DiMeLo-seq: a long-read, single-molecule method for mapping protein–DNA interactions genome wide

Nicolas Altemose, Annie Maslan, Owen K. Smith, Kousik Sundararajan, Rachel R. Brown, Reet Mishra, Angela M. Detweiler, Norma Neff, Karen H. Miga, Owen K. Smith, Kousik Sundararajan and Aaron F. Straight

Genomic DNA needs to be decoded and maintained by proteins that read, regulate, replicate, recombine and repair it. Mapping where and how 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, chromatin immunoprecipitation with sequencing (ChIP–seq), CUT&RUN, and their derivatives. 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 next-generation sequencing (NGS), and mapping those sequences back to a reference genome, using sequencing coverage as a measure of protein–DNA interaction frequency. While these methods have been developed to map where individual target proteins interact with DNA genome wide, including DamID, chromatin immunoprecipitation with sequencing (ChIP–seq), CUT&RUN, and their derivatives, they suffer from several limitations.

Studies of genome regulation routinely use high-throughput DNA sequencing approaches to determine where specific proteins interact with DNA, and they rely on DNA amplification and short-read sequencing, limiting their quantitative application in complex genomic regions. To address these limitations, we developed directed methylation with long-read sequencing (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 localizes within highly repetitive regions that were unmappable with short sequencing reads, and we estimated the density of centromere protein A molecules along single chromatin fibers. DiMeLo-seq is a versatile method that provides multimodal, genome-wide information for investigating protein–DNA interactions.

Furthermore, these approaches rely on digesting or shearing DNA into short fragments for enrichment, followed by NGS, which produces short sequencing reads typically under 250 base pairs (bp) in length. Short fragment lengths are often necessary for achieving adequate binding-site resolution with these techniques. Although it is possible to map multiple protein–DNA interactions on short reads, shearing the DNA into short fragments can destroy joint long-range binding information, and it 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. These obstacles hinder 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 interrogate assembled repetitive regions and to read out DNA modifications directly. To address this need, we developed DiMeLo-seq (from ‘dimelo’, 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. These features provide an opportunity to study genome variation.

1Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA. 2UC Berkeley-UCSF Graduate Program in Bioengineering, University of California, Berkeley, Berkeley, CA, USA. 3Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA. 4Department of Biochemistry, Stanford University, Stanford, CA, USA. 5Department of Chemical and Systems Biology, Stanford University, Stanford, CA, USA. 6Chan Zuckerberg Biohub, San Francisco, CA, USA. 7Department of Molecular & Cell Biology, University of California, Santa Cruz, Santa Cruz, CA, USA. 8UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. 9Present address: Department of Molecular & Cell Biology, University of California, Berkeley, Berkeley, CA, USA. 10These authors contributed equally: Nicolas Altemose, Annie Maslan, Owen K. Smith, Kousik Sundararajan. These authors jointly supervised this work: Aaron F. Straight, Aaron Streets. E-mail: astraigh@stanford.edu; astreets@berkeley.edu
DNA methylation marks directly. Turning advantage of the low abundance of N\textsuperscript{6}-methyl-deoxyadenosine (hereafter mA) in human DNA, we fused the antibody-binding protein A to the nonspecific deoxyadenosine methyltransferase Hia5 (pA–Hia5) to catalyze the formation of mA in the DNA proximal to targeted chromatin-associated proteins (Fig. 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-adenosylmethionine (SAM) to activate adenine methylation in the vicinity of the protein of interest. Finally, genomic DNA is isolated and sequenced using modification-sensitive, long-read sequencing with mA base calls providing a readout of the sites of protein–DNA interactions (Fig. 1a and Supplementary Fig. 1).

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 (Fig. 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 (Fig. 1c). Modification-sensitive readout allows for the simultaneous detection of both exogenous antibody-directed adenine methylation and endogenous CpG methylation on single molecules (Fig. 1d). Additionally, DiMeLo-seq’s long sequencing reads often overlap multiple heterozygous sites, enabling phased and measurement of haplo-typic-specific protein–DNA interactions (Fig. 1e). Finally, long reads enable mapping of protein–DNA interactions within highly repetitive regions of the genome (Fig. 1f).

**Results**

**DiMeLo-seq workflow.** DiMeLo-seq combines elements of antibody-directed protein–DNA mapping approaches\textsuperscript{6,15,16} to deposit methylation marks near a specific target protein, then uses long-read sequencing to read out these exogenous methylation marks directly.\textsuperscript{10–14} Taking advantage of the low abundance of N\textsuperscript{6}-methyl-deoxyadenosine (hereafter mA) in human DNA,\textsuperscript{15,16} we fused the antibody-binding protein A to the nonspecific deoxyadenosine methyltransferase Hia5 (pA–Hia5)\textsuperscript{11,18} to catalyze the formation of mA in the DNA proximal to targeted chromatin-associated proteins (Fig. 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-adenosylmethionine (SAM) to activate adenine methylation in the vicinity of the protein of interest. Finally, genomic DNA is isolated and sequenced using modification-sensitive, long-read sequencing with mA base calls providing a readout of the sites of protein–DNA interactions (Fig. 1a and Supplementary Fig. 1).

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 (Fig. 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 (Fig. 1c). Modification-sensitive readout allows for the simultaneous detection of both exogenous antibody-directed adenine methylation and endogenous CpG methylation on single molecules (Fig. 1d). Additionally, DiMeLo-seq’s long sequencing reads often overlap multiple heterozygous sites, enabling phased and measurement of haplo-typic-specific protein–DNA interactions (Fig. 1e). Finally, long reads enable mapping of protein–DNA interactions within highly repetitive regions of the genome (Fig. 1f).

**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 and Extended Data Fig. 1a,b). To test the ability of pA–Hia5 to target chromatin and methylate accessible DNA in vitro, we reconstituted chromatin containing the histone variant centromere protein A (CENP-A) using the nucleosome-positioning DNA sequence referred to as ‘601’ (ref. 15; Extended Data Fig. 1c,d and Supplementary Note 2). Incubating mononucleosomes together with free-floating pA–Hia5 and SAM, followed by long-read sequencing and methylation-sensitive base calling, showed methylation on 97.1% ± 0.8% of reads (mean ± s.e.m., n = 3; Supplementary Notes 3 and 4, Fig. 1c–k; Extended Data Fig. 1e–k). Moreover, we observed almost no methylation at the expected nucleosome-protected region (Fig. 1c,d and Extended Data Fig. 1j).

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 before activating methylation with SAM.
Fig. 2 | Application of DiMeLo-seq in artificial chromatin. a, Schematic of antibody-directed methylation of artificial chromatin. b, Heat map of 5,000 individual 1 × 601 reads from chromatin containing CENP-A mononucleosomes methylated with CENP-A-directed pA–Hia5 (red dashed line indicates 601 dyad position). c, Plots of A or T density (top) and average mA/A on base position of 1 × 601 containing DNA (bottom; red dashed line indicates 601 dyad position). d, Plot of percentage of reads with methylation as a function of the distance from dyad axis. e, Schematic of directed methylation of 18 × 601 chromatin array. f,g, Heat map of 2,000 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 18 × 601 array (f) or a subset 4 × 601 region (g) along with cartoons depicting predicted nucleosome positions (red circles). Insets (below) show average mA/A on every base position of 18 × 601 array or 4 × 601 portion (red dashed line indicates the 601 dyad position).
Fig. 2b–d and Extended Data Fig. 1e–h,k), with methylation levels decaying with distance from the nucleosome footprint (Fig. 2c). 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; Fig. 2d and Extended Data Fig. 1e,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; Fig. 2c,d and Extended Data Fig. 1j), ~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 first assessed the ability of pA–Hia5 to methylate accessible regions of DNA on in vitro reconstituted chromatin assembled on an 18× array of the 601 nucleosome-positioning sequence (Extended Data Fig. 2a–c). Co-incubation of chromatin together with free-floating pA–Hia5 and SAM resulted in structured patterns of oligonucleosome footprinting (Extended Data Fig. 2b,g,h), as reported previously for reconstituted chromatin incubated with another exogenous methyltransferase, EcoGII16.

We then tested antibody-directed methyltransferase of chromatin arrays reconstituted with nucleosomes containing either CENP-A or histone H3. We incubated chromatin with CENP-A antibody and pA–Hia5, washed away unbound antibody, and activated methyltransferase with SAM (Fig. 2c,e). Following activation, we immunostained chromatin-conjugated beads with an anti-mA antibody, demonstrating an increase in mA signal when CENP-A chromatin, but not H3 chromatin, was incubated with pA–Hia5 and CENP-A antibody (Extended Data Fig. 2d,e and 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) (Extended Data Fig. 2f). 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 (Fig. 2g). On individual reads, we observed protection from methylation centered at 601 dyad positions, consistent with nucleosome occupancy protecting the DNA from antibody-directed methylation (Fig. 2f,g) and similar to the free pA–Hia5 condition (Extended Data Fig. 2g,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 approximately fourfold lower average probability of methylation (Fig. 2f and Extended Data Fig. 2g), 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 (Extended Data Fig. 2i). We observed similar results for H3-antibody-directed methylation of H3 chromatin using pAG–Hia5 (Extended Data Fig. 2j–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.

Optimization of lamin B1 mapping in situ. We next optimized DiMeLo-seq for mapping protein–DNA interactions in situ in permaobilized 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 LADs20. 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; Fig. 3a). 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 (1) cLADs and ciLADs provide well-characterized on-target and off-target control regions, respectively; (2) LMNB1 has a very large binding footprint (LADs have a median size of 500 kb and cover roughly 30% of the genome17), so DNA–LMNB1 interactions can be detected even with very low sequencing coverage; (3) 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 (Extended Data Fig. 3c,d); and (4) we have previously generated LMNB1 DamID data from HEK293T cells using bulk and single-cell protocols, providing ample reference materials21.

To assess the performance of the LMNB1-targeted DiMeLo-seq protocol, we quantified the proportion of adenines that were called as methylated (ma/A) 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 ratio), 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 h (Methods and Supplementary Notes 6–8). A protocol optimization pipeline, we tested over 200 different conditions (Fig. 3b), varying the following: methyltransferase type (Hia5 versus 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 and Supplementary Table 1). We validated an initial version of the protocol (v1; https://doi.org/10.17504/protocols.io.bv8tn9wn/), and then further optimized the methyltransferase activation conditions to increase the amount of on-target methylation by 50–60% without sacrificing specificity (v2; https://doi.org/10.17504/protocols.io.h2u8qezw/); Extended Data Figs. 3 and 4, Supplementary Note 8 and Fig. 3b). 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 (Extended Data Fig. 5a).

We also verified that there is very little loss of performance when using cells that were cryopreserved in dimethylsulfoxide-containing medium or lightly fixed in paraformaldehyde (PFA), when using between 1 and 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, Supplementary Notes 9 and 10 and Supplementary Table 1). To confirm antibody specificity, we performed experiments using IgG isotype controls and free-floating Hia5 controls to measure nonspecific methylation and DNA accessibility, respectively (Methods and Supplementary Table 2). We also generated a stably transduced line expressing a direct fusion between EcoGII and LMNB1 in vivo, as in MadID23, and then we detected mAs with nanopore sequencing (Extended Data Fig. 4a and Supplementary Note 10). This in vivo approach produced threefold more on-target methylation compared to in situ DiMeLo-seq with pAG–EcoGII (Fig. 3b), although this performance is expected to vary with different fusion proteins and their expression levels (Supplementary Note 10).

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 (Fig. 3c). 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 (Fig. 3c)\(^23\). 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 (Fig. 3c,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, that is, 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...
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. 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 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 (Extended Data Fig. 5b). We also measured a sixfold mA/A enrichment in the free pA–Hia5 control in CTCF ChIP–seq peaks, reflecting that many CTCF-binding sites overlap with accessible regions of the genome where pA–Hia5 can methylate more easily. However, both the free pA–Hia5 and the IgG controls produced significantly less on-target methylation than the CTCF-targeted sample (Extended Data Fig. 5b). We confirmed that signal enrichment is caused by CTCF-targeted methylation and not accessibility of CTCF sites by measuring a 1.8 times greater proportion of mA in ChIP–seq peaks compared to regions of open chromatin measured by the assay for transposable-accessible chromatin with sequencing (ATAC–seq; Extended Data Fig. 5c).

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 (Fig. 4a). 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 (Extended Data Fig. 5d). The 88-bp dip at the center of the binding peak reflects CTCF’s binding footprint27–29 and is evident even on single molecules. CTCF binds to ~50 bp of DNA as determined by DNase I footprinting and ChIP–exo20,30,31. 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 Fig. 4, and an antibody targeting the N terminus of CTCF. We observed methylation enrichment 5′ to the binding motif with C-terminal targeting and 3′ to the motif with N-terminal targeting (P value = 0.0010; Supplementary Note 11 and Extended Data Fig. 5e). The free pA–Hia5 control profile supports this finding that the antibody-binding site is causing the peak asymmetry, as there is no notable asymmetry in this untargeted case (Extended Data Fig. 6).

To evaluate the use of DiMeLo-seq for de novo peak detection, we called CTCF peaks using DiMeLo-seq data alone and created receiver operating characteristic curves at increasing sequencing depths using ChIP–seq peaks as ground truth (Supplementary Note 11 and Extended Data Fig. 5f). At ~25× coverage, we detected 60% of ChIP–seq peaks (false positive rate of 1.6%) and measured an area under the curve value of 0.92 (Supplementary Note 11). Among the peaks detected with DiMeLo-seq that were not annotated ChIP–seq peaks, 10% overlapped 1-kb marker deserts and gaps in the hg38 reference and were 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 (Fig. 4a). This inverse relationship between CpG methylation and CTCF-targeted methylation reflects previous findings that mCpG inhibits CTCF binding32. We measured both mA and mCpG on the same single molecules and also observed that both A and CpG were preferentially methylated in linker DNA (Fig. 4b). 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 sites33. 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 6bp 5′ of the CTCF motif center, and the peak center on approximately 70% of the reads fell within ±200bp of the motif center (Extended Data Fig. 5g). This systematic bias toward 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 169bp (Extended Data Fig. 5d).
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% false positive rate; Extended Data Fig. 5h,i and 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 (Fig. 4c). 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 chromosome 6 where haplotype-specific single nucleotide polymorphisms (SNPs) within the CTCF-binding motif prevent CTCF binding at one of the two neighboring sites (Extended Data Fig. 7a). 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 (Fig. 4d). We also reported haplotype-specific CTCF-binding profiles at specific sites and broadly across the active and inactive X chromosomes (Extended Data Fig. 7b–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 (Extended Data Fig. 8), indicating that DiMeLo-seq is compatible with PacBio’s circular consensus sequencing technique. However, while PacBio sequencing has reported improved base-calling accuracy\(^1\), this approach detected more methylation in the unmethylated control than nanopore sequencing, slightly reducing the signal-to-noise ratio of the measurement (Extended Data Fig. 8).

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 trimethylated histone H3 lysine 9 (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\(^4\), and many different sequencing data types are available for it\(^1\). To validate the specificity of targeted methylation, we calculated the fraction of adenines methylated within HG002 CUT&RUN H3K9me3 peaks\(^5\) compared to the fraction of adenines methylated outside 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 (Fig. 5a), 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)\(^35–38\). 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-fold enrichment in active (kinetochore-binding) HOR arrays\(^35\) over non-centromeric regions (Fig. 5b). We next looked at HOR array boundaries and observed a decrease in H3K9me3 across the boundary moving from within to outside HOR arrays (Fig. 5c). In contrast, for the free pA–Hia5 control, mA/A increased moving from within to outside the HOR array, as chromatin becomes more accessible (Extended Data Fig. 9a)\(^36\).

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 (Fig. 5d). An IgG isotype control confirmed that adenine methylation in the H3K9me3-targeted sample was not caused by background methylation (Fig. 5d and Extended Data Fig. 9b). Unlike methods that 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 to 68 Mb, where H3K9me3 signal drops as CpG methylation increases (Fig. 5d). 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 (Extended Data Fig. 9c). 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

---

**Fig. 5 | Detecting H3K9me3 in centromeres.** a, The proportion of adenines methylated within CUT&RUN peaks relative to the proportion of adenines methylated outside 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 fractions 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 the HOR array. HOR array boundaries that transition quickly into non-repetitive sequences were considered: 1p, 2p, 6p, 9p, 13q, 14q, 15q, 16p, 17p, 18p, 20p, 21q and 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. Gray bars below the centromere annotation indicate regions with >20-kb marker deserts.
that CENP-A nucleosomes are found at the core of these repeat units where the repeats are the most homogeneous\textsuperscript{35,39–41}. 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 to avoid sequencing the entire genome.

Our enrichment strategy, called AlphaHOR-RES (alpha HOR restriction and enrichment by size; from ‘alfajores’), is based on classic centromere enrichment strategies\textsuperscript{42} that involve digesting the genome with restriction enzymes that cut frequently. 
outside centromeric regions but rarely inside them, then removing short DNA fragments (Methods and Extended Data Fig. 10a). 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; Fig. 6a,b, Extended Data Fig. 10b–d and Methods). Thus, this enrichment strategy drastically 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.

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 (Fig. 6a,b). In an alignment-independent manner\(^1\), we classified DiMeLo-seq reads based on the presence or absence of CENP-A-enriched k-mers from an available short-read sequencing dataset\(^2\). CENP-A-directed DiMeLo-seq reads with CENP-A-enriched k-mers had roughly sevenfold more adenine methylation when compared to reads without CENP-A-enriched k-mers (Fig. 6c). 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)\(^3,4\). 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\(^3,5,6\). We confirmed low endogenous CpG methylation within the CDR as expected (Fig. 6d). 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 (Fig. 6e,f)\(^3,6\). We found that the density of detected CENP-A nucleosomes increases fivefold within chromosome X CDRs compared to neighboring regions (Fig. 6g). We estimate that 26% ± 5% of nucleosomes contain CENP-A within the chromosome X CDR, whereas only 5% ± 2% of nucleosomes contain CENP-A within a neighboring region (mean ± s.d.; Supplementary Note 14 and Fig. 6g) 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 ± s.d.) of nucleosomes detected on IgG control reads within chromosome X CDR (Fig. 6g and Extended Data Fig. 10e)). 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\(^6\). In contrast, we estimated that at least 1 in 4 nucleosomes contain CENP-A within the smaller ~100-kb CDR on chromosome X. This demonstrates that CENP-A nucleosome occupancy varies considerably across a human centromere. Further, we showed that the region with the highest CENP-A density coincides with the CDR. We observed 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 (Extended Data Fig. 10f,g). These observations support the finding of one active HOR array for each chromosome\(^3,4\). These findings illuminate the density and positioning of CENP-A nucleosomes within HOR segments on individual chromatin fibers, which was not previously attainable with existing techniques.

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 chromatins 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 chromosome X CDR are lower limits, and the actual CENP-A density within CDRs could be even higher than ~25% (Fig. 6g). A source of variation in CENP-A positions is the cell cycle state of chromatin. Because preexisting 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% and 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 trade-offs 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 or to use DNA enrichment methods like AlphaHOR-RES, the centromere enrichment method 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 AlphaHOR-RES. 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 measured 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 three-dimensional (3D) space, as does CUT&RUN. These 3D interactions, and the factors that mediate them, can potentially be investigated by perturbing 3D chromatin structure before 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 fusions between proteins and MTases, as in conventional DamID or MadID, 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 PacBio high fidelity (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.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01475-6.

Received: 3 August 2021; Accepted: 24 March 2022; Published online: 8 April 2022

References
1. van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nat. Biotechnol. 18, 424–428 (2000).
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. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. Cell 129, 823–837 (2007).
6. Shen, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife 6, e21856 (2017).
7. Rivera, C. M. & Ren, B. Mapping human epigenomes. Cell 155, 39–55 (2013).
8. Sönmez, C. et al. Molecular co-occupancy identifies transcription factor binding cooperativity in vivo. Mol. Cell 81, 255–267 (2021).
9. Nunk, S. et al. The complete sequence of a human genome. Science 376, 44–53 (2022).
10. Abdulhail, N. J. et al. Massively multiplex single-molecule oligonucleosome footprinting. Elife 9, e59404 (2020).
11. 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).
12. Lee, I. et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat. Methods 17, 1191–1199 (2020).
13. Shipony, Z. et al. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. Nat. Methods 17, 319–327 (2020).
14. Wang, Y. et al. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. Genome Res. 29, 1329–1342 (2019).
15. Schmid, M., Durussel, T. & Laemmli, U. K. ChIC and ChEC: genomic mapping of chromatin proteins. Mol. Cell 16, 147–157 (2004).
16. van Schaik, T., Vos, M., Peric-Hupkes, D., Hn Celie, P. & van Steensel, B. Cell cycle dynamics of lamina-associated DNA. EMBO Rep. 21, e56536 (2020).
17. 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).
18. Drozdz, M., Piekarowicz, A., Bujnicki, J. M. & Radilinska, M. Novel nonspecific DNA adenosine methyltransferases. Nucleic Acids Res. 40, 2119–2130 (2012).
19. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to octamer and sequence-directed nucleosome positioning. J. Mol. Biol. 276, 19–42 (1998).
20. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453, 948–951 (2008).
21. 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).
22. Altemose, N. et al. μDamID: a microfluidic approach for joint imaging and sequencing of protein–DNA interactions in single cells. Cell Syst. 11, 354–368 (2020).
23. Sobek, M. et al. MadID, a versatile approach to map protein–DNA interactions, highlights telomere-nucleus envelope contact sites in human cells. Cell Rep. 25, 2891–2903 (2018).
24. Kind, J. et al. Genome-wide maps of nuclear lamina interactions in single human cells. Cell 163, 134–147 (2015).
25. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature 405, 482–485 (2000).
26. Song, L. et al. Open chromatin defined by DNase I and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 21, 1757–1767 (2011).
27. 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).
28. Klenova, E. M. et al. CTCF; a conserved nuclear factor required for optional 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).
29. Lobanenkov, V. V. et al. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CTCCTC motif in the 5′-flanking sequence of the chicken c-myc gene. Oncogene 5, 1743–1753 (1990).
30. Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTFC is a uniquely versatile transcription regulator linked to epigenetics and disease. Trends Genet. 17, 520–527 (2001).
31. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. Cell 147, 1408–1419 (2011).
32. Kelly, T. K. et al. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome Res. 22, 2422–2426 (2012).
33. 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).
34. Gershman, A. et al. Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
35. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eab4178 (2022).
36. McNulty, S. M. & Sullivan, B. A. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res.* **26**, 115–138 (2018).
37. 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).
38. Willard, H. F. & Waye, J. S. Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **3**, 192–198 (1987).
39. Miga, K. H. et al. Centromere reference models for human chromosomes X and Y satellite arrays. * Genome Res.* **24**, 697–707 (2014).
40. Hayden, K. E. et al. Sequences associated with centromere competency in the human genome. *Mol. Cell. Biol.* **33**, 763–772 (2013).
41. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
42. Lica, L. & Hamkalo, B. Preparation of centromeric heterochromatin by restriction endonuclease digestion of mouse L929 cells. *Chromosoma* **88**, 42–49 (1983).
43. Smith, O. K. et al. Identification and characterization of centromeric sequences in *Xenopus laevis*. *Genome Res.* **31**, 958–967 (2021).
44. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
45. Bodor, D. L. et al. The quantitative architecture of centromeric chromatin. *Elife* **3**, e02137 (2014).
46. 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).
47. Gilpatrick, T. et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* **38**, 433–438 (2020).
48. 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).
49. Gamba, R. et al. A method to enrich and purify centromeric DNA from human cells. Preprint at bioRxiv https://doi.org/10.1101/2021.09.24.461328 (2021).

*Publisher’s note* Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022
**Methods**

Protocol and material availability. For detailed and updated protocols, please refer to the following:

DiMeLo-seq v1: https://doi.org/10.17504/protocols.io.bvbrtrw9

DiMeLo-seq v2: https://doi.org/10.17504/protocols.io.bzq8pzwr

pA–Hia5 protein purification: https://doi.org/10.17504/protocols.io.bov2n9yc

AlphaHOR-RES: https://doi.org/10.17504/protocols.io.bv9n966.

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 Supplementary Tables 1–3 and Supplementary Fig. 2.

**Cell culture.** HEK293T cells (CRL-3216, American Type Culture Collection; validated by microsatellite typing and mycoplasma test) were maintained in DMEM (high glucose, with Glutamax, with phenol red, without sodium pyruvate; Gibco, 10560016) supplemented with 10% FBS (VWR 89510-186) and 1% penicillin–streptomycin (Gibco, 15070063) at 37°C in 5% CO₂. GM12878 cells (GM12878, Coriell Institute; mycoplasma tested) and HG002 cells (GM24385, Coriell Institute; mycoplasma tested) were maintained in DME/F12 media (Gibco, 11095065) supplemented with 10% FBS (VWR 89510-186) and 1% penicillin–streptomycin (Gibco, 15070063) at 37°C in 5% CO₂.

Cloning of pET–pA–Hia5 and pET–pAG–Hia5. The pHiA5ET vector was generically provided by Staphylococcus aureus (SAur) pA. pA was amplified from pA19/p-MNase (ref. 50). The pHia5ET vector was linearized via NdeI restriction digest. pA or pAG was amplified from pAG/MNase (ASP4154, Addgene plasmid no. 123461; ref. 50). The pHia5ET vector was linearized via NdeI restriction digest. pA or pAG was inserted in front of the Hia5 cassette using Gibson Assembly. Peptide linker between pA (or pA/G) and Hia5 in pET-pA–Hia5 and pET-pAG–Hia5 was inserted in front of the Hia5 cassette using Gibson Assembly. Nuclei were pelleted and resuspended in 200 µl of l-glutamine (Gibco, 11875093) supplemented with 15% FBS (VWR 89510-186) and 1% penicillin–streptomycin (Gibco, 15070063) at 37°C in 5% CO₂.

**Sequencing.** Sequencing was performed on an Oxford Nanopore MinION sequencer with the following modifications: (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 (DiMeLo-seq 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. Enzyme incubation time was increased to 10 min. A total of 1 pg of RNA were used for each sample and Supplementary Fig. 2 for read length distributions. For sequencing, 500 ng of the final library was loaded, with a wash using the Flow Cell Wash Kit (ON EXP-WSSH04) and reloaded every 24h. Other approaches, such as using Zymo Genomic DNA Clean & Concentrator (D4063) for cleanup between reaction steps in the LSK110 protocol and using the Rapid Barcodekit (ON SQK-RBK-004) for cleanup after barcode ligation. For sequencing, 500 ng of the final library was loaded, with a wash using the Flow Cell Wash Kit (ON EXP-WSSH04) and reloaded every 24h. Other approaches, such as using Zymo Genomic DNA Clean & Concentrator (D4063) for cleanup between reaction steps in the LSK110 protocol and using the Rapid Barcodekit (ON SQK-RBK-004) for cleanup after barcode ligation. For sequencing, 500 ng of the final library was loaded, with a wash using the Flow Cell Wash Kit (ON EXP-WSSH04) and reloaded every 24h. Other approaches, such as using Zymo Genomic DNA Clean & Concentrator (D4063) for cleanup between reaction steps in the LSK110 protocol and using the Rapid Barcodekit (ON SQK-RBK-004) for cleanup after barcode ligation. For sequencing, 500 ng of the final library was loaded, with a wash using the Flow Cell Wash Kit (ON EXP-WSSH04) and reloaded every 24h. Other approaches, such as using Zymo Genomic DNA Clean & Concentrator (D4063) for cleanup between reaction steps in the LSK110 protocol and using the Rapid Barcodekit (ON SQK-RBK-004) for cleanup after barcode ligation. For sequencing, 500 ng of the final library was loaded, with a wash using the Flow Cell Wash Kit (ON EXP-WSSH04) and reloaded every 24h. Other approaches, such as using Zymo Genomic DNA Clean & Concentrator (D4063) for cleanup between reaction steps in the LSK110 protocol and using the Rapid Barcodekit (ON SQK-RBK-004) for cleanup after barcode ligation.
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 S2000) with 60-s spins at 4,200 r.p.m. 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 and 5067-5366), and DNA quantification was performed using the Qubit (Invitrogen, Q32583).

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 (v1.0) was added for the 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 h 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 2h, as recommended for the HiFi sequencing application.

Circular consensus sequencing (CCS) was performed in SMRT Link v10.1 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; the percentage 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 12,048 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 was 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 (Extended Data Fig. 10a). 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 cutoff.

Genomic DNA was extracted from ~25 million cells using an NEB HMBW DNA extraction kit using rotation at 300 r.p.m. during lysis (T3050L). The DNA was eluted in a total of 300 µl EB and allowed to relax at 4 °C for 2 d, although it remained viscous until it was solubilized. In total, 37 µl NEBuffer 2.1 was added, along with 100 units of MscI and 100 units of AseI (NEB nos. R0334M and R0256M) to a total volume of 370 µl in a 1.5-ml lo-bind microcentrifuge tube. This was placed on a rotator at 12 r.p.m. at 37 °C overnight. DNA concentration was then quantified using a Qubit Broad Range DNA kit (Thermo Fisher, Q32580). 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 10 µl SYBRSafe gel stain (Thermo Fisher, SS3102) per 100 µl gel. A GeneRuler High Range Marker Ladder (Thermo Fisher, SM3181) 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⁻¹ for 1 h and imaged over a blue-light transilluminator. The gel was cut to remove fragments smaller than 20kb, 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 rotor at 12 r.p.m., and DNA was eluted from the column twice with the EB heated to 70 °C. The DNA was then quantified by Qubit again. DNA was prepared for sequencing using an ONT LSK109 Native Library Prep Kit, and sequenced on a v4.9 MinION flow cell. CENP-A-targeted DiMeLo-seq was performed on unfixed H1299 cell extracts targeted, free-floating pA–HIA, and untreated samples. For each treatment, ~25 million cells were processed in five 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 µg of recovered DNA from each sample was then processed for nanopore sequencing using method 1, described above.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
All raw fastq sequencing data are available in the Sequence Read Archive (SRA) under BioProject accession PRJNA752170. These data were used to produce Figs. 2–6, Extended Data Figs. 1–10, Supplementary Tables 1–3 and Supplementary Fig. 2. The CTCF ChIP–seq peak bed file for GM12878 is available from ENCODE Project Consortium under accession ENCSR797DL1. The ATAC–seq peak bed file for GM12878 is available from ENCODE Project Consortium under accession ENCSR784UHZ1. Bulk and scDamiD data were obtained from the Gene Expression Omnibus (GEO) under accession GSE156150. H3K9me3 CUT&RUN data are from Altomese et al. and accessible in the SRA with BioProject accession PRJNA527955. Data for Fig. 6c used CHM13 CENP-A ChIP–seq data for CENP-A k-mer analyses, which are available under BioProject accession PRJNA559484 from Logsdon et al. Centromere and HOR definition bed files from the telomere-to-telomere consortium can be found at https://github.com/marbl/chr13. Known CTCF motifs are from http://compbio.mit.edu/encode-motifs/ matches.txt.gz. Data for the CpG methylation track in Fig. 6d were obtained from datasets available at https://github.com/nanopore-wgs-consortium/CHM13 (ref. 35). Source data are provided with this paper.

Code availability
The code to reproduce the results in this manuscript is available on https://github.com/amaslan/dime-lo-seq/.

References
50. Meers, M. P., Bryson, T. D., Henikoff, J. G. & Henikoff, S. Improved CUT&RUN chromatin profiling tools. Elife 8, e46314 (2019).
51. Cao, S., Zhou, K., Zhang, Z., Lugter, 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).
52. Zhou, K. et al. CENP-N promotes the compaction of centromeric chromatin. bioRxiv https://doi.org/10.1101/2021.06.14.448351 (2021).
53. Kim, B. Y. et al. Highly contiguous assemblies of 101 drosophilid genomes. Preprint at bioRxiv https://doi.org/10.1101/2020.12.14.422775 (2020).
54. Hellman, A. & Chess, A. Gene body-specific methylation on the active X chromosome. Science 315, 1141–1143 (2007).

Acknowledgements
We thank A. Stergachis for the plasmid encoding Hia5, G. Caldas for experimental training, and G. Karpen for helpful discussions. We thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. We thank M. Tan for contributions to sequencing. This work was supported by the Chan Zuckerberg Biohub and by the NIGMS of the National Institutes of Health under award no. R35GM124916 to A.S. and R01 GM074728 to A.F.S. K.H.M. is supported by R21HG010548-01. O.K.S. and R.R.B. are supported by the National Institutes of Health T32 award, nos. GM113854-02 and GM007279-45, respectively. A.M., O.K.S. and R.R.B. are supported by NSF GRFP awards. N.A. is an HHMI Hanna H. Gray Fellow. A.S. is a Chan Zuckerberg Biohub Investigator, and a Pew Scholar in the Biomedical Sciences.

Author contributions
N.A., O.K.S., K.S., A.F.S. and A.S. designed the study. N.A., A.M., O.K.S., K.S. and R.R.B. performed the experiments. A.M.D. and N.N. assisted with sequencing and provided feedback. K.H.M. provided unpublished datasets and feedback. R.M. assisted with analysis software development. N.A., A.M., O.K.S. and K.S. analyzed and interpreted the data. N.A., A.M., O.K.S. and K.S. wrote the manuscript, with input from R.R.B., A.F.S. and A.S. and A.S. supervised the study.

Competing interests
N.A., O.K.S., 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.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41592-022-01475-6. Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01475-6. Correspondence and requests for materials should be addressed to Aaron F. Straight or Aaron Streets. Peer review information Nature Methods thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Lei Tang, in collaboration with the Nature Methods team. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | 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 indicates 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 Fig. 2d).
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | In vitro assessment of methylation of 18×601 array chromatin by pA-Hia5 and pAG-Hia5. a, Schematic showing the location of 601 sequences (grey boxes) and Aval digestion sites (dashed line) in between 601 sequences on the 18×601 array. Yellow hexagons indicate biotinylation. b, Schematic of methylation of 18×601 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 Aval digested naked 18×601 array DNA or 18×601 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 18×601 array (g) or a subset 4×601 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 18×601 array or 4×601 portion.(red dashed line indicates 601 dyad position). i, Violin plot of nucleosomes detected per read on reads from CENP-A or H3 18×601 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,i, Same as g,h, but corresponding to H3 chromatin methylation with H3-directed pAG-Hia5.
**Extended Data Fig. 3 | Assessment of mA calling and LMNB1 targeting.**

**a.** The proportion of all adenines called as methylated at each possible 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 HiaS in vitro. The untreated DNA provides a measure of the false positive rate (FPR) at each score, since it contains few or no methyladenines. The HiaS 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 HiaS-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 HiaS-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.
Extended Data Fig. 4 | See next page for caption.
Extended Data Fig. 4 | 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 Supplementary Table 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 = 2.8 M) or ciLADs (blue, representing off-target methylation; n = 2.1 M). 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 Fig. 3e, 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.
Extended Data Fig. 5 | See next page for caption.
Extended Data Fig. 5 | 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-HiaS; opt3: 2 hour activation, 0.05 mM spermidine at activation, replenish SAM, pA-HiaS 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 noticeably 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 Fig. 4a.
Extended Data Fig. 6 | See next page for caption.
Extended Data Fig. 6 | Control mA and mCpG profiles at CTCF peaks. Profiles at CTCF ChIP-seq peaks for free pA-HiaS, 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 Fig. 4a, 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_SmC_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.
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | 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 Fig. 4d. d, Aggregate enrichment profiles from DiMeLo-seq reads across all CTCF sites on chrX are shown, as in Fig. 4b. 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. 

12, 54.
Extended Data Fig. 8 | See next page for caption.
Extended Data Fig. 8 | 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.
Extended Data Fig. 9 | 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 Fig. 5d.
Extended Data Fig. 10 | See next page for caption.
Extended Data Fig. 10 | 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 MscI and AseI 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.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

☐ n/a Confirmed

☐ □ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐ □ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ □ The statistical test(s) used AND whether they are one- or two-sided

☐ □ Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ □ A description of all covariates tested

☐ □ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ □ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ □ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

Give P values as exact values whenever suitable.

☐ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ □ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection MinKNOW 21.02.1; softWoRx 4.1.0 [Applied Precision]

Data analysis Guppy [v4.4.2, v4.5.4, v5.0.7], Megalodon [v2.2.9, v2.3.1], Winnowmap [v2.03], bedtools [v2.28.0], pysam [v0.15.3], scpy [v1.4.1], macs2 [v2.1.1], plotly [v4.5.2], bamCoverage [v3.3.1], pygenometracks [v3.6], samtools [v1.8], matplotlib [v3.3.4]; methplotlib [v0.17.0] from De Coster, Wouter, Endre Bakken Stovner, and Mojca Strazisar. 2020. “Methplotlib: Analysis of Modified Nucleotides from Nanopore Sequencing.” Bioinformatics 36 (10): 3236–38. Custom scripts that are available on Github: https://github.com/amaslan/dimelo-seq.

For manuscripts utilizing custom algorithms or software that are not central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All raw fastq sequencing data are available in the SRA with BioProject accession PRJNA752795. These data were used to produce Figures 2-6, Extended Data Figures 1-10, Supplementary Tables 1-3, and Supplementary Figure 2. CTCF ChIP-seq peak bed file for GM12878 is available from ENCODE Project Consortium with accession code ENCF797SOL. ATAC-seq peak bed file for GM12878 is available from ENCODE Project Consortium with accession code ENCF748UZH. Bulk and single-cell DamID data were obtained from GEO with accession GSE156150. H3K9me3 CUT&RUN data are from Altsemou et al. 2022 and accessible in the SRA with BioProject accession PRJNA752795. Data for Figure 6c used CHM13 CENP-A ChIP-seq data for CENP-A kmer analyses which are available at Bioproject accession
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-list.pdf](http://nature.com/documents/nr-reporting-summary-list.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
No sample-size calculations were performed. Sample size was determined to be adequate based on the magnitude and consistency of measurable differences between replicates.

Data exclusions
No data were excluded.

Replication
Attempts at replication were successful. For CENP A targeting and corresponding controls, samples were processed in duplicate and reads were pooled. For LMNB1, CTGF, and H3K9me3 targeting, at least two samples were run per target and results were reproducible. For in vitro experiments, results of CENP-A or H3 directed methylation, and non-targeted (free pA(G)-H4A5) methylation on CENP-A or H3 chromatin were reproducible. In vitro experiments were repeated at least 3 times with reproducible results. The results were repeated with different preparations/reconstitutions of DNA, chromatin, purified pA4-H4A5, and buffers. The experiments were performed independently by at least two authors, OKS and KS, to reduce researcher/technical biases.

Randomization
Randomization was not relevant to this study. All experiments were performed on chromatin or DNA from a mixed population of cells or in vitro reconstituted reagents. For a given experiment, the same material (cells or reconstituted chromatin) were split equally amongst the different conditions being tested.

Blinding
Blinding was not relevant to this study. For sequencing experiments reported in this manuscript, samples were barcoded for multiplexed long-read sequencing. The barcoded samples were pooled and sequenced together. Following basecalling, the libraries from these sequences were de-multiplexed and analyzed identically allowing comparisons between libraries.

For imaging based analyses of chromatin-coated beads, following treatment of beads according to the different conditions, all beads were processed simultaneously on coverslips. Coverslips were then treated identically, numbered, and imaged identically irrespective of conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☑   | Antibodies            |
| ☑   | Eukaryotic cell lines |
| ☑   | Paleontology and archaeology |
| ☑   | Animals and other organisms |
| ☑   | Human research participants |
| ☑   | Clinical data         |
| ☑   | Dual use research of concern |

### Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☑   | ChIP-seq              |
| ☑   | Flow cytometry        |
| ☑   | MRI-based neuroimaging |

### Antibodies

**Antibodies used**

- Anti-Lamin B1 antibody - Nuclear Envelope Marker [Abcam; ab36048; lots GR3369248-1, GR338070-1, GR328478-1], Recombinant Anti-CTCF antibody [EPR18253] - CHIP Grade [Abcam; ab188408; lot GR295192-3; clone EPR18253], Histone H3K9me3 antibody (pAb) [Active Motif 39162; lot 2002301], CENP-A (Aaron Straight, Stanford University; Cao et al. 2018, Zhou et al.), rabbit IgG, polyclonal - Isotype Control (CHP Grade) [Abcam; ab17180], lot GR3353004-2, Goat anti-Mouse IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor Plus 647 (Invitrogen A32728, lot U0729293), FITC Anti-V5 tag antibody (Abcam; ab1274, GR3350632-I), Anti-N6-methyladenosine (m6A) Antibody (Millipore Sigma A303174; lot Q3517443), Histone H3 antibody (mAb) (Active Motif 39064, lot 19500024, clone MAB0301), ChromPure Rabbit IgG, whole molecule (Jackson Immuno Research 011-000-003, lot 137605), Goat anti-Rabbit IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 647 (Invitrogen A21245, lot 18085235), ChromPure Mouse IgG, whole molecule (Jackson Immuno Research 015-000-003, lot 1329980), CTFE antibody (pAb) (Active Motif 61312; lot 09121009), FITC Anti-GX His tag antibody (Abcam; ab3554; lot GR3351791-2)
For all antibodies except CENP-A, validation was performed by the provider and refer to manufacturers’ websites. Rabbit anti-CENP-A antibody was generated in Aaron Straight’s lab. The antibody was verified and validated in Aaron Straight’s lab using western bloting against purified CENP-A protein, western blotting against endogenous CENP-A in cultured cells, and immunofluorescence localization at centromeres in human cells (unpublished results; Cao et al. 2018; Zhou et al.).

### Eukaryotic cell lines

#### Policy information about cell lines

| Cell line source(s) | HEK293T cells (CRL-3216, ATCC, Manassas, VA); GM12878 cells (Coriell Institute, Camden, NJ); HG002 cells (Coriell Institute, Camden, NJ) |
|---------------------|-----------------------------------------------------------------------------------------------------------------------------------|

| Authentication | For HEK293T, the Certificate of Analysis provided states that STR analysis was performed and that the identified cell type was 293T/17 with 100% match. For GM12878 and HG002 cells from Coriell, the quality control documentation states: “Family relationships have been verified by Southern blot hybridization with minisatellite (VNTR) probes or by PCR using a panel of microsatellite markers. Gender has been verified by PCR with a Y chromosome-specific primer pair. Replicate cultures or matched cultures of differing cell types from the same individual have been analyzed by PCR using microsatellite and Y chromosome-specific primer pairs to assure cell culture identity.” (https://www.coriell.org/0/sections/support/global/ QCcells.aspx?gid=409) |

| Mycoplasma contamination | All cell lines tested negative for mycoplasma contamination. |

| Commonly misidentified lines (See ITAC register) | No commonly misidentified lines were used. |