the heat map, there were many regions displaying significant changes, in red and yellow, with up to 40% difference in contact score in response to the DAC demethylating agent. The DAC caused LAD repositioning in Chr 5, observed by a sinking at 20-30 Mb to below under OE = 1, an erosion on the left and right sides of the crack just past 100 Mb, and an erosion on the left side of the LAD at 120 Mb **(Fig. 52b)**. In Chr 13, there was LAD repositioning in the form of sinking near 65-70 Mb, where the crack in the LAD at 67 and 70 Mb dropped below OE = 1, a small and wide cracking of the LAD at 82.5 and then 85 Mb, and an overall sinking between 85 to 95 Mb **(Fig. 52c)**. LAD repositioning occurs in each chromosome **(Fig. 53)** and those with major changes were highlighted in the heat map. Chr 17 **(Fig. 52a)** did not have major repositioning as with the other chromosomes.

To better understand the effects of DAC, specific genes that repositioned to the nuclear interior were examined. On Chr 12, a crack appears in the LAD at 83.1-83.2 Mb **(Fig. 54a)**, which corresponds with the *TMTC3* gene, and there is an erosion to the nuclear interior on the right side of the LAD at 89.3-89.4 Mb, which corresponds to the gene *KITLG*. The gene *TMTC3* is not directly cancer related, although it is involved in the control of ER stress response (*TMTC3 Gene, GeneCards*). However, *KITLG* is essential in regulating cell survival and proliferation, hematopoiesis, and migration, suggesting it is a cancer related gene that may be dysregulated in this cell line as cancer progresses (*KITLG Gene, GeneCards*)[114]. Its repositioning suggests the gene may have been abnormally expressed or suppressed in the cancer state, and it could have been anchored to the NL because it was close to a LAD, acting as a straggler. By adding DAC, the expression was corrected because it was weakly attached. Another example of a gene that repositioned to the nuclear interior

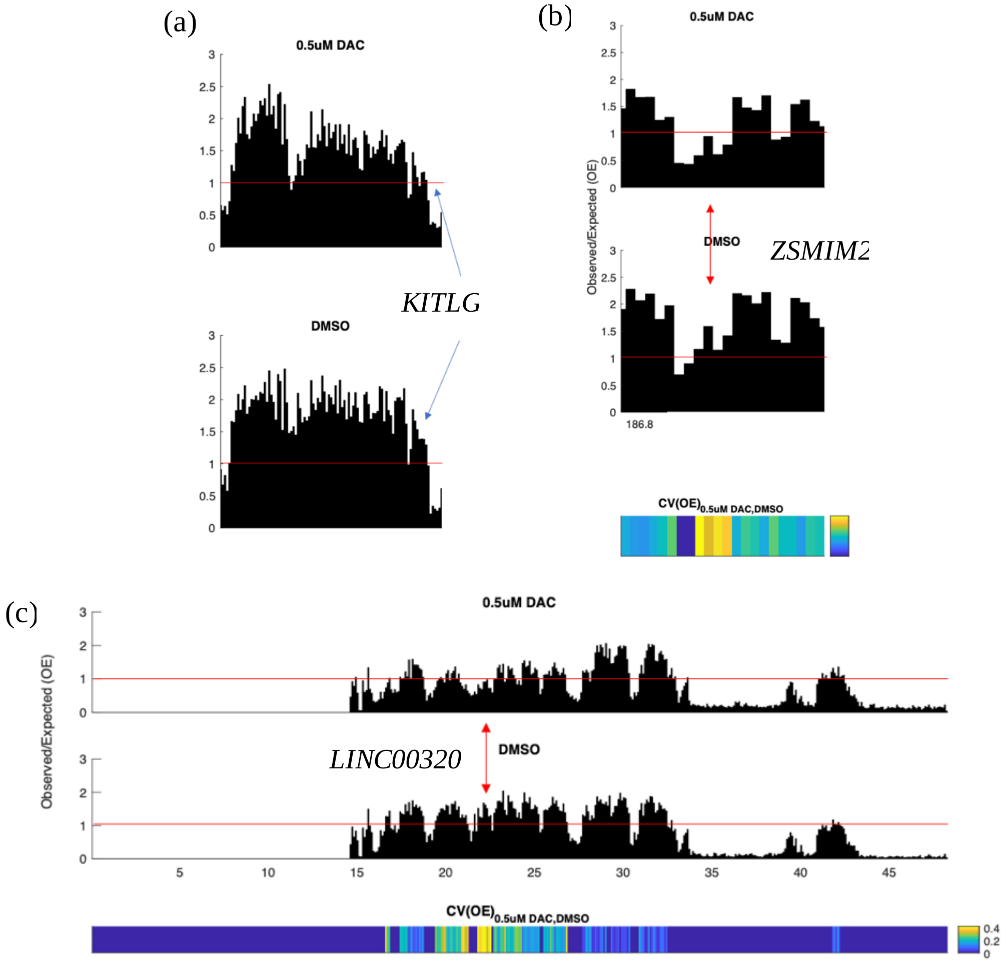


Fig. 54 | A few genes that relocated to nuclear interior upon DAC treatment. (a) *KITLG* on chromosome 12, 89.3-89.4 Mb, gene "erodes" alongside *TMTC3* "cracking" at 83.1-83.2 Mb. **(b)** *ZSWIM2* on chromosome 2 "sinks", 187.6-187.7 Mb. **(c)** *LINC00320* on Chr 21, gene "sinks".

was *LINC00320* on Chr 21 (**Fig. 54c**), which is tumor suppressive lncRNA, known to down-regulated in glioma tissues[115]. Considering it is harbored at the nuclear lamina before DAC is added, it may represent a gene that is supposed to be expressed at the nuclear interior in a normal cell, but it is turned off in a cancer cell. Another example of a repositioned gene is *ZSWIM2* on Chr 2 **(Fig. 54b)**, which is involved in regulating apoptosis, and DAC treatment relocated this gene to the nuclear interior for expression (*ZSWIM2 Gene, GeneCards*).

Given that the global demethylation caused loss of contact in LADs, contact run analysis was performed to measure how much contact was being lost and if there were any additional factors that contributed to why certain LADs had lost contact. Of the 621 LADs present in non-DAC treated KBM7, 249 of these LADs (or 40.1%) lost contact with the NL when treated with 0.5 μM DAC **(Fig. 55a)**. Some of these LADs lost up to 40% of their contact, many losing 15-20%, **(Fig. 55b)** and some even lost most of their contact. The loss of contact was dose dependent, in that higher noncontact fractions were frequently observed in the same LAD when the higher concentration of 5 μM DAC was used, as opposed to 0.5 μM of DAC **(Fig. 55c)**. LAD contact score was measured between the DAC treated condition and the control, showing a greater change in contact occurred for fewer LADs, and this tend levelled off after a 40% higher contact score in the DAC condition **(Fig. 55d)**. Correlations can be drawn along this curve, such as 10% of LADs had a 30% higher contact score in the control than in the DAC condition. LAD properties including LAD length, contact frequency, and contact score were further correlated with the amount of repositioning. The noncontact fraction in the 0.5 μM DAC condition at LAD regions defined by the LAD coordinates of the DMSO control, was measured as a function of LAD length, contact frequency, and contact score **(Fig. 55e-g)**. The ratio of change (DAC noncontact fraction > 0) to no change (DAC noncontact fraction = 0) (depicted in **Fig. 55h**), was used to determine the percentage of nonchanging LADs in response to global demethylation **(Fig. 55i-k)**. Longer LADs were more likely to undergo some changes than shorter LADs, most likely because there are more sites that could reposition, meaning there are more potential regions for cracks to occur, for example. Using contact frequency measurements from scDamID data, consistent LADs did not tend to reposition when DAC was added, however more variable domains were likely to

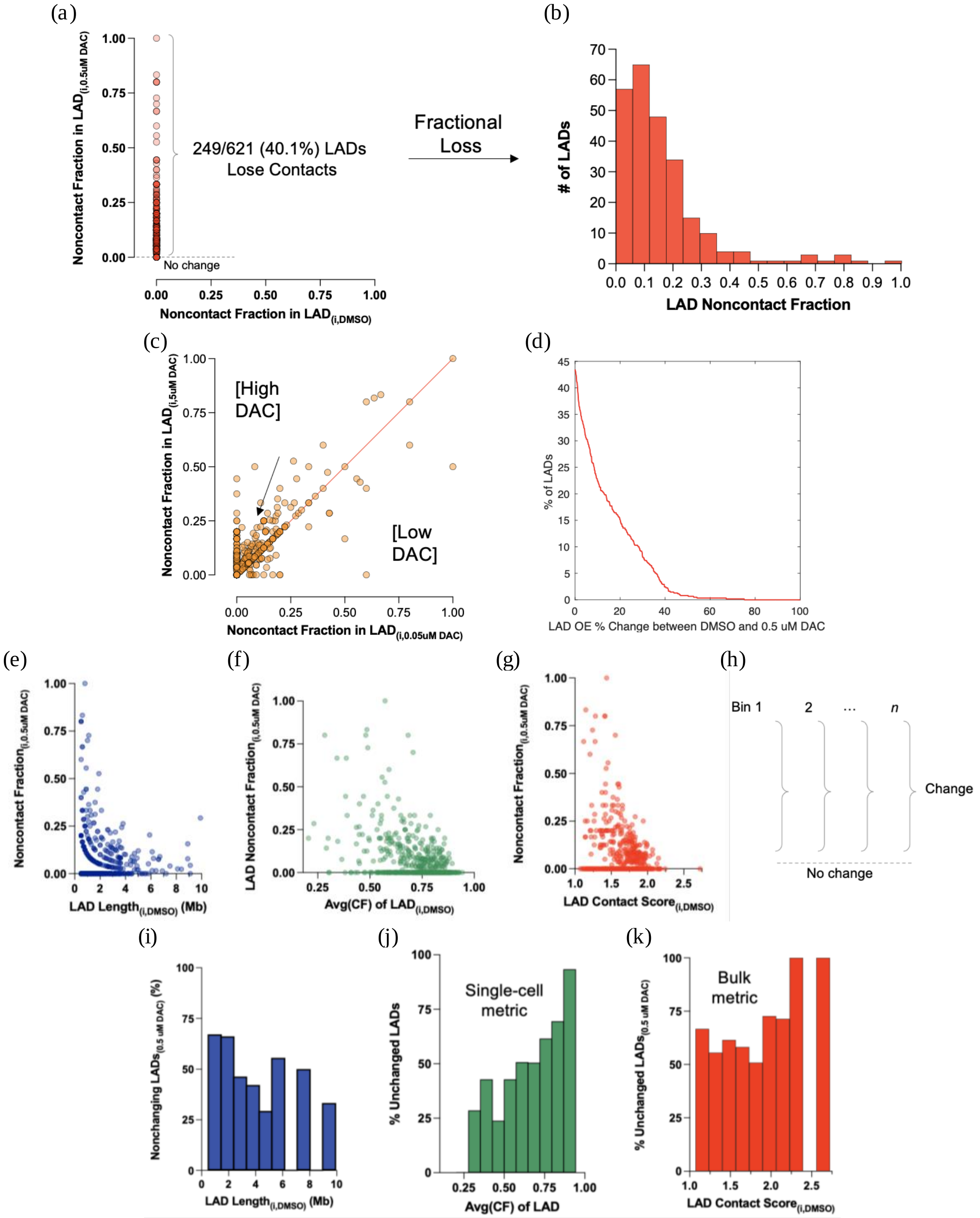


Fig. 55 | Global Demethylation Causes Loss of Contact in LADs. (a) Noncontact fraction in LAD in 0.5 uM DAC condition vs the noncontact fraction of the same LAD in the DMSO condition. **(b)** Number of LADs with a certain noncontact fraction. **(c)** Noncontact fraction of LADs in 5 uM vs. 0.05 uM DAC treated cells. Each data point is the same LAD. **(d)** Percentage of LADs displaying certain OE percentage changes between 0.5 uM DAC and DMSO treated cells. **(e)** 0.5 uM DAC treaded cell LAD noncontact fraction as a function of the LAD length, **(f)** average contact frequency of the LAD, and **(g)** the LAD contact score of the corresponding LAD in the DMSO treated cells. **(h)** Method for calculating nonchanging LAD percentage. **(i)** Percentage of non-repositioning LADs in the 0.5 uM DAC condition as a function of the LAD length, **(j)** the average contact frequency of the LAD, and **(k)** the LAD contact score of the same LAD in the DMSO condition.

reposition to the nuclear interior. This result elucidates the definitions of consistent and variable LADs, in that consistent LADs may overall create stronger attachments to the nuclear lamina across all of the cells than variable LADs. This suggests when all LADs are subject to perturbation by global demethylation, the variable LADs will most often lose contact, and crack first per se. Likewise, a similar trend was observed with contact score and percentage of unchanged LADs; LADs with lower average contact scores tended to change more than LADs with much higher contact scores that did not change. Since it takes an average of the whole LAD's contact score, it does not consider the variability in the contact score, which would indicate if there were weaker regions that would be more prone to cracking. This likely explains why the contact frequency metric reached 25% unchanged LADs at low contact frequency verses the contact score metric only reached 50% unchanged LADs at low contact scores. This further supports the need for single-cell measurements, such as contact frequency, over bulk measurements, such as OE score, since they can reveal stronger correlations because LAD changes are occurring at the single-cell level.

C. Activating and Repressing Histone Modifications and DAC LAD Repositioning

Besides LAD metrics, there are other contributors to why certain LADs reposition, such as histone modification. Publicly available KBM7 ChIP-Seq data[108,109] was used to determine the coordinates of activating and repressive histone modifications, which were then compared to the locations of contacts and newly formed noncontacts within DMSO called LADs in cells that were treated with DAC. The amounts of histone modifications at each genomic bin were measured against how much that location's contact score changed when DAC was added **(Fig. 56)**. For activating histone modifications (H3K4me1, H3K4me2,

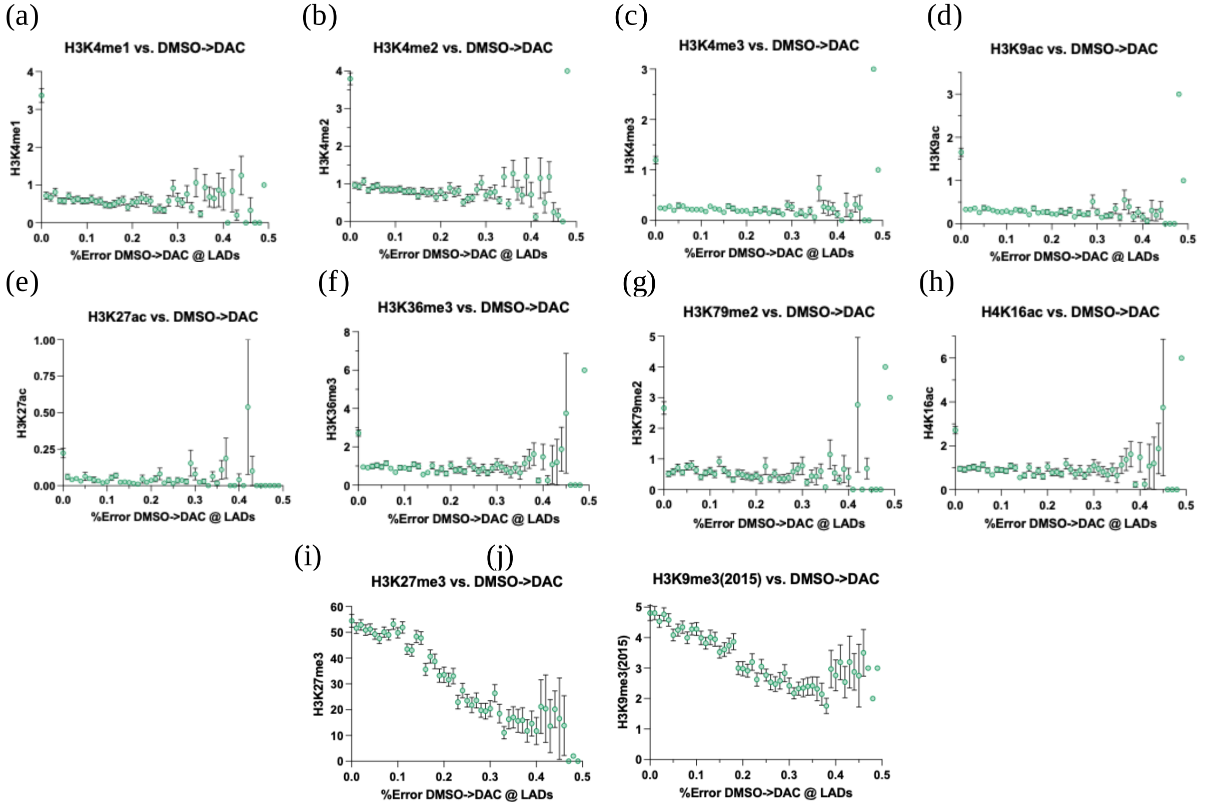


Fig. 56 | Activating and repressing histone modifications and DAC LAD repositioning. (a) Amount of H3K4me1, **(b)** H3K4me2, **(c)** H3K4me3, **(d)** H3K9ac, **(e)** H3K27ac, **(f)** H3K36me3, **(g)** H3K79me2, **(h)** H4K16ac, **(i)** H3K27me3, and **(j)** H3K9me3 as a function the percent change between the genome-NL contact OE score in LADs between 0.5 uM and DMSO treated cells.

H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K79me2, H4K16ac), the trend was subtle, however LADs that had repositioned the most, represented by 30-40% difference in contact score between the DMSO and DAC conditions, tended to have more activating histone modifications **(Fig. 56a-h)**. The repressive histone modifications (H3K27me3 and H3K9me3) had a much more significant trend (anticorrelation) with the DAC mediated LAD repositioning than the activating histone modifications **(Fig. 56i,j)**. LADs that didn't reposition upon global demethylation had high levels of repressive marks, and those that did reposition had much fewer of these marks. H3K9 methylation is known to be associated with gene repression and NL tethering, which was seen by these data. When many H3K9me3 marks were present, there was a lack of LAD repositioning, however fewer of these marks were associated with less genome-NL contacts remaining after DAC treatment. These results suggest that histone modifications do contribute to which parts of LADs reposition, which helps to elucidate the bigger picture of how multiple components of the epigenome work together to influence gene expression.

D. Global Demethylation Causes Gains of Contact in iLADs

Similar to how LADs repositioned to the nuclear interior in response to the DAC treatment, iLADs also repositioned from the nuclear interior to the nuclear periphery, suggesting a genome wide rearrangement. Global demethylation by DAC treatment caused iLADs to gain contacts, and 21.7% of iLADs had regions that moved towards and contacted the nuclear lamina compared to the DMSO control **(Fig. 57a)**. Of these repositioned iLADs, many gained 10-20% contact with the nuclear lamina, and a significant amount gained 40% contact **(Fig. 57b)**. The properties of iLADs, such as the length, average noncontact frequency, and contact score, mostly correlate with the amount of repositioning towards the nuclear lamina, however, the correlations are not as strong as those observed with the LAD regions moving towards the nuclear interior **(Fig. 55a,b)**. The ratio of change (DAC contact fraction > 0) to no change (DAC contact fraction = 0) (depicted in **Fig. 57c-e**), was used to determine the percentage of nonchanging iLADs in response to global demethylation. Typically, the longer iLADs were more likely to experience some change, gaining contact, compared to the smaller iLADs, which were more likely to remain unchanged **(Fig. 57f)**. iLADs known to always be situated away from the nuclear lamina, those having an average noncontact frequency close to 1, were unchanged by the global demethylation, and did not

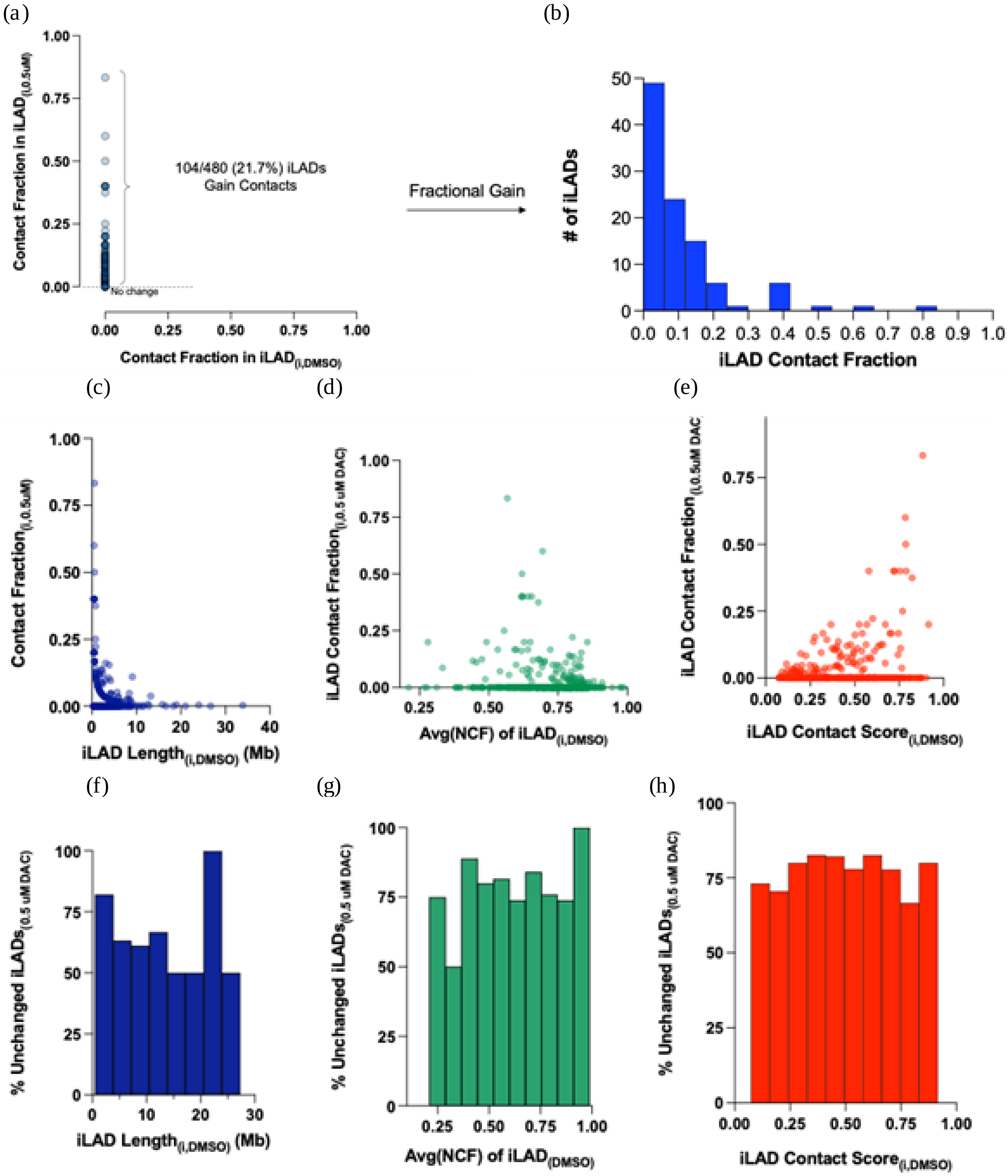


Fig. 57 | Global demethylation causes gain of contact in iLADs. **(a)** Contact fraction in iLADs in 0.5 uM DAC condition vs. the contact fraction of the same iLAD in the DMSO condition. **(b)** Number of iLADs with a certain contact fraction. **(c)** 0.5 uM DAC treaded cell **i**LAD contact fraction as a function of the iLAD length, **(d)** average noncontact frequency of the iLAD, and **(e)** the iLAD contact score of the corresponding iLAD in the DMSO treated cells. **(f)** Percentage of non-repositioning iLADs in the 0.5 uM DAC condition as a function of the iLAD length, **(g)** the average noncontact frequency of the iLAD, and **(h)** the iLAD contact score of the same iLAD in the DMSO condition.

gain contacts with the nuclear lamina **(Fig. 57g)**. iLADs that contained regions that sometimes contact the nuclear lamina (average noncontact frequency < 1), repositioned in response to the DAC, however it appeared to be more of binary event than dependent on the noncontact frequency of the iLAD. This could be explained by whenever a space opened at the nuclear periphery after a LAD moved to the nuclear interior, a nearby iLAD region makes contact and starts condensing into heterochromatin to prevent accumulation of uncondensed DNA in the cell nucleus. By this rationale, iLADs display a binary preference for reorganization; either most of their domain was away from the nuclear lamina (NCF = 1) and they won't reposition, or when some regions of iLADs were close to the nuclear lamina in an ensemble of single cells (NCF < 1), repositioning occurs if space to harbor becomes available. iLAD contact score had no strong correlation with iLAD repositioning, and this is likely because contact score is a metric meant for detecting LADs and not iLADs, and is used to determine if contact is made based on how much m^{6}A was added to regions that contacted the Dam-LaminB1 fusion. Any observed/expected m^{6}A score less than 1 doesn't have considerable meaning, other than statistically, contacts weren't made at that region **(Fig. 57h)**. These results together suggest that in addition to LADs repositioning in response to DAC treatment, iLADs reposition as well. However, LAD repositioning appears to be the main event due to stronger correlations with LAD properties such as length, contact frequency, and contact score than iLADs with their iLAD properties, implying iLADs likely reposition in response to LADs repositioning.

E. Proposed Schematic of Genome Rearrangement

From the results of the DAC experiment that demonstrated a simultaneous repositioning of LADs and iLADs in response to global demethylation, a schematic was drawn to describe the full picture of what was occurring. Consider a genome arrangement before (State 1) and after (State 2) the response to the DAC treatment **(Fig. 58a)**, in which certain regions start at the nuclear periphery (drawn in red) and others start at the nuclear interior (blue). When parts (or all) of certain LADs decondense, and these newly unrepressed genes reposition towards the nuclear interior for their expression (State 2), the quantity of space that the DNA takes up in the small nucleus reaches a maximum. In order to decrease this amount back to baseline, new contacts form with the nuclear lamina, corresponding to parts of the iLADs, to limit the amount of uncondensed DNA, effectively repositioning iLADs to the nuclear periphery. The parts of LADs and iLADs that do reposition display certain characteristics, mainly a lower contact and noncontact frequency, respectively. This is depicted in **Fig. 58b**, where genome-NL contact profiles for State 1, the (-) DAC condition, and State 2, the (+) DAC treated condition are shown. Consistent LADs (those with high contact frequencies) are less likely to have much or any of their domain moved towards the nuclear interior, represented by a low or no noncontact fraction, compared to the more variable LADs, that will lose more contact with the nuclear lamina, displaying a higher noncontact fraction. The iLADs behave similarly, in that they gain contacts with the nuclear lamina as the LADs reposition, however they don’t reposition proportionally to their noncontact frequency in the way that LADs reposition proportionally to their contact frequency. The most consistent iLADs won’t gain any contacts, however any with less than 90.9% average noncontact frequency will gain contacts, based on **Fig. 57g**. This idea of genome repositioning parallels the “Cat’s Cradle”

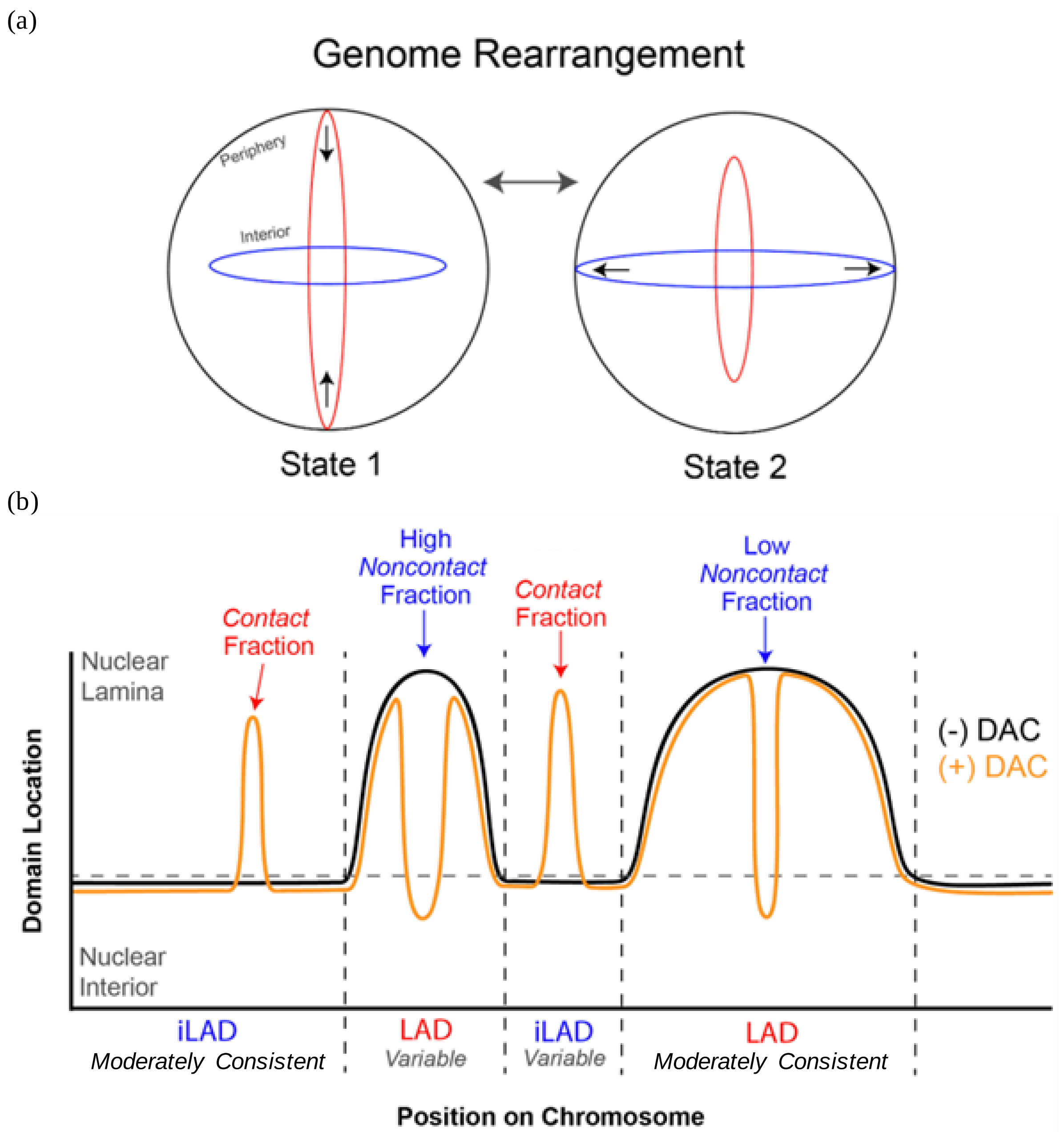


Fig. 58: Proposed schematic of simultaneous repositioning of LADs and iLADs. (a) State 1 has red region (LAD) at nuclear periphery, and blue region (iLAD) at nuclear interior. Upon addition of DAC, state 2 has red region repositions and becomes iLAD and blue region becomes LAD. **(b)** Hypothetical LAD and iLAD profiles along chromosome based on *Chapter IV* trends before and after DAC addition. Moderately consistent defined as more frequent than variable, around 90%, but not 100% consistent, as those LAD/iLADs don't tend to reposition.

model of gene expression, in which transcription and spatial organization of the genome are connected, since expressed lncRNA acting as handles for nuclear proteins to reshape the genomic architecture[3]. Although gene expression was not also directly measured in this experiment (though it is possible with the protocol, described in *Chapter V*), it is conceivable that some of the genes that repositioned to the nuclear interior for expression via global demethylation were in fact lncRNA that could further facilitate repositioning the LADs and iLADs into their final reorganized state.

V. Adding Transcriptome Measurements and Epitranscriptome

A. Schematic and Sequencing Results

For the complete picture of how epigenetic features such as DNA modification and spatial organization of the genome are correlated and influence gene expression, the transcriptome, or the ensemble of all mRNA expressed by the cell, can be measured simultaneously with 5mC and genome-NL contacts, at the single-cell level. Transcriptome measurements can be incorporated into both version 1 (DpnI + MspJI) and version 2 (DpnI + TET2/APOBEC) of the sc5mC+DamID protocol by adding a few more initial steps prior to the DNA digestion, making the technique versatile. Before adding a modified lysis buffer and sorting Shield1+, G1/S phase cells into 384 well plates, single stranded CEL-Seq barcoded adapters are added to each well. These adapters are used to capture and later amplify the mRNA of each cell. On the first day of library preparation, the protocol starts with reverse transcribing (RT) all the mRNA **(Fig. 59,1)** into cDNA (first-strand synthesis) utilizing the polyA tail endogenously present on the mRNA molecules and a complementary

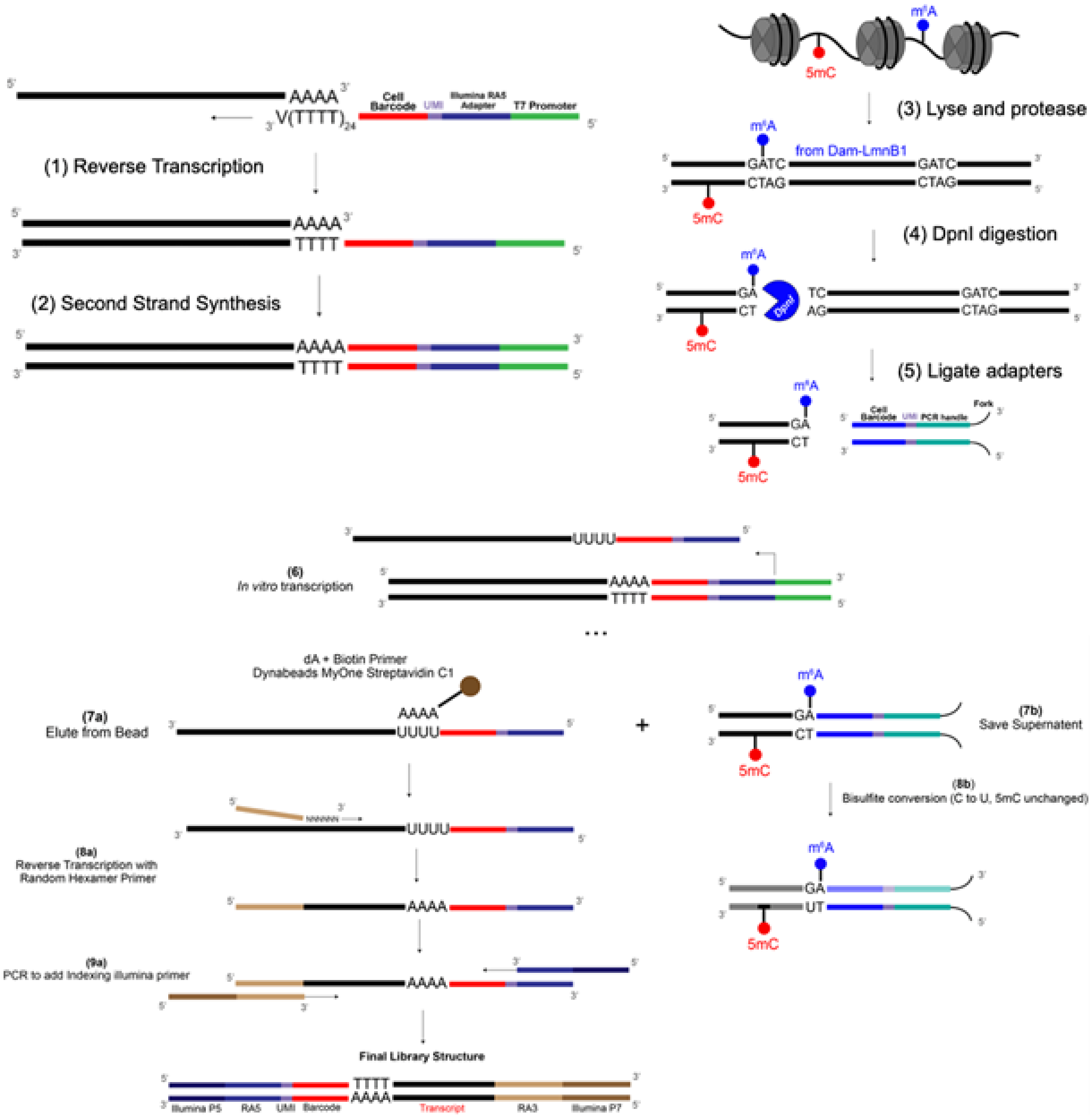
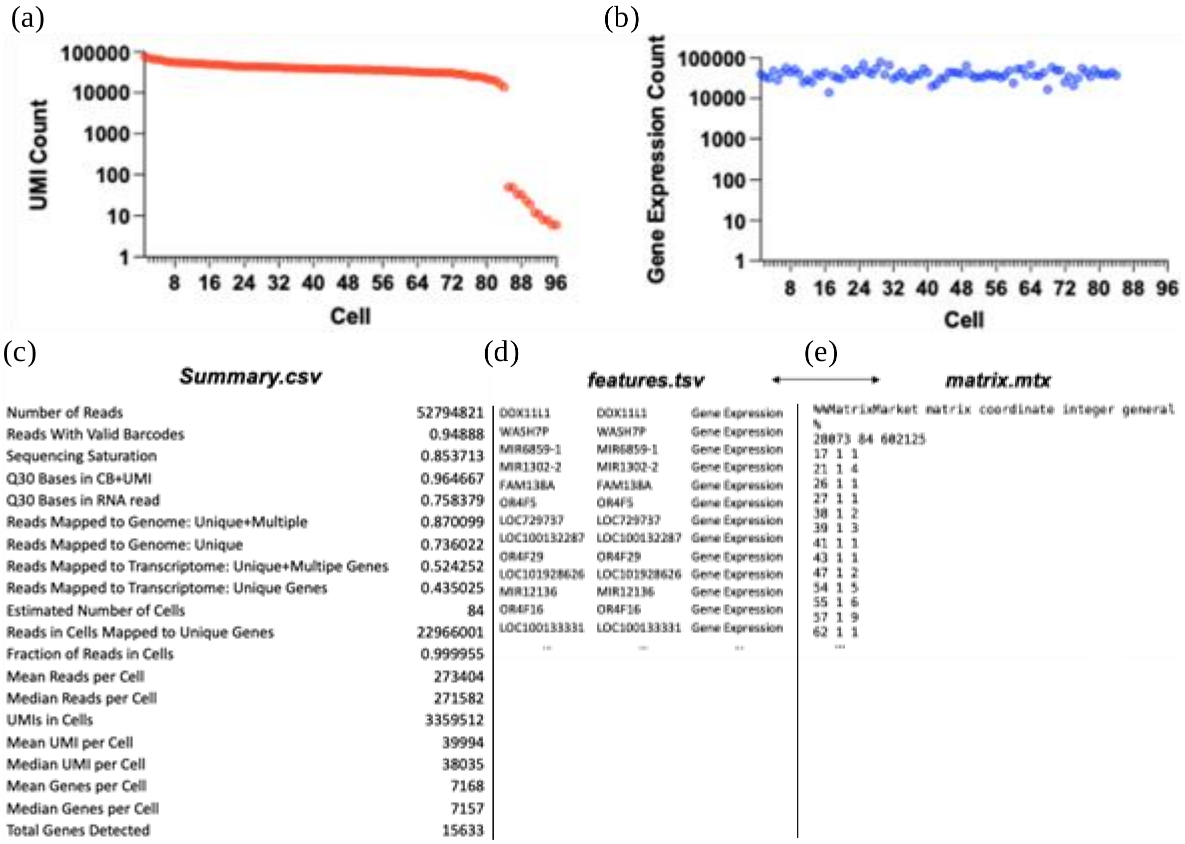


Fig. 59 | sc5mC+DamID+Transcriptome schematic. (1) mRNA molecules from the single cell are reverse transcribed using CEL-Seq barcoded adapter. **(2)** Second strand synthesis. **(3)** Lyse cells and protease nucleosomes, as in sc5mC+DamIDv1/2. **(4)** Digest DNA m^{6}A mark on GATC with DpnI. **(5)** Ligate barcoded Dam adapters. **(6)** Amplify mRNA component with *in vitro* transcription. **(7a)** Enrich for mRNA and save **(7b)** DNA flow through. **(8a)** Add RA3 Illumina adapter to mRNA with random hexamer priming. **(9a)** Amplify library adding both the 5' and 3' Illumina adapters to generate final library. **(8b)** Deamination of DNA component with TET2/APOBEC and generate final library with strategies discussed in *Chapter II* (v2.0-v2.3).

polyT tail on the CEL-Seq adapter. The attached adapter has a cell specific barcode to identify which RNA molecule originated from each cell, a unique molecular identifier (UMI) used for determining transcript counts, the Illumina RA5 adapter, which is part of 5' Illumina sequencing adapter that allows it to bind to the flow cell for sequencing, and lastly a T7 promoter to amplify the mRNA in each cell with *in vitro* transcription (IVT) to a higher concentration to meet the threshold needed for sequencing. Next, second strand synthesis (SSS) is performed to generate the full double-stranded cDNA product **(Fig. 59,2)**, protecting it from degradation by RNases at ambient temperatures that would reduce the yield if it were RNA. At this point, the protocol continues into the standard protocol, described in *Chapter II*, in which the cells are lysed to expose the chromatin, and the DNA is unwound from the histones using protease to make it accessible for enzymatic digestion **(Fig. 59,3)**. These steps can all be performed on the same day, with RT and SSS in the morning followed by protease in the afternoon, rather than only protease in the afternoon as in the sc5mC+DamID protocol, meaning up to this point the transcriptome measurement incorporation doesn't add an extra day to the protocol. The following day, the DNA is reacted with either MspJI and DpnI (v1), or just DpnI (v2) to digest the m^{6}A contact mark added to LADs by the Dam-LaminB1 fusion **(Fig. 59,4)**. For the sake of simplifying the schematic, only version 2 of the method will be depicted, however it is compatible with both versions. Double-stranded Dam adapters, containing the cell/mark specific barcode, a PCR handle for attaching the 5' Illumina adapter for sequencing, and a fork to prevent end to end adapter blunt ligations, are ligated **(Fig. 59,5)** overnight to the DpnI cut sites. The 384 well plate is pooled into 4×96 barcoded libraries, each containing all the single-cell material with unique barcodes for the DNA and RNA components. The samples are processed as bulk and IVT is performed **(Fig.**

59,6) on all the material using the T7 promoter on the CEL-Seq adapter to amplify the RNA component, while leaving DpnI cut DNA unchanged. The DNA and RNA components are then separated with an RNA enrichment strategy that uses dA+Biotin primers attached to streptavidin beads **(Fig. 59,7a)**. The dA component of the beads binds to the uracil repeat that was generated by IVT, corresponding to the polyA tail of the mRNA molecules. The mRNA derived eluate is further processed the same as the steps post IVT in the sc5mC+DamID v1 protocol, and the flow through supernatant is saved for the TET2/APOBEC conversion **(Fig. 59,7b)**. On the mRNA derived eluted material **(Fig. 59,8a)**, the Illumina RA3 adapter is attached with random hexamer priming, generating a complementary DNA strand. Exponential PCR is used to attach the complete Illumina P5 and P7 indexing primers (Fig 51, 9a), producing a final library structure containing the UMI, cell/transcriptome specific barcode, ultimately followed by the transcript sequence. The protocol is continued on the DNA component from the bead flow through, performing the enzymatic version of bisulfite conversion with TET2/APOBEC **(Fig. 59,8b)**, and then the RA3 Illumina sequence is attached with either random priming (v2.0), TACS ligation (v2.1), A-Tailing with dATP and Poly-T Priming (v2.2), or adaptase (v2.3), followed by their corresponding amplification scheme to generate the final DNA library structures.

Single-cell 5mC+DamID+Transcriptome sequencing was performed on 84 cell libraries and the transcriptome reads were aligned with Star aligner[116,117]. Analysis post alignment indicated 84 cells were detected based on UMI count, ranging from 10,000-100,000 per cell for 84 of the cell barcodes, and less than 100 for the remaining cells **(Fig. 60a)**. The aligner outputted Summary.csv, a sequencing statistics file that indicates the efficiency of mapping and sample quality, as well features.tsv and matrix.mtx, which are critical for determining



Summary.csv

Number of Reads	52794821
Reads With Valid Barcodes	0.94888
Sequencing Saturation	0.853713
Q30 Bases in CB+UMI	0.964667
Q30 Bases in RNA read	0.758379
Reads Mapped to Genome: Unique+Multiple	0.870099
Reads Mapped to Genome: Unique	0.736022
Reads Mapped to Transcriptome: Unique+Multipe Genes	0.524252
Reads Mapped to Transcriptome: Unique Genes	0.435025
Estimated Number of Cells	84
Reads in Cells Mapped to Unique Genes	22966001
Fraction of Reads in Cells	0.999955
Mean Reads per Cell	273404
Median Reads per Cell	271582
UMIs in Cells	3359512
Mean UMI per Cell	39994
Median UMI per Cell	38035
Mean Genes per Cell	7168
Median Genes per Cell	7157
Total Genes Detected	15633

features.tsv ⟷ *matrix.mtx*

```
DDX11L1        DDX11L1        Gene Expression
WASH7P         WASH7P         Gene Expression
MIR6859-1      MIR6859-1      Gene Expression
MIR1302-2      MIR1302-2      Gene Expression
FAM138A        FAM138A        Gene Expression
OR4F5          OR4F5          Gene Expression
LOC729737      LOC729737      Gene Expression
LOC100132287   LOC100132287   Gene Expression
OR4F29         OR4F29         Gene Expression
LOC101928626   LOC101928626   Gene Expression
MIR12136       MIR12136       Gene Expression
OR4F16         OR4F16         Gene Expression
LOC100133331   LOC100133331   Gene Expression
     ...            ...            ...
```

```
%%MatrixMarket matrix coordinate integer general
%
28873 84 602125
17 1 1
21 1 4
26 1 1
27 1 1
38 1 2
39 1 3
41 1 1
43 1 1
47 1 2
54 1 5
55 1 6
57 1 9
62 1 1
 ...
```

Fig. 60 | Simultaneous sc5mC+DamID+Transcriptome sequencing, statistics of RNA component. **(a)** UMI count per cell. **(b)** Total gene expression scores per cell. Same cell order as in (a). STAR output files: **(c)** Summary.csv, **(d)** features.tsv, and **(e)** matrix.mtx.

both variable and highly expressed genes across all the cells in the sample **(Fig. 60c-e)**. The matrix.mtx file has three columns: (1) the gene identity number, which corresponds to the line in the features.tsv file that has the index of the possible genes transcribed, (2) the cell barcode representing which cell the transcript originated from, and (3) the gene expression score, representing how many times that gene was expressed in a single cell. As an example, [1 2 1] would represent geneID = 1, which is DDX11L1 **(Fig. 60d)**, being transcribed by cell = 2, at a frequency = 1 time. By summing the transcript counts per cell barcode, the total gene expression count was determined per cell **(Fig. 60b)**. Each cell numbers corresponds to the same in **Fig. 60a**, indicating the genes were filtered from their original UMI count to produce the final expression count. Cell 1-84 range from 10,000-100,000 transcripts per cell, and cell 85-96 have less than 1 gene expressed per cell, confirming those cells were not present in the library.

By utilizing the matrix.mtx file, the top genes expressed across the cells were determined. This is done systematically by first determining the top 100-200 genes per cell based on gene score. Next, my algorithm counts how frequently each of those genes occur across all the cells in the data set (the number of gene intersections, representing the frequently expressed genes), and subsequently sums the scores of each gene across the cells in the data set (the magnitude of expression). The genes are then sorted by a two-level hierarchy either by (1) ordering by the number of times a gene occurs across the cells (the most intersections), followed by the sum of each gene score across the cells, or (2) sorting first by the sum of the scores for each gene across the cells, followed by the number of times each gene appears in the cells in the data set. Sorting strategy (1) allows the most frequently expressed genes to appear at the top of the list, whereas strategy (2) puts the genes with the highest total scores at the top of the list, which can reveal genes that have high expression scores in some cells, that may not be expressed in all the cells, signifying expression variability. After the genes have been ordered, the top 50-100 are selected and the gene score for each cell is plotted in a heatmap **(Fig. 61)**. To account for cells that have lower total gene expression, the score of each gene in each cell is normalized by the total (overall) expression score in that cell **(Fig. 60b)**. It is necessary to normalize by the total gene score in each cell because certain cells have lower total expression, such as cell 17 and 68, causing their expression to not be visible in the heatmap **(Fig. 61a vs. b, c vs. d)**. The different sorting strategies reveal genes such as *EEF1A1*, *PTMA*, and *PABPC1* were most common across all of the cells **(Fig. 61b)** and

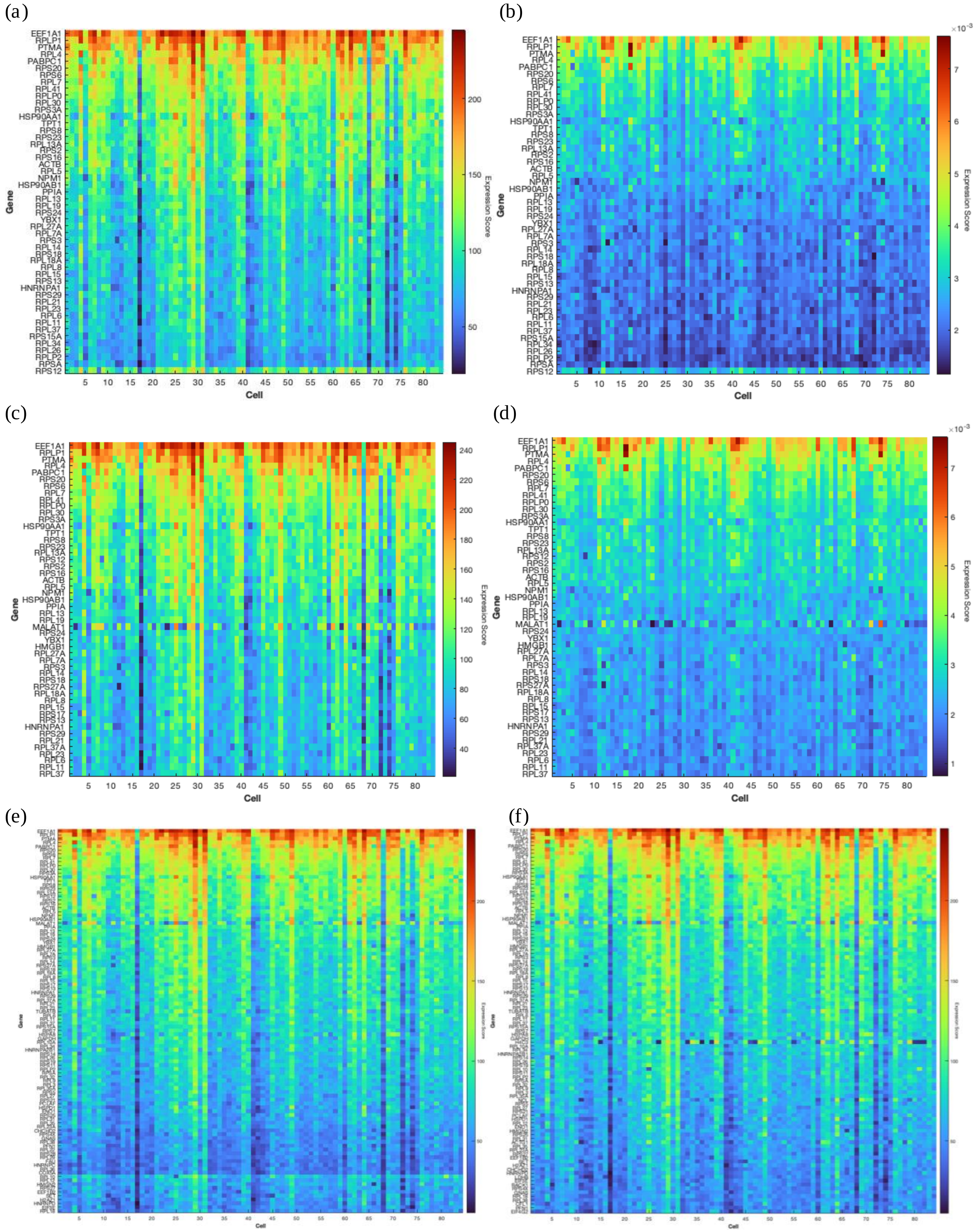


Fig. 61 | Single-cell gene expression score heat maps. (a) 100_1_50. **(b)** 100_1_50,norm. **(c)** 100_2_50. **(d)** 100_2_50,norm. **(e)** 200_1_100. **(f)** 200_2_100. For (a-f), the naming is: xx_y_zz,norm. xx = number of top genes surveyed from each cell, y = sorting strategy (1) or (2), zz = number of genes across the cells plotted, "norm" = if individual gene score is normalized by total gene expression score in that cell.

genes such as *MALAT1* and *HMGB1* had the highest overall expression across all of the cells, but they were not present in every cell **(Fig. 61d)**. Numerous ribosomal genes were frequently expressed, including those with the prefix RPL or RPS, which could additionally be filtered out of the gene list. With single-cell 5mC+DamID+Transcriptome analysis, the top expressed genes can be referenced to their coordinates on the chromosomes, and the corresponding LAD and 5mC profiles can be observed to see how they are correlated with the amount of gene expression within the same cell.

B. Epitranscriptomic and Epigenetic Links in Chronic Myeloid Leukemia Cells

In addition to the field of epigenetics that aims to understand the modifications to DNA and proteins that affect gene expression, modifications can also be made to RNA, known as the emerging field of epitranscriptomics. Features of the epigenome, such as 5mC, genome organization (seen via genome-NL contacts), and histone modifications such as the repressive marks H3K27me3 and H3K9me3, were seen to be linked in KBM7 in the DAC experiment, because global DNA demethylation repositioned primarily LAD regions with lower repressive histone marks. Therefore, it is plausible that epitranscriptomic and epigenetic links may exist as well in chronic myeloid leukemia cells to influence gene expression **(Fig. 62a)**. It has been found that m^6A marks on RNA are able to influence DNA methylation in proximity to the location of the transcription. As RNA polymerase II (Pol II) transcribes the DNA into RNA, m^6A methyltransferase METTL3 co-transcriptionally adds m^6A marks to the newly formed RNA, which is interpreted by m^6A reader FXR1, which recruits Ten-eleven translocation methylcytosine dioxygenase 1 (TET1) to the site to demethylate the 5mC **(Fig. 62b)**. It was found that RNA m^6A was anticorrelated with DNA

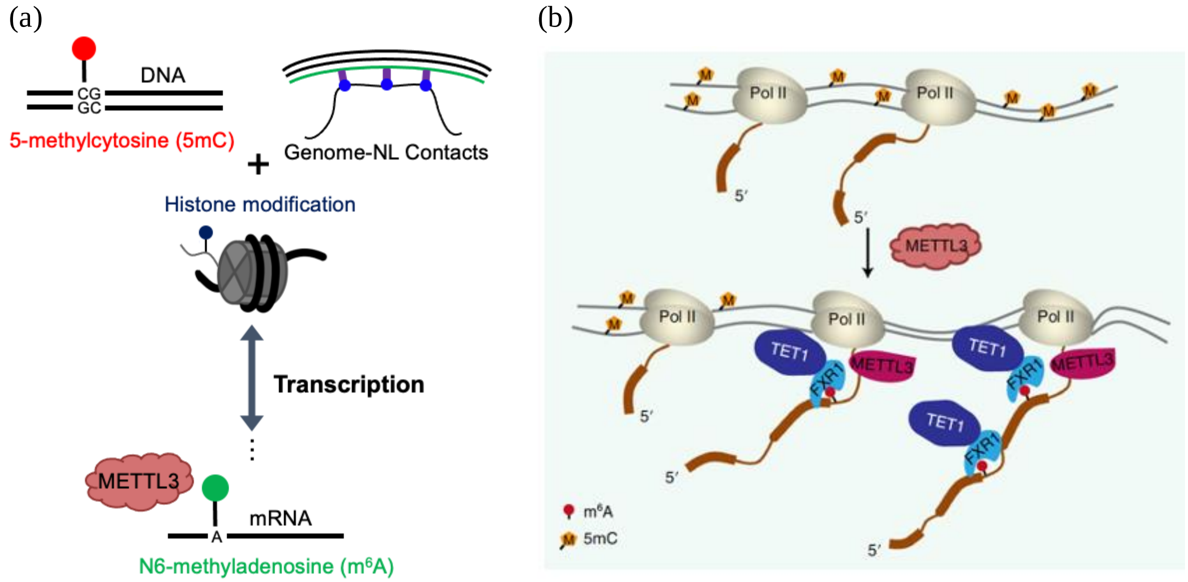
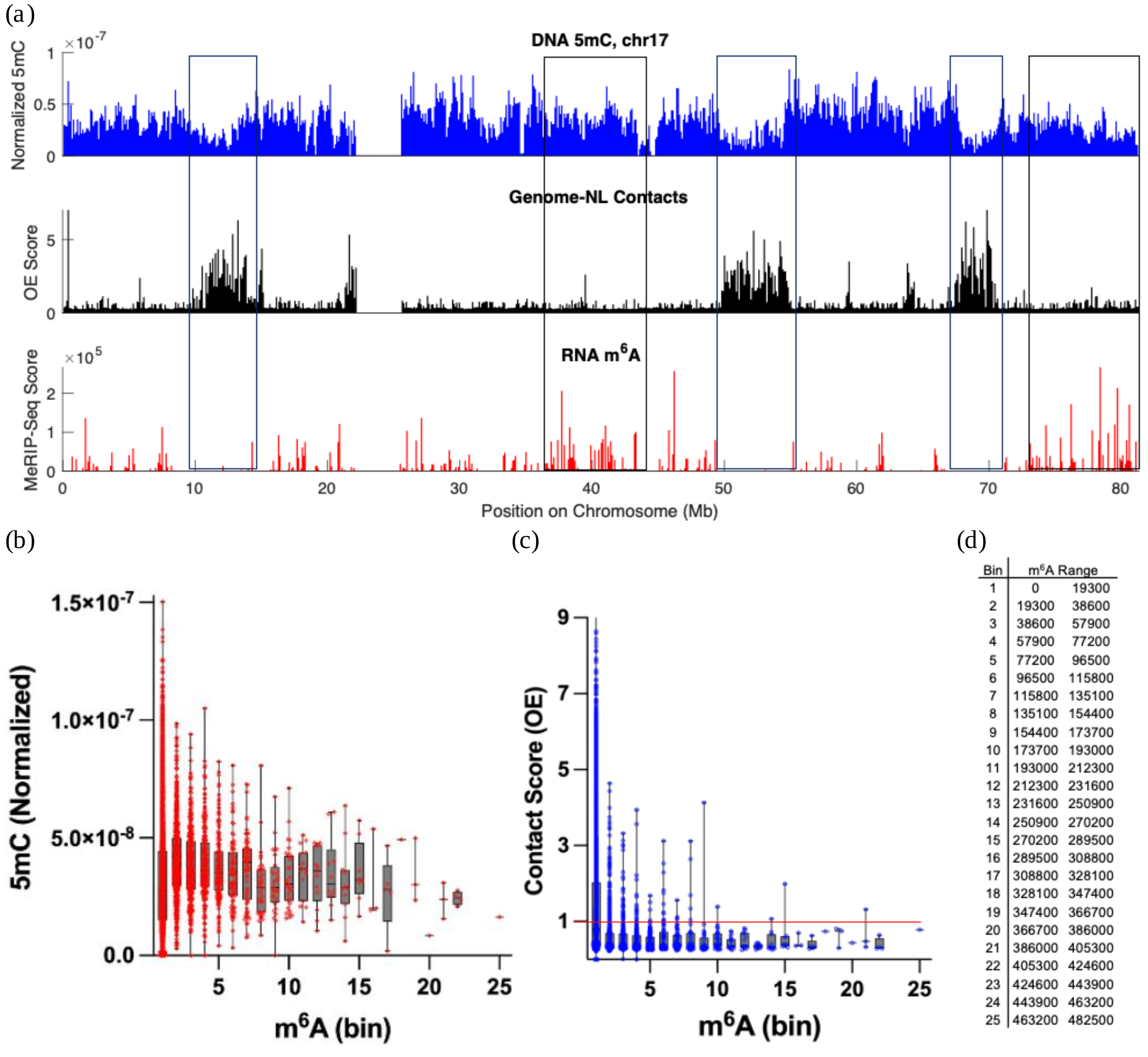


Fig. 62 | Features of epigenome, transcriptome and epitranscriptome display links. (a) 5-methylcytosine, histone modification, and genome-NL contacts may all interact synergistically to influence gene expression, which is further modified my m^{6}A. Likewise, gene expression may influence these epigenetic and epitranscriptomic features. **(b)** RNA m^{6}A formation-coupled DNA 5mC demethylation. METTL3 co-transcriptionally recruits TET1 by m^{6}A reader FXR1. Figure adapted from S. Deng *et al. Nat. Genet* (2022).

5mC due to the co-transcriptional TET1 activity, which raises the question, does m^{6}A on RNA influence other epigenetic features, such as genome organization? One limitation to the study was the m^{6}A level was not normalized to gene expression, which could interfere with the results if more m^{6}A was appearing in the immunoprecipitation for a specific gene only because that gene was more frequently transcribed, allowing more m^{6}A marks to be placed co-transcriptionally by METTL3.

To study the relationships between the epigenome and the epitranscriptome, publicly available bulk K562 MeRIP-seq (m^{6}A immunoprecipitation) data[118] was used and compared with single-cell 5mC and DamID data. There was no KBM7 data publicly available at the time, however both cell lines are chronic myelogenous leukemia (CML), meaning that comparisons can still be made. As shown previously in *Chapter III*, 5mC is anti-correlated with LADs, displaying hypomethylation at the single-cell level **(Fig. 63a, top two tracks)**. However, LADs don’t have RNA m^{6}A **(Fig. 63a, bottom two tracks)**, suggesting the genes in LADs have been shut down and the lower gene expression levels prevent the m^{6}A from being co-transcriptionally added by METTL3 and later detected by MeRIP-seq. This supports a hierarchy of epigenetic control with spatial organization of the genetic material



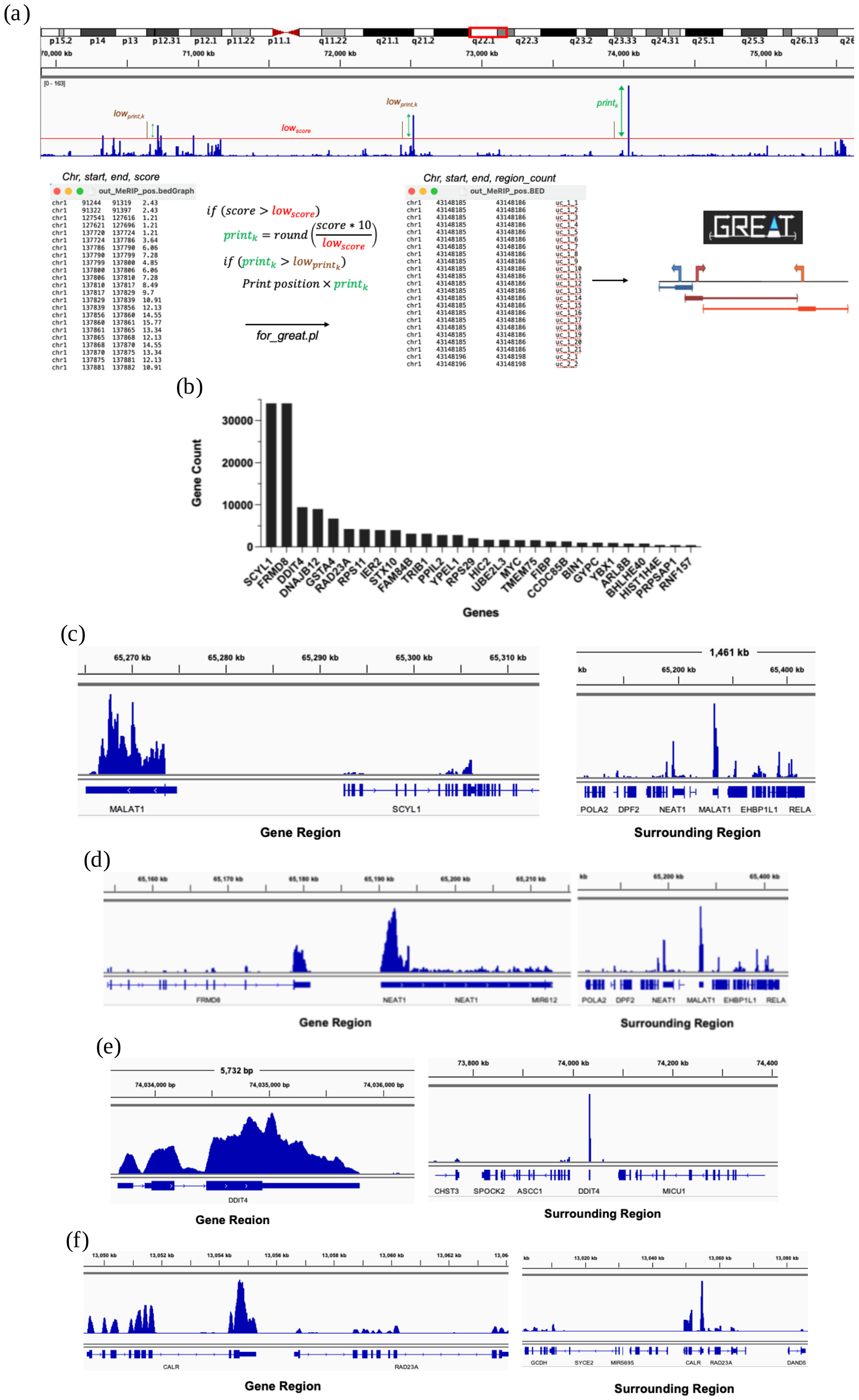
Bin	m^{6}A Range	
1	0	19300
2	19300	38600
3	38600	57900
4	57900	77200
5	77200	96500
6	96500	115800
7	115800	135100
8	135100	154400
9	154400	173700
10	173700	193000
11	193000	212300
12	212300	231600
13	231600	250900
14	250900	270200
15	270200	289500
16	289500	308800
17	308800	328100
18	328100	347400
19	347400	366700
20	366700	386000
21	386000	405300
22	405300	424600
23	424600	443900
24	443900	463200
25	463200	482500

Fig. 63 | RNA m^{6}A anticorrelated with genome-NL contacts and DNA 5mC in iLADs. (a) 5mC (top) and genome-NL contacts (middle) from the same KBM7 cell (sc5mC+DamIDv1) and bulk RNA m^{6}A (bottom) in K562. **(b)** 5mC vs. m6A levels. **(c)** Genome-NL contact score vs m^{6}A levels. In (b,c) the ranges of m^{6}A were put into bins for box plot (min/mix) **(d)** Range of bin and m^{6}A amount.

having the most influence over gene expression. In iLAD regions, m^{6}A may be anti-correlated with 5mC levels, seen by the iLAD highlighted at 73-83 Mb that contains high m^{6}A and low 5mC, at 36-45 Mb with higher m^{6}A and lower 5mC than the adjacent iLAD coordinates of 34-36 Mb with less m^{6}A and more 5mC. Position 32-33 Mb is a variable LAD **(Fig. 40d)** that doesn't appear in this single-cell, based on the absence of Dam signal at that position, however there is low m^{6}A from the bulk profiled cells. What is likely occurred was the cells that were making the contact had no m^{6}A at that region, and the cells that didn't make that contact may have had some, however, it is obscured by the bulk m^{6}A measurement. This supports the need to measure epigenetics and epi-transcriptomics within the same cell to get the full picture of how these features are related. Quantitatively, the 5mC and contact score (OE) of single cells were compared to the RNA m^{6}A in the same genomic bins **(Fig. 63)**. For the LADs **(Fig. 63c)**, the first bin reached the highest contact scores, supporting there was low m^{6}A at LAD regions, confirmed by later bins 2-25, containing more m^{6}A and having low contact scores, with the upper quartile less than OE of 1. This implies that RNA m^{6}A is found mostly in iLADs. For the 5mC **(Fig. 63b)**, the first bin had a wide range of 5mC levels, reaching the highest and lowest levels, which is most likely explained by LADs with no m^{6}A having low amounts of 5mC because they are hypomethylated, and iLADs with no m^{6}A having high amounts of 5mC because TET1 is not co-transcriptionally recruited to demethylate the region. Bin two and onwards, until there were few data points, displayed decreasing lower quartiles of 5mC levels as the amount of m^{6}A increased, suggesting more m^{6}A is associated with reaching lower amounts of 5mC through DNA demethylation. Future analysis would include segregating the 5mC and m^{6}A levels into which data points originated from LADs and iLADs to fully explain the wide

5mC range in the first bin. The expected outcome is there would be low amounts of 5mC in the first bin for the LADs, and high amounts of 5mC in the first bin for the iLADs. Additionally, the m^{6}A levels could be normalized to the average expression along each gene region, as the paper also included the corresponding RNA-seq data. Highly expressed genes, typically containing more 5mC in the gene body, may appear to have more RNA m^{6}A in the immunoprecipitation from their high expression, therefore acting as a confounding variable. To summarize, the data seems to suggest RNA m^{6}A is anticorrelated with genome-NL contacts and iLAD DNA 5mC, and sequencing both the epigenome and the epitranscriptome within the same single cell would prove useful in better understanding the correlations hidden by bulk measurements, such as how the segments of noncontact in LADs affect the transcription and m^{6}A levels in corresponding locations in the same cell.

To identify the genes that have high m^{6}A levels, an online tool called Genomic Regions Enrichment of Annotations Tool (GREAT) was used to associate the genomic regions from the K562 MeRIP-seq data with nearby genes[119]. GREAT works by first assigning each gene a regulatory domain that extends in both directions from its transcription start site (TSS) until reaching the next nearest gene, and then each inputted coordinate is associated with the regulatory domains it overlaps to determine which genes it could have originated from. Since the tool only takes the coordinates as input and not the m^{6}A score, the coordinates were inputted multiple times into the program proportional to the score magnitude, thresholding the input above a low m^{6}A score. This approach can be visualized in **Fig. 64a**, where each position on the chromosomes was surveyed for its m^{6}A score, and scores greater than an initial low score threshold (red horizontal line) were assigned an input count to GREAT (proportional to the size of the green arrow over the red baseline). The input count was



out_MeRIP_pos.bedGraph

```
chr1    91244   91319   2.43
chr1    91322   91397   2.43
chr1    127541  127616  1.21
chr1    127621  127696  1.21
chr1    137720  137724  1.21
chr1    137724  137786  3.64
chr1    137786  137790  6.06
chr1    137790  137799  7.28
chr1    137799  137800  4.85
chr1    137800  137806  6.06
chr1    137806  137810  7.28
chr1    137810  137817  8.49
chr1    137817  137829  9.7
chr1    137829  137839  10.91
chr1    137839  137856  12.13
chr1    137856  137860  14.55
chr1    137860  137861  15.77
chr1    137861  137865  13.34
chr1    137865  137868  12.13
chr1    137868  137870  14.55
chr1    137870  137875  13.34
chr1    137875  137881  12.13
chr1    137881  137882  10.91
```

$$if\ (score > low_{score})$$
$$print_k = round\left(\frac{score * 10}{low_{score}}\right)$$
$$if\ (print_k > low_{print_k})$$
$$Print\ position \times print_k$$

out_MeRIP_pos.BED

```
chr1    43148185        43148186        uc_1_1
chr1    43148185        43148186        uc_1_2
chr1    43148185        43148186        uc_1_3
chr1    43148185        43148186        uc_1_4
chr1    43148185        43148186        uc_1_5
chr1    43148185        43148186        uc_1_6
chr1    43148185        43148186        uc_1_7
chr1    43148185        43148186        uc_1_8
chr1    43148185        43148186        uc_1_9
chr1    43148185        43148186        uc_1_10
chr1    43148185        43148186        uc_1_11
chr1    43148185        43148186        uc_1_12
chr1    43148185        43148186        uc_1_13
chr1    43148185        43148186        uc_1_14
chr1    43148185        43148186        uc_1_15
chr1    43148185        43148186        uc_1_16
chr1    43148185        43148186        uc_1_17
chr1    43148185        43148186        uc_1_18
chr1    43148185        43148186        uc_1_19
chr1    43148185        43148186        uc_1_20
chr1    43148185        43148186        uc_1_21
chr1    43148196        43148198        uc_2_1
chr1    43148196        43148198        uc_2_2
```

Fig. 64 | Identify genes with high m^{6}A RNA modification using GREAT. (a) Pipeline for identifying genes. Scores multiplied by 10, otherwise rounded input counts will be the same. **(b)** Nearest genes to locations with high m^{6}A levels. **(c)** Genes of interest and surrounding regions in Integrative Genomics Viewer (IGV)[121], with peaks corresponding to m^{6}A magnitudes: MALAT1, **(d)** NEAT1, **(e)** DDIT4, **(f)** CALR.

further subject to a second round of thresholding (brown vertical line to tune how far above the initial threshold) to include only genes with the highest m^{6}A scores. After running the thresholded, score-replicated coordinates through GREAT, it was counted how many times coordinates were associated with a gene. Numerous genes were found to be associated with the inputted coordinates **(Fig. 64b)**. The gene with the most m^{6}A was actually *MALAT1* (Metastasis Associated Lung Adenocarcinoma Transcript 1), a lncRNA just upstream of *SCYL1* (**Fig. 64c**, likely misidentified because it is lncRNA rather than a standard gene), which acts as transcriptional regulator for genes that play a role in cell migration, metastasis, and cell cycle regulation (*MALAT1 Gene, GeneCards*). This lncRNA is of utmost importance because when it is upregulated in cancer, the tumor cells experience metastasis and have increased proliferation. The gene with the second most m^{6}A was actually *NEAT1* (Nuclear Paraspeckle Assembly Transcript 1), a lncRNA just downstream of *FRMD8* (**Fig. 64d**, likely misidentified because it is lncRNA rather than a standard gene), which is significant because it acts as a transcriptional regulator for genes that affect cancer progression (*NEAT1 Gene, GeneCards*). The gene with the third most m^{6}A was *DDIT4* (DNA Damage Inducible Transcript 4) **(Fig. 64e)** which plays a role in regulating p53/TP53-mediated apoptosis when DNA damage occurs, as well as regulating cell survival and proliferation by inhibiting mammalian target of rapamycin complex 1 (mTORC1) activity (*DDIT4 Gene, GeneCards*). One of the next genes with the most m^{6}A was *CALR* (Calreticulin), just upstream of *RAD23A* **(Fig. 64f)**, which is known to regulate how much calcium is stored within cells, in turn influencing gene activity, cell growth and movement, and cell death (*CALR gene, NCI Dictionary of Cancer Terms*). Further analysis would include normalizing the m^{6}A scores by gene expression from the RNA-seq data to account for the possibility that the m^{6}A amount in the immunoprecipation may be influenced by how much mRNA is available for methylation by METTL3. Besides using GREAT for only m^{6}A, any coordinates can be inputted, such as the regions within LADs that reposition after global DNA demethylation via DAC, to better understand the genes that are influenced by epigenetic regulation.

VI. Summarizing Discussion

The epigenome can influence gene expression in cells by chemically modifying DNA with 5-methylcytosine and spatially reorganizing genetic material, such as anchoring regions of DNA to the nuclear lamina to repress genes. Although 5mC and genome-NL interactions have both been independently examined at the single-cell level, and display correlation in bulk studies, they had yet to be measured simultaneously within the same cell, making it unknown if they were truly related within the same cell or were a result of the bulk measurement. Here we developed two methods for their simultaneous measurement in single cells, demonstrating that in the KBM7 cell line, regions that contact the nuclear lamina coincide with loss of 5mC, suggesting these epigenetic features are anticorrelated. Furthermore, LAD methylation is dependent on the contact frequency of that domain, such that more consistent LADs have a greater amount and noise in their 5mC, than variable LADs, which contain higher levels of 5mC. More variable LADs tend to have higher levels of activating histone modifications compared to consistent LADs that have more repressive histone modifications, supporting the role of the repressive H3K9 methylation in NL tethering. At the single-cell level, LADs have regions of noncontact, suggesting a multivalent coordination of contacts to anchor the genome to the NL. Longer segments of contact tend to have less average 5mC than shorter segments, and a similar trend was observed for the noncontact segments in the LAD, however most of the 5mC density is located within the noncontact regions. This contact length methylation trend was also observed in iLADs, and longer contact runs display less average methylation levels.

Further studies were aimed at determining whether 5mC and genome-NL contacts display causality, which could be pivotal for better understanding how epigenetic features combine to regulate gene expression. The genome was globally demethylated using decitabine, which caused rearrangement of LADs to the nuclear interior and to a lesser degree iLADs to the nuclear periphery. The regions of LADs that repositioned (in the form of "cracking", "sinking", and "eroding") were not random, and in fact were based upon the contact frequency of the region and the amount of repressive histone modifications (H3K9me3 and H3K27me3). Regions with lower contact frequently and lesser amounts of repressive histone marks were more likely to reposition, suggesting multiple epigenetic features working together to control the spatial organization of the genome, which in turn influences the gene expression.

A method for measuring the transcriptome simultaneously with the epigenetic marks within the same cell was implemented by reverse transcribing the mRNA with barcoded CEL-seq adapters prior to digestion with DpnI, allowing for determining how the epigenome regulates the transcriptome. m^6A RNA modifications were compared to the LAD profiles, indicating that this epitranscriptomic feature does not exist in the LAD regions, only in the iLADs, and the amount of iLAD RNA m^6A is anti-correlated with DNA 5mC levels, where more RNA m^6A is associated with less DNA 5mC. This is supported by co-transcriptional interactions between PolII, METTL3, FXR1, and TET1, to demethylate DNA near RNA that contains m^6A marks. Together this supports a larger mechanism of coordination between

5mC, spatial organization of the genetic material, histone modification, and RNA methylation, to influence gene expression. While the method was implemented using a cancer cell line, this technology could be applied to any cell line, allowing scientists to better understand how the epigenome is dysregulated in a broad range of illnesses.

A. Application

An application of this technology is characterizing epigenetic heterogeneity in tumors. Cancer is one of the primary causes of death in people under 70 years old globally and curing the disease is considered the most important factor to increase life expectancy in the next century[120]. However, the path to treating cancer has been challenging due to a growing understanding of the hallmarks of the disease and how genome instability elicits tumor proliferation and evasion of growth suppressors[121,122]. Although changes to the epigenome of both individual cancer cells and cells within the surrounding extracellular matrix environment likely influence the hallmarks of cancer, the interactions of epigenetic components in these single cells have yet to be fully explored. Since the heterogeneity in 5mC and genome-NL contacts can be characterized in a cell line, and each are known to regulate gene expression in cancer, I would expect in a complex system, such as a tumor or the tumor microenvironment, there will be even more epigenetic heterogeneity between the cells that can be characterized. Previously, much cancer research has focused on whole population (bulk) studies, effectively treating tumors as homogenous, and using therapeutics to target variants that are only detectable above background[123]. However, the driver of the cancer may not be detectable above background, such as in the case of rare circulating tumor cells (CTCs), which are released from primary tumors to seed metastasis, in parallel with epithelial-mesenchymal transition (EMT)[124,125]. Single-cell sequencing is therefore crucial for characterizing the complete heterogeneity of the cells in tumors and the tumor microenvironment since cancer is comprised of many cell states with divergent gene expression. Development of this technology to study the dysregulated state of the cancer epigenome will allow for characterizing which cell types in the tumor, and which regions of the cancer epigenome, have misregulated gene expression due to aberrant 5mC and genome-NL contacts. The impact this will have on the field is expanding our understanding of how multiple epigenetic features are related and dysregulated in cancer cells and underlying causes of the disease at the molecular level in individual cells.

B. Future Directions

1. Increase Throughput with Single-Cell Combinatorial Indexing

To increase the number of cells that can be sequenced in one experiment, a single-cell combinatorial indexing (sci) strategy could be used. Sci methods have been implemented to measure chromatin accessibility, DNA methylation, transcriptional dynamics, and profile the genome with a potential throughput of up to one million single cells by a three-level barcoding strategy[94,126–128]. Combinatorial indexing allows sequencing of many cells by creating unique barcodes for each cell that are a fusion of multiple rounds of barcoding. The method typically starts by isolating nuclei from cells, crosslinking them with formaldehyde to keep the nuclei intact, and then depleting the nucleosomes with SDS **(Fig. 65a)**. Next, nuclei are distributed into individual wells and are each molecularly tagged with well specific barcodes, using restriction enzymes to create the sites for adapter attachment. Nuclei are pooled, a limited number are redistributed into the wells to prevent cells from receiving the

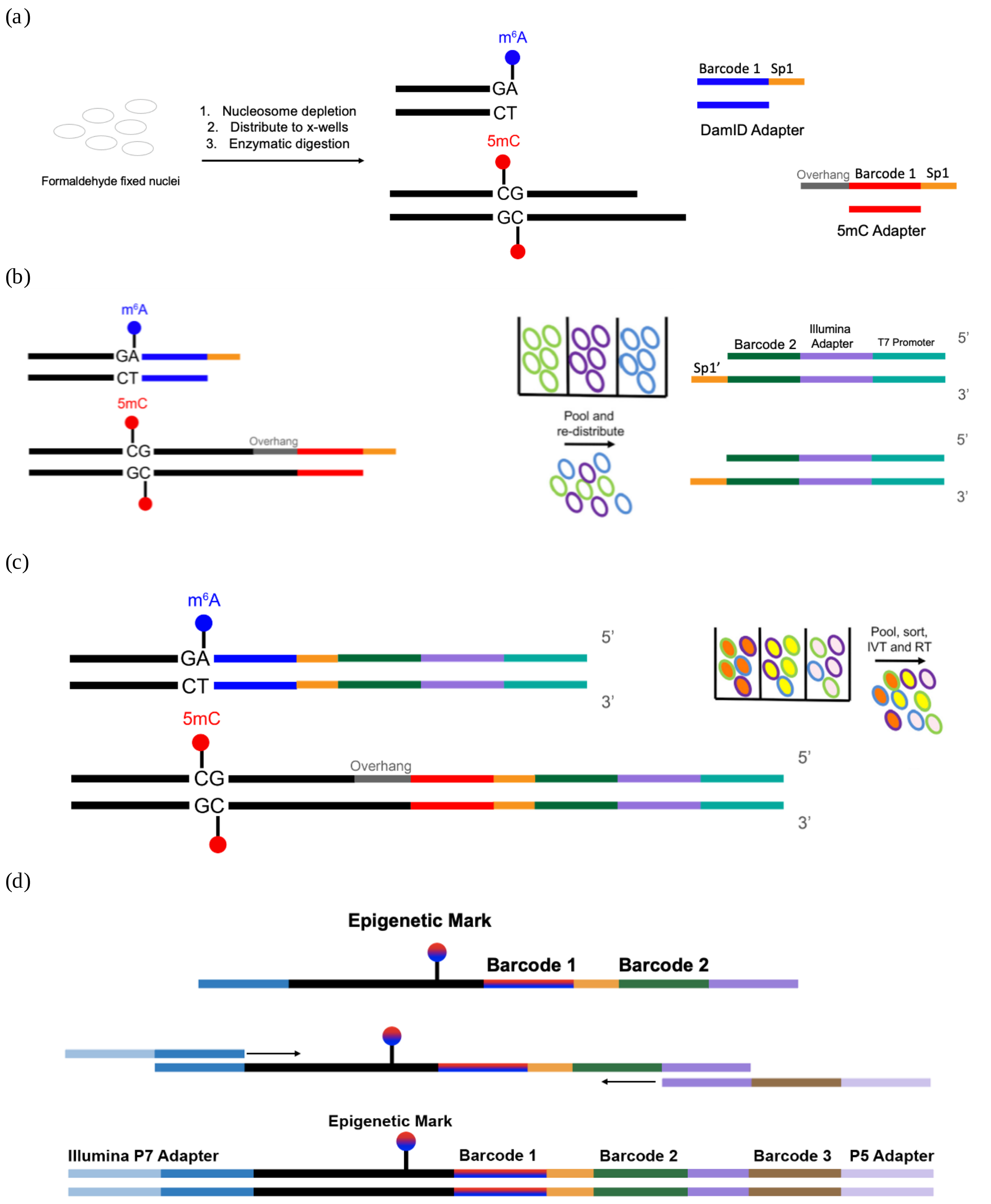


Fig. 65 | Increase cell throughput with single-cell combinatorial indexing (sci). (a) Formaldehyde fixed nuclei are depleted of their nucleosomes, distributed to the well plate, enzymatically digested with DpnI and MspJI, and then well/mark specific barcodes, containing a spacer, are ligated to the fragments. **(b)** Nuclei are pooled and redistributed into wells to attach the second set of barcoded adapters, containing similar components as in sc5mC+DamID v1. **(c)** Nuclei are pooled and retributed a final time and IVT and RT are performed. **(d)** The third set of barcodes are added via PCR to produce the final library.

same barcode, and then a second round of barcodes is added by ligation **(Fig. 65b)**. Nuclei are pooled one last time and redistributed, lysed, IVT and RT are performed **(Fig. 65c)**, and the third barcode is added by PCR using modified Illumina adapters, increasing the unique combinations of barcodes **(Fig. 65d)**. This strategy may be implemented as sci-5mC+DamID sequencing by first modifying the 5mC and DamID adapter ligation step, seen in **Fig. 9b**, to initially attach only a cell barcode that contains an overhang for ligating an adapter onto it. The adapter will be a modified version of the adapters in **Fig. 9b**, containing a complementary overhang to attach to the first barcode. By attaching the first barcode to the restriction sites, pooling, splitting, and then ligating to it the modified version of the 5mC/DamID adapter, two barcodes will be added in series to each cell. The third barcode will be added in the PCR step, seen in **Fig. 9d**, in which the Illumina P5 PCR primer will be modified to contain a barcode after the sequence complementary to RA5. A barrier to the sci method may be keeping the nuclei intact, to ensure they can be sorted without their contents leaking. This will likely involve optimization of the reagents and concentrations used in the protocol. Currently our method only allows for sequencing up to 384 cells per plate, however we are still able to see heterogeneities in 5mC and LAD profiles with even a few cells. A sci technique would prove particularly useful for dissociating a tumor and profiling the epigenomes of all the cell types it contains as well as profiling rare cell types within the tumor.

2. Reposition LADs by Perturbing Boundaries to Change 5mC and the full hierarchical mechanisms of epigenetic regulation

To further investigate and subsequently prove the causation between changes in 5mC and genome organization, individual LADs could be repositioned with the notion that this will impact genome methylation. A method to repositioning LADs is perturbing their boundaries by adding methylation to boundary elements. Editing DNA methylation has been achieved by using a dCas9-DNMT3A/TET complex and guiding it to a DNA target site with guide RNA to add/remove 5mC **(Fig. 66a)**[129]. LADs could be repositioned to the nuclear interior using dCas9 tethered to an enzyme that changes DNA methylation, such as DNMT, to target the borders of LADs **(Fig. 66b)**, containing the insulator protein CTCF which is thought to play a role in LAD confinement, and maintaining boundary and structure[22]. By adding 5mC to the boundary of the LAD, the potential outcomes are it will destabilize the LAD and reposition it to the nuclear interior, or changing the methylation of only the boundary elements is not sufficient to reposition the LAD. If the LADs do reposition **(Fig. 66b,**

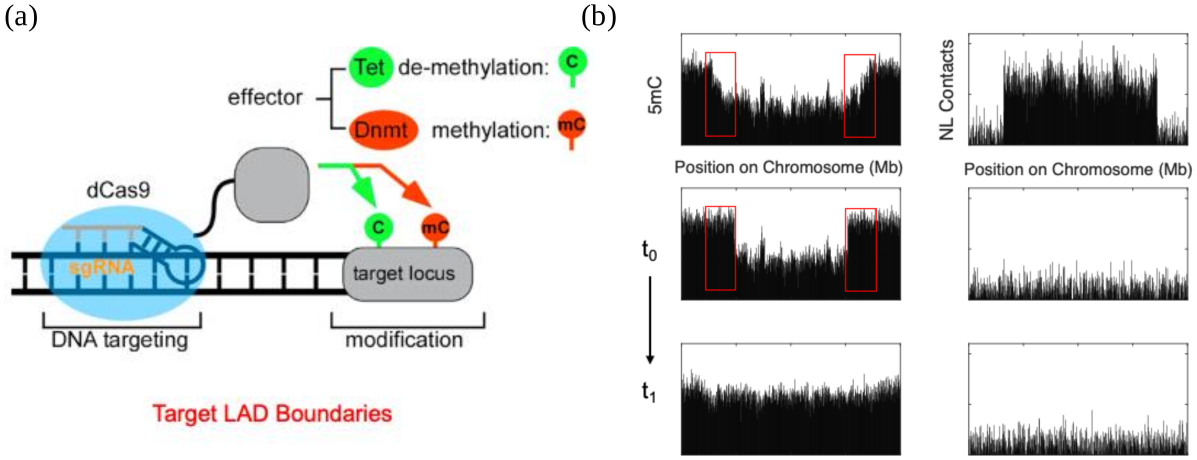


Fig. 66: Repositioning LADs by perturbing boundaries to change 5mC. (a) dCas9-DNMT3A/TET complex for modifying DNA methylation at target locus. Figure adapted from X. S. Liu *et al. Cell* (2016). **(b)** Hypothetical 5mC (left) and Dam profiles (right) before LAD boundary elements modification (top), after modification (t_0) and after methylation response to genome reorganization (t_1).

middle), a corresponding methylation change in the region in most of the cells will occur if these epigenetic features display causation **(Fig. 66b, bottom)**, or the repositioning may not change the methylation, which suggests the observed repositioning of LADs and iLAD in response to DAC global demethylation may be a part of a higher mechanism of epigenetic regulation. The genome-NL contact profiles and methylation would be measured in a time course experiment to observe the stages of repositioning and to determine the response time delay in methylation change. A backup strategy if adding the methylation to the boundary elements isn't sufficient to reposition the LADs is using siRNA to knockdown nuclear envelope transmembrane proteins to reposition LADs, which has been shown to reversibly reposition the normal peripheral positioning of chromosome 5 in liver cells[130]. Alternatively, the dCas9-TET system could be used to demethylate a tumor suppressor, which could in turn cause global reorganization of LADs/iLADs, which may result in global methylation changes. The CRISPRa (dCas9 activation) system should not be used initially however, because although it can activate genes and reposition them to the nuclear interior, it could change more epigenetic features than 5mC, such as histone modification as well, obscuring the direct relationships between genome-NL contacts and DNA methylation. The full causation of DNA methylation and genome organization will be proved if in addition to changing the 5mC repositions the LADs, as observed in the DAC experiments, repositioning the LADs changes 5mC at the single-cell level.

To better understand the full hierarchical mechanisms of epigenetic regulation within the single cell, additional components can be measured simultaneously by developing more advanced sequencing strategies. One such method could be measuring 5mC, genome-NL contacts, histone modification, and transcriptome all within the same cell for the larger picture of the causation between epigenetic features and the transcriptome. Measuring genome-NL contacts and transcriptome within the same cell was already achieved with A-Tailing and Poly-T Priming (v2.2 of the protocol), which will be described further in *Chapter V*, and measuring 5mC and genome-NL contacts within the same cell is doable, suggesting measuring the additional features is possible. A way to observe this with the methods already developed is by adding a polycomb repressive complex 2 (PRC2) inhibitor to the cell media, preventing methylation of histone H3 on lysine 27[131,132]. Since the H3K27me3 histone modification is associated with high contact frequency (consistent) LAD regions **(Fig. 46i)** and the locations in LADs that don't reposition when DAC is added **(Fig. 56i)**, if DAC and PRC2 inhibitor are added together, there would likely be an increase in genome reorganization compared to adding only DAC, which could result in further methylation changes. By using this PRC2 inhibitor alongside DAC and performing the sc5mC+DamID+Transcriptome protocol in *Chapter V*, the connections between DNA methylation, genome-NL contacts, transcriptome, and histone modification (H3K27me3 in this example) would be revealed **(Fig. 67)**.

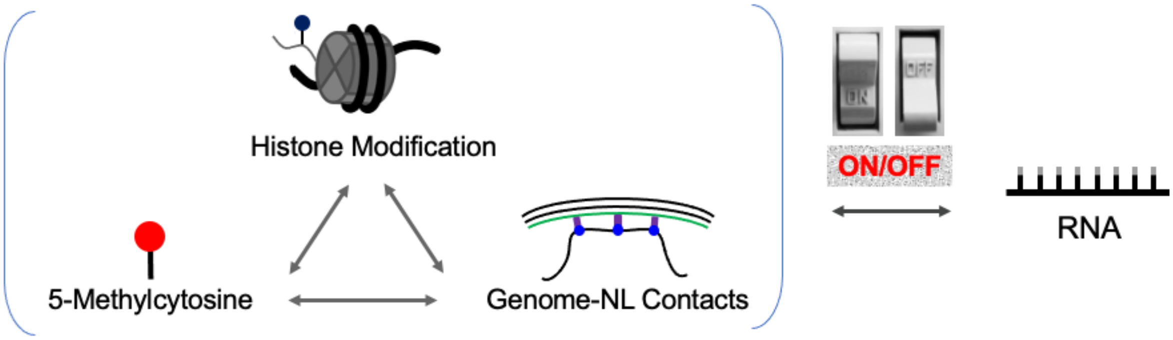


Fig. 67 | Causality of epigenetic features and transcriptome. 5-methylcytosine, histone modification and genome-NL contacts may all interact synergistically to influence gene expression. Likewise, gene expression may influence these epigenetic features.

References:

(1) Berman, B. P.; Weisenberger, D. J.; Aman, J. F.; Hinoue, T.; Ramjan, Z.; Liu, Y.; Noushmehr, H.; Lange, C. P. E.; van Dijk, C. M.; Tollenaar, R. A. E. M.; Van Den Berg, D.; Laird, P. W. Regions of Focal DNA Hypermethylation and Long-Range Hypomethylation in Colorectal Cancer Coincide with Nuclear Lamina–Associated Domains. *Nat. Genet.* **2012**, *44* (1), 40–46. https://doi.org/10.1038/ng.969.

(2) Pakneshan, P. Demethylation of Urokinase Promoter as a Prognostic Marker in Patients with Breast Carcinoma. *Clin. Cancer Res.* **2004**, *10* (9), 3035–3041. https://doi.org/10.1158/1078-0432.CCR-03-0545.

(3) Melé, M.; Rinn, J. L. "Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription. *Mol. Cell* **2016**, *62* (5), 657–664. https://doi.org/10.1016/j.molcel.2016.05.011.

(4) Gatticchi, L.; de las Heras, J. I.; Roberti, R.; Schirmer, E. C. Optimization of DamID for Use in Primary Cultures of Mouse Hepatocytes. *Methods* **2019**, *157*, 88–99. https://doi.org/10.1016/j.ymeth.2018.11.005.

(5) Vogel, M. J.; Peric-Hupkes, D.; van Steensel, B. Detection of in Vivo Protein–DNA Interactions Using DamID in Mammalian Cells. *Nat. Protoc.* **2007**, *2* (6), 1467–1478. https://doi.org/10.1038/nprot.2007.148.

(6) Singh, A. M.; Trost, R.; Boward, B.; Dalton, S. Utilizing FUCCI Reporters to Understand Pluripotent Stem Cell Biology. *Methods* **2016**, *101*, 4–10. https://doi.org/10.1016/j.ymeth.2015.09.020.

(7) Pauklin, S.; Vallier, L. The Cell-Cycle State of Stem Cells Determines Cell Fate Propensity. *Cell* **2013**, *155* (1), 135–147. https://doi.org/10.1016/j.cell.2013.08.031.

(8) Kind, J.; Pagie, L.; de Vries, S. S.; Nahidiazar, L.; Dey, S. S.; Bienko, M.; Zhan, Y.; Lajoie, B.; de Graaf, C. A.; Amendola, M.; Fudenberg, G.; Imakaev, M.; Mirny, L. A.; Jalink, K.; Dekker, J.; van Oudenaarden, A.; van Steensel, B. Genome-Wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell* **2015**, *163* (1), 134–147. https://doi.org/10.1016/j.cell.2015.08.040.

(9) Smith, Z. D.; Meissner, A. DNA Methylation: Roles in Mammalian Development. *Nat. Rev. Genet.* **2013**, *14* (3), 204–220. https://doi.org/10.1038/nrg3354.

(10) Bird, A. DNA Methylation Patterns and Epigenetic Memory. *Genes Dev.* **2002**, *16* (1), 6–21. https://doi.org/10.1101/gad.947102.

(11) Jones, P. A. Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond. *Nat. Rev. Genet.* **2012**, *13* (7), 484–492. https://doi.org/10.1038/nrg3230.

(12) Lyko, F. The DNA Methyltransferase Family: A Versatile Toolkit for Epigenetic Regulation. *Nat. Rev. Genet.* **2018**, *19* (2), 81–92. https://doi.org/10.1038/nrg.2017.80.

(13) Wu, X.; Zhang, Y. TET-Mediated Active DNA Demethylation: Mechanism, Function and Beyond. *Nat. Rev. Genet.* **2017**, *18* (9), 517–534. https://doi.org/10.1038/nrg.2017.33.

(14) Robertson, K. D.; Uzvolgyi, E.; Liang, G.; Talmadge, C.; Sumegi, J.; Gonzales, F. A.; Jones, P. A. The Human DNA Methyltransferases (DNMTs) 1, 3a and 3b: Coordinate mRNA Expression in Normal Tissues and Overexpression in Tumors. *Nucleic Acids Res.* **1999**, *27* (11), 2291–2298. https://doi.org/10.1093/nar/27.11.2291.

(15) Song, C.-X.; Szulwach, K. E.; Dai, Q.; Fu, Y.; Mao, S.-Q.; Lin, L.; Street, C.; Li, Y.; Poidevin, M.; Wu, H.; Gao, J.; Liu, P.; Li, L.; Xu, G.-L.; Jin, P.; He, C. Genome-Wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. *Cell* **2013**, *153* (3), 678–691. https://doi.org/10.1016/j.cell.2013.04.001.

(16) Ito, S.; Shen, L.; Dai, Q.; Wu, S. C.; Collins, L. B.; Swenberg, J. A.; He, C.; Zhang, Y. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science* **2011**, *333* (6047), 1300–1303. https://doi.org/10.1126/science.1210597.

(17) Lu, X.; Zhao, B. S.; He, C. TET Family Proteins: Oxidation Activity, Interacting Molecules, and Functions in Diseases. *Chem. Rev.* **2015**, *115* (6), 2225–2239. https://doi.org/10.1021/cr500470n.

(18) Xu, W.; Yang, H.; Liu, Y.; Yang, Y.; Wang, P.; Kim, S.-H.; Ito, S.; Yang, C.; Wang, P.; Xiao, M.-T.; Liu, L.; Jiang, W.; Liu, J.; Zhang, J.; Wang, B.; Frye, S.; Zhang, Y.; Xu, Y.; Lei, Q.; Guan, K.-L.; Zhao, S.; Xiong, Y. Oncometabolite 2-Hydroxyglutarate Is a Competitive Inhibitor of α-Ketoglutarate-Dependent Dioxygenases. *Cancer Cell* **2011**, *19* (1), 17–30. https://doi.org/10.1016/j.ccr.2010.12.014.

(19) Yang, H.; Liu, Y.; Bai, F.; Zhang, J.-Y.; Ma, S.-H.; Liu, J.; Xu, Z.-D.; Zhu, H.-G.; Ling, Z.-Q.; Ye, D.; Guan, K.-L.; Xiong, Y. Tumor Development Is Associated with Decrease of TET Gene Expression and 5-Methylcytosine Hydroxylation. *Oncogene* **2013**, *32* (5), 663–669. https://doi.org/10.1038/onc.2012.67.

(20) Prokocimer, M.; Davidovich, M.; Nissim-Rafinia, M.; Wiesel-Motiuk, N.; Bar, D. Z.; Barkan, R.; Meshorer, E.; Gruenbaum, Y. Nuclear Lamins: Key Regulators of Nuclear Structure and Activities. *J. Cell. Mol. Med.* **2009**, *13* (6), 1059–1085. https://doi.org/10.1111/j.1582-4934.2008.00676.x.

(21) Chubb, J. R.; Boyle, S.; Perry, P.; Bickmore, W. A. Chromatin Motion Is Constrained by Association with Nuclear Compartments in Human Cells. *Curr. Biol.* **2002**, *12* (6), 439–445. https://doi.org/10.1016/S0960-9822(02)00695-4.

(22) Guelen, L.; Pagie, L.; Brasset, E.; Meuleman, W.; Faza, M. B.; Talhout, W.; Eussen, B. H.; de Klein, A.; Wessels, L.; de Laat, W.; van Steensel, B. Domain Organization of Human Chromosomes Revealed by Mapping of Nuclear Lamina Interactions. *Nature* **2008**, *453* (7197), 948–951. https://doi.org/10.1038/nature06947.

(23) van Steensel, B.; Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **2017**, *169* (5), 780–791. https://doi.org/10.1016/j.cell.2017.04.022.

(24) Peric-Hupkes, D.; Meuleman, W.; Pagie, L.; Bruggeman, S. W. M.; Solovei, I.; Brugman, W.; Gräf, S.; Flicek, P.; Kerkhoven, R. M.; van Lohuizen, M.; Reinders, M.; Wessels, L.; van Steensel, B. Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. *Mol. Cell* **2010**, *38* (4), 603–613. https://doi.org/10.1016/j.molcel.2010.03.016.

(25) Osborne, C. S.; Chakalova, L.; Brown, K. E.; Carter, D.; Horton, A.; Debrand, E.; Goyenechea, B.; Mitchell, J. A.; Lopes, S.; Reik, W.; Fraser, P. Active Genes Dynamically Colocalize to Shared Sites of Ongoing Transcription. *Nat. Genet.* **2004**, *36* (10), 1065–1071. https://doi.org/10.1038/ng1423.

(26) Petruk, S.; Sedkov, Y.; Riley, K. M.; Hodgson, J.; Schweisguth, F.; Hirose, S.; Jaynes, J. B.; Brock, H. W.; Mazo, A. Transcription of Bxd Noncoding RNAs Promoted by Trithorax Represses Ubx in Cis by Transcriptional Interference. *Cell* **2006**, *127* (6), 1209–1221. https://doi.org/10.1016/j.cell.2006.10.039.

(27) Kind, J.; Pagie, L.; Ortabozkoyun, H.; Boyle, S.; de Vries, S. S.; Janssen, H.; Amendola, M.; Nolen, L. D.; Bickmore, W. A.; van Steensel, B. Single-Cell Dynamics of Genome-Nuclear Lamina Interactions. *Cell* **2013**, *153* (1), 178–192. https://doi.org/10.1016/j.cell.2013.02.028.

(28) Perovanovic, J.; Dell'Orso, S.; Gnochi, V. F.; Jaiswal, J. K.; Sartorelli, V.; Vigouroux, C.; Mamchaoui, K.; Mouly, V.; Bonne, G.; Hoffman, E. P. Laminopathies Disrupt Epigenomic Developmental Programs and Cell Fate. *Sci. Transl. Med.* **2016**, *8* (335), 335ra58-335ra58. https://doi.org/10.1126/scitranslmed.aad4991.

(29) Moss, S. F.; Krivosheyev, V.; de Souza, A.; Chin, K.; Gaetz, H. P.; Chaudhary, N.; Worman, H. J.; Holt, P. R. Decreased and Aberrant Nuclear Lamin Expression in Gastrointestinal Tract Neoplasms. *Gut* **1999**, *45* (5), 723–729. https://doi.org/10.1136/gut.45.5.723.

(30) Venables, R. S.; McLean, S.; Luny, D.; Moteleb, E.; Morley, S.; Quinlan, R. A.; Lane, E. B.; Hutchison, C. J. Expression of Individual Lamins in Basal Cell Carcinomas of the Skin. *Br. J. Cancer* **2001**, *84* (4), 512–519. https://doi.org/10.1054/bjoc.2000.1632.

(31) Ehrlich, M. DNA Hypomethylation in Cancer Cells. *Epigenomics* **2009**, *1* (2), 239–259. https://doi.org/10.2217/epi.09.33.

(32) Jones, P. A. DNA Methylation Errors and Cancer. *Cancer Res* **1996**, *56* (11), 2463–2467.

(33) Mizuno, S.; Chijiwa, T.; Okamura, T.; Akashi, K.; Fukumaki, Y.; Niho, Y.; Sasaki, H. Expression of DNA Methyltransferases DNMT1,3A, and 3B in Normal Hematopoiesis and in Acute and Chronic Myelogenous Leukemia. *Blood* **2001**, *97* (5), 1172–1179. https://doi.org/10.1182/blood.V97.5.1172.

(34) Pogribny, I. P.; Miller, B. J.; James, S. J. Alterations in Hepatic P53 Gene Methylation Patterns during Tumor Progression with Folateimethyl Deficiency in the Rat. *Cancer Lett.* **1997**, *115*, 31–38.

(35) Frank, S. A. Somatic Mutation: Early Cancer Steps Depend on Tissue Architecture. *Curr. Biol.* **2003**, *13* (7), R261–R263. https://doi.org/10.1016/S0960-9822(03)00195-7.

(36) Vogelstein, B.; Kinzler, K. W. Cancer Genes and the Pathways They Control. *Nat. Med.* **2004**, *10* (8), 789–799. https://doi.org/10.1038/nm1087.

(37) Agrelo, R.; Setien, F.; Espada, J.; Artiga, M. J.; Rodriguez, M.; Pérez-Rosado, A.; Sanchez-Aguilera, A.; Fraga, M. F.; Piris, M. A.; Esteller, M. Inactivation of the *Lamin A/C* Gene by CpG Island Promoter Hypermethylation in Hematologic Malignancies, and Its Association With Poor Survival in Nodal Diffuse Large B-Cell Lymphoma. *J. Clin. Oncol.* **2005**, *23* (17), 3940–3947. https://doi.org/10.1200/JCO.2005.11.650.

(38) Tomlinson, M. J.; Tomlinson, S.; Yang, X. B.; Kirkham, J. Cell Separation: Terminology and Practical Considerations. *J. Tissue Eng.* **2013**, *4*, 204173141247269. https://doi.org/10.1177/2041731412472690.

(39) Latt, S. A.; Stetten, G. Spectral Studies on 33258 Hoechst and Related Bisbenzimidazole Dyes Useful for Fluorescent Detection of Deoxyribonucleic Acid Synthesis. *J. Histochem. Cytochem.* **1976**, *24* (1), 24–33. https://doi.org/10.1177/24.1.943439.

(40) Walter, J.; Schermelleh, L.; Cremer, M.; Tashiro, S.; Cremer, T. Chromosome Order in HeLa Cells Changes during Mitosis and Early G1, but Is Stably Maintained during Subsequent Interphase Stages. *J. Cell Biol.* **2003**, *160* (5), 685–697. https://doi.org/10.1083/jcb.200211103.

(41) Vazquez, J.; Belmont, A. S.; Sedat, J. W. Multiple Regimes of Constrained Chromosome Motion Are Regulated in the Interphase Drosophila Nucleus. *Curr. Biol.* **2001**, *11* (16), 1227–1239. https://doi.org/10.1016/S0960-9822(01)00390-6.

(42) Banaszynski, L. A.; Chen, L.; Maynard-Smith, L. A.; Ooi, A. G. L.; Wandless, T. J. A Rapid, Reversible, and Tunable Method to Regulate Protein Function in Living Cells Using Synthetic Small Molecules. *Cell* **2006**, *126* (5), 995–1004. https://doi.org/10.1016/j.cell.2006.07.025.

(43) Rooijers, K.; Markodimitraki, C. M.; Rang, F. J.; de Vries, S. S.; Chialastri, A.; de Luca, K. L.; Mooijman, D.; Dey, S. S.; Kind, J. Simultaneous Quantification of Protein–DNA Contacts and Transcriptomes in Single Cells. *Nat. Biotechnol.* **2019**, *37* (7), 766–772. https://doi.org/10.1038/s41587-019-0150-y.

(44) Altemose, N.; Maslan, A.; Lai, A.; White, J. A.; Streets, A. M. *μDamID: A Microfluidic Approach for Imaging and Sequencing Protein-DNA Interactions in Single Cells*; preprint; Bioengineering, 2019. https://doi.org/10.1101/706903.

(45) Li, Z.; Jiao, X.; Di Sante, G.; Ertel, A.; Casimiro, M. C.; Wang, M.; Katiyar, S.; Ju, X.; Klopfenstein, D. V.; Tozeren, A.; Dampier, W.; Chepelev, I.; Jeltsch, A.; Pestell, R. G. Cyclin D1 Integrates G9a-Mediated Histone Methylation. *Oncogene* **2019**, *38* (22), 4232–4249. https://doi.org/10.1038/s41388-019-0723-8.

(46) van Schaik, T.; Vos, M.; Peric-Hupkes, D.; van Steensel, B. *Cell Cycle Dynamics of Lamina Associated DNA*; preprint; Molecular Biology, 2019. https://doi.org/10.1101/2019.12.19.881979.

(47) Gossen, M.; Bujard, H. Tight Control of Gene Expression in Mammalian Cells by Tetracycline-Responsive Promoters. *Proc. Natl. Acad. Sci.* **1992**, *89* (12), 5547–5551. https://doi.org/10.1073/pnas.89.12.5547.

(48) Achwal, C. W.; Chandra, H. S. A Sensitive Immunochemical Method for Detecting 5mC in DNA Fragments. *FEBS Lett.* **1982**, *150* (2), 469–472. https://doi.org/10.1016/0014-5793(82)80791-6.

(49) Braunschweig, M. H. Quantification of Global DNA Methylation with Infrared Fluorescence in Liver and Muscle Tissues of Differentially Fed Boars. *Luminescence* **2009**, *24* (4), 213–216. https://doi.org/10.1002/bio.1098.

(50) Balaghi, M.; Wagner, C. DNA Methylation in Folate Deficiency- Use of CpG Methylase. *Biochem. Biophys. Res. Commun.* **1993**, *193* (3), 1184–1190.

(51) Frommer, M.; McDonald, L. E.; Millar, D. S.; Collis, C. M.; Watt, F.; Grigg, G. W.; Molloy, P. L.; Paul, C. L. A Genomic Sequencing Protocol That Yields a Positive Display of 5-Methylcytosine Residues in Individual DNA Strands. *Proc. Natl. Acad. Sci.* **1992**, *89* (5), 1827–1831. https://doi.org/10.1073/pnas.89.5.1827.

(52) Harrison, J.; Stirzaker, C.; Clark, S. J. Cytosines Adjacent to Methylated CpG Sites Can Be Partially Resistant to Conversion in Genomic Bisulfite Sequencing Leading to Methylation Artifacts. *Anal. Biochem.* **1998**, *264* (1), 129–132. https://doi.org/10.1006/abio.1998.2833.

(53) Cohen-Karni, D.; Xu, D.; Apone, L.; Fomenkov, A.; Sun, Z.; Davis, P. J.; Morey Kinney, S. R.; Yamada-Mabuchi, M.; Xu, S. -y.; Davis, T.; Pradhan, S.; Roberts, R. J.; Zheng, Y. The MspJI Family of Modification-Dependent Restriction Endonucleases for Epigenetic Studies. *Proc. Natl. Acad. Sci.* **2011**, *108* (27), 11040–11045. https://doi.org/10.1073/pnas.1018448108.

(54) Li, Y.; Miyanari, Y.; Shirane, K.; Nitta, H.; Kubota, T.; Ohashi, H.; Okamoto, A.; Sasaki, H. Sequence-Specific Microscopic Visualization of DNA Methylation Status at Satellite Repeats in Individual Cell Nuclei and Chromosomes. *Nucleic Acids Res.* **2013**, *41* (19), e186–e186. https://doi.org/10.1093/nar/gkt766.

(55) Cheetham, S. W.; Jafrani, Y. M. A.; Andersen, S. B.; Jansz, N.; Kindlova, M.; Ewing, A. D.; Faulkner, G. J. *Single-Molecule Simultaneous Profiling of DNA Methylation and DNA-Protein Interactions with Nanopore-DamID*; preprint; Genomics, 2021. https://doi.org/10.1101/2021.08.09.455753.

(56) Rand, A. C.; Jain, M.; Eizenga, J. M.; Musselman-Brown, A.; Olsen, H. E.; Akeson, M.; Paten, B. Mapping DNA Methylation with High-Throughput Nanopore Sequencing. *Nat. Methods* **2017**, *14* (4), 411–413. https://doi.org/10.1038/nmeth.4189.

(57) Altemose, N.; Maslan, A.; Smith, O. K.; Sundararajan, K.; Brown, R. R.; Mishra, R.; Detweiler, A. M.; Neff, N.; Miga, K. H.; Straight, A. F.; Streets, A. DiMeLo-Seq: A Long-Read, Single-Molecule Method for Mapping Protein–DNA Interactions Genome Wide. *Nat. Methods* **2022**, *19* (6), 711–723. https://doi.org/10.1038/s41592-022-01475-6.

(58) Lowe, R.; Shirley, N.; Bleackley, M.; Dolan, S.; Shafee, T. Transcriptomics Technologies. *PLOS Comput. Biol.* **2017**, *13* (5), e1005457. https://doi.org/10.1371/journal.pcbi.1005457.

(59) Hashimshony, T.; Wagner, F.; Sher, N.; Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep.* **2012**, *2* (3), 666–673. https://doi.org/10.1016/j.celrep.2012.08.003.

(60) Hashimshony, T.; Senderovich, N.; Avital, G.; Klochendler, A.; de Leeuw, Y.; Anavy, L.; Gennert, D.; Li, S.; Livak, K. J.; Rozenblatt-Rosen, O.; Dor, Y.; Regev, A.; Yanai, I. CEL-Seq2: Sensitive Highly-Multiplexed Single-Cell RNA-Seq. *Genome Biol.* **2016**, *17* (1), 77. https://doi.org/10.1186/s13059-016-0938-8.

(61) Picelli, S.; Faridani, O. R.; Björklund, Å. K.; Winberg, G.; Sagasser, S.; Sandberg, R. Full-Length RNA-Seq from Single Cells Using Smart-Seq2. *Nat. Protoc.* **2014**, *9* (1), 171–181. https://doi.org/10.1038/nprot.2014.006.

(62) Macosko, E. Z.; Basu, A.; Satija, R.; Nemesh, J.; Shekhar, K.; Goldman, M.; Tirosh, I.; Bialas, A. R.; Kamitaki, N.; Martersteck, E. M.; Trombetta, J. J.; Weitz, D. A.; Sanes, J. R.; Shalek, A. K.; Regev, A.; McCarroll, S. A. Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **2015**, *161* (5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002.

(63) Erhard, F.; Baptista, M. A. P.; Krammer, T.; Hennig, T.; Lange, M.; Arampatzi, P.; Jürges, C. S.; Theis, F. J.; Saliba, A.-E.; Dölken, L. scSLAM-Seq Reveals Core Features of Transcription Dynamics in Single Cells. *Nature* **2019**, *571* (7765), 419–423. https://doi.org/10.1038/s41586-019-1369-y.

(64) Dey, S. S.; Kester, L.; Spanjaard, B.; Bienko, M.; Van Oudenaarden, A. Integrated Genome and Transcriptome Sequencing of the Same Cell. *Nat. Biotechnol.* **2015**, *33* (3), 285–289. https://doi.org/10.1038/nbt.3129.

(65) Wiener, D.; Schwartz, S. The Epitranscriptome beyond m6A. *Nat. Rev. Genet.* **2021**, *22* (2), 119–131. https://doi.org/10.1038/s41576-020-00295-8.

(66) Liu, J.; Dou, X.; Chen, C.; Chen, C.; Liu, C.; Xu, M. M.; Zhao, S.; Shen, B.; Gao, Y.; Han, D.; He, C. *N* ^{6} -Methyladenosine of Chromosome-Associated Regulatory RNA Regulates Chromatin State and Transcription. *Science* **2020**, *367* (6477), 580–586. https://doi.org/10.1126/science.aay6018.

(67) Huang, H.; Weng, H.; Chen, J. m6A Modification in Coding and Non-Coding RNAs: Roles and Therapeutic Implications in Cancer. *Cancer Cell* **2020**, *37* (3), 270–288. https://doi.org/10.1016/j.ccell.2020.02.004.

(68) Meyer, K. D.; Saletore, Y.; Zumbo, P.; Elemento, O.; Mason, C. E.; Jaffrey, S. R. Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3′ UTRs and near Stop Codons. *Cell* **2012**, *149* (7), 1635–1646. https://doi.org/10.1016/j.cell.2012.05.003.

(69) Meyer, K. D. DART-Seq: An Antibody-Free Method for Global m6A Detection. *Nat. Methods* **2019**, *16* (12), 1275–1280. https://doi.org/10.1038/s41592-019-0570-0.

(70) Hu, J. F.; Yim, D.; Ma, D.; Huber, S. M.; Davis, N.; Bacusmo, J. M.; Vermeulen, S.; Zhou, J.; Begley, T. J.; DeMott, M. S.; Levine, S. S.; De Crécy-Lagard, V.; Dedon, P. C.; Cao, B. Quantitative Mapping of the Cellular Small RNA Landscape with AQRNA-Seq. *Nat. Biotechnol.* **2021**, *39* (8), 978–988. https://doi.org/10.1038/s41587-021-00874-y.

(71) Barski, A.; Cuddapah, S.; Cui, K.; Roh, T.-Y.; Schones, D. E.; Wang, Z.; Wei, G.; Chepelev, I.; Zhao, K. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **2007**, *129* (4), 823–837. https://doi.org/10.1016/j.cell.2007.05.009.

(72) Greer, E. L.; Shi, Y. Histone Methylation: A Dynamic Mark in Health, Disease and Inheritance. *Nat. Rev. Genet.* **2012**, *13* (5), 343–357. https://doi.org/10.1038/nrg3173.

(73) Park, P. J. ChIP–Seq: Advantages and Challenges of a Maturing Technology. *Nat. Rev. Genet.* **2009**, *10* (10), 669–680. https://doi.org/10.1038/nrg2641.

(74) 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* **2012**, *13* (1), 424. https://doi.org/10.1186/1471-2164-13-424.

(75) Zhou, V. W.; Goren, A.; Bernstein, B. E. Charting Histone Modifications and the Functional Organization of Mammalian Genomes. *Nat. Rev. Genet.* **2011**, *12* (1), 7–18. https://doi.org/10.1038/nrg2905.

(76) Shogren-Knaak, M.; Ishii, H.; Sun, J.-M.; Pazin, M. J.; Davie, J. R.; Peterson, C. L. Histone H4-K16 Acetylation Controls Chromatin Structure and Protein Interactions. *Science* **2006**, *311* (5762), 844–847. https://doi.org/10.1126/science.1124000.

(77) Kim, J.; Kim, H. Recruitment and Biological Consequences of Histone Modification of H3K27me3 and H3K9me3. *ILAR J.* **2012**, *53* (3–4), 232–239. https://doi.org/10.1093/ilar.53.3-4.232.

(78) Pal, D.; Patel, M.; Boulet, F.; Sundarraj, J.; Grant, O. A.; Branco, M. R.; Basu, S.; Santos, S. D. M.; Zabet, N. R.; Scaffidi, P.; Pradeepa, M. M. H4K16ac Activates the Transcription of Transposable Elements and Contributes to Their Cis-Regulatory Function. *Nat. Struct. Mol. Biol.* **2023**, *30* (7), 935–947. https://doi.org/10.1038/s41594-023-01016-5.

(79) Kaya-Okur, H. S.; Wu, S. J.; Codomo, C. A.; Pledger, E. S.; Bryson, T. D.; Henikoff, J. G.; Ahmad, K.; Henikoff, S. CUT&Tag for Efficient Epigenomic Profiling of Small Samples and Single Cells. *Nat. Commun.* **2019**, *10* (1), 1930. https://doi.org/10.1038/s41467-019-09982-5.

(80) Zhu, C.; Zhang, Y.; Li, Y. E.; Lucero, J.; Behrens, M. M.; Ren, B. Joint Profiling of Histone Modifications and Transcriptome in Single Cells from Mouse Brain. *Nat. Methods* **2021**, *18* (3), 283–292. https://doi.org/10.1038/s41592-021-01060-3.

(81) Klemm, S. L.; Shipony, Z.; Greenleaf, W. J. Chromatin Accessibility and the Regulatory Epigenome. *Nat. Rev. Genet.* **2019**, *20* (4), 207–220. https://doi.org/10.1038/s41576-018-0089-8.

(82) Chen, C.; Xing, D.; Tan, L.; Li, H.; Zhou, G.; Huang, L.; Xie, X. S. Single-Cell Whole-Genome Analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* **2017**, *356* (6334), 189–194. https://doi.org/10.1126/science.aak9787.

(83) DeAngelis, M. M.; Wang, D. G.; Hawkins, T. L. Solid-Phase Reversible Immobilization for the Isolation of PCR Products. *Nucleic Acids Res.* **1995**, *23* (22), 4742–4743. https://doi.org/10.1093/nar/23.22.4742.

(84) Quail, M. A.; Swerdlow, H.; Turner, D. J. Improved Protocols for the Illumina Genome Analyzer Sequencing System. *Curr. Protoc. Hum. Genet.* **2009**, *62* (1). https://doi.org/10.1002/0471142905.hg1802s62.

(85) Greenleaf, W. J.; Sidow, A. The Future of Sequencing: Convergence of Intelligent Design and Market Darwinism. *Genome Biol.* **2014**, *15* (3), 303. https://doi.org/10.1186/gb4168.

(86) Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* **2009**, *25* (14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

(87) Li, H.; Durbin, R. Fast and Accurate Long-Read Alignment with Burrows–Wheeler Transform. *Bioinformatics* **2010**, *26* (5), 589–595. https://doi.org/10.1093/bioinformatics/btp698.

(88) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25* (16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

(89) Markodimitraki, C. M.; Rang, F. J.; Rooijers, K.; de Vries, S. S.; Chialastri, A.; de Luca, K. L.; Lochs, S. J. A.; Mooijman, D.; Dey, S. S.; Kind, J. Simultaneous Quantification of Protein–DNA Interactions and Transcriptomes in Single Cells with scDam&T-Seq. *Nat. Protoc.* **2020**, *15* (6), 1922–1953. https://doi.org/10.1038/s41596-020-0314-8.

(90) Siwek, W.; Czapinska, H.; Bochtler, M.; Bujnicki, J. M.; Skowronek, K. Crystal Structure and Mechanism of Action of the N6-Methyladenine-Dependent Type IIM Restriction Endonuclease R.DpnI. *Nucleic Acids Res.* **2012**, *40* (15), 7563–7572. https://doi.org/10.1093/nar/gks428.

(91) Horton, J. R.; Mabuchi, M. Y.; Cohen-Karni, D.; Zhang, X.; Griggs, R. M.; Samaranayake, M.; Roberts, R. J.; Zheng, Y.; Cheng, X. Structure and Cleavage Activity of the Tetrameric MspJI DNA Modification-Dependent Restriction Endonuclease. *Nucleic Acids Res.* **2012**, *40* (19), 9763–9773. https://doi.org/10.1093/nar/gks719.

(92) Moffatt, B. A.; Dunn, J. J.; Studier, F. W. Nucleotide Sequence of the Gene for Bacteriophage T7 RNA Polymerase. *J. Mol. Biol.* **1984**, *173* (2), 265–269. https://doi.org/10.1016/0022-2836(84)90194-3.

(93) Vaisvila, R.; Ponnaluri, V. K. C.; Sun, Z.; Langhorst, B. W.; Saleh, L.; Guan, S.; Dai, N.; Campbell, M. A.; Sexton, B. S.; Marks, K.; Samaranayake, M.; Samuelson, J. C.; Church, H. E.; Tamanaha, E.; Corrêa, I. R.; Pradhan, S.; Dimalanta, E. T.; Evans, T. C.; Williams, L.; Davis, T. B. Enzymatic Methyl Sequencing Detects DNA Methylation at Single-Base Resolution from Picograms of DNA. *Genome Res.* **2021**, *31* (7), 1280–1289. https://doi.org/10.1101/gr.266551.120.

(94) Mulqueen, R. M.; Pokholok, D.; Norberg, S. J.; Torkenczy, K. A.; Fields, A. J.; Sun, D.; Sinnamon, J. R.; Shendure, J.; Trapnell, C.; O'Roak, B. J.; Xia, Z.; Steemers, F. J.; Adey, A. C. Highly Scalable Generation of DNA Methylation Profiles in Single Cells. *Nat. Biotechnol.* **2018**, *36* (5), 428–431. https://doi.org/10.1038/nbt.4112.

(95) Bernstein, N.; Karimi-Busheri, F.; Rasouli-Nia, A.; Mani, R.; Dianov, G.; Glover, J.; Weinfeld, M. Polynucleotide Kinase as a Potential Target for Enhancing

Cytotoxicity by Ionizing Radiation and Topoisomerase I Inhibitors. *Anticancer Agents Med. Chem.* **2008**, *8* (4), 358–367. https://doi.org/10.2174/187152008784220311.

(96) Miura, F.; Shibata, Y.; Miura, M.; Sangatsuda, Y.; Hisano, O.; Araki, H.; Ito, T. Highly Efficient Single-Stranded DNA Ligation Technique Improves Low-Input Whole-Genome Bisulfite Sequencing by Post-Bisulfite Adaptor Tagging. *Nucleic Acids Res.* **2019**, *47* (15), e85–e85. https://doi.org/10.1093/nar/gkz435.

(97) Revathidevi, S.; Murugan, A. K.; Nakaoka, H.; Inoue, I.; Munirajan, A. K. APOBEC: A Molecular Driver in Cervical Cancer Pathogenesis. *Cancer Lett.* **2021**, *496*, 104–116. https://doi.org/10.1016/j.canlet.2020.10.004.

(98) Šponer, J.; Jurečka, P.; Hobza, P. Accurate Interaction Energies of Hydrogen-Bonded Nucleic Acid Base Pairs. *J. Am. Chem. Soc.* **2004**, *126* (32), 10142–10151. https://doi.org/10.1021/ja048436s.

(99) Penázová, H.; Vorlicková, M. Guanine Tetraplex Formation by Short DNA Fragments Containing Runs of Guanine and Cytosine. *Biophys. J.* **1997**, *73* (4), 2054–2063. https://doi.org/10.1016/S0006-3495(97)78235-3.

(100) Motea, E. A.; Berdis, A. J. Terminal Deoxynucleotidyl Transferase: The Story of a Misguided DNA Polymerase. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **2010**, *1804* (5), 1151–1166. https://doi.org/10.1016/j.bbapap.2009.06.030.

(101) Integrated DNA Technologies, Inc. Tail Trimming for Better Data, Technical Note, 04/22.

(102) Integrated DNA Technologies, Inc. xGen Methyl-Seq DNA Library Prep Kit Protocol.Pdf, 1/22.

(103) Integrated DNA Technologies, Inc. xGen Adaptase Module Protocol.Pdf, 1/22.

(104) Chandrasekhar, K. Restriction Enzyme HincII Is Sensitive to Methylation of Cytosine That Occurs 5' to the Recognition Sequence. *Nucleic Acids Res.* **1996**, *24* (6), 1045–1046. https://doi.org/10.1093/nar/24.6.1045.

(105) Doublet, V.; Souty-Grosset, C.; Bouchon, D.; Cordaux, R.; Marcadé, I. A Thirty Million Year-Old Inherited Heteroplasmy. *PLoS ONE* **2008**, *3* (8), e2938. https://doi.org/10.1371/journal.pone.0002938.

(106) Cremer, M.; Fischer, C.; Solovei, I.; Cremer, C.; Cremer, T. Non-Random Radial Higher-Order Chromatin Arrangements in Nuclei of Diploid Human Cells. 28.

(107) Croft, J. A.; Bridger, J. M.; Boyle, S.; Perry, P.; Teague, P.; Bickmore, W. A. Differences in the Localization and Morphology of Chromosomes in the Human Nucleus. *J. Cell Biol.* **1999**, *145* (6), 1119–1131. https://doi.org/10.1083/jcb.145.6.1119.

(108) Tchasovnikarova, I. A.; Timms, R. T.; Matheson, N. J.; Wals, K.; Antrobus, R.; Göttgens, B.; Dougan, G.; Dawson, M. A.; Lehner, P. J. Epigenetic Silencing by the HUSH Complex Mediates Position-Effect Variegation in Human Cells. *Science* **2015**, *348* (6242), 1481–1485. https://doi.org/10.1126/science.aaa7227.

(109) Tchasovnikarova, I. A.; Marr, S. K.; Damle, M.; Kingston, R. E. TRACE Generates Fluorescent Human Reporter Cell Lines to Characterize Epigenetic Pathways. *Mol. Cell* **2022**, *82* (2), 479-491.e7. https://doi.org/10.1016/j.molcel.2021.11.035.

(110) Towbin, B. D.; González-Aguilera, C.; Sack, R.; Gaidatzis, D.; Kalck, V.; Meister, P.; Askjaer, P.; Gasser, S. M. Step-Wise Methylation of Histone H3K9 Positions Heterochromatin at the Nuclear Periphery. *Cell* **2012**, *150* (5), 934–947. https://doi.org/10.1016/j.cell.2012.06.051.

(111) Bian, Q.; Khanna, N.; Alvikas, J.; Belmont, A. S. β-Globin Cis-Elements Determine Differential Nuclear Targeting through Epigenetic Modifications. *J. Cell Biol.* **2013**, *203* (5), 767–783. https://doi.org/10.1083/jcb.201305027.

(112) Hollenbach, P. W.; Nguyen, A. N.; Brady, H.; Williams, M.; Ning, Y.; Richard, N.; Krushel, L.; Aukerman, S. L.; Heise, C.; MacBeth, K. J. A Comparison of Azacitidine and Decitabine Activities in Acute Myeloid Leukemia Cell Lines. *PLoS ONE* **2010**, *5* (2), e9001. https://doi.org/10.1371/journal.pone.0009001.

(113) Lavelle, D.; DeSimone, J.; Hankewych, M.; Kousnetzova, T.; Chen, Y.-H. Decitabine Induces Cell Cycle Arrest at the G1 Phase via p21WAF1 and the G2/M Phase via the P38 MAP Kinase Pathway. *Leuk. Res.* **2003**, *27* (11), 999–1007. https://doi.org/10.1016/S0145-2126(03)00068-7.

(114) Yang, S.; Li, W.; Dong, F.; Sun, H.; Wu, B.; Tan, J.; Zou, W.; Zhou, D. KITLG Is a Novel Target of *miR-34c* That Is Associated with the Inhibition of Growth and Invasion in Colorectal Cancer Cells. *J. Cell. Mol. Med.* **2014**, *18* (10), 2092–2102. https://doi.org/10.1111/jcmm.12368.

(115) Tian, S.; Liu, W.; Pan, Y.; Zhan, S. Long Non-Coding RNA Linc00320 Inhibits Glioma Cell Proliferation through Restraining Wnt/β-Catenin Signaling. *Biochem. Biophys. Res. Commun.* **2019**, *508* (2), 458–464. https://doi.org/10.1016/j.bbrc.2018.11.101.

(116) Dobin, A.; Davis, C. A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T. R. STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* **2013**, *29* (1), 15–21. https://doi.org/10.1093/bioinformatics/bts635.

(117) Kaminow, B.; Yunusov, D.; Dobin, A. *STARsolo: Accurate, Fast and Versatile Mapping/Quantification of Single-Cell and Single-Nucleus RNA-Seq Data*; preprint; Bioinformatics, 2021. https://doi.org/10.1101/2021.05.05.442755.

(118) Xiong, F.; Wang, R.; Lee, J.-H.; Li, S.; Chen, S.-F.; Liao, Z.; Hasani, L. A.; Nguyen, P. T.; Zhu, X.; Krakowiak, J.; Lee, D.-F.; Han, L.; Tsai, K.-L.; Liu, Y.; Li, W. RNA m6A Modification Orchestrates a LINE-1–Host Interaction That Facilitates Retrotransposition and Contributes to Long Gene Vulnerability. *Cell Res.* **2021**, *31* (8), 861–885. https://doi.org/10.1038/s41422-021-00515-8.

(119) McLean, C. Y.; Bristor, D.; Hiller, M.; Clarke, S. L.; Schaar, B. T.; Lowe, C. B.; Wenger, A. M.; Bejerano, G. GREAT Improves Functional Interpretation of Cis-Regulatory Regions. *Nat. Biotechnol.* **2010**, *28* (5), 495–501. https://doi.org/10.1038/nbt.1630.

(120) Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R. L.; Torre, L. A.; Jemal, A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **2018**, *68* (6), 394–424. https://doi.org/10.3322/caac.21492.

(121) Hanahan, D.; Weinberg, R. A. The Hallmarks of Cancer. *Cell* **2000**, *100* (1), 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9.

(122) Hanahan, D.; Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **2011**, *144* (5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013.

(123) Lawson, D. A.; Kessenbrock, K.; Davis, R. T.; Pervolarakis, N.; Werb, Z. Tumour Heterogeneity and Metastasis at Single-Cell Resolution. *Nat. Cell Biol.* **2018**, *20* (12), 1349–1360. https://doi.org/10.1038/s41556-018-0236-7.

(124) Keller, L.; Pantel, K. Unravelling Tumour Heterogeneity by Single-Cell Profiling of Circulating Tumour Cells. *Nat. Rev. Cancer* **2019**, *19* (10), 553–567. https://doi.org/10.1038/s41568-019-0180-2.

(125) Cheng, Y.-H.; Chen, Y.-C.; Lin, E.; Brien, R.; Jung, S.; Chen, Y.-T.; Lee, W.; Hao, Z.; Sahoo, S.; Min Kang, H.; Cong, J.; Burness, M.; Nagrath, S.; S. Wicha, M.; Yoon, E. Hydro-Seq Enables Contamination-Free High-Throughput Single-Cell RNA-Sequencing for Circulating Tumor Cells. *Nat. Commun.* **2019**, *10* (1), 2163. https://doi.org/10.1038/s41467-019-10122-2.

(126) Cao, J.; Zhou, W.; Steemers, F.; Trapnell, C.; Shendure, J. Sci-Fate Characterizes the Dynamics of Gene Expression in Single Cells. *Nat. Biotechnol.* **2020**. https://doi.org/10.1038/s41587-020-0480-9.

(127) Cusanovich, D. A.; Daza, R.; Adey, A.; Pliner, H. A.; Christiansen, L.; Gunderson, K. L.; Steemers, F. J.; Trapnell, C.; Shendure, J. Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing. *Science* **2015**, *348* (6237), 910–914. https://doi.org/10.1126/science.aab1601.

(128) Yin, Y.; Jiang, Y.; Lam, K.-W. G.; Berletch, J. B.; Disteche, C. M.; Noble, W. S.; Steemers, F. J.; Camerini-Otero, R. D.; Adey, A. C.; Shendure, J. High-Throughput

Single-Cell Sequencing with Linear Amplification. *Mol. Cell* **2019**, *76* (4), 676-690.e10. https://doi.org/10.1016/j.molcel.2019.08.002.

(129) Liu, X. S.; Wu, H.; Ji, X.; Stelzer, Y.; Wu, X.; Czauderna, S.; Shu, J.; Dadon, D.; Young, R. A.; Jaenisch, R. Editing DNA Methylation in the Mammalian Genome. *Cell* **2016**, *167* (1), 233-247.e17. https://doi.org/10.1016/j.cell.2016.08.056.

(130) Zuleger, N.; Boyle, S.; Kelly, D. A.; de las Heras, J. I.; Lazou, V.; Korfali, N.; Batrakou, D. G.; Randles, K. N.; Morris, G. E.; Harrison, D. J.; Bickmore, W. A.; Schirmer, E. C. Specific Nuclear Envelope Transmembrane Proteins Can Promote the Location of Chromosomes to and from the Nuclear Periphery. *Genome Biol.* **2013**, *14* (2), R14. https://doi.org/10.1186/gb-2013-14-2-r14.

(131) Liu, K.-L.; Zhu, K.; Zhang, H. An Overview of the Development of EED Inhibitors to Disable the PRC2 Function. *RSC Med. Chem.* **2022**, *13* (1), 39–53. https://doi.org/10.1039/D1MD00274K.

(132) Huang, Y.; Sendzik, M.; Zhang, J.; Gao, Z.; Sun, Y.; Wang, L.; Gu, J.; Zhao, K.; Yu, Z.; Zhang, L.; Zhang, Q.; Blanz, J.; Chen, Z.; Dubost, V.; Fang, D.; Feng, L.; Fu, X.; Kiffe, M.; Li, L.; Luo, F.; Luo, X.; Mi, Y.; Mistry, P.; Pearson, D.; Piaia, A.; Scheufler, C.; Terranova, R.; Weiss, A.; Zeng, J.; Zhang, H.; Zhang, J.; Zhao, M.; Dillon, M. P.; Jeay, S.; Qi, W.; Moggs, J.; Pissot-Soldermann, C.; Li, E.; Atadja, P.; Lingel, A.; Oyang, C. Discovery of the Clinical Candidate MAK683: An EED-Directed, Allosteric, and Selective PRC2 Inhibitor for the Treatment of Advanced Malignancies. *J. Med. Chem.* **2022**, *65* (7), 5317–5333. https://doi.org/10.1021/acs.jmedchem.1c02148.