Methylation-based methods for studying chromatin structure

by

Anne M. Maslan

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Joint Doctor of Philosophy
with the University of California, San Francisco

in

Bioengineering

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Aaron Streets, Chair
Professor Ian Holmes
Professor Jasper Rine
Professor James Wells

Summer 2022

Abstract

Methylation-based methods for studying chromatin structure

by

Anne M. Maslan

Joint Doctor of Philosophy in Bioengineering

and Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley and University of California, San Francisco

Associate Professor Aaron Streets, Chair

DNA is decorated with chemical marks and proteins that allow our genome to encode countless distinct cellular states. Proteins in the nucleus interact with DNA and mediate methylation of cytosine bases, modifications to the histones around which DNA wraps, the degree of compaction of DNA, and the spatial localization of DNA within the nucleus. These epigenetic regulatory elements determine the binding of transcriptional machinery that control gene expression and give rise to cellular diversity. Measuring where proteins bind to the genome and how the epigenome regulates gene expression can reveal underlying mechanisms that control cell state and can help uncover how these mechanisms are modified in diseased states. In this dissertation, I describe a suite of tools that measure protein-DNA interactions by encoding these epigenetic features in exogenous methylation of genomic DNA and detecting these marks with DNA sequencing platforms.

First, I extend and optimize short-read sequencing techniques for measuring protein-DNA interactions with the goal of applying these methods to single cells. Single-cell measurements capture dynamics occurring in small populations of cells and can reveal coordinated features within a cell that cannot be measured with bulk methods. Short-read methods map features of chromatin structure through base conversion, fragmentation, or selective enrichment of short sequences of DNA and leverage high-throughput DNA sequencing to detect enrichment of these features genome-wide. Next, I detail a method that I developed collaboratively that leverages long-read sequencing to measure protein binding events. Long-read sequencing provides a new dimension in which to encode information about the chromatin structure of a cell because long-read sequencers can read out not only nucleotide bases, but also modifications to those bases. With this new encoding space, we take multi-omic measurements of endogenous DNA methylation directly together with other elements of chromatin structure (e.g., histone modifications, DNA accessibility, DNA spatial localization, and protein binding events) by encoding these elements as DNA base modifications.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Colleagues

I am so grateful to have been in the Streets Lab for my PhD. Aaron provided me space for creativity and project ownership. I left every meeting with Aaron eager to do the next experiment or try the newest analysis he suggested. When experiments were failing for months, Aaron would celebrate my careful troubleshooting experiments rather than pressure me to get results. Aaron has created a lab environment where lab members cheer each other on and support one another, and I feel so lucky to have been part of this group of researchers.

I worked most closely with Nick Altemose during my PhD. Nick is the best teacher I have ever had. The words that come to mind when I think of Nick are integrity, persistence, and curiosity. Nick does not drop a problem until every detail has been explored. We tested over 100 conditions for the DiMeLo-seq protocol to ensure we had the best assay, and Nick still to this day is testing additional optimizations on the side. Nick's curiosity is unmatched and once you experience it, he passes it on to you. I know he is going to make the best PI, instilling in his students his passion for science, being their biggest advocate, and having brilliant project ideas. Above all, Nick has become a lifelong friend. I will never forget one evening in lab telling Nick I wasn't excited about my wardrobe, and next thing I knew later that night he had created a whole clothing mood board for me. He is so supportive and takes the time to help with anything, whether it is learning to speak up, troubleshooting a failing experiment, or helping with a wardrobe refresh.

Carolina Rios-Martinez worked with me in the lab for three years. Mentoring Carolina was one of the most rewarding parts of my graduate school experience. I am so proud of the scientist Carolina is becoming. Carolina's optimism is infectious, and she taught me that every small victory, even just a good Qubit reading, should be celebrated. So many of my grad school memories are with Carolina… late night experiments listening to Shawn Mendes and Taylor Swift, zoom yoga together when Carolina was home her junior year because of the pandemic, and watching Carolina's K-Pop dance crew perform. Carolina spent months learning Korean because she loves K-Pop and Korean culture. I really admire this enthusiasm Carolina brings to every aspect of her life and her projects in lab.

I also had the privilege of mentoring Reet and Denise. Reet set up the backend for the dimelo Python package and always came to meetings with new ideas inspired by what she was learning in her classes. We shared our love of Taylor Swift and had so much fun at our weekend hackathons together. Denise was a joy to watch grow; she started in the lab never having done lab work and became the BluePippin expert. I am so excited to see what these young scientists do next!

I would like to thank all the other Streets lab members – Gabriel Dorlhiac, Anushka Gupta, Zoë Steier, Adam Gayoso, Rodrigo Cotrim Chaves, Ryan Keivanfar, Kim de Luca, Jeremy Marcus, and Boden Eakins. You all made the lab such a fun and supportive environment.

The Straight lab members – Kousik Sundararajan, Owen Smith, Rachel Brown, and Aaron Straight – were the most amazing collaborators. We all worked so efficiently and seamlessly together, and working on DiMeLo-seq with all of you was my favorite grad school project. That week in July when we were finishing the DiMeLo-seq preprint together, all working late and chatting over slack, stands out in my mind. It felt like real teamwork, and you all were the best teammates to have. I just can't believe many of us haven't met in person!

I would like to thank my thesis committee – Aaron Streets, Jim Wells, Jasper Rine, and Ian Holmes. Our annual meetings were always so helpful, and I appreciate your guidance during my PhD.

Friends

Ivana, I feel so lucky to have had you as my cohort buddy during the PhD. I remember sitting with you in our first-year home and you were sharing that you felt like we weren't having enough fun. I had never reflected on and thought about whether the amount of fun I was having was sufficient, and you really taught me to make sure I have things throughout the week to look forward to and that time with friends is really energizing. I am so excited about your grad school discoveries. Kelsey, my dog-walking yoga buddy, I am so happy we got randomly paired up during your visit weekend and that you eventually became my roommate. You are the most loyal and driven person I know. Chris, I loved our daily lunches during my UCSF rotation and always love talking to you. You really understand the experience of a #6. Katie Boisseree is such a thoughtful friend. Our SF park hangouts with blue bottle were the best weekend mornings. You are really like a sister to me. Thank you to Katelyn and Farrell Stein. Ben and I absolutely love spending time with you, and we love your daily snaps. Thank you to Taylor Bradley for fetching me from Stanley Hall for our evening fire trail hikes with Sadie.

Family

My grandparents – Poppy, Grandy, Grandma Patty, and Grandma Judy – were always the first people I'd email when my work was published. They are the best cheerleaders! I have the most amazing parents. Whenever I needed help, I just showed up at their house. During the DiMeLo-seq submission crunch, my mom made me food for a weekend and my dad set me up at his big desk. You've supported me in becoming a scientist from the very beginning. Having a home base with supportive parents is the biggest gift. Thank you to Colin for always bringing in humor when I get too serious and for sharing your love of dogs with the whole family. Thank you to Madeline, my friend turned sister-in-law, who taught me to get outside and adventure each day.

The best part of grad school was of course meeting my fiancée Ben. I will never forget that Santa Cruz retreat, spending time together for the first time on the beach with Ivana, Jon, Kelsey, and Nick. You are the kindest person I know, and I feel incredibly lucky to have met you. We have so many grad school memories together from coming by your poster at retreat to cheering you on at your marathon to pandemic road trips to Park City, Netarts, and San Diego to moving into our first home to raising a puppy together. Thank you to our puppy Tillie for having the cutest, most loving little personality. Family couch cuddles are one of the highlights of my day. Ben, I always look forward to coming home to you at the end of the day. I can't express how grateful I am to have met you. Ben, I love you!

Chapter 1

Introduction

Nearly all cells in the human body have the same DNA. How do cells with the exact same blueprint behave so differently? While the DNA sequence across distinct cells may be indistinguishable, there is a layer of information above DNA sequence that regulates gene expression. Components of this regulatory layer include chemical modifications to the DNA itself (e.g., cytosine methylation), chemical modifications to the proteins around which DNA wraps (e.g., histone methylation and acetylation), the accessibility of DNA, and the binding of key proteins. My dissertation focuses on new tools to explore this layer above DNA sequence that I refer to as chromatin structure. Chromatin structure encompasses modifications to the DNA itself, to the proteins around which DNA wraps, to the accessibility of DNA, to the 3D organization of DNA within the nucleus, to the DNA binding patterns of proteins (Figure 1.1).

In this dissertation, I develop sequencing-based methods for studying chromatin structure, where the goal is to measure the location of an epigenomic feature genome-wide with sequencing. I start with optimizing a method that relies on targeted methylation followed by methylation-sensitive digestion, selective amplification, and short-read sequencing. Short-read methods have been the gold standard and have been used to profile the binding patterns of thousands of proteins in human cells.[1] In order to extend these techniques to single-cell resolution, I modified existing protein-DNA interaction mapping assays to improve sensitivity and specificity and to allow for simultaneous measurement of multiple features of chromatin structure. Then, I collaboratively developed a new method that was enabled by the development of long-read, native DNA sequencing technologies, which allow us to encode an additional layer of information by depositing targeted exogenous methylation that is directly detected with sequencing.

1.1 The landscape of methods for studying chromatin structure with short-read sequencing

To study chromatin structure with short-read sequencing, the location of epigenetic features must be converted to ACTG. Many short-read methods enrich for genomic regions where features of interest are localized, which improves sensitivity and makes these methods well suited for genome-wide, single-cell analysis. For measuring protein-DNA interactions, protein-bound genomic regions are selectively amplified, and coverage serves as a proxy for protein binding. To measure chromatin accessibility, methods selectively cut and amplify accessible DNA. To detect CpG methylation, methods perform methylation-sensitive base conversion. Variations of these techniques combine base conversion for methylation detection together with the methods for measuring protein binding or chromatin accessibility. In this section, I provide an overview of current approaches and discuss the limitations of short-read methods.

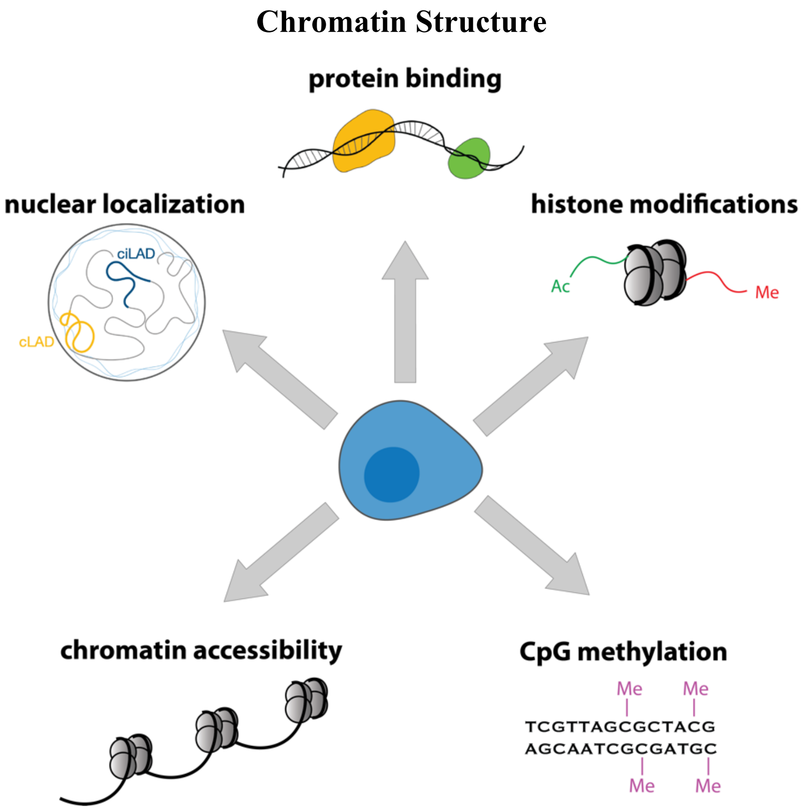


Figure 1.1 Overview of regulatory features that compose chromatin structure. Certain regions of the genome reside at the periphery of the nucleus (constitutive lamina associated domains, cLADs), and tend to be gene-poor and heterochromatic, while other more highly expressed and accessible regions of the genome reside at the interior (constitutive inter lamina associated domains, ciLADs). The accessibility of DNA, the endogenous CpG methylation, and the modifications on the histones that compose nucleosomes all influence the ability of key proteins to bind, ultimately mediating gene expression and cell state.

1.1.1 Methods to measure protein-DNA interactions

Methods for measuring protein-DNA interactions include ChIP-seq,[2–4] CUT&RUN,[5] CUT&Tag,[6] and DamID-seq.[7] These methods selectively amplify fragments of DNA from protein-bound regions and then map the fragments back to the reference to determine where on the genome the protein was bound. ChIP-seq, CUT&RUN, and CUT&Tag require an antibody to target a protein of interest. In ChIP-seq, protein-DNA interactions are detected by crosslinking the interacting protein and DNA, fragmenting the DNA, and then performing immunoprecipitation (IP) using a target-protein-specific antibody to pull down protein-DNA crosslinked complexes. IP enrichment yields short fragments of DNA are then amplified and sequenced. Protein-DNA complexes must survive shearing or digesting followed by many washing and purification steps, resulting in loss of sensitivity. In CUT&RUN and CUT&Tag, an antibody is also used to target a protein of interest, similar to ChIP-seq, but then DNA in the vicinity of a protein of interest is cut, either with MNase or Tn5. These methods use protein A, a cell wall protein from Staphylococcus aureus that binds a range of IgG antibodies, to recruit a DNA-modifying enzyme to the antibody-tagged protein. As a result, DNA is cut in the vicinity of the protein binding event and subsequently enriched before sequencing. These approaches improve sensitivity compared to ChIP-seq because physical separation of protein-bound DNA regions and extensive washes are not required. While using antibodies to target a protein *in situ* enables detection of post-translational modifications and makes these methods more modular for targeting a range of proteins from a variety of cell types, all these methods are limited by antibody availability and performance.

DamID-seq uses a different approach that does not require antibodies. Instead, DamID-seq deposits targeted methylation to mark protein-DNA interactions. DNA adenine methyltransferase (Dam) is an enzyme from bacteria that methylates adenines in the GATC sequence context. In DamID-seq, a fusion between a protein of interest (POI) and Dam is introduced to live cells. When the POI binds DNA, the DNA is methylated by Dam that is tethered to the POI. These ^{m6}A marks are highly stable in eukaryotic cells, which do not tend to methylate (or demethylate) adenines.[8] A methylation-sensitive restriction enzyme DpnI, which cuts at methylated adenines within the GATC context, is then used to convert this methyl mark into a signal that can be detected by Illumina sequencing. Digested fragments are selectively amplified, and fragment ends mark where methylation, and therefore protein binding, occurred. As with CUT&RUN and CUT&Tag but unlike ChIP-seq, no physical separation of protein-DNA complexes is required, allowing for improved sensitivity that makes the method well suited for single-cell analysis.

Beyond the requirement of a genetically tractable system for expressing a Dam-POI fusion, optimizing Dam fusion expression levels and induction times present major hurdles in adopting this method compared to antibody-based approaches. Importantly, DamID-seq does not profile the endogenous protein's binding patterns and instead measures an introduced fusion protein that may behave differently than the endogenous protein. Binding patterns for proteins with specific post-translational modifications cannot be measured, unless further genetic engineering is performed with approaches like EpiDamID.[9] Despite these limitations, a key feature of DamID-seq is that it leaves a biorthogonal mark of protein binding and does not fragment the genome until methylation-sensitive digestion after DNA extraction. DamID-seq also produces an integrated signal over time of where a protein has been bound during the induction period. These features further make DamID-seq particularly compatible with single-cell measurements and paired measurements of protein binding together with other modalities like imaging. In the first section of my dissertation, I detail work I did to profile different Dam mutants for expression level optimization, to verify minimal gene expression changes with Dam expression, and to improve a system for imaging Dam methylation. I then extended DamID-seq to more closely resemble antibody-based approaches that do not require genetic manipulation and that allow for profiling post-translational modifications.

1.1.2 Methods to measure chromatin accessibility

Methods for measuring chromatin accessibility include ATAC-seq,[10] DNase-seq,[11] MNase-seq,[12] and a modified DamID-seq protocol using untethered Dam.[13] In ATAC-seq, a hyperactive transposase, Tn5, is used to both cut and add adapters for sequencing to accessible DNA. DNase-seq and MNase-seq similarly cut accessible DNA; however, these methods require additional steps to attach adapters needed for sequencing. In DamID-seq with untethered Dam, Dam methylates accessible DNA that is then enriched with the DamID-seq assay. These methods have been used to profile accessibility in individual single cells. Single-cell ATAC-seq has gained widespread adoption and commercial single-cell platforms now offer scATAC-seq assays.[14]

1.1.3 Methods to measure endogenous CpG methylation

Methods for measuring endogenous CpG methylation include bisulfite sequencing,[15] Enzymatic Methyl-seq (EM-seq),[16] and TAPS.[17] These methods rely on methylation-sensitive base conversion. Bisulfite sequencing uses bisulfite treatment to deaminate unmethylated cytosines, converting them to uracil; methylated cytosines remain unconverted. In PCR, uracil is converted to thymine. At sequencing, locations where thymine is detected instead of the reference cytosine are marked as unmethylated, while bases that match the cytosine reference are considered methylated because they were protected from base conversion. In EM-seq, TET2 oxidizes methylated bases, protecting them from deamination by APOBEC, while unmethylated cytosines are deaminated to uracil. TAPS combines TET oxidation of methylated cytosines and pyridine borane reduction of the oxidation product to dihydrouracil. PCR then converts dihydrouracil to thymine so methylated cytosines are detected as thymine at sequencing. While these three methods all rely on base conversion, there are some alternatives, such as MspJI-seq,[18] that use selective fragmentation with a methylation-sensitive restriction enzyme, similar to the other methods discussed above for studying chromatin structure.

1.1.4 Multi-omic methods

Combinations of these methods have been developed to take multi-omic measurements of protein binding or DNA accessibility together with CpG methylation (BisChIP-Seq/ChIP-BS-Seq,[19,20] methyl-ATAC-seq,[21] EpiMethylTag[22]). However, these multi-omic measurements have not been taken robustly in single cells because they rely on lossy and harsh bisulfite treatment that degrades DNA. With bulk methods, one could just as easily perform the two assays independently to recover the same information. New information about cell states and about the relationship between regulatory elements lies in the linked information about multiple facets of chromatin structure from the same cell or molecule.

1.2 Long-read, native sequencing technologies enable new methods for studying chromatin structure

While methods that rely on amplification and short-read sequencing are constrained to storing epigenomic information in DNA sequence space, new sequencing technologies that sequence native, long molecules of DNA can encode information in an additional layer – methylation. These long-read sequencing technologies take a fundamentally different measurement of DNA compared to short-read sequencers. Short-read sequencers create a discrete digital sequence of A, C, T, G by building a complementary strand to DNA amplicons using fluorescently labeled bases. Data science advances for working with digital data and strategies for improving experimental throughput using DNA sequence barcoding have transformed biology. Now, long-read sequencing methods and machine learning approaches are poised to do the same. Long-read sequencers make a direct physical measurement of the DNA molecule, taking an analytical rather than digital measurement. These analytical measurements of DNA can now be made in a high-throughput way and can be decoded using advances in machine learning. Taking physical measurements of DNA molecules directly as opposed to amplicons provides an opportunity to read out other chemical properties of the molecule, including chemical modifications to DNA bases. We take advantage of this new channel to encode additional layers of information on DNA. In particular, we encode features of chromatin structure as methylation marks that are detected directly with sequencing.

1.2.1 Oxford Nanopore Technologies

Oxford Nanopore Technologies (ONT) sequencers detect base-specific disturbances in electric current as DNA is passed through a narrow pore. Each device contains thousands of microwells, each with a single nanopore embedded within an electrically resistant membrane. As DNA passes through the pore, the bases present within the pore produce a base-specific disturbance in the electric current across the membrane. Importantly, methylated bases produce a disturbance that is distinct from their unmethylated counterpart. The fact that native DNA molecules, as opposed to amplified clones, are sequenced directly is just one of the key features of ONT devices. ONT sequencers can also sequence much longer reads compared to those generated with Illumina, reaching lengths on the scale of millions of bases.

1.2.2 Pacific Biosciences

The Pacific Biosciences (PacBio) Sequel IIe relies on sequencing by synthesis, just as in Illumina sequencing. However, rather than performing clonal amplification, a single native DNA molecule is immobilized in a zero-mode waveguide and a single anchored polymerase incorporates labeled bases. Importantly, if a modified base is encountered, the polymerase stalls when trying to incorporate the complementary base. This pause is used to determine the methylation status of a given base. During preparation of DNA for sequencing, the template can be circularized to allow the polymerase to make multiple passes synthesizing the DNA. A consensus sequence is then determined to produce more accurate basecalling than could be achieved with a single pass of DNA synthesis. Just as with ONT sequencers, PacBio can also sequence long reads of DNA; however, read lengths are limited to tens of kilobases.

1.2.3 Methods for studying chromatin structure with long-read, native sequencing

With the development of new native sequencing technologies from ONT and PacBio, new methods for detecting chromatin structure have been developed that make use of the ability to detect methylation on DNA. Endogenous DNA methylation marks can be read on these sequencers without any additional base conversion or amplification steps.[23] As an extension, exogenous methylation marks that are deposited on the DNA as part of an assay can also be detected. A series of methods – Fiber-seq,[24] SAMOSA,[25] SMAC-seq,[26] NanoNOMe,[27] and MeSMLR-seq[28] – were developed to measure chromatin accessibility through deposition of exogenous DNA methylation. These methods expose chromatin to methyltransferases that preferentially methylate accessible DNA. Methylated bases are then detected directly with ONT and PacBio sequencers. With these methods, both endogenous CpG methylation and chromatin accessibility can be measured on the same molecules. While these methods can indirectly measure protein binding in a protein-agnostic way by detecting the “shadow” (exogenous methylation depletion) of where a protein is bound, they cannot determine protein binding events for a specific protein of interest.

1.2.4 Long-read sequencing technologies expose new regions of the genome for exploration

Long-read sequencing techniques have not only enabled new methods for studying the epigenome but have also exposed new regions of the genome with unknown epigenomic features. The Telomere-to-Telomere consortium recently released a new human reference genome that is 5-10% larger than the previous reference genome size.[29] These newly sequenced regions could not be sequenced with short-read methods because the regions are so repetitive that no sequence variation exists to anchor short reads during mapping. This new reference now enables exploration of the epigenome within these previously understudied regions of the genome.

1.3 Simultaneous measurements of multiple chromatin features on single molecules

While short-read methods for probing chromatin structure have achieved single-cell resolution, the data are so sparse that analyses often cluster cells in the context of the whole feature profile for that cell. Many single-cell multi-omic tools make use of a second or third modality for the sole purpose of classifying cell type and state rather than to understand mechanisms within a cell. Taking measurements on single molecules instead switches the focus to the interplay between features on a single molecule derived from a given locus within a cell. Encoding multiple aspects of chromatin structure as base modifications on single molecules allows for the interactions between molecular layers to be studied.

1.4 Dissertation Overview

In this dissertation, I detail three projects I completed during my PhD, all of which were collaborative efforts. First, I optimized an existing method for measuring protein-DNA interactions with targeted methylation and short-read sequencing, called DamID-seq, and extended its sensitivity in single-cell applications. I characterized Dam mutants to increase sensitivity and specificity, evaluated gene expression effects of Dam and Dam fusion proteins, and improved imaging methods for determining spatial localization of protein-DNA interactions. In an effort to improve signal-to-noise and to perform protein-DNA interaction mapping without genetic manipulation, I also developed an *in situ* extension of DamID-seq. To allow for CpG methylation detection in addition to either protein binding or DNA accessibility, I also extended DamID-seq to include mCpG-sensitive fragmentation and enrichment steps. Next, I developed Directed Methylation with Long-read sequencing (DiMeLo-seq). This method allows for simultaneous measurement of protein-DNA interactions, chromatin accessibility, and endogenous CpG methylation. Because DiMeLo-seq uses long-read sequencing technologies, reads can be mapped to repetitive regions of the genome and can be phased to determine haplotype-specific protein-DNA interactions. Finally, I’ve developed an analysis software package for analysis of DiMeLo-seq data and have applied this analysis framework to study various histone modifications. In ongoing efforts, I am working on multiplexing DiMeLo-seq to measure protein-DNA interactions for two proteins simultaneously.

Chapter 2

Short-read methods for mapping chromatin structure

2.1 Introduction

Each time a cell divides, the careful organization of six meters of DNA must be re-established in the nuclei of each daughter cell. Levels of DNA organization exist from the location of a given locus within the nucleus to topologically associated domains and loops to specific contacts between genomic regions like enhancers and promoters. Here, we study one broad level of genome organization by measuring whether genomic regions are at the periphery or interior of the nucleus. The radial location of a given region of the genome within the nucleus acts as a regulatory mark and is correlated with gene content and gene expression. To study this, we performed DamID targeting LMNB1, a component of the nuclear lamina. Previous studies have published Lamina-Associated Domain (LAD) maps in various human cell lines.[30,31] LADs are large (median 500 kb) genomic stretches that reside at the periphery of the nucleus and comprise up to 30% of the genome.[32] LADs tend to be gene-poor, heterochromatic, and transcriptionally less active compared to regions at the interior of the nucleus (reviewed by van Steensel and Belmont, 2017 and Buchwalter et al., 2019).[33,34]

While certain genomic regions reside at the periphery or interior of the nucleus across cell types, there are regions that are variably associated with the lamina across cells and across cell cycle phase. Bulk methods average populations of cells and fail to capture dynamics occurring in small populations of cells and in single cells during differentiation and the cell cycle. Single-cell resolution is required to capture heterogeneity in lamina association. Because DamID avoids antibody binding, physical separations, and intermediate purification steps, it lends itself to single-cell applications. DamID has been successfully applied to sequence lamina-associated domains (LADs) in single cells in a one-pot reaction, recovering hundreds of thousands of unique fragments per cell.[30] Nicolas Altemose developed a microfluidic device for paired imaging of lamina association and sequencing of lamina-DNA interactions with scDamID-seq performed on the same single cells.[35] LADs are visualized by expressing ^{m6}A-Tracer, which contains green fluorescent protein and a domain that binds specifically to methylated GATC sites. Imaging with ^{m6}A-Tracer produces a characteristic ring at the nuclear periphery in confocal fluorescence imaging.[36] This paired imaging technology allowed us to select cells for sequencing based on morphological features of LADs as determined by imaging. Beyond this application, this platform for pairing imaging and sequencing data could be applied to study, for example, how the dynamic remodeling of chromatin proteins across the genome in developing cells relates to the localization of those proteins in the nucleus. Imaging prior to sequencing also allows for the identification and sorting of complex cytological phenotypes in cells, such as the presence of micronuclei and other nuclear abnormalities that would be difficult or impossible using common fluorescence-activated sorting methods.

Here, I focus on my primary contributions to this study which were to: (i) generate and analyze bulk DamID-seq data to validate single-cell LAD maps and characterize the performance of different Dam mutants, (ii) perform RNA-seq experiments and analysis to determine the effects on gene expression of expressing Dam and Dam-LMNB1, and (iii) develop an improved molecule for imaging Dam methylation. Much of this work was published in Altemose et al., 2020.[35] In an effort to create a method for similarly marking protein-bound regions with methylation, but allowing one to do so without performing genetic manipulation, I then developed an *in situ* version of the DamID-seq protocol. I further extended *in situ* DamID-seq to create a multi-omic measurement of chromatin accessibility or protein binding together with CpG methylation using untethered Dam to mark open chromatin or targeted Dam to mark protein binding, followed by joint digestion with mA-sensitive and mC-sensitive restriction enzymes.

2.2 Validating lamina contact maps from single-cell DamID-seq

To validate single-cell DamID-seq performance on the microfluidic device Nicolas Altemose developed, I produced benchmark bulk DamID-seq datasets in HEK293T. I transiently transfected HEK293T cells with plasmids encoding drug-inducible Dam-LMNB1 and Dam only, induced expression for ~18 hours, performed DamID-seq, and then developed a Snakemake pipeline for automated and reproducible analysis of DamID-seq data. The Dam only condition provides an accessibility control.[37] The unfused Dam enzyme, when not tethered to the nuclear lamina, preferentially methylates open chromatin and is a useful control for estimating the propensity of each genomic region to be methylated.[31,38,39] To report the degree to which a region of the genome interacts with the lamina, we binned the genome into 100-kb bins and performed differential expression analysis using DESeq2, reporting the log ratio between the coverage in the Dam-LMNB1-expressing cells and the coverage from the Dam-only-expressing cells (Methods: Bulk and Single-cell DamID analysis). This value is a measure of the targeted increase in coverage over background. Then, for each single cell, we made binary calls for whether a bin was in contact with the lamina (Methods: Bulk and Single-cell DamID analysis). Visually the binary LAD maps from single cells roughly correspond with the bulk DamID-seq results (Figure 2.1a). Aggregating across the single cells expressing Dam-LMNB1, we reported strong correlation with the bulk DamID coverage (Figure 2.1b).

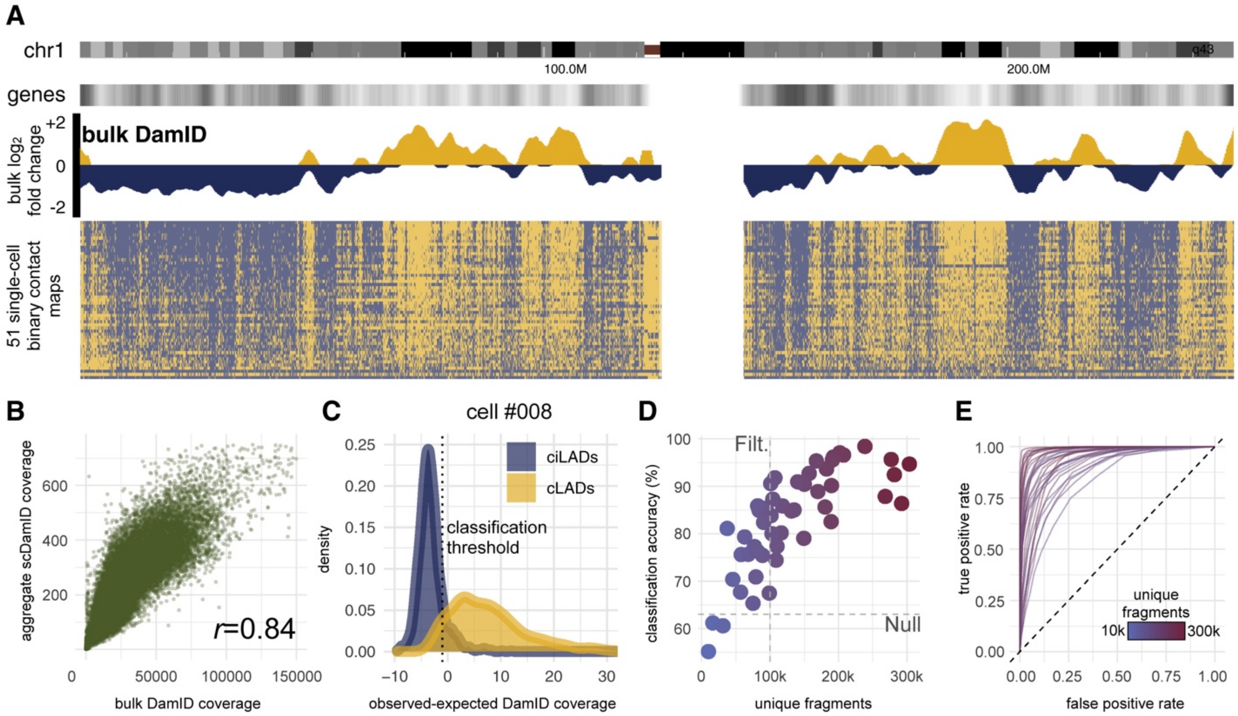


Figure 2.1 Validation of μDamID Sequencing Data. **a,** Comparison of bulk and single-cell LMNB1 DamID sequencing data across all of human chromosome 1 (chr1 ideogram in first track from top), normalized to bulk Dam-only data. Positive values (gold) represent regions associated with the nuclear lamina, which tend to have lower relative gene density (shown in second track). The bulk data (third track) are shown as $\log_2$-fold-change values between Dam-LMNB1 and Dam-only samples. Each row of the binary contact map (fourth track) represents a single cell, sorted from top to bottom by genome-wide control set classification accuracy. **b,** Scatterplot comparing raw Dam-LMNB1 sequencing coverage in bulk versus aggregated single-cell samples in 100-kb bins. **c,** Normalized coverage distributions within positive (cLADs, gold) and negative (ciLADs, blue) control sets in one cell (#008) expressing Dam-LMNB1. The threshold that distinguishes these sets with maximal accuracy is shown as a vertical dotted line. **d,** The maximum control set classification accuracy for each of 50 Dam-LMNB1 cells versus the number of unique DpnI fragments sequenced for each cell (also indicated by color gradient; outlier cell #007 was excluded). A coverage threshold of 100k fragments used for downstream analyses is indicated, as well as the null accuracy achieved after scrambling values in all bins across the genome (63%). **e,** Receiver-operator characteristic curves for 31 Dam-LMNB1 cells above the 100k coverage threshold.

We defined positive and negative control genomic sets using both previously annotated lamina contact maps across many cell lines and our own bulk DamID-seq data. Positive controls were derived from 100-kb bins across the genome that were previously annotated in other human cell lines to be strongly associated with the nuclear lamina (referred to as constitutive LADs or cLADS) and further filtered to have the highest bulk DamID scores in HEK293T cells from our bulk DamID experiments. These bins are therefore most likely to have high contact frequencies (CFs) in individual cells.[30] Negative controls were similarly determined using bulk data to be consistently not associated with the nuclear lamina across cell types and in our cells (referred to as constitutive inter-LADs or ciLADS), making them most likely to have low CFs in individual cells.[30] These stringent control sets constitute roughly 4% of the genome each.

For each single cell expressing Dam-LMNB1, we computed the distribution of its normalized sequencing coverage in bins from the positive (cLAD) and negative (ciLAD) control regions (Figure 2.1c), with the expectation that cLADs have high coverage, and ciLADs have little or no coverage in each cell. Given these control distributions, we chose a coverage threshold to maximally separate the known cLADs and ciLADs. Across the 51 Dam-LMNB1 cells, we determined thresholds that distinguish the known cLADs and ciLADs with a median accuracy of 85% before any filtering (versus 63% if all bins are scrambled), which correlates positively with the number of unique DpnI fragments sequenced per cell, a measure of library complexity (Figure 2.1d). Because we used a transient transfection system, expression levels of Dam-LMNB1 varied widely from cell to cell, reducing classification accuracy in some cells with high noise levels due to background methylation. We filtered higher-noise cells using a threshold of unique covered fragments, leaving 31 Dam-LMNB1 cells with a median classification accuracy of 90% (range 74%–98%, Figure 2.1d). Our classification approach enables inference of expected error rates for each bin's coverage level in each cell, providing a framework for data normalization, interpretation, and further inference. These error rates can be represented with receiver operating characteristic (ROC) curves for each cell, showing the empirical trade-off between false-positive and false-negative classifications at varying normalized coverage thresholds (Figure 2.1e).

We next computed pairwise correlations between the raw coverage for all single cells with each other, with the bulk data, with aggregated published single-cell DamID data (from Kind et al., 2015),[30] and with the number of annotated genes across 100-kb bins genome wide. After removing low-complexity cells, we performed unsupervised hierarchical clustering on these datasets and produced a heatmap of their pairwise correlations (Figure 2.2a). We found that the 3 Dam-only single cells cluster with each other, along with the bulk Dam-only data, with the Kind et al. Dam-only data, and with the number of genes, as expected. The Dam-LMNB1 cells cluster separately with each other, with the bulk Dam-LMNB1 data, and with the Kind et al. Dam-LMNB1 data, confirming that these sequencing data are measuring meaningful biological patterns in single cells. These clusters also reflect expected nuclear spatial distributions of methylation reported by ^{m6}A-Tracer fluorescence (Figure 2.2b-d). Notably, one Dam-LMNB1 cell with unexpectedly high fluorescence signal in the nuclear interior contained a methylation profile that appeared intermediate between the Dam-only and other Dam-LMNB1 cells, perhaps owing to high Dam-LMNB1 expression (Figure 2.2a). This illustrates how spatial information can be used to validate DamID with joint single-cell imaging and sequencing measurements.

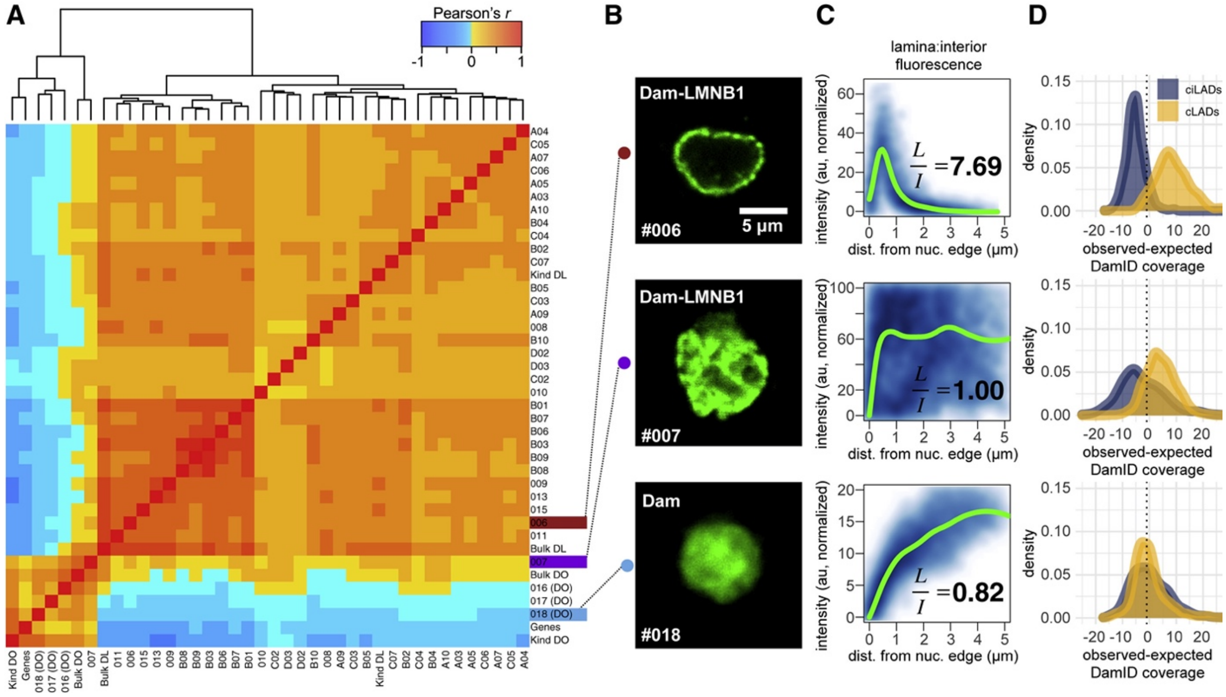


Figure 2.2 Genome-wide Comparisons of Sequencing Data and Relation to Imaging Data. **a,** Pairwise cell-cell Pearson correlation heatmap for raw sequencing coverage in 100-kb bins genome-wide, with dendrogram indicating hierarchical clustering results. Cell identifiers label each row (first batch 00**, second batch A–D**). DL, Dam-LMNB1; DO, Dam-only; Genes, number of RefSeq genes in each bin; Kind, aggregated single-cell data from Kind et al., 2015;[30] Bulk, bulk HEK293T DamID data from this study. **b,** Confocal fluorescence microscopy images of ^{m6}A-Tracer GFP signal from 3 cells: one expressing Dam-only (#018), one expressing Dam-LMNB1 but showing high interior fluorescence (#007), and one expressing Dam-LMNB1 and showing the expected ring-like fluorescence at the nuclear lamina (#006). **c,** Normalized pixel intensity values plotted as a function of their distance from the nuclear edge (blue), with a fitted loess curve overlaid (green). Ratios of the mean normalized pixel intensities in the lamina (<1 µm from the edge) versus the Interior (>3.5 µm from the edge) are printed on each plot. **d,** DamID sequencing coverage distributions for each of the cLAD or ciLAD control sets (as in Figure 2.1c).

2.3 Evaluation of Dam mutants

With the goal of increasing sensitivity and specificity of protein-DNA contact detection, we compared bulk DamID results with wild type Dam and with a mutant of Dam. We used a mutant of Dam (V133A),[40] which is predicted to have weaker methylation activity than the wild-type allele on unmethylated DNA, and we hypothesized that it would reduce background methylation, similar to weakened Dam mutants previously engineered to improve methylation specificity.[41] To test this, we performed bulk DamID experiments comparing the mutant and wild-type alleles and found that the V133A mutant allele provides more than 2-fold greater signal-to-background compared to the wild-type allele (Figure 2.3).

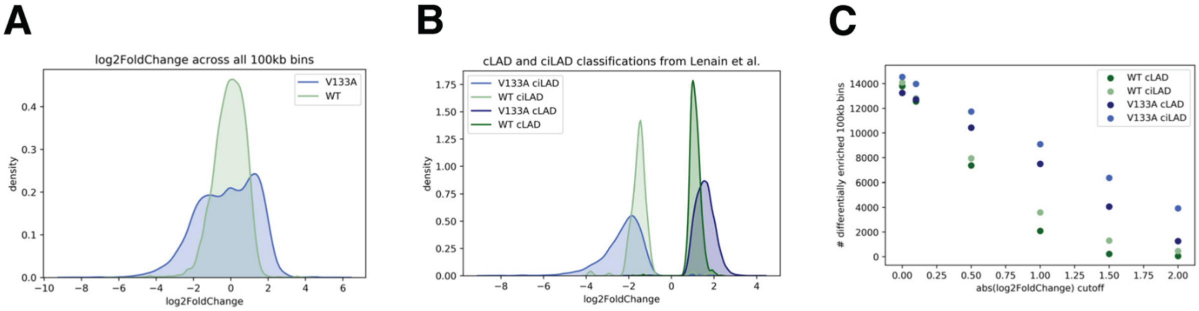


Figure 2.3 Comparing Dam mutants. **a,** Kernel density estimate of log2FoldChange from DESeq2 differential enrichment analysis of Dam-LMNB1 coverage compared to Dam-only as reference. With V133A, more extreme log2FoldChange values are observed with greater separation between the positive and negative log2FoldChange peaks. In other words, compared to wild-type, the V133A Dam-LMNB1 and Dam-only signals are more distinct. **b,** Kernel density estimate of log2 Fold Change, with cLAD/ciLAD classification from Lenain et al. 2017 indicated, shows greater separation for cLAD and ciLAD signal with V133A. **c,** V133A has higher sensitivity than WT, with more differentially enriched regions at each log2FoldChange threshold for calling significant differential enrichment.

2.4 Gene expression effects of Dam expression and cell-type-specific vLAD characterization

We performed RNA sequencing in cells that were untreated or transfected with Dam-only, Dam-LMNB1, or ^{m6}A-Tracer, and we found only two differentially expressed genes (Figure 2.4). This corroborates similar published findings by others showing that Dam expression and adenine methylation have little or no effect on gene expression in HEK293T cells.[41]

We further used this RNA-seq data to explore regions of the genome that are variably associated with the lamina. In any given cell, only a subset of potential LADs come into contact with the lamina, and this subset can vary stochastically between cells.[36] While most LADs at the lamina appear to remain in stable contact with the lamina throughout interphase, some LADs have been shown to move dynamically short distances toward and away from the lamina within the same cell over time,[36] also potentially contributing to cell-to-cell variability in LADs. Single-cell DamID provides a unique opportunity to identify LADs that vary within a population of cells of the same type. Our RNA-seq data allows us to characterize these regions in terms of the expression levels of the genes within these regions.

To measure this variability, at each bin in the genome, we counted the number of Dam-LMNB1 cells (out of 31 total cells) in which that bin was classified as having laminar contact to estimate its contact frequency (CF),[30] and we developed a method for propagating measurement and sampling uncertainty when inferring the true CF of each bin (Methods: Calling vLADs, Figure 2.5a, Figure 2.5b). As expected, bins belonging to the cLAD control sets have high CFs and lower gene expression while those in the ciLAD control sets have low CFs and higher gene expression (Figure 2.5a,c,d and Figure 2.6a). Furthermore, we found that CFs for each bin correlated well overall with published single-cell CFs from a different cell line, KBM7 (r = 0.8, Figure 2.6a; Kind et al, 2015).[30]

To identify variable LADs, we defined a conservative set of bins with intermediate CFs between 33% and 66% (Methods: Calling vLADs). We hypothesized that these stringently defined regions, which comprise 8% of the genome, would be more gene rich and have higher gene expression than cLADs, given their dynamic positioning in cells. Indeed, these variable LADs show intermediate gene density and intermediate bulk gene expression levels compared with the control sets of cLADs and ciLADs (Figure 2.5c,d), consistent with these regions being variably active within different cells.

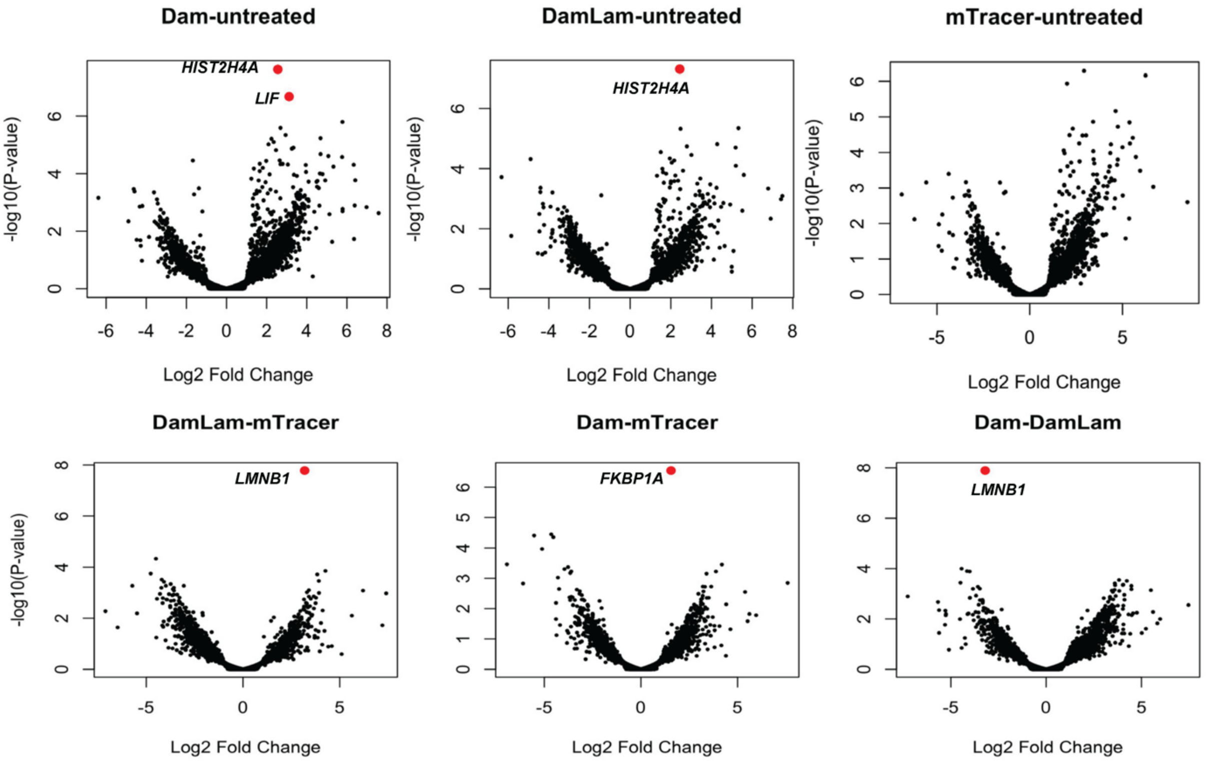


Figure 2.4 Examining effect of Dam on gene expression. Significantly differentially expressed genes (logFC significantly > 1 and adjusted p-value < 0.01) are indicated in red for bulk HEK293T cells transfected with Dam, Dam-LMNB1, m6ATracer, or no treatment control. Differentially expressed genes compared to no treatment control are HIST2H4A and LIF for Dam, HIST2H4A for Dam-LMNB1, and no genes for ^{m6}A-Tracer. When comparing Dam to ^{m6}A-Tracer, the only differentially expressed gene is FKBP1A, which is expected given the mutated FKBP1A-derived destabilization domain tethered to Dam in our construct. When comparing Dam-LMNB1 to ^{m6}A-Tracer, the only differentially expressed gene is LMNB1, which is again expected given LMNB1 is expressed from the Dam-LMNB1 construct itself.

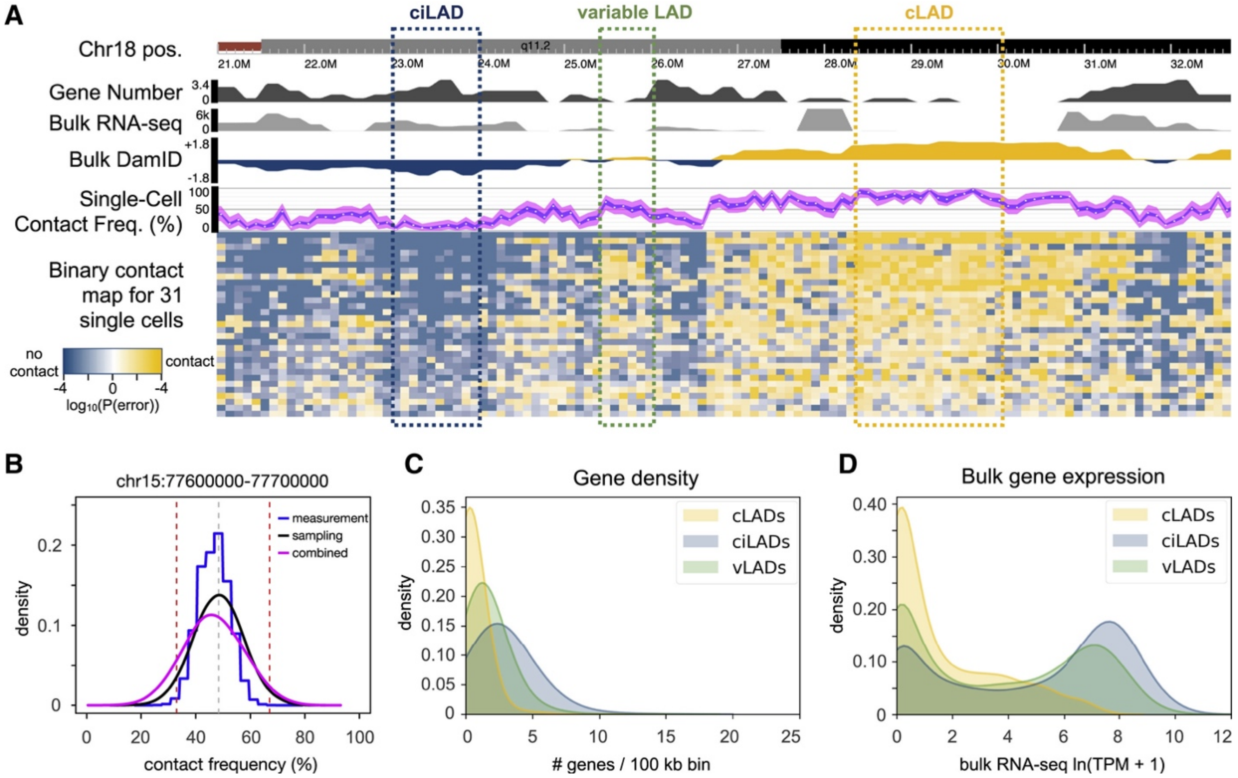


Figure 2.5 Identification and Characterization of Variable LADs in HEK293T Cells. **a,** browser screenshot from chr18:21–33 Mb. The first track shows the chromosome ideogram and coordinates. The second track reports the number of RefSeq genes falling in each 100-kb bin. The third track reports the mean transcripts per million (TPM) value for each gene within each bin from bulk RNA sequencing data from untreated HEK293T cells. The fourth track reports the bulk DamID $\log_2$ fold change values as in Figure 2.1a. The fifth track indicates the CF estimate for each bin (white point), with a blue ribbon indicating the 95% confidence interval for the sample CF (measurement error), and the magenta ribbon indicating the 95% confidence interval for the population CF (measurement + sampling error). The sixth track shows binary contact calls for each bin (columns) in each cell (rows). Shades of gold and blue indicate bins classified as having lamina contact or no lamina contact, respectively, with darker shades indicating higher confidence in the classification (smaller measurement error probability). Annotated cLADs and ciLADs are indicated by gold and blue boxes, respectively, with a variable LAD region (vLAD) in green. **b,** For one bin in a different region, a comparison of measurement (blue) and sampling (black) distributions, along with a combined distribution (magenta) used for CF inference with propagated measurement uncertainty (as shown in A track 5). The gray vertical dotted line is the point estimate for that bin, and red dotted vertical lines are drawn at the vLAD CF thresholds (33% and 66%). **c-d,** Distributions of the number of genes (c) or mean TPM per gene (d) per 100-kb bin for each of the sets of cLADs, ciLADs, or vLADs.

We then explored whether these variable LADs were conserved in another human cell type. We found that the CFs of bins containing variable LADs identified in HEK293T cells varied widely in KBM7 cells (Figure 2.6a), suggesting only a small subset of these LADs are variable in both cell types, consistent with prior observations that regions with intermediate CFs are more likely to have different bulk DamID signals across cell types.[30] Comparison of bulk RNA expression levels in bins that were classified as high, intermediate, or low CF in each cell type corroborated the inverse relationship between single-cell CF and bulk gene expression observed previously (Kind et al., 2015;[30] Figure 2.6b-h). For example, as regions shift from intermediate CFs to high CFs in one cell type as compared with the other, we observe a corresponding decrease of gene expression (Figure 2.6e,h). These observations support the notion that the nuclear lamina serves as a dynamic regulatory element, not only between cell types but within a given cell type.[13]

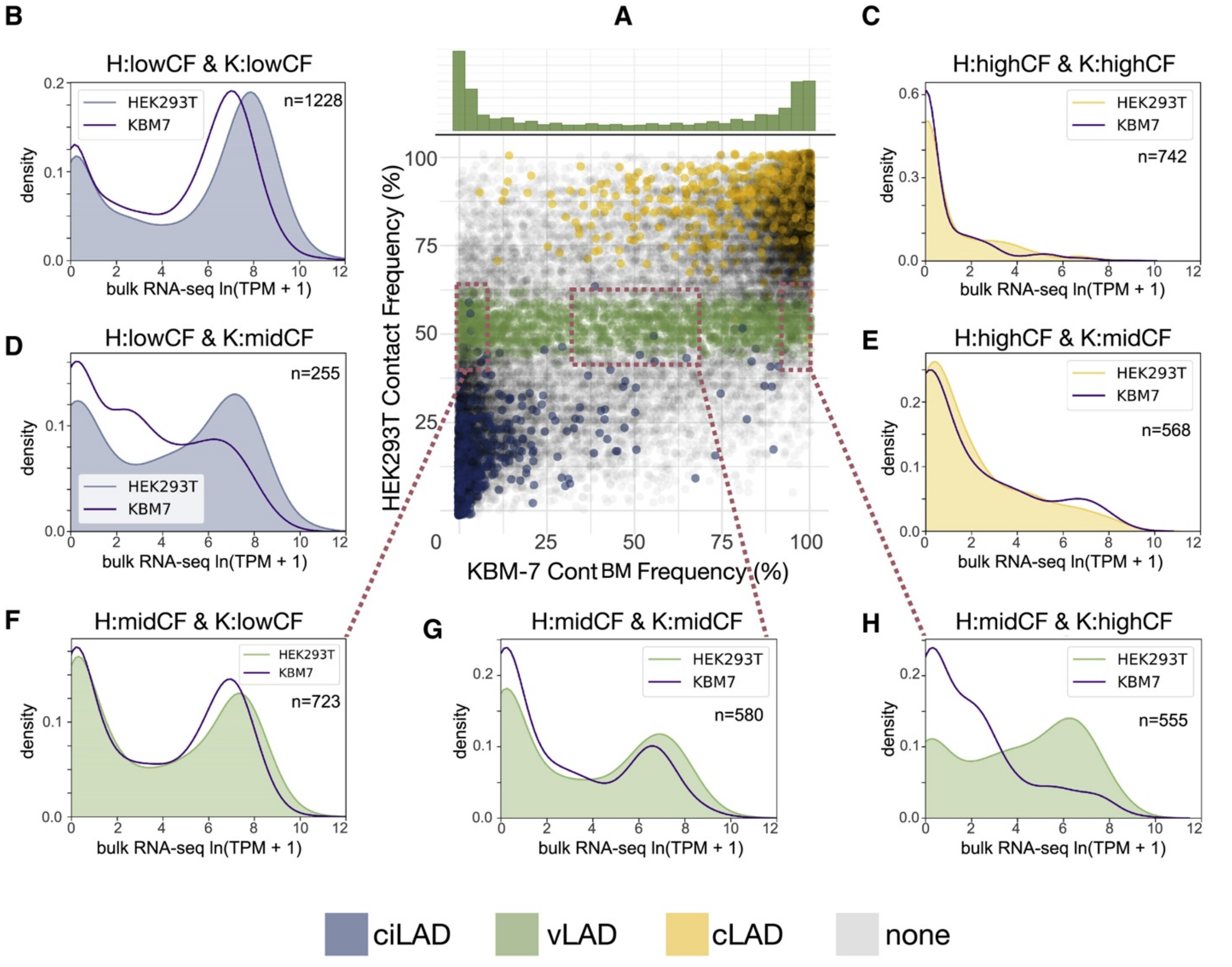


Figure 2.6 Comparing Single-Cell CFs between Cell Types. **a,** A scatterplot of the CF estimates in HEK293T cells (this study) versus KBM-7 cells (Kind et al., 2015)[30] across all bins in the genome. Each point is colored if the corresponding bin belongs to the cLAD (gold), ciLAD (blue), or vLAD (green) sets defined in HEK293T. Above the scatterplot is a histogram showing the KBM-7 CF distribution for all bins defined as vLADs in HEK293T, illustrating vLAD differences between cell types. **b-h,** Density plots indicating the relative distributions of bulk RNA sequencing

coverage (transcripts per million) in each cell type, within bins classified as low CF (<5% CF, with high expression), middle CF (33%–66% CF, with intermediate expression), or high CF (>90% CF, with low expression) in each cell type, allowing for comparison of cell-type specific expression levels in bins that have low CF in both HEK293T cells and KBM7 cells (B), high CF in both HEK293T and KBM7 (C), low CF in HEK293T and middle CF in KBM7 (D), high CF in HEK293T and middle CF in KBM7 (E), middle CF in HEK293T and low CF in KBM7 (F), middle CF in both HEK293T and KBM7 (G), and middle CF in HEK293T and high CF in KBM7 (H).

2.5 Improved ^{m6}A-Tracer for imaging protein-DNA interactions

While the ^{m6}A-Tracer technology was key to creating this multimodal measurement of paired imaging and sequencing data and provided an important quality control check to select cells for trapping and sequencing, we realized an important limitation of the ^{m6}A-Tracer technology, which is that the ^{m6}A-Tracer protein localizes to the nucleus even in cells expressing no Dam (Figure 2.7a,b). One consequence is that cells with Dam and cells without Dam are nearly indistinguishable (Figure 2.7b), and cells with overexpressed ^{m6}A-Tracer show high background fluorescence levels in the nuclear interior even when co-expressing Dam-LMNB1 (Figure 2.7b). The only way to prevent this background issue is to carefully tune the expression level of ^{m6}A-Tracer so that the copy number of ^{m6}A-Tracer proteins does not exceed the number of available methylated GATC sites. This tuning would have to occur separately for any new Dam fusion protein. In a heterogeneous expression system like the one used here, since ^{m6}A-Tracer and Dam are expressed from separate plasmids, only a small fraction of cells have the correct ratios of expression to produce sharp laminar rings with low background in the nuclear interior (Figure 2.7b).

No cryptic nuclear localization sequences were detected in ^{m6}A-Tracer (Methods: ^{m6}A-Tracer-NES) nor are human cells likely to contain any significant background levels of ^{m6}A without Dam.[8] Instead, its default nuclear localization may arise from a weak interaction between genomic DNA and the DNA-binding domain of ^{m6}A-Tracer, combined with the ability of ^{m6}A-Tracer to diffuse freely through nuclear pores given its small size (Figure 2.7a). We hypothesized that adding a nuclear export signal (NES) to ^{m6}A-Tracer might overcome its weak affinity for DNA and keep any unbound copies of the protein sequestered in the cytoplasm. We found that the HIV-1 Rev NES sequence fused to either terminus resulted in robust localization of ^{m6}A-Tracer to the cytoplasm in cells not expressing Dam (Figure 2.7c and Figure 2.8), and for downstream experiments, we proceeded to use the C-terminal fusion, which we call ^{m6}A-Tracer-NES.

While the NES appears to prevent nonspecific ^{m6}A-Tracer interactions with DNA, it does not overcome on-target binding to Dam-methylated DNA. When Dam was co-expressed, the localization of ^{m6}A-Tracer-NES shifted almost entirely from the cytoplasm to the nucleus (Figure 2.7b). When Dam-LMNB1 was co-expressed, ^{m6}A-Tracer-NES shifted to the nuclear lamina, with excess copies remaining in the cytoplasm in a subset of cells with especially high expression (Figure 2.7b and Figure 2.8). This shift in localization began within 2–3 h of Dam-LMNB1 induction and produced visible rings in the majority of transfected cells within 5 h (Figure 2.8). Because ^{m6}A-Tracer-NES only binds methylated sites in the nucleus, it solves two major problems: (1) ^{m6}A-Tracer fluorescence in the nucleus is no longer ambiguous and can be interpreted as a signal of methylation, and (2) high contrast between the nuclear lamina and the nuclear interior can be achieved for a much wider range of ^{m6}A-Tracer expression levels. ^{m6}A-Tracer-NES will allow for more sensitive imaging of other classes of protein-DNA interactions in the nucleus, and it could potentially also be utilized in synthetic genetic and epigenetic circuits[41] to reduce off-target effects, or to serve as a nuclear localization switch.

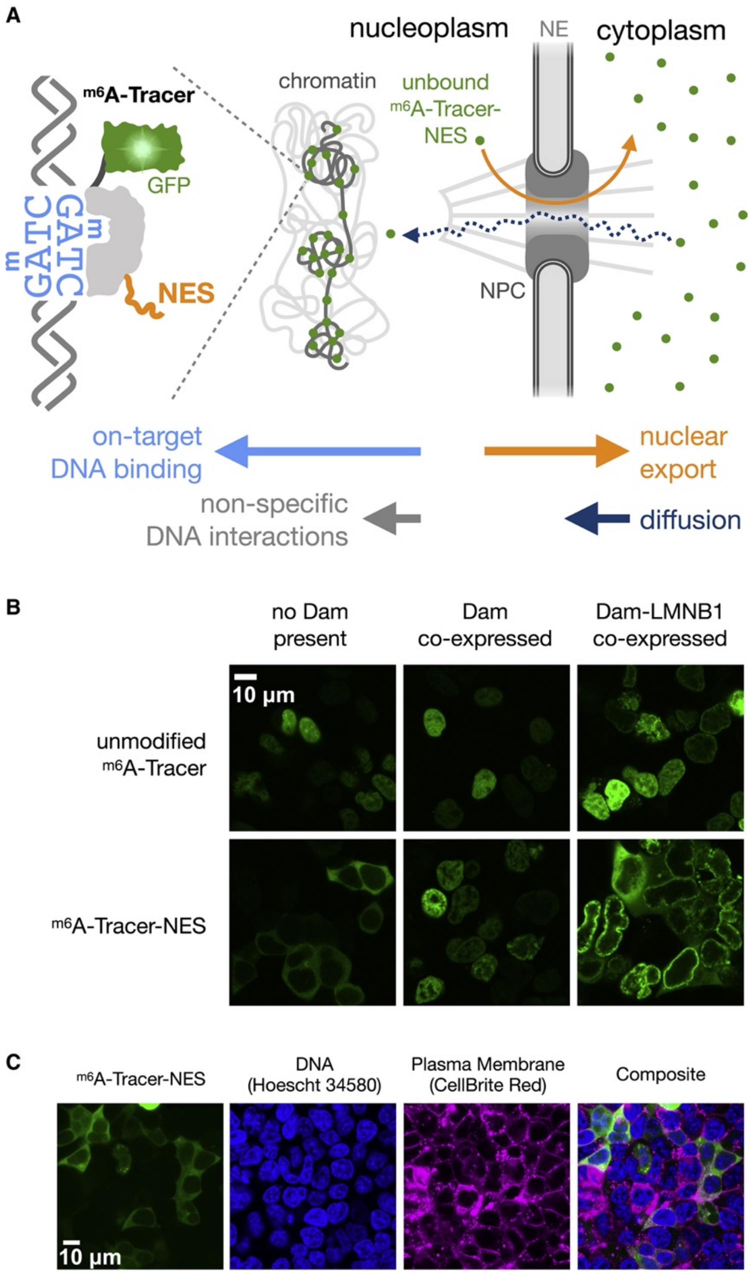


Figure 2.7 Improved Imaging of Protein-DNA Interactions with ^{m6}A-Tracer-NES. **a,** Illustration of potential mechanism by which ^{m6}A-Tracer-NES (^{m6}A-Tracer with a C-terminal HIV-1 Rev Nuclear Export Signal) reduces background fluorescence in the nucleus caused by non-specific DNA interactions, due to the relative rates of export, diffusion, and DNA binding (indicated by horizontal arrows). NPC, nuclear pore complex; NE, nuclear envelope. **b,** Confocal images of ^{m6}A-Tracer-NES expressing cells co-stained with Hoescht 34580 to label DNA and CellBrite Red to

label plasma membranes, showing cytoplasmic localization when Dam is not co-expressed. **c,** Confocal fluorescent microscope images revealing the different localization patterns of ^{m6}A-Tracer[36] with or without a NES and with or without Dam or Dam-LMNB1 co-expression.

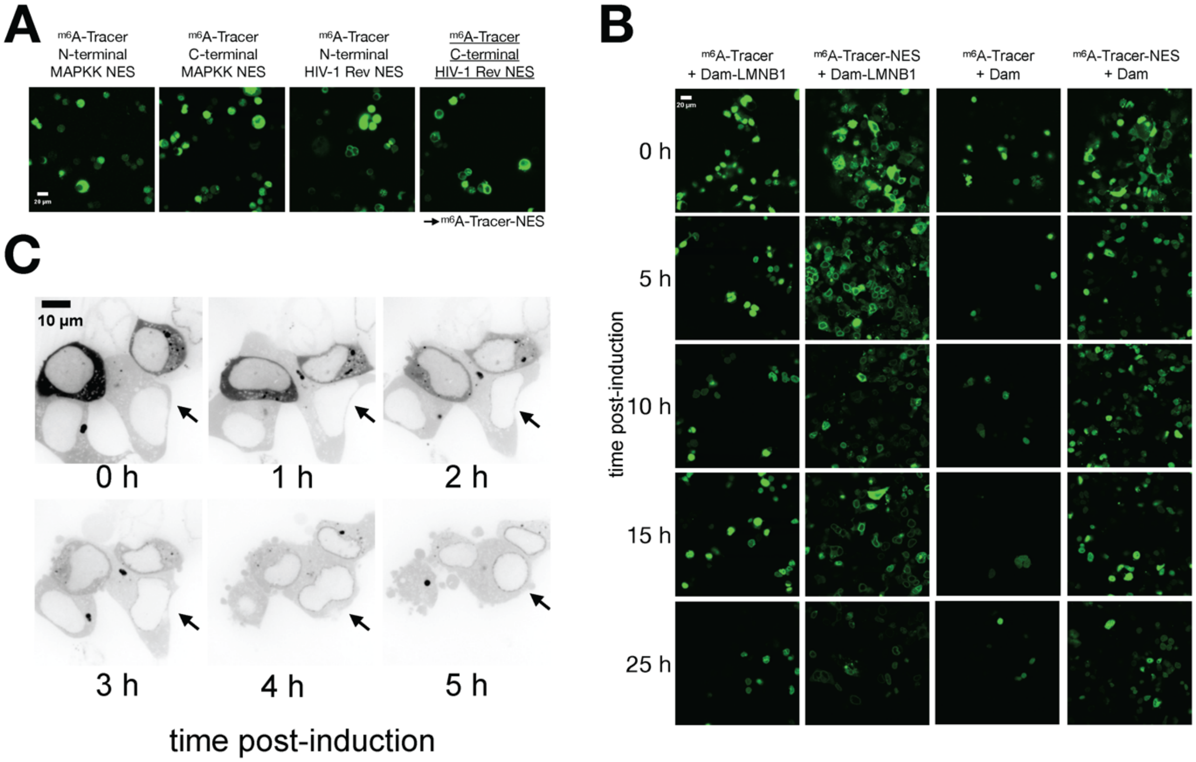


Figure 2.8 Additional characterization of ^{m6}A-Tracer-NES constructs. **a,** Confocal microscope images showing the localization of ^{m6}A-Tracer fluorescence when fused to one of two different Nuclear Export Signals on either terminus, in cells not expressing Dam. The HIV-1 Rev NES worked on either terminus and the C-terminal fusion was selected for downstream experiments. **b,** Time-lapse confocal images of ^{m6}A-Tracer-NES or unmodified ^{m6}A-Tracer fluorescence in different fields of cells, in cells co-expressing either Dam or Dam-LMNB1. Some nuclear localization is visible at time 0 in ^{m6}A-Tracer-NES + Dam cells, likely owing to leaky expression prior to induction. **c,** Time-lapse confocal microscope images of ^{m6}A-Tracer-NES fluorescence in the same field of cells at timepoints after Dam-LMNB1 expression. An inverted lookup table is used, and an arrow points to the nucleus of the same cell, which begins to show laminar signal around 2h post-induction.

2.6 In situ DamID-seq

There are some important limitations of using DamID-seq as a general method for mapping protein-DNA interactions in single cells: (1) Genetic manipulations are required to express a protein-Dam fusion specific to the protein of interest using a transfection method that works in the cell type of interest; (2) There is background from Dam methylating any accessible DNA it contacts; (3) DamID involves expressing an exogenous protein fusion and may not reflect the true biology of the endogenous protein; (4) Imaging protein-DNA interactions requires yet another genetic manipulation with transfecting ^{m6}A-Tracer. To address these limitations, I explored an *in situ* version of the DamID protocol (Methods: *In situ* DamID-seq). Using this *in situ* version of DamID makes genetic manipulation unnecessary, has the potential to reduce background because wash steps are performed to remove unbound antibody and methyltransferase before supplying the methyl donor, measures the endogenous protein's binding patterns and does not affect cell behavior and gene expression, can be used to target post-translational modifications, and allows for imaging with fluorescent secondary antibodies rather than requiring ^{m6}A-Tracer.

For this *in situ* protocol, instead of expressing a Dam fusion protein *in vivo*, I instead bound an antibody to the protein of interest, LMNB1, within permeabilized nuclei (Methods: *In situ* DamID-seq). I next recruited a methyltransferase that methylates adenines, Hia5, to the antibody using a protein A – Hia5 (pA-Hia5) fusion. Hia5 was chosen because we had demonstrated that it methylates efficiently *in situ* (Chapter 3), although Dam may be another good option because DpnI digestion in the DamID-seq protocol already limits analysis to GATC sites. We also tested EcoGII, which did not methylate as efficiently as Hia5 *in situ* (Chapter 3). Finally, I supplied the methyl donor S-adenosylmethionine (SAM) to activate methylation and then extracted the DNA and followed the conventional bulk DamID-seq protocol without the pre-PCR DpnII digestion. I excluded the DpnII digestion because it obliterated signal following *in situ* methylation. Following PCR, there was very little DNA recovered when the pre-PCR DpnII step was included. I hypothesize this is because adenines blocked by nucleosomes or other proteins are excluded from methylation more strongly *in situ*, leaving unmethylated gaps between neighboring GATC sites; amplification requires digestion and ligation to methylated GATC sites on both ends of a molecule. While with *in vivo* methylation DpnII digestion increases the signal specificity by making fragments that have an unmethylated GATC site unable to be amplified, this step is too selective following *in situ* methylation.

To evaluate the efficiency and specificity of DamID-seq following *in vivo* vs. *in situ* targeted methylation, I performed bulk DamID-seq on DNA extracted from HEK293T cells expressing Dam-LMNB1 and from HEK293T nuclei that were *in situ* methylated by incubating with an anti-LMNB1 antibody and pA-Hia5. I chose to target LMNB1 in HEK293T cells because we have previously identified genomic regions that we know should and should not be in contact with the nuclear lamina, which allowed me to compare performance between *in vivo* and *in situ* methylation. Positive and negative control sets of cLAD and ciLAD bins were defined as in the previous analyses for this chapter. DamID-seq experiments include the important control untethered Dam to measure the propensity of different genomic regions to be methylated. For the *in vivo* methylation experiments, the control was free Dam, not tethered to LMNB1, just as in DamID-seq experiments. For *in situ* methylation, this background control is instead free Hia5 added during activation, with the antibody and pA binding steps omitted. I followed the traditional DamID-seq protocol, as described in our recent manuscript;[35] however, as discussed above, I omitted the pre-PCR DpnII digestion.

To address efficiency, I calculated the read-depth-normalized coverage in cLADs, which is a measure of on-target signal. To evaluate the specificity, I computed the ratio of the coverage in cLADs (on-target) to the coverage in ciLADs (off-target). The *in situ* approach has both higher efficiency and specificity (Table 2.1). I also plotted the coverage in 100-kb bins in the free methyltransferase controls versus the targeted methylation samples (Figure 2.9). For both *in vivo* and *in situ* methylation, coverage in cLADs is higher for the targeted samples, while coverage in ciLADs is higher for the non-targeted controls. Some of the ciLAD bins have higher coverage in the *in situ* targeted sample relative to the *in vivo* targeted sample; however, overall, the coverage in cLADs relative to ciLADs is higher with *in situ* targeted methylation. These analyses suggest that when performing DamID-seq for profiling protein-DNA interactions, antibody-directed *in situ* methylation can be used instead of performing genetic manipulation to methylate *in vivo*. There are still scenarios where *in vivo* methylation may be preferred, such as if an integrated signal over time of all DNA a protein contacts is desired or if a protein does not have a high-quality antibody for targeting.

condition	**Read-depth-normalized coverage in cLADs**	**Coverage in cLADs / coverage in ciLADs**
in vivo Dam-LMNB1	1.440 x 10^{-4}	10.33
in situ pA-Hia5 targeting LMNB1	1.468 x 10^{-4}	16.23

Table 2.1 Read-depth-normalized coverage in cLADs and coverage ratio in cLADs and ciLADs for DamID-seq data following *in vivo* and *in situ* targeted methylation.

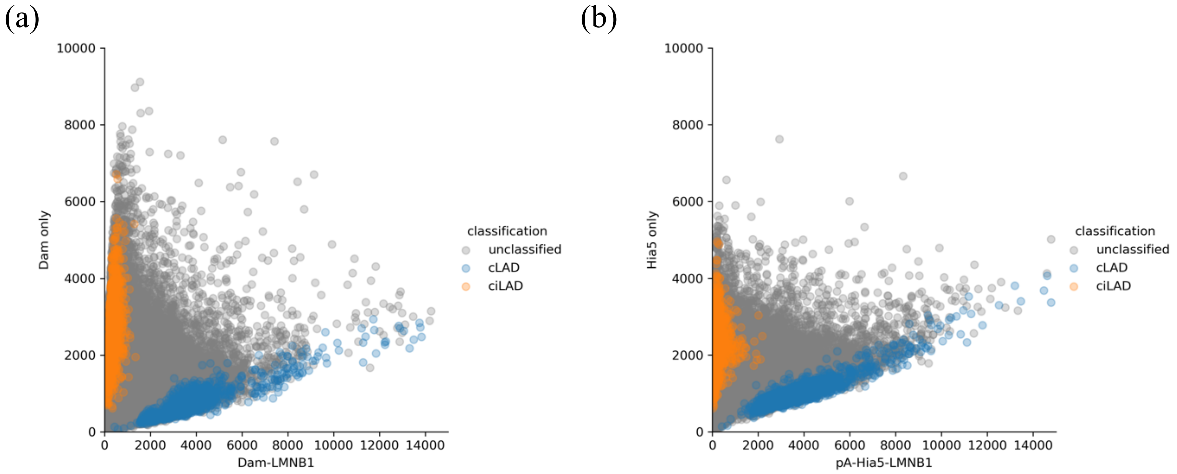


Figure 2.9 Coverage in 100-kb bins, colored by LAD classification. **a,** Coverage for Dam only control vs. coverage for Dam-LMNB1 sample. **b,** Coverage for Hia5 only control vs. coverage for pA-Hia5 targeting LMNB1.

2.7 *Chromatin accessibility or protein binding & CpG methylation measurements in parallel*

I developed a further modified DamID-seq workflow that enriches for regions of open chromatin or regions where protein-DNA interactions occur together with CpG-methylated regions by digesting at both such sites. Following either *in vivo* methylation or *in situ* methylation, the assay outlined in Figure 2.10 can be performed to simultaneously measure chromatin accessibility or protein binding together with CpG methylation (Methods: DpnI & MspJI joint assay). To measure chromatin accessibility, I transfected and induced expression of a Dam only plasmid *in vivo* or I treated nuclei *in situ* with the Dam enzyme. To measure protein binding, I transfected and induced expression of a Dam-LMNB1 plasmid or performed *in situ* methylation with pA-Hia5 targeting an antibody specific to LMNB1. Following deposition of exogenous mA through one of these methods, the DNA is then marked both with mA and CpG methylation. Just as in DamID-seq, a methylation-sensitive restriction enzyme digestion is performed; however, in addition to DpnI which cuts at GmATC, the restriction enzyme MspJI is added to cut around mCpG sites.[18] Adapters with sample barcodes and barcodes to mark whether a fragment is derived from a DpnI cut or an MspJI cut are then ligated to the digested DNA. The DNA is then pooled, cleaned, and prepared for sequencing.

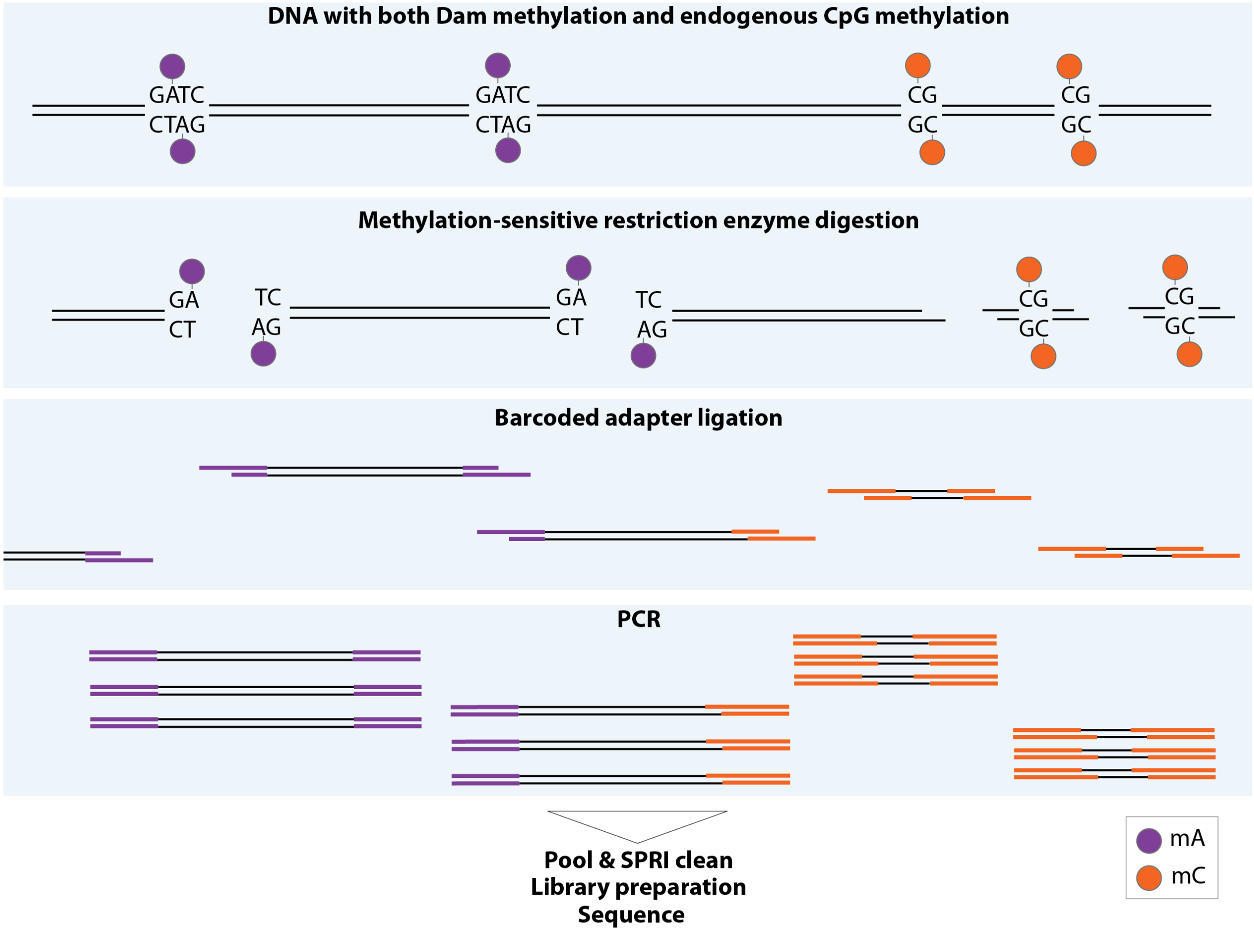

DNA with both Dam methylation and endogenous CpG methylation
GATC
CTAG
GATC
CTAG
CG
GC
CG
GC
Methylation-sensitive restriction enzyme digestion
GA
CT
TC
AG
GA
CT
TC
AG
CG
GC
CG
GC
Barcoded adapter ligation
PCR
Pool & SPRI clean
Library preparation
Sequence
mA
mC

Figure 2.10 Joint accessibility or protein-DNA interactions and CpG methylation profiling assay. DNA is methylated with an adenine methyltransferase like Dam or Hia5 either to mark open

chromatin or to mark protein binding events. The endogenous CpG methylation is maintained on the DNA. Methylation-sensitive restriction enzyme digestion is performed to cut at GmATC and mCG sites. Adapters are ligated that contain a sample identifying barcode and a barcode to mark whether the fragment originated from a DpnI or MspJI cut. Digested fragments are selectively amplified and prepared for sequencing.

As a first demonstration of using free Dam to measure chromatin accessibility, I performed standard bulk DamID-seq on DNA that was methylated with Dam *in vivo* or *in situ* (Methods: Cell Transfection and Harvesting, Methods: *In situ* DamID-seq). To benchmark performance, I looked at coverage enrichment at transcriptions start sites (TSSs) for highly expressed genes in HEK293T cells relative to TSSs for genes with low expression (Figure 2.11). TSSs for highly expressed genes are more open, and we see increased Dam signal for those TSSs. Importantly, this enrichment is not evident at TSSs for genes with low expression, which are inaccessible. I similarly compared to DNase-seq peaks (ENCODE ENCFF127KSH)[42] as an orthogonal method for measuring accessibility and observed increased coverage at high confidence DNase-seq peaks. Comparing the *in vivo* and *in situ* signal, it is evident that *in vivo* methylation has a broader signal. This is likely because *in vivo*, ATP is present, chromatin remodeling can occur, and methylation is occurring over a period of ~18 hours. In contrast, the *in situ* method takes a snapshot measurement in time from a nucleus devoid of ATP. These data demonstrate the ability to use free methyltransferase *in vivo* or *in situ* together with DamID-seq to measure chromatin accessibility.

To validate the joint use of DpnI and MspJI for measuring protein-DNA interactions together with CpG methylation, I targeted LMNB1 *in vivo* using a Dam-LMNB1 fusion. I then performed the modified DamID-seq workflow described in Figure 2.10 to cut both at Dam-methylated and endogenously methylated sites. I observed the expected increased coverage in cLADs relative to ciLADs when considering DpnI-cut fragments (Figure 2.12a). When analyzing MspJI-cut reads, we looked at TSSs of highly expressed genes, which should have a depletion of mCpG, relative to genes with low expression (Figure 2.12b). We observed the expected depletion of mCpG at highly expressed TSSs only. We also analyzed MspJI-cut read coverage in ciLAD and cLADs. While the cLAD signal is noisy because few TSSs reside in cLADs, we see the expected depletion in CpG methylation at TSSs in ciLADs, where genes are more highly expressed and TSSs are more accessible.

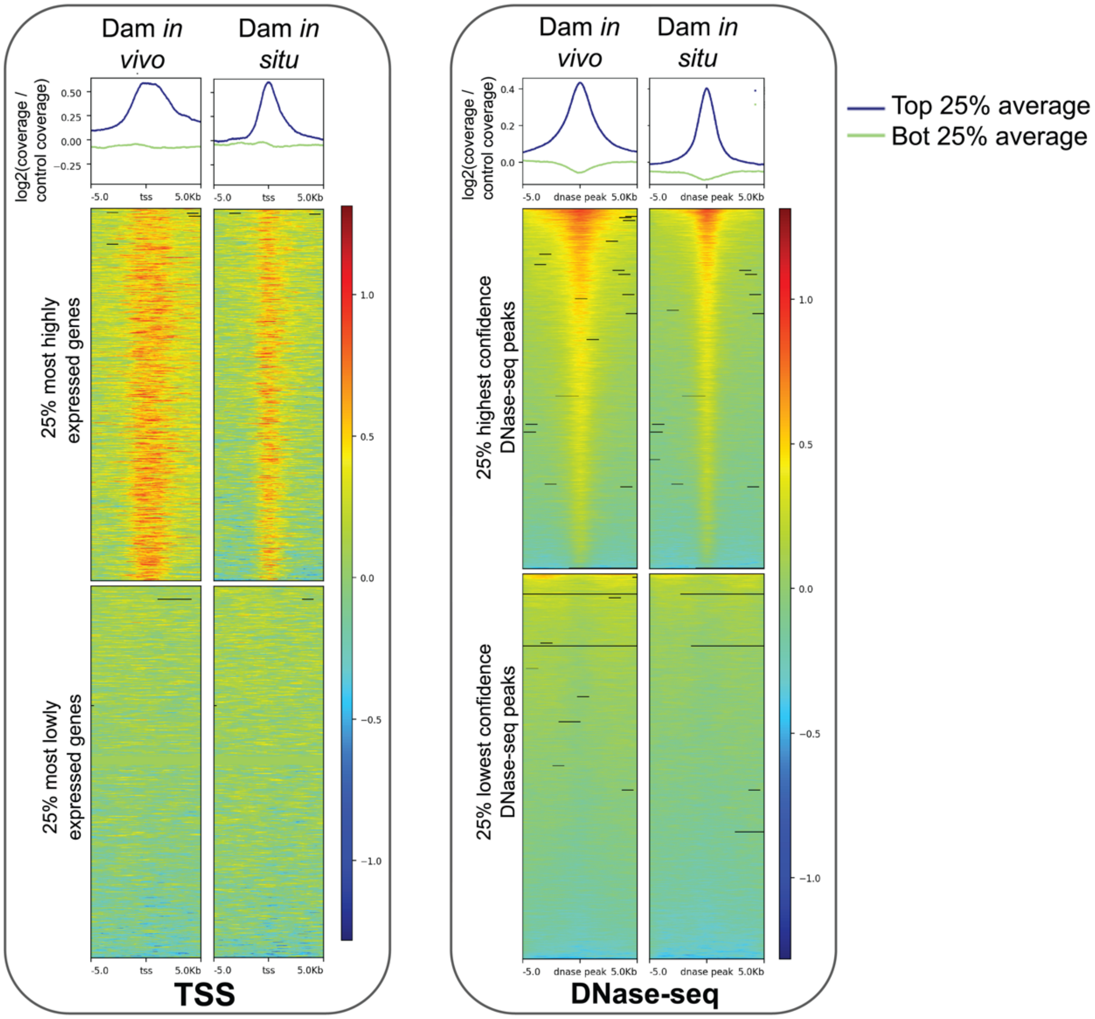


Figure 2.11 Open chromatin signal from *in vivo* and *in situ* methylation. DNA was methylated with Dam either *in vivo* or *in situ*. DamID-seq was then performed. Regions with enriched coverage overlap known regions of the genome that are accessible, such as TSSs of highly expressed genes and strong DNase-seq peaks. Importantly, coverage is not enriched at TSS for genes with low expression and at weak DNase-seq peaks.

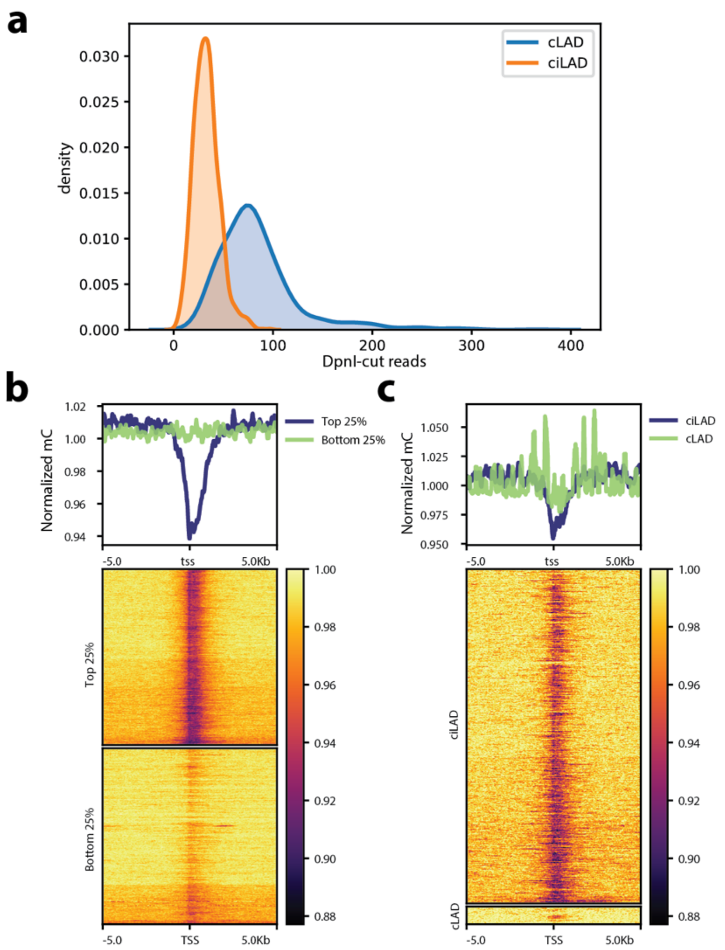


Figure 2.12 DpnI & MspJI assay for protein-DNA interactions & CpG methylation. **a,** Coverage by LAD classification. Higher coverage of reads derived from DpnI cuts is observed in constitutive lamina-associated domains (cLADs) compared to constitutive inter-lamina-associated domains (ciLADs). **b,** Depletion of CpG methylation at transcription start sites (TSSs) of highly expressed genes only is observed. Normalized signal is calculated by taking the ratio of the coverage and MspJI motif abundance. **c,** A larger dip in CpG methylation is evident in ciLADs, where more highly expressed genes reside.

This method does not require damaging and lossy bisulfite sequencing or physical separation of protein-DNA complexes, and therefore has the potential to enable multi-omic measurements in single cells. The work is ongoing; next steps involve improving sensitivity of this assay to produce sufficient signal in single cells.

2.8 Discussion

This chapter detailed optimizations and extensions of short-read methods for measuring chromatin structure in single cells. I characterized DamID-seq and improved the protocol and tools for imaging protein-DNA interactions with this method. I then modified DamID-seq to work *in situ*, making it possible to target proteins in genetically intractable systems and making the method more easily adaptable to new protein targets and systems. Next, I extended DamID-seq to include a measurement of CpG methylation towards the goal of creating multi-omic measurements of chromatin structure. This work set the stage for the development of DiMeLo-seq, an *in situ* method for measuring multiple features of chromatin structure on single molecules, in the next chapter.

2.9 Acknowledgements & author contributions

Acknowledgments

We would like to thank Anushka Gupta, Gabriel Dorlhiac, Zoë Steier, Adam Gayoso, Tyler Chen, Xinyi Zhang, and Sophie Dumont for their helpful feedback on this work. We are grateful to Carolyn de Graaf for providing us with LAD coordinates, to Bo Huang and Bianxiao Cui for providing guidance and materials, and to Bas van Steensel for providing us with plasmids. N.A. is supported by a Howard Hughes Medical Institute Gilliam Fellowship for Advanced Study. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (grant number R35GM124916). A.S. is a Chan Zuckerberg Biohub Investigator.

Author Contributions

N.A. and A.S. conceived of and designed the study and the microfluidic device; N.A. and A.L. fabricated and optimized operation of the device; A.M. performed bulk cell experiments and data processing; C.R.M. performed ^{m6}A-Tracer-NES experiments with supervision from A.M. and N.A.; N.A. performed all other experiments, analysis, and hardware construction; J.A.W. developed the microfluidic control platform and thermal cycling software, with minor modifications by N.A.; N.A. wrote the manuscript with contributions from A.M., C.R.M., and A.S.; A.S. supervised the study.

2.10 Methods

Data and Code Availability

The sequencing data generated during this study are available at GEO (accession GSE156150). The imaging data generated during this study are available at FigShare: https://doi.org/10.6084/m9.figshare.12798158. Analysis code, control software, device design files, and plasmid sequences are freely available for download on GitHub: https://github.com/altemose/microDamID. Source data for bulk KBM-7 RNA-seq were obtained from SRA (accession SRP044391), and source data for KBM-7 scDamID were obtained from GEO (accession GSE69423).

Cell Transfection and Harvesting

HEK293T cells were seeded in 24-well plates at 50000 cells per well in 0.5-ml media (see above for culturing and media details). The next day, cells were transfected using FuGene HD transfection reagent according to their standard protocol for HEK293 cells (Promega, Madison, WI). DNA plasmids were cloned in Dam-negative *E. coli* (New England Biolabs, Ipswitch, MA) to reduce sequencing reads originating from plasmid. Dam-LMNB1 and ^{m6}A-Tracer plasmids were obtained from Bas van Steensel (from Kind et al., 2013);[36] Dam-LMNB1 was modified to replace GFP with mCherry and to produce a Dam-only version, as well as to create a Dam-tdTomato-LMNB1 fusion for batch 2 experiments; their sequences are available in the accompanying GitHub repository. 250 ng Dam construct DNA plus 250 ng ^{m6}A-Tracer DNA were used per well. As controls to validate transfection, additional wells were left untransfected, transfected with ^{m6}A-Tracer only, or transfected with Dam construct only. The following day, successful transfection was validated by widefield fluorescence microscopy, seeing GFP signal in wells containing ^{m6}A-Tracer, and mCherry signal in all wells containing Dam construct only. Cells were harvested 72 hours after transfection. 20 hours before harvesting, the media was replaced and 0.5 µl Shield-1 ligand (0.5 mM stock, Takara Bio USA, Inc., Mountain View, CA) was added to each well to stabilize protein expression. Cells transfected with Dam-LMNB1 were inspected by fluorescence microscopy to look for the characteristic signal at the nuclear lamina, indicating proper expression and protein activity. To harvest the cells and prepare them for loading on the device, the cells were washed with PBS, then incubated at room temperature with 1X TrypLE Select (ThermoFisher Scientific, Waltham, MA) for 5 minutes to dissociate them from the plate. Cells were pipetted up and down to break up clumps, then centrifuged at 300xg for 5 minutes, resuspended in PBS, centrifuged again, and resuspended in 500 µl Pick Buffer (50 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM MgCl2, 137 mM NaCl), achieving a final cell concentration of roughly 500,000 cells per ml. Cells were passed through a 40 µm cell strainer before loading onto the device.

Confocal Imaging

Fluorescence confocal imaging of cells was performed in the trapping region using an inverted scanning confocal microscope with a 488 nm Ar/Kr laser (Leica, Germany) for excitation, with a bandpass filter capturing backscattered light from 500-540 nm at the primary photomultiplier tube (PMT), with the pinhole set to 1 Airy unit, with a transmission PMT capturing widefield unfiltered forward-scattered light, and with a 63X 0.7 NA long-working-distance air objective with a correction collar, zoomed by scanning 4X. For batch 2 imaging, a 63X 1.2 NA water immersion objective was used, with a 6X scanning zoom. The focal plane was positioned in the middle of each nucleus, capturing the largest-circumference cross-section, and final images were averaged over 10 frames to remove noise. For batch 2 cells, 10 confocal z slices were taken for each cell, and the slice with the largest nuclear perimeter was selected for image processing. The 3 cells expressing Dam-only that were sequenced in this study were imaged with a widefield CCD camera. Other Dam-only cells were imaged with confocal microscopy and showed similar relatively homogenous fluorescence throughout the nucleus, and never the distinct ‘ring’ shape found in Dam-LMNB1 expressing cells (Kind et al., 2013;[36] Figure 2.2b). No image enhancement methods were used prior to quantitative image processing. Images in Figure 2.2 have been linearly thresholded to diminish background signal.

Bulk DamID

Genomic DNA was isolated from ~3.7 x 10^6 transfected HEK293T cells using the DNeasy Blood & Tissue kit (Qiagen) following the protocol for cultured animal cells with the addition of RNase A. The extracted gDNA was then precipitated by adding 2 volumes of 100% ethanol and 0.1 volume of 3 M sodium acetate (pH 5.5) and storing at -20 °C for 30 minutes. Next, centrifugation for 30 minutes at 4 °C, >16,000 x g was performed to spin down the gDNA. The supernatant was removed, and the pellet was washed by adding 1 volume of 70% ethanol. Centrifugation for 5 minutes at 4 °C, >16,000 x g was performed, the supernatant was removed, and the gDNA pellets were air-dried. The gDNA was dissolved in 10 mM Tris-HCl pH 7.5, 0.1 mM EDTA to 1 µg/µl, incubating at 55 °C for 30 minutes to facilitate dissolving. The concentration was measured using Nanodrop.

The following DpnI digestion, adaptor ligation, and DpnII digestion steps were all performed in the same tube.[37] Overnight DpnI digestion at 37 °C was performed with 2.5 µg gDNA, 10 U DpnI (NEB), 1X CutSmart (NEB), and water to 10 µl total reaction volume. DpnI was then inactivated at 80 °C for 20 minutes. Adaptors were ligated by combining the 10 µl of DpnI-digested gDNA, 1X ligation buffer (NEB), 2 µM adaptor dsAdR, 5 U T4 ligase (NEB), and water for a total reaction volume of 20 µl. Ligation was performed for 2 hours at 16 °C and then T4 ligase was inactivated for 10 minutes at 65 °C. DpnII digestion was performed by combining the 20 µl of ligated DNA, 10 U DpnII (NEB), 1X DpnII buffer (NEB), and water for a total reaction volume of 50 µl. The DpnII digestion was 1 hour at 37 °C followed by 20 minutes at 65 °C to inactivate DpnII.

Next, 10 µl of the DpnII-digested gDNA was amplified using the Takara Advantage 2 PCR Kit with 1X SA PCR buffer, 1.25 µM Primer Adr-PCR, dNTP mix (0.2 mM each), 1X PCR advantage enzyme mix, and water for a total reaction volume of 50 µl. PCR was performed with an initial extension at 68 °C for 10 minutes; one cycle of 94 °C for 1 minute, 65 °C for 5 minutes, 68 °C for 15 minutes; 4 cycles of 94 °C for 1 minute, 65 °C for 1 minute, 68 °C for 10 minutes; 21 cycles of 94 °C for 1 minute, 65 °C for 1 minute, 68 °C for 2 minutes. Post-amplification DpnII digestion was performed by combining 40 µl of the PCR product with 20 U DpnII, 1X DpnII buffer, and water to a total volume of 100 µl. The DpnII digestion was performed for 2 hours at 37 °C followed by inactivation at 65 °C for 20 minutes. The digested product was purified using QIAquick PCR purification kit. The purified PCR product (1 µg brought up to 50 µl in TE) was sheared to a target size of 200 bp using the Bioruptor Pico with 13 cycles with 30”/30” on/off cycle time. DNA library preparation of the sheared DNA was performed using NEBNext Ultra II DNA Library Prep Kit for Illumina using AMPure XP beads (Beckman Coulter Life Sciences, Indianapolis, IN).

Bulk DamID, Comparing Dam Mutants

Bulk DamID for comparing the wild-type allele and V133A mutant allele was performed as outlined in the Bulk DamID section above with the following modifications. Genomic DNA was extracted from ~ 2.4 x 10^5transfected HEK293T cells. A cleanup before methylation-specific amplification was included to remove unligated Dam adapter before PCR. The Monarch PCR & DNA Cleanup Kit with 20 µl DpnII-digested gDNA input and an elution volume of 10 µl was used. Shearing with the Bioruptor Pico was performed for 20 total cycles with 30”/30” on/off cycle time. Paired-end 2 x 75 bp sequencing was performed on an Illumina NextSeq with a mid output kit. Approximately 3.8 million read pairs per sample were obtained.

Bulk RNA-seq

RNA was extracted from ~1.9 x 10^6 transfected HEK293T cells using the Rneasy Mini Kit from Qiagen with the QIAshredder for homogenization. RNA library preparation was performed using the NEBNext Ultra II RNA Library Prep Kit for Illumina with the NEBNext Poly(A) mRNA Magnetic Isolation Module. Paired-end 2 x 150 bp sequencing for both DamID-seq and RNA-seq libraries was performed on 1 lane of a NovaSeq S4 run. Approximately 252 million read pairs were obtained for each DamID-seq sample, and roughly 64 million read pairs for each RNA sample.

^{m6}A-Tracer-NES

To reduce background fluorescence due to ^{m6}A-Tracer, we fused its N or C terminus to one of two different nuclear export signals (NES): HIV-1 Rev (LQLPPLERLTLD) or MAPKK (LQKKLEELEL).[43] We compared the localization of each of the 4 resulting constructs by imaging HEK293T cells transiently transfected with ^{m6}A-tracer-NES by itself or with Dam. Negative controls included transfection with unmodified ^{m6}A-Tracer only or Dam-only, and no transfection. The MAPKK NES did not appreciably reduce nuclear localization of ^{m6}A-tracer-NES in the absence of Dam (Figure 2.8). However, the HIV-1 Rev NES, in either the N- or C-terminal configuration, showed significant improvement in localizing signal to the cytoplasm in the absence of Dam, while permitting nuclear localization in the presence of Dam (Figure 2.7b and Figure 2.8). We proceeded to use the C-terminal HIV-1 Rev ^{m6}A-Tracer construct for downstream experiments. Co-transfection with Dam-LMNB1 resulted in a greater proportion of transiently transfected cells having visible laminar rings than with unmodified ^{m6}A-Tracer. Timelapse imaging of the same field of Dam-LMNB1 + ^{m6}A-Tracer-NES cells over time or different fields at each timepoint (Figure 2.8) demonstrated that laminar rings become visible within 2-3 hours and reach full intensity around 5 hours after Dam-LMNB1 induction with Shield-1 ligand. To test the possibility that unmodified ^{m6}A-Tracer localizes to the nucleus due to a cryptic Nuclear Localization Signal, we searched for NLS motifs using NLSdb[44] but found no matches.

Bulk RNA-seq analysis

Adapters were trimmed using trimmomatic (v0.39; Bolger et al., 2014;[45] ILLUMINACLIP:adapters-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36, where adapters-PE.fa is:

```
>PrefixPE/1

TACACTCTTTCCCTACACGACGCTCTTCCGATCT

>PrefixPE/2

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT).
```

Transcript quantification was performed using Salmon[46] with the GRCh38 transcript reference. Differential expression analysis was performed using the voom function in limma.[47] Differential expression was called based on logFC significantly greater than 1 and adjusted p-value < 0.01.
For KBM7 bulk gene expression analysis, publicly available single-end RNA sequencing data (SRA accession SRP044391[48]) from two replicates were processed. For adapter trimming, trimmomatic was used in the SE mode with the adapter file ILLUMINACLIP:TruSeq3-SE. All other trimmomatic parameters were the same as were used in the HEK293T RNA-seq data processing, and Salmon was used for transcript quantification in single-end mode.

Bulk and Single-cell DamID analysis

Bulk and single-cell DamID reads were demultiplexed using Illumina's BaseSpace platform to obtain fastq files for each sample. DamID and Illumina adapter sequences were trimmed off using trimmomatic[45] (v0.39; ILLUMINACLIP:adapters-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:20, where adapters-PE.fa is:

```
>PrefixPE/1

TACACTCTTTCCCTACACGACGCTCTTCCGATCT

>PrefixPE/2

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

>Dam

GGTCGCGGCCGAGGA

>Dam_rc

TCCTCGGCCGCGACC
```

Trimmed reads were aligned to a custom reference (hg38 reference sequence plus the Dam-LMNB1 and ^{m6}A-Tracer plasmid sequences) using BWA-MEM[49] (v0.7.15-r1140). Alignments with mapping quality 0 were discarded using samtools[50] (v1.9). The hg38 reference sequence was split into simulated DpnI digestion fragments by reporting all intervals between GATC sites (excluding the GATC sites themselves), yielding 7180359 possible DpnI fragments across the 24 chromosome assemblies. The number of reads overlapping each fragment was counted using bedtools[51] (v2.28). For single-cell data, the number of DpnI fragments with non-zero coverage was reported within each non-overlapping bin in the genome (28163 total 100 kb bins, after excluding unmappable regions with zero coverage in any cell). For bulk data, the number of read pairs overlapping each 100 kb bin was reported. The same exact pipeline was applied to the raw reads from Kind et al., 2015 (GEO accession GSE69423).[30] RefSeq gene positions were downloaded from the UCSC Genome Browser and counted in each bin. For bulk data, Dam-LMNB1 vs DamOnly enrichment was computed using Deseq2 in each 100 kb bin.[52] For single-cell data, the expected background coverage in each bin was computed as n(m/t), where n is the number of unique fragments sequenced from that cell, m is the number of bulk Dam-only read pairs mapping to that bin, and t is the total number of mapped bulk Dam-only read pairs. Single-cell normalization was computed either as a ratio of observed to expected coverage (for browser visualization and comparison to bulk data), or as their difference (for classification and coverage distribution plotting). Positive and negative control sets of cLAD and ciLAD bins were defined under the assumption that genomic regions that have high bulk DamID signal and that are lamina associated across many cell types are likely to be in contact with the lamina in the vast majority of single cells, which is supported by previous scDamID data.[30] Specifically, we defined them as bins with a bulk Dam-LMNB1:Dam-only DESeq2 p-value smaller than 0.05/28760, that intersected published cLADs and ciLADs in other cell lines,[31] and that were among the top 1200 most differentially enriched bins in either direction (positive or negative log fold change for cLADs and ciLADs, respectively). Normalized coverage thresholds for LAD/iLAD (i.e., contact vs. no contact) classification were computed for each cell to maximize accuracy on the cLAD and ciLAD control sets. To examine whether using the full control sets to set thresholds and define classification error was resulting in substantial overfitting, we split the control sets into training and test sets for threshold setting and accuracy determination, respectively, and only observed a 0.7% mean drop in accuracy relative to using the full sets. Signal-to-noise ratios were computed for each cell using the normalized coverage distributions in the cLAD and ciLAD control sets as $(\mu_{cLAD} - \mu_{ciLAD})/\sigma_{ciLAD}$. For most downstream analyses, we chose to exclude 20 cells with fewer than 100,000 unique covered fragments, which includes cells with poor laminar rings and lower DNA yields (Figure 2.1d). For any given application of µDamID, this threshold will depend on the level of noise due to background methylation in the biological system being used, which is expected to depend in part on the expression level of the Dam fusion protein. In a transiently transfected cell population, this expression level is expected to vary widely, which motivated the use of data to explore this as a cause of variable classification accuracy between cells. The remaining 31 Dam-LMNB1 cells had a median classification accuracy of 90% (range 74%-98%).

Calling vLADs

Variable LADs were defined as bins called as LADs in 33-66% of cells and conservatively filtered to remove regions resulting from sampling error. This was done by computing, for each bin and for each cell, the probability that the true sample contact frequency lies outside the interval (33%, 66%). We estimated this probability using a Poisson-binomial distribution, a generalization of the binomial distribution allowing individual samples to have varying success probabilities. Specifically, each bin in the genome has k cells called as LADs and n-k cells called as iLADs, with n=31 in this study. For the k LADs we generated a vector of k false-positive probabilities, with each probability estimated as the fraction of negative-control ciLADs with coverage greater than the observed coverage in that bin. We used this probability vector to parameterize a Poisson-binomial distribution with k draws, providing the distribution of false-positive calls in the bin. We repeated this for the n-k iLAD bins, with each false-negative probability estimated as the fraction of positive-control cLADs with coverage lower than the observed coverage in that bin. These two distributions were combined into a single density by reflecting the false-positive distribution about the y axis, scaling each one according to its mean, and adding k. Only regions with $p<10^{-3}$ for both tails were called as variable LADs. We then generated 10,000 samples of the sample contact frequency, c, from this distribution and used each one to generate a single binomial (n=31, p=c/31) sample, generating a combined measurement and sampling distribution with greater variance than either alone (Figure 2.5b), from which we generated 95% confidence intervals for the population contact frequency in each bin (Figure 2.5a). Statistical analyses and plots were made in R (v4.0.0) using the ggplot2 (v3.3.0), gplots (v3.0.3), colorRamps (v2.3), reshape2 (v1.4.4), ggextra (v0.9) and poisbinom (v1.0.1) packages. Browser figures were generated using the WashU Epigenome Browser.[53]

In situ DamID-seq

For the *in situ* DamID-seq protocol, *in situ* methylation targeting LMNB1 in HEK293T cells was performed as is described in the DiMeLo-seq protocol in Chapter 3 Methods: DiMeLo-seq. Once DNA was extracted following that protocol, the standard bulk DamID-seq protocol described in Methods: Bulk DamID was performed. One modification was made to exclude the pre-PCR DpnII digestion. Libraries were prepared for sequencing using the NEB Ultra II FS DNA Library Prep Kit for Illumina (E7805S) and were sequenced with NextSeq High Output.

DpnI & MspJI joint assay

For the DpnI and MspJI joint assay, DNA from HEK293T cells transfected with a Dam-LMNB1 plasmid was digested with DpnI and MspJI. A 40 µl reaction with 40 ng DNA, 4 µl CutSmart buffer, 1 µl DpnI (R0176S), 1 µl MspJI (R0661S), and 1.33 µl MspJI activator was set up. Digestion was performed for 4 hours at 37 °C, followed by a 20 min 65 °C deactivation step for MspJI. For ligation, 10 µl of digest was combined with 2 µl total adapters (50 µM, see below), 2 µl 10X ligation buffer, and 2 µl Roche ligase (5U/µl, 10799009001) in a total reaction volume of 20 µl.

DpnI fragment adapter:
5' AGTGGCTATCCTGTCTGACTG
5' CAGTCAGA/3ddC/

MspJI fragment adapter:
5' TACCGCTATCCTGCTGTCAGT
5' N*N*N*N*ACTGACAG/3ddC/

Ligation was performed at 16 °C for 16 hours, followed by 10 minutes at 65 °C to deactivate the ligase. PCR was performed by adding 5 µl water and 25 µl MyTaq Master Mix (BIO-25041) directly to the ligation reaction. An initial extension step at 72 °C for 8 min, followed by 21 cycles of 94 °C for 20 sec, 58 °C for 30 sec, and 72 °C for 20 sec was performed. Then a final extension at 72 °C for 2 minutes and finally a hold at 4 °C was performed. The amplified material was cleaned with a 2.5X SPRI cleanup and input into NEBNext Ultra II DNA Library Prep Kit for Illumina (E7645S) and sequenced with NextSeq High Output.

Chapter 3

DiMeLo-seq: Directed Methylation with Long-Read Sequencing

3.1 Overview

Our initial *in situ* DamID-seq experiments demonstrated that *in situ* methylation could provide comparable protein-DNA interaction profiles without the need for genetic manipulation. This modification of the "recording" process (methylation deposition) in DamID makes the protocol more adaptable for new targets and cell types. We next considered how the "reading" aspect (i.e., the methylation detection) could be improved to create richer multi-omic datasets. Rather than digesting DNA with a methylation-sensitive restriction enzyme and selectively amplifying these digested fragments, we aimed to read out methylation directly with long-read, single-molecule sequencing. In this approach, protein-DNA interactions are also encoded through targeted adenine methylation, but we do not subsequently fragment or amplify the DNA and instead measure methylation on the native DNA molecules themselves. This work was published in Altemose, Maslan, Smith, Sundararajan, et al., 2022.[54]

3.2 Abstract

Studies of genome regulation routinely use high-throughput DNA sequencing approaches to determine where specific proteins interact with DNA. These methods rely on DNA amplification and short-read sequencing, limiting their quantitative application in complex genomic regions. To address these limitations, we developed **Di**rected **Me**thylation with **Lo**ng-read **seq**uencing (DiMeLo-seq), which uses antibody-tethered enzymes to methylate DNA near a target protein's binding sites *in situ*. These exogenous methylation marks are then detected simultaneously with endogenous CpG methylation on unamplified DNA using long-read, single-molecule sequencing technologies. We optimized and benchmarked DiMeLo-seq by mapping chromatin-binding proteins and histone modifications across the human genome. Furthermore, we identified where centromere protein A (CENP-A) localizes within highly repetitive regions that are unmappable with short sequencing reads, while also estimating the density of CENP-A molecules along single chromatin fibers. DiMeLo-seq is a versatile method that provides multimodal, genome-wide information for investigating protein-DNA interactions.

3.3 Introduction

Genomic DNA must be decoded and maintained by proteins that read, regulate, replicate, recombine, and repair it. Mapping where and how these proteins interact with DNA can provide key insights into how they function or malfunction in healthy and diseased cells. Several powerful approaches have been developed to map where individual target proteins interact with DNA genome-wide, including DamID, ChIP-seq, CUT&RUN, and related methods.[2–4,7,55,56] These approaches involve selectively amplifying short DNA fragments from regions bound by a particular protein of interest, determining the sequence of those DNA molecules using high-throughput DNA sequencing, and mapping those sequences back to a reference genome, using sequencing coverage as a measure of protein-DNA interaction frequency. While these methods have proven to be extremely useful for studying DNA-binding proteins and chromatin modifications,[57] they rely on strategies for enrichment and detection that present important limitations.

Firstly, the process of DNA amplification fails to copy DNA modification information, like methylation and oxidation, from the native DNA molecules to the amplified and sequenced library DNA. This prevents simultaneous measurement of protein-DNA interactions and DNA modifications and limits the amount of information that can be gleaned about the relationship between these regulatory elements. Secondly, amplification-based enrichment methods often have intrinsic biases related to base composition, and they typically rely on nonlinear methods like PCR to yield enough amplified product for sequencing. Therefore, the sequencing coverage produced by these techniques provides only a semi-quantitative readout of protein-DNA interaction frequencies.

Furthermore, these approaches rely on digesting or shearing DNA into short fragments for enrichment, followed by high-throughput DNA sequencing for detection, which produces short sequencing reads typically under 250 bp in length. Short fragment lengths are often essential for achieving adequate binding site resolution with these techniques, and although it is possible to map multiple protein-DNA interactions on short reads,[58] shearing the DNA into short fragments destroys joint long-range binding information and hinders the ability to phase reads to measure haplotype-specific protein-DNA interactions. Additionally, repetitive regions of the human genome have presented a major challenge for genome assembly and mapping methods due to the difficulty of unambiguously assigning short DNA sequencing reads to their unique positions in the genome. Recent efforts using long-read sequencing technologies have provided a complete assembly of repetitive regions across the human genome.[59] However, unambiguously mapping short-read sequencing data remains impossible in many repetitive genomic regions, limiting our ability to address lingering biological questions about the roles of repetitive sequences in cell division, protein synthesis, aging, and genome regulation.

These limitations underline the need for protein-DNA interaction mapping methods that fully leverage the power of long-read, single-molecule sequencing technologies, including their ability to read out DNA modifications directly. To address this need, we developed Directed Methylation with Long-read sequencing (DiMeLo-seq; from *dímelo*, pronounced DEE-meh-low). DiMeLo-seq provides the ability to map protein-DNA interactions with high resolution on native, long, single, sequenced DNA molecules, while simultaneously measuring endogenous DNA modifications and sequence variation. Each of these features provides an opportunity to study genome regulation in unprecedented ways. Recent technologies have begun to take advantage of long-read sequencing to identify accessible regions and CpG methylation on native single molecules, but they cannot directly target specific protein-DNA interactions.[24–28] Here we extend these capabilities to map specific regulatory elements and demonstrate the advantages of DiMeLo-seq by mapping lamina-associated domains, CTCF binding sites, histone modifications/variants, and CpG methylation across the genome and through complex repetitive domains.

3.4 Results

1. DiMeLo-seq workflow

DiMeLo-seq combines elements of antibody-directed protein-DNA mapping approaches[56,60,61] to deposit methylation marks near a specific target protein, then uses long-read sequencing to read out these exogenous methylation marks directly.[24–28] Taking advantage of the lack of N^6-methyl-deoxyadenosine (hereafter mA) in human DNA,[8] we fused the antibody-binding Protein A to the nonspecific deoxyadenosine methyltransferase Hia5[24,62] (pA-Hia5) to catalyze the formation of mA in the DNA proximal to targeted chromatin-associated proteins (Figure 3.1a). First, nuclei are permeabilized, primary antibodies are bound to the protein of interest, and any unbound antibody is washed away. Next, pA-Hia5 is bound to the antibody, and any unbound pA-Hia5 is washed away. The nuclei are then incubated in a buffer containing the methyl donor S-adenosyl methionine (SAM) to activate adenine methylation in the vicinity of the protein of interest.[61] Finally, genomic DNA is isolated and sequenced using modification-sensitive, long-read sequencing with mA basecalls providing a readout of the sites of protein-DNA interactions (Figure 3.1a; full workflow in Figure 3.2). This approach provides a distinct advantage in the ability to detect multiple binding events by the target protein on each long, single DNA molecule, which would not be possible with short-read sequencing (Figure 3.1b). This protocol also avoids amplification biases, enabling improved estimation of absolute protein-DNA interaction frequencies at each site in the genome across a population of cells (Figure 3.1c). Modification-sensitive readout allows for the simultaneous detection of both exogenous antibody-directed adenine methylation and endogenous CpG methylation on single molecules (Figure 3.1d). Additionally, DiMeLo-seq's long sequencing reads often overlap multiple heterozygous sites, enabling phasing and measurement of haplotype-specific protein-DNA interactions (Figure 3.1e). Finally, long reads enable mapping of protein-DNA interactions within highly repetitive regions of the genome (Figure 3.1f). Overall, these improvements allow investigation of protein-DNA interactions on single-molecules, including in challenging genomic regions, with resolution and specificity that was not previously possible.

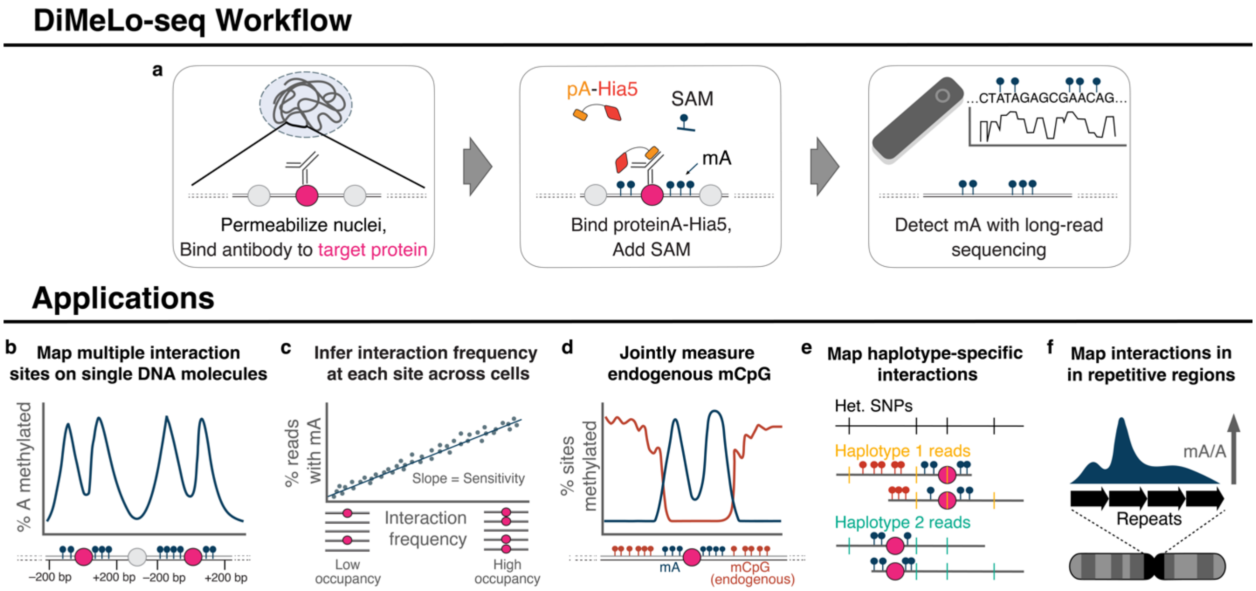


Figure 3.1 High-resolution, genome-wide mapping of protein-DNA interactions with DiMeLo-seq. **a**, Schematic of the DiMeLo-seq workflow for the mapping of protein-DNA interactions. **b**, DiMeLo-seq can be used to map multiple interaction sites for the target protein on each single DNA molecule, enabling estimation of the density and spacing of interaction sites on long single chromatin fibers. **c**, Because DiMeLo-seq avoids biases associated with DNA amplification, the fraction of reads with methyladenines can be used to better estimate the absolute protein-DNA interaction frequency at each site in the genome, with a scaling factor related to the single-molecule sensitivity. **d**, Endogenous CpG methylation information is preserved, enabling studies of the joint relationship between DNA methylation and protein binding. **e**, Since long sequencing reads tend to overlap multiple heterozygous sites, they can be reliably phased, enabling studies of how protein binding is affected by genetic or epigenetic differences between parental haplotypes. **f**, Long reads enable mapping of protein-DNA interactions across the entire genome, including in repetitive regions that remain challenging to uniquely align short reads to. The applications and advantages illustrated in b-f can also be used in combination with one another.

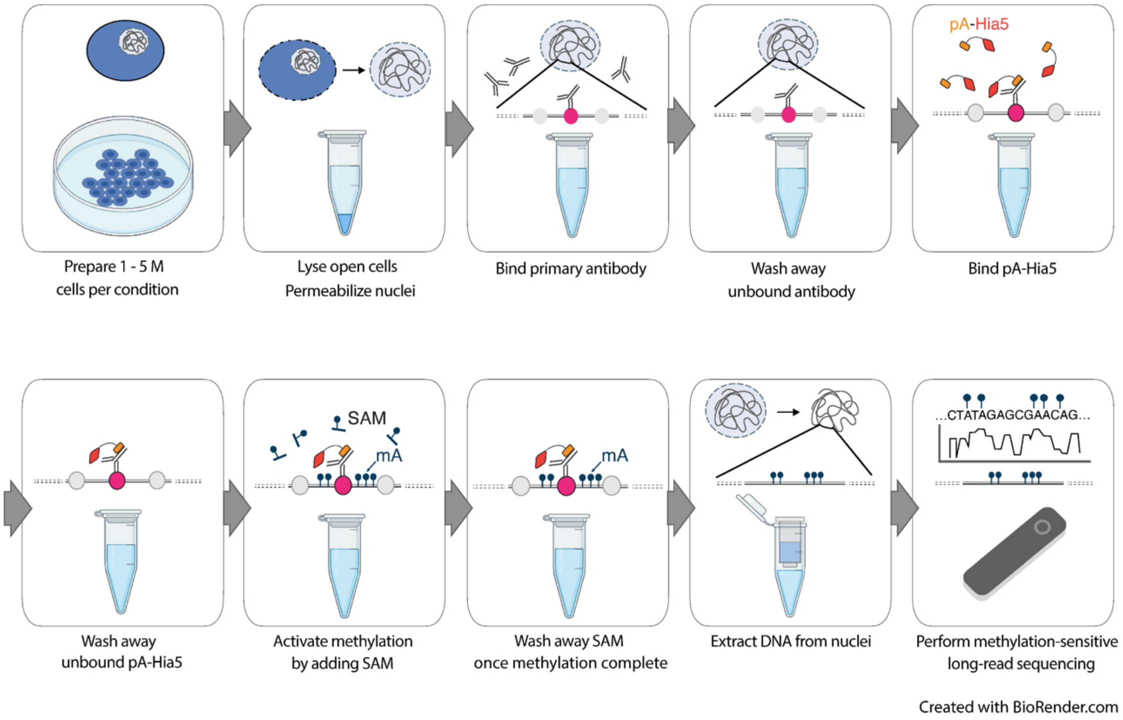


Figure 3.2 Workflow of DiMeLo-seq *in situ* methylation, DNA extraction, and sequencing. Schematic of the DiMeLo-seq *in situ* methylation protocol, which involves a series of binding steps and washes followed by DNA extraction and sequencing.

2. Antibody-directed histone-specific DNA adenine methylation of reconstituted chromatin **in vitro**

We expressed and purified recombinant pA-Hia5 and tested its methylation activity on purified DNA using the methylation-sensitive restriction enzyme DpnI, which only cuts GATC sites when adenine is methylated. DNA incubated with Hia5, pA-Hia5, or Protein A/G Hia5 (pAG-Hia5) in the presence of SAM became sensitive to DpnI digestion, confirming the methyltransferase activity of the purified fusion proteins (Supplementary Note 1, Figure 3.3a,b). To test the ability of pA-Hia5 to target chromatin and methylate accessible DNA *in vitro*, we reconstituted chromatin containing the histone variant CENP-A using the nucleosome-positioning DNA sequence referred to as “601”[63] (Figure 3.3c,d, Supplementary Note 2). Incubating mononucleosomes together with free-floating pA-Hia5 and SAM, followed by long-read sequencing and methylation-sensitive basecalling, showed methylation on 97.1 ± 0.8% of reads (mean ± s.e.m., n=3) (Supplementary Notes 3,4, Figure 3.4c,d, Extended Data Fig. 1e-k). Moreover, we observed almost no methylation at the expected nucleosome-protected region (Figure 3.4c,d, Figure 3.3j).

We reconstituted CENP-A chromatin on biotinylated DNA, bound it to streptavidin-coated magnetic beads, incubated it with CENP-A antibody and pA-Hia5, and washed away any unbound antibody and pA-Hia5 prior to activating methylation with SAM (Figure 3.4a, Figure 3.3c). We observed methylation on 65.0 ± 10.0% of CENP-A DiMeLo-seq reads (mean ± s.e.m., n=3)

(Figure 3.4b-d, Figure 3.3e-h,k), with methylation levels decaying with distance from the nucleosome footprint (Figure 3.4c). We observed only background levels of methylation on IgG control DiMeLo-seq reads (5.1 ± 0.6% of IgG reads, (mean ± s.e.m., n=2), compared to 4.1 ± 0.5 % of untreated reads, (mean ± s.e.m., n=3)) (Figure 3.4d, Figure 3.3e,k). While reads from either free-floating pA-Hia5 or antibody-tethered pA-Hia5 conditions showed nucleosome-sized protection from methylation (~150 - 180 bp centered at the dyad, Figure 3.4c,d, Figure 3.3j), ~70% of all methylation on reads from antibody-tethered pA-Hia5 fell within 250 bp on either side of the dyad. This result demonstrates that antibody tethered pA-Hia5 can methylate accessible DNA close to target nucleosomes *in vitro*.

To test the specificity of DiMeLo-seq to identify target nucleosomes on chromatin fibers, we assessed the ability of pA-Hia5 to methylate accessible regions of DNA on *in vitro* reconstituted chromatin assembled on an 18x array of the 601 nucleosome positioning sequence (Figure 3.5a-c). Co-incubation of chromatin together with free-floating pA-Hia5 and SAM resulted in structured patterns of oligonucleosome footprinting (Figure 3.5b,g,h), as reported previously for reconstituted chromatin incubated with another exogenous methyltransferase, EcoGII.[25]

We tested antibody-directed methylation of chromatin arrays reconstituted with either CENP-A or histone H3 containing nucleosomes. We incubated chromatin with CENP-A antibody and pA-Hia5, washed away unbound antibody, and activated methylation with SAM (Figure 3.4e). Following activation, we immunostained chromatin-conjugated beads with an anti-mA antibody, demonstrating a significant increase in mA signal when CENP-A chromatin, but not H3 chromatin, was incubated with pA-Hia5 and CENP-A antibody (Figure 3.5d,e, Supplementary Note 5), indicating antibody-directed methylation. Long-read sequencing detected mA on DNA after CENP-A-directed methylation of CENP-A chromatin (but not H3 chromatin) (Figure 3.5f). On average, CENP-A-directed methylation of CENP-A chromatin was depleted at the central axis of the nucleosome where the 601 sequence positions the nucleosome dyad (Figure 3.4f,g). On individual reads, we observed protection from methylation centered at 601 dyad positions, consistent with nucleosome occupancy protecting the DNA from antibody-directed methylation (Figure 3.4f,g) and similar to the free pA-Hia5 condition (Figure 3.5g,h). In contrast to the free pA-Hia5 condition, for which we observed a high prevalence of methylation on any region not protected by nucleosomes, in the antibody-directed pA-Hia5 condition, we observed ~4-fold lower average probability of methylation (Figure 3.4f (inset), Figure 3.5g (inset)), consistent with the expectation that tethering of pA-Hia5 produces preferential methylation of deoxyadenosines closest to the antibody-bound nucleosome. Despite this reduction in total methylation of accessible DNA in CENP-A DiMeLo-seq reads compared to free pA-Hia5 treated reads, we detect a similar distribution of nucleosome densities in our chromatin array population (Figure 3.5i). We observed similar results for H3-antibody-directed methylation of H3 chromatin using pAG-Hia5 (Figure 3.5j-l). We conclude that directing pA-Hia5 activity using a histone-specific antibody targets specific methylation in proximity to the nucleosome of interest *in vitro.*

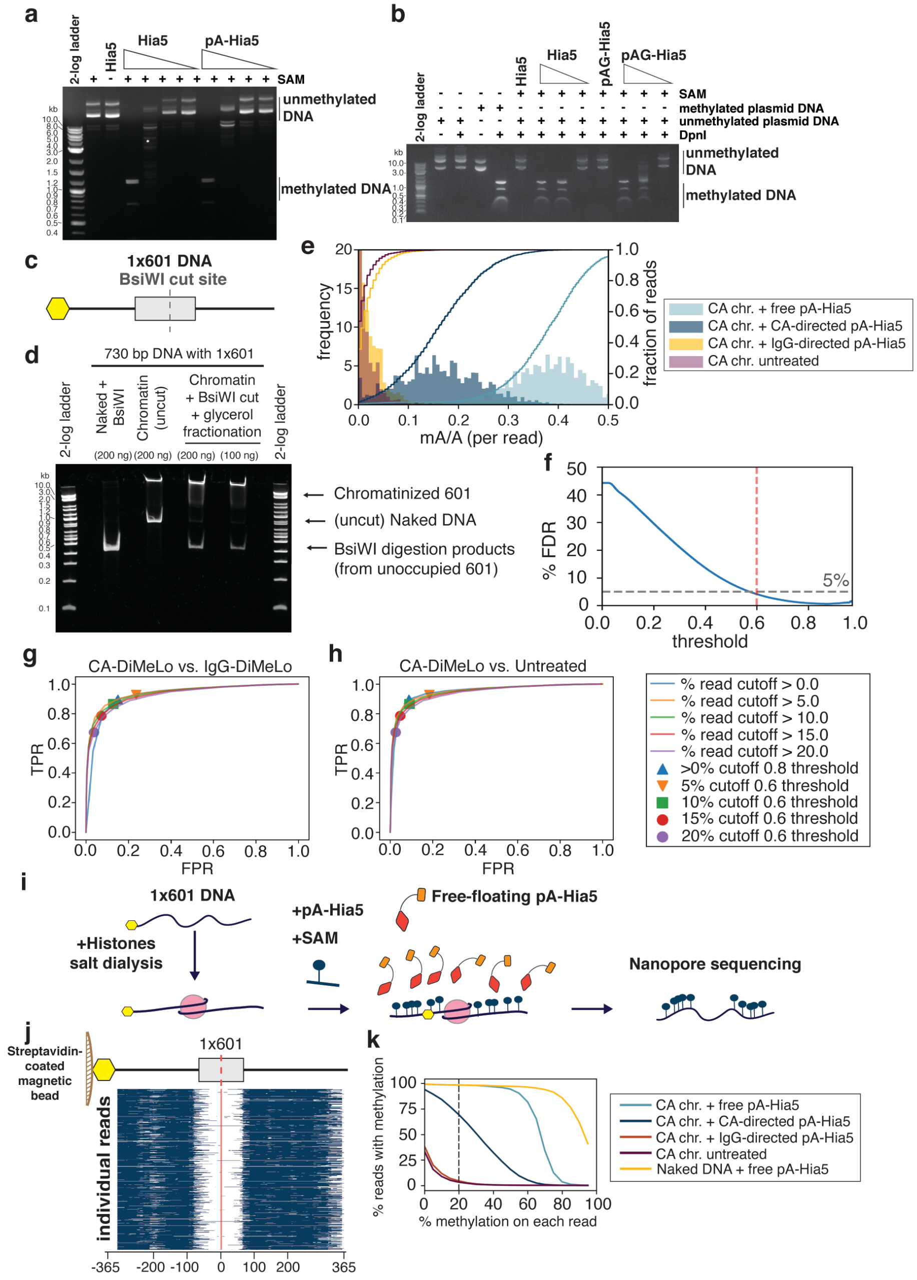


Figure 3.3 In vitro assessment of methylation of DNA and chromatin by pA-Hia5 and pAG-Hia5. **a**,**b**, Agarose gel electrophoresis image of DpnI digestion of (unmethylated) plasmid DNA following incubation with Hia5, pA-Hia5 (a), or pAG-Hia5 (b) (Supplementary Note 1). Representative images of at least 2 replicates. **c**, Schematic of 1x601 DNA sequence. Grey box indicates 601 sequence, Yellow hexagon indicates end with biotin. **d**, Native polyacrylamide gel electrophoresis of naked 1x601 DNA or chromatinized 1x601 DNA before and after BsiWI digestion and glycerol gradient fractionation. Representative image of at least 2 replicates. **e**, Histogram (filled bars, left axis) and cumulative distribution (line traces, right axis) of fraction of methylation (mA/A) on reads from CENP-A 1x601 chromatin methylated with free pA-Hia5, CENP-A-directed pA-Hia5, IgG-directed pA-Hia5, or untreated. Left y-axis is truncated at 20 for

better visualization. **f**, Plot showing percentage false discovery rate plotted against binned minimum mA probability score (Supplementary Note 4). Dotted lines indicate threshold - 0.6, 5% FDR. **g**,**h**, Receiver Operator Characteristic (ROC) curves comparing fraction of methylated reads from 1x601 CENP-A chromatin after CENP-A-directed methylation (True Positive Rate) to IgG-directed methylation (g) or no treatment (h) (False Positive Rate). Areas under the curves (AUC) for the ROC curves range between 0.92 and 0.94 for (g), and between 0.92 and 0.95 for (h). **i**, Schematic of methylation of accessible DNA on 1x601 CENP-A chromatin co-incubated with free pA-Hia5 and SAM. **j**, Heatmap showing methylation on 5000 individual reads from CENP-A chromatin following incubation with free pA-Hia5. Blue indicates methylation above threshold (0.6). **k**, Line plot showing percentage of reads with methylation as a function of the minimum percentage of methylation on each read. (methylation threshold - 0.6). Dotted line corresponds to methylation on at least 20% of each read (used in Figure 3.4d).

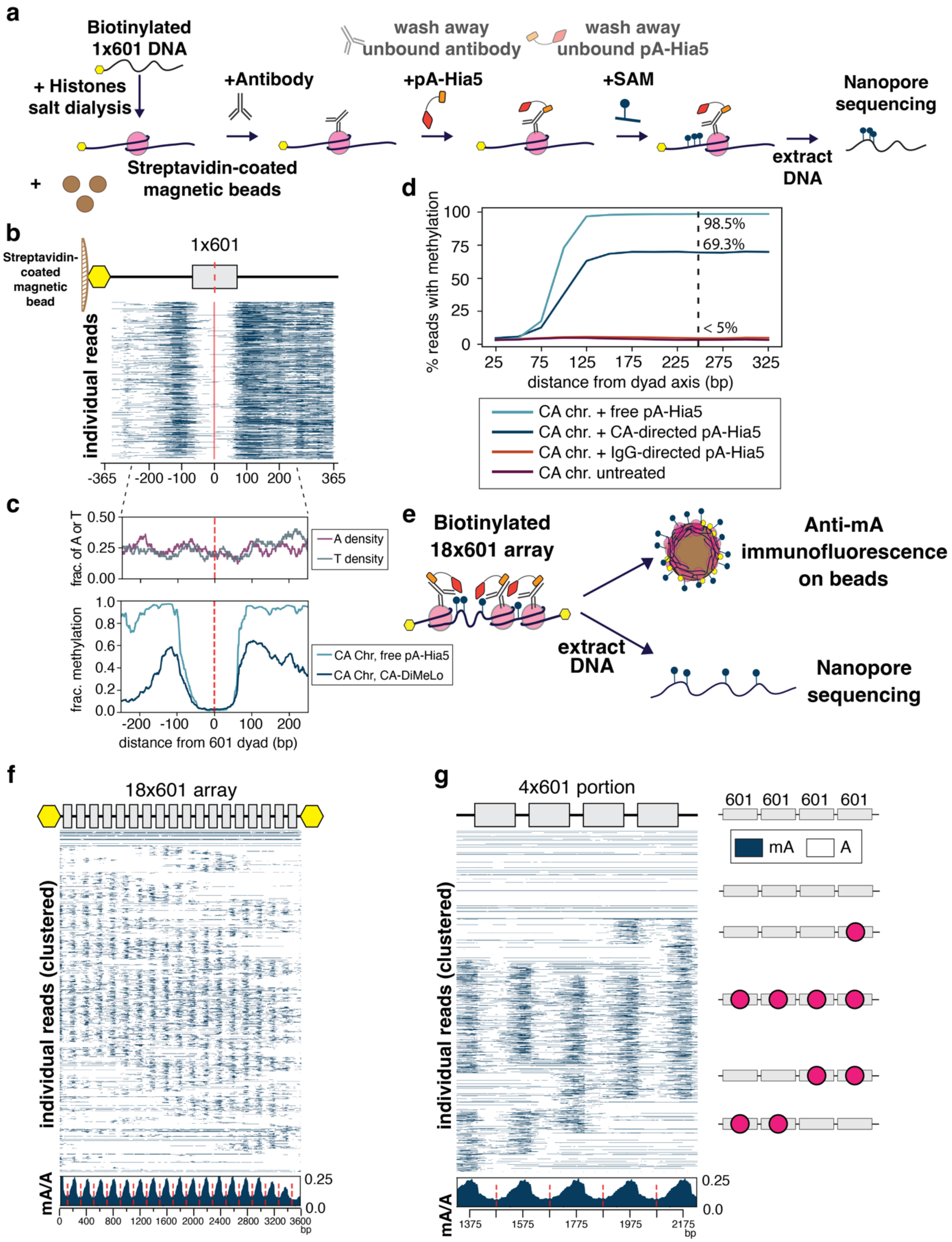


Figure 3.4 Antibody-directed methylation of artificial chromatin and long-read sequencing. **a**, Schematic of directed methylation of artificial chromatin depicting biotinylated chromatin

reconstitution on DNA containing 1x601 positioning sequence, specific antibody binding, pA-Hia5 targeting, SAM addition and activation, followed by long-read sequencing of methylated DNA extracted from chromatin. **b.** Heatmap showing methylation on 5000 individual 1x601 reads from chromatin containing CENP-A mononucleosomes methylated with CENP-A-directed pA-Hia5 (red dashed line indicates 601 dyad position). **c**, Plots showing A or T density (top) and average mA/A on every base position of 1x601 containing DNA (bottom) (red dashed line indicates 601 dyad position). **d**, Plot showing percentage of reads that have at least 20% methylation (Megalodon probability score threshold of 0.6) at increasing distance from the 601 dyad position (representative plot from one of three replicates). Dashed line at 250 bp from 601 dyad position shows percentage of reads with at least 20% methylation from replicate displayed. **e**, Schematic of directed methylation of 18x601 chromatin array depicting methylation by pA-Hia5 tethered to CENP-A nucleosomes by CENP-A-antibody followed by anti-methyladenine immunofluorescence or long-read sequencing of methylated DNA extracted from chromatin. **f**,**g**, Heatmap showing methylation on 2000 individual reads from CENP-A chromatin methylation with CENP-A-directed pA-Hia5, hierarchically clustered by jaccard distances of inferred nucleosome positions over the entire 18x601 array (f) or a subset 4x601 region (g) along with cartoons depicting predicted nucleosome positions (red circles). Insets below heatmaps show average mA/A on every base position of 18x601 array or 4x601 portion (red dashed line indicates 601 dyad position).

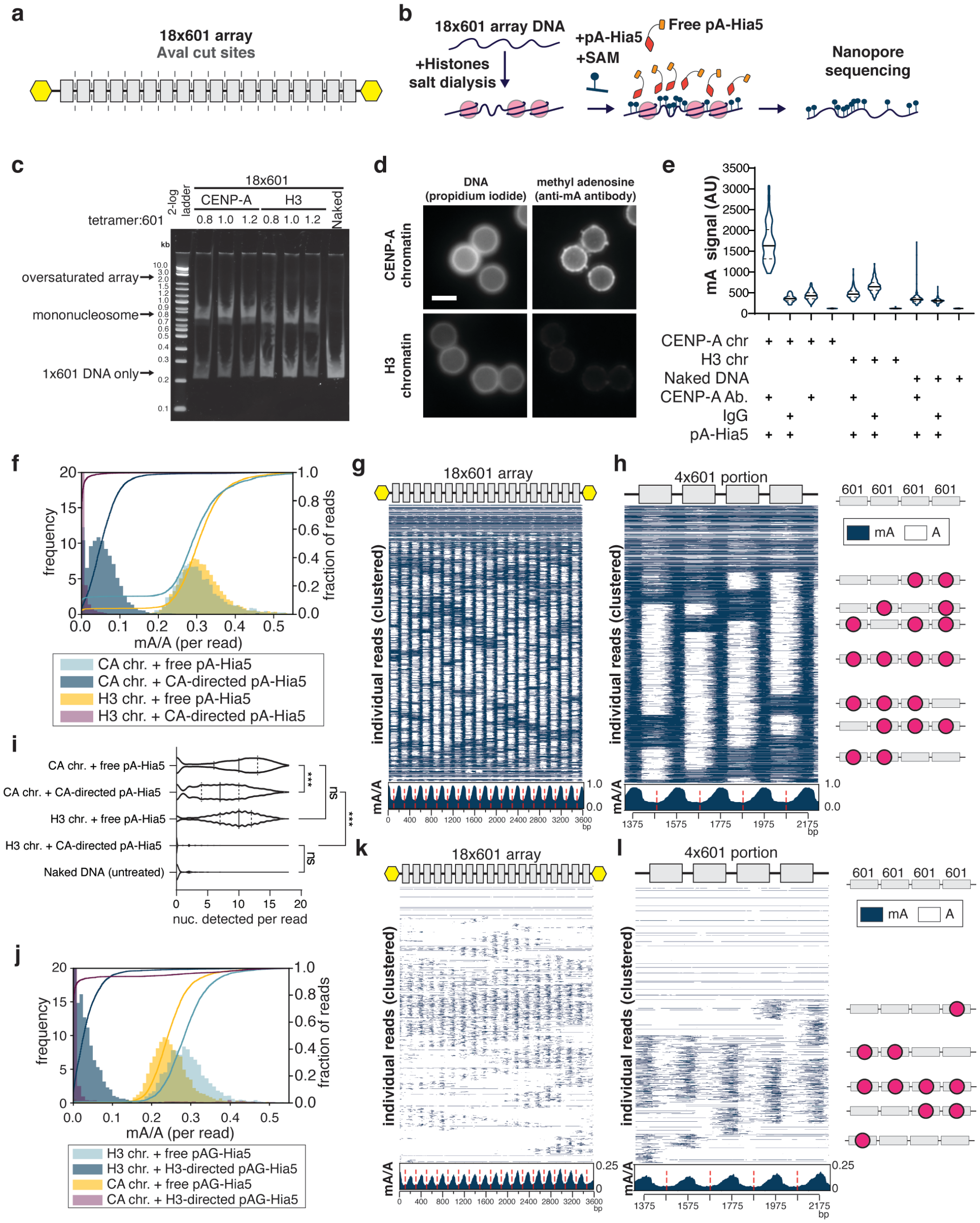


Figure 3.5 In vitro assessment of methylation of 18x601 array chromatin by pA-Hia5 and pAG-Hia5. **a**, Schematic showing the location of 601 sequences (grey boxes) and AvaI digestion sites (dashed line) in between 601 sequences on the 18x601 array. Yellow hexagons indicate biotinylation. **b**, Schematic of methylation of 18x601 chromatin reconstitution, incubation with free pA-Hia5 and SAM, and long-read sequencing of methylated DNA extracted from chromatin. **c**, Native polyacrylamide gel electrophoresis showing AvaI digested naked 18x601 array DNA or 18x601 chromatin array reconstituted with CENP-A or H3 (Supplementary Note 2). Representative gel image of at least 3 replicates. **d**, Representative immunofluorescence images of

chromatin-coated beads following methylation using CENP-A-directed pA-Hia5. Scale bar - 3 microns. **e**, Violin plots of immunofluorescence signal on (denatured) chromatin-coated beads following antibody-directed methylation. Solid line - median, dashed line - quartiles. n > 90 beads/condition. (Supplementary Note 5) **f**, Histogram (filled bars, left axis) and cumulative distribution (line traces, right axis) of fraction of methylation (mA/A) on reads from CENP-A or H3 chromatin methylated with free pA-Hia5 or CENP-A-directed pA-Hia5. Left y-axis is truncated at 20 for better visualization. **g**,**h**, Heatmap showing methylation on 2000 individual reads from CENP-A chromatin methylation with free pA-Hia5, clustered over the entire 18x601 array (g) or a subset 4x601 region (Supplementary Note 4) along with cartoons depicting predicted nucleosome positions (red circles) (h). Insets below heatmaps show average mA/A on every base position of 18x601 array or 4x601 portion. (red dashed line indicates 601 dyad position). **i**, Violin plot of nucleosomes detected per read on reads from CENP-A or H3 18x601 chromatin array methylated with free pA-Hia5, or CENP-A-directed pA-Hia5. Solid line - median, dashed lines - quartiles. n = 3000 reads. Statistical significance was calculated using Kruskal-Wallis test. *** - P-value < 0.0001 ns - P-value > 0.05. **j**, Histogram (filled bars, left axis) and cumulative distribution (line traces, right axis) of fraction of methylation (mA/A) on reads from CENP-A or H3 chromatin methylated with free pA-Hia5 or CENP-A-directed pA-Hia5. Left y-axis is truncated at 20 for better visualization. **k**,**l**, Same as g,h, but corresponding to H3 chromatin methylation with H3-directed pAG-Hia5.

3. Optimization of LMNB1 mapping **in situ**

We next optimized DiMeLo-seq for mapping protein-DNA interactions *in situ* in permeabilized nuclei from a human cell line (HEK293T). To do this, we mapped the interaction sites of lamin B1 (LMNB1), which is often targeted in DamID studies to profile lamina associated domains (LADs).[32] Large regions of the genome that are almost always in contact with the nuclear lamina across cell types are called constitutive lamina associated domains (cLADs). Regions that are rarely in contact with the nuclear lamina across cell types and instead reside in the nuclear interior are called constitutive inter-LADs (ciLADs) (Figure 3.6a). Other regions can vary in their lamina contact frequency between cell types and/or between cells of the same type. We chose LMNB1 as an initial target because (i) cLADs and ciLADs provide well-characterized on-target and off-target control regions, respectively; (ii) LMNB1 has a very large binding footprint (LADs have a median size of 500 kb and cover roughly 30% of the genome),[64] so DNA-LMNB1 interactions can be detected even with very low sequencing coverage; (iii) LMNB1 localization at the nuclear lamina can be easily visualized by immunofluorescence, allowing for intermediate quality control using microscopy during each step of the protocol (Figure 3.7c,d); and (iv) we have previously generated LMNB1 DamID data from HEK293T cells using bulk and single-cell protocols, providing ample reference materials.[35]

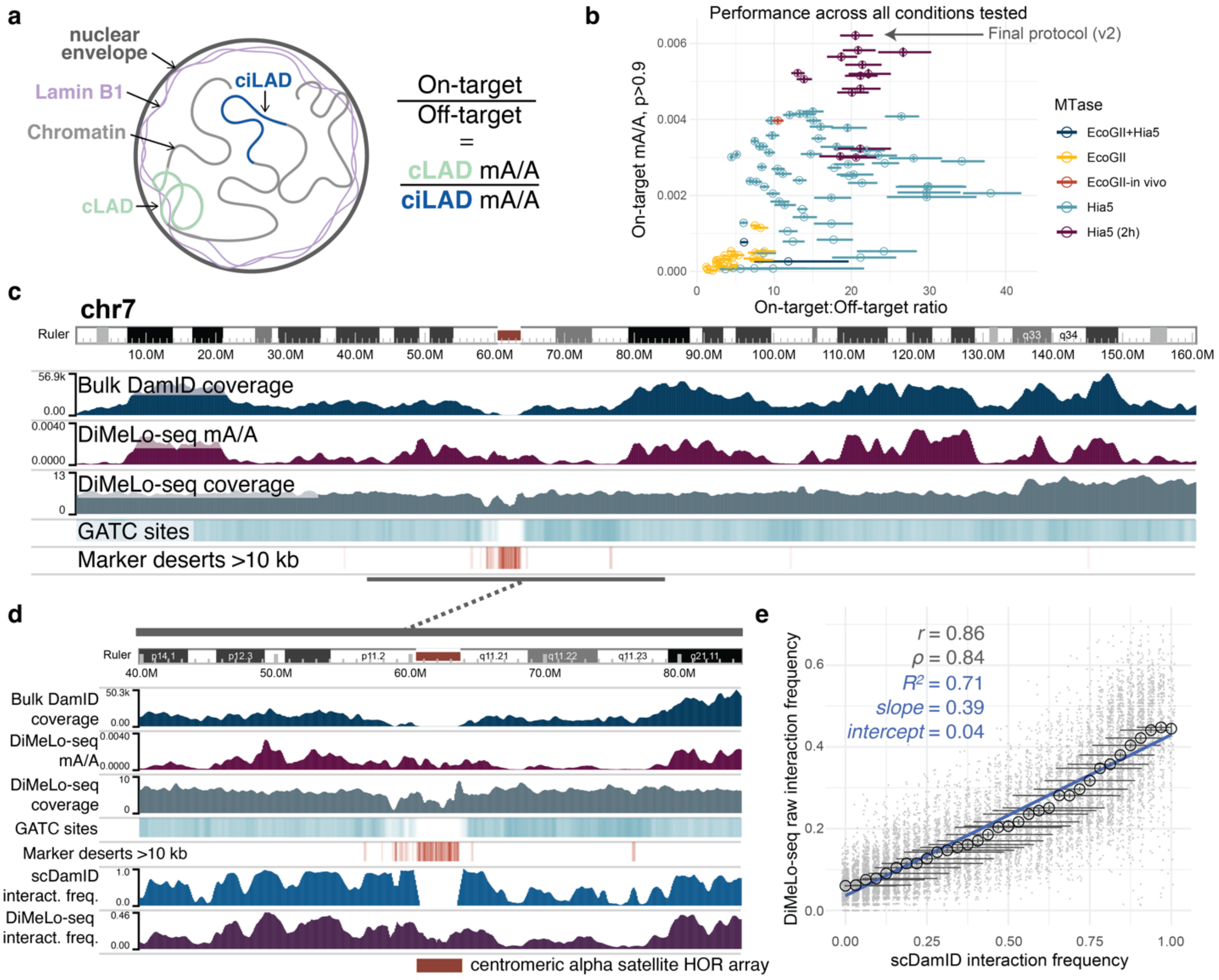


Figure 3.6 Optimization of DiMeLo-seq targeting Lamin B1 *in situ*. **a,** Schematic of interactions between LMNB1 and lamina-associated domains, and the use of mA levels in cLADs and ciLADs to estimate on-target and off-target mA. **b,** Scatterplot showing for each protocol condition tested the proportion (y-axis) of all A bases (basecalling q>10, n = min 1.4M, max 28M A bases per condition) called as methylated (stringent threshold p>0.9; abbreviated mA/A) across all reads in on-target (cLAD) regions and the ratio (x axis) of these mA levels compared to off-target (ciLAD) regions. Circles are colored by the methyltransferase condition used. Error bars provide a measure of uncertainty due to each condition's sequencing coverage (described below). Complete data are in Table 3.1. **c,** A browser image across all of chromosome 7 comparing *in situ* LMNB1-targeted DiMeLo-seq (protocol v1) to *in vivo* LMNB1-tethered DamID data (blue) [22]. The coverage of each region by simulated DpnI digestion fragments (splitting reference at GATC sites) between 150 and 750 bp (sequenceable range) is indicated by a teal heatmap track (range 0 to 0.7). The presence of intervals longer than 10 kb between unique 51-mers in the reference, a measure of mappability, is indicated with an orange heatmap track. **d,** A closer view of the centromere on chr7, with added tracks at the bottom illustrating LMNB1 interaction frequencies from single-cell DamID data [22], as well as from DiMeLo-seq data (protocol v1). **e,** For a quality-filtered set of 100 kb genomic bins (gray points, Supplementary Note 7, n = 11292 total bins), a comparison of LMNB1 interaction frequency estimates from DiMeLo-seq (protocol v1; black circles indicate mean across

n = 94 to 663 genomic bins, computed for each genomic bin as the prop. of n = 61 to 335 overlapping reads with at least 1 mA call with p>0.9) versus scDamID (prop. of n = 32 cells with detected interactions in each genomic bin). A linear regression line computed across all bins is overlaid (blue). Error bars in (b) and (e) represent 95% credible intervals determined for each proportion, mean of proportions, or ratio of proportions by sampling from posterior beta distributions computed using uninformative priors.

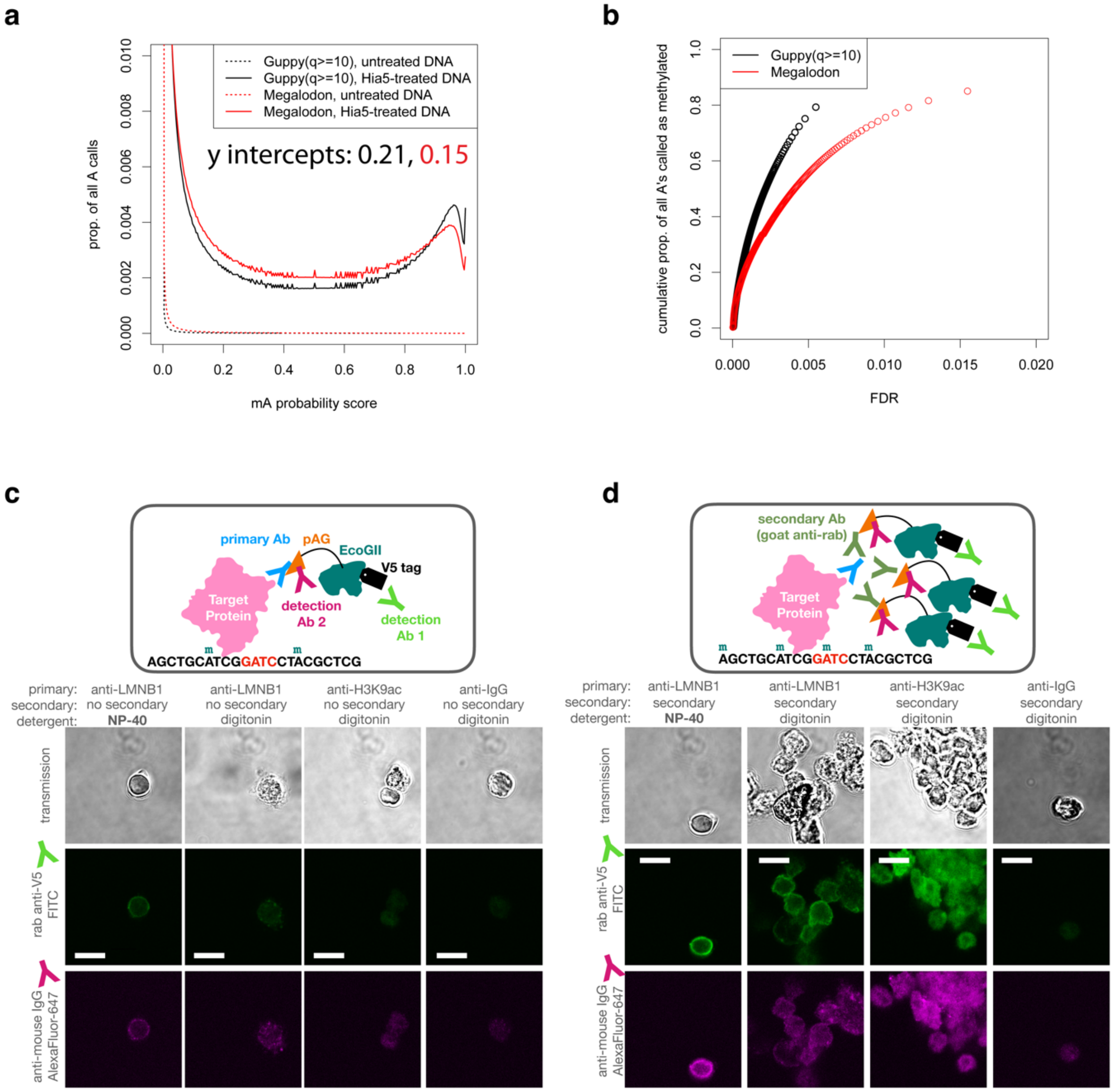


Figure 3.7 Assessment of mA calling and LMNB1 targeting. **a**, The proportion of all adenines called as methylated at each possible probability mA probability score using two different software packages on ONT reads from two GM12878 DNA samples: untreated genomic DNA and purified genomic DNA methylated by Hia5 *in vitro*. The untreated DNA provides a measure of the false positive rate (FPR) at each score, since it contains few or no methyl adenines. The Hia-5 treated

DNA provides a lower bound on the true positive rate (TPR) at each threshold. **b**, Estimates of the proportion of As methylated in the Hia5-treated DNA sample at each false discovery rate (FDR) threshold (FDR=FPR/(TPR+FPR), determined from a). At least 80% of the adenines on the Hia5-treated DNA appear to be methylated. **c-d**, In the DiMeLo-seq workflow, following the primary antibody and pA/G-MTase binding and wash steps, a sample of nuclei can be taken for quality assessment by immunofluorescence. One can determine the locations and relative quantity of pA/G-MTase molecules using fluorophore-conjugated antibodies that bind to the pA/G-MTase but not to the primary antibody. In these representative images, the results for pAG-EcoGII are shown, comparing different antibodies, detergents, and samples with (d) and without (c) the use of an unconjugated secondary antibody to recruit more pA/G-MTase molecules to the target protein. Scale bars representing 10 microns are shown in the FITC channel images as white lines.

To assess the performance of the LMNB1-targeted DiMeLo-seq protocol, we quantified the proportion of adenines that were called as methylated across all reads mapping to cLADs (on-target regions), and across all reads mapping to ciLADs (off-target regions). We evaluated the performance of each iteration of the protocol using both the on-target methylation rate (as a proxy for sensitivity) and the on-target:off-target ratio (as a proxy for signal-to-background), aiming to increase both. We developed a rapid pipeline for testing variations of many components of the protocol, allowing us to go from harvested cells to fully analyzed data in under 60 hours (Methods: DiMeLo-seq, Methods: Nanopore library preparation and sequencing, and Supplementary Notes 6-8). With this optimization pipeline, we tested over 100 different conditions (Figure 3.6b), varying the following: methyltransferase type (Hia5 vs. EcoGII), input cell numbers, detergents, primary antibody concentrations, the use of secondary antibodies, enzyme concentrations, incubation temperatures, methylation incubation times, methylation buffers, and SAM concentrations (Supplementary Note 8, Table 3.1). We validated an initial version of the protocol (v1), and then further optimized the methyltransferase activation conditions to increase the amount of on-target methylation 50-60% without sacrificing specificity (v2; see Figure 3.7, Figure 3.8, Supplementary Note 8, and Figure 3.6b). To confirm that this optimization would apply to other types of proteins, we also examined the results of different protocol variations targeting the protein CTCF and found them to be concordant (Figure 3.9a).

We also verified that there is very little loss of performance when using cells that were cryopreserved in DMSO-containing media or lightly fixed in paraformaldehyde, when using between 1-5 million cells per replicate, or when using concanavalin-A coated magnetic beads to carry out cell washing steps by magnetic separation instead of centrifugation (Methods: DiMeLo-seq, Supplementary Notes 9-10, Table 3.1). To confirm antibody specificity, we performed IgG isotype controls and free-floating Hia5 controls to measure nonspecific methylation and DNA accessibility, respectively (Methods: DiMeLo-seq, Table 3.2). We also generated a stably transduced line expressing a direct fusion between EcoGII and LMNB1 *in vivo*, as in MadID,[65] then we detected mAs with nanopore sequencing (Figure 3.8a and Supplementary Note 10). This *in vivo* approach produced threefold more on-target methylation compared to *in situ* DiMeLo-seq with pAG-EcoGII (Figure 3.6b), though this performance is expected to vary with different fusion proteins and their expression levels (Supplementary Note 10).

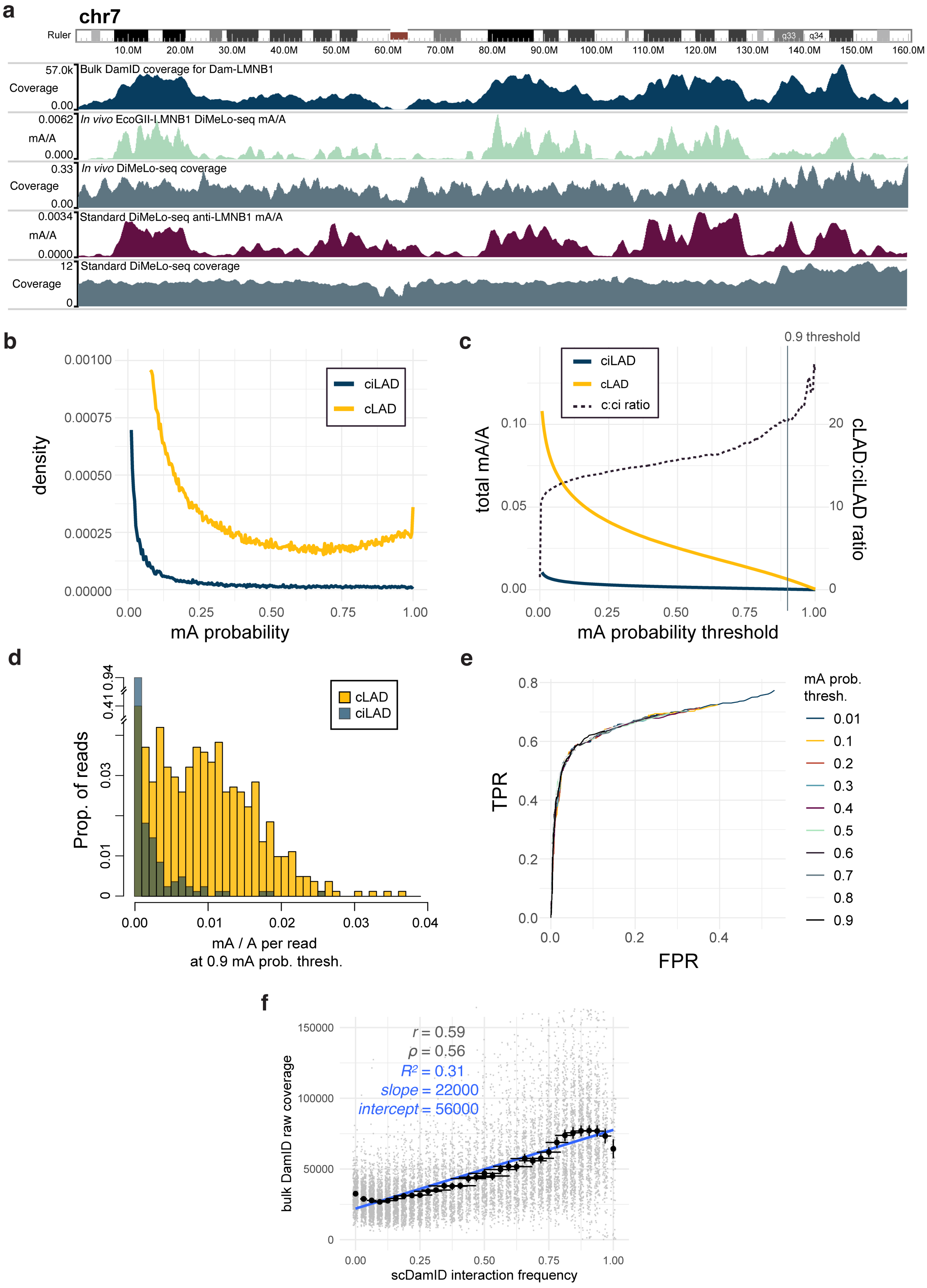

a
chr7
Ruler
q33
q34
10.0M
20.0M
30.0M
40.0M
50.0M
60.0M
70.0M
80.0M
90.0M
100.0M
110.0M
120.0M
130.0M
140.0M
150.0M
160.0M
57.0k
Bulk DamID coverage for Dam-LMNB1
Coverage
0.00
0.0062
In vivo EcoGII-LMNB1 DiMeLo-seq mA/A
mA/A
0.000
0.33
In vivo DiMeLo-seq coverage
Coverage
0.00
0.0034
Standard DiMeLo-seq anti-LMNB1 mA/A
mA/A
0.0000
12
Standard DiMeLo-seq coverage
Coverage
0
b
density
0.00100
0.00075
0.00050
0.00025
0.00000
0.00
0.25
0.50
0.75
1.00
mA probability
ciLAD
cLAD
c
0.9 threshold
ciLAD
cLAD
c:ci ratio
total mA/A
0.10
0.05
0.00
20
10
0
cLAD:ciLAD ratio
0.00
0.25
0.50
0.75
1.00
mA probability threshold
d
0.94
0.41
0.03
0.01
0
Prop. of reads
cLAD
ciLAD
0.00
0.01
0.02
0.03
0.04
mA / A per read
at 0.9 mA prob. thresh.
e
0.8
0.6
0.4
0.2
0.0
TPR
0.0
0.2
0.4
FPR
mA prob.
thresh.
0.01
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
f
150000
100000
50000
0
bulk DamID raw coverage
r = 0.59
ρ = 0.56
R² = 0.31
slope = 22000
intercept = 56000
0.00
0.25
0.50
0.75
1.00
scDamID interaction frequency

Figure 3.8 Demonstration of in vivo LMNB1-targeting and estimation of in situ sensitivity and specificity. **a**, A browser view of chr7 comparing in vivo EcoGII-LMNB1 DamID (second track, green) to conventional LMNB1 in vivo DamID (first track, blue), and to LMNB1-targeted in situ DiMeLo-seq (fourth track, dark red). **b**, For an in situ LMNB1-targeting experiment using the final v2 protocol (#120 in Table 3.1), the distributions of guppy mA probability scores across all A bases (q>10) on all reads mapping to cLADs (gold, representing on-target methylation; n = 2.8M) or ciLADs (blue, representing off-target methylation; n = 2.1M). **c**, As in b, but showing the cumulative distributions for all mA calls above each probability score threshold, with the ratio between these plotted as a dotted line (using the right-hand y-axis). Vertical line indicates the stringent threshold of 0.9, at which cLADs have 20 times more mA as a proportion of all As (0.6%) than do ciLADs. If the threshold is reduced to 0.5, the fraction of As called as methylated increases to 2.5% but the cLAD:ciLAD ratio decreases to 15.6. **d**, On a per-read basis, for all reads with at least 500 A basecalls (q>10) and using a mA probability threshold of 0.9, the distribution of mA/A called on each read for cLADs (n = 812 reads) vs. ciLADs (n = 827 reads). **e**, Receiver-Operator Characteristic (ROC) curve showing, for different mA calling thresholds, the ability to classify individual reads from (d) as originating from cLADs or ciLADs using a simple linear threshold on mA/A. At a false positive rate of 6%, reads can be classified with a true positive rate of 59%, and this is similar for all mA thresholds used. The total Area Under the Curve (AUC) for the p>0.9 curve is 0.78. **f**, As in Figure 3.6e, but for bulk conventional DamID raw coverage. The y axis is truncated to omit outliers for visualization (max = 300000), but these were not omitted for linear model and correlation computation. Error bars in x represent the proportion of 32 cells +/- 2 standard errors of the proportion. Error bars in y represent the mean of n = 94 to 663 genomic bins +/- 2 standard errors of the mean.

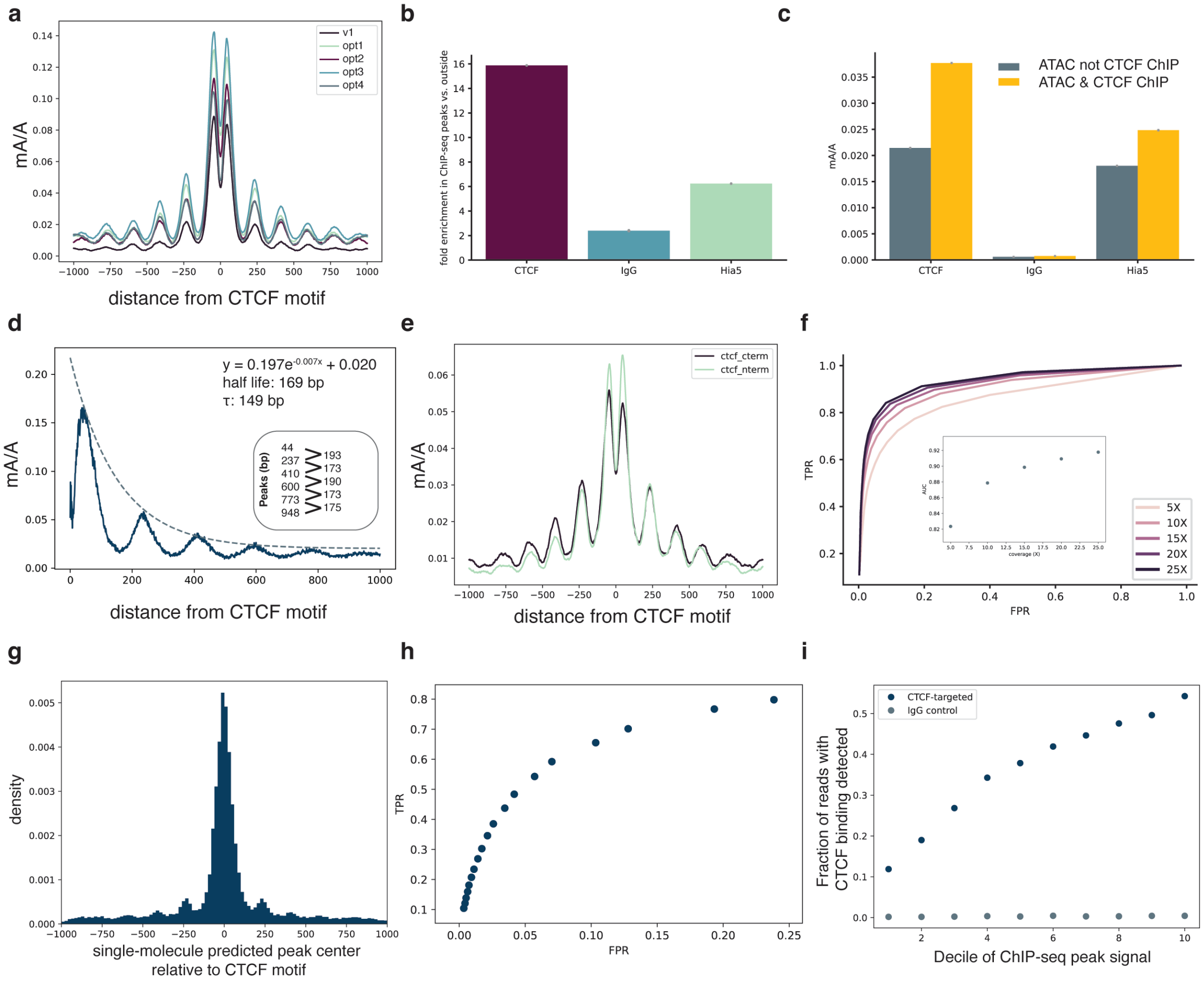
We found that DiMeLo-seq and conventional bulk DamID are highly concordant in the non-repetitive parts of the genome (Spearman correlation = 0.71 in 1 Mb bins), but conventional DamID achieves little-to-no coverage across pericentromeric regions (Figure 3.6c). This is due in part to the low availability of unique sequence markers to map short reads to in the pericentromere, but also to the low frequency of GATC (the binding motif for Dam and DpnI in the DamID protocol) within centromeric repeats (Figure 3.6c).[65] DiMeLo-seq, unlike DamID, produces long reads that can be uniquely mapped across the centromeric region of chromosome 7, revealing that this region has an intermediate level of contact with the nuclear lamina (Figure 3.6c,d).

Because DiMeLo-seq directly probes unamplified genomic DNA, each sequencing read represents a single, native DNA molecule from a single cell, sampled independently and with near-uniform probability from the population of cells. This allows for estimation of absolute protein-DNA interaction frequencies, i.e., the proportion of cells in which a site is bound by the target protein, without needing to account for the amplification bias inherent to other protein-DNA mapping methods. We leveraged single-cell Dam-LMNB1 DamID data from the same cell line[35] to assess the relationship between DiMeLo-seq methylation and an orthogonal estimate of protein-DNA interaction frequencies. This revealed a nearly linear relationship between the two interaction frequency estimates, with a simple linear model achieving an R^2 of 0.71, compared to an R^2 of 0.31 when scDamID-based interaction frequencies are compared to bulk conventional DamID coverage (Figure 3.6e, Figure 3.8f). We note that scDamID tends to slightly overestimate intermediate interaction frequencies compared to DiMeLo-seq, attributable to the *in vivo* vs. *in situ* nature of the two protocols,[61] as well as to the fact that homolog-specific information is collapsed within each hypotriploid HEK293T cell.[30,35] This analysis demonstrates that DiMeLo-seq is capable of estimating absolute protein-DNA interaction frequencies without needing to account for amplification bias, while capturing heterogeneity in protein-DNA interactions at the single-cell level.

4. Joint analysis of CTCF binding and CpG methylation on single molecules

DiMeLo-seq measures protein-DNA interactions in the context of the local chromatin environment by simultaneously detecting endogenous CpG methylation, nucleosome occupancy, and protein binding. To highlight this feature of DiMeLo-seq, we targeted CTCF, a protein that strongly positions surrounding nucleosomes and whose binding is inhibited by CpG methylation.[66] We first validated that targeted methylation is specific to CTCF in GM12878 cells by calculating the fraction of adenines that are methylated within GM12878 CTCF ChIP-seq peaks relative to the fraction of adenines methylated outside of these peaks. We chose to target CTCF in GM12878 cells because GM12878 is an ENCODE Tier 1 cell line with abundant ChIP-seq reference datasets. We measured a 16-fold increase in targeted methylation over background in our CTCF-targeted sample (Figure 3.9b). We also measured a 6-fold mA/A enrichment in the free pA-Hia5 control in CTCF ChIP-seq peaks, which reflects the fact that many CTCF binding sites overlap with accessible regions of the genome where pA-Hia5 can methylate more easily.[67] However, both the free pA-Hia5 and the IgG controls produced significantly less targeted methylation than the CTCF-targeted sample (Figure 3.9b). We confirmed that signal enrichment is caused by CTCF-targeted methylation and not accessibility of CTCF sites by measuring a 1.8X greater proportion of mA in ChIP-seq peaks compared to regions of open chromatin measured by ATAC-seq (Figure 3.9c).

As further validation of DiMeLo-seq's concordance with ChIP-seq data and to visualize protein binding on single molecules, we analyzed mA and mCpG across individual molecules spanning CTCF motifs within ChIP-seq peaks of various strengths (Figure 3.10a). DiMeLo-seq signal tracks with ChIP-seq signal strength, with mA density decreasing from the top to bottom quartiles of ChIP-seq peak signal. We observed an increase in local mA surrounding the binding motif, with a periodic decay in methylation from the peak center, indicating methylation of neighboring linker DNA between strongly positioned nucleosomes (Figure 3.9d). The 88 bp dip at the center of the binding peak reflects CTCF's binding footprint[68–70] and is evident even on single molecules. CTCF binds to ~50 bp of DNA as determined by DNase I footprinting and ChIP-exo.[68,71,72] The larger footprint observed with DiMeLo-seq is likely due to steric hindrance with Hia5 unable to methylate DNA within ~20 bp of the physical contact between CTCF and DNA as efficiently. We also observed an asymmetric methylation profile, with stronger methylation 5' of the CTCF motif. This increased methylation relative to 3' of the motif extends beyond the central peak to the neighboring linker DNA. We hypothesized that this asymmetry was a result of the antibody binding the C-terminus of CTCF, thereby positioning pA-Hia5 closer to the 5' end of the binding motif. To test this hypothesis, we compared DiMeLo-seq binding profiles in top quartile ChIP-seq peaks when using an antibody targeting the C-terminus of CTCF, as is used in Figure 3.10, and an antibody targeting the N-terminus of CTCF. We observed methylation enrichment 5' to the binding motif with C-terminus targeting and 3' to the motif with N-terminus targeting (p-value: 0.00010, Supplementary Note 11, Figure 3.9e). The free pA-Hia5 control profile supports this finding that the antibody binding site is causing the peak asymmetry, as there is no significant asymmetry in this untargeted case (Figure 3.11).



j

Quartile	Number of motifs in ChIP-seq peaks	Number of reads	Number of reads with at least one mA >thresh75	Number of reads with at least one mCpG >thresh75
4	7901	110657	94437	67167
3	7902	119002	92004	69396
2	7902	122162	81783	70007
1	7902	132063	67238	81949

Figure 3.9 Analysis of CTCF targeting performance. **a**, Enrichment profiles with mA probability threshold of 0.75 at the top quartile of ChIP-seq peaks for the DiMeLo-seq protocol v1 compared to four optimization conditions (opt1: 2 hour activation, 0.05 mM spermidine at activation, replenish SAM; opt2: 2 hour activation, 0.05 mM spermidine at activation, replenish SAM, 500 nM pA-Hia5; opt3: 2 hour activation, 0.05 mM spermidine at activation, replenish SAM, pA-Hia5 binding at 4°C for 2 hours; opt4: 2 hour activation, no spermidine, 1 mM Ca++ and 0.5 mM Mg++ buffer) (Supplementary Note 11). **b,** Fold enrichment over background of mA/A in ChIP-seq peak regions. Error bars represent the 95% credible interval for each ratio of proportions determined by

sampling proportions from posterior beta distributions computed with uninformative priors. **c,** mA/A in ATAC-seq peaks that do not overlap CTCF ChIP-seq peaks (grey) and mA/A in ATAC-seq peaks that do overlap CTCF ChIP-seq peaks (yellow). Error bars are computed as in (b) **d**, Methylation decay from the CTCF motif center for the top decile of ChIP-seq signal is fit with an exponential decay function. The positions of the peaks are indicated, with the spacing between peaks also noted. **e,** Methylation profiles at top quartile of ChIP-seq peaks when targeting the C-terminus or N-terminus of CTCF. The difference between antibody binding site produces significantly different profiles (Supplementary Note 11). **f,** Receiver-Operator Characteristic (ROC) curves from aggregate peak calling with DiMeLo-seq targeting CTCF at 5-25X coverage using ChIP-seq as ground truth. Inset shows Area Under the Curve (AUC) as a function of coverage. **g,** The distribution of differences between our single-molecule predicted peak center and the known CTCF motif are plotted for single molecules within top decile ChIP-seq peaks. **h,** ROC curve for binary classification of CTCF-targeted DiMeLo-seq reads to identify CTCF-bound molecules based on each read's proportion of methylated adenines in peak regions (Supplementary Note 11). At a FPR of 5.7%, a TPR of 54% is achieved. **i**, Fraction of reads that have a CTCF binding event detected in the peak region for each decile of ChIP-seq peak strength for the CTCF-targeted sample and IgG control. Calculated using thresholds determined from analysis in (h). Error bars do not extend beyond the points themselves so are not shown. **j**, Number of motifs and reads displayed in Figure 3.10a.

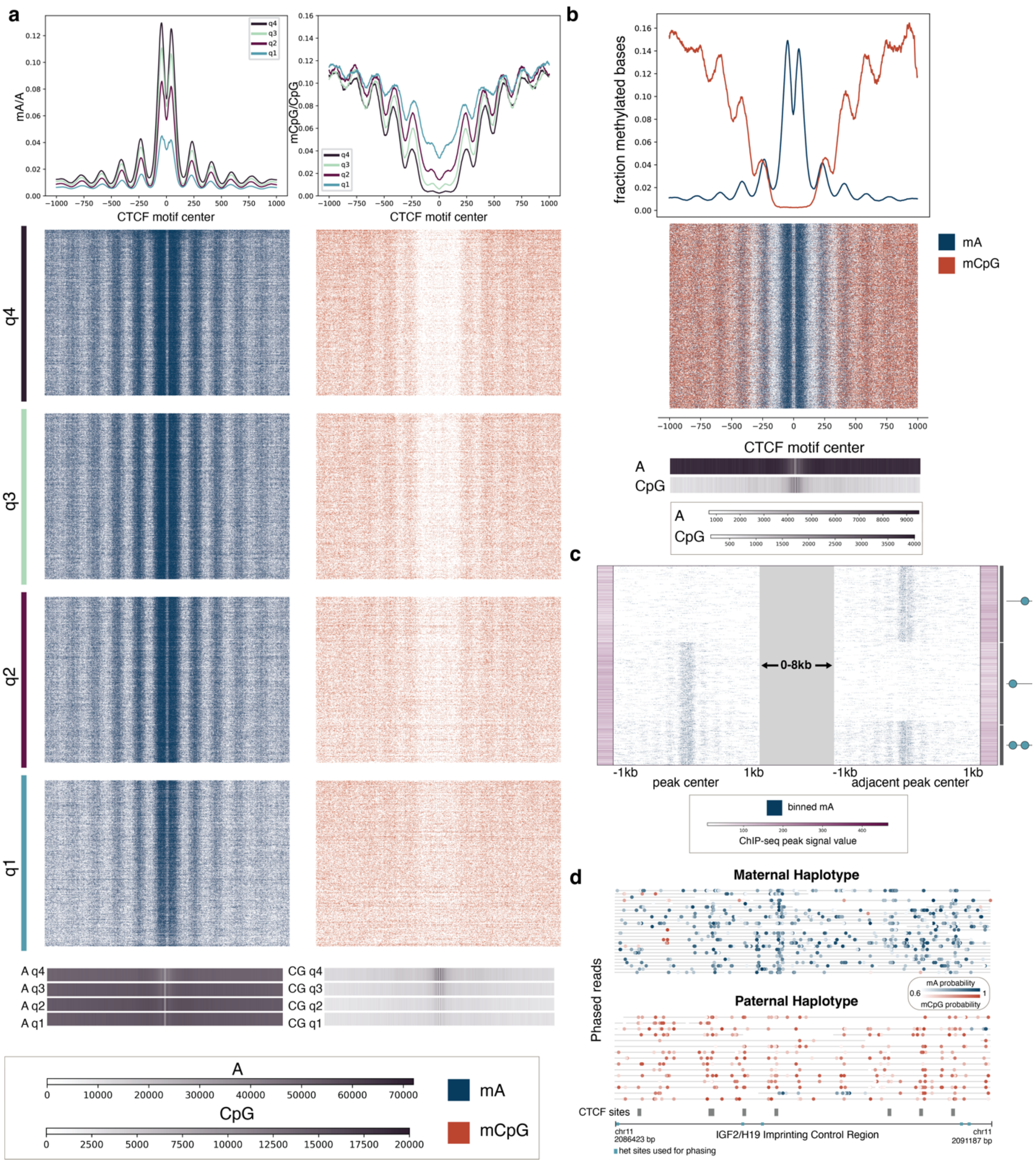
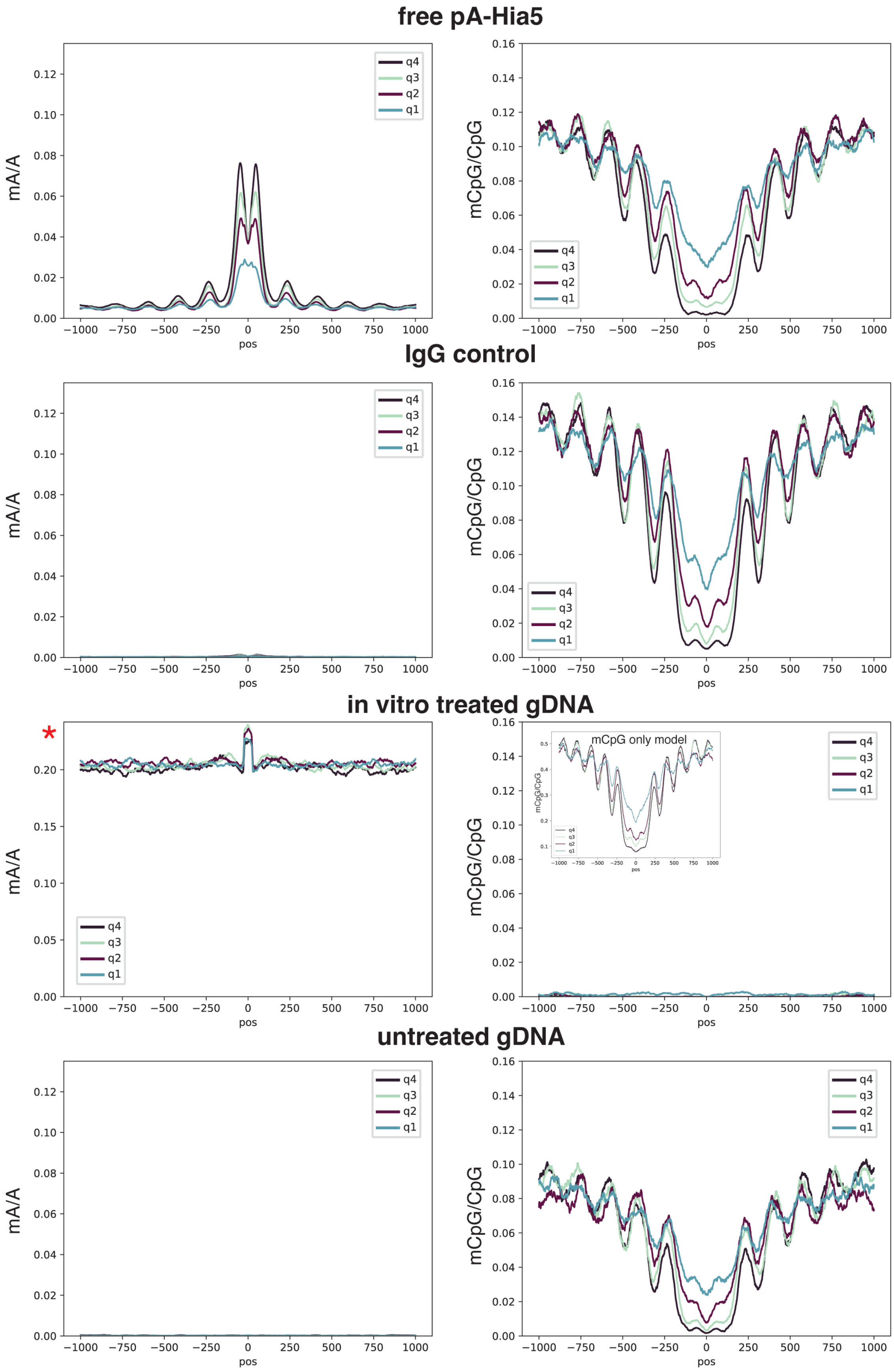


Figure 3.10 Single-molecule CTCF binding and CpG methylation profiles. **a,** Single molecules spanning CTCF ChIP-seq peaks are shown across quartiles of ChIP-seq peak strength within 1 kb of the peak center. Q4, quartile 4, are peaks with the strongest ChIP-seq peak signal, while Q1, quartile 1, are peaks with the weakest ChIP-seq peak signal. Blue points indicate mA called with probability ≥ 0.75, while orange points indicate mCpG called with probability ≥ 0.75. Aggregate curves for each quartile were created with a 50 bp rolling window. Base density across the 2 kb region for each quartile is indicated in the 1D heatmaps; the scale bar indicates the number of adenine bases and CG dinucleotides sequenced at each position relative to the motif center. **b,** Joint

mA and mCpG calls on the same individual molecules spanning the upper decile of ChIP-seq peaks are displayed. Molecules displayed have at least one mA called and one mCpG called with probability ≥ 0.75. Aggregate curves were created with a 50 bp rolling window. Base density is indicated as in (a). **c,** CTCF site protein occupancy is measured on single molecules spanning two neighboring CTCF motifs within 2-10 kb of one another. CTCF motifs are selected from all ChIP-seq peaks, and molecules are shown that have a peak at least one of the two motifs. Each row is a single molecule, and the molecules are anchored on the peaks that they span, with a variable distance between the peaks indicated by the grey block. ChIP-seq peak signal for each of the motif sites is indicated with the purple bars. The graphic on the side illustrates the CTCF binding pattern for each cluster. **d,** Phased reads across the IGF2/H19 Imprinting Control Region with CTCF sites indicated in grey. Blue dots represent mA calls and orange dots represent mCpG calls. Heterozygous sites used for phasing are indicated in turquoise.



free pA-Hia5

IgG control

in vitro treated gDNA

untreated gDNA

Figure 3.11 Control mA and mCpG profiles at CTCF peaks. Profiles at CTCF ChIP-seq peaks for free pA-Hia5, IgG control, *in vitro* treated genomic DNA, and untreated genomic DNA. Quartiles indicate rank of ChIP-seq peak strength. All axes are the same scaling as in Figure 3.10a, except for mA/A of *in vitro* treated gDNA. With high mA levels achieved only with this *in vitro* methylated control, mC basecalling fails. However, if the Rerio model res_dna_r941_min_modbases_5mC_CpG_v001.cfg is used for calling mCpG separately from mA, the mCpG profile is restored, as seen in the inset for the *in vitro* treated gDNA sample. Importantly, as indicated by the y-axis scale in the inset, if mCpG is called separately from mA, the detected mCpG levels are higher.

To evaluate the use of DiMeLo-seq for de novo peak detection, we called CTCF peaks using DiMeLo-seq data alone and created ROC curves at increasing sequencing depth using ChIP-seq peaks as ground truth (Supplementary Note 11, Figure 3.9f). At ~25X coverage, we detected 60% of ChIP-seq peaks (FPR 1.6%) and measured an AUC of 0.92 (Supplementary Note 11). Among the peaks detected with DiMeLo-seq that were not annotated ChIP-seq peaks, ten percent overlapped 1 kb marker deserts and gaps in the hg38 reference and are undetectable by ChIP-seq. Another 12% of these peaks fell within 500 bp of a known CTCF motif.

We next probed the relationship between CTCF binding and endogenous CpG methylation. Single molecules spanning CTCF binding sites in stronger ChIP-seq peaks exhibited a larger dip in mCpG around the motif compared to the shallower dip in weaker ChIP-seq peaks (Figure 3.10a). This inverse relationship between CpG methylation and CTCF-targeted methylation reflects previous findings that mCpG inhibits CTCF binding.[66] We measured both mA and mCpG on the same single molecules and also observed that both A and CpG are preferentially methylated in linker DNA (Figure 3.10b). The increased methylation of CpG in linker DNA relative to nucleosome-bound DNA surrounding CTCF sites is supported by previous studies that have similarly reported higher levels of mCpG in linker DNA than nucleosomal DNA around CTCF sites.[73]

CTCF's known binding motif and abundance genome-wide make it a good target for characterizing the resolution of DiMeLo-seq. To characterize resolution, we estimated the peak center on single molecules spanning the top decile of CTCF ChIP-seq peaks (Supplementary Note 11). The mean single-molecule peak center was 6 bp 5' of the CTCF motif center, and the peak center on approximately 70% of the reads fell within +/- 200 bp of the motif center (Figure 3.9g). This systematic bias towards predicting the peak center 5' of the motif can be explained by the observed asymmetry in methylation when targeting the C-terminus of CTCF. Another factor that impacts the resolution of DiMeLo-seq is the reach of the methyltransferase, which can be characterized by measuring the decay rate of methylation density from the peak center. To do this, we fit the average adenine methylation density with respect to the motif center to an exponential function and calculated a half-life of 169 bp (Figure 3.9d). Together, this analysis suggests that DiMeLo-seq can resolve binding events to within about 200 bp; however, this metric is likely dependent on the protein target and influenced by the local chromatin environment.

To characterize the sensitivity of DiMeLo-seq for detecting CTCF binding events on single molecules, we performed a binary classification of individual CTCF-targeted DiMeLo-seq reads based on each read's proportion of methylated adenines within CTCF peak regions, defined as +/- 150 bp around the CTCF binding motif center. For top-decile ChIP-seq peaks, which are regions that are most likely to contain CTCF binding, we classified reads containing CTCF binding events with 54% sensitivity (5.7% FPR, Figure 3.9h,i, Supplementary Note 11).

We next investigated the ability of DiMeLo-seq to measure protein binding at adjacent sites on single molecules. We first characterized CTCF occupancy across two binding sites that were spanned by a single molecule. We were able to detect neighboring CTCF motifs that are bound by CTCF at both sites or just one of the two sites, and the detected binding appears to track with ChIP-seq peak strength (Figure 3.10c). This analysis demonstrates the potential of DiMeLo-seq to analyze coordinated binding patterns on long single molecules, which is not possible with short-read methods. We further investigated this potential within a specific HLA locus on chr6 where haplotype-specific SNPs within the CTCF binding motif prevent CTCF binding at one of the two neighboring sites (Figure 3.12a). DiMeLo-seq can map haplotype-specific interactions because long reads often span multiple heterozygous sites, allowing reads to be phased. Importantly, at 25X coverage, we were able to detect the binding patterns of both sites on the same single molecule and could attribute the lack of detected binding at one of the two sites to a mutation within the binding motif. The ability to map haplotype-specific interactions is also useful in studying imprinted genomic regions such as the IGF2/H19 Imprinting Control Region, where CpG methylation on the paternal allele prevents CTCF binding, while on the maternal allele, CTCF is able to bind (Figure 3.10d). We also reported haplotype-specific CTCF binding profiles at specific sites and broadly across the active and inactive X chromosomes (Figure 3.12b-d). These results demonstrate that DiMeLo-seq can measure the effect of haplotype-specific genetic or epigenetic variation on protein binding.

To test the compatibility of DiMeLo-seq with other long-read sequencing platforms capable of modification calling, we performed Pacific Biosciences (PacBio) sequencing of DNA from a CTCF-targeted DiMeLo-seq sample and from an unmethylated control (Supplementary Note 12). We found similar enrichment profiles using both methods (Figure 3.13), indicating that DiMeLo-seq is compatible with PacBio's circular consensus sequencing technique. However, while PacBio sequencing has reported improved base calling accuracy,[74] this approach detected more methylation in the unmethylated control than Nanopore, slightly reducing the signal-to-noise ratio of the measurement (Figure 3.13).

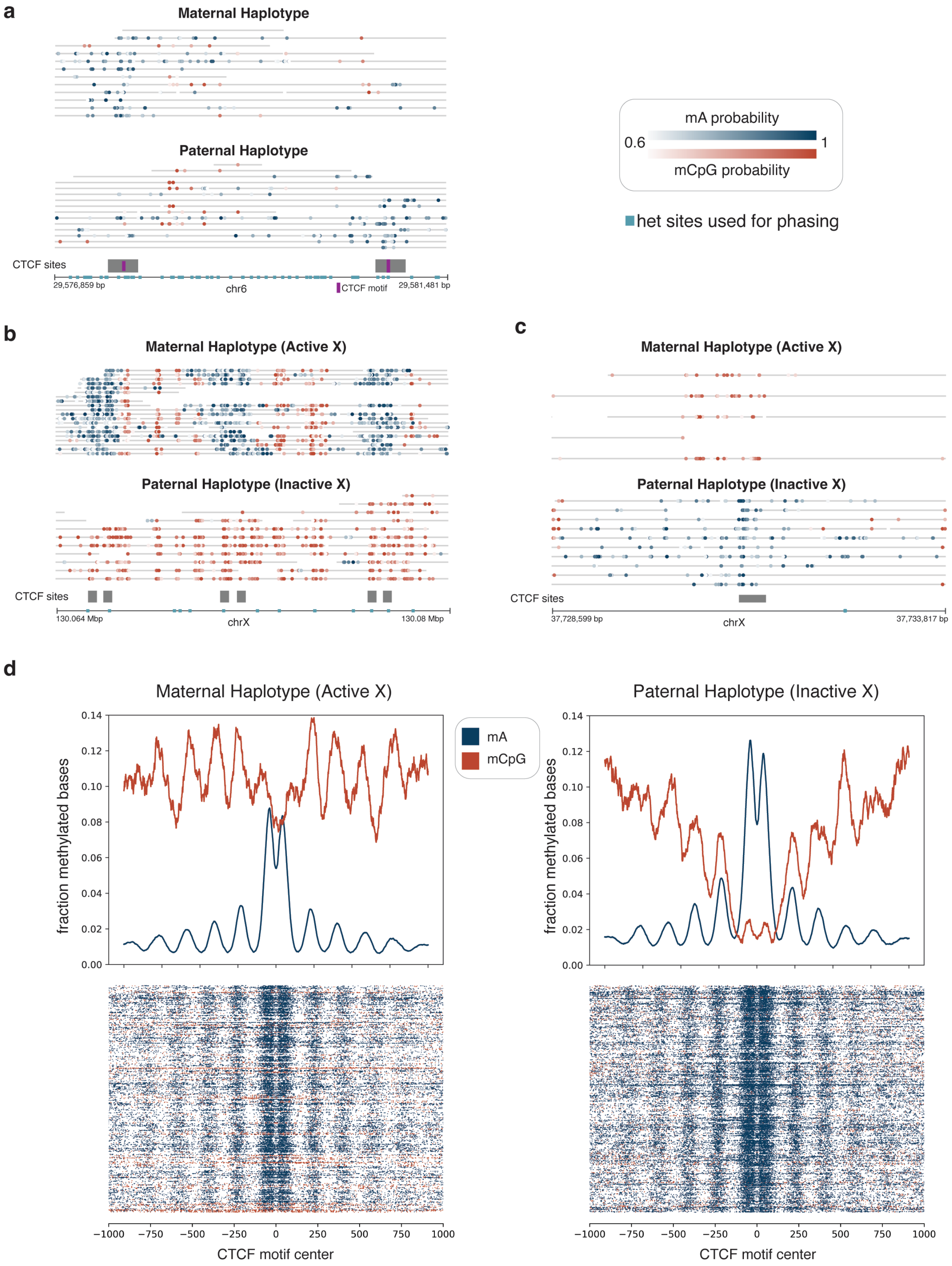

a
Maternal Haplotype
Paternal Haplotype
CTCF sites
29,576,859 bp
chr6
CTCF motif
29,581,481 bp
mA probability
0.6
1
mCpG probability
het sites used for phasing
b
Maternal Haplotype (Active X)
Paternal Haplotype (Inactive X)
CTCF sites
130.064 Mbp
chrX
130.08 Mbp
c
Maternal Haplotype (Active X)
Paternal Haplotype (Inactive X)
CTCF sites
37,728,599 bp
chrX
37,733,817 bp
d
Maternal Haplotype (Active X)
Paternal Haplotype (Inactive X)
fraction methylated bases
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
mA
mCpG
−1000
−750
−500
−250
0
250
500
750
1000
CTCF motif center

Figure 3.12 Phased CTCF-targeted DiMeLo-seq reads. Phased reads across one region on chr6 and two regions on chrX illustrate haplotype-specific CTCF binding due to genetic and epigenetic differences between haplotypes. **a,** A region on chr6 within the human leukocyte antigen (HLA) locus which contains two CTCF binding sites and many heterozygous SNPs useful for phasing reads. Both CTCF binding sites overlap a het SNP within their binding motif. At the first CTCF site, the paternal SNP allele within the motif is associated with weak or no CTCF binding on the paternal haplotype, and the opposite is true at the second CTCF site. Thus, only one of these two neighboring sites tends to be bound on each haplotype, which is clearly visible on reads spanning both CTCF sites. Further, because CpG methylation patterns are similar between the two haplotypes, these binding differences likely owe to the genetic differences present in/near the CTCF binding motifs themselves. **b-c,** Because the GM12878 cell line has two X chromosomes and was clonally derived, one X homolog (the paternally inherited X homolog for this cell line) has undergone X inactivation and remains inactive in all cells. Shown here are one region with CTCF binding on the active X only (b) and one region with CTCF binding on the inactive X only (c). The haplotype-specific CTCF binding patterns in these chrX regions appear to be associated with haplotype-specific CpG methylation, as similarly seen for the imprinted H19 locus shown in Figure 3.10d. **d,** Aggregate enrichment profiles from DiMeLo-seq reads across all CTCF sites on chrX are shown, as in Figure 3.10b. Each row in the heatmaps below the aggregate plots represents a single molecule centered at the CTCF motif. Notable strips of CpG hypermethylated reads are visible on the active X, as observed previously [27,75].

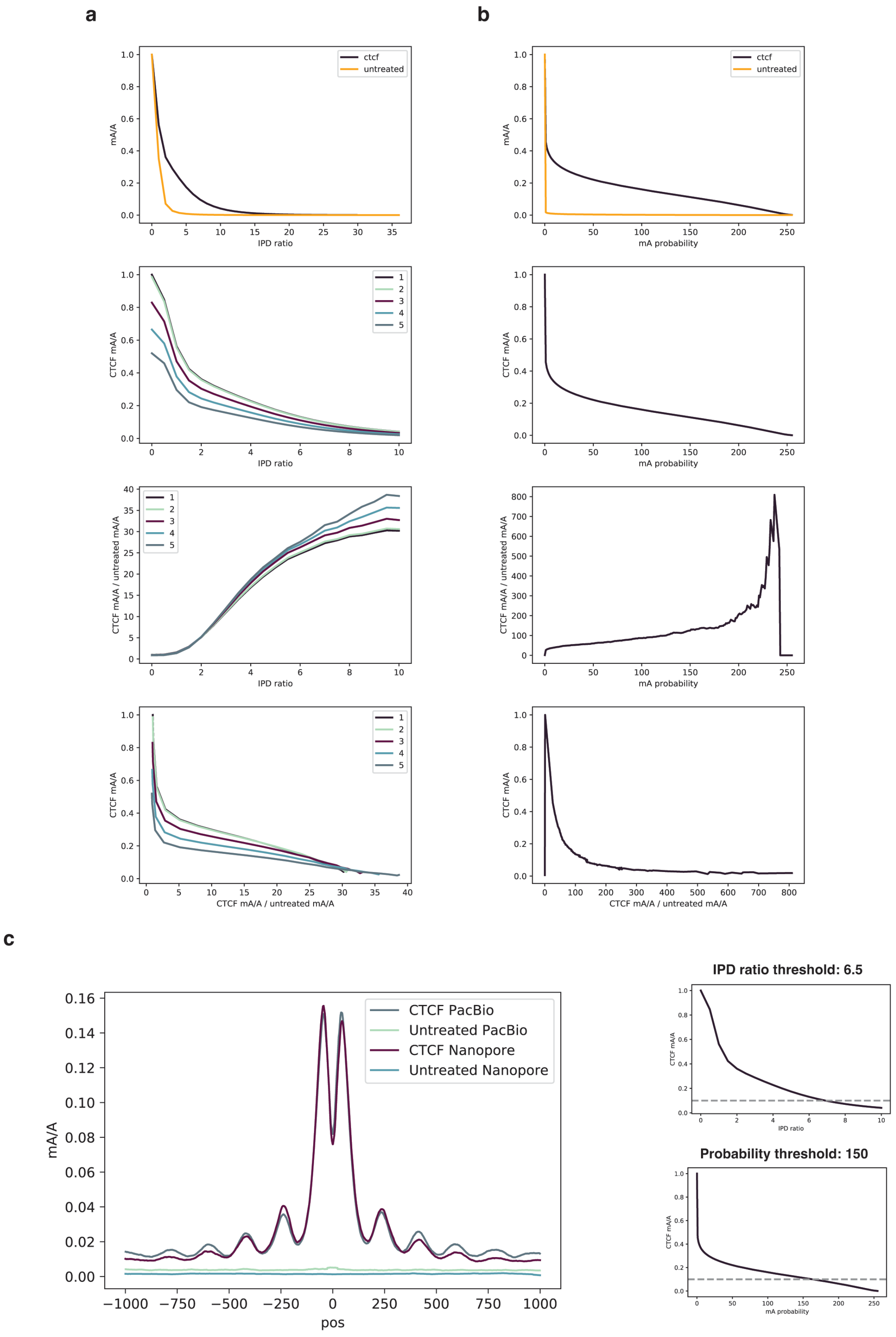

a
mA/A
IPD ratio
ctcf
untreated
0 5 10 15 20 25 30 35
CTCF mA/A
IPD ratio
1
2
3
4
5
CTCF mA/A / untreated mA/A
IPD ratio
CTCF mA/A
CTCF mA/A / untreated mA/A
b
mA/A
mA probability
ctcf
untreated
0 50 100 150 200 250
CTCF mA/A
mA probability
CTCF mA/A / untreated mA/A
mA probability
CTCF mA/A
CTCF mA/A / untreated mA/A
0 100 200 300 400 500 600 700 800
c
CTCF PacBio
Untreated PacBio
CTCF Nanopore
Untreated Nanopore
mA/A
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16
pos
−1000 −750 −500 −250 0 250 500 750 1000
IPD ratio threshold: 6.5
CTCF mA/A
IPD ratio
Probability threshold: 150
CTCF mA/A
mA probability

Figure 3.13 Comparison of PacBio and Nanopore sequencing platforms for detecting mA from DiMeLo-seq. The same DNA from a DiMeLo-seq experiment targeting CTCF in GM12878 cells was sequenced on both PacBio and Nanopore. The same untreated GM12878 DNA was also sequenced on both platforms. Methylated base calls for reads spanning the top decile of CTCF ChIP-seq peaks are analyzed. **a,** PacBio data. (i) Fraction of adenines methylated +/- 100 bp ("peak region") from CTCF motif center as a function of IPD ratio for the CTCF-targeted sample and the untreated control. (ii) Fraction of adenines methylated for CTCF-targeted sample in the peak region for various IPD ratio thresholds and number of pass thresholds (indicated in legend from 1 to 5). (iii) Fraction of adenines methylated in the peak region for CTCF-targeted sample over the fraction for the untreated control as a function of IPD ratio and number of passes (indicated in legend from 1 to 5). (iv) Fraction of adenines methylated in the peak region for CTCF-targeted sample versus the enrichment of CTCF-targeted methylation over the untreated control. **b,** Nanopore data. Same as in (a), but probability of methylation is the threshold that varies rather than IPD ratio and number of passes. **c,** For a given fraction of adenines methylated in the peak region, here 0.1 for illustration, the PacBio and Nanopore enrichment profiles are overlaid. The thresholds for each platform for 10% peak methylation are indicated and the number of passes threshold for PacBio is one.

5. Mapping protein-DNA interactions in centromeric regions

Mapping histone modifications in heterochromatin with DiMeLo-seq

To test DiMeLo-seq's ability to measure protein occupancy in heterochromatic, repetitive regions of the genome we targeted H3K9me3, which is abundant in pericentric heterochromatin. We chose to target H3K9me3 in HG002 cells because the chromosome X centromere has been completely assembled for this male-derived lymphoblast line,[59] and many different sequencing data types are available for it.[23] To validate the specificity of targeted methylation, we calculated the fraction of adenines methylated within HG002 CUT&RUN H3K9me3 peaks[76] compared to the fraction of adenines methylated outside of broadly defined peaks (Supplementary Note 13). For H3K9me3 targeting in HG002 cells, the enrichment of mA/A in CUT&RUN peaks was 3.6-fold over background (Figure 3.14a), indicating enrichment of methylation within expected H3K9me3-containing regions of the genome.

Human centromeres are located within highly repetitive alpha-satellite sequences, which are organized into higher order repeats (HORs).[76–79] To validate enrichment of H3K9me3-directed mA signal in centromeres, and in particular in HOR arrays, we similarly calculated the fold increase in mA/A and found 1.9-fold enrichment in centromeres and 3.0-fold enrichment in active (kinetochore-binding) HOR arrays[76] over non-centromeric regions (Figure 3.14b). We next looked at HOR array boundaries and observed a decrease in H3K9me3 across the boundary moving from within to outside of HOR arrays (Figure 3.14c). In contrast, for the free pA-Hia5 control, mA/A increases moving from within to outside of the HOR array, as chromatin becomes more accessible (Figure 3.15a).[23]

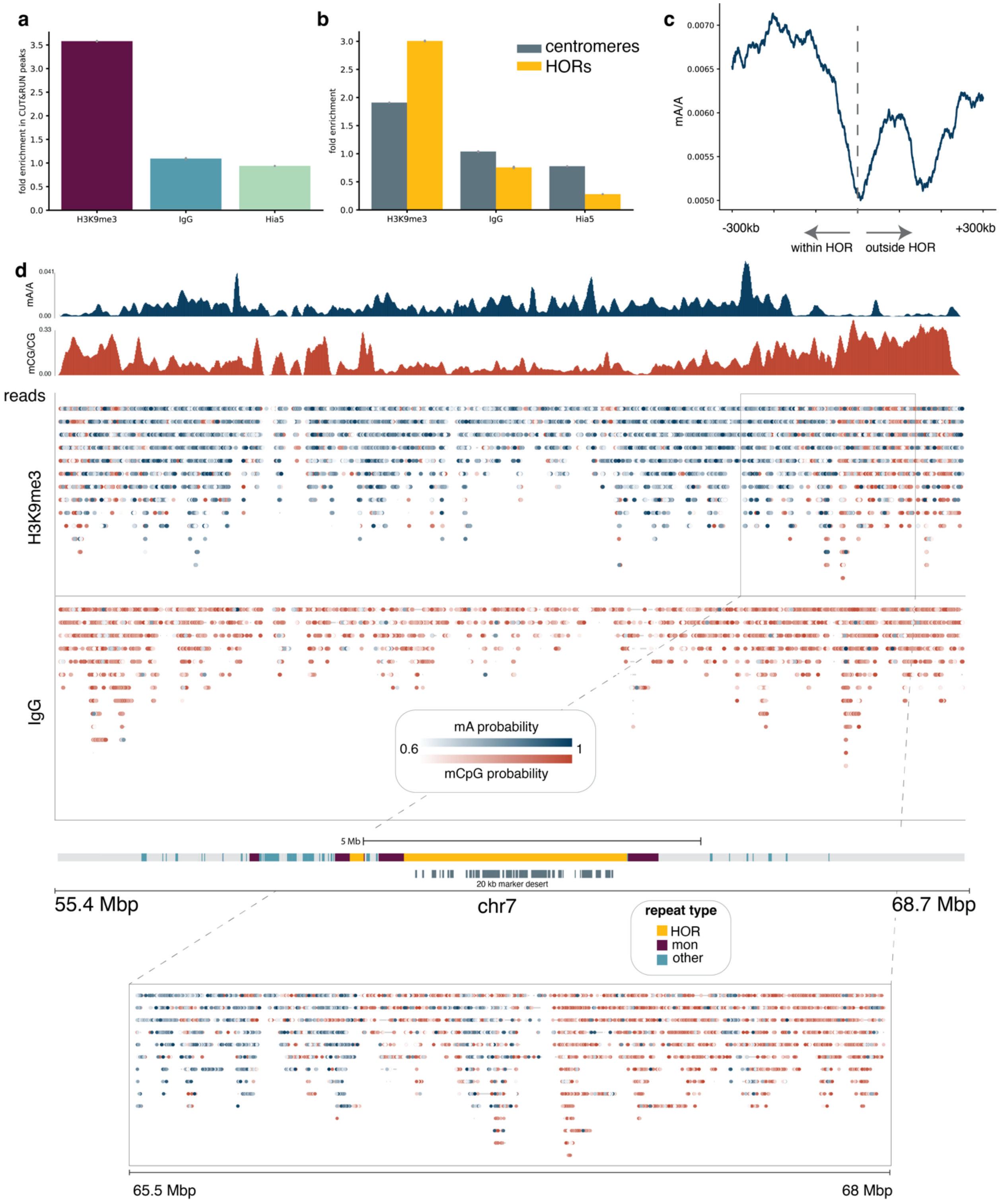


Figure 3.14 Detecting H3K9me3 in centromeres. **a,** The proportion of adenines methylated within CUT&RUN peaks relative to the proportion of adenines methylated outside of CUT&RUN broad peak regions is reported for the H3K9me3-targeted sample as well as IgG and free pA-Hia5 controls. Error bars represent 95% credible intervals determined for each ratio by sampling from posterior beta distributions computed with uninformative priors. **b,** The fraction of adenines

methylated within centromeres relative to non-centromeric regions, and similarly the fraction of adenines methylated within active HOR arrays relative to non-centromeric regions are displayed for the H3K9me3-targeted sample as well as the IgG and free pA-Hia5 controls. Error bars are defined as in (a). **c,** The decline in mA/A for the H3K9me3-targeted sample in a rolling 100 kb window from -300 kb within the HOR array to 300 kb outside of the HOR array. HOR array boundaries that transition quickly into non-repetitive sequences were considered: 1p, 2pq, 6p, 9p, 13q, 14q, 15q, 16p, 17pq, 18pq, 20p, 21q, 22q. **d,** Single molecules are displayed across the centromere of chromosome 7 for the H3K9me3-targeted sample and the IgG control. Reads mapping to the same position are displayed vertically, and modified bases are colored by the probability of methylation at that base for probabilities ≥ 0.6. Aggregate tracks show mA/A and mCpG/CpG in the H3K9me3-targeted sample in 10 kb bins. Grey bars below centromere annotation indicate regions with >20 kb marker deserts.

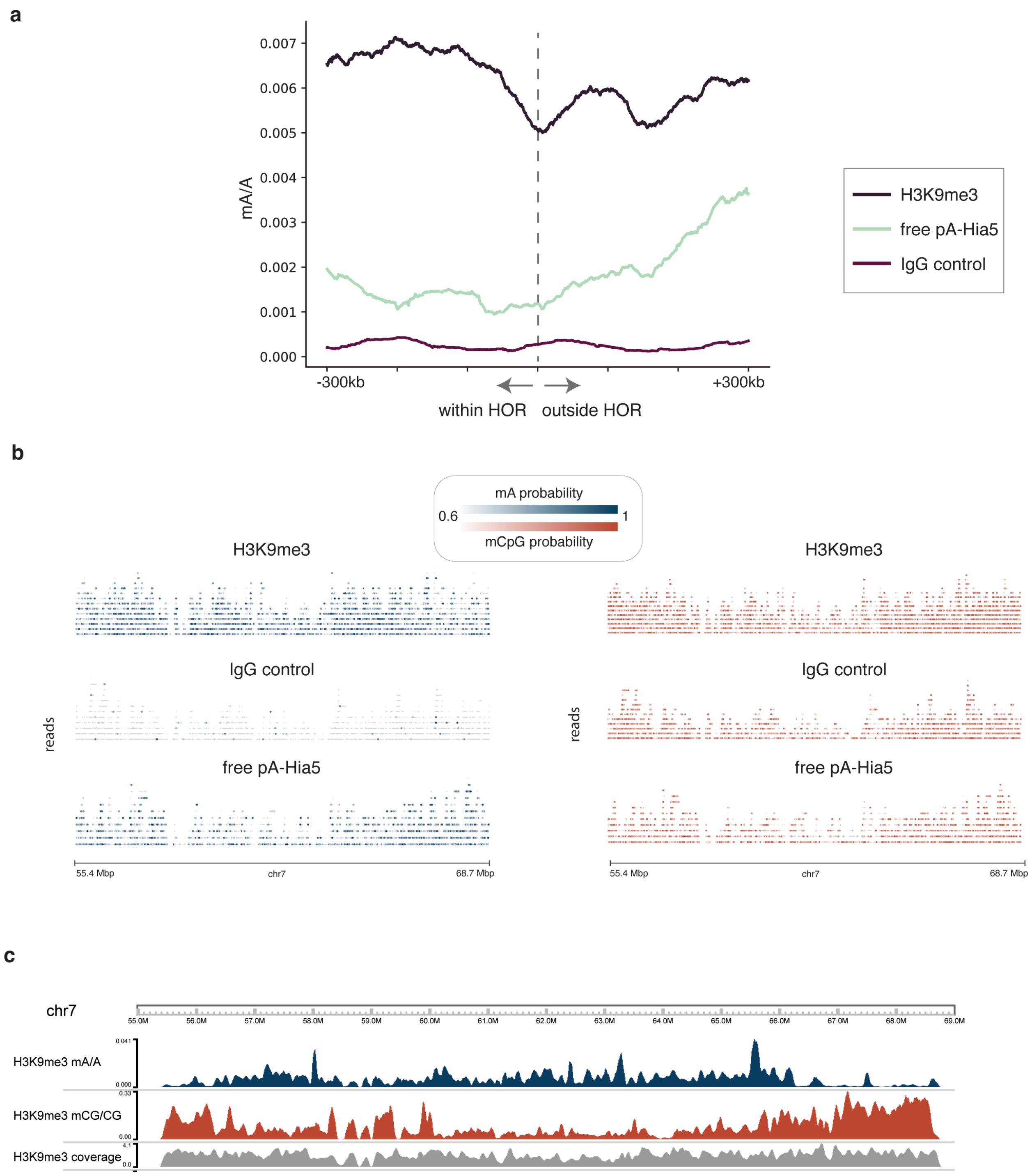


Figure 3.15 H3K9me3 control analysis at HOR boundaries and in centromere 7. **a**, Density of methylated adenines for the H3K9me3-targeted sample and IgG and free pA-Hia5 controls in 100 kb sliding window across HOR boundaries 1p, 2pq, 6p, 9p, 13q, 14q, 15q, 16p, 17pq, 18pq, 20p, 21q, 22q. **b**, Centromere 7 single molecule browser tracks for H3K9me3-targeted sample, IgG control, and free pA-Hia5. The same molecules are shown in both plots, with mA calls indicated in the first, and mCpG calls indicated in the second. **c**, Coverage tracks in 10-kb bins to accompany mA/A and mCpG/CpG tracks from Figure 3.14d.

We mapped heterochromatin not only in aggregate across HOR array boundaries, but also in single molecules across the centromere. H3K9me3-targeted DiMeLo-seq reads map across the centromere of chromosome 7, even in regions with over 20 kb between unique markers (Figure 3.14d). An IgG isotype control confirmed that adenine methylation in the H3K9me3-targeted sample was not caused by background methylation (Figure 3.14d, Figure 3.15b). Unlike methods which rely on amplifying short DNA fragments, such as ChIP-seq and CUT&RUN, we are able to detect single-molecule heterogeneity in chromatin boundaries, as highlighted in the transition from 65.5 Mb to 68 Mbp, where H3K9me3 signal drops as CpG methylation increases (Figure 3.14d). However, lower methylation efficiency in heterochromatin and the challenges of mapping even moderately long reads in repetitive regions can still lead to uneven and low coverage in these regions (Figure 3.15c). To improve sensitivity for targeted DiMeLo-seq applications in the centromere, we developed a centromere enrichment method to enhance coverage in active HOR arrays and applied this method to study CENP-A.

Restriction-based enrichment strategy improves centromere coverage

Within alpha satellite HOR arrays, the centromere-specific histone variant CENP-A delineates the site where the functional centromere and kinetochore will form. Population-level studies demonstrate that CENP-A nucleosomes are found at the core of these repeat units where the repeats are the most homogeneous.[76,80–82] However, it has not been possible to resolve the positions of CENP-A nucleosomes on single chromatin fibers to determine the one-dimensional organization and density of CENP-A at centromeres. To map the positions of CENP-A nucleosomes at centromeres using DiMeLo-seq, we developed a strategy to enrich specifically for human centromeric DNA in order to avoid sequencing the entire genome.

Our enrichment strategy, called AlphaHOR-RES (alpha higher-order repeat restriction and enrichment by size; from *alfajores*), is based on classic centromere enrichment strategies[83] that involve digesting the genome with restriction enzymes that cut frequently outside centromeric regions but rarely inside them, then removing short DNA fragments (Methods: Centromere enrichment using AlphaHOR-RES, Figure 3.16a). We added AlphaHOR-RES to our DiMeLo-seq workflow and observed at least 20-fold enrichment of sequencing coverage at centromeres while preserving relatively long read lengths (mean ~8 kb; Figure 3.17a,b, Figure 3.16b-d, Methods: Centromere enrichment using AlphaHOR-RES). Thus, this enrichment strategy significantly increases the proportion of molecules sequenced that are useful for investigating CENP-A distribution, saving substantial sequencing time and costs. Furthermore, because AlphaHOR-RES targets the DNA and not the protein in the protein-DNA interaction, and because it is performed after directed methylation is complete, it is unlikely to bias our inferences of protein-DNA interaction frequencies in these regions.

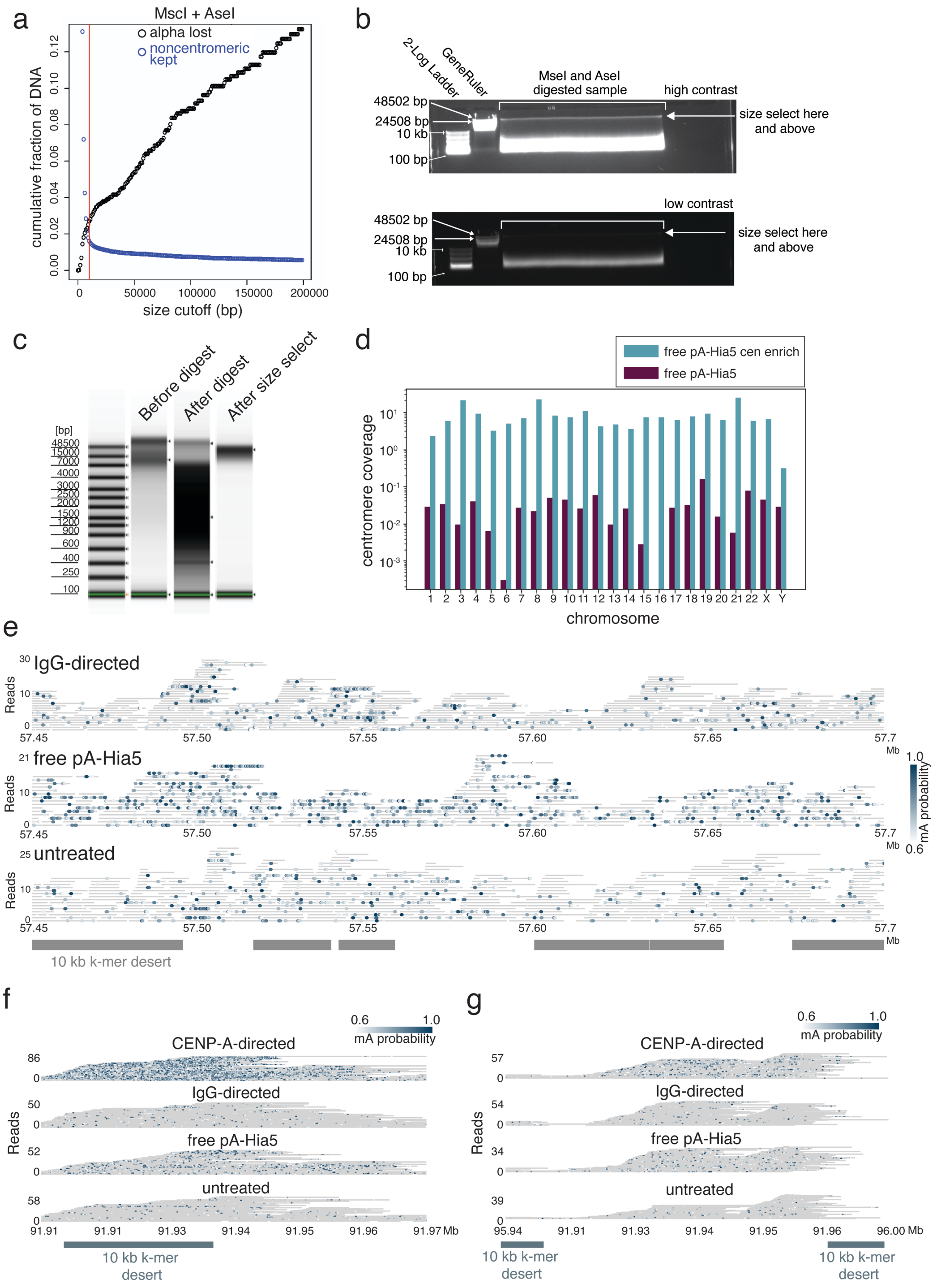


Figure 3.16 AlphaHOR-RES centromere enrichment and methylation within chromosome X and chromosome 3 HORs. **a**, Simulated cumulative distribution of the proportion of alpha-satellite DNA lost (black) and non-centromeric DNA kept (blue) after MscI and AseI digestion of the T2T chm13 genome at different size selection cutoffs. **b**, High (top) and low contrast (bottom) images

of agarose gel run on total genomic DNA after Msc1 and Ase1 digestion. Sample recovered from above cut site (arrow). Representative image of at least 4 replicates. **c**, genomic DNA tapestation gel image of sample before digestion, after digestion, and after size selection. Representative image of at least 3 replicates. **d**, Coverage of the active HOR on each chromosome from the CHM13+HG002X+hg38Y reference genome from free floating pA-Hia5 DiMeLo-seq libraries with and without AlphaHOR-RES. **e-g**, Single molecule view with individual reads in gray and mA depicted as dots for the indicated conditions. Scale bar indicates the probability of adenine methylation (from Guppy) between 0.6 and 1. Regions with at least 10 kb without unique 51 bp k-mers shown in grey to illustrate difficult to map locations for short-read sequencing. e. ChrX CDR (57.45 - 57.7 Mb), f. chromosome 3 HOR between 91.91 and 91.97 Mb, g. chromosome 3 HOR between 95.94 and 96.00 Mb.

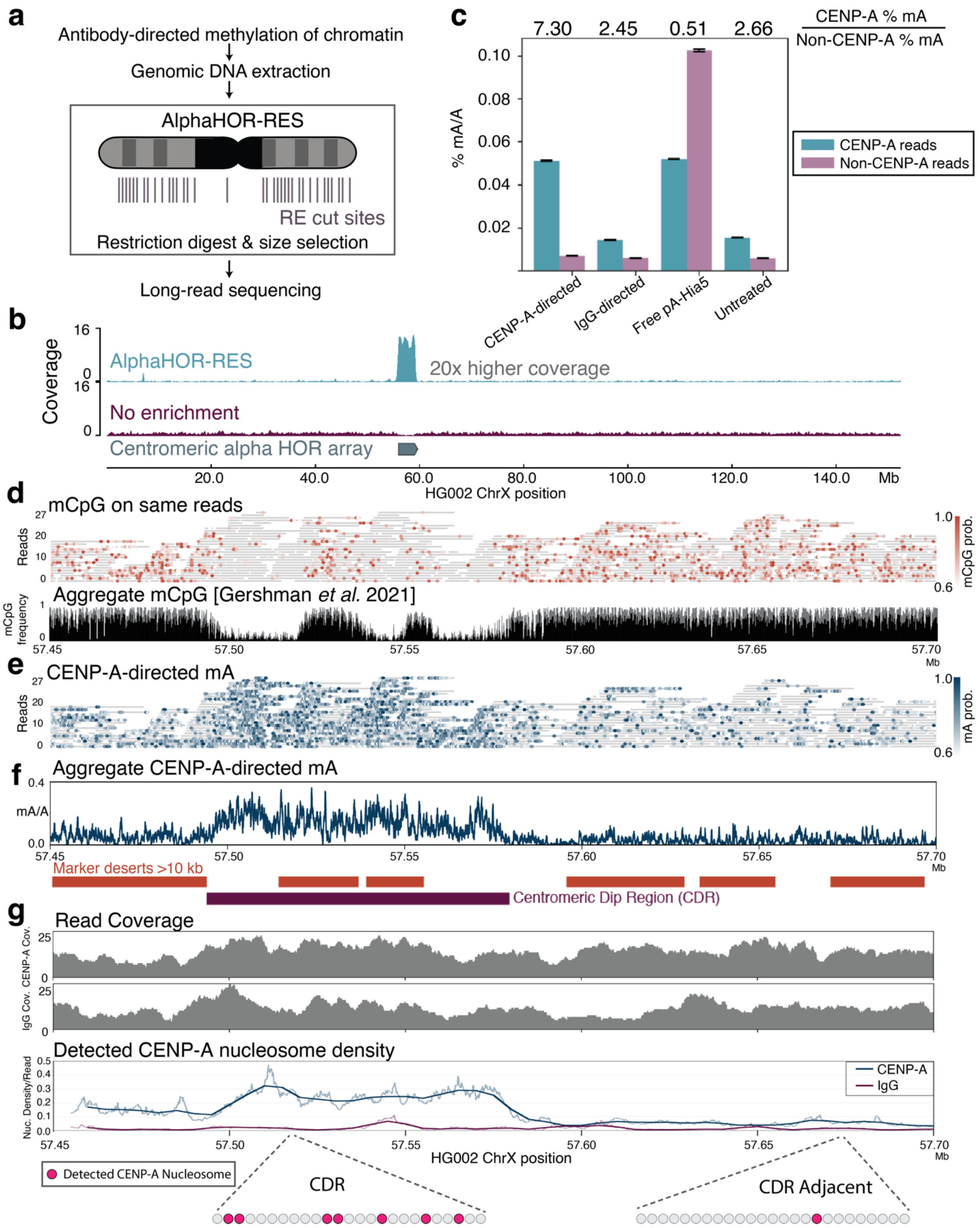


Figure 3.17 CENP-A-directed methylation within chromosome X centromeric higher order repeats. **a**, Schematic of DiMeLo-seq with AlphaHOR-RES centromere enrichment. **b**, Genome browser plot on HG002 chromosome X of read coverage from DiMeLo-seq libraries with

centromere enrichment (top) or without (middle). Bottom track depicts the region of the alpha satellite array. **c**, Barplot of percentage mA/A (using a very stringent Guppy probability threshold of 0.95) for reads from each library that contain or do not contain CENP-A enriched *k*-mers. Fold enrichment of methylation percentage on CENP-A reads over Non-CENP-A reads reported on top. Error bars represent 95% confidence intervals. **d**, View of a 250 kb region spanning the CDR within the active chrX HOR array. (top) Single-molecule view, with individual reads as gray lines and mCpG positions as orange dots shaded by Guppy's methylation probability. (bottom) CpG methylation frequency from nanopore sequencing reported in Gershman et al., 2022.[23] **e**, Single-molecule view of reads in (**d**). mA positions are depicted as blue dots shaded by Guppy's methylation probability. **f**, Aggregate view of mA. mA/A plot indicates the fraction of reads with a Guppy methylation probability above 0.6 at each adenine position (averaged over a 250 bp rolling window for visualization). Marker deserts (regions of at least 10 kb without unique 51 bp *k*-mers) are shown in orange to illustrate difficult-to-map locations for short-read sequencing. **g**, For CENP-A or IgG control DiMeLo-seq, read coverage (top plots) and average fraction of nucleosomes detected as CENP-A (bottom plot) per read in sliding 5 kb windows (step size 1 bp), providing a measure of the density of CENP-A nucleosomes within single DNA molecules across the region. Thick lines indicate a 25 kb rolling average. Cartoon below shows representations of detected CENP-A nucleosomes within a 5 kb region corresponding to the CDR or CDR-adjacent region.

DiMeLo-seq reveals variable CENP-A nucleosome density across centromeres

We performed CENP-A-directed DiMeLo-seq on HG002 cells. After extraction of total genomic DNA, we used AlphaHOR-RES to enrich centromeric sequences before sequencing (Figure 3.17a,b). In an alignment-independent manner,[84] we classified DiMeLo-seq reads based on the presence or absence of CENP-A-enriched *k*-mers from an available short-read sequencing dataset.[82] CENP-A-directed DiMeLo-seq reads with CENP-A enriched *k*-mers had ~7 fold more adenine methylation when compared to reads without CENP-A-enriched *k*-mers (Figure 3.17c). We observed similar absolute methylation levels in DiMeLo-seq reads containing CENP-A *k*-mers when comparing CENP-A-targeted samples to free pA-Hia5 samples. However, the free pA-Hia5 samples also had a higher percentage of mA/A in reads that did not contain CENP-A *k*-mers, indicating a lack of CENP-A specificity in the absence of targeting.

To examine the positions of CENP-A nucleosomes within centromeric repeat arrays, we aligned our reads to a hybrid complete human assembly containing a fully assembled chromosome X from the HG002 cell line (Supplementary Note 14).[23,29] We investigated the recently described chromosome X centromere dip region (CDR), a hypomethylated region in the centromeric alpha HOR array where short-read CENP-A datasets align.[23,76,82,85] We confirmed low endogenous CpG methylation within the CDR as expected (Figure 3.17d). CENP-A-directed mA was found to be higher within both large and small CDRs compared to their adjacent CpG methylated regions, consistent with short-read data for this cell line (Figure 3.17e,f).[23,76] We found that the density of detected CENP-A nucleosomes increased 5-fold within ChrX CDRs compared to neighboring regions (Figure 3.17g). We estimate that 26 ± 5 % of nucleosomes contain CENP-A within the ChrX CDR, whereas only 5 ± 2 % of nucleosomes contain CENP-A within a neighboring region (mean ± standard deviation) (Supplementary Note 14, Figure 3.17g) confirming what ensemble short-read methods cannot: the *density* of CENP-A nucleosomes on single DNA molecules increases in CDRs. IgG isotype controls confirm that this signal is not due to background methylation (2 ± 1 % (mean ± standard deviation) of nucleosomes detected on IgG control reads within ChrX CDR (Figure 3.17g, Figure 3.16e)). A previous study estimated the average CENP-A density across endogenous human centromeres to be 1 in 25 nucleosomes, assuming a mean centromere size of ~1 Mb.[86] In contrast, we estimate that at least 1 in 4 nucleosomes contains CENP-A within the smaller ~100 kb CDR on ChrX. This demonstrates that CENP-A nucleosome occupancy varies considerably across a human centromere, and further we show that the region with the highest CENP-A density coincides with the CDR. The sensitivity of CENP-A DiMeLo-seq on CENP-A chromatin *in vitro* (~65%, Figure 3.4d) suggests that the actual CENP-A density within ChrX CDR could be even higher. We observe a similar distribution of CENP-A-directed methylation on chromosome 3, where only one of the two HOR arrays was observed to have clear CENP-A-directed methylation (Figure 3.16f,g). These observations support the finding of one active HOR array per chromosome.[76,87] These findings illuminate the density and positioning of CENP-A nucleosomes within HOR sequences on individual chromatin fibers, which was not previously attainable with existing techniques.

3.5 Discussion

Here, we have developed, optimized, and validated DiMeLo-seq, a long-read method for mapping protein-DNA interactions genome-wide. DiMeLo-seq can map a protein's binding sites within hundreds of base pairs at multiple loci on single molecules of sequenced DNA up to hundreds of kilobases in length. This long read length improves mappability in highly repetitive regions of the genome, opening them up for future studies of their regulation and function. Because DiMeLo-seq involves no amplification, it can be used to better estimate the absolute protein-DNA interaction frequency at each site in the genome. It also provides joint information about endogenous CpG methylation and protein-DNA interactions on the same long single molecules, which can be phased to reveal haplotype-specific binding and methylation patterns.

By mapping individual CENP-A nucleosomes on long, sequenced DNA molecules, we found that CENP-A nucleosome density increases on single chromatin fibers in mCpG depleted regions within centromeres. The sensitivity of CENP-A DiMeLo-seq on CENP-A chromatin *in vitro* was measured to be ~65%, suggesting that the estimates of CENP-A nucleosome densities within the ChrX CDR are lower limits, and the actual CENP-A density within CDRs could be even higher than ~25% (Figure 3.17g). A source of variation in CENP-A positions is the cell cycle state of chromatin. Because pre-existing CENP-A nucleosomes are thought to epigenetically direct the assembly of new CENP-A nucleosomes in each cell cycle, it will be interesting to understand how CENP-A density varies along the sequence of the active centromere after cell cycle synchronization. We estimated the single-molecule sensitivity of DiMeLo-seq to be between 54-59% for CTCF and LMNB1, at thresholds that achieve 94% specificity compared to off-target regions. However, sensitivity may vary by target protein and antibody, perhaps owing to differences in local steric effects, or to differences in the binding strength of the target protein, antibody, or pA.

This study also allowed us to characterize the benefits and tradeoffs of using DiMeLo-seq compared to short-read ensemble methods. Because DiMeLo-seq is an amplification-free method that sequences single native DNA molecules, and because it relies on centrifugation for washing steps, it requires a relatively large amount of starting material to produce cell pellets big enough to easily handle (1-2 million cells per replicate). Using concanavalin-A coated magnetic beads, which we demonstrated to be compatible with the DiMeLo-seq protocol, may help to reduce these cell input requirements in the future (Supplementary Note 9). Additionally, the standard DiMeLo-seq protocol requires the entire genome to be sequenced uniformly, potentially wasting sequencing reads in regions of the genome that are irrelevant for the target protein's binding domain. For proteins that only target small regions, it is possible to perform targeted DNA sequencing[88,89] or to use DNA enrichment methods like AlphaHOR-RES, the centromere enrichment method we demonstrated here. Another group recently described a complementary approach using a distinct set of restriction enzymes to enrich for centromeric DNA, which may serve as an important alternative to Alpha-HOR-RES.[90] It is also possible to use immunoprecipitation to enrich for methylated DNA or DNA bound to a protein of interest, but this would no longer sample DNA molecules uniformly from the cell population, potentially diminishing the ability to infer protein-DNA interaction frequencies from read methylation frequencies.

Because Hia5 tends to methylate unbound linker DNA, DiMeLo-seq provides information about local nucleosome occupancy along with the target protein's footprint. This also means that highly inaccessible regions can be more difficult to methylate, and they may require higher sequencing coverage. Additionally, because DiMeLo-seq is performed *in situ* in conditions meant to preserve chromatin conformation, it may methylate unbound DNA in *trans* if it is close enough to the target protein's binding sites in 3D space, as does CUT&RUN.[56] These 3D interactions, and the factors that mediate them, can potentially be investigated by perturbing 3D chromatin structure prior to performing DiMeLo-seq, which may also be a useful approach for improving DNA accessibility in highly condensed regions.

We anticipate that DiMeLo-seq will be useful for investigating a wide range of biological questions. For example, because it can allow one to explore the density of a protein's binding along a single chromatin fiber from a single cell, it can be used to investigate how the exact boundaries between chromatin states vary among single cells, or perhaps how the stoichiometry of a DNA-binding protein in enhancers affects the transcription of nearby genes. We also demonstrated that DiMeLo-seq can read out methyladenines deposited by *in vivo* expression of protein-MTase fusions, as in conventional DamID[7] or MadID,[65] instead of antibody targeting *in situ*. This may prove useful for investigating more transient protein-DNA interactions, or proteins that lack suitable antibodies, in cases where the biological system being studied can be readily genetically modified. One can also imagine adding exogenous cytosine methylation marks to provide joint information about DNA accessibility or about a second protein's binding profile. Although we primarily used Oxford Nanopore Technologies sequencing in this study, we also demonstrated that DiMeLo-seq is compatible with Pacific Biosciences HiFi sequencing, which may be preferred for applications that require highly accurate base calls, such as genome assembly. With this study, we show that DiMeLo-seq provides a versatile approach for characterizing protein-DNA interactions on individual molecules spanning difficult-to-interrogate genomic regions.

3.6 Acknowledgements

We thank Andrew Stergachis for the plasmid encoding Hia5, Gina Caldas for experimental training, and Gary Karpen for helpful discussions. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. We would like to thank Michelle Tan for her contributions to sequencing. This work was supported by the Chan Zuckerberg Biohub and by the NIGMS of the National Institutes of Health under award number R35GM124916 to AS and R01 GM074728 to AFS. KHM is supported by R21HG010548-01. OKS and RRB are supported by an NIH T32 award, numbers GM113854-02 and GM007279-45 respectively. AM, OKS, and RRB are supported by NSF GRFP awards. NA is an HHMI Hanna H. Gray Fellow. AS is a Chan Zuckerberg Biohub Investigator, and a Pew Scholar in the Biomedical Sciences.

3.7 Author Contributions Statement

NA, AM, OKS, KS, AFS, and AS designed the study. NA, AM, OKS, KS, and RRB performed the experiments. AD and NN assisted with sequencing. KHM provided unpublished datasets and feedback. RM assisted with analysis software development. NA, AM, OKS, and KS analyzed and interpreted the data. NA, AM, OKS, KS, and RRB made the figures. NA, AM, OKS, and KS wrote the manuscript, with input from RRB, AFS, and AS. AFS and AS supervised the study.

3.8 Competing Interests Statement

NA, AM, OKS, KS, AFS, and AS are co-inventors on a patent application related to this work.

3.9 Methods

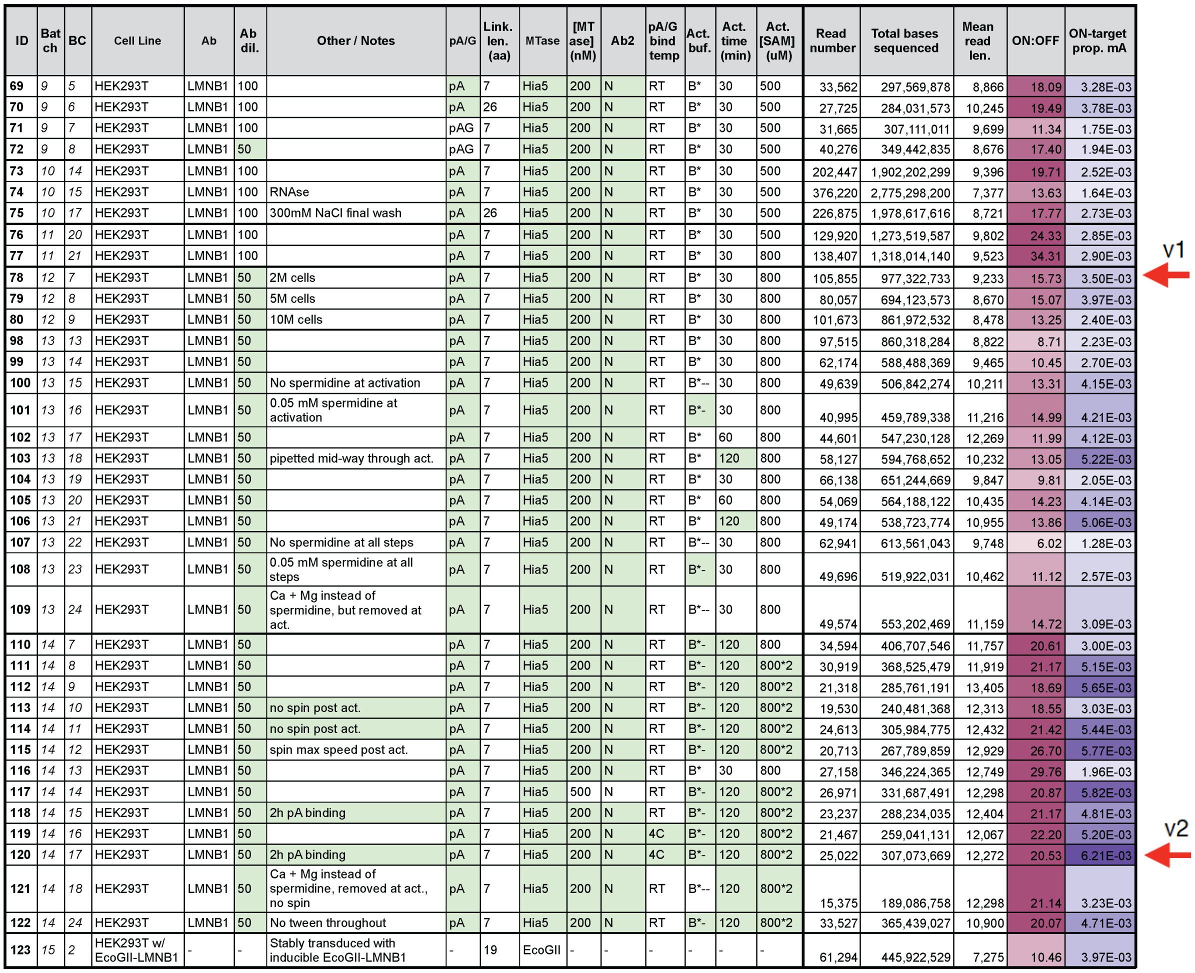
Protocols/Materials availability
For detailed and updated protocols, please refer to the following protocols.io web pages:

DiMeLo-seq v1: dx.doi.org/10.17504/protocols.io.bv8tn9wn
DiMeLo-seq v2: dx.doi.org/10.17504/protocols.io.b2u8qezw
pA-Hia5 Protein Purification: dx.doi.org/10.17504/protocols.io.bv82n9ye
AlphaHOR-RES: dx.doi.org/10.17504/protocols.io.bv9vn966

Plasmids are available on Addgene: pA-Hia5 expression plasmid (pET-PA-Hia5, Addgene #174372) and pAG-Hia5 expression plasmid (pET-pAG-Hia5, Addgene #174373).

Sample summary metrics
Sequencing summary metrics for samples included in this study can be found in Table 3.1, Table 3.2, Table 3.3, and Figure 3.18.

ID	Batch	BC	Cell Line	Ab	Ab dil.	Other / Notes	pA/G	Link. len. (aa)	MTase	[MTase] (nM)	Ab2	pA/G bind temp	Act. buf.	Act. time (min)	Act. [SAM] (uM)	Read number	Total bases sequenced	Mean read len.	ON:OFF	ON-target prop. mA
1	*1*	*6*	HEK293T	LMNB1	500	SRE XL	pAG	29	EcoGII	50	N	4C	A	30	500	86,481	1,976,058,348	22,850	1.67	4.46E-05
2	*2*	*9*	HEK293T	LMNB1	500	SRE XL, Ab2 bind at 4C	pAG	29	EcoGII	150	Y	4C	A	30	500	38,539	847,065,308	21,979	2.93	1.57E-04
3	*2*	*10*	HEK293T	LMNB1	500	SRE XL	pAG	29	EcoGII	150	N	4C	A	30	500	36,921	856,905,359	23,209	2.50	1.37E-04
4	*2*	*14*	HEK293T	LMNB1	500	NP40 0.5%,SRE XL,Ab2@4C	pAG	29	EcoGII	150	Y	4C	A	30	500	60,602	1,490,024,882	24,587	1.21	1.18E-04
5	*2*	*15*	HEK293T	LMNB1	500	NP40 0.5%, SRE XL	pAG	29	EcoGII	150	N	4C	A	30	500	38,772	943,586,111	24,337	1.35	7.23E-05
6	*3*	*16*	HEK293T	LMNB1	500		pAG	29	EcoGII	527	N	RT	A	30	500	162,252	1,752,177,765	10,799	1.38	5.32E-05
7	*3*	*17*	HEK293T	LMNB1	500		pAG	29	EcoGII	527	Y	RT	A	30	500	198,264	2,102,108,611	10,603	3.03	9.52E-05
8	*3*	*18*	HEK293T	LMNB1	500		pAG	29	EcoGII	150	N	RT	A	30	500	106,892	1,227,114,851	11,480	2.05	2.99E-05
9	*3*	*19*	HEK293T	LMNB1	100		pAG	29	EcoGII	527	N	RT	A	30	500	238,437	2,415,676,861	10,131	4.80	1.74E-04
10	*4*	*20*	HEK293T	LMNB1	100		pAG	29	EcoGII	527	Y	4C	A	30	500	121,419	1,261,478,986	10,389	4.73	4.01E-04
11	*4*	*21*	HEK293T	LMNB1	100		pAG	29	EcoGII	527	N	4C	A	30	500	123,908	1,276,684,561	10,303	3.98	2.99E-04
12	*4*	*23*	HEK293T	LMNB1	100		pAG	29	EcoGII	527	N	4C	A	60	500	112,198	1,162,823,359	10,364	3.73	3.23E-04
13	*4*	*24*	HEK293T	LMNB1	100	SAM replenished	pAG	29	EcoGII	527	N	4C	A	60	500*2	119,153	1,207,239,004	10,132	3.05	3.13E-04
14	*4*	*1*	HEK293T	LMNB1	100	SAM replenished	pAG	29	EcoGII	527	N	4C	A	60	500*2	96,385	1,076,704,141	11,171	2.65	2.87E-04
15	*5*	*4*	HEK293T	LMNB1	100		pAG	29	EcoGII	527	Y	RT	A	30	500	80,135	839,079,448	10,471	3.56	4.22E-04
16	*5*	*5*	HEK293T	LMNB1	100		pAG	29	EcoGII	527	Y	RT	A	15	500	101,911	1,067,611,582	10,476	3.87	3.25E-04
17	*5*	*6*	HEK293T	LMNB1	100	30C act	pAG	29	EcoGII	527	Y	RT	A	30	500	68,779	651,361,202	9,470	7.54	4.95E-04
18	*5*	*7*	HEK293T	LMNB1	100		pAG	29	EcoGII	527	Y	RT	A	30	500	83,487	902,381,502	10,809	4.85	4.40E-04
19	*5*	*8*	HEK293T	LMNB1	100		pA	7	Hia5	527	Y	RT	A	30	500	26,351	227,185,814	8,622	5.64	7.04E-05
20	*5*	*9*	HEK293T	LMNB1	100		pA	26	Hia5	527	Y	RT	A	30	500	19,112	169,551,133	8,871	7.42	8.39E-05
21	*5*	*10*	HEK293T	LMNB1	100		pA	mix	Hia5	527	Y	RT	A	30	500	15,778	151,798,963	9,621	9.84	7.88E-05
22	*5*	*11*	HEK293T	LMNB1	100	pAG-EcoGII+pA-Hia5-both	mix	mix	Both	527	Y	RT	A	30	500	15,975	149,309,499	9,346	11.82	2.65E-04
23	*5*	*12*	HEK293T	LMNB1	100	NP40 0.1%	pAG	29	EcoGII	527	Y	RT	A	30	500	86,033	860,416,335	10,001	2.48	2.50E-04
24	*5*	*14*	HEK293T	LMNB1	100		pAG	29	EcoGII	527	N	RT	A	30	500	101,662	993,731,709	9,775	2.51	2.97E-04
25	*5*	*15*	HEK293T	LMNB1	100		pA	7	Hia5	527	N	RT	A	30	500	106,830	985,205,512	9,222	12.43	7.89E-04
26	*5*	*16*	HEK293T	LMNB1	100		pA	26	Hia5	527	N	RT	A	30	500	74,873	793,137,043	10,593	11.72	1.06E-03
27	*5*	*17*	HEK293T	LMNB1	100		pA	mix	Hia5	527	N	RT	A	30	500	93,269	988,096,097	10,594	17.68	8.36E-04
28	*5*	*18*	HEK293T	LMNB1	100	pAG-EcoGII+pA-Hia5-both	mix	mix	Both	527	N	RT	A	30	500	92,725	945,637,416	10,198	6.09	7.71E-04
29	*5*	*19*	HEK293T	LMNB1	100	**light fixation**	pAG	29	EcoGII	527	Y	RT	A	30	500	91,520	686,940,638	7,506	5.37	4.31E-04
30	*6*	*21*	HEK293T	LMNB1	100	Triton 0.1%	pA	26	Hia5	527	N	RT	A	30	500	100,177	1,126,513,813	11,245	4.66	9.10E-05
31	*6*	*22*	HEK293T	LMNB1	100	Triton 0.1%	pA	26	Hia5	527	N	RT	B	30	500	114,462	1,105,844,974	9,661	21.18	3.68E-04
32	*6*	*23*	HEK293T	LMNB1	100	NP40 0.1%	pA	26	Hia5	527	N	RT	A	30	500	105,547	934,810,760	8,857	3.68	6.12E-05
33	*6*	*24*	HEK293T	LMNB1	100	NP40 0.1%	pAG	29	EcoGII	527	N	RT	A	30	500	84,310	868,314,441	10,299	4.31	1.11E-04
34	*6*	*1*	HEK293T	LMNB1	100		pAG	29	EcoGII	527	N	RT	A	30	500	91,978	1,189,823,234	12,936	3.41	1.49E-04
35	*6*	*2*	HEK293T	LMNB1	100		pA	26	Hia5	527	N	RT	A	30	500	105,110	1,264,716,882	12,032	24.21	5.33E-04
36	*6*	*3*	HEK293T	LMNB1	100		pA	26	Hia5	527	N	RT	B	30	500	103,978	1,311,352,305	12,612	38.05	2.06E-03
37	*7*	*4*	HEK293T	LMNB1	100		pA	26	Hia5	527	N	RT	B*	30	500	90,785	752,583,312	8,290	26.47	4.08E-03
38	*7*	*5*	HEK293T	LMNB1	100		pAG	7	Hia5	527	N	RT	B*	30	500	48,122	459,680,632	9,552	29.83	2.23E-03
39	*7*	*6*	HEK293T	LMNB1	100		pAG	29	EcoGII	527	N	RT	B*	30	500	81,695	727,342,635	8,903	7.48	1.21E-03
40	*7*	*7*	HEK293T	LMNB1	100		pA	26	Hia5	527	N	RT	B	30	500	65,117	583,081,677	8,954	23.49	3.05E-03
41	*7*	*8*	HEK293T	LMNB1	100		pAG	7	Hia5	527	N	RT	B	30	500	56,317	540,388,209	9,595	29.75	2.08E-03
42	*7*	*9*	HEK293T	LMNB1	100		pAG	29	EcoGII	527	N	RT	B	30	500	47,684	449,643,915	9,430	8.22	1.15E-03
43	*7*	*10*	HEK293T	LMNB1	100	30C act	pA	26	Hia5	527	N	RT	B	30	500	58,489	523,464,949	8,950	29.98	2.24E-03
44	*7*	*11*	HEK293T	LMNB1	100	30C act	pAG	7	Hia5	527	N	RT	B	30	500	57,609	511,675,484	8,882	17.42	1.26E-03
45	*7*	*12*	HEK293T	LMNB1	100	30C act	pAG	29	EcoGII	527	N	RT	B	30	500	61,562	551,698,966	8,962	4.40	5.33E-04
46	*7*	*13*	HEK293T	LMNB1	100	30C act	pAG	29	EcoGII	527	N	RT	A	30	500	71,900	628,811,974	8,746	8.18	3.02E-04
47	*7*	*14*	HEK293T	LMNB1	100	Ab2 (GP)	pA	26	Hia5	527	Y	RT	B	30	500	99,602	843,096,447	8,465	21.37	2.33E-03
48	*7*	*15*	HEK293T	LMNB1	100	Ab2 (GP)	pA	26	Hia5	527	Y	RT	A*	30	500	54,677	484,736,060	8,865	8.61	4.85E-04
49	*7*	*16*	HEK293T	LMNB1	100	Ab2 (GP)	pAG	7	Hia5	527	Y	RT	B	30	500	63,400	571,575,224	9,015	13.88	1.43E-03
50	*7*	*17*	HEK293T	LMNB1	100	Ab2 (GP), 30C act	pAG	29	EcoGII	527	Y	RT	A	30	500	60,290	517,001,320	8,575	7.27	3.43E-04
51	*7*	*18*	HEK293T	LMNB1	100	pA-Hia5 short+long	pA	mix	Hia5	527	N	RT	B	30	500	47,923	468,996,420	9,786	21.99	3.19E-03
52	*7*	*19*	HEK293T	LMNB1	100	pA-Hia5 200 nM, Ab2 (GP)	pA	26	Hia5	200	Y	RT	B	30	500	52,514	513,948,158	9,787	19.59	2.82E-03
53	*7*	*20*	HEK293T	LMNB1	100	pA-Hia5 50 nM, Ab2 (GP)	pA	26	Hia5	50	Y	RT	B	30	500	48,723	441,526,404	9,062	20.30	2.55E-03
54	*7*	*22*	HEK293T	LMNB1	100	**frozen**	pA	26	Hia5	527	N	RT	B	30	500	71,276	488,718,798	6,857	28.06	2.98E-03
55	*7*	*23*	HEK293T	LMNB1	100	Ab2 (Goat), 30C act	pAG	29	EcoGII	527	Y	RT	A	30	500	61,444	554,745,469	9,028	8.71	5.32E-04
56	*8*	*1*	HEK293T	LMNB1	100	poor batch	pA	7	Hia5	200	N	RT	B*	30	500	45,362	345,941,926	7,626	7.68	2.36E-03
57	*8*	*2*	HEK293T	LMNB1	100	poor batch	pA	7	Hia5	200	N	RT	B*	30	500	40,187	353,173,477	8,788	6.86	2.38E-03
58	*8*	*3*	HEK293T	LMNB1	50	poor batch	pA	7	Hia5	200	N	RT	B*	30	500	33,387	278,128,346	8,330	9.62	3.97E-03
59	*8*	*7*	GM12878	LMNB1	100	poor batch	pA	7	Hia5	200	N	RT	B*	30	500	55,380	462,838,466	8,358	8.39	2.75E-03
60	*8*	*8*	HG002	LMNB1	100	poor batch	pA	7	Hia5	200	N	RT	B*	30	500	59,508	439,509,594	7,386	9.34	3.13E-03
61	*8*	*9*	Hap1	LMNB1	100	poor batch	pA	7	Hia5	200	N	RT	B*	30	500	59,648	424,749,575	7,121	7.47	3.43E-03
62	*8*	*15*	HEK293T	LMNB1	100	**light fixation**, poor batch	pA	7	Hia5	200	N	RT	B*	30	500	52,734	315,838,720	5,989	9.74	3.58E-03
63	*8*	*17*	Hap1	LMNB1	100	primary at RT, poor batch	pA	7	Hia5	200	N	RT	B*	30	500	54,120	354,171,388	6,544	8.12	3.29E-03
64	*8*	*19*	HEK293T	LMNB1	100	**conA beads**, poor batch	pA	7	Hia5	200	N	RT	B*	30	500	61,614	494,542,424	8,026	8.56	3.29E-03
65	*8*	*20*	GM12878	LMNB1	100	**conA beads**, poor batch	pA	7	Hia5	200	N	RT	B*	30	500	64,420	461,182,886	7,159	5.15	3.08E-03
66	*8*	*24*	Hap1	LMNB1	100	**conA beads**, poor batch	pA	7	Hia5	200	N	RT	B*	30	500	54,354	404,350,726	7,439	4.45	3.01E-03
67	*9*	*1*	HEK293T	LMNB1	100		pA	7	Hia5	200	N	RT	B*	30	500	45,503	360,723,378	7,927	10.65	1.84E-03
68	*9*	*4*	HEK293T	LMNB1	100		pA	26	Hia5	527	N	RT	B*	30	500	62,799	407,120,248	6,483	16.01	3.81E-03



ID	Batch	BC	Cell Line	Ab	Ab dil.	Other / Notes	pA/G	Link. len. (aa)	MTase	[MTase] (nM)	Ab2	pA/G bind temp	Act. buf.	Act. time (min)	Act. [SAM] (uM)	Read number	Total bases sequenced	Mean read len.	ON:OFF	ON-target prop. mA
69	9	5	HEK293T	LMNB1	100		pA	7	Hia5	200	N	RT	B*	30	500	33,562	297,569,878	8,866	18.09	3.28E-03
70	9	6	HEK293T	LMNB1	100		pA	26	Hia5	200	N	RT	B*	30	500	27,725	284,031,573	10,245	19.49	3.78E-03
71	9	7	HEK293T	LMNB1	100		pAG	7	Hia5	200	N	RT	B*	30	500	31,665	307,111,011	9,699	11.34	1.75E-03
72	9	8	HEK293T	LMNB1	50		pAG	7	Hia5	200	N	RT	B*	30	500	40,276	349,442,835	8,676	17.40	1.94E-03
73	10	14	HEK293T	LMNB1	100		pA	7	Hia5	200	N	RT	B*	30	500	202,447	1,902,202,299	9,396	19.71	2.52E-03
74	10	15	HEK293T	LMNB1	100	RNAse	pA	7	Hia5	200	N	RT	B*	30	500	376,220	2,775,298,200	7,377	13.63	1.64E-03
75	10	17	HEK293T	LMNB1	100	300mM NaCl final wash	pA	26	Hia5	200	N	RT	B*	30	500	226,875	1,978,617,616	8,721	17.77	2.73E-03
76	11	20	HEK293T	LMNB1	100		pA	7	Hia5	200	N	RT	B*	30	500	129,920	1,273,519,587	9,802	24.33	2.85E-03
77	11	21	HEK293T	LMNB1	100		pA	7	Hia5	200	N	RT	B*	30	800	138,407	1,318,014,140	9,523	34.31	2.90E-03
78	12	7	HEK293T	LMNB1	50	2M cells	pA	7	Hia5	200	N	RT	B*	30	800	105,855	977,322,733	9,233	15.73	3.50E-03
79	12	8	HEK293T	LMNB1	50	5M cells	pA	7	Hia5	200	N	RT	B*	30	800	80,057	694,123,573	8,670	15.07	3.97E-03
80	12	9	HEK293T	LMNB1	50	10M cells	pA	7	Hia5	200	N	RT	B*	30	800	101,673	861,972,532	8,478	13.25	2.40E-03
98	13	13	HEK293T	LMNB1	50		pA	7	Hia5	200	N	RT	B*	30	800	97,515	860,318,284	8,822	8.71	2.23E-03
99	13	14	HEK293T	LMNB1	50		pA	7	Hia5	200	N	RT	B*	30	800	62,174	588,488,369	9,465	10.45	2.70E-03
100	13	15	HEK293T	LMNB1	50	No spermidine at activation	pA	7	Hia5	200	N	RT	B*--	30	800	49,639	506,842,274	10,211	13.31	4.15E-03
101	13	16	HEK293T	LMNB1	50	0.05 mM spermidine at activation	pA	7	Hia5	200	N	RT	B*-	30	800	40,995	459,789,338	11,216	14.99	4.21E-03
102	13	17	HEK293T	LMNB1	50		pA	7	Hia5	200	N	RT	B*	60	800	44,601	547,230,128	12,269	11.99	4.12E-03
103	13	18	HEK293T	LMNB1	50	pipetted mid-way through act.	pA	7	Hia5	200	N	RT	B*	120	800	58,127	594,768,652	10,232	13.05	5.22E-03
104	13	19	HEK293T	LMNB1	50		pA	7	Hia5	200	N	RT	B*	30	800	66,138	651,244,669	9,847	9.81	2.05E-03
105	13	20	HEK293T	LMNB1	50		pA	7	Hia5	200	N	RT	B*	60	800	54,069	564,188,122	10,435	14.23	4.14E-03
106	13	21	HEK293T	LMNB1	50		pA	7	Hia5	200	N	RT	B*	120	800	49,174	538,723,774	10,955	13.86	5.06E-03
107	13	22	HEK293T	LMNB1	50	No spermidine at all steps	pA	7	Hia5	200	N	RT	B*--	30	800	62,941	613,561,043	9,748	6.02	1.28E-03
108	13	23	HEK293T	LMNB1	50	0.05 mM spermidine at all steps	pA	7	Hia5	200	N	RT	B*-	30	800	49,696	519,922,031	10,462	11.12	2.57E-03
109	13	24	HEK293T	LMNB1	50	Ca + Mg instead of spermidine, but removed at act.	pA	7	Hia5	200	N	RT	B*--	30	800	49,574	553,202,469	11,159	14.72	3.09E-03
110	14	7	HEK293T	LMNB1	50		pA	7	Hia5	200	N	RT	B*-	120	800	34,594	406,707,546	11,757	20.61	3.00E-03
111	14	8	HEK293T	LMNB1	50		pA	7	Hia5	200	N	RT	B*-	120	800*2	30,919	368,525,479	11,919	21.17	5.15E-03
112	14	9	HEK293T	LMNB1	50		pA	7	Hia5	200	N	RT	B*-	120	800*2	21,318	285,761,191	13,405	18.69	5.65E-03
113	14	10	HEK293T	LMNB1	50	no spin post act.	pA	7	Hia5	200	N	RT	B*-	120	800*2	19,530	240,481,368	12,313	18.55	3.03E-03
114	14	11	HEK293T	LMNB1	50	no spin post act.	pA	7	Hia5	200	N	RT	B*-	120	800*2	24,613	305,984,775	12,432	21.42	5.44E-03
115	14	12	HEK293T	LMNB1	50	spin max speed post act.	pA	7	Hia5	200	N	RT	B*-	120	800*2	20,713	267,789,859	12,929	26.70	5.77E-03
116	14	13	HEK293T	LMNB1	50		pA	7	Hia5	200	N	RT	B*	30	800	27,158	346,224,365	12,749	29.76	1.96E-03
117	14	14	HEK293T	LMNB1	50		pA	7	Hia5	500	N	RT	B*-	120	800*2	26,971	331,687,491	12,298	20.87	5.82E-03
118	14	15	HEK293T	LMNB1	50	2h pA binding	pA	7	Hia5	200	N	RT	B*-	120	800*2	23,237	288,234,035	12,404	21.17	4.81E-03
119	14	16	HEK293T	LMNB1	50		pA	7	Hia5	200	N	4C	B*-	120	800*2	21,467	259,041,131	12,067	22.20	5.20E-03
120	14	17	HEK293T	LMNB1	50	2h pA binding	pA	7	Hia5	200	N	4C	B*-	120	800*2	25,022	307,073,669	12,272	20.53	6.21E-03
121	14	18	HEK293T	LMNB1	50	Ca + Mg instead of spermidine, removed at act., no spin	pA	7	Hia5	200	N	RT	B*--	120	800*2	15,375	189,086,758	12,298	21.14	3.23E-03
122	14	24	HEK293T	LMNB1	50	No tween throughout	pA	7	Hia5	200	N	RT	B*-	120	800*2	33,527	365,439,027	10,900	20.07	4.71E-03
123	15	2	HEK293T w/ EcoGII-LMNB1	-	-	Stably transduced with inducible EcoGII-LMNB1	-	19	EcoGII	-	-	-	-	-	-	61,294	445,922,529	7,275	10.46	3.97E-03

Table 3.1 All LMNB1-directed conditions tested. Conditions are specified on the left side of the table, and outputs are summarized on the right hand side. For each protocol parameter (columns), the option that was selected for the final protocol is highlighted in green. ON:OFF represents the ratio of the proportion of adenines (q>=10) methylated (p>=0.9) in cLADs (on-target regions) to the proportion of adenines methylated in ciLADs (off-target regions), and cells are colored by the magnitude of this ratio (white=low, magenta=high). The neighboring column is shaded white to purple to correspond to the proportion of adenines methylated in cLADs. Each batch was run on a separate day. A list of abbreviations follows. Ab: Primary Antibody, Ab2: Secondary Antibody (either Goat if not specified or GP for guinea pig), Ab. Dil.: the dilution factor of the antibody (50 = 1:50), BC: barcode, RT: Room Temperature, Buffer A: same as wash buffer in final protocol, Buffer A*: same as A but with 75 mM NaCl, Buffer B: activation buffer (15 mM Tris, pH 8.0, 15 mM NaCl, 60 mM KCl, 1 mM EDTA, pH 8.0, 0.5 mM EGTA, pH 8.0, 0.5 mM Spermidine, 800 µM SAM) without BSA, Buffer B*: Buffer B with 0.1% BSA, Buffer B*-: Buffer B* with 0.05 mM spermidine, Buffer B*--: Buffer B* with no spermidine, conA: concanavalin A beads, MTase: methyltransferase, RNAse: the nuclei were treated with RNAse prior to antibody binding, SAM: S-adenosylmethionine (methyl donor), SRE XL: the Circulomics Short Read Eliminator XL kit

was used to select longer fragments prior to sequencing. The red arrows indicate instances of the v1 and v2 protocols (conditions 78 and 120).

ID	Batch	BC	Cell Line	Ab	Ab dil.	Other / Notes	pA/G	Link. len. (aa)	MTase	[MTase] (nM)	Ab2	pA/G bind temp	Act. buf.	Act. time (min)	Act. [SAM] (uM)	Read number	Total bases sequenced	Mean read len.	ON:OFF	ON-target prop. mA
81	*4*	*22*	HEK293T	LMNB1	100	**no SAM**	pAG	29	EcoGII	527	N	4C	A	30	**0**	144,251	1,297,558,669	8,995	1.22	3.83E-05
82	*8*	*14*	HEK293T	LMNB1	100	**no SAM**, poor batch	pA	7	Hia5	200	N	RT	B*	30	**0**	37,978	301,486,227	7,938	2.58	5.13E-04
83	*1*	*7*	HEK293T	**IgG**	500	SRE XL	pAG	29	EcoGII	50	N	4C	A	30	500	39,167	872,653,096	22,280	1.11	2.22E-05
84	*2*	*13*	HEK293T	**IgG**	500	SRE XL	pAG	29	EcoGII	150	N	4C	A	30	500	66,878	1,500,527,829	22,437	0.98	4.86E-05
85	*4*	*2*	HEK293T	**IgG**	100		pAG	29	EcoGII	527	Y	4C	A	30	500	165,186	1,608,748,957	9,739	0.82	6.05E-05
86	*4*	*3*	HEK293T	**IgG**	100		pAG	29	EcoGII	527	N	4C	A	30	500	95,567	982,777,929	10,284	0.84	7.71E-05
87	*8*	*10*	GM12878	**IgG**	100		pA	7	Hia5	200	N	RT	B*	30	500	35,649	304,927,237	8,554	1.77	4.57E-04
88	*8*	*11*	HG002	**IgG**	100		pA	7	Hia5	200	N	RT	B*	30	500	58,023	418,761,436	7,217	1.45	4.44E-04
89	*8*	*12*	Hap1	**IgG**	100		pA	7	Hia5	200	N	RT	B*	30	500	73,437	453,120,139	6,170	1.86	4.77E-04
90	*8*	*13*	HEK293T	**IgG**	100		pA	7	Hia5	200	N	RT	B*	30	500	50,184	395,339,186	7,878	1.56	4.15E-04
91	*12*	*16*	GM12878	**IgG**	50		pA	7	Hia5	200	N	RT	B*	30	800	169,811	1,344,179,912	7,916	1.47	1.50E-04
92	*5*	*13*	HEK293T	-	-	**free floating EcoGII**	pAG	29	EcoGII	527	-	-	A	30	500	102,650	983,478,388	9,581	0.57	2.80E-04
93	*5*	*20*	HEK293T	-	-	**light fix., free floating EcoGII**	pAG	29	EcoGII	527	-	-	A	30	500	118,802	260,574,348	2,193	0.57	8.96E-05
94	*8*	*16*	HEK293T	-	-	**free floating Hia5**	-	-	Hia5	200	-		B*	30	500	49,149	363,924,351	7,405	1.10	7.55E-03
95	*10*	*21*	HEK293T	-	-	**free floating Hia5**	-	-	Hia5	200	-		B*	30	500	224,739	1,954,375,136	8,696	1.17	6.34E-03
96	*10*	*22*	HEK293T	-	-	**free floating Hia5** w/ RNAse	-	-	Hia5	200	-		B*	30	500	192,453	1,755,187,759	9,120	1.23	6.96E-03
97	*12*	*17*	GM12878	-	-	**free floating Hia5**	-	-	Hia5	200	-		B*	30	800	94,126	831,589,618	8,835	1.12	7.98E-03

Table 3.2 Control conditions tested. Same as Table 3.1 but for 3 different negative control conditions (not expected to show methylation, or not expected to show enrichment in cLADs): no SAM added at activation/methylation step, nonspecific IgG isotype control antibody used, or free-floating enzyme was added at the activation/methylation step to methylate all accessible DNA.

Sample number	Target / description	Cell line	Reads	Bases	Mean read length	Protocol version
1	CTCF	GM12878	729097	8824849107	12104	v1
2	CTCF	GM12878	209699	2864997263	13662	v1
3	CTCF	GM12878	215664	7052382103	32701	v1
4	CTCF	GM12878	161144	1961203749	12170	v1
5	CTCF	GM12878	194417	2202030238	11326	v2 - opt1
6	CTCF	GM12878	231715	2395286461	10337	v2 - opt2
7	CTCF	GM12878	201517	1719224973	8531	v2 - opt3
8	CTCF	GM12878	191839	2218327728	11563	v2
9	CTCF	GM12878	1781121	23009057167	12918	v2
10	CTCF	GM12878	1820026	22882150932	12572	v2
11	free pA-Hia5	GM12878	970282	10371517658	10689	v1
12	IgG	GM12878	1421526	11531597797	8112	v1
13	H3K9me3	HG002	896511	8057506035	8988	v1
14	H3K9me3	HG002	155656	1847831908	11871	v1
15	H3K9me3	HG002	233920	5798858000	24790	v1
16	free pA-Hia5	HG002	71713	1433794520	19994	v1
17	free pA-Hia5	HG002	204504	3771902818	18444	v1
18	free pA-Hia5	HG002	64758	1219166120	18826	v1
19	free pA-Hia5	HG002	145380	2878390589	19799	v1
20	IgG	HG002	306270	7857066233	25654	v1
21	IgG	HG002	132045	2685488328	20338	v1
22	in vitro methylated genomic DNA	GM12878	330573	2785142705	8425	v1
23	unmethylated genomic DNA	GM12878	437135	4040533340	9243	v1
24	CENP-A	HG002	258948	2278497590	8799	v1
25	CENP-A	HG002	325860	2490299385	7642	v1
26	CENP-A	HG002	65205	454714846	6974	v1
27	CENP-A	HG002	320201	6719916952	20986	v1
28	CENP-A	HG002	1669667	5013914326	3003	v1
29	free pA-Hia5	HG002	230216	2018755168	8769	v1
30	free pA-Hia5	HG002	114364	848652995	7421	v1
31	free pA-Hia5	HG002	257095	1740051056	6768	v1
32	IgG	HG002	252694	2349991082	9300	v1
33	IgG	HG002	235264	2068374069	8792	v1
34	IgG	HG002	84856	624068929	7354	v1
35	untreated	HG002	260016	2167665966	8337	v1
36	untreated	HG002	436472	3262286345	7474	v1
37	untreated	HG002	86613	573212999	6618	v1

Table 3.3 Sequencing summary metrics. The number of reads, bases, and mean read length are indicated for CTCF-, H3K9me3-, and CENP-A-directed DiMeLo-seq, along with accompanying controls. Protocol version indicates whether the standard protocol (v1) or the protocol for optimized methylation efficiency (v2) was used (Methods: DiMeLo-seq). For CTCF samples used in optimization for v2 protocol development, the optimization conditions are: opt1: 2 hour activation, 0.05 mM spermidine at activation, replenish SAM; opt2: 2 hour activation, 0.05 mM spermidine at activation, replenish SAM, 500 nM pA-Hia5; opt3: 2 hour activation, no spermidine, 1 mM Ca++ and 0.5 mM Mg++ buffer.

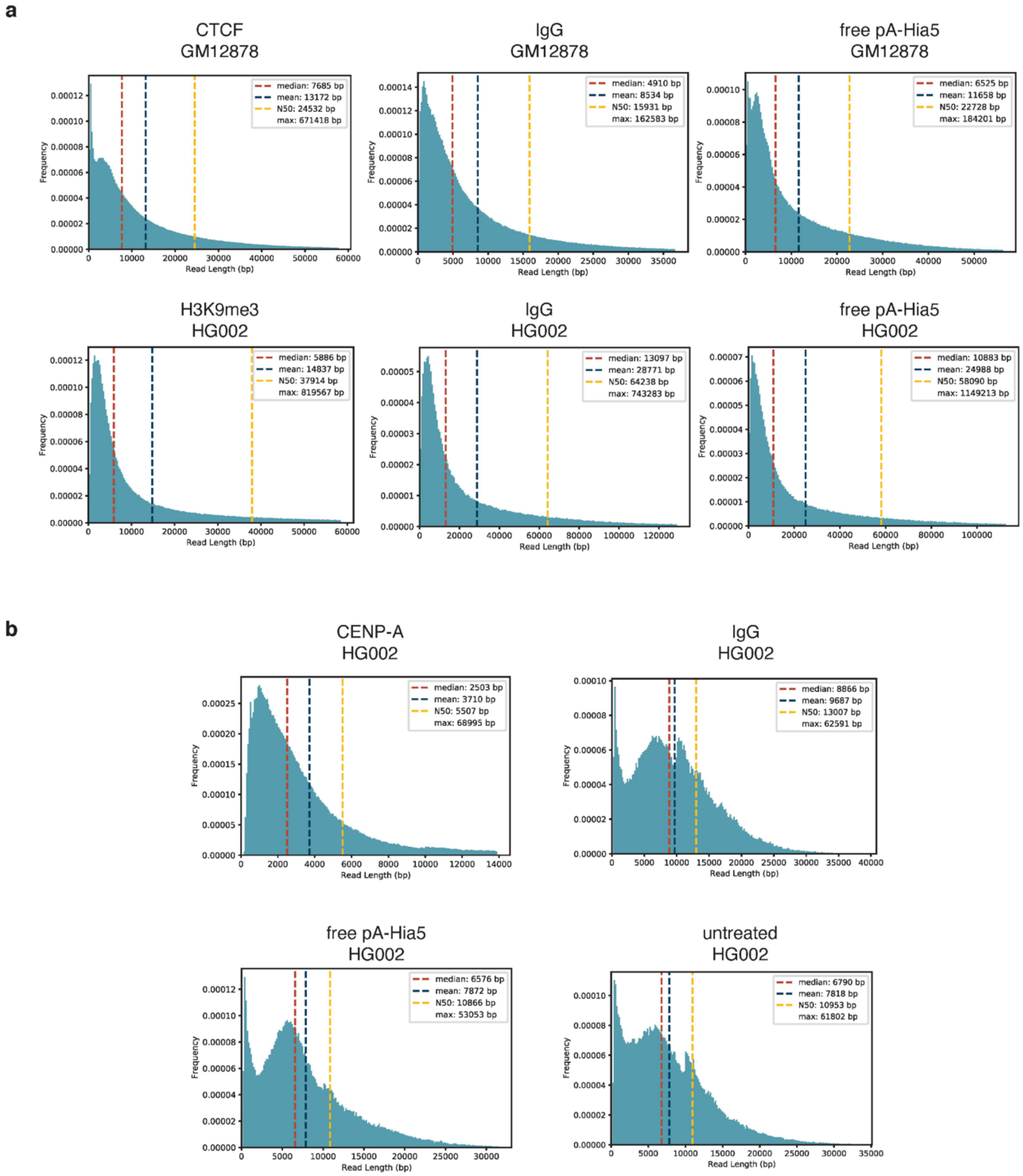


Figure 3.18 Read length distribution histograms. Read length distributions for mapped reads passing quality filters with median, mean, N50, and max read length indicated. Outliers above Q3 + 3*IQR are not shown. Distributions contain reads merged across all experiments for a given target and cell line described in Table 3.3. **a,** Histograms from samples prepared with standard DiMeLo-seq with variable library preparation cleanup methods (Methods: Nanopore library preparation and sequencing). Larger fragments are maintained in HG002 samples where we pelleted and resuspended the DNA between library preparation steps rather than performing traditional bead-based cleanups. **b,** Histograms from samples prepared with DiMeLo-seq with AlphaHOR-RES.

Cell culture

HEK293T cells (CRL-3216, ATCC, Manassas, VA; validated by microsatellite typing and mycoplasma tested) were maintained in DMEM (high glucose, with GlutaMAX, with phenol red, without sodium pyruvate; Gibco 10566016) supplemented with 10% Fetal Bovine Serum (VWR 89510-186) and 1% Pen Strep (Gibco 15070063) at 37°C in 5% CO_2. GM12878 cells (GM12878, Coriell Institute, Camden, NJ; mycoplasma tested) and HG002 cells (GM24385, Coriell Institute, Camden, NJ; mycoplasma tested) were maintained in RPMI-1640 with L-glutamine (Gibco 11875093) supplemented with 15% Fetal Bovine Serum (VWR 89510-186) and 1% Pen Strep (Gibco 15070063) at 37°C in 5% CO_2.

Cloning of pET-pA-Hia5 and pET-pAG-Hia5

The pHia5ET vector was generously provided by Andrew Stergachis and John Stamatoyannopoulos.[24] Protein A (pA) was amplified from pK19pA-MN (ASP4062, Addgene plasmid #86973, Schmid et al, 2004 [60]) and Protein AG (pAG) was amplified from pAG/MNase (ASP4154, Addgene plasmid #123461, Meers et al, 2019[91]). The pHia5ET vector was linearized via NdeI restriction digest. pA or pAG was inserted in front of the Hia5 cassette in pHia5ET using Gibson Assembly. Peptide linker between protein A (or protein A/G) and Hia5 in pET-pA-Hia5 and pET-pAG-Hia5 plasmids is DDDKEFA. All plasmid sequences were verified using Sanger sequencing. Plasmids pET-pA-Hia5 and pET-pAG-Hia5 are available from Addgene (plasmid number 174372 and 174373 respectively).

Purification of Hia5, pA-Hia5 and pAG-Hia5

pA-Hia5, pAG-Hia5, and Hia5 purification were adapted from Stergachis et al., 2020.[24] Please refer to Supplementary Note 15 for detailed protocol.

DiMeLo-seq

All reagents were prepared fresh, syringe filtered through a 0.2 µm filter, and kept on ice. Cells (1M-5M per condition) were pelleted at 300 x g for 5 minutes and washed with PBS. While live cells were used for experiments targeting CTCF, H3K9me3, CENP-A, and the accompanying controls, both frozen and fixed cells are also compatible with the DiMeLo-seq protocol. Frozen cells stored in freezing medium with DMSO in liquid nitrogen should be thawed on ice and prepared with the same protocol as fresh cells. For optional light fixation, cells can be fixed with 0.1% PFA for 2 minutes with gentle vortexing, followed by the addition of 1.25 M glycine to twice the molar concentration of PFA, a 3 minute spin at 500 x g at 4°C, and then continuation with standard DiMeLo-seq protocol's nuclear isolation. Pelleted cells were resuspended in 1 ml of Dig-Wash buffer (0.02% digitonin, 20 mM HEPES-KOH, pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1 Roche Complete tablet -EDTA (11873580001) per 50 ml buffer, 0.1% BSA) and incubated on ice for 5 minutes. Note: use of detergents other than digitonin and Tween may reduce methylation efficiency (Supplementary Note 8). The nuclei suspension was then split into separate tubes for each condition and spun down at 4°C at 500 x g for 3 minutes. All subsequent spins were performed with these same conditions, and all steps involving pipetting nuclei were performed with wide bore tips. The supernatant was removed and the pellet was gently resolved in 200 µl Tween-Wash (0.1% Tween-20, 20 mM HEPES-KOH, pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1 Roche Complete tablet -EDTA per 50 ml buffer, 0.1% BSA) containing the primary antibody at a 1:50 dilution.

Note: ensure primary antibody species is compatible with protein A. Antibodies targeted the following: LMNB1 (ab16048), CTCF (targeting C-terminus, ab188408), CTCF (targeting N-terminus, Active Motif 61312), H3K9me3 (Active Motif 39162), CENP-A (targeting Cenp-A N-terminus (amino acids 1-42), Aaron Straight, Stanford University),[92,93] and rabbit IgG isotype control (ab171870). Samples were placed on a rotator at 4°C for 2 hours. Nuclei were then pelleted and washed twice with 0.95 ml Tween-Wash. For each wash, the pellet was completely resolved by pipetting up and down ~10 times and placed on a rotator at 4°C for 5 minutes before spinning down. Following the second wash, the nuclei pellet was gently resolved in 200 µl Tween-Wash containing 200 nM pA-Hia5. pA-Hia5 concentration was measured using the Qubit Protein Assay Kit (Q33211). For pA-Hia5 binding, the nuclei were placed on the rotator at room temperature for 1 hour. Nuclei were then spun down and washed twice with 0.95 ml Tween-Wash with a 4°C rotating incubation for 5 minutes between spins, as in the wash following antibody binding. For the free pA-Hia5 control, nuclei were kept on the rotator at 4°C during antibody binding and pA-Hia5 binding steps, and pA-Hia5 was added at the time of activation. Nuclei were then resuspended in 100 µl of Activation Buffer (15 mM Tris, pH 8.0, 15 mM NaCl, 60 mM KCl, 1 mM EDTA, pH 8.0, 0.5 mM EGTA, pH 8.0, 0.5 mM Spermidine, 0.1% BSA, 800 µM SAM) and incubated at 37°C for 30 minutes before spinning and resuspending in 100 µl of cold PBS. To increase methylation efficiency, the following protocol changes were made and used when targeting LMNB1 and CTCF for experiments indicated in Table 3.1 and Table 3.3: (1) changed pA-Hia5 binding to 2 hours at 4°C, (2) increased activation time to 2 hours, (3) replenished SAM halfway through activation by adding an additional 800 µM final concentration, (4) reduced Spermidine in the activation buffer from 0.5 mM to 0.05 mM. We refer to the protocol with these changes as protocol v2. DiMeLo-seq protocol v2 requires ordinary lab equipment to prepare sequencing libraries (Figure 3.2). This protocol is also compatible with cryogenically frozen and lightly fixed samples, expanding the range of potential samples and targets (Table 3.1; interactive, updated protocol on protocols.io).

Depending on the desired read length, either the NEB Monarch Genomic DNA Purification Kit (T3010S) or the NEB Monarch HMW DNA Extraction Kit (T3050L) with 2000 rpm agitation was used to extract DNA from the nuclei. If fixation was performed, the incubation was performed at 56°C for 1 hour for lysis to reverse crosslinks. For T3050L, we agitated the sample for the first 10 minutes of lysis and then kept the samples at 56°C without agitation for 50 minutes. DNA yield was quantified using the Qubit dsDNA BR Assay Kit (Q32850).

Immunofluorescence imaging following binding with pA/G-MTase (i.e., pA-Hia5 or pAG-Hia5 or pAG-EcoGII) was performed to evaluate cell permeabilization, nuclear integrity, primary antibody on-target and background binding. For detection of pA/G-MTase binding, two different fluorophore-conjugated antibodies were used: a goat anti-mouse IgG antibody conjugated to AlexaFluor647 (Invitrogen A32728), which is not expected to bind to the rabbit primary or goat secondary antibodies but is expected to be bound by pA/G, and a goat anti-V5 antibody conjugated to FITC (Abcam 1274), which is expected to bind to the C-terminal V5 tag on pA/G-MTase. It is also possible to use a chicken anti-HisTag FITC-conjugated antibody (Abcam 3554) to avoid any binding by pA or pG. All antibodies were diluted 1:1000 for immunofluorescence imaging.

Nanopore library preparation and sequencing

For each sample, 3 μg DNA was input into library preparation using one of the following library preparation kits: (1) Ligation Sequencing Kit (ON SQK-LSK109) with Native Barcoding Expansion 1-12 (ON EXP-NBD104) and Native Barcoding Expansion 13-24 (ON EXP-NBD114) for optimization experiments and CENP-A targeted experiments after AlphaHOR-RES, or (2) Ligation Sequencing Kit (ON SQK-LSK110) for CTCF targeting, H3K9me3 targeting, and the corresponding IgG and free pA-Hia5 controls in GM12878 and HG002.

For method (1), the protocol was performed as described in the LSK109 documentation with the following modifications. End repair incubation time was increased to 10 minutes. 1 μg of end repaired DNA was loaded into barcode ligation. All ligation incubation times were increased to at least 20 minutes. Elution following barcode ligation reaction cleanup was decreased to 18 μl to allow for loading 3 μg of pooled barcoded material into the final ligation. If DNA was not sufficiently concentrated, the speedvac was used to concentrate the DNA. LFB was used for the final cleanup and elution was performed with 13 μl EB. 1 μg of DNA was loaded onto the sequencer.

For method (2), initial runs following high molecular weight extraction using NEB Monarch HMW DNA Extraction Kit with 2000 rpm agitation during lysis suffered from bead clumping during library preparation cleanups, resulting in low yields and reduced fragment size. To preserve longer fragments with the LSK110 kit, the following modifications were made to the standard LSK110 protocol.[94] End preparation incubation time was increased to 1 hour with a 30 minute deactivation. The cleanup following end preparation was performed by combining 60 μl SRE buffer (Circulomics SS-100-101-01) with the 60 μl end prep reaction, centrifuging at 10,000 x g at room temperature for 30 minutes, or until the DNA had pelleted, and washing with 150 μl 70% ethanol two times with a 2 minute spin at 10,000 x g between washes. The pellet was resuspended in 31 μl EB, and incubated at 50°C for 1 hour and then 4°C for at least 48 hours. Ligation volume was reduced by half for a total of 30 μl DNA in a 50 μl reaction volume. The ligation incubation was increased to 1 hour. The DNA was pelleted at 10,000 x g at room temperature for 30 minutes. The pellet was washed twice with 100 μl LFB, with a 2 minute spin at 10,000 x g between washes. The pellet was resuspended in 31 μl EB and incubated at least 48 hours at 4°C. For sequencing, 500 ng of the final library was loaded, with a wash using the Flow Cell Wash Kit (ON EXP-WSH004) and reload every 24 hours. Other approaches, such as using Zymo Genomic DNA Clean & Concentrator (D4065) for cleanup between reaction steps in the LSK110 protocol and using the Rapid Barcoding Kit (ON SQK-RBK004) were performed; however, LSK110 with pelleting DNA for cleanup resulted in the best throughput with the longest reads.

Sequencing was performed on an Oxford Nanopore MinION sequencer with v9.4 flow cells (ON FLO-MIN106.1) with MinKNOW software (v21.02.1). N50 varied with library preparation method, with a range from ~20 kb with LSK110 without modification to ~50 kb with LSK110 with the modifications for pelleting for DNA cleanup. See Table 3.3 for summary sequencing metrics for each sample and Figure 3.18 for read length distributions.

PacBio library preparation and sequencing

We performed PacBio sequencing on a DiMeLo-seq sample targeting CTCF in GM12878 and on unmethylated GM12878 DNA as a control. To fragment the DNA before library preparation, we targeted 20 kb fragments using a g-Tube (Covaris 520079) with 60-second spins at 4200 rpm. We prepared PacBio libraries for sequencing using the SMRTbell® Express Template Prep Kit 2.0 (100-938-900) with 1 μg input to library preparation. DNA size was determined using the TapeStation Genomic DNA ScreenTape Analysis (Agilent 5067-5365 & 5067-5366) and DNA quantification was performed using the Qubit (Invitrogen Q32853).

Primer annealing and polymerase binding to the SMRTbell libraries were performed using the Sequel II® Binding Kit 2.2 (102-089-000). An internal control complex (v 1.0) was added for sequencing quality control check. Each library was sequenced on a single SMRT cell at a loading concentration of 70 pM, as recommended for HiFi sequencing on a PacBio Sequel IIe. Sequencing runs were set up with a movie time of 30 hrs per SMRT Cell. The new adaptive loading feature in SMRTLink v10.1 was set to a loading target (P1+P2) of 0.75 and a maximum loading time of 2 hrs, as recommended for the HiFi sequencing application. CCS analysis was performed in SMRT Link v 10.1 to generate consensus reads, with the option to include kinetics information for further analysis. SMRT Cell runs produced 19.6 GB (CTCF-targeted) and 21.9 GB (untreated) of HiFi data, with a high productivity rate (P1)(% of zero-mode-waveguides with a high quality read detected) of 77.2% and 82.7%, respectively. For the CTCF-targeted sample, we sequenced 1,399,946 reads with a mean read length of 13,972 bp and a median quality score of Q33. For the untreated sample, we sequenced 1,817,035 reads with a mean read length of 12,048 bp and a median quality score of Q35.

Centromere enrichment using AlphaHOR-RES

The T2T-CHM13v1.0 reference genome was *in silico* digested with all 4-6 bp restriction enzymes available from New England Biolabs annotated as insensitive to dam or CpG methylation. A subset of these enzymes were selected based on the criteria of having less than 5% of the generated fragments map back to the alpha-satellite region of the genome and for which the genome was fragmented into at least 200,000 total fragments. Centromere enrichment was calculated after artificially removing fragments under 20 kb to simulate a size selection step and determining the fraction of remaining fragments that map to centromeric regions, as well as the loss of alpha satellite containing sequences (Figure 3.16a). Combinations of digests were then evaluated and MscI and AseI were identified as an optimal pair for centromere enrichment, predicted to yield over 20-fold enrichment when using a 20 kb size cutoff.

Genomic DNA was extracted from ~25 million cells using an NEB HMW DNA extraction kit using 300 rpm rotation during lysis (#T3050L). The DNA was eluted in a total of 300 μl elution buffer and allowed to relax at 4 °C for 2 days, although it remained viscous until it was solubilized. 37 μl NEBuffer 2.1 was added, along with 100 units of MscI and 100 units of AseI (NEB #R0534M and #R0526M) to a total volume of 370 μl in a 1.5 ml lo-bind Eppendorf tube. This was placed on a rotator at 12 rpm at 37 °C overnight. DNA concentration was then quantified using a Qubit Broad Range DNA kit (Thermo Fisher #Q32850). DNA was then mixed with orange loading buffer and loaded on a 0.3% TAE agarose gel made with Lonza SeaKem Gold agarose (# 50512) and 15 μl SYBRSafe gel stain (Thermo Fisher #S33102) per 100 ml gel. A GeneRuler High Range DNA Ladder (Thermo Fisher SM1351) was loaded in an adjacent lane. To avoid overloading, DNA was loaded with no more than 250 ng per mm of lane width (~30 μg per sample). The gel was run at 2 V/cm for 1 hour and imaged over a blue light transilluminator. The gel was cut to remove fragments smaller than 20 kb, while keeping everything larger, up to the well itself. DNA was purified from the resulting gel slice using a Zymoclean Large Fragment DNA Recovery Kit (Zymo

D4045), with modifications: the gel slice was melted at room temperature on a rotator at 12 rpm, and DNA was eluted from the column twice with the elution buffer heated to 70 °C. The DNA was then quantified by Qubit again. DNA was prepared for sequencing using an ONT LSK-109 native library prep kit, and sequenced on a v9.4 MinION flow cell. CENP-A-targeted DiMeLo-seq was performed on unfixed HG002 cells processed in parallel with IgG-targeted, free-floating pA-Hia5, and untreated samples. For each treatment ~25 million cells were processed in 5 tubes of ~5 million cells each. DiMeLo-seq was initially performed as described above. AlphaHOR-RES was performed on these samples and 250 ng to 1 ug of recovered DNA from each sample was then processed for Nanopore sequencing using method (1), described above.

Data availability

All raw fast5 sequencing data are available in the SRA with BioProject accession PRJNA752170. These data were used to produce Figure 3.4, Figure 3.6, Figure 3.10, Figure 3.14, Figure 3.17, Figure 3.3, Figure 3.5, Figure 3.7, Figure 3.8, Figure 3.9, Figure 3.11, Figure 3.12, Figure 3.13, Figure 3.15, Figure 3.16, Table 3.1, Table 3.2, Table 3.3, and Figure 3.18. CTCF ChIP-seq peak bed file for GM12878 is available from ENCODE Project Consortium with accession code ENCFF797SDL. ATAC-seq peak bed file for GM12878 is available from ENCODE Project Consortium with accession code ENCFF748UZH. Bulk and single-cell DamID data were obtained from GEO with accession GSE156150. H3K9me3 CUT&RUN data are from Altemose et al.[36] and accessible in the SRA with BioProject accession PRJNA752795. Data for Figure 3.17c used CHM13 CENP-A ChIP-seq data for CENP-A kmer analyses which are available at Bioproject accession number PRJNA559484 from Logsdon et al.[82] Data for the CpG methylation track in Figure 3.17d were obtained from data available at https://github.com/nanopore-wgs-consortium/CHM13.[23]

Code availability

The code to reproduce the results in this manuscript is available on Github: https://github.com/amaslan/dimelo-seq

3.10 Supplementary Information

Supplementary Note 1
In vitro DNA methylation assay

Hia5, pA-pHia5, and pAG-pHia5 concentrations were estimated using the extinction coefficients. Serial dilutions (100, 10, 1, 0.1 nM for Hia5 and pA-Hia5 comparison, 30, 3, 0.3 nM for Hia5 and pAG-Hia5 comparison) were made using Buffer A (15 mM Tris-HCl, pH 8.0; 15 mM NaCl, 60 mM KCl, 1 mM EDTA, 0.5 mM EGTA, and 0.5 mM spermidine). Proteins were then mixed with Buffer A supplemented with S-adenosyl-methionine (NEB B9003S) containing 1 ug of either naked unmethylated DNA (Plasmid ASP3552, 2x601, prepared from GM2163 dam- *E. coli* strain) or methylated DNA (Plasmid ASP3552, 2x601, prepared from DH5a *E. coli* strain). Reactions were incubated for 1 hr at 37°C. PCR purification was performed to extract DNA which was then digested with DpnI for 1.5 hours at 37°C and run on an agarose gel to assess the degree of methylation.

Supplementary Note 2
Reconstituted chromatin experiments

1x601 DNA containing plasmid was obtained from Addgene (pGEM-3z/601 Plasmid #26656). A 730 bp region containing 1x601 sequence in the middle was amplified (Forward primer - CAGGTTTCCCGACTGGAAAG, Reverse primer with NheI and AscI sites- GATCGCTAGCGGCGCGCCATTCAGGCTGCGCAAC) and digested with NheI. After digestion, the DNA was then biotinylated by filling in NheI 5'overhang with dGTP (NEB), α-thio-dCTP (Chemcyte), α-thio-dTTP (Chemcyte), and biotin-14-dATP (Thermo Fisher Scientific) using Large Klenow fragment 3'-5' exo- (NEB).

18x601 DNA array was obtained as previously described.[95] To summarize, puC18 vector with 18 repeats of the "601" nucleosome positioning sequence[63] (ASP 696) was transformed into competent dam- *E. coli* strain, GM2163, and purified using a QIAGEN Gigaprep kit. The unmethylated 18x601 plasmid was digested with EcoRI, XbaI, DraI, and HaeII. Array DNA used in directed methylation experiments was then biotinylated by filling in EcoRI and XbaI 5'overhangs with dGTP (NEB), α-thio-dCTP (Chemcyte), α-thio-dTTP (Chemcyte), and biotin-14-dATP (Thermo Fisher Scientific) using Large Klenow fragment 3'-5' exo- (NEB).

Histones for chromatin assembly (CENP-A, H3, H4, H2A, and H2B) were purified as previously described[95,96]. Chromatin was reconstituted using salt dialysis as described previously[95]. 1x601 or 18x601 biotinylated DNA, H2A/H2B histone dimer, and tetramer (H3/H4 or CENP-A/H4 histone tetramer) were added to high salt buffer (10 mM Tris-HCl, pH 7.5; 0.25 mM EDTA; 2 M NaCl). The mixture was gradually dialyzed over the course of ~67 hours at a rate of 0.5 mL/min from high salt buffer into low salt buffer (10 mM Tris-HCl, pH 7.5; 0.25 mM EDTA; 2.5 mM NaCl). CENP-A/H4 or H3/H4 tetramer concentrations were titrated to obtain chromatin of varying saturation. Nucleosome assembly on 1x601 DNA was verified using overnight digestion at room temperature with BsiWI (restriction site at the center of 601 sequence) followed by native acrylamide gel shift analyses and agarose gels. After digestion, intact nucleosome occupied 1x601 chromatin (uncut by BsiWI, 730bp) was separated from digested nucleosome unoccupied 1x601

DNA (cut by BsiWI, 360bp and 370bp) using glycerol gradient ultracentrifugation and fractionation. 50 ul of overnight BsiWI digested chromatin was pipetted on top of a 5 mL 5-30% (w/v) linear glycerol gradient in buffer containing 20 mM Tris pH 7.5, 0.25 mM EDTA, 0.1% Igepal/NP40, 10 mM KCl, and spun for 16 hours at 35000 rpm at 4 °C in a SW-55 Ti swinging-bucket rotor (Beckman Coulter). After the spin, 100 ul fractions were collected from the top of glycerol gradient. Fractions containing high migrating nucleosome bands (> 730 bp) were collected and concentrated for 1x601 experiments.

Chromatin assembly on 18x601 array was verified using a native acrylamide gel shift analysis after overnight restriction digestion using AvaI at room temperature (18x601 array DNA contains engineered AvaI recognition sites between adjacent 601 positions) (Figure 3.5a,c).[95]

Supplementary Note 3
In vitro chromatin methylation

In experiments involving free pA-Hia5 (non-targeted) methylation on chromatin, reconstituted chromatin was incubated in activation buffer (15 mM Tris-Cl pH 8.0, 15 mM NaCl, 60 mM KCl, 0.1% w/v Bovine Serum Albumin (BSA)) containing 0.8 mM SAM and 25 nM pA-Hia5 or pAG-Hia5 for 30 minutes at 37 °C. In antibody-directed methylation experiments, chromatin reconstituted on biotinylated 18x601 array DNA was used. In DNA LoBind Eppendorf tubes, M-280 Streptavidin-coated Dynabeads (Invitrogen) were washed in bead buffer (50 mM Tris, pH 7.4, 75 mM NaCl, 0.25 mM EDTA, 0.05% Triton X-100, and 2.5% polyvinyl alcohol (30kDa - 70 kDa)) and incubated with biotinylated CENP-A or H3 containing chromatin (at 12.5 nM 601 concentration or 0.7 nM 18x601 concentration) for 1 hour at room temperature with constant agitation. Chromatin-coated beads were then magnetically separated and washed twice with Chromatin wash buffer (CWB)-75 (20 mM HEPES pH 7.5, 75 mM KCl, 0.05% Triton X-100, 0.1% BSA) and then incubated in CWB-75 containing 1 ug/mL of rabbit anti-CENP-A antibody[97], mouse anti-H3 antibody (MABI 0301, Active Motif), or rabbit or mouse IgG (Jackson Immuno Research) control for 30 minutes with agitation at room temperature. Beads were then washed twice with CWB-75 and incubated in CWB-75 for 30 minutes with agitation, followed an additional wash in CWB-75 before incubation in CWB-75 containing 25 nM pA-Hia5 (in CENP-A-directed methylation experiments) or pAG-Hia5 (in H3-directed methylation experiments). After incubation with pA-Hia5 or pAG-Hia5, beads were washed twice with CWB-100 (20 mM HEPES pH 7.5, 100 mM KCl, 0.05% Triton X-100, 0.1% BSA) to remove unbound pA-Hia5 and then resuspended in activation buffer (15 mM Tris-Cl pH 8.0, 15 mM NaCl, 60 mM KCl, 0.1% w/v BSA) containing 0.8 mM SAM for 30 minutes at 37°C. Beads were then split into two tubes and processed separately for immunostaining (with anti-mA antibody) and library preparation (for long-read sequencing). For library preparation, chromatin was released from beads using BamHI and KpnI digestion (cuts near biotinylated ends of 18x601 array DNA), or using AscI digestion (cuts near biotinylated end of 1x601), DNA was extracted, and processed using Oxford Nanopore Technology native barcoding (PCR-free) kit (EXP-NBD104 or EXP-NBD114) with the Ligation Sequencing Kit (SQK-LSK109).

Supplementary Note 4
In vitro chromatin DiMeLo-seq analyses

Reads from *in vitro* experiments were initially basecalled with Guppy (4.4.2) using the fast basecalling model (dna_r9.4.1_450bps_fast.cfg). After initial basecalling reads were demultiplexed and split by barcode using the guppy_barcoder and fast5_subset from ont_fast5_api. Fast5s for each barcode were then aligned and modification basecalled with Megalodon (2.2.9) using the rerio all-context basecalling model (res_dna_r941_min_modbases-all-context_v001.cfg) with --guppy_params "trim_barcodes" and --mod_min_prob 0. In experiments involving 1x601 array, reads less than 700 bp, representing BsiWI digestion products of unoccupied 1x601, were removed from downstream analyses. In experiments involving 18x601 array, reads less than 3.6 kb (i.e., 18x 200 bp repeats of 601), representing partial arrays, were removed from downstream analyses. Modification basecalled reads were smoothed by calculating rolling average over a 50 bp window in a NaN-sensitive manner (averaging only over adenine bases). Following smoothing, adenine bases with methylation probability score > 0.6 were assigned as methylated (mA). The threshold of 0.6 was empirically determined by comparing pA-Hia5 treated and untreated naked 1x601 DNA, false detection rate (FDR) < 5%, (Extended Data Fig. 1f). FDR was estimated using binned probability scores for reads from naked DNA methylated with free pA-Hia5 or untreated, corresponding to True Positive or True Negative respectively. For a given cutoff, a read is classified as methylated if the percentage of methylation (i.e., % mA/A) on that read is greater than the cutoff. FDR was calculated as (FPR/(FPR+TPR)). For classifying 1x601 reads as methylated, we empirically determined the minimum percentage of each read to be methylated above a given threshold (0.6) from Receiver Operator Characteristic curves comparing binned methylation on reads from 1x601 CENP-A chromatin after CENP-A-directed methylation (as TPR) to IgG-directed methylation (Extended Data Fig. 1g) or no treatment (Extended Data Fig. 1h) (as FPR). In experiments estimating extent of methylation on 1x601 reads (Figure 3.4d), we classify a portion of the read centered at the 601 dyad as methylated if 20% of its length is methylated above the threshold of 0.6 (Purple dot in Extended Data Fig. 1g, h, Dotted line on Extended Fig. 1k).

For clustering and visualizing methylation on individual 18x601 reads, we first classified each 601 position as with or without nucleosome. A region spanning 400 bp centered at each theoretical 601 dyad position was classified as containing a nucleosome if > 10% and < 60% of that region was methylated. (< 60% is used to filter out regions that do not show nucleosome protection). Reads were then clustered by performing hierarchical clustering of jaccard distances of inferred nucleosome positions on either 18x601 or 4x601 region.

Supplementary Note 5
Chromatin-coated beads immunostaining and imaging

Following incubation in activation buffer, chromatin coated beads were incubated in CWB-2M (20 mM HEPES pH 7.5, 2M NaCl, 0.05% Triton X-100, 0.1% BSA) for 1 hour at 55 °C to denature protein. Beads were then washed twice in CWB-2M to remove denatured protein while retaining biotinylated DNA on beads. (Anti-CENP-A antibody and anti-methyladenine antibody are both derived from rabbit, therefore, to avoid cross-reactivity with anti-rabbit conjugated secondary antibody used for immunofluorescence, chromatin coated beads were washed with CWB-2M as mentioned above to remove CENP-A antibody prior to staining with anti-N6-methyladenosine antibody.) Beads were then washed twice with Antibody dilution buffer or AbDil (20 mM Tris-HCl, pH 7.4, 150 mM NaCl with 0.1% Triton X-100, and 2% BSA) and dropped onto poly-Llysine-coated coverslips and allowed to attach for 30 minutes. Coverslips were incubated with AbDil containing 1 ug/ml rabbit anti-N6-methyladenosine antibody (Millipore Sigma ABE572) for 30 minutes, washed twice with AbDil, and incubated with AbDil containing 2 ug/ml Alexa 647 fluorophore conjugated goat anti-rabbit secondary antibody (Molecular Probes) for 30 minutes. Coverslips were then washed twice with AbDil, incubated with AbDil containing 1 ug/ml propidium iodide (Sigma) for 10 minutes, washed twice with AbDil and phosphate buffered saline (PBS), blotted gently, mounted in 90% glycerol, 10 mM Tris-Cl pH 8.8, and 0.5% *p*-phenylenediamine, and sealed using clear nail polish.

Imaging was performed using IX70 (Olympus) microscope with a DeltaVision core system (Applied precision) with a Sedat quad-pass filter set (Semrock) and monochromatic solid-state illuminators, controlled via softWoRx 4.1.0 software (Applied Precision). Images were acquired using a 100x 1.4 NA Plan Apochromatic oil immersion objective (Olympus) and captured using a CoolSnap HQ CCD camera (Photometrics). Z-stacks were acquired at 0.2 uM intervals over a total 3 uM total axial distance. Bead images were analyzed using custom ruby software.[96] At least 50 beads were analyzed for each condition per experiment.

Supplementary Note 6
Modification calling thresholds

Basecalling was performed using Oxford Nanopore Technologies's Guppy software (v4.5.4) and Megalodon software (v2.3.1) with the Rerio res_dna_r941_min_modbases-all-context_v001.cfg basecalling model. To estimate false positive rates (FPR) at each mA probability score threshold, we counted the fraction of As called as mA on untreated GM12878 genomic DNA, which should lack any methyladenines (Figure 3.7). To provide a lower bound on the true positive rate (TPR), we counted the fraction of As called as mA on purified GM12878 genomic DNA treated with pA-Hia5 *in vitro* (Figure 3.7). Using these values, we could estimate a lower bound on the FDR (FPR/(FPR+TPR)). For Guppy modified base calls, we used a modification probability threshold of 0.6 (basecalling 0.0009 FDR, 0.000245 FPR, 0.281 TPR lower bound). For Megalodon's modified base calls, we used a modification probability threshold of 0.75 (basecalling 0.0008 FDR, 0.000159 FPR, 0.203 TPR lower bound). For some analyses higher thresholds were used; for example, a stringent Guppy threshold of 0.9 was used for LMNB1 analyses (Figure 3.8). We note that the predominant source of background noise in DiMeLo-seq stems from off-target methylation, as opposed to false-positive methylation calls. Using higher mA score thresholds effectively serves as a threshold on higher mA density, to distinguish on-target methylation from off-target methylation.

Supplementary Note 7
LMNB1 data analysis

All sequencing was performed on ONT MinION v9.4 flow cells. Basecalling and modification calling were performed on Amazon Web Services g4dn.metal instances, which have 8 NVIDIA T4 GPUs, 96 CPUs, 384 Gb memory, and 2x900 Gb local solid-state storage; this configuration allows for efficient parallelization and high basecalling speed. Basecalling was first performed using Oxford Nanopore Technologies's Guppy software (v4.5.4), using a Rerio res_dna_r941_min_modbases-all-context_v001.cfg basecalling model, and demultiplexing when appropriate. Modification calls were extracted from fast5 output files using ont-pyguppy-client-api. Basecalled reads were aligned to the T2T-CHM13v1.0 reference sequence using Winnowmap (v2.03), which is adapted to perform better than other long-read aligners in repetitive regions[98]. Fast5 files were split by barcode using fast5_subset then re-basecalled using ONT's Megalodon software (v2.3.1), using the same reference and model file. Custom code was used to parse output files and is available on Github. To evaluate performance, cLAD and ciLAD coordinates[35] were lifted over from hg38 to the T2T-CHM13v1.0 reference.[59] Single-cell Dam-LMNB1 data were re-mapped to T2T-CHM13v1.0 and processed as described in Altemose et al.[35] Browser plots were made using the WashU Epigenome Browser.[53]

Figure 3.6e: In single-cell DamID, each 100 kb bin of the genome is given a binary classification indicating whether it was in contact with the nuclear lamina or not in that particular cell during an ~18 hour incubation period when Dam-LMNB1 is expressed *in vivo.*[30] Across a sample of 32 single cells, we used these binary classifications to estimate a scDamID-based LMNB1 interaction frequency for each bin of the genome across the sample of cells.[30,35] We then performed a similar binary classification of individual LMNB1-targeted DiMeLo-seq reads based on each read's proportion of methylated adenines, determining a lamina interaction threshold (mA/A > 0.001 with a stringent mA calling threshold of 0.9) to identify reads from cLADs with 59% sensitivity and 94% specificity (Figure 3.8). The DiMeLo-seq-based interaction frequency for each bin was then computed as the proportion of overlapping reads with mA/A above the lamina interaction threshold. A read was determined to overlap a bin if it aligned to it with more than 50% of its length, and any mA calls on that read were assigned to that bin for browser plotting. 100 kb genomic bins were filtered to those with at least 60 overlapping DiMeLo-seq reads, and with a single-cell combined mean-squared-error estimate <0.004, to select for regions with higher-confidence interaction frequency estimates.

Supplementary Note 8
LMNB1 optimization

We found that we could reliably estimate protocol performance parameters (on-target methylation and on-target:off-target methylation) using only ~0.2X genome-wide coverage per sample, allowing us to multiplex several conditions on the same MinION flow cell and achieve sufficient coverage after only 24 hours of sequencing. Using the v2 protocol and applying a stringent methylation score threshold of 0.9 (Figure 3.7a,b-Figure 3.8, Supplementary Note 6), we regularly achieve on-target methylation of 0.3-0.6% of adenines in cLADs, with an on-target:off-target ratio in the range of 15-30 (Table 3.1). These performance metrics depend on the choice of mA score threshold (Figure 3.8c), which was chosen to balance sensitivity and specificity in distinguishing regions with on-target and off-target methylation. We note that this threshold does not primarily serve to reduce false-positive mA calls, which occur at an extremely low rate (Figure 3.7a,b; see full discussion of threshold evaluation in Supplementary Note 6). Unlike other protein-DNA mapping methods, which use sequencing coverage as a readout of interaction frequency, DiMeLo-seq sequences the entire genome without enrichment for interacting regions. Thus, as further validation we can plot DiMeLo-seq's coverage and methylation frequency as separate tracks in a browser representing the T2T-CHM13 complete reference sequence, and we can compare these to the results obtained for the same protein target in the same cells by conventional bulk DamID (Figure 3.6c).

Surprisingly, we found no improvement in on-target methylation when using a secondary antibody to recruit more methyltransferase molecules to each site, perhaps due to steric effects, and we saw no improvement when increasing the linker length between pA and Hia5 (Figure 3.7 and Table 3.1). We saw a slight drop in performance when using pAG-Hia5 compared to pA-Hia5, also potentially due to steric effects. We also found that cell permeabilization with NP40 or Triton X-100 (vs. standard digitonin) actively reduces methylation downstream (Table 3.1). While optimization was carried out in HEK293T cells, we also validated that the protocol worked in other human cell lines: Hap1, GM12878, and HG002.

Supplementary Note 9
DiMeLo-seq with Concanavalin A coated magnetic beads & input considerations

Concanavalin A coated magnetic beads (Bangs Laboratories BP531) were tested as an alternative to centrifugation for cell pelleting throughout the protocol, adapted from the CUT&RUN protocol.[5,99] To equilibrate beads, conA bead slurry was resuspended by gentle vortexing and 10 µL of bead slurry per sample was added to 1.5 ml conA binding buffer (20 mM HEPES pH 7.5, 10 mM KCl, 1 mM CaCl2, 1 mM MnCl2), then placed on a magnet. Supernatant was removed and beads were resuspended in 1.5 ml conA binding buffer again, then cleared on a magnet again. Washed beads were resuspended in 20 ul conA binding buffer per sample. For the experiments numbered 64-66 in our Table 3.1, we used conA beads with 500k, 430k, and 500k cells each for HEK293T (~triploid), GM12878 (diploid), and Hap1 (haploid) cells, respectively. Prior to the permeabilization step, cells were first washed with PBS 3x by centrifugation, then resuspended in 1 ml wash buffer (20 mM HEPES-KOH, pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1 Roche Complete tablet -EDTA per 50 ml buffer). 10 µl of equilibrated bead slurry was added while gently vortexing the cell suspension, and the beads + cells were incubated for 10 minutes at room temperature on a rotator. Bead-bound cells were pelleted on a magnet and resuspended in Dig-Wash buffer to begin the permeabilization step, and the remainder of the protocol was carried out as described above, substituting magnetic separation for centrifugation. Note: conA beads may interfere with the ability to perform quality control by IF. At the DNA extraction step, cells were lysed on beads, and beads were separated from lysate on a magnet prior to proceeding with DNA precipitation. The final DNA yield was 20% for HEK293T, 39% for GM12878, and 75% for Hap1, relative to what one would theoretically expect from the input number of cells, after accounting for ploidy (estimated as 3*ploidy pg per cell). This variance in DNA recovery may have to do with the propensity for each cell type to bind conA beads and resist nuclear envelope rupture, or possibly to do with relative cell sizes and the binding capacity of the conA beads.

The input requirements for DiMeLo-seq ultimately depend on a multitude of factors: the desired coverage, the desired fragment length distribution, the genome size, the ploidy of the cell type, and the efficiency of the DNA extraction and library prep protocols being used. For the conA bead experiment with 430k GM12878 cells, we yielded 500 ng of DNA after extraction, and ~200 ng after library prep. If prepared with an lsk-110 kit, this would be enough to load a minION flowcell twice while maintaining high pore occupancy (100 ng per loading). Each loading of a flowcell yields ~9 Gb on average, so this amount of DNA would provide 6x coverage of the human genome. Thus, based on our empirical results from this replicate, we can estimate that around 200k diploid cells are needed for 3x human genome coverage. For a line/protocol with higher recovery efficiency, this could fall closer to 100k cells per 3x human genome coverage. These input requirements may continue to decrease with flow cell designs and new library prep chemistries.

Supplementary Note 10
Creation and induction of stable cell lines for *in vivo* DiMeLo-seq

Stable HEK293T cell lines were created by retroviral transduction followed by drug selection. Retroviral plasmids containing DDdegron-EcoGII-V5linker-LMNB1 were obtained from Addgene (#122083; Sobecki et al.[65]). Retroviruses were produced in the Phoenix Ampho packaging cell line (obtained from the UC Berkeley cell culture facility). Phoenix cells were seeded in standard growth medium (DMEM with 10% FBS and 1X P/S) in a T75 flask 24 hours before transfection, aiming for 70% confluence at the time of transfection. 25 µg of plasmid DNA was combined with 75 µl FUGENE-HD transfection reagent in 1200 µl optiMEM and incubated for 10 minutes, then added to the media. After 12 hours, the media was replaced with fresh media, and the cells were incubated at 32 °C with 5% CO_2 and 100% humidity to help preserve viral particles. 36 hours later, the virus-containing media was harvested and centrifuged at 1800 rpm for 5 minutes to remove any Phoenix cells. The media was supplemented with 10 µl/ml of 1 M HEPES and 4 µg/ml of polybrene. For HEK293T cells, 2.5 ml of this media was added to each well of a 6-well plate containing adhered cells at 40-50% confluence. Plates were spinoculated in a centrifuge with a swinging-bucket plate rotor at 1300x*g* for 1 hour at room temperature, then incubated at 37 °C overnight. The media was replaced the next morning. After 24 hours, puromycin was added to the media at a concentration of 1 µg/ml and the media was replenished every 48 hours for 10 days. Surviving cells were expanded and frozen for later use. 15 hours prior to harvesting, 1 µM Aqua-Shield-1 reagent (AOBIOUS AOB6677, made to 0.5 mM stock) was added to the media to stabilize protein expression. DNA was harvested using an NEB Monarch Genomic DNA Purification Kit (T3010S), sheared to a target of 8 kb using a Covaris g-tube (Covaris 520079), and purified with a Circulomics SRE XS kit (SS-100-121-01), then barcoded and library prepped with method 1 described below.

The higher methylation observed in the *in vivo* sample likely owes to the effectively longer incubation time during which methyl groups can be deposited on adenines *in vivo* (15 h) compared to *in situ* (2 h), as well as to chromatin dynamics *in vivo* that may make a greater fraction of the genome accessible to the methyltransferase[61]. However, compared to *in situ* DiMeLo-seq with pA-Hia5, the *in vivo* EcoGII-LMNB1 approach produced 36% less on-target methylation and 25% more off-target methylation.

Supplementary Note 11
CTCF data analysis

For GM12878 samples (CTCF-targeted, IgG control, free pA-Hia5, in vitro methylated genomic DNA, and untreated genomic DNA), Megalodon modified basecalls were used for analysis. Reference GM12878 ChIP-seq peaks (ENCFF797SDL, ENCODE Project Consortium[1]) were lifted over from hg38 to T2T chm13v1.0. These peaks were intersected with known CTCF motifs that were also lifted over to T2T chm13v1.0.[100] Reference GM12878 ATAC-seq peaks (ENCFF748UZH, ENCODE Project Consortium[1]) were also lifted over from hg38 to T2T chm13v1.0. Enrichment in CTCF ChIP-seq peaks and ATAC-seq peaks was calculated using bedtools (v2.28.0). The anti-CTCF antibody (ab188408) was confirmed by personal correspondence to bind a peptide in the first 600 C-terminal amino acids of the protein.

For analysis of CTCF-targeted DiMeLo-seq data, modified basecalls for reads spanning CTCF ChIP-seq motifs were extracted within -1 kb to 1 kb of the motif center. To extract single molecules spanning peaks, pysam (v0.15.3) was used with code adapted from De Coster et al.[101] If the motif was on the - strand, the positions of bases relative to the motif center were flipped. Only non-overlapping CTCF sites within the -1 kb to 1 kb display were considered, and only mA called with probability ≥ 0.75 were plotted. Aggregate profiles were plotted with a moving average of 50 bp. For peak and read counts considered, see Figure 3.9j. For joint analysis of mA and mCpG on the same molecules, only molecules spanning motifs in the top decile of ChIP-seq peaks that have at least one mA called with probability ≥ 0.75 and one mCpG called with probability ≥ 0.75 were considered, resulting in 23,147 reads considered.

To test that C-terminal and N-terminal CTCF targeting produce significantly different methylation enrichment patterns, we performed a Fisher's exact test comparing the fraction of methylated adenines 3' to the motif center (-300 bp to 0 bp) to the fraction of methylated adenines 5' to the motif center (0 bp to +300 bp) for the two samples (p-value of 5.9×10^{-8}). We used the fisher_exact function from scipy (v1.4.1) with the alternative hypothesis "greater" for this comparison of C-terminal to N-terminal targeting. The significance still held when narrowing the test region to include only the central peak (-100 bp to 100 bp), and in this region, the p-value was 0.00010.

We detected peaks in our DiMeLo-seq data in aggregate to create Figure 3.9f. First, we took the average probability of methylation reported across all reads for a given base in the reference. We then computed the mean methylation probability in a 200 bp sliding window with a 20 bp step size. Next, for various average methylation cutoff thresholds from 0 to 50, we classified 200 bp bins as true positive (above threshold in DiMeLo-seq, overlapping ChIP-seq peak), false negative (below threshold in DiMeLo-seq, overlapping ChIP-seq peak), false positive (above threshold in DiMeLo-seq, not overlapping ChIP-seq peak), or true negative (below threshold in DiMeLo-seq, not overlapping ChIP-seq peak). We then created the ROC curves having performed this peak calling method with 25X, 20X, 15X, 10X, and 5X coverage for our CTCF-targeted sample. The area under the curve was calculated for the various sequencing depths using sklearn.metrics.auc function (v.0.24.2).

To call CTCF peaks on single molecules, all molecules spanning top decile ChIP-seq peaks that had at least one mA detected with probability ≥ 0.9 were considered (25,122 reads). We used a more stringent probability threshold to select reads that contained confident methylation because the goal was to determine where the peak center is detected on reads that call a peak. Reads were also filtered to require they span the CTCF motif with at least 100 bp covered on each side of the motif. With a sliding window of 20 A's, the probability that at least one A was methylated within the bin was computed by calculating 1-exp(sum(log(1-p))) for each mA with probability > 0.5. We used a lower probability cutoff for calculating binned probabilities to detect peaks on single molecules because on single molecules, we wanted to capture any mA calls, even lower confidence calls, to increase our sensitivity at the cost of specificity. Reads that had at least one binned probability ≥ 0.8 have a called peak, and the peak center was calculated as the midpoint of the longest stretch of mA with binned probability ≥ 0.8 (1.2% FDR). The FDR was calculated as the fraction of adenines methylated in the unmethylated control divided by the fraction of adenines methylated in the CTCF-targeted sample using these same filtering criteria.

To estimate the single-molecule sensitivity for detecting CTCF binding events, we performed a binary classification of CTCF-targeted DiMeLo-seq reads based on the proportion of adenines methylated. We quantified this proportion in both on-target peak regions, defined as +/- 150 bp of the CTCF binding motif center in top decile ChIP-seq peaks and in off-target regions, defined as -2000 to -1850 and +1850 to +2000 bp of the top decile motif center, once baseline background *in situ* methylation levels have been reached. Using this approach, we calculated TPR and FPR as a function of the number of mA in these 300 bp regions required to consider a CTCF binding event detected, and approximated the single-molecule sensitivity to be 54% (FPR 5.7%) when requiring 6 mA with probability ≥ 0.75 in 300 bp for a CTCF binding event detection.

For analysis of single molecules spanning two CTCF sites, peak pairs that were 2 to 10 kb apart were selected from all CTCF ChIP-seq peaks). As in peak calling, binned qualities in bins of 20 A's were computed. Here, if a binned probability > 0.9 fell within 100 bp on either side of at least one of the two CTCF binding sites, the read was considered to have a called peak and the molecule was included in Figure 3.10c. A total of 1959 peak pairs were considered with a total of 3036 reads spanning these peaks with a peak detected at at least one of the two sites (4207 total reads spanned these pairs of sites). Reads were clustered using k-means clustering (scikit-learn v0.24.2) with 3 clusters.

A vcf file containing high-quality phased heterozygous polymorphisms in GM12878 were obtained from https://hgdownload.soe.ucsc.edu/gbdb/hg38/platinumGenomes/hg38.hybrid.vcf.gz, which combines variant calls from the Platinum Genomes and Genome in a Bottle projects[102,103]. This vcf was lifted over from hg38 to CHM13v1.0 using VCF-liftover (https://github.com/hmgu-itg/VCF-liftover) with a chain file from http://t2t.gi.ucsc.edu/chm13/hub/t2t-chm13-v1.0/hg38Lastz/hg38.t2t-chm13-v1.0.over.chain.gz. DiMeLo-seq alignments were phased using NanoMethPhase v1.0[104] with parameters --mapping_quality 10 --min_SNV 1 --average_base_quality 10. Because NanoMethPhase requires base quality values, the input bam files for CTCF phasing were obtained by merging guppy output bam file information with alignment position information from winnowmap using custom in-house code available on github. IGV v2.11.4[105] was used for initial data exploration (note: for megalodon mod_mappings bam files, "C+Z" was replaced with "C+m", and "A+Y" was replaced with "A+a" in each line of the bam file for proper visualization). Final single-molecule plots were made with custom in-house code available on github. CTCF site coordinates within the H19/IGF2 Imprinting Control Region were obtained from Ulaner et al.[106]

Supplementary Note 12
PacBio data analysis

Starting with the hifi_reads.bam file output from the sequencer, we used SMRTLink (v10) command-line tools to process the data. First we used ccs-kinetics-bystrandify to create a bam file with forward and reverse strands as separate reads. We aligned this bam file using pbmm2 align to the T2T-CHM13v1.0 reference and extracted reads that overlapped the top decile of CTCF

ChIP-seq peaks using bedtools (v2.28.0) intersect. We then ran a custom script provided by PacBio to compute an IPD ratio for each base. We aligned this output using pbmm2 align to the T2T-CHM13v1.0 reference. We then used custom scripts to extract single base IPD ratios for comparison to nanopore for Figure 3.13. In particular, methylated base calls +/- 100 bp around the CTCF motif center for top decile CTCF ChIP-seq peaks were extracted. For PacBio, we plotted the fraction of adenines methylated in this peak region as a function of IPD ratio and number of passes. For Nanopore, we plotted the peak methylation as a function of mA probability. For both, we compared to the methylation detected in the untreated control in this same peak region. We then selected a constant peak methylation level of 10% of adenines methylated and compared the profiles for PacBio and CTCF with thresholds corresponding to a peak methylation rate of 10%.

Supplementary Note 13
H3K9me3 data analysis

For all HG002 samples (H3K9me3-targeted, IgG control, and free pA-Hia5) a merged bam file was created with samtools (v1.8) from the Guppy bam and winnowmap outputs aligned to a special male reference genome (CHM13+HG002X+hg38Y: autosomes from the T2T chm13v1.0 genome combined with a T2T assembly of HG002 chromosome X[59] and the chrY sequence from hg38), and a mapping quality threshold of 10 was applied. To compare to CUT&RUN, broad peaks were called using macs2 (v2.1.1) on a H3K9me3 CUT&RUN bam file from HG002.[76] Regions outside H3K9me3 CUT&RUN regions for Figure 3.14a were defined as regions of the genome outside of the called broad H3K9me3 peaks with 10 kb buffer on each side of the called peaks. Centromere and HOR boundaries were defined from the T2T centromere annotation.[76] Enrichment in CUT&RUN peaks, centromeres, and active HOR arrays was computed using bedtools (v2.28.0). For analyzing mA signal at HOR boundaries, the mean mA/A in a 100 kb rolling window from -300 kb within the HOR to 300 kb outside of the HOR was computed. A total of 2,359 reads spanned this region. HOR boundaries considered were those that transition quickly into non-repetitive sequences: 1p, 2pq, 6p, 9p, 13q, 14q, 15q, 16p, 17pq, 18pq, 20p, 21q, 22q. For single molecule browser visualization, modified bases were extracted as in CTCF analysis using custom python scripts, and modified bases with probability ≥ 0.6 were displayed. Single-molecule browser plots were generated using plotly (v4.5.2) with code adapted from De Coster et al.[101]

Supplementary Note 14
CENP-A data analysis

Basecalling for centromere enriched samples was performed twice both times using Guppy (5.0.7). The first basecalling used the "super accuracy" basecalling model (dna_r9.4.1_450bps_sup.cfg), followed by alignment to the CHM13+HG002X+hg38Y reference genome using Winnowmap (v2.03). These alignments were then filtered for only primary alignments and mapq score greater than 10 using samtools view -F 2308 -q10. A second round of basecalling was then performed again using Guppy (5.0.7) but now with the rerio all-context basecalling model (res_dna_r941_min_modbases-all-context_v001.cfg) with --bam_out and --bam_methylation_threshold 0.0. Modified basecalls were then merged by read id with winnowmap alignments to generate bam files with high confidence alignments combined with modification calls for downstream processing. For CENP-A-directed experiments four independent biological replicates were used, and for controls (IgG-directed, free-floating pA-Hia5, and untreated), two independent biological replicates were used. For all samples the first replicate was sequenced on two separate flow cells and all sequencing runs were merged for the final analysis.

To calculate centromere enrichment samtools bedcov was used to calculate the total bases that mapped to alpha satellite active HORs in free-floating Hia5 treated samples treated with and without centromere enrichment[107]. Reported coverage at each centromeric region is relative to the length of that region. Chromosomes with more than one active HOR had the mean value of length-normalized coverage reported. deepTools2 bamCoverage (v3.3.1)[108] was used to generate bigWigs with 10 kb bin size, that were plotted on HG002 chromosome X using pygenometracks (v3.6)[109] to compare chromosome-wide coverage between centromere enriched and unenriched samples.

A *k*-mer counting pipeline was used to identify CENP-A enriched *k*-mers from chm13 Native ChIP-seq experiment.[82,84] After separating DiMeLo-seq reads into those that did and did not have a CENP-A enriched *k*-mer, methylation frequency for each subset was calculated, as well as the fold enrichment for percentage mA/A of reads containing CENP-A enriched *k*-mers over those that did not.

For single molecule browser visualization, modified bases were extracted as in CTCF analysis using custom python scripts, and modified bases with mA probability (from Guppy) > 0.6 and mCpG probability (from Guppy) > 0.6 were displayed.

Average fraction of mA or mCpG methylation for aggregate views were calculated as the fraction of reads at each adenine or CpG that have a probability score (from Guppy) greater than 0.6. Representative plots show average fraction of reads at each adenine or CpG with methylation probability score above threshold binned by smoothing over a rolling window of 250 bp for better visualization. Coverage plots indicate the number of reads that are aligned within the region.

For estimating the density of CENP-A-containing nucleosomes per read, a 5 kb window was slid across each read (step size 1 bp), and within that 5 kb window, the proportion of all 200 bp windows (step size 1 bp) containing at least 3 mAs was computed (using a Guppy mA probability threshold of 0.6). On average within alpha satellite, this threshold of 3 mAs corresponds to 5% of all A bases within a 200 bp window. These values were then averaged across all 5 kb read windows overlapping each 5 kb reference window to produce the density plot in Figure 3.17g. The thresholds for bin size, minimum percentage methylation, and probability score were empirically determined to produce a 5% FDR, using IgG DiMeLo-seq reads as a true negative control set.

Supplementary Note 15
Protein purification

Histones for chromatin assembly (CENP-A, H3, H4, H2A, and H2B) were purified as previously described.[95,96] pA-Hia5, pAG-Hia5, or Hia5 purification were purified according to Stergachis et al.[24] with a few modifications. Plasmids were transformed into T7 Express lysY competent *E. coli* cells (NEB #C3010I) for recombinant protein expression. 200 mL starter culture was grown in LB broth at 37°C with 50 ug/mL kanamycin and 34 ug/mL chloramphenicol to an OD_{600} of 0.6. Starter cultures were then diluted to 2L culture in LB broth at 37°C with 50 ug/mL kanamycin to an OD_{600} of 0.8 - 1.0. Protein expression was induced using a final concentration of 1mM IPTG (Isopropyl beta-D-1-thiogalactopyranoside) for 4 hours at 20°C with shaking. Cells were then pelleted at 5000 x g for 15 minutes at 4°C. Pelleted cells were resuspended in 35 mL lysis buffer (50 mM HEPES, pH 7.5; 300 mM NaCl; 10% glycerol; 0.5% Triton X-100). Resuspended cell pellets were flash frozen in liquid nitrogen and stored at -80°C until purification. After thawing frozen cell pellets, EDTA-free protease inhibitor tablets (Roche 11873580001) and 10 mM β-mercaptoethanol were added. Cells were lysed by probe sonication (6 pulses, 30s on, 1 min off at 200W). Lysed cells were centrifuged for 1 hour at 40,000 x g (4°C) in 50 mL Oakridge tubes. Ni-NTA agarose was prepared with 2x washes of 30 mL equilibration buffer (50 mM HEPES, pH 7.5; 300 mM NaCl; 20 mM imidazole) per 5 mL of slurry. Cell lysate was incubated with Ni-NTA agarose and rotated for 1 hour at 4°C. Mixture was poured onto a gravity column, then washed with 40 mL buffer 1 (50 mM HEPES, pH 7.5; 300 mM NaCl; 50 mM imidazole), 30 mL of buffer 2 (50 mM HEPES, pH 7.5; 300 mM NaCl; 70 mM imidazole), and eluted with 30 mL of elution buffer (50 mM HEPES, pH 7.5; 300 mM NaCl; 250 mM imidazole). Eluted protein was filtered with a 0.2 μm filter, then loaded onto a HiPrep 26/10 Desalting column (Cytiva) to buffer exchange eluate into FPLC buffer A (50 mM Tris-HCl, pH 8.0; 100 mM NaCl; 1mM DTT). Following buffer exchange, the sample was applied in tandem onto HiTrap Q HP and HiTrap SP HP (Cytiva) columns. Both columns were washed with 5x combined column volumes of FPLC buffer A. The HiTrap Q HP (Cytiva) column was removed and protein was eluted from the SP column using a linear gradient of 20 column volumes with increasing linear gradient of FPLC buffer B (50 mM Tris-HCl, pH 8.0; 1 M NaCl; 1 mM DTT). Fractions were collected and quantified using A280 absorbance. Elution peak fractions were concentrated using a 10K Amicon Ultra-15 tube to final protein concentration > 5 uM. The final concentrated protein was supplemented with 10% glycerol final concentration, aliquoted, and stored at -80°C.

Chapter 4

DiMeLo-seq benchmarking & software package

4.1 Overview

In distributing DiMeLo-seq to other labs, it quickly became apparent that the analysis was the largest hurdle to adopting the protocol. Therefore, I led the development of an analysis software package to make quality control, data visualization, and modified basecall parsing simple for users with limited programming experience. We applied this package to deeply sequenced DiMeLo-seq data targeting H3K27ac, H3K27me3, and H3K4me3 in GM12878. We also targeted H3K9me3 in Drosophila melanogaster embryos, the first application of DiMeLo-seq to primary cells. Overall, this work provides an analysis framework for analyzing DiMeLo-seq data, an augmented protocol, and benchmarking and analysis of new protein targets. A preprint of this work was published in Maslan et al., 2022.[110]

4.2 Abstract

We recently developed **Di**rected **Me**thylation with **Lo**ng-read **seq**uencing (DiMeLo-seq) to map protein-DNA interactions genome wide. DiMeLo-seq maps multiple interaction sites on single DNA molecules, profiles protein binding in the context of endogenous DNA methylation, and maps protein-DNA interactions in repetitive regions of the genome that are difficult to study with short-read methods. Adenines in the vicinity of a protein of interest are methylated in situ by tethering the Hia5 methyltransferase to an antibody using protein A. Protein-DNA interactions are then detected by direct readout of adenine methylation with long-read, single-molecule, DNA sequencing platforms such as Nanopore sequencing. Here, we present a detailed protocol and guidance for performing DiMeLo-seq. This protocol can be run on nuclei from fresh, lightly fixed, or frozen cells. The protocol requires 1 day for performing in situ targeted methylation, 1-5 days for library preparation depending on desired fragment length, and 1-3 days for Nanopore sequencing depending on desired sequencing depth. The protocol requires basic molecular biology skills and equipment, as well as access to a Nanopore sequencer. We also provide a Python package, *dimelo*, for analysis of DiMeLo-seq data.

4.3 Introduction

Common methods for mapping protein-DNA interactions rely on selective amplification and sequencing of short DNA fragments from regions bound by the protein of interest.[2–7,55] These short-read methods for profiling protein-DNA interactions are powerful and have been used to map the binding patterns of thousands of proteins in human cells.[1] However, because the measurement is on short, amplified fragments of DNA, these methods dissociate joint binding information at neighboring sites, remove endogenous DNA methylation, and are limited in detecting haplotype-specific interactions and interactions in repetitive regions. DiMeLo-seq addresses these limitations by recording protein binding through the deposition of targeted methyladenine (mA) marks that are read out with long-read, single-molecule sequencing (Figure 4.1).[54]

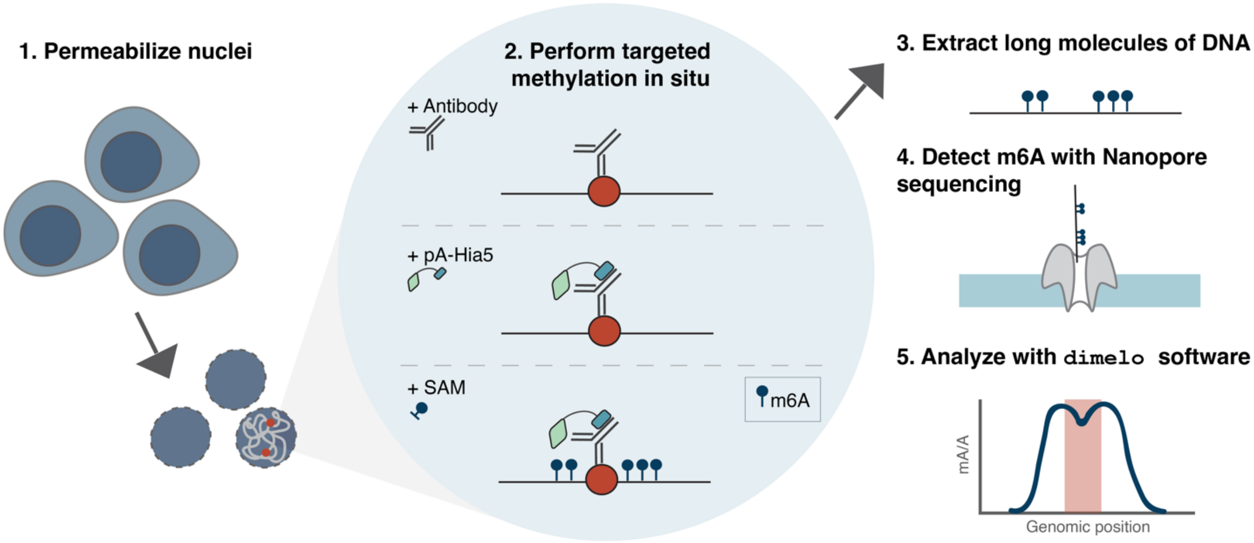

1. Permeabilize nuclei
2. Perform targeted methylation in situ
+ Antibody
+ pA-Hia5
+ SAM
m6A
3. Extract long molecules of DNA
4. Detect m6A with Nanopore sequencing
5. Analyze with dimelo software
mA/A
Genomic position

Figure 4.1 DiMeLo-seq protocol overview. 1. Permeabilize nuclei from fresh, frozen, or fixed cells. 2. Perform a series of steps within the permeabilized nuclei: (i) bind primary antibody to the protein of interest, (ii) bind pA-Hia5 to the primary antibody, (iii) add S-adenosylmethionine (SAM), the methyl donor, to activate methylation. 3. Extract long molecules of DNA. 4. Sequence this DNA with a Nanopore sequencer to detect m6A directly. 5. Analyze modified basecalls from sequencing using the *dimelo* software package.

Development of the protocol

DiMeLo-seq is a method for profiling protein-DNA interactions that relies on in situ antibody-targeted DNA methylation followed by direct readout of methylation with single-molecule, native DNA sequencing. The method was inspired by the targeted methylation strategy used in DamID-seq[7] and builds from short-read techniques for mapping protein-DNA interactions (e.g. CUT&Tag,[6] CUT&RUN,[5] and pA-DamID[61]), as well as recent work that implements long-read sequencing and detection of exogenous methylation to profile chromatin accessibility (Fiber-seq,[24] SMAC-seq,[26] SAMOSA,[25] NanoNOMe[27], MeSMLR-seq[28]). Developing DiMeLo-seq required substantial optimization, with over 100 conditions tested.[54] These optimization experiments revealed critical components that improved efficiency including the following: (1) Hia5 performed significantly better than EcoGII in situ; (2) compared to other detergents for nuclear permeabilization, digitonin and Tween-20 dramatically increased methylation levels; and (3) a low salt concentration, including BSA, increasing incubation time, and replenishing the methyl donor during activation all improved methylation levels.

Applications of the method

DiMeLo-seq can be used to profile the genome localization of any DNA-binding protein for which there is a specific, high-quality antibody. The protocol can be used on fresh, fixed, or frozen cells from culture or from primary tissue. Because DiMeLo-seq uses antibody-based targeting, DiMeLo-seq is also able to profile post-translational modifications like protein acetylation, methylation, and phosphorylation. In our previous work,[54] we demonstrated application of DiMeLo-seq for targeting LMNB1, CTCF, H3K9me3, and CENPA in cultured human cells and here we applied DiMeLo-seq to profile H3K27ac, H3K27me3, and H3K4me3 in cultured human cells, and H3K9me3 in *D. melanogaster* embryos.

Comparison with other methods

The key distinguishing feature of DiMeLo-seq compared to other methods is that protein-DNA interactions are measured on native, long molecules of DNA. Native DNA molecules allow for binding assessment in the context of endogenous CpG methylation. Long reads facilitate interaction mapping in highly repetitive regions of the genome, measurement of multiple binding events on the same chromatin fiber, and haplotype-specific interaction detection.

ChIP-seq, CUT&RUN, CUT&Tag, and DamID-seq all rely on amplification of short sequencing reads. These methods use coverage as a proxy for binding, and resolution is determined by the size of the fragments sequenced. With DiMeLo-seq, the resolution is tied to adenine density and the reach of the methyltransferase. Short reads often preclude haplotype phasing, mapping to repetitive regions, and measuring coordinated binding. These short-read methods require amplification, thereby losing the endogenous mCpG marks. Joint protein binding and mCpG measurement can be done with BisChIP-Seq/ChIP-BS-Seq, but this requires lossy and harsh bisulfite conversion that degrades DNA.[19,20] Similar to ChIP-seq, CUT&Tag, and CUT&RUN, DiMeLo-seq is compatible with primary cells and can be used to target post-translational modifications.

DiMeLo-seq requires substantial input to generate sufficient material for sequencing because there is no amplification, so typical experiments have ~1M cells as input. With short-read methods such as CUT&Tag and DamID-seq, protein-bound regions are selectively amplified and sequenced, thereby enriching for protein-bound regions in the final library for sequencing and allowing for input as low as a single cell. Without enrichment for regions of interest, DiMeLo-seq will sequence the whole genome uniformly, requiring deep sequencing to achieve sufficient coverage of specific target regions. However, there are options to enrich for regions of interest with DiMeLo-seq like AlphaHOR-RES,[54] the Oxford Nanopore Technologies Cas9-based targeted library preparation kit (SQK-CS9109), or m6A-IP as in MadID.[54,65]

Experimental design

While the standard DiMeLo-seq protocol described here has performed consistently for all cell types tested, application to other cell types and primary tissue may require tuning of digitonin concentration. A digitonin concentration of 0.02% has worked well for human GM12878, HG002, Hap1, HEK293T, and Drosophila S2. The optimal digitonin concentration can be determined using

Trypan Blue (Figure 4.2a). Primary tissue also requires upstream processing for nuclear extraction before the nuclear permeabilization step (Supplementary Methods: DiMeLo-seq with *D. melanogaster* embryos). It is important that the nuclear extraction method does not contain NP-40, as we have found this detergent can significantly reduce methylation.

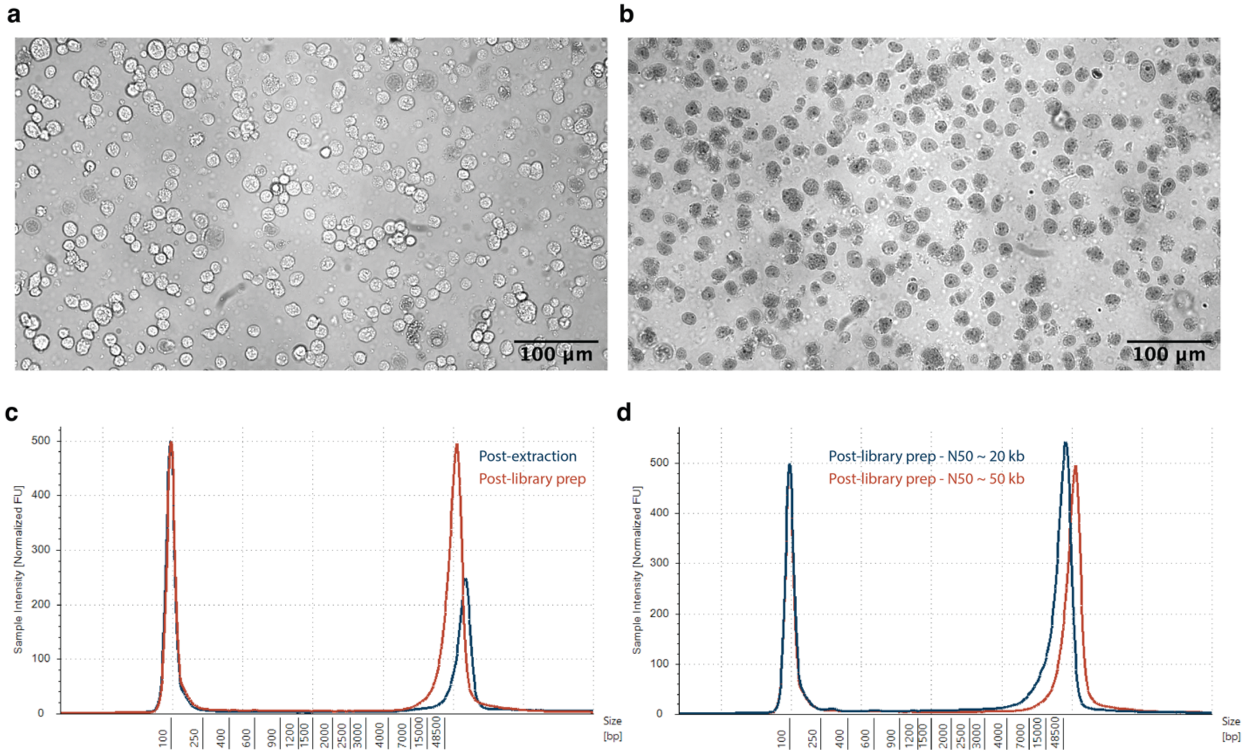


Figure 4.2 Experimental quality control. **a-b,** To determine successful permeabilization, cells are stained with Trypan Blue before (a) and after (b) digitonin treatment. Successful permeabilization allows Trypan Blue to enter the nuclei, while still maintaining high recovery of nuclei from cells. Over-permeabilization results in lower recovery of nuclei. Under-permeabilization does not allow Trypan Blue to enter the nuclei. **c-d,** After the DiMeLo-seq in situ protocol and DNA extraction, DNA is sized using the TapeStation. Representative traces from ligation-based library preparation are show in (c) for the fragment size distribution after extraction and after library preparation. In (d), the size distribution after library preparation for the two ligation-based methods presented in this protocol are shown. The blue curve results in N50 ~20 kb, while the red curve results in N50 ~50 kb. Larger fragment sizes can be achieved with other ultra long kits.

Key variables to optimize for a new target protein are the antibody concentration and the extent of fixation for targets with low binding affinity. An antibody dilution of 1:50 has worked well for all targets reported here and in Altemose et al.[54] Extensive washes are performed after antibody binding to remove any unbound antibody, making excess antibody less detrimental. We have demonstrated that light fixation is compatible with the DiMeLo-seq workflow. If targeting a protein that binds transiently, including fixation may improve signal by preventing the protein from dissociating from the DNA during the DiMeLo-seq protocol.

Typical controls include an IgG isotype control and a free pA-Hia5 control. The IgG isotype control measures nonspecific antibody binding. The free pA-Hia5 control measures chromatin accessibility, similar to Fiber-seq and related methods, and is analogous to the Dam only control used in DamID-seq.[24,37] This control is performed by excluding the primary antibody and pA-Hia5 binding steps and instead adding pA-Hia5 at activation at 200 nM. While these controls are not required, they provide a useful measure of background methylation and bias caused by variable chromatin accessibility. Excluding pA-Hia5 as a control to account for modified basecalling errors can also be included. If troubleshooting a DiMeLo-seq experiment, using one of the antibodies and cell lines validated here and in Altemose et al.[54] may also be a useful control.

Commercially available kits and techniques for DNA extraction, library preparation, and sequencing are rapidly improving. The protocol described here produces consistent localization profiles shown below and in Altemose et al.,[54] but it is important to note that after the in situ methylation steps, any DNA extraction method, library preparation kit, flow cell chemistry, and sequencing device can be used as long as m6A is maintained (no amplification is performed) and the flow cell and device have basecalling models available for calling m6A. We have also demonstrated sequencing of DiMeLo-seq samples with Pacific Biosciences's Sequel IIe.[54]

The key considerations for sequencing are fragment size and sequencing depth. The target N50 (half of sequenced bases are from a fragment size of N50 or larger) varies by application. Longer reads may be desired when mapping to repetitive regions, probing coordinated binding events at longer distances, or phasing reads, longer reads may be desired. With ligation-based library preparation we typically target an N50 of ~20 kb to ~50 kb, which results in fragment size distributions as in Figure 4.2b. With other library preparation kits (e.g., SQK-ULK001), much larger fragments can be sequenced; however, there is a tradeoff between fragment length and throughput. The target sequencing depth also varies by application and will depend on the binding footprint of the protein, the mappability of the region of interest, and the biological question at hand. Final libraries can be saved and flow cells can be reloaded, so it is recommended to do an initial pilot run with shallow sequencing followed by deeper sequencing as needed. For example, for initial tests of new protein targets in human cells, we typically sequence to ~1-3 Gb, or 0.3-1X coverage, to validate and determine optimal experimental conditions, and then sequence more deeply to ~5-45X coverage depending on the analysis we are performing. See the sequencing saturation analysis with CTCF-targeted DiMeLo-seq in Altemose et al.[54]

Expertise needed to implement the protocol

To perform DiMeLo-seq and analyze the data produced, basic molecular biology skills and basic command line skills are required.

Limitations

The performance of DiMeLo-seq is strongly dependent on the quality of the antibody used to target the protein of interest. For proteins that do not have a specific, high-quality antibody compatible with protein A, one could consider epitope tagging or performing in vivo expression of a protein-MTase fusion.[111] DNA must be accessible for Hia5 to methylate in situ. Thus, targets in less accessible regions of the genome may require longer incubations or deeper sequencing. Hia5 may methylate DNA in trans if close enough to the target protein in 3D space. DiMeLo-seq experiments typically require ~1M cells as input, although Concanavalin A beads (which we previously showed are compatible with DiMeLo-seq) and lower-input library preparation kits can reduce required input material.[54]

Here and in our previous study, we have benchmarked DiMeLo-seq's performance in targeting LMNB1, CTCF, H3K9me3, CENPA, H3K27ac, H3K27me3, and H3K4me3.[54] When targeting CTCF and LMNB1, we estimated 54% and 59% sensitivity (94% specificity), but this is dependent on the protein, antibody, and chromatin environment and must be evaluated for new targets.[54] Transiently bound proteins may benefit from the optional fixation step at the start of the DiMeLo-seq protocol.

4.4 Materials

REAGENTS

A. Reagents for in situ protocol

- HEPES-KOH 1 Molarity (M) pH 7.5 (Boston BioProducts BBH-75-K)
- NaCl 5 M (Sigma-Aldrich 59222C-500ML)
- Spermidine 6.4 M (Sigma-Aldrich S0266-5G)
- Roche cOmplete™ EDTA-free Protease Inhibitor Tablet (Sigma-Aldrich 11873580001)
- Bovine Serum Albumin (Sigma-Aldrich A6003-25G)
- Digitonin (Sigma-Aldrich 300410-250MG)
  CAUTION acute toxic and health hazard; work in fume hood when making digitonin solution.
- Tween-20 (Sigma-Aldrich P7949-100ML)
- Tris-HCl 1M pH 8.0 (Invitrogen 15568025)
- KCl (Sigma-Aldrich PX1405-1)
- EDTA 0.5 M pH 8.0 (Invitrogen 15575-038)
- EGTA 0.5 M pH 8.0 (Fisher 50-255-956)
- S-Adenosylmethionine 32 mM (NEB B9003S)
- PFA, 16% (if performing fixation) (EMS 15710)
- Glycine (if performing fixation) (Fisher BP381-1)
- Eppendorf DNA LoBind tubes 1.5 mL (Fisher 022431021)
- Wide bore 200 µl and 1000 µl tips (e.g., USA Scientific 1011-8810, VWR 89049-168)
- pA-Hia5 (see https://www.protocols.io/view/pa-hia5-protein-expression-and-purification-x54v9j56mg3e/v1 for expression and purification protocol. The pET-pA–Hia5 (pA-Hia5) plasmid is available from Addgene (cat no. 174372)).
- Primary antibody for protein target of interest, from species compatible with pA (e.g., Abcam ab16048)
- Secondary antibody for immunofluorescence quality control (e.g., Abcam ab3554)
- Trypan Blue (Fisher T10282)
- Qubit dsDNA BR Assay Kit (Fisher Q32850)
- Qubit Protein Assay Kit (Fisher Q33211)

B. Reagents for extraction, library preparation, and sequencing

N.B. We have validated the following reagents, but extraction, library preparation, and sequencing reagents are improving rapidly. The important considerations are to choose a DNA extraction method that maintains long DNA molecules, to perform amplification-free library preparation, and to use a flow cell that is compatible with mA calling.

- Monarch Genomic DNA Purification Kit (NEB T3010S)
- Monarch HMW DNA Extraction Kit (NEB T3050L)
- Agencourt AMPure XP beads (Beckman Coulter A63881)
- Blunt/TA Ligase Master Mix (NEB M0367S)
- NEBNext quick ligation module (NEB E6056S)
- NEBNext End Repair dA-tailing Module (NEB E7546S)
- NEBNext FFPE DNA repair kit (NEB M6630S)
- Ligation Sequencing Kit (ON SQK-LSK109, ON SQK-LSK110, or latest kit compatible with m6A calling)
- Native Barcoding Expansion 1-12 (ON EXP-NBD104, or latest kit compatible with m6A calling)
- Native Barcoding Expansion 13-24 (ON EXP-NBD114, or latest kit compatible with m6A calling)
- Circulomics Short Read Eliminator Kit (SS-100-101-01)
- Flow Cell Wash Kit (ON EXP-WSH004)
- Flow cells (ON FLO-MIN106D or ON FLO-PRO002, or latest flow cells compatible with m6A calling)

EQUIPMENT

- Centrifuge that can hold 4ºC
- Rotator (e.g., Millipore Sigma Z740289)
- Oxford Nanopore Technologies Nanopore sequencer (e.g., MIN-101B)
- Magnetic separation rack (if targeting N50 ~20 kb) (e.g., NEB S1515S)
- Qubit (e.g., Thermo Fisher Scientific Q33238)
- Tapestation (not required; for quality control) (e.g., Agilent G2992AA)
- Microscope (not required; for quality control)

REAGENT SETUP

A. Buffer preparation

Prepare all buffers fresh, filter buffers through a 0.2 µm filter, and keep buffers on ice.

Digitonin

Solubilize digitonin in preheated 95ºC Milli-Q water to create a 5% digitonin solution (e.g., 10 mg/200 µl).

Wash Buffer

Prepare wash buffer according to the following table.

Component	Amount	Final Concentration
HEPES-KOH, 1 M, pH 7.5	1 ml	20 mM
NaCl, 5 M	1.5 ml	150 mM
Spermidine, 6.4 M	3.91 µl	0.5 mM
Roche Complete tablet -EDTA	1 tablet	-
BSA	50 mg	0.1%
H2O	up to 50 ml	-

Dig-Wash Buffer

Add 0.02% digitonin to wash buffer. For example, add 20 µl of 5% digitonin solution to 5 ml wash buffer. The optimal concentration of digitonin may vary by cell type.

Tween-Wash Buffer

Add 0.1% Tween-20 to wash buffer. For example, add 50 µl Tween-20 to 50 ml wash buffer.

Activation Buffer

Prepare the activation buffer but wait to add SAM until the activation step.

Component	Amount	Concentration
Tris, pH 8.0 1 M	750 µL	15 mM
NaCl 5 M	150 µL	15 mM
KCl 1 M	3 mL	60 mM
EDTA, pH 8.0 0.5 M	100 µL	1 mM
EGTA, pH 8.0 0.5 M	50 µL	0.5 mM
Spermidine, 6.4 M	0.391 µL*	0.05 mM
BSA	50 mg	0.1%
H2O	up to 50 mL	-
SAM, 32 mM	(add at activation step)	800 µM

*To reduce pipetting error, first perform a 1:10 dilution of Spermidine in H2O by adding 1µLof 6.4 M Spermidine to 9 µL H2O. Mix well and then add 3.91 µL of this dilution to the Activation Buffer.

4.5 Procedure

General notes

- The protocol will be kept up-to-date at: https://www.protocols.io/view/dimelo-seq-directed-methylation-with-long-read-seq-n2bvjxe4wlk5/v2
- All spins are at 4°C for 3 minutes at 500 g.
- Spinning in swinging bucket rotor can help pellet the nuclei.
- To prevent nuclei from lining the side of the tube, break all spins into two parts: 2 minutes with the tube hinge facing inward, followed by 1 minute with the tube hinge facing outward. This two-part spin is not needed if using a swinging bucket rotor.
- Working with Eppendorf DNA LoBind tubes can reduce loss of material.
- Use wide bore tips when working with nuclei.
- Do not use NP-40 or Triton-X100 for nuclear extraction, permeabilization, or any other stage of the protocol, as they appear to dramatically reduce methylation activity.
- The best digitonin concentration may vary by cell type. For HEK293T, GM12878, HG002, Hap1, and S2 cells, 0.02% works well. You can test different concentrations of digitonin and verify permeabilization and nuclear integrity by Trypan blue staining. For example, you may try 0.02% to 0.1% digitonin.
- We use Tween to reduce hydrophilic non-specific interactions and BSA to reduce hydrophobic non-specific interactions. We also found that including BSA at the activation step significantly increases methylation activity as well.
- The best primary antibody concentration may vary by protein target of interest. A 1:50 dilution works well for targeting LMNB1, CTCF, and histone modifications, and is likely a good starting point for most antibodies.
- Binding a secondary antibody after the primary antibody but before pA-Hia5 reduced total methylation and specificity. Including a secondary antibody binding step is not recommended. For pA-incompatible antibodies, a secondary antibody can be used as a bridging antibody, but performance is diminished; instead, we recommend using pAG-Hia5 for pA-incompatible antibodies.

(Optional fixation)
TIMING 10 minutes

1. Resuspend cells in PBS (1 million to 5 million cells per condition).
2. Add PFA to 0.1% (e.g., 6.2 µl of 16% PFA to 1 ml cells) for 2 minutes while gently vortexing.
3. Add 1.25 M glycine (sterile; 0.938 g in 10 ml) to twice the molar concentration of PFA to stop the crosslinking (e.g., 60 µl of 1.25 M glycine to 1 ml).
4. Centrifuge 3 minutes at 500 x g at 4°C and remove the supernatant.

5. Continue with Nuclear Isolation, starting with step 8.

Nuclear isolation

TIMING 15 minutes

6. Prepare cells (1M-5M per condition).
7. Wash cells in PBS. Spin and remove supernatant.
8. Resuspend cells in 1 ml Dig-Wash buffer. Incubate for 5 minutes on ice.
   TROUBLESHOOTING
9. Split nuclei suspension into separate tubes for each condition.
10. Spin and remove supernatant.
11. Quality control: Check permeabilization was successful by taking 1 µl of the nuclei following the 5-minute incubation on ice, diluting to 10 µl with PBS, and staining with Trypan Blue.

Primary antibody binding

TIMING 2.5 hours

12. Gently resolve each pellet in 200 µl Tween-Wash containing primary antibody at 1:50 or the optimal dilution for your antibody and target.
13. Place on rotator at 4°C for ~2 hr.
    PAUSE POINT - Samples can be left overnight on the rotator at 4°C.
14. Spin and remove supernatant.
15. Wash twice with 0.95 ml Tween-Wash. For each wash, gently and completely resolve the pellet. This may take pipetting up and down ~10 times. Following resuspension, place on rotator at 4°C for 5 minutes before spinning down.

Quantify pA-Hia5 concentration

TIMING 30 minutes

16. Thaw protein from -80°C at room temperature and then move to ice immediately.
17. Spin at 4°C for 10 minutes at 10,000 x g or higher to remove aggregates.
18. Transfer the supernatant to a new tube and save it, discarding the previous tube.
19. Use Qubit with 2 µl sample volume to quantify protein concentration.

pA-Hia5 binding

TIMING 2.5 hours

20. Gently resolve pellet in 200 µl Tween-Wash containing 200 nM pA-Hia5.
21. Place on rotator at 4°C for ~2 hr.
22. Spin and remove supernatant.
23. Wash twice with 0.95 ml Tween-Wash. For each wash, gently and completely resolve the pellet. Following resuspension, place on rotator at 4°C for 5 minutes before spinning down.

Quality control (optional)

TIMING 1 hour

24. Add 1.6 µl of 16% PFA to 25 µl of nuclei in Tween-Wash (taken from the 0.95 ml final wash) for 1% total PFA concentration.
25. Incubate at room temperature for 5 minutes.
26. Add 975 µl of Tween-Wash to stop the fixation by dilution.
27. Add 1 µl fluorophore-conjugated secondary antibody.
28. Put on rotator for 30 minutes at room temperature, protected from light.
29. Wash 2 times (or just once). Pellet likely won't be visible.
30. Resuspend in mounting media after last wash. Use as little as possible, ideally 5 µl.
31. Put 5 µl on a slide, make sure there are no bubbles, and put on a coverslip.
32. Seal with nail polish along the edges.
33. Image or put at -20°C once the nail polish has dried.
    TROUBLESHOOTING

Activation

TIMING 2.5 hours

34. Gently resolve pellet in 100 µl of Activation Buffer per sample. Be sure to add SAM to a final concentration of 800 µM in the activation buffer at this step! In 100 µl of Activation Buffer, this means adding 2.5 µl of the SAM stock that is at 32 mM.
35. Incubate at 37°C for 2 hours. Replenish SAM by adding an additional 800 µM at 1 hour. This means adding an additional 2.5 µl of the SAM stock that is at 32 mM to the 100 µl reaction. Pipet mix every 30 minutes. Tapping to mix also works.
36. Spin and remove supernatant.
37. Resuspend in 100 µl cold PBS.
38. Check nuclei by Trypan blue staining to determine recovery and check integrity of nuclei if desired.

Depending on desired fragment size, either follow Method A for N50 ~20 kb or Method B for N50 ~50 kb.

N.B. We have validated the following reagents, but extraction, library preparation, and sequencing reagents are improving rapidly. The important considerations are to choose a DNA extraction method that maintains long DNA molecules, to perform amplification-free library preparation, and to use a flow cell that is compatible with mA calling. These are workflows we have validated and modifications we have made.

A. DNA extraction for N50 of ~20 kb

TIMING 1 hour

39. Use the Monarch Genomic DNA Purification Kit. Follow protocol for genomic DNA isolation using cell lysis buffer. Include RNase A. NB. If fixation was performed, be sure to do the 56°C incubation for lysis for 1 hour (not just 5 minutes) to reverse crosslinks.
40. Perform two elutions: 100 µl and then 35µl.
    PAUSE POINT - Samples can be stored at 4°C or -20°C.
41. Quantify DNA yield by Qubit dsDNA BR Assay Kit.
42. Concentrate by speedvac if necessary for 1-3 µg DNA in 48 µl for input to library prep.

B. DNA extraction for N50 ~50 kb

TIMING 1 hour

43. Use the NEB Monarch HMW DNA Extraction Kit. Follow protocol for genomic DNA isolation using cell lysis buffer. Include RNase A. Perform lysis with 2000 rpm agitation. We have validated 2000 rpm gives N50 ~50-70 kb but if longer reads are desired we expect 300 rpm would work. Apart from using a different kit, all of the steps for the long fragment DNA extraction are the same as the general protocol. To reiterate, make the following changes to the protocol outlined in the following steps. If fixation was performed, be sure to do the 56°C incubation for lysis for 1 hour (not just 10 minutes) to reverse crosslinks. Agitate for 10 minutes and then keep at 56ºC without agitation for 50 minutes.

    PAUSE POINT - Samples can be stored at 4°C or -20°C.
44. Quantify DNA yield by Qubit dsDNA BR Assay Kit.
45. Concentrate by speedvac if necessary to obtain 1-3 µg DNA in 48 µl for input to library prep.

(Optional enrichment)

46. If the sequencing cost and time for sufficient coverage becomes prohibitive, a few enrichment strategies can be used. A restriction enzyme-based approach like AlphaHOR-RES relies on preferential digestion of DNA outside of target regions followed by size selection to maintain larger on-target fragments.[54] The ONT Cas9 Sequencing Kit (SQC-CS9109) is another option to selectively ligate adapters to targeted regions during library preparation. To enrich for methylated regions, m6A-specific immunoprecipitation can be used, as in MadID.[65]

A. Library preparation & sequencing for N50 ~20 kb

TIMING 3 hours

47. If multiplexing samples on a flow cell, follow Nanopore protocol for Native Barcoding Ligation Kit 1-12 and Native Barcoding Ligation Kit 13-24 with ON SQK-LSK109. If not multiplexing, use ON SQK-LSK110. We recommend the following modifications:
    a. Load ~3 µg DNA into end repair.
    b. Incubate for 10 minutes at 20°C for end repair instead of 5 minutes.
    c. Load ~ 1 µg of end repaired DNA into barcode ligation.
    d. Double the ligation incubation time(s) to at least 20 minutes.
    e. Elute in 18 µl instead of 26 µl following barcode ligation reaction cleanup to allow for more material to be loaded into the final ligation.
    f. Load ~3 µg of pooled barcoded material into the final ligation. If needed, concentrate using speedvac to be able to load 3 µg into the final ligation.
    g. Perform final elution in 13 µl EB. Take out 1 µl to dilute 1:5 for quantification by Qubit (and size distribution analysis by TapeStation / Bioanalyzer if desired).
    h. Load ~1 µg of DNA onto the sequencer. Input requirements vary by sequencing kit and are becoming lower.

TROUBLESHOOTING

B. Library preparation & sequencing for N50 ~50 kb

TIMING 5 days

48. Follow Nanopore protocol for ON SQK-LSK110 (method validated with this kit only, not with multiplexing with ON SQK-LSK109) with the following modifications (inspired by Kim et al, dx.doi.org/10.17504/protocols.io.bdfqi3mw):[94]
    a. Increase end preparation time to 1 hour with a 30-minute deactivation.
    b. Following end preparation, perform a cleanup by combining 60 µL SRE buffer from Circulomics (SS-100-101-01) with the 60 µL end prep reaction.
    c. Centrifuge this reaction at 10,000 x g at room temperature for 30 minutes (or until DNA has pelleted).
    d. Wash pelleted DNA with 150 µL of 70% ethanol two times, using a 2 minute spin at 10,000 x g between washes.
    e. Resuspend the pellet in 31 µL EB.
    f. Incubate at 50ºC for 1 hour. Incubate at 4ºC for at least 48 hours.
    g. For the ligation step, reduce ligation volume by half (total of 30 µL DNA in a 50 µL reaction volume). Increase the ligation incubation to 1 hour.
    h. Pellet DNA at 10,000 x g at room temperature for 30 minutes.
    i. Wash the pellet twice with 100 µL LFB, using a 2 minute spin at 10,000 x g between washes.
    j. Resuspend the pellet in 31 µL EB.
    k. Incubate at least 48 hours at 4ºC.
    l. Load 500 ng of DNA onto the sequencer. Input requirements vary by sequencing kit and are becoming lower.
    m. If you see the number of active pores has dropped considerably after 24 hours, you can recover pore activity using the flow cell wash kit, then loading additional library material.

TROUBLESHOOTING

Analysis

After sequencing, the raw output files produced by the Nanopore sequencer must be converted to BAM files for input to a Python package called *dimelo* that we have created for analysis of DiMeLo-seq data (https://github.com/streetslab/dimelo). Recommendations for the basecalling and alignment steps, which will create an aligned BAM file with "Mm" and "Ml" tags that describe methylation calls, can be found in the package documentation (https://streetslab.github.io/dimelo/). Basecalling is being rapidly improved by ONT and others, so basecalling suggestions are likely to become outdated quickly. After basecalling and alignment, the resulting BAM file is the input to the quality control, visualization, and custom analysis functions from the *dimelo* software package. This analysis software can either be run as an imported Python module or can be run from the command line. A summary of the functions is in Figure 4.3.

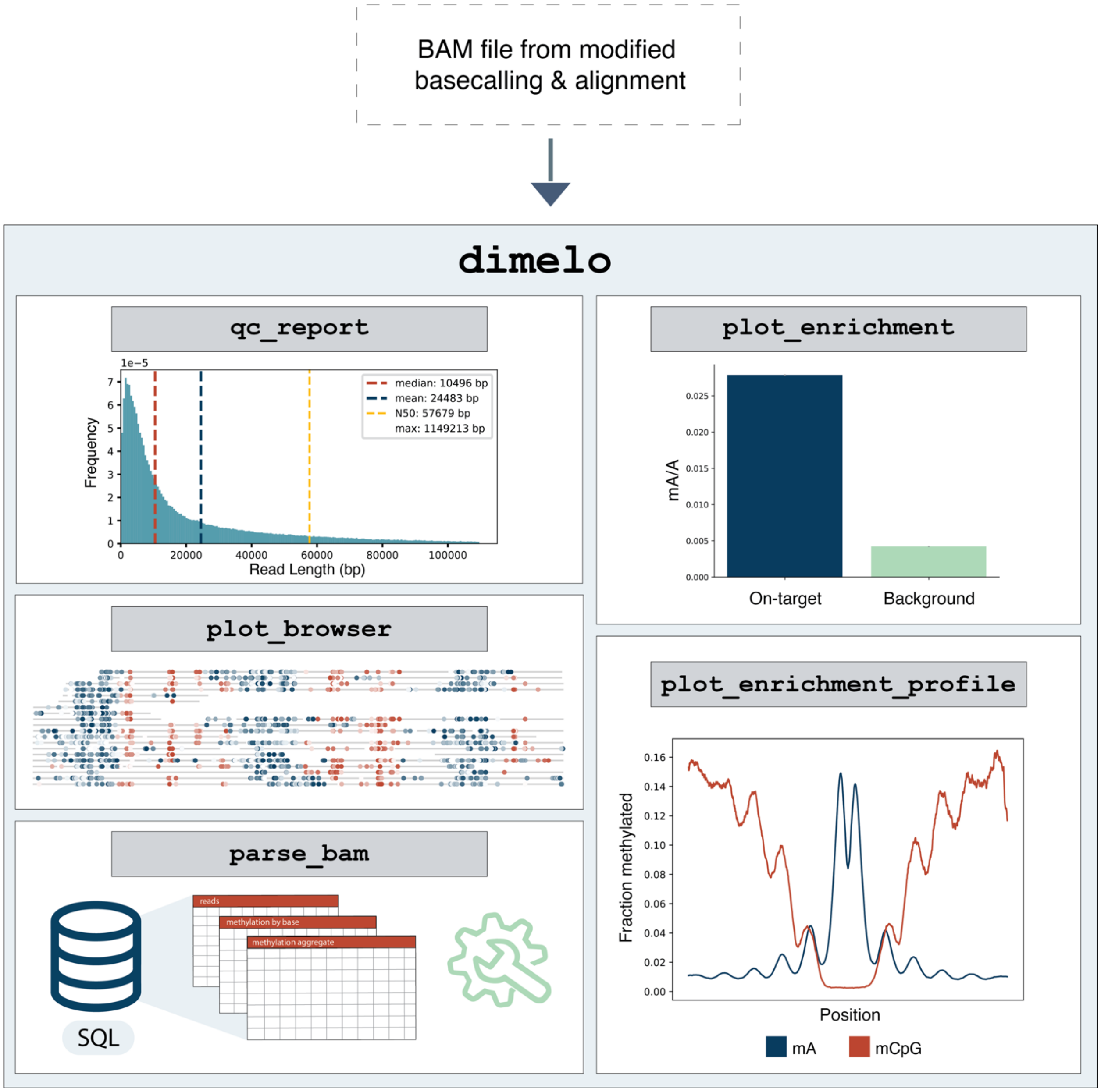
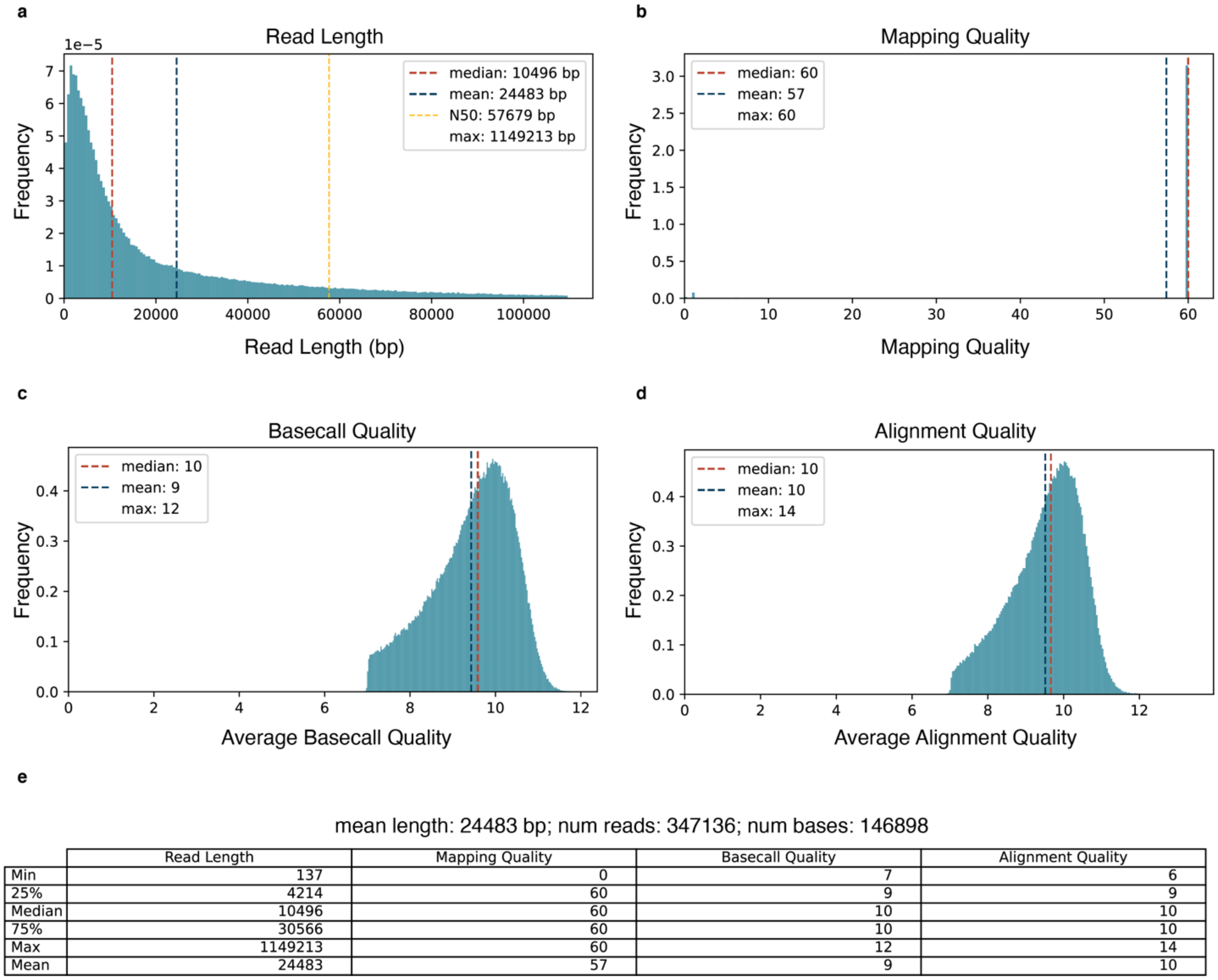


Figure 4.3 Analysis pipeline overview. Basecalling and alignment are performed on the fast5 output from the Nanopore sequencer. The resulting bam that contains the modified base information is then input to the *dimelo* software package. A recommended workflow involves quality control with `qc_report` followed by visualization with `plot_browser`, `plot_enrichment`, and `plot_enrichment_profile`. For custom analysis, `parse_bam` creates a SQL database with base modification calls in a format that makes it easier to manipulate for downstream analysis.



mean length: 24483 bp; num reads: 347136; num bases: 146898

	Read Length	Mapping Quality	Basecall Quality	Alignment Quality
Min	137	0	7	6
25%	4214	60	9	9
Median	10496	60	10	10
75%	30566	60	10	10
Max	1149213	60	12	14
Mean	24483	57	9	10

Figure 4.4 Sequencing quality control. The `qc_report` function takes in one or more bam files and for each, outputs a QC report including the following 5 features. **a,** Histogram of read lengths with the median, mean, N50, and max value annotated. **b,** Histogram of mapping quality. **c-d,** Both the basecall quality and alignment quality scores are present in bam outputs from Guppy but not from Megalodon. **c,** Histogram of average basecall quality per read. Here, the mean indicates our sample's average basecall quality is Q10 which is equivalent to 90% accuracy. **d,** Histogram of average alignment quality per read. While mapping quality provides the accuracy of the read mapping to specific genomic coordinates, the average alignment quality provides the quality of matching between the read and the reference sequence. For example, if a read almost perfectly matches multiple genomic coordinates, it will have a low mapping quality but a high alignment quality. **e,** Summary table with descriptive statistics of each feature (a-d), in addition to highlighting important values such as mean length of reads, total number of reads, and total number of bases sequenced. Example data used in this figure are from targeting H3K9me3 in *D. melanogaster* embryos.

A recommended workflow is to first run `qc_report` to generate summary statistics and histograms for metrics such as coverage, read length, mapping quality, basecall quality, and alignment quality (Figure 4). Next, three functions are provided for visualization. All functions take BAM file(s) as input and region(s) of interest defined as a string or bed file.

The `plot_enrichment` function compares methylation levels across samples or across different genomic regions. This tool is useful for looking at overall on- vs. off-target methylation and for comparing methylation levels in regions of interest across samples.

The `plot_browser` function allows the user to view single molecules with base modifications colored according to the probability of methylation within a region of interest. This function can either produce a static PDF of the single molecules or an interactive HTML file that allows the user to zoom and pan around the browser plot, using plotting code adapted from De Coster et al.[101] Plots of aggregate coverage and the fraction of methylated bases over the window of interest are also generated with this function.

The `plot_enrichment_profile` function creates single-molecule plots and an aggregate plot of the fraction of methylated bases centered at features of interest defined in a bed file. For example, one may enter a bed file with the locations of the binding motif for a given protein or with transcription start site coordinates to view the methylation profiles around these features of interest. Inputting multiple BAM files creates an overlay of the methylation profiles across samples and inputting multiple bed files creates an overlay of methylation profiles for a given sample across the different sets of regions defined in the bed files.

The `parse_bam` function converts the base modification information stored in the BAM file into a SQL database to give users the option to create custom figures or analysis with the data in an easier format to manipulate.

For all functions, the user can specify the modification(s) of interest to extract - “A”, “CG”, or “A+CG”. The probability threshold for calling a base as modified is also a parameter to each function. For discussion of threshold determination see Supplementary Note 6 of Altemose et al.[54]

Timing

<u>N50 ~20 kb</u>
Day 0
Steps 1-38, perform in situ targeted methylation: 10 h
Steps 39-42, DNA extraction: 1 h (2 h if fixation was performed)

Day 1
Step 47, perform library preparation & start sequencing: 3 h

Day 2
Step 47, re-load sequencer if necessary: 1 h

Day 3

Step 47, re-load sequencer if necessary: 1 h

<u>N50 ~50 kb</u>
Day 0
Steps 1-38, perform in situ targeted methylation: 10 h
Steps 43-45, DNA extraction: 2 h (3 h if fixation was performed)

Day 1
Step 48, perform library preparation end repair and clean: 2 h

Day 3
Step 48, perform library preparation ligation and clean: 2 h

Day 5
Step 48, start sequencing

Day 6
Step 48, re-load sequencer if necessary: 1 h

Day 7
Step 48, re-load sequencer if necessary: 1 h

Troubleshooting
Troubleshooting advice can be found in Table 4.1.

Step	Problem	Possible Reason	Solution
8	Few intact nuclei	Digitonin concentration is not optimal for the cell type	Try a range of digitonin concentrations and perform QC with Trypan Blue stain
33	No difference in fluorescence between IgG control and targeted methylation	Target abundance is low and/or target is diffuse	IF may not be a good quality control step for your target
		Insufficient washing	Add another wash step after secondary antibody binding
		Antibody concentration is not optimal	Try a range of primary and secondary antibody concentrations
		Primary or secondary antibody is not working	Try different antibody
		Permeabilization failure	To confirm permeabilization, perform Trypan Blue quality

			control step with varying digitonin concentrations.
47, 48	Unable to pipette viscous DNA	DNA is too long	Fragment DNA or follow library preparation protocol for persevering longer fragments (Step 48)
47	Bead clumping	DNA is too long for bead-based cleanup	Fragment DNA or follow library preparation protocol for preserving longer fragments (Step 48)
47	Low recovery from bead cleanup	DNA is too long for bead-based cleanup	Fragment DNA or follow library preparation protocol for preserving longer fragments (Step 48)
		DNA is too short for long fragment buffer (LFB) used in bead cleanup.	Handle HMW DNA carefully with wide bore tips and ensure your DNA extraction method maintains long DNA fragments.
47, 48	Short reads	DNA sheared during library preparation	Handle HMW DNA carefully with wide bore tips; follow library preparation protocol for preserving longer fragments (Step 48)
		Too much DNA loaded onto sequencer	Repeat qubit of final library. For target N50 ~20 kb, load ~1 µg of library; for target N50 ~50 kb, load 300 - 500 ng of library.
47, 48	Low yield from sequencer	Low input and long DNA fragments cause pores to become inactive quickly	Perform flow cell wash and reload every ~24 hours and/or load more DNA onto the flow cell. Washing and reloading becomes very important with larger fragment sizes.
		Bubbles destroy pores.	Use a new flow cell and be sure not to introduce bubbles during the flow cell loading process.

Table 4.1 Troubleshooting tips.

4.6 Anticipated results

One person can collect and analyze sequencing data within 3-8 days of beginning the DiMeLo-seq protocol. In this section, we show representative data from DiMeLo-seq experiments targeting H3K27ac, H3K27me3, and H3K4me3 in GM12878 cells and H3K9me3 in *D. melanogaster* embryos (Table 4.2). We use these targets to provide example output from the *dimelo* package and include suggested figures to evaluate performance and to perform exploratory analysis with DiMeLo-seq data.

Target	**Cell type**	**Antibody**	**Library prep kit**	**Flow cell chemistry**	**Device**	**Gb**	**Coverage**	**N50 (bp)**
H3K27ac	GM12878	Active Motif 39133	SQK-LSK110	R9.4.1	PromethION	124	41X	25,536
H3K27me3		Active Motif 39055				122	41X	27,226
H3K4me3		Active Motif 39916				124	41X	25,163
H3K9me3	*D. melanogaster* embryo	Active Motif 39062			MinION	8.24	46X	27,843

Table 4.2 Experimental overview. Summary of experimental specifications for histone modifications profiled using DiMeLo-seq.

The specificity and efficiency of methylation vary by target, depending on the antibody quality, how broad the binding domain is, and the chromatin environment, among other factors. The on-target and off-target methylation levels when targeting H3K27ac, H3K27me3, and H3K4me3 with DiMeLo-seq are shown in Figure 4.5a-c. These plots are generated from the `plot_enrichment` function. For H3K27ac, to define on-target regions, we used top ChIP-seq peaks for H3K27ac (ENCODE ENCFF218QBO).[42] For off-target, we used top ChIP-seq peaks for H3K27me3 (ENODE ENCFF119CAV).[42] We similarly analyze on- and off-target for H3K27me3 with H3K27me3 top ChIP-seq peaks for on-target and H3K27ac top ChIP-seq peaks for off-target regions. For H3K4me3, to define on-target regions, we used top ChIP-seq peaks for H3K4me3 (ENCODE ENCFF228TWF);[42] for off-target, we used transcription start sites of unexpressed genes where H3K4me3 is not expected to be present. The on-target methylation levels are higher for H3K27me3 compared to H3K27ac, despite H3K27me3 being a repressive mark in a less accessible genomic context. This is likely because it binds a broader genomic region, allowing a larger methylated footprint. The performance difference could also occur if the anti-H3K27me3 antibody performs better than the anti-H3K27ac antibody used in these experiments. The off-target methylation level is also higher in H3K27me3 compared to H3K27ac. This is likely because the off-target region used in this analysis is H3K27ac ChIP-seq peaks, which are in very accessible regions of the genome, and off-target methylation with DiMeLo-seq occurs preferentially within open chromatin. Again, higher off-target methylation can also be caused by differences in antibody performance.

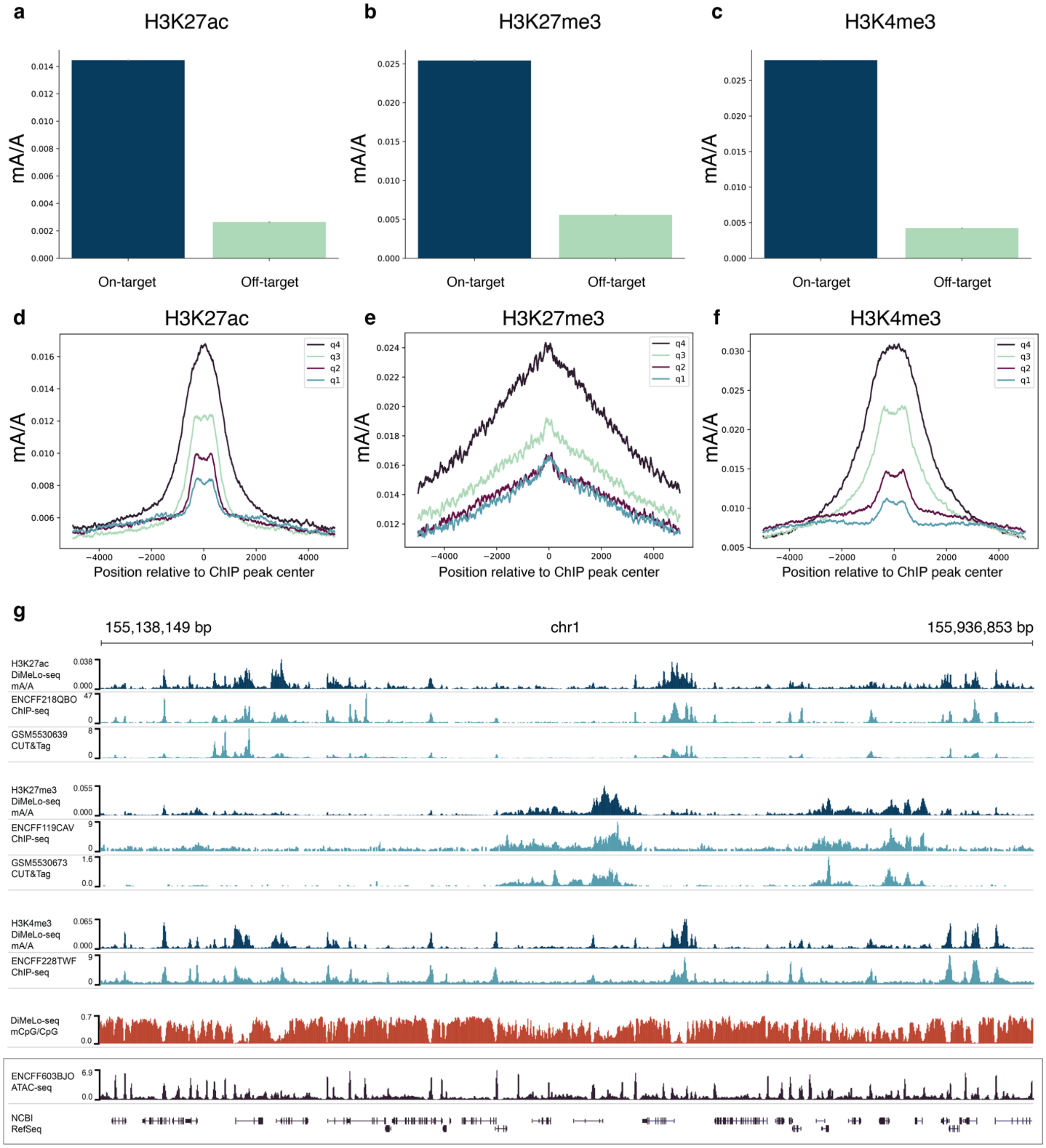


Figure 4.5 Validation of targeted methylation. **a-c,** Using bed files defining on- and off-target regions, the `plot_enrichment` function can be used to determine whether methylation is concentrated within expected regions. We've defined on-target regions using ChIP-seq peaks for

the corresponding histone marks. We defined off-target regions when targeting H3K27ac as H3K27me3 ChIP-seq peaks and when targeting H3K27me3 as H3K27ac ChIP-seq peaks; for off-target regions for H3K4me3 we use transcription start sites for unexpressed genes. Methylation probability threshold of 0.75 was used. Error bars represent 95% credible intervals determined for each ratio by sampling from posterior beta distributions computed with uninformative priors. **d-f,** Methylation profiles centered at ChIP-seq peaks for H3K27ac-, H3K27me3-, and H3K4me3-targeted DiMeLo-seq are plotted using `plot_enrichment_profile`. The quartiles (q4 to q1) indicate the strength of the ChIP-seq peaks which the DiMeLo-seq reads overlap. Methylation probability threshold of 0.75 was used. **g,** Aggregate browser traces comparing DiMeLo-seq signal to ChIP-seq and CUT&Tag. BED files used for creating aggregate curves are generated either from `parse_bam` or `plot_browser`. CpG methylation signal is aggregated from the H3K27ac-, H3K27me3-, and H3K4me3-targeted DiMeLo-seq experiments. Methylation probability threshold of 0.8 was used. ATAC-seq and NCBI RefSeq annotations are also shown.

The methylation profile centered at features of interest can be visualized using the `plot_enrichment_profile` function. Here, we show profiles from H3K27ac-, H3K27me3-, and H3K4me3-targeted DiMeLo-seq with aggregate methylation curves from reads centered at ChIP-seq peaks of varying strength (Figure 4.5d-f) (ENCODE ENCFF218QBO, ENCFF119CAV, ENCFF228TWF).[42] H3K27ac and H3K4me3 have narrow peaks, while H3K27me3 has a broader peak. Signals for all three marks track with ChIP-seq peak strength, indicating concordance between DiMeLo-seq and ChIP-seq in aggregate.

To further analyze the concordance between DiMeLo-seq and other methods for measuring protein-DNA interactions - here ChIP-seq and CUT&Tag - we created aggregate browser tracks across a stretch of chromosome 1 (Figure 4.5g) (ENCODE ENCFF218QBO, ENCFF119CAV, ENCFF228TWF; GEO GSM5530639, GSM5530673).[42,112] DiMeLo-seq signal for all three histone marks tracks with ChIP-seq and CUT&Tag profiles. These curves were generated using the bed file output from the `plot_browser` function with smoothing in a 100 bp window. DiMeLo-seq also measures endogenous CpG methylation together with protein binding. An aggregate mCpG signal from the three DiMeLo-seq samples is shown, and dips in mCpG are evident where H3K27ac and H3K4me3 signals are highest. H3K27ac and H3K4me3 are both marks of open chromatin and have peaks overlapping accumulations in ATAC-seq signal (ENCODE ENCFF603BJO).[42]

In addition to comparing DiMeLo-seq to other methods, we also evaluated methylation profiles around genomic features where our targets are expected to localize. In particular, both H3K27ac and H3K4me3 are found at transcription start sites (TSSs).[113] Using the `plot_enrichment_profile` function, we created the aggregate methylation and single-molecule methylation plots shown in Figure 4.6a. As expected, both marks have enrichment at the TSSs, with the highest methylation levels at the TSSs for the genes with highest expression.[113] The periodicity from positions 0 bp to 500-1000 bp with respect to the TSS indicate preferential methylation of linker DNA between strongly positioned nucleosomes downstream from the TSSs for both targets. For genes that are not expressed (quartile 1), no significant enrichment at TSSs is evident.

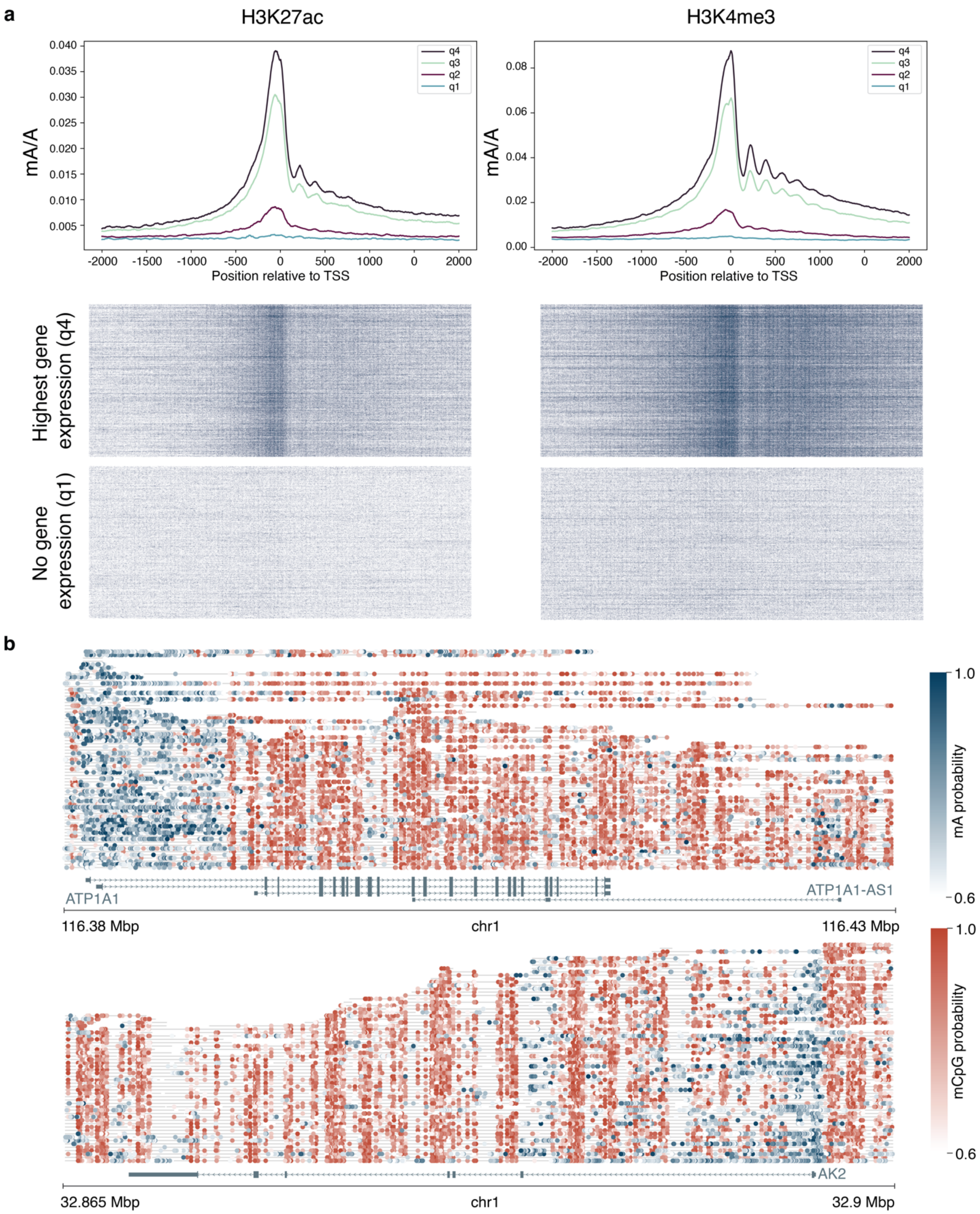


Figure 4.6 Evaluating protein binding at regions of interest. Both H3K27ac and H3K4me3 are found at transcription start sites (TSS). **a,** Signal from H3K27ac- and H3K4me3-targeted DiMeLo-seq at TSS. Reads overlapping TSS, gated by gene expression level from highest gene expression

(quartile 4 (q4)) to lowest gene expression (quartile 1 (q1)). Aggregate mA/A profiles are shown for all reads spanning these TSSs. Single molecules are shown below with blue representing mA calls for TSS for the highest gene expression (q4) and for no gene expression (q1). Aggregate and single-molecule plots were produced with `plot_enrichment_profile`. Methylation probability threshold of 0.75 was used. **b,** Single-molecule browser plots produced from `plot_browser` from H3K4me3-targeted DiMeLo-seq experiment. Each grey line represents a read, blue circles indicate mA, and red circles indicate mCpG. NCBI RefSeq genes are shown below. Methylation probability threshold of 0.6 was used.

Using the `plot_browser` function, single molecules are shown from H3K4me3-targeted DiMeLo-seq in a window around a few highly expressed genes in GM12878 (Figure 4.6b). Methylated adenines are enriched around the TSSs for these highly expressed genes ATP1A1 and AK2. Together with mA, the endogenous mCpG can also be analyzed. Here, it is evident that mCpG is depleted in the regions around TSSs where H3K4me3 is enriched, as has been previously reported.[114] Multiple TSSs are spanned by some of the molecules in the region from 116.38 Mbp to 116.43 Mbp on chromosome 1, highlighting DiMeLo-seq's ability to probe multiple binding events on a single molecule.

DiMeLo-seq can be used to target proteins in nuclei isolated not only from cultured cells but also from primary tissue or intact organisms. We mapped H3K9me3 distributions in *D. melanogaster* embryos across the genome and show that averaging methylation signal from single molecules generates profiles consistent with previously published ChIP-seq data (Figure 4.7).[115] DiMeLo-seq coverage is consistent across the entire *D. melanogaster* genome because DiMeLo-seq's long reads can be mapped in repetitive regions of the genome. We highlight a transition on chr3L where H3K9me3 accumulates and show that the accumulation is evident on single molecules using the `plot_browser` function.

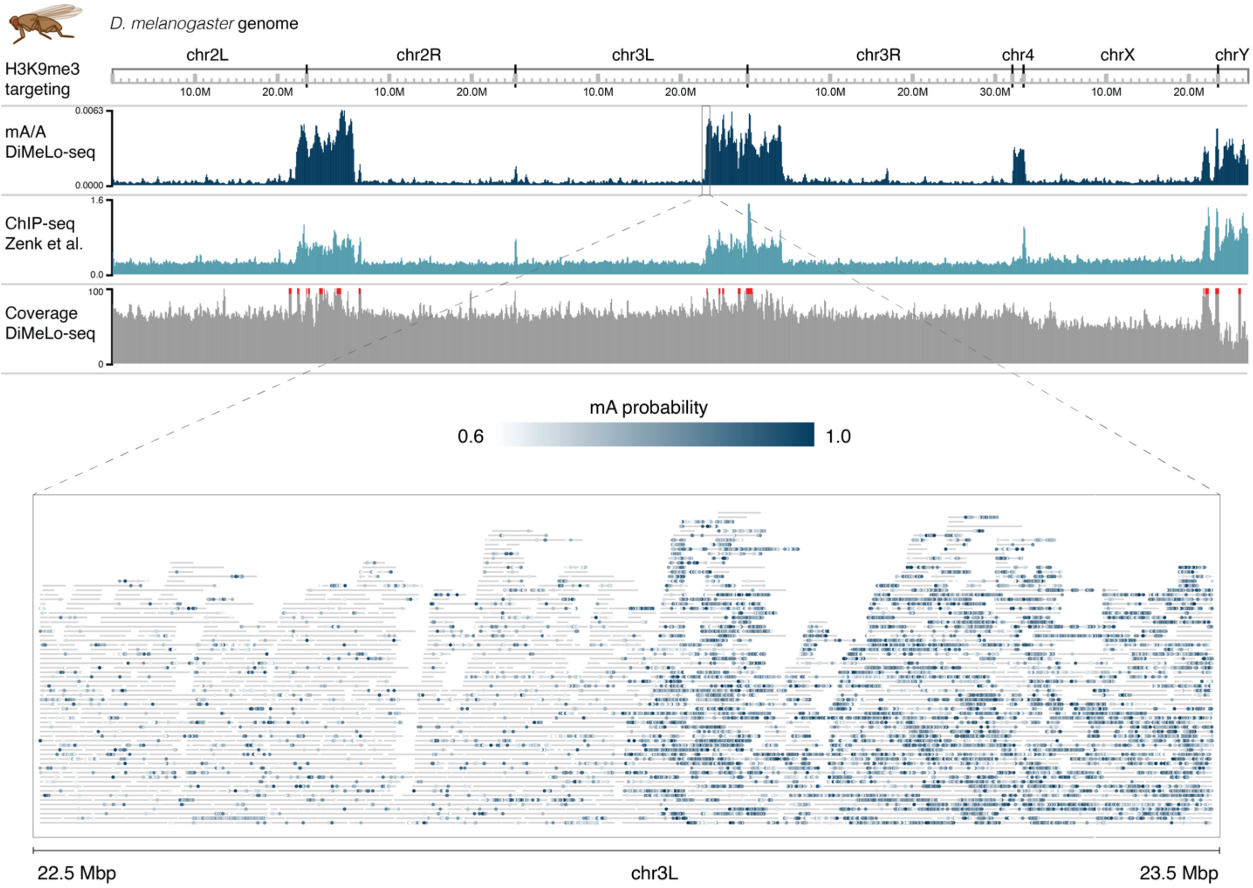


Figure 4.7 H3K9me3-targeted DiMeLo-seq in *D. melanogaster* embryos. Aggregate mA/A across the entire *D. melanogaster* genome from a DiMeLo-seq experiment targeting H3K9me3 is shown in dark blue. H3K9me3 ChIP-seq data in *D. melanogaster* embryos is shown in light blue.[115] Coverage from the DiMeLo-seq experiment is shown in grey. A region on chr3L where a transition from H3K9me3 depletion to H3K9me3 enrichment is highlighted with a single-molecule browser plot generated from `plot_browser`. Grey lines indicate reads and blue dots indicate mA calls with intensity colored by probability of methylation. An alignment length filter of 10 kb was applied. Methylation probability threshold of 0.6 was used.

The DiMeLo-seq protocol described here enables profiling of protein-DNA interactions in repetitive regions of the genome, makes phasing easier for determining haplotype-specific interactions,[54] detects joint binding events on single molecules of DNA, and captures protein binding together with endogenous CpG methylation. Performance varies by protein target, antibody quality, and chromatin environment; therefore, methylation sensitivity and specificity must be evaluated for each new target. The *dimelo* software package provides tools for quality control and data exploration for the multimodal datasets that DiMeLo-seq produces.

4.7 Data availability, code availability, author contributions, acknowledgements

Data availability

Raw sequencing data are available in the Sequence Read Archive (SRA) under BioProject accession PRJNA855257 and processed data are available on Gene Expression Omnibus (GEO) under accession GSE208125 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE208125). All raw fast5 sequencing data from the accompanying Altemose et al. manuscript are available in the Sequence Read Archive (SRA) under BioProject accession PRJNA752170. H3K27ac ChIP-seq data in GM12878 available from ENCODE Project Consortium under accession ENCFF218QBO (https://www.encodeproject.org/files/ENCFF218QBO/). H3K27me3 ChIP-seq data in GM12878 available from ENCODE Project Consortium under accession ENCFF119CAV (https://www.encodeproject.org/files/ENCFF119CAV/). H3K4me3 ChIP-seq data in GM12878 available from ENCODE Project Consortium under accession ENCFF228TWF (https://www.encodeproject.org/files/ENCFF228TWF/). H3K27ac CUT&Tag data in GM12878 available on GEO under accession GSM5530639 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5530639). H3K27me3 CUT&Tag data in GM12878 available on GEO under accession GSM5530673 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5530673).
ATAC-seq data in GM12878 available from ENCODE Project Consortium under accession ENCFF603BJO (https://www.encodeproject.org/files/ENCFF603BJO/). Transcription start site and gene annotations from NCBI RefSeq downloaded from UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=refSeqComposite&db=hg38). RNA-seq data in GM12878 available from ENCODE Project Consortium under accession ENCFF978HIY (https://www.encodeproject.org/files/ENCFF978HIY/). *D. melanogaster* H3K9me3 ChIP-seq data available on GEO under accession GSE140539 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140539). File GSE140539_H3K9me3_sorted_deepnorm_log2_smooth.bw was used.

Code availability

The *dimelo* Python package for analysis of DiMeLo-seq data is available on Github: https://github.com/streetslab/dimelo.

Author contributions

AM, NA, and AS designed the study. AM, NA, and LDB performed the experiments. AM, RM, and JM developed *dimelo* software package. AM and NA analyzed and interpreted the data. AM, NA, and RM made the figures. AM wrote the manuscript, with input from NA, RM, JM, LDB, KS, GK, AFS, and AS. AS and NA supervised the study.

Acknowledgements

This work was supported by the Chan Zuckerberg Biohub and by the NIGMS of the National Institutes of Health under award number R01HG012383 to AS, R01GM074728 to AFS, and R35GM139653 to GHK. AM is supported by a NSF GRFP award. NA is an HHMI Hanna H. Gray Fellow. AS is a Chan Zuckerberg Biohub Investigator, and a Pew Scholar in the Biomedical Sciences. The sequencing was carried out by the DNA Technologies and Expression Analysis Core at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01.

Competing interests

N.A., A.M., K.S., A.F.S. and A.S. are co-inventors on a patent application related to this work. The remaining authors declare no competing interests.

4.8 Supplementary Methods

Additional protocol and material availability

DiMeLo-seq: https://doi.org/10.17504/protocols.io.b2u8qezw; pA–Hia5 protein purification: https://doi.org/10.17504/protocols.io.bv82n9ye; AlphaHOR-RES: https://doi.org/10.17504/protocols.io.bv9vn966. Plasmids are available on Addgene: pA–Hia5 expression plasmid (pET-pA–Hia5; Addgene, 174372) and pAG–Hia5 expression plasmid (pET-pAG–Hia5; Addgene, 174373).

Cell culture

GM12878 cells (GM12878, Coriell Institute; mycoplasma tested) were maintained in RPMI-1640 with L-glutamine (Gibco, 11875093) supplemented with 15% FBS (VWR 89510-186) and 1% penicillin–streptomycin (Gibco, 15070063) at 37 °C in 5% CO2.

Antibody information

Antibodies used in this study were: (1) Histone H3K27ac antibody (pAb) (Active Motif 39133), (2) Histone H3K27me3 antibody (pAb) (Active Motif 39055), (3) Histone H3K4me3 antibody (pAb) (Active Motif 39916), (4) Histone H3K9me3 antibody (pAb) (Active Motif 39062).

DiMeLo-seq in GM12878 cells

For each target, 3.24 M cells from fresh culture were input to DiMeLo-seq. Antibody dilutions were all 1:50. DNA extraction was performed using the Monarch Genomic DNA Purification Kit (NEB T3010S).

DiMeLo-seq with D. melanogaster *embryos*

Timed embryo collections for downstream DiMeLo-seq

Approximately 200-400 OregonR flies were maintained on standard molasses medium before transfer to embryo collection cages with apple-juice plates with yeast paste. A heterogeneous mixture of embryos were collected overnight from 2-4 cages and pooled together. Embryos were rinsed from the apple-juice plates with DI water and collected in a mesh sieve, the chorion was removed by soaking in 50% bleach for 90 seconds, and then rinsed with water to remove the bleach. Embryos were transferred to a 1.5 mL eppendorf tube, allowed to settle, and the water was replaced with 1 mL Embryo Storage Buffer. Embryos were frozen in a Mr. Frosty isopropanol bath at -80ºC overnight, then stored at -80ºC. Nuclei were prepped for DiMeLo-seq by thawing the embryos at room temperature, removing the storage buffer and replacing it with 1mL of 1xPBS. Embryos were transferred to a 1 mL glass Dounce homogenizer and lysed with 10-15 strokes of a loose-fitting pestle. Nuclei were gently pelleted at 600xg for 3 minutes at 4ºC, the supernatant was removed, and the pellet was resuspended in Dig-wash buffer for downstream DiMeLo-seq. DNA extraction was performed using the Monarch HMW DNA Extraction Kit (NEB T3050L) with 2000 rpm at lysis.

Materials
Apple juice plates
Yeast paste
Oregon R flies (young) ~ 200 per bottle
Bleach
DI water

Buffers:
Embryo Storage Buffer: 80% Schneider's S2 media (Thermo 21720024), 10% FBS, 10% DMSO
1X PBS
50% bleach solution

Equipment
Small embryo sieve
Paintbrush
Squirt bottle (with DI water)
1 mL pipette and tips
Mr. Frosty isopropanol freezing bath
1.5 mL Eppendorf tubes

Library preparation and sequencing

All library preparation was performed using ON SQK-LSK110 with the standard protocol's bead-based cleanup protocol. Targets in GM12878 were sequenced on PromethION with R9.4.1 flow cells (ON FLO-PRO002). *D. melanogaster* embryo experiments were sequenced on MinION with R9.4.1 flow cells (ON FLO-MIN106D).

Chapter 5

Conclusion

5.1 Summary

This dissertation focused on methylation-based methods for measuring chromatin structure. I first optimized DamID-seq – a methylation-based, short-read method for mapping protein-DNA interactions – and extended it for single-cell multi-omic measurements. I modified DamID-seq to be compatible with *in situ* methylation and simultaneously added an additional measure of CpG methylation. The bulk of this dissertation was focused on the development of DiMeLo-seq, a method inspired by my earlier work that involved targeted *in situ* methylation followed by direct readout of both exogenous and endogenous methylation with long-read sequencing. To increase adoption of DiMeLo-seq by other labs, I created a detailed protocol manuscript with a Python package for analysis of DiMeLo-seq data.

5.2 Future directions

Native sequencing technologies are poised to further transform measurements of chromatin structure because these technologies can encode information in other DNA modifications beyond methylation. Immediate extensions of DiMeLo-seq include using the method to simultaneously profile multiple proteins on single molecules and applying DiMeLo-seq for measurement of topological association.

5.2.1 Multiplexed DiMeLo-seq

Our current efforts towards encoding additional layers of information involve multiplexing for measurement of DNA interactions with two proteins or with both DNA accessibility and a protein of interest. We introduce two methyltransferases – Hia5 and M.CviPI – which methylate different DNA bases in different sequence contexts. Hia5 methylates adenine, while M.CviPI methylates cytosine in the GC context. Importantly, endogenous methylation in the human genome is mostly in the CG context. We have demonstrated the compatibility of the two methyltransferases and the ability to simultaneously measure chromatin accessibility and protein binding by targeting Hia5 to LMNB1 and introducing untethered free M.CviPI to methylate open chromatin (Figure 5.1a,b). We are currently extending this method using nanobody-Hia5 and nanobody-M.CviPI fusions with nanobodies targeting different species to simultaneously target two proteins. The performance of the nanobody-Hia5 fusion is demonstrated in Figure 5.1c. Next steps involve validating nanobody-M.CviPI performance and improving assay conditions for optimized methyltransferase activity.

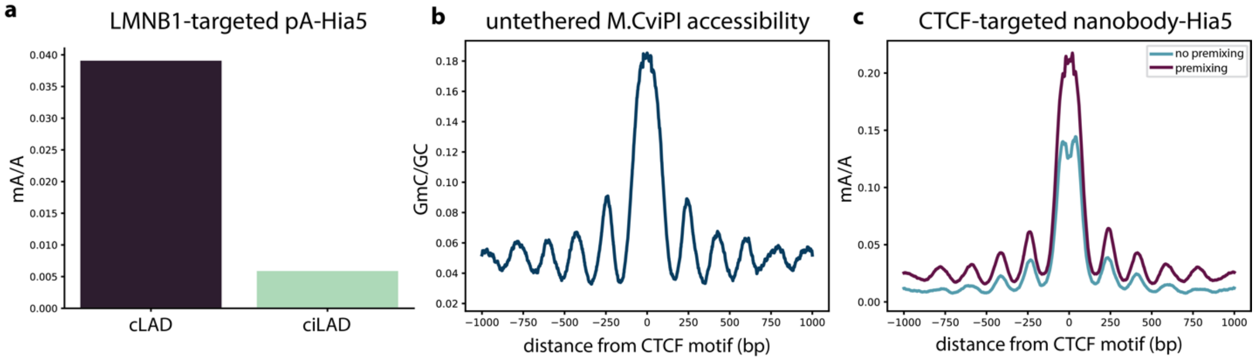


Figure 5.1 Initial validation of DiMeLo-seq with multiplexing. In a single sample, we simultaneously perform DiMeLo-seq targeting pA-Hia5 to LMNB1 and measure chromatin accessibility using M.CviPI. **a,** mA levels in cLADs (on-target) are higher compared to ciLADs (off-target) when targeting LMNB1 with pA-Hia5. **b,** GmC signal at top CTCF sites when adding untethered M.CviPI during activation in the same sample as (a) shows successful measurement of chromatin accessibility. CTCF strongly positions nucleosomes creating this periodic methylation pattern from preferential methylation of linker DNA relative to nucleosome-bound DNA. **c,** Using a nanobody instead of protein A to target Hia5 to an antibody works. Here, an anti-mouse nanobody is tethered to Hia5 and directed to CTCF. Premixing involves combining the nanobody-Hia5 fusion and primary antibody before adding the complex to the permeabilized nuclei. When premixing is not performed, primary antibody binding and nanobody-Hia5 binding are performed sequentially. Premixing results in more targeted methylation relative to the condition without premixing and has the added benefits of reduced experimental time, fewer wash steps, and higher recovery.

5.2.2 Profiling topological association on long single molecules

The leading methods for measuring genome organization are Chromatin Conformation Capture (3C)-based.[116] These methods rely on proximity ligation, so resolution is fundamentally linked to fragment size. Using an extension of DiMeLo-seq, we can create a new 4C-like[117,118], "one vs. all" measurement of genome organization that is instead limited in resolution by adenine density. We guide nuclease defective Cas9 (dCas9) to a region of interest by transiently transfecting plasmids encoding gRNA to our target (Figure 5.2a). We then perform standard DiMeLo-seq targeting dCas9 with the idea that regions in the vicinity of dCas9 both in linear space beside the molecule and in 3D space will be methylated. This measurement then results in all regions that interact with the targeted locus being methylated. With HiC, interactions in repetitive regions cannot be mapped with high resolution because long reads are needed to map to these regions. Encoding interactions in methylation space instead of fragmentation followed by proximity ligation allows for finer resolution in studying interactions in repetitive regions. We are performing initial validation experiments in *D. melanogaster* S2 cells because the genome is 6% the size of the human genome; this allows for shallower sequencing to validate and optimize performance. Initial imaging results showing targeted recruitment of dCas9 to a single locus in the *D. melanogaster* genome are shown in Figure 5.2b.

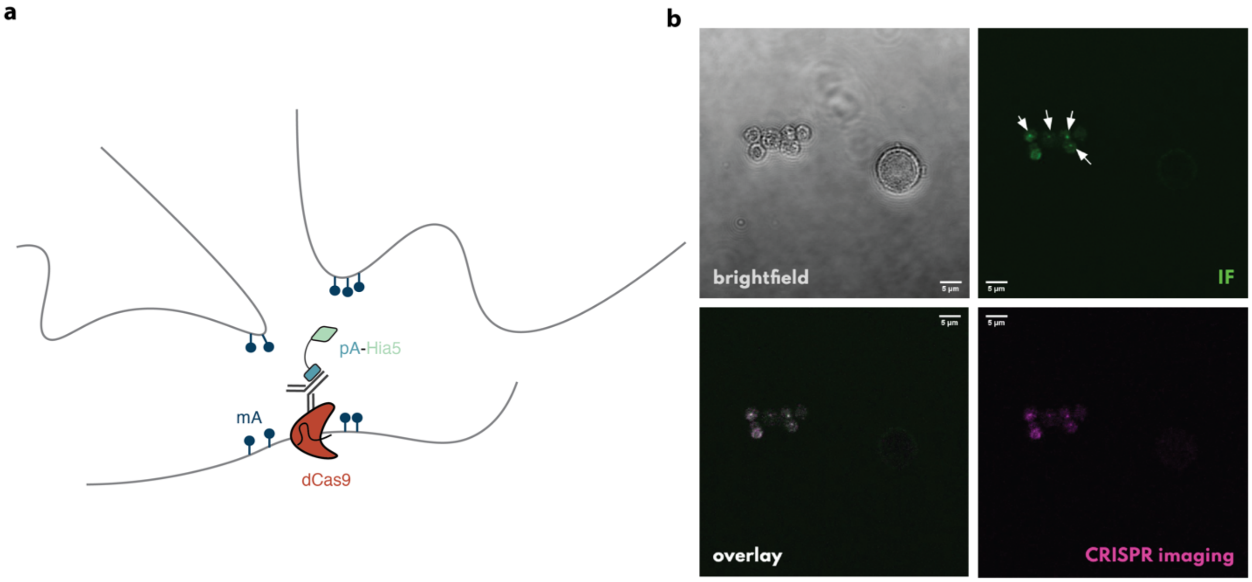


Figure 5.2 DiMeLo-seq targeting dCas9 to profile topological association. **a,** Schematic of the approach illustrating how methylation is targeted to the 3D vicinity of a locus of interest. **b,** Imaging validation in *D. melanogaster* S2 cells showing targeted recruitment of dCas9-mScarlet protein (red) and an antibody targeting this fusion protein (green). This antibody is then used as a handle to recruit pA-Hia5 and methylate in the vicinity of the dCas9-bound locus.

5.2.3 Beyond methylation for base modification encoding of chromatin structure

Other marks can be deposited on DNA and read out with long-read sequencing to store information about chromatin structure. One strategy could be to modify the methyl donor to have the methyltransferase deposit an azide group instead of a methyl group. Creating synthetic SAM analogs and modifying methyltransferases to deposit an azide group has been demonstrated.[119] With click chemistry, other groups could then be added with a DBCO-azide reaction as well. The key will be producing a distinct Nanopore signal relative to the unmodified base without having too bulky of a group that clogs the pores. While the space for potential modifications is vast, the bases themselves can still be converted too, as in short-read approaches. For example, an approach could involve using a cytosine deaminase to perform a C to U base conversion.[120] With long-read sequencers producing direct physical measurements of the DNA molecules themselves, countless possibilities for targeted modifications to these molecules to encode features of interest can be imagined.

5.3 Concluding remarks

The research described in this dissertation has expanded the toolkit of methods for studying chromatin structure, and the methods I've developed here can be used to study the relationship between multiple features of chromatin on single molecules and in regions of the genome previously understudied with short-read sequencing methods. These new methods leverage sequencing technologies that take an analytical measurement of physical properties of DNA, thereby creating new dimensions for encoding information about chromatin structure on the DNA molecules themselves. Ultimately, these multi-omic, genome-wide measurements of chromatin structure can be used to study the layers of regulatory elements that are controlling gene expression and to better understand how these elements go awry in diseased states.

References

1. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
3. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
4. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
5. Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat. Protoc.* **13**, 1006–1019 (2018).
6. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
7. van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.* **18**, 424–428 (2000).
8. O'Brown, Z. K. *et al.* Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics* **20**, 445 (2019).
9. Rang, F. J. *et al.* Single-cell profiling of transcriptome and histone modifications with EpiDamID. *Mol. Cell* **82**, 1956-1970.e14 (2022).
10. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1-21.29.9 (2015).

11. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **2010**, db.prot5384 (2010).

12. Zaret, K. Micrococcal nuclease analysis of chromatin structure. *Curr. Protoc. Mol. Biol.* **Chapter 21**, Unit 21.1 (2005).

13. Rooijers, K. *et al.* Simultaneous quantification of protein-DNA contacts and transcriptomes in single cells. *Nat. Biotechnol.* **37**, 766–772 (2019).

14. Genomics, 10x. Chromium Single Cell ATAC. *Single Cell ATAC - 10x Genomics* https://www.10xgenomics.com/products/single-cell-atac (2022).

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

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

17. Liu, Y. *et al.* Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome Biol.* **21**, 54 (2020).

18. Huang, X. *et al.* High-throughput sequencing of methylated cytosine enriched by modification-dependent restriction endonuclease MspJI. *BMC Genet.* **14**, 56 (2013).

19. Statham, A. L. *et al.* Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Res.* **22**, 1120–1127 (2012).

20. Brinkman, A. B. *et al.* Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.* **22**, 1128–1138 (2012).

21. Spektor, R., Tippens, N. D., Mimoso, C. A. & Soloway, P. D. methyl-ATAC-seq measures DNA methylation at accessible chromatin. *Genome Res.* **29**, 969–977 (2019).

22. Lhoumaud, P. *et al.* EpiMethylTag: simultaneous detection of ATAC-seq or ChIP-seq signals with DNA methylation. *Genome Biol.* **20**, 248 (2019).

23. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).

24. Stergachis, A. B., Debo, B. M., Haugen, E., Churchman, L. S. & Stamatoyannopoulos, J. A. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449–1454 (2020).

25. Abdulhay, N. J. *et al.* Massively multiplex single-molecule oligonucleosome footprinting. *Elife* **9**, e59404 (2020).

26. Shipony, Z. *et al.* Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat. Methods* **17**, 319–327 (2020).

27. Lee, I. *et al.* Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat. Methods* **17**, 1191–1199 (2020).

28. Wang, Y. *et al.* Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res.* **29**, 1329–1342 (2019).

29. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

30. Kind, J. *et al.* Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134–147 (2015).

31. Lenain, C. *et al.* Massive reshaping of genome-nuclear lamina interactions during oncogene-induced senescence. *Genome Res.* **27**, 1634–1644 (2017).

32. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).

33. van Steensel, B. & Belmont, A. S. Lamina-associated domains: Links with chromosome architecture, heterochromatin, and gene repression. *Cell* **169**, 780–791 (2017).

34. Buchwalter, A., Kaneshiro, J. M. & Hetzer, M. W. Coaching from the sidelines: the nuclear periphery in genome regulation. *Nat. Rev. Genet.* **20**, 39–50 (2019).

35. Altemose, N. *et al.* μDamID: A Microfluidic Approach for Joint Imaging and Sequencing of Protein-DNA Interactions in Single Cells. *Cell Syst* **11**, 354-366.e9 (2020).

36. Kind, J. *et al.* Single-cell dynamics of genome-nuclear lamina interactions. *Cell* **153**, 178–192 (2013).

37. Vogel, M. J., Peric-Hupkes, D. & van Steensel, B. Detection of in vivo protein-DNA interactions using DamID in mammalian cells. *Nat. Protoc.* **2**, 1467–1478 (2007).

38. Singh, J. & Klar, A. J. Active genes in budding yeast display enhanced in vivo accessibility to foreign DNA methylases: a novel in vivo probe for chromatin structure of yeast. *Genes Dev.* **6**, 186–196 (1992).

39. Aughey, G. N., Estacio Gomez, A., Thomson, J., Yin, H. & Southall, T. D. CATaDa reveals global remodelling of chromatin accessibility during stem cell differentiation in vivo. *Elife* **7**, (2018).

40. Elsawy, H. & Chahar, S. Increasing DNA substrate specificity of the EcoDam DNA-(adenine N(6))-methyltransferase by site-directed mutagenesis. *Biochemistry (Mosc.)* **79**, 1262–1266 (2014).

41. Park, M., Patel, N., Keung, A. J. & Khalil, A. S. Engineering epigenetic regulation using synthetic read-write modules. *Cell* vol. 176 227-238.e20 (2019).

42. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).

43. Kakar, M., Davis, J. R., Kern, S. E. & Lim, C. S. Optimizing the protein switch: altering nuclear import and export signals, and ligand binding domain. *J. Control. Release* **120**, 220–232 (2007).

44. Bernhofer, M. *et al.* NLSdb—major update for database of nuclear localization signals and nuclear export signals. *Nucleic Acids Res.* **46**, D503–D508 (2018).

45. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

46. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

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

48. Essletzbichler, P. *et al.* Megabase-scale deletion using CRISPR/Cas9 to generate a fully haploid human cell line. *Genome Res.* **24**, 2059–2065 (2014).

49. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

50. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

51. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

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

53. Li, D., Hsu, S., Purushotham, D., Sears, R. L. & Wang, T. WashU Epigenome Browser update 2019. *Nucleic Acids Res.* **47**, W158–W165 (2019).

54. Altemose, N. *et al.* DiMeLo-seq: a long-read, single-molecule method for mapping protein–DNA interactions genome wide. *Nat. Methods* 1–13 (2022).

55. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).

56. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, (2017).

57. Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39–55 (2013).

58. Sönmezer, C. *et al.* Molecular co-occupancy identifies transcription factor binding cooperativity in vivo. *Mol. Cell* **81**, 255-267.e6 (2021).

59. Nurk, S. *et al.* The complete sequence of a human genome. *bioRxiv* 2021.05.26.445798 (2021) doi:10.1101/2021.05.26.445798.

60. Schmid, M., Durussel, T. & Laemmli, U. K. ChIC and ChEC; genomic mapping of chromatin proteins. *Mol. Cell* **16**, 147–157 (2004).

61. van Schaik, T., Vos, M., Peric-Hupkes, D., Hn Celie, P. & van Steensel, B. Cell cycle dynamics of lamina-associated DNA. *EMBO Rep.* **21**, e50636 (2020).

62. Drozdz, M., Piekarowicz, A., Bujnicki, J. M. & Radlinska, M. Novel non-specific DNA adenine methyltransferases. *Nucleic Acids Res.* **40**, 2119–2130 (2012).

63. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of Molecular Biology* vol. 276 19–42 (1998).

64. Meuleman, W. *et al.* Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res.* **23**, 270–280 (2013).

65. Sobecki, M. *et al.* MadID, a Versatile Approach to Map Protein-DNA Interactions, Highlights Telomere-Nuclear Envelope Contact Sites in Human Cells. *Cell Rep.* **25**, 2891-2903.e5 (2018).

66. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**, 482–485 (2000).

67. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* **21**, 1757–1767 (2011).

68. Boyle, A. P. *et al.* High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.* **21**, 456–464 (2011).

69. Klenova, E. M. *et al.* CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Mol. Cell. Biol.* **13**, 7612–7624 (1993).

70. Lobanenkov, V. V. *et al.* A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene* **5**, 1743–1753 (1990).

71. Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.* **17**, 520–527 (2001).

72. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).

73. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22**, 2497–2506 (2012).

74. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).

75. Hellman, A. & Chess, A. Gene body-specific methylation on the active X chromosome. *Science* **315**, 1141–1143 (2007).

76. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **375**, (2022).

77. McNulty, S. M. & Sullivan, B. A. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res.* **26**, 115–138 (2018).

78. Rudd, M. K., Schueler, M. G. & Willard, H. F. Sequence organization and functional annotation of human centromeres. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 141–149 (2003).

79. Willard, H. F. & Waye, J. S. Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **3**, 192–198 (1987).

80. Miga, K. H. *et al.* Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014).

81. Hayden, K. E. *et al.* Sequences associated with centromere competency in the human genome. *Mol. Cell. Biol.* **33**, 763–772 (2013).

82. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).

83. Lica, L. & Hamkalo, B. Preparation of centromeric heterochromatin by restriction endonuclease digestion of mouse L929 cells. *Chromosoma* **88**, 42–49 (1983).

84. Smith, O. K. *et al.* Identification and characterization of centromeric sequences in Xenopus laevis. *Genome Res.* **31**, 958–967 (2021).

85. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).

86. Bodor, D. L. *et al.* The quantitative architecture of centromeric chromatin. *Elife* **3**, e02137 (2014).

87. Aldrup-MacDonald, M. E., Kuo, M. E., Sullivan, L. L., Chew, K. & Sullivan, B. A. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Res.* **26**, 1301–1311 (2016).

88. Gilpatrick, T. *et al.* Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* **38**, 433–438 (2020).

89. Kovaka, S., Fan, Y., Ni, B., Timp, W. & Schatz, M. C. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat. Biotechnol.* **39**, 431–441 (2021).

90. Gamba, R. *et al.* A method to enrich and purify centromeric DNA from human cells. *bioRxiv* (2021) doi:10.1101/2021.09.24.461328.

91. Meers, M. P., Bryson, T. D., Henikoff, J. G. & Henikoff, S. Improved CUT&RUN chromatin profiling tools. *Elife* **8**, (2019).

92. Cao, S., Zhou, K., Zhang, Z., Luger, K. & Straight, A. F. Constitutive centromere-associated network contacts confer differential stability on CENP-A nucleosomes in vitro and in the cell. *Mol. Biol. Cell* **29**, 751–762 (2018).

93. Zhou, K. *et al.* CENP-N promotes the compaction of centromeric chromatin. *Nat. Struct. Mol. Biol.* **29**, 403–413 (2022).

94. Kim, B. Y. *et al.* Highly contiguous assemblies of 101 drosophilid genomes. *Elife* **10**, (2021).

95. Guse, A., Fuller, C. J. & Straight, A. F. A cell-free system for functional centromere and kinetochore assembly. *Nat. Protoc.* **7**, 1847–1869 (2012).

96. Westhorpe, F. G., Fuller, C. J. & Straight, A. F. A cell-free CENP-A assembly system defines the chromatin requirements for centromere maintenance. *J. Cell Biol.* **209**, 789–801 (2015).

97. Carroll, C. W., Silva, M. C. C., Godek, K. M., Jansen, L. E. T. & Straight, A. F. Centromere assembly requires the direct recognition of CENP-A nucleosomes by CENP-N. *Nat. Cell Biol.* **11**, 896–902 (2009).

98. Jain, C. *et al.* Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).

99. Janssens, D. & Henikoff, S. CUT&RUN: Targeted in situ genome-wide profiling with high efficiency for low cell numbers v3. (2019) doi:10.17504/protocols.io.zcpf2vn.

100. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).

101. De Coster, W., Stovner, E. B. & Strazisar, M. Methplotlib: analysis of modified nucleotides from nanopore sequencing. *Bioinformatics* **36**, 3236–3238 (2020).

102. Eberle, M. A. *et al.* A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* **27**, 157–164 (2017).

103. Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **37**, 561–566 (2019).

104. Akbari, V. *et al.* Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. *Genome Biol.* **22**, 68 (2021).

105. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

106. Ulaner, G. A. *et al.* CTCF binding at the insulin-like growth factor-II (IGF2)/H19 imprinting control region is insufficient to regulate IGF2/H19 expression in human tissues. *Endocrinology* **144**, 4420–4426 (2003).

107. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).

108. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-5 (2016).

109. Lopez-Delisle, L. *et al.* pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics* **37**, 422–423 (2021).

110. Maslan, A. et al. Mapping protein-DNA interactions genome-wide with DiMeLo-seq. *bioRxiv* (2022) doi:10.1101/2022.07.03.498618.

111. Brothers, M. & Rine, J. Distinguishing between recruitment and spread of silent chromatin structures in Saccharomyces cerevisiae. *Elife* **11**, (2022).

112. Zhao, L. *et al.* FACT-seq: profiling histone modifications in formalin-fixed paraffin-embedded samples with low cell numbers. *Nucleic Acids Res.* **49**, e125 (2021).

113. Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 2926–2931 (2010).

114. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* **10**, 295–304 (2009).

115. Zenk, F. *et al.* HP1 drives de novo 3D genome reorganization in early Drosophila embryos. *Nature* **593**, 289–293 (2021).

116. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).

117. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38**, 1341–1347 (2006).

118. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).

119. Gabrieli, T. *et al.* Chemoenzymatic labeling of DNA methylation patterns for single-molecule epigenetic mapping. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkac460.

120. Gallagher, L. A. *et al.* Genome-wide protein-DNA interaction site mapping in bacteria using a double-stranded DNA-specific cytosine deaminase. *Nat. Microbiol.* **7**, 844–855 (2022).