Integrated analysis of multimodal long-read epigenetic assays

Jeremy Marcus*[,1], Oberon Dixon-Luinenburg*[,1,2], Nathan Gamarra[3], Jacob P. Schwartz[4], Michal Rozenwald[1], Annie Maslan[2], Fyodor D. Urnov[5,6], Aaron F. Straight[7], Nicolas Altemose[3,8], Nilah M. Ioannidis[1,8,9], Aaron Streets[1,8,10]

1. Center for Computational Biology, University of California, Berkeley, California 94720, USA.
2. UC Berkeley–UCSF Graduate Program in Bioengineering, University of California, Berkeley, California 94720, USA.
3. Department of Genetics, Stanford University, Palo Alto, California 94304, USA.
4. Department of Molecular and Cellular Physiology, Stanford University, Palo Alto, California 94304, USA.
5. Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA.
6. Innovative Genomics Institute, University of California, Berkeley, California 94720, USA.
7. Department of Biochemistry, Stanford University, Palo Alto, California 94304, USA.
8. Chan Zuckerberg Biohub – San Francisco, San Francisco, California 94158, USA.
9. Department of Electrical Engineering and Computer Sciences, University of California Berkeley, California 94720, USA
10. Department of Bioengineering, University of California Berkeley, Berkeley, California 94720, USA.

* These authors contributed equally to this work

Corresponding author Aaron Streets, astreets@berkeley.edu

Abstract

Long-read sequencing assays that detect base modifications are becoming increasingly important research tools for the study of epigenetic regulation, especially with the development of DiMeLo-seq and similar methods that deposit non-native base modifications to mark a range of epigenetic features such as protein-DNA interactions and chromatin accessibility. A main benefit of these methods is their inherent capacity for multimodality, enabling the encoding of multiple genomic signals onto single nucleic acid molecules. However, there are limited tools available for visualization and statistical analysis of this type of multimodal data. Here we introduce *dimelo-toolkit*, a python package built to enable flexible visualizations and easy integration into custom data processing workflows. We demonstrate the utility of *dimelo-toolkit*'s preset visualizations of multiple base modifications in long-read single-molecule sequencing data with a novel extension of the DiMeLo-seq protocol that can capture three separate aspects of chromatin state on the same single reads: target protein binding, CpG methylation, and chromatin accessibility. We apply this multimodal method to simultaneously map chromatin accessibility, CpG methylation, and LMNB1 and CTCF binding patterns, respectively, in

GM12878 cells. Additionally, we use *dimelo-toolkit* to investigate technical biases that arise when working with this type of multimodal data. This software tool will pave the way for developing well-optimized protocols and help unlock previously inaccessible biological insights.

Introduction

The genome in higher eukaryotes encodes a vast set of distinct cellular phenotypes, requiring sophisticated machinery to regulate gene activity and maintain cell state. This epigenetic regulation relies on the coordination of direct nucleotide modifications in the form of DNA methylation, binding and modification of proteins such as transcription factors and histones, and larger-scale structural features including chromatin accessibility and topologically associated domains[1]. Genome-wide measurement of these epigenetic features has proven transformative for understanding the structure and function of the genome[2–4]. However, performing multiple epigenetic measurements simultaneously in the same cells remains difficult using standard technologies.

The most widely used methods for assessing epigenetic state currently rely on short-read sequencing. Methods for measuring protein-DNA interactions or chromatin accessibility rely on fragmentation of the genome and subsequent fragment enrichment. Following amplification, sequencing, and alignment, sequencing coverage at genomic loci is used as a proxy for the phenomenon of interest. Protein-DNA interaction assays such as DamID[5,6], ChIP-seq[7], and CUT&TAG[8] enrich for reads which are bound by a target protein of interest, while chromatin accessibility assays such as DNAse-seq[9] and ATAC-seq[10] enrich for reads from accessible regions of the genome. These short-read assays all rely on the same final output: the number of reads which map to each genomic location. As a result, they are mutually incompatible and must be performed separately. Additionally, short-read sequencing technologies have important limitations. These methods do not directly measure DNA modifications, instead requiring base conversion protocols such as bisulfite sequencing[11] or EM-seq[12]. Furthermore, the short length of the sequenced fragments makes it difficult to map reads to repetitive sequences or perform high-confidence haplotype phasing.

A new class of assays utilize third-generation long-read sequencing and DNA modifications to interrogate epigenetic state. Assays such as DiMeLo-seq[13,14], Fiber-seq[15], SAMOSA[16], nanoNOME-seq[17], and others[18–27] write information about protein-DNA interactions or chromatin accessibility directly onto the DNA itself through the introduction of base modifications in exogenous motif contexts proximal to phenomena of interest. These modifications can be detected directly alongside primary DNA sequence and endogenous methylation on reads many kilobases in length using Oxford Nanopore Technologies (ONT) or Pacific Biosciences HiFi sequencing technologies. Long-read data enables applications including quantitative haplotype-specific chromatin actuation, mechanistic understanding of the ultra-repetitive regions of the centromere, and assessment of coordinated binding events on single DNA molecules.

As these experimental methods have advanced, software tools to visualize and analyze this new data modality remain an area of active work. Visualizing local and global patterns and performing genome-wide statistical analyses requires tools beyond those typically used for read-coverage-based short-read assays. Furthermore, the wide array of possible exogenous modifications, such as methylation and hydroxymethylation, in different motif contexts, such as CpG, GpC, and CpC, requires a flexible and modular approach. We previously developed a software package for visualization of DiMeLo-seq data named *dimelo*[14], which provided plotting tools hardcoded for adenine and CpG methylation. To support further development of modification-based long-read assays, we have expanded the scope of past software built for DiMeLo-seq and released the new *dimelo-toolkit*. By using *modkit*[28] to parse sequencing data and then optimizing data loading and processing within our core package, key operations are sped up from hours or days with the previous tool to seconds or minutes with *dimelo-toolkit*. We have maintained the easy-to-use interfaces of the previous tool while implementing a new modular and flexible design to facilitate both end-to-end pipelines for high-quality visualizations and vectorized data extraction for more advanced workflows. Additionally, we have added the capability to process any type of base modification in any sequence context, enabling a wider range of experimental applications. Speeding up parsing and loading enables quick iteration in analysis, while the flexibility enables compatibility with a wide range of assays and pipeline configurations.

dimelo-toolkit facilitates the analysis of datasets containing an arbitrary number of nucleotide modifications, creating the opportunity to build experimental methods which encode multiple epigenetic features onto single chromatin fibers. These multimodal assays could enable study of the biological interactions between epigenetic phenomena, such as the function of co-occurring protein-DNA binding events or relationships between binding events and chromatin structure. Here, we leverage the multimodal capabilities of *dimelo-toolkit* to drive analysis for a new protocol that simultaneously measures protein-DNA interaction, chromatin accessibility, and CpG methylation as part of a single assay. We characterize these multiplexed measurement signals and correct for signal crosstalk, thus enabling high-read-depth, single-molecule, and haplotype-resolvable information for all three signals.

Results

dimelo-toolkit workflow

dimelo-toolkit defines a systematic workflow for working with modified nucleotide data in python and producing publication-quality visualizations (Figure 1). This allows users to easily interpret and analyze the results of experiments which rely on nucleotide methylation state as the primary output modality, whether through measurement of endogenous methylation patterns or the introduction of exogenous methylation signals.

As a primary input, the package accepts aligned and base-called long-read sequencing data in a BAM format file containing the appropriate base modification tags (MM and ML) for an

arbitrary number of endogenous and exogenous modification types, such as 5mC and 6mA. To disambiguate between methylation signals in different contexts (i.e, 5mC in CpG vs. GpC motifs, or 6mA specific to GATC motifs), the user can specify the genomic motifs at which to extract methylation events using IUPAC standard nucleotide codes, a zero-indexed modification position, and optionally a mod code. For example, GCH,1,m specifies a motif with guanine at position 0, a 5-methylcytosine at position 1, and an adenine, cytosine, or thymine at position 2. Additionally, many of the methods in *dimelo-toolkit* allow the user to separately specify genomic regions of interest. Given these inputs, *dimelo-toolkit* defines two major workflow branches. In the pileup operation, reads are collapsed into a single value for each genomic coordinate, representing the fraction of the reads covering each position which exhibit a given modification. This enables comparison of relative methylation rates in different genomic regions, or visualization of general methylation profiles across regions of interest. In the extract operation, modification data is tabulated in a given genomic region for each read, reformatting single-read data for more efficient downstream processing. Both operation branches utilize modkit, an optimized tool created and published by ONT for parsing modified bases stored in BAM format [28]. Outputs from these initial parsing operations are saved as compressed intermediate files to enable reuse in downstream analyses. The package provides methods for producing preset plots from these intermediate files, covering several common use cases for visualization of modified base data. Additionally, the methods for loading parsed data are made available to the user to enable the building of custom python analysis pipelines.

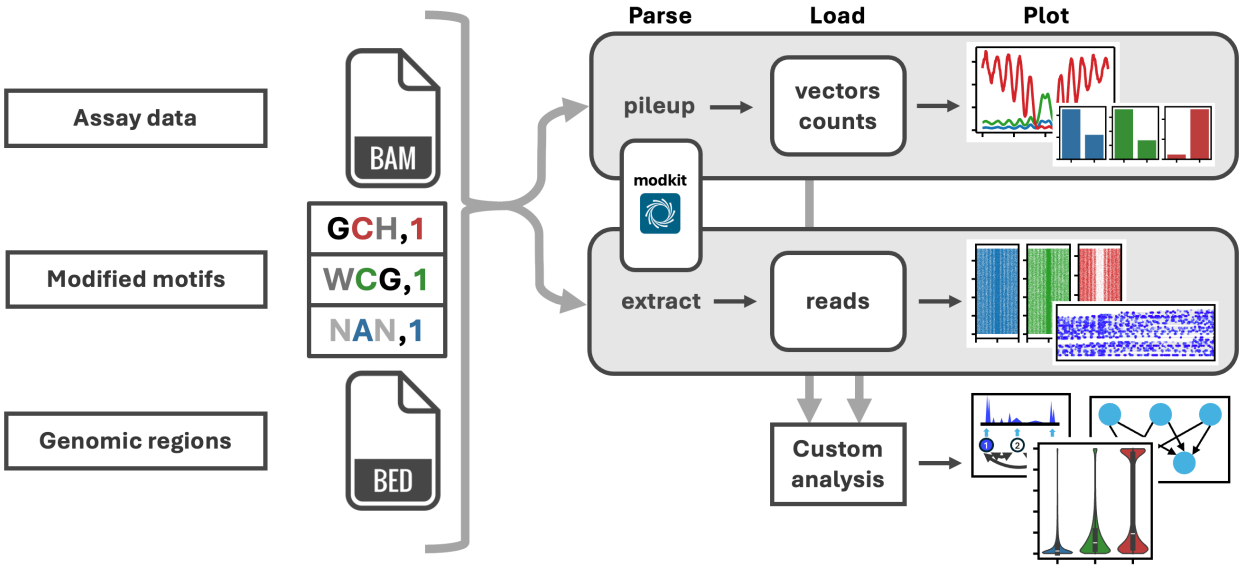


Figure 1. dimelo-toolkit workflow. *dimelo-toolkit accepts base-called, modification-tagged, and aligned sequencing data in a BAM format file as primary input, along with base modification codes and sequence contexts for an arbitrary number of modification types. The user may specify genomic regions of interest via a BED file or region strings. Given these inputs, dimelo-toolkit can either collapse modification information into a single value for each genomic coordinate (pileup), representing the fraction of the reads covering each position which exhibit a given modification, or tabulate information for each read in each region (extract). Parsing*

operations create compressed and indexed intermediate files from outputs generated by ONT modkit, enabling quick iteration when prototyping new downstream analyses. The package provides methods for producing preset plots from these intermediate files. Additionally, the methods for loading parsed data are made available to the user to enable the building of custom python analysis pipelines, ranging from custom data visualizations to training models on large volumes of data.

Integrated multimodal analysis with *dimelo-toolkit*

We extended the original DiMeLo-seq protocol[13,14] to make a multimodal version of the method, adapting it to enable simultaneous targeting of protein-DNA interactions and chromatin accessibility using distinct methyltransferases (Figure 2a). Combining antibody-targeted methylation with existing untargeted methylation techniques for measuring accessibility[15–17] enables measurement of protein-DNA interaction, chromatin accessibility, and endogenous methylation on the same single-molecules in a single experiment. To mark protein-DNA interactions, we use a fusion of the nonspecific deoxyadenosine methyltransferase Hia5 and the antibody-binding protein A (pA-Hia5) to deposit 6mA near targets of interest. To mark accessible chromatin, we use the GpC methyltransferase M.CviPI to deposit 5mC at accessible GpC motifs. Methylation events representing each distinct epigenetic feature can be read in tandem using modification-sensitive long-read sequencing technologies such as Nanopore sequencing.

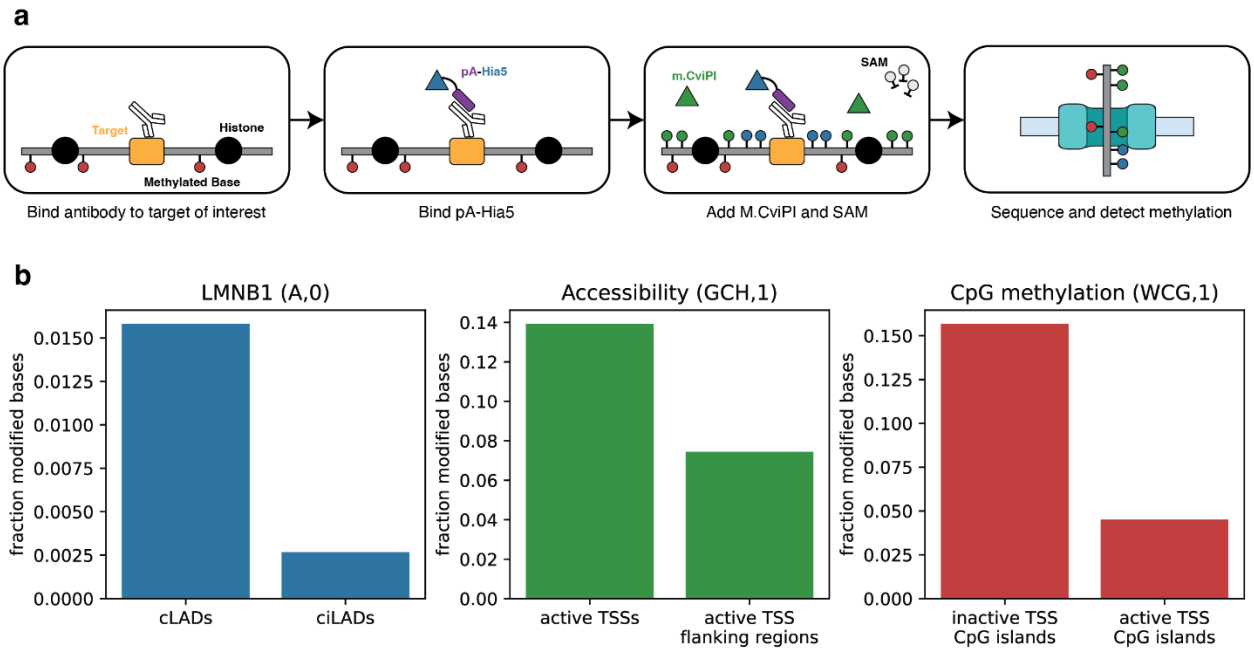


Figure 2. Multimodal DiMeLo-seq. a*, Schematic of the multimodal DiMeLo-seq protocol for simultaneously mapping protein-DNA interactions, chromatin accessibility, and endogenous methylation. Epigenetic features are encoded on DNA using exogenous methylation, which can be read alongside endogenous methylation using long-read sequencing technologies.* ***b****, Barplots of modified base fractions in known on- and off-target regions for LMNB1 occupancy*

(blue), accessibility (green), and endogenous CpG methylation (red), measured simultaneously in an LMNB1-targeting multimodal DiMeLo-seq experiment in GM12878 cells. Pileups were performed with a modification probability threshold of 0.95.

To achieve optimal methylation performance, M.CviPI requires the reducing agent DTT, which is not present in the Hia5 buffer formulation described in the previously published DiMeLo-seq protocol[13,14]. It has been shown previously that antibodies can tolerate high concentrations of DTT and retain binding capability, and that Hia5 methylates efficiently in DTT-containing buffers[29]. We additionally found that Hia5 and M.CviPI can methylate simultaneously in Hia5 buffer supplemented with DTT, suggesting that there is minimal biochemical impact on enzyme performance when combining these methyltransferases (Supplementary Figure 1).

We applied the multimodal DiMeLo-seq method to simultaneously profile interaction sites of proteins of interest along with chromatin accessibility and endogenous CpG methylation in GM12878 cells. To validate the performance of this approach, we first targeted lamin B1 (LMNB1), which was previously used to optimize and validate the performance of DiMeLo-seq because lamina associated domains (LADs) are well-characterized genome-wide[6]. For each signal, we quantified the proportion of methylated motifs across all reads mapping to appropriate on-target and off-target regions and plotted the relative enrichment using the *plot_enrichment* method of *dimelo-toolkit* (Figure 2b). Relative methylation levels for each methylation signal followed the expected pattern of enrichment in on- and off-target regions, indicating that multimodal DiMeLo-seq can interrogate large-scale genome structure in the context of chromatin accessibility and endogenous methylation on single chromatin fibers.

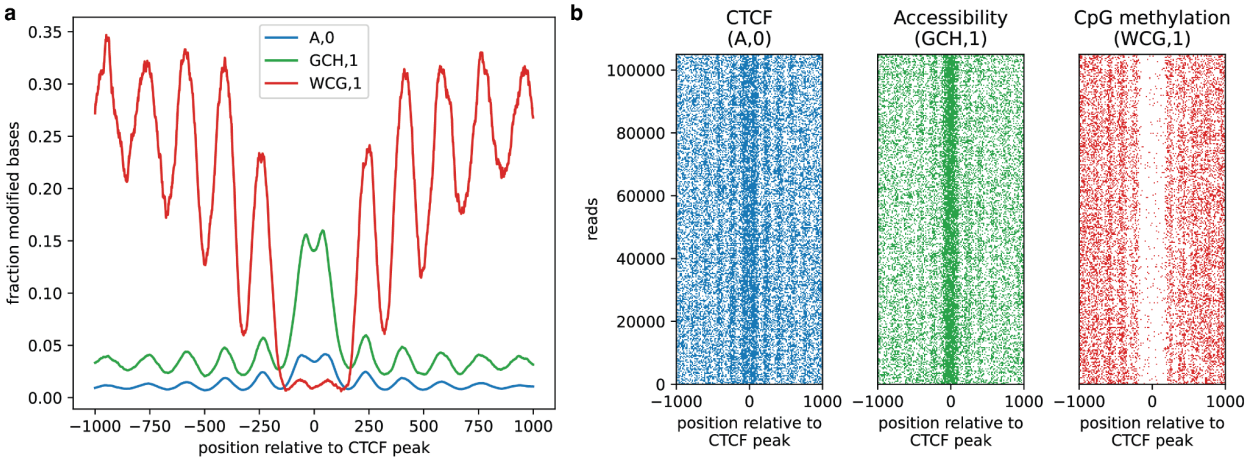


***Figure 3. Multimodal analysis of CTCF binding sites. a**, Enrichment profiles describing patterns of CTCF occupancy (blue), accessibility (green), and endogenous CpG methylation (red), as measured simultaneously in a CTCF-targeting multimodal DiMeLo-seq experiment in GM12878 cells. Modified bases reported with probability >0.95 were aggregated at 3,000 high-signal CTCF ChIP-seq peaks overlapping known CTCF motifs and plotted using a 25-bp rolling window. **b**, Single molecule plots for the reads used to generate the aggregate profiles in **(a)**.*

To demonstrate the ability to visualize colocalization of protein interactions, accessibility, and endogenous CpG features, we chose to examine binding sites for CCCTC-Binding factor (CTCF). CTCF binding is inhibited by the presence of CpG methylation, and results in strong positioning of surrounding nucleosomes. We targeted CTCF with 6mA, measured chromatin accessibility with GpC methylation, and detected endogenous CpG methylation in GM12878 cells. We then investigated the resulting signals in aggregate at strong CTCF binding sites defined by ChIP-seq, using the *plot_enrichment_profile* and *plot_reads* methods of dimelo-toolkit to visualize patterns in aggregate and on single molecules respectively (Figure 3a-b). We observed on-target increase of 6mA signal directly adjacent to the binding motif, inversely correlated to the CpG signal moving away from the binding sites. All three signals exhibit a periodic pattern consistent with the periodic methylation of linker DNA between strongly-positioned nucleosomes. This analysis demonstrates the ability to detect and visualize three distinct epigenetic features on long, single molecules of genomic DNA.

We then expanded our analysis to demonstrate *dimelo-toolkit*'s support for other long-read assays and flexible data export for downstream analysis. We processed data from a Fiber-seq assay on the same GM12878 cell line from another study[30] and exported all the different channels from both Fiber-seq and multimodal DiMeLo-seq in a genome-browser-compatible format using the *pileup_to_bigwig* method. This method allows the export of context-disambiguated modification profiles (e.g. WCG vs GCH), which is not natively supported in commonly used genome browsers such as IGV[31]. In Figure 4 we explore these tracks, along with data from short-read assays and ENCODE annotations, to show concordance and differences between the measurements and alignment to annotated features. This pipeline endpoint makes it easy to explore genomic loci of interest and cross-compare for more robust biological insights.

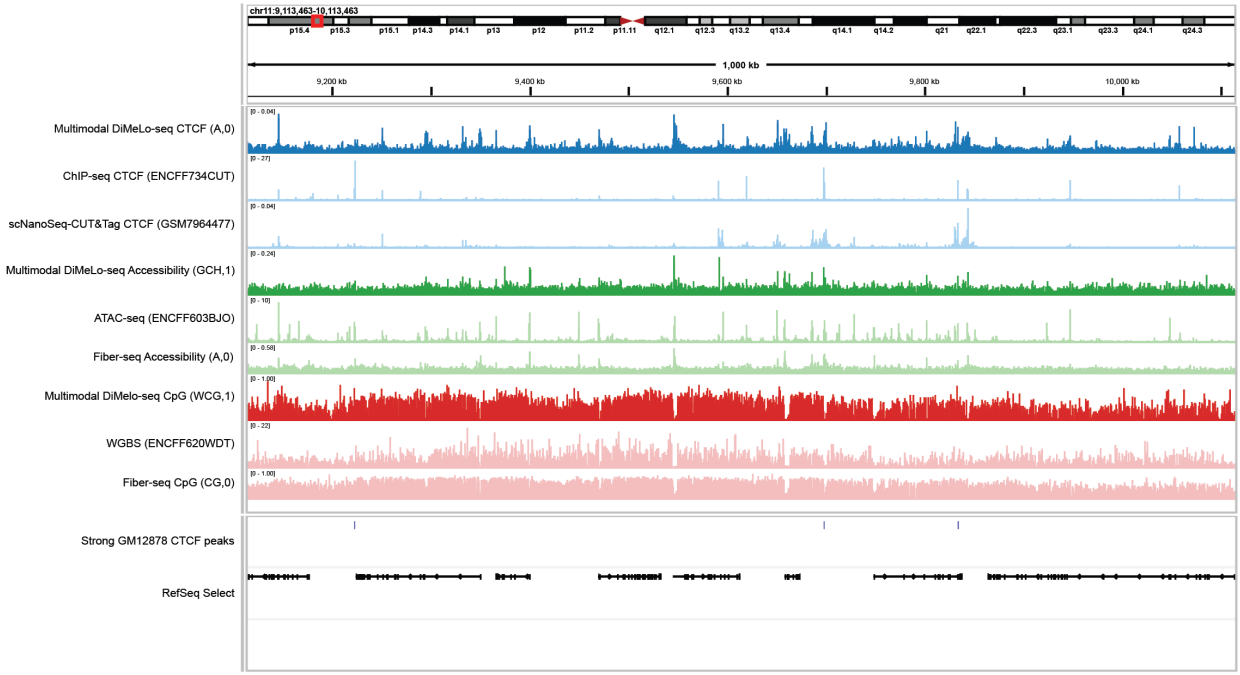


Figure 4. Genome browser comparison of multimodal DiMeLo-seq and orthogonal measurements. *Aggregate browser traces from a CTCF-targeting multimodal DiMeLo-seq experiment in GM12878 cells (dark tracks) alongside reference data from comparable orthogonal measurements (ChIP-seq, scNanoSeq-CUT&Tag, ATAC-seq, WGBS, and Fiber-seq; light tracks), visualized in a representative 1,000 kb region of chromosome 11. Multimodal DiMeLo-seq pileups with a methylation probability threshold of 0.95 were exported in bigWig format using dimelo-toolkit.*

dimelo-toolkit facilitates assessment of bias in multimodal assays

Increasing the number of nucleotide modification signals creates unexpected technical problems which must be addressed to ensure accurate interpretation of the multimodal assays discussed here. First, if two or more signals use the same type of modification in different sequence motifs to represent distinct epigenetic features, care must be taken to exclude ambiguous motifs which could be interpreted as coming from either feature. For example, in the multimodal assay used in this study, methylation at GCG motifs can ambiguously be interpreted as both biologically relevant endogenous CpG methylation and exogenous GpC methylation representing chromatin accessibility; GCG motifs must therefore be omitted from analysis. Second, basecallers are usually not trained on data containing multiple types of exogenous modifications. Nanopore sequencing devices generate electrical traces representing multiple nucleotides passing through the pore at any given time. The presence of unexpected modifications in the pore may therefore confuse the basecaller, impacting modification detection in a model-specific way. For this assay, heavy GpC methylation is an out-of-distribution signal which empirically affects 6mA and CpG methylation levels reported by the models evaluated in this study.

To assess the impacts of adding a GpC methylation step to the existing DiMeLo-seq protocol, we first evaluated the effect of different M.CviPI incubation conditions on the reported signals for CpG and 6mA in a CTCF-targeting DiMeLo-seq experiment. We used the *plot_enrichment_profile* function to visualize all three signals and make direct comparisons between conditions (Figure 5). With appropriate motif disambiguation, CpG signal was overall unperturbed by the addition of GpC methylation. 6mA signal, while reduced in the presence of GpC methylation, maintained clear peak structure and enrichment over baseline in all conditions. We selected a protocol which balances the quality of the targeted 6mA signal with enrichment for GpC accessibility and histone patterning, to maximize the useful biological insights that can be drawn from the data. To select appropriate disambiguated motifs for CpG and GpC signals for a multimodal application, we used the *plot_enrichment_profile* method to evaluate 5mC methylation detection in a variety of sequence contexts (Supplementary Figure 2). We found that the motifs with the best performance were sometimes stricter than those suggested by the methyltransferase specificities, supporting the notion that out-of-distribution modifications can perturb modification calling for nearby sites. These results led us to select a set of sequence motif specifications that minimized the overlap between accessible and endogenous methylation signals. Finally, we used the *read_vectors_from_hdf5* function to load python vectors for modification scores within reads localized to on- and off-target regions,

allowing us to easily visualize the distributions of modification probabilities reported by the *dorado* basecaller (Supplemental Figure 3, Supplemental Figure 4) and how these distributions shift in different contexts. This aided us in setting reasonable modification probability thresholds. By allowing for flexible data processing and visualization, dimelo-toolkit enabled us to differentiate biologically meaningful signal from technical bias for our multiplexed assay. The experimental conditions, motif specifications, and modification thresholds uncovered in this analysis were used throughout this study.

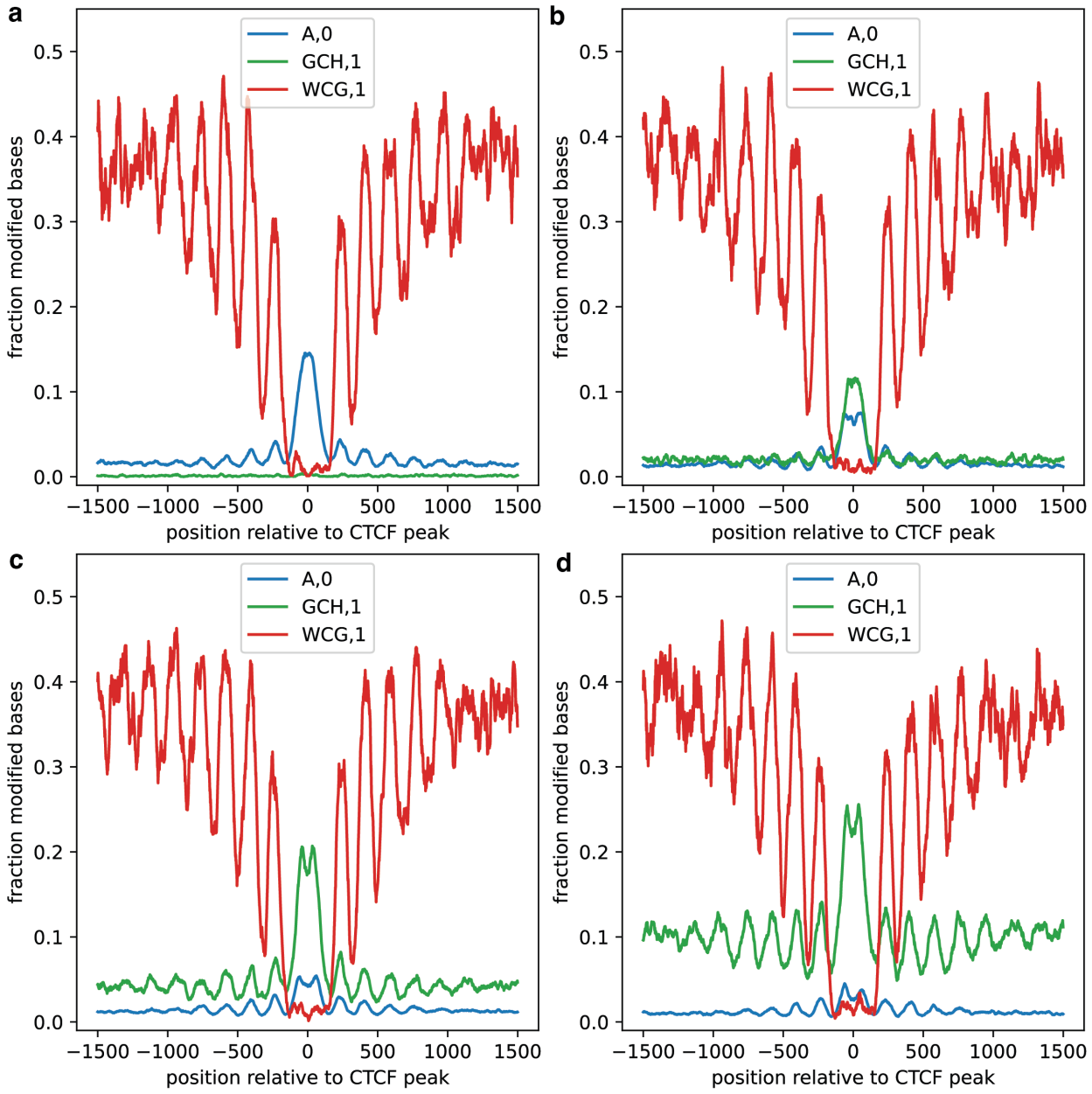


Figure 5. Evaluation of M.CviPI incubation protocols for multimodal DiMeLo-seq experiments. a-d*, Enrichment profiles describing CTCF occupancy (blue), accessibility (green), and endogenous CpG methylation (red), as measured simultaneously in CTCF-targeting multimodal DiMeLo-seq experiments in GM12878 cells. M.CviPI incubation protocols varied*

across each of the four conditions: ***(a)*** *no M.CviPI incubation,* ***(b)*** *incubation with 0.12 U/µl M.CviPI for the final 15 minutes of Hia5 incubation,* ***(c)*** *incubation with 0.12 U/µl M.CviPI for the full 2 hours of Hia5 incubation, and* ***(d)*** *incubation with 0.4 U/µl M.CviPI in a separate 15 minute incubation using M.CviPI buffer as adapted from the nanoNOMe-seq protocol*[17]*.* ***(c)*** *matches the protocol chosen for the multimodal DiMeLo-seq experiments in the remainder of this publication.*

Discussion

Here, we present *dimelo-toolkit*, a python package to facilitate analysis of assays reliant on long-read modification-aware nucleic acid sequencing technologies, such as Oxford Nanopore or PacBio HiFi. A key use case is to assist in developing and processing data from assays that involve adding exogenous nucleotide modifications to encode biological signals such as protein-DNA binding and chromatin accessibility. Our software aids users by providing a streamlined pipeline for downstream analyses and ready-made tools for producing both exploratory and publication-ready plots. We used this software package to analyze data from a new version of the DiMeLo-seq protocol which allows us to measure protein-DNA interactions and accessible chromatin on single DNA molecules in a single assay. We further demonstrated that *dimelo-toolkit* can be used to analyze and cross-compare data from many different protocols and chemistries, such as Fiber-seq sequenced with PacBio.

When analyzing data from long-read methylation assays, it is important to consider factors such as the selection of genomic regions of interest, read filtering strategies, selection of thresholds for modified basecall probabilities, and specification of modified sequence contexts. Providing an end-to-end python solution for this data makes it simple for researchers with varying levels of computational proficiency to build analysis pipelines and explore these parameters, helping users rapidly iterate, test hypotheses, and make informed data-driven assay development decisions. This process can be challenging for users because each factor cannot be considered independently. Score distributions, and thus appropriate thresholds, are impacted by sequence and modification contexts. Furthermore, protocol optimization results and disambiguation motif results can change based on the threshold used. *dimelo-toolkit* is therefore a key tool to use in this fundamentally iterative process.

Additionally, as new experimental methods are developed, a need arises for novel analytical methods for making meaningful biological inference. Providing direct access to the underlying data in python facilitates prototyping and development of tools to measure the distribution of data coming out of an assay and about the statistical confidence of biological conclusions. For example, exposing modified and unmodified counts from specified regions and sequence motifs makes it trivial to calculate p-values for differential states across conditions, paving the way for development of peak callers, comparing between cellular or chromosomal states, or integrating targeted and untargeted signals.

As new targets are explored with novel protocols for highly multiplexed analysis of epigenetic regulatory factors, *dimelo-toolkit* will accelerate data exploration and pipeline building through a unified framework for both one-click and bespoke analysis. By enabling researchers to better

work with and understand multimodal long-read epigenetic data, this toolkit stands to accelerate the field and build towards a deeper understanding of regulatory genomics.

Methods

Methyltransferase functional validation assay

Assays were performed on 100 ng of a 360 bp dsDNA oligo containing a single restriction site each for AciI and MboI. Restriction sites were spaced such that unique combinations of methylation and digestion events result in fragments of unique lengths distinguishable using a High Sensitivity D1000 Tapestation kit. Methylation reactions using three different buffer conditions: 1X GC Reaction Buffer (NEB), DiMeLo-seq Activation Buffer (15 mM Tris, pH 8.0, 15 mM sodium chloride, 60 mM potassium chloride, 1 mM EDTA, pH 8.0, 0.5 mM EGTA, pH 8.0, 0.05 mM spermidine, 0.1% BSA), and DiMeLo-seq Activation Buffer containing 10mM DTT. All reactions were carried out in 30 µl volumes containing 800 µM SAM and incubated for 2 hours at 37 °C with SAM replenishment of 800 µM at the 1 hour mark. Following methyltransferase incubation, each condition was subjected to a restriction enzyme digestion for 2 hours in CutSmart buffer at 37 °C using AciI and MboI (NEB). A control sample of unmethylated oligo was subjected to the same digestion protocol. All digestion products were resolved using a High Sensitivity D1000 Tapestation kit.

DiMeLo-seq + mGpC accessibility

Unfixed GM12878 cells were treated as described in past work, with the following modifications to accommodate the addition of an exogenous accessibility signal. Following primary and secondary antibody incubations, nuclei were resuspended in 100 µl of DiMeLo-seq Activation Buffer with the addition of 10mM DTT and 12 U of M.CviPI (NEB). Below, we provide the full protocol as adapted from previous publications[13,14].

GM12878 cells (GM12878, Coriell Institute; mycoplasma tested) were maintained in RPMI-1640 with l-glutamine (Gibco, 11875093) supplemented with 15% FBS (VWR 89510-186) and 1% penicillin–streptomycin (Gibco, 15070063) at 37 °C in 5% CO_2. After pelleting cells at 300g for 5 minutes and washing with PBS, nuclei were isolated by resuspending cells in 1 ml of Dig-Wash buffer (0.02% digitonin, 20 mM HEPES-potassium hydroxide buffer, pH 7.5, 150 mM sodium chloride, 0.5 mM spermidine, 1 Roche cOmplete EDTA-free tablet (11873580001) per 50 ml buffer and 0.1% BSA) and incubating on ice for 5 min. The resulting nuclei suspension was spun down at 4 °C at 500g for 3 min. All subsequent nuclei spins were performed with these same conditions, and all steps involving pipetting nuclei were performed with wide-bore tips. The supernatant was removed, and the pellet was gently resolved in 200 µl Tween-Wash (0.1% Tween-20, 20 mM HEPES-potassium hydroxide, pH 7.5, 150 mM sodium chloride, 0.5 mM spermidine, 1 Roche cOmplete EDTA-free tablet per 50 ml buffer and 0.1% BSA) containing the primary antibody at a 1:50 dilution. Antibodies targeted the following: LMNB1 (ab16048), CTCF (targeting C terminus, ab188408). Primary antibody incubation took place at 4 °C on a rotator

overnight. Nuclei were then pelleted and washed twice with 0.95 ml Tween-Wash, with a 5 minute rotation at 4 °C between each spin. Following the second wash, the nuclei pellet was gently resolved in 200 µl Tween-Wash containing 200 nM pA–Hia5. Secondary antibody incubation took place at 4 °C on a rotator for 2 hours. Nuclei were then washed twice with 0.95 ml Tween-Wash in the same fashion as after the primary incubation. Nuclei were resuspended in 100 µl of DiMeLo-seq Activation Buffer containing DTT and M.CviPI (15 mM Tris, pH 8.0, 15 mM sodium chloride, 60 mM potassium chloride, 1 mM EDTA, pH 8.0, 0.5 mM EGTA, pH 8.0, 0.05 mM spermidine, 0.1% BSA, 10mM DTT, 12 U M.CviPI (NEB), and 800 µM SAM) , followed by a 2 hour incubation at 37 °C with SAM replenishment of 800 µM at the 1 hour mark. DNA was extracted using the NEB Monarch Spin gDNA Extraction Kit (T3010S) and quantified using the Qubit 1x dsDNA BR Assay Kit (Q33265).

Nanopore library preparation and sequencing

For each sample, 3 µg DNA was input into library preparation using one of the following kits: (1) Ligation Sequencing Kit V14 (ON SQK-LSK114) for single-library experiments, or (2) Native Barcoding Kit 24 V14 (ON SQK-NBD114.24) for optimization experiments and paired controls. For preparations using both kits, the protocol was performed as described in the gDNA version of the documentation, with the increased incubation times recommended in the original DiMeLo-seq protocol[13,14]. Sequencing was performed on an ONT P2Solo sequencer with v10.4.1 flow cells (ON FLO-PRO114M) with MinKNOW software (versions ranging from 24.06.14 to 24.11.10).

Basecalling and postprocessing

Basecalling was performed using ONT dorado[32] (v1.0.2), using super-accuracy models for canonical bases, 5mC_5hmC, and 6mA (model complex: sup@v5.0.0,5mC_5hmC@v3,6mA@v3). Post-processing of basecalled data (trimming, aligning, demultiplexing, and summarization) was also performed using dorado, using base settings. Basecalled reads were mapped to the T2T-CHM13v2.0 reference sequence[33]. Uniquely-mapping reads of high average quality were selected based on the dorado summary (alignment_mapq == 60, mean_qscore_template > 10) and subsequently filtered using samtools[34] to arrive at the final BAM files. For all analyses, bases with a methylation probability score >0.95 were assigned as methylated.

Genomic data file manipulation

Liftover operations from grch38 to T2T-CHM13v2.0 for publicly available files (BED, bigWig, etc.) were performed using the UCSC command line utilities[35]. Additional operations (e.g. region merging, filtration, etc.) were performed using BEDtools[36] (v2.31.1) and custom code available on Github.

LMNB1 data analysis

Reference TSSs[37,38] were ranked by expression levels using GM12878 bulk RNA-seq data (ENCFF978HIY, ENCODE Project Consortium). Top quartile TSSs by expression were labeled as active and bottom quartile TSS were labeled as inactive. Accessible target regions were defined as 1000 bp windows centered at active TSSs, and inaccessible target regions were defined as two 500 bp windows flanking each accessible target region. High and low CpG target regions were respectively defined as CpG islands[37] located within 1000 bp of inactive and active TSSs. On- and off-target LMNB1 regions were respectively defined as cLADs and ciLADs identified by Altemose et al[6].

CTCF data analysis

Strong CTCF binding sites were defined as the top 3000 peaks from a set of reference GM12878 ChIP-seq peaks (ENCFF797SDL, ENCODE Project Consortium) which overlapped known CTCF motifs[39]. On-target CTCF sites were defined as 400 bp windows centered at strong CTCF binding sites, and off-target CTCF sites were defined as two 1300 bp windows flanking each on-target CTCF site.

Data Availability

Raw .pod5 files and basecalled and aligned BAM files available upon request for GM12878 LMNB1 antibody+pA-hia5+M.CviPI (Figure 2), GM12878 CTCF antibody+M.CviPI high read coverage (Figure 3,4), and CTCF antibody+pA-hia5+M.CviPI incubation time experiments (Figure 5).

Code Availability

dimelo-toolkit, along with installation instructions, documentation, and tutorials, can be found at https://github.com/streetslab/dimelo-toolkit. Code and instructions for generating the figures for this paper can be found at https://github.com/streetslab/multimodal-dimelo.

Acknowledgments

Research reported in this publication was supported by the National Human Genome Research Institute and the National Institute of General Medical Sciences of the National Institutes of Health under award number R01HG012383 to A.S. O.D.L. was supported by the National Science and Engineering Research Council of Canada. A.S. and N.A. are Chan Zuckerberg Biohub – San Francisco Investigators. A.S. is supported by the Chan Zuckerberg Initiative SDL Award and the Harvey and Leslie Wagner Foundation. N.A. is an HHMI Hanna H. Gray Faculty Fellow and a Pew Biomedical Scholar. The authors would like to thank Karen Miga, Sofia

Lundqvist, Denise Robles, Juliana Lee, and Pragya Sidhwani, Gary Karpen, and Arthur Rand for testing the software package and providing helpful discussion.

Author Contributions

J.M., O.D.L., and A.S. designed the study. J.M. and A.M. performed the sequencing assays and experiments. J.M. and O.D.L. developed the software package, performed the analysis, generated figures, and wrote the manuscript. N.G. and M.R. beta tested the software package and helped conceive of new features. J.P.S. provided essential materials. A.F.S., N.A., N.M.I., F.D.U., and A.S. supported the work and provided guidance. A.S. supervised the study.

Competing Interests

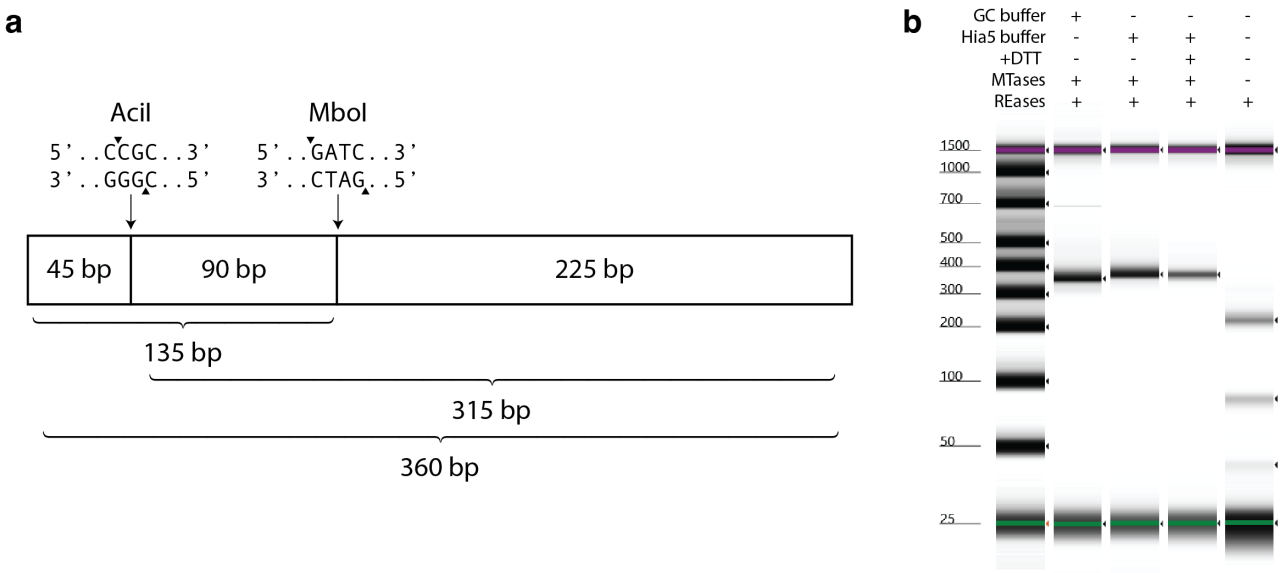
N.A., A.S., & A.F.S. are co-inventors on a patent filing related to the DiMeLo-seq method. The remaining authors declare no competing interests.

References

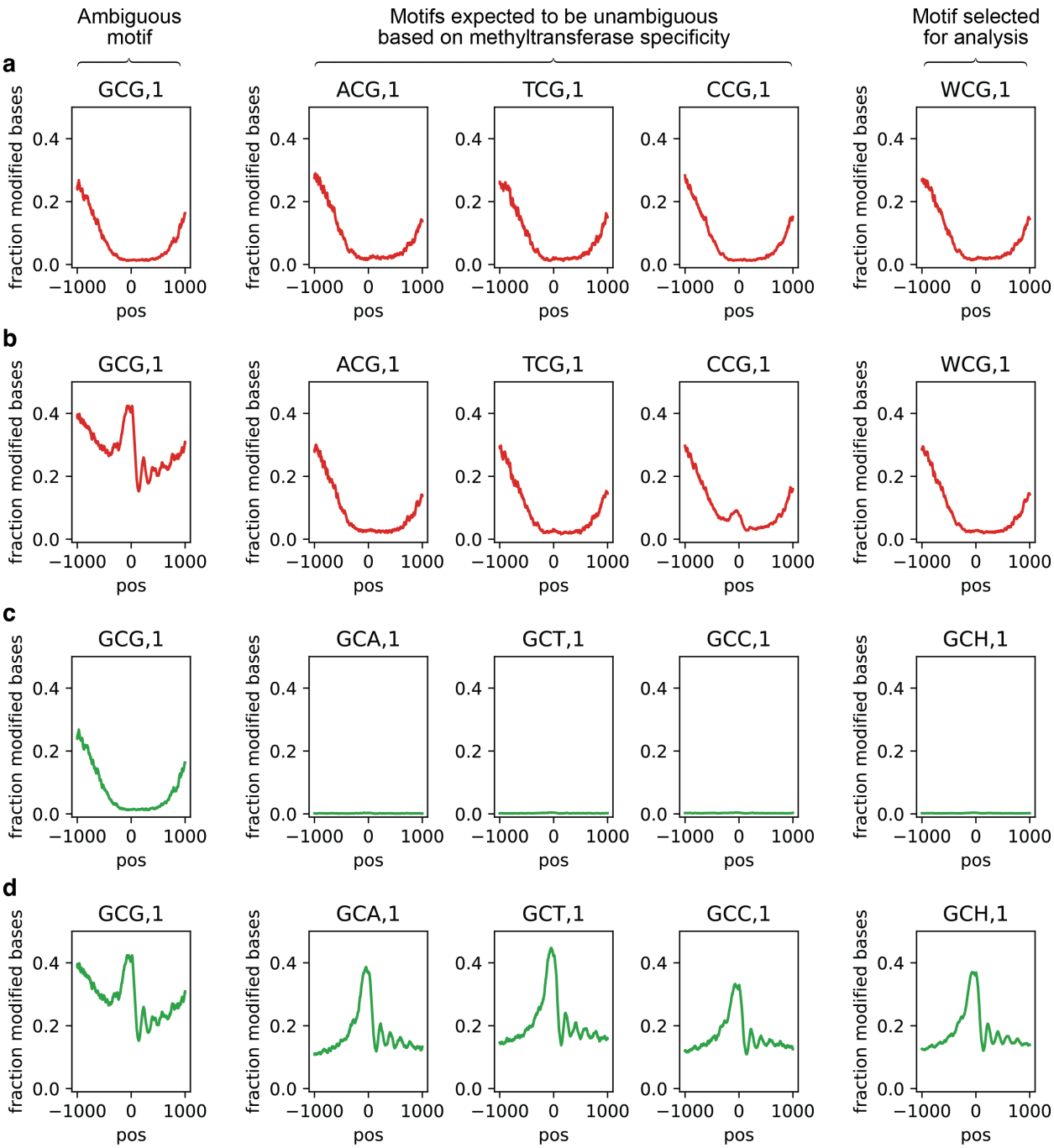
1. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17**, 487–500 (2016).
2. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
3. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
4. Hitz, B. C. *et al.* The ENCODE Uniform Analysis Pipelines. 2023.04.04.535623 Preprint at https://doi.org/10.1101/2023.04.04.535623 (2023).
5. Steensel, B. V. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nat. Biotechnol.* **18**, 424–428 (2000).
6. Altemose, N. *et al.* µDamID: A Microfluidic Approach for Joint Imaging and Sequencing of Protein-DNA Interactions in Single Cells. *Cell Syst.* **11**, 354-366.e9 (2020).
7. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**, 1497–1502 (2007).
8. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
9. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311–322 (2008).
10. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
11. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci.* **89**, 1827–1831 (1992).

12. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* **31**, 1280–1289 (2021).
13. Altemose, N. *et al.* DiMeLo-seq: a long-read, single-molecule method for mapping protein–DNA interactions genome wide. *Nat. Methods* **19**, 711–723 (2022).
14. Maslan, A. *et al.* Mapping protein–DNA interactions with DiMeLo-seq. *Nat. Protoc.* **19**, 3697–3720 (2024).
15. 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).
16. Abdulhay, N. J. *et al.* Massively multiplex single-molecule oligonucleosome footprinting. *eLife* **9**, e59404 (2020).
17. Lee, I. *et al.* Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat. Methods* **17**, 1191–1199 (2020).
18. Weng, Z. *et al.* BIND&MODIFY: a long-range method for single-molecule mapping of chromatin modifications in eukaryotes. *Genome Biol.* **24**, 61 (2023).
19. Yue, X. *et al.* Simultaneous profiling of histone modifications and DNA methylation via nanopore sequencing. *Nat. Commun.* **13**, 7939 (2022).
20. Brothers, M. & Rine, J. Distinguishing between recruitment and spread of silent chromatin structures in Saccharomyces cerevisiae. *eLife* **11**, e75653 (2022).
21. Wang, Y. *et al.* Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res.* **29**, 1329–1342 (2019).
22. Shipony, Z. *et al.* Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat. Methods* **17**, 319–327 (2020).
23. Battaglia, S. *et al.* Long-range phasing of dynamic, tissue-specific and allele-specific regulatory elements. *Nat. Genet.* **54**, 1504–1513 (2022).
24. Abdulhay, N. J. *et al.* Nucleosome density shapes kilobase-scale regulation by a mammalian chromatin remodeler. *Nat. Struct. Mol. Biol.* **30**, 1571–1581 (2023).
25. Leduque, B., Edera, A., Vitte, C. & Quadrana, L. Simultaneous profiling of chromatin accessibility and DNA methylation in complete plant genomes using long-read sequencing. *Nucleic Acids Res.* **52**, 6285–6297 (2024).
26. Mo, W. *et al.* Single-molecule targeted accessibility and methylation sequencing of centromeres, telomeres and rDNAs in Arabidopsis. *Nat. Plants* **9**, 1439–1450 (2023).
27. Chanou, A. *et al.* Single molecule MATAC-seq reveals key determinants of DNA replication origin efficiency. *Nucleic Acids Res.* **51**, 12303–12324 (2023).
28. nanoporetech/modkit. Oxford Nanopore Technologies (2025).
29. Gamarra, N. *et al.* DiMeLo-cito: a one-tube protocol for mapping protein-DNA interactions reveals CTCF bookmarking in mitosis. Preprint at https://doi.org/10.1101/2025.03.11.642717 (2025).
30. Vollger, M. R. *et al.* Synchronized long-read genome, methylome, epigenome and transcriptome profiling resolve a Mendelian condition. *Nat. Genet.* **57**, 469–479 (2025).
31. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
32. nanoporetech/dorado. Oxford Nanopore Technologies (2025).
33. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).

34. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
35. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
36. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
37. Perez, G. *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* **53**, D1243–D1249 (2025).
38. Goldfarb, T. *et al.* NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res.* **53**, D243–D257 (2025).
39. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).

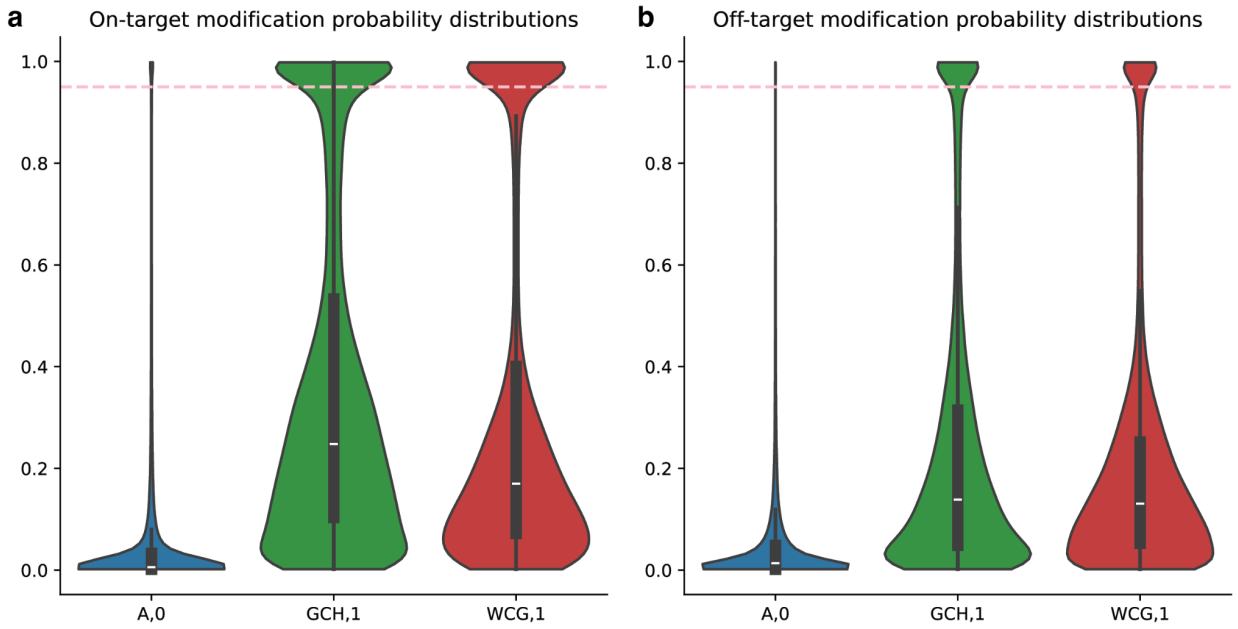


Supplementary Figure 1. Validation of methyltransferase function using restriction enzyme digestion. a, *Diagram of DNA template used for digestion assay. A 360 bp long unmethylated template was designed to contain a single restriction site each for AciI and MboI. AciI activity is blocked by cytosine methylation and MboI activity is blocked by adenine methylation. Restriction sites were placed such that unique combinations of methylation and digestion events result in fragments of unique lengths.* ***b,*** *Restriction enzyme digestion experiment testing methylation buffer conditions. DNA template from* ***(a)*** *was incubated with both M.CviPI and pA-Hia5 for 2 hours in different buffers, then subjected to restriction digest by AciI and MboI. GC buffer is NEB GC Reaction Buffer supplied with M.CviPI. Hia5 buffer is standard DiMeLo-seq methylation buffer. +DTT represents the addition of 10mM DTT. MTases and REases represent the addition of methyltransferases and restriction enzymes respectively.*

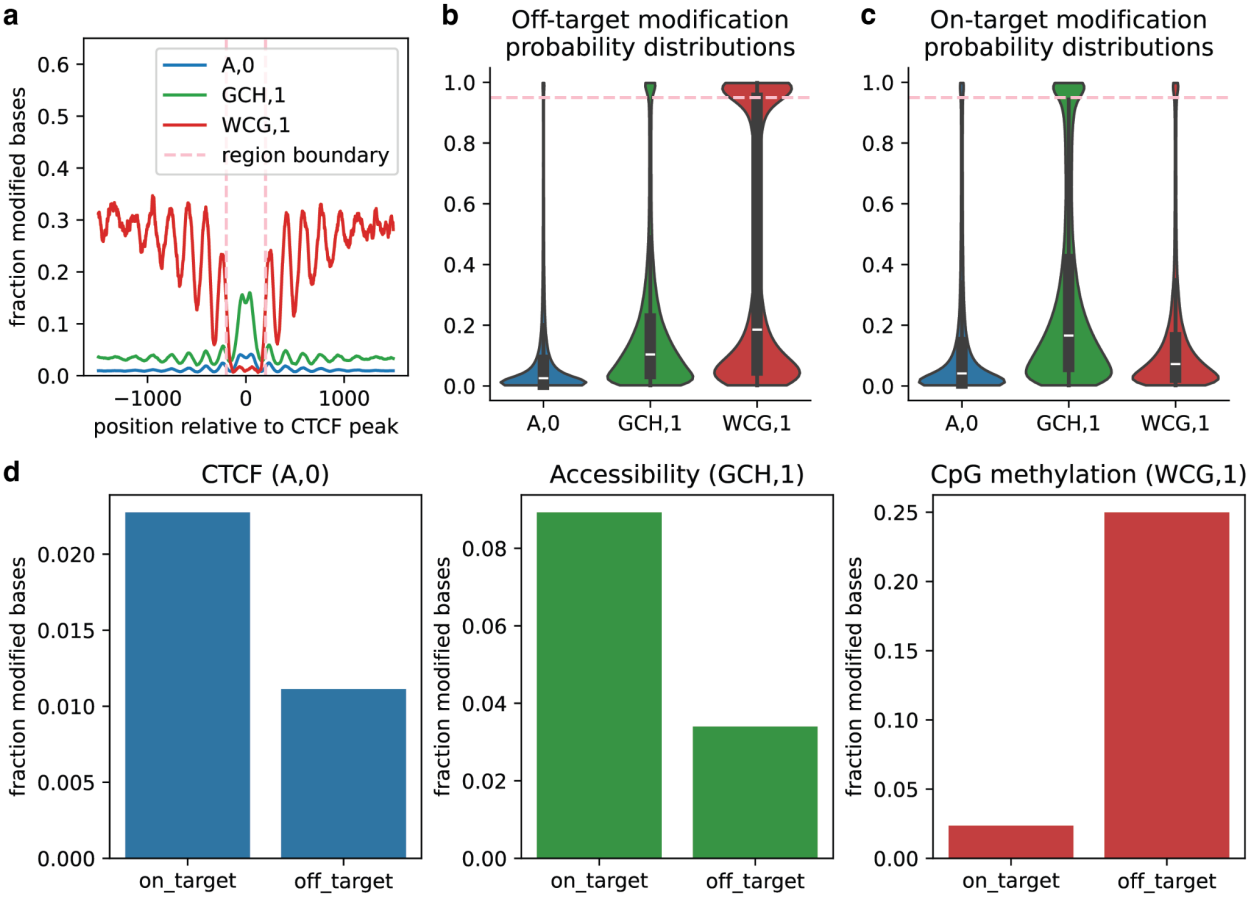


Supplementary Figure 2. Motif disambiguation for 5mC detection in the presence of exogenous GpC methylation. a-b*, CpG methylation profiles (red) for high-expression TSSs, at CG motifs with different preceding bases, in untreated* ***(a)*** *and M.CviPI-treated* ***(b)*** *conditions.* ***c-d****, GpC methylation profiles (green) for high-expression TSSs, at GC motifs with different succeeding bases, in untreated* ***(c)*** *and M.CviPI-treated* ***(d)*** *conditions. All profiles were generated using a modification probability threshold of 0.75 and a 25 bp rolling window. Subplot titles represent the motif codes used for each pileup operation. Left column represents the GCG motif which is shared between CG and GC methyltransferases. Middle columns represent*

motifs which are not expected to be shared between the methyltransferases. Right column represents the disambiguated motif selected for the analyses in this study.



Supplementary Figure 3. Modification probability distributions in LMNB1-targeting multimodal DiMeLo-seq experiment in GM12878 cells. a-b, *Distributions of modified base probabilities reported by the dorado basecaller for 6mA (blue; targeted to LMNB1), 5mC in GC motifs (green; accessible chromatin), and 5mC in CG motifs (red; endogenous CpG methylation). The pink dashed line represents the modification probability threshold of 0.95 selected for subsequent analysis of this data. Distributions are shown for on-target regions* ***(a)*** *and off-target regions* ***(b)*** *for each methylation signal. On- and off-target regions are defined the same way as in Figure 2: for 6mA, on-target regions are cLADs and off-target regions are ciLADs. For 5mC in GC motifs, on-target regions are active TSSs and off-target regions are the flanking regions of those TSSs. For 5mC in CG motifs, on-target regions are CpG islands overlapping inactive TSSs and off-target regions are CpG islands overlapping active TSSs.*



Supplementary Figure 4. Modification probability distributions in CTCF-targeting multimodal DiMeLo-seq experiment in GM12878 cells. a*, Enrichment profiles describing CTCF occupancy (blue), accessibility (green), and endogenous CpG methylation (red). Pink dashed lines represent the boundary defining on-target regions (+/- 200 bp from peak center) and off-target regions (1,300 bp flanking on-target regions on both sides).* ***b-c****, Distributions of modified base probabilities reported by the dorado basecaller for 6mA (blue; targeted to CTCF), 5mC in GC motifs (green; accessible chromatin), and 5mC in CG motifs (red; endogenous CpG methylation). The pink dashed line represents the modification probability threshold of 0.95 selected for subsequent analysis of this data. Distributions are shown in the on-target* ***(b)*** *and off-target* ***(c)*** *regions defined in* ***(a)****.* ***d****, Barplots of modified base fractions in regions defined in* ***(a)*** *for modifications reported with probability >0.95.*