Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells

Sebastian Pott[1]

1 University of Chicago, Department of Human Genetics, Chicago, IL, United States

Correspondence:

Sebastian Pott

University of Chicago

Department of Human Genetics

920 E. 58th Street, CLSC 317

Chicago, IL 60637

spott@uchicago.edu

Introductory paragraph

Gaining insights into the regulatory mechanisms that underlie the pervasive transcriptional variation observed between individual cells[1,2] necessitates the development of methods that measure chromatin organization in single cells. *N*ucleosome *O*ccupancy and *Me*thylome-sequencing (NOMe-seq) employs a GpC methyltransferase to detect accessible chromatin and has been used to map nucleosome positioning and DNA methylation genome-wide in bulk samples[3,4]. Here I provide proof-of-principle that NOMe-seq can be adapted to measure chromatin accessibility and endogenous DNA methylation in single cells (scNOMe-seq). scNOMe-seq recovered characteristic accessibility and DNA methylation patterns at DNase hypersensitive sites (DHSs) and enabled direct estimation of the fraction of accessible DHSs within an individual cell. In addition, scNOMe-seq provided high resolution of chromatin accessibility within individual loci which was exploited to detect footprints of CTCF binding and to estimate the average nucleosome phasing distances in single cells. This approach could be applied to characterize the chromatin organization of single cells in heterogeneous mixtures of cells, for example to samples of primary cancer cells.

Main

A number of methods that map chromatin organization in populations of cells previously have been adapted for single cells, including ATAC-seq[5,6] , DNase-seq[7], methylome sequencing[8,9], and ChIP-seq[10]. Interpretation of these data in single cells is complicated because the resulting signal is near binary and extremely sparse[5,7,11–13]. *N*ucleosome *O*ccupancy and *Me*thylome-sequencing (NOMe-seq)[3] employs the GpC methyltransferase (MTase) from *M.CviPI* to probe chromatin accessibility[3,4]. The GpC MTase methylates cytosines in GpC di-nucleotides in non-nucleosomal DNA *in vitro.* Combined with high-throughput bisulfite sequencing this approach has been used to characterize nucleosome positioning and endogenous methylation in human cell lines[3,14] and in

selected promoters of single yeast cells[15]. NOMe-seq data have several unique features that are advantageous in light of the challenges associated with single cell measurements **(Fig. 1a)**. First, NOMe-seq simultaneously measures chromatin accessibility (through GpC methylation) and endogenous CpG DNA methylation. Chromatin accessibility indicates whether a putative regulatory region might be utilized in a given cell[16], while endogenous DNA methylation in regulatory regions has been connected to a variety of regulatory processes often associated with repression[17]. The ability to combine complementary assays within single cells is essential for a comprehensive genomic characterization of individual cells since each cell represents a unique biological sample which is almost inevitably destroyed in the process of the measurement**.** Second, each sequenced read might contain several GpCs which independently report the accessibility status along the length of that read. NOMe-seq therefore captures additional information compared to purely count-based methods, such as ATAC-seq and DNase-seq, which increases the confidence associated with the measurements and allows detection of footprints of individual transcription factor (TF) binding events in single cells. Third, the DNA is recovered and sequenced independently of its methylation status, which is a pre-requisite to distinguish between true negatives (i.e. closed chromatin) and false negatives (i.e. loss of DNA) when assessing accessibility at specified locations in single cells. NOMe-seq can therefore measure the fraction of accessible regions among a set of covered, pre-defined genomic locations.

To adapt the NOMe-seq protocol[3,18] to single cells (scNOMe-seq), individual nuclei were isolated using fluorescence-activated cell sorting (FACS) and sorted into wells of a 96-well plate following the incubation with the GpC MTase (**Fig. 1b and Supplemental Fig. 1**). DNA from isolated nuclei was subjected to bisulfite conversion and sequencing libraries were prepared using a commercial kit for amplification of low input bisulfite converted DNA (**Methods)**. In this proof-of-concept study, I used the well-characterized lymphoblast cell lines GM12878 and K562 to assess the feasibility and performance of scNOMe-seq. The scNOMe-seq datasets in this study represent 19 individual GM12878 cells and 12 individual K562 cells. The set of GM12878 cells included seven control

cells that were not incubated with GpC MTase (**Supplemental Fig. 2**). Each GpC MTase treated library was sequenced to at least 16 M 100 bp reads, of which 37%- 64% aligned to the human genome using the bisulfite aligner Bismark[19] (**Supplemental Table 1**). Genome-wide, the number of cytosines covered in GpC and CpG contexts averaged 6,679,864 (2.9%) of all GpCs and 1,291,180 (3.6%) of all CpGs per cell (**Supplemental Fig. 3 and Supplemental Table 1**).

To test whether the GpC methylation observed in GpC MTase treated samples (**Supplemental Fig. 4)** captured chromatin accessibility at specific genomic features, I focused on DNase hypersensitive sites (DHSs) previously identified in GM12878 and K562 cell lines[16]. DHSs were associated with strong enrichment of GpC methylation. This was observed both in data from pooled and individual GM12878 cells (**Fig. 1 c, d and Supplemental Fig. 5**), and K562 cells (**Supplemental Fig. 6, 7**). Conversely, endogenous DNA methylation decreased around the center of the DHSs in agreement with previous reports[20,21] (**Figure 1c and Supplemental Fig. 6**). These data show that scNOMe-seq detects chromatin accessibility at DHSs. In principle, the frequent occurrence of GpC di-nucleotides renders the majority of DHSs detectable by NOMe-seq (**Supplemental Fig. 8, 9**). On average in 10.6% (20388/191566) and 17.3% (33182/191598) of the DHSs with one or more GpC at least one GpC was covered by a sequencing read, and in 5.2% (9083/174896) and 9.5% (16608/174828) of the DHSs with four or more GpC at least four GpCs were covered in individual GM12878 cells and K562 cells, respectively (**Fig. 1 e**). Chromatin accessibility signal can vary along the length of a given DHSs due to binding of transcription factors[22] and the specific position of a GpC within a DHSs will thus affect its chance of being methylated. To account for this variability and to obtain more robust estimates of GpC methylation only DHSs with at least 4 covered GpC were used for the subsequent analyses and referred to as 'covered DHSs'. Average GpC methylation of covered DHSs in single cells was strongly associated with the observed DNaseI accessibility at these sites in bulk populations (**Fig. 1f and Supplemental Fig. 10)**. The opposite trend was observed for endogenous CpG methylation which was lowest for DHSs with the highest DNaseI accessibility (**Fig. 1g and Supplemental Fig. 10).** At the level of individual sites the distribution of GpC

methylation suggested that around 50% of the covered DHS showed no or low accessibility (i.e. less than 25% GpC methylation) in individual cells (**Supplemental Fig. 11)**. To estimate the proportion of covered DHSs that were concurrently accessible in a single cells I applied a fixed threshold of 40% GpC methylation above which sites were considered accessible (**Methods**). At this GpC methylation threshold 32%-44% and 26%-37% of all covered DHSs were determined accessible in single GM12878 and K562 cells, respectively. As expected these result depended to some degree on the GpC methylation and the numbers of GpCs required to include a DHSs in the analysis (**Supplemental Fig. 12**). However, even under the most lenient conditions less than 50% of DHSs were accessible in most individual cells. Grouping the DHSs based on DNaseI accessibility measured in bulk samples revealed that the degree of DNaseI accessibility closely related to the frequency DHSs accessibility in single cells (**Fig. 1h)**. This analysis leveraged the NOMe-seq-specific property that the DNA sequence is recovered independently of its accessibility status. It provided direct evidence for the notion that the degree of DNaseI accessibility observed in DNase-seq of bulk samples reflects the frequency with which a particular region is accessible in individual cells.

A potentially powerful application for single cell genomic approaches is the label-free classification of single cells from heterogeneous mixtures of cells solely based on the measured feature[5,11,23]. Of note, using a union set of DHSs from both cell types was sufficient to classify individual GM12878 and K562 cells into their respective cell types based on GpC methylation (**Fig. 1 i)**. While this assessment might have been influenced in part by the separate processing of both cell types, both cell types showed preferential enrichment of GpC methylation at their respective DHSs compared to DHSs identified in the other cell type **(Supplemental Fig. 13**). Thus, this approach should be extendable to scNOMe-seq data from samples containing mixtures of cell types and endogenous CpG methylation could be included in such analyses to provide additional information[9].

To examine in detail whether scNOMe-seq captures features of chromatin accessibility that are specifically associated with transcription factor binding I analyzed scNOMe-seq data at

transcription factor binding sites (TFBS). The average GpC methylation around CTCF ChIP-seq peaks[16] in single cells recapitulated the accessibility previously observed in NOMe-seq bulk samples[3]: Accessibility increased strongly towards the CTCF binding sites while the location of the CTCF motif at the center of the region showed low accessibility suggesting that CTCF binding protected from GpC MTase activity and thus creating a footprint of CTCF binding, both when averaged across data from all single cells (**Fig. 2a and Supplemental Fig. 14**) and in individual cells (**Fig. 2b and Supplemental Fig. 15**). In contrast, endogenous CpG methylation was generally depleted around the center of CTCF binding sites **(Fig. 2a and Supplemental Fig. 14)**. Similar accessibility profiles, albeit less pronounced compared to CTCF were observed for additional transcription factors (**Supplemental Fig. 16**). These analyses provided evidence, that in aggregate, scNOMe-seq detected CTCF DNA binding events from single cells. In addition to aggregated binding sites, scNOMe-seq data should also provide information about transcription factor binding at individual loci. To test whether scNOME-seq data detected CTCF footprints at individual motifs, GpC methylation at motifs within CTCF ChIP-seq peaks that contained at least one GpC was compared to the GpC methylation level in the regions flanking each motif **(Fig. 2c)**. On average, two-thirds of CTCF motif instances within these accessible regions showed no GpC methylation, or lower GpC methylation than the flanking regions suggesting that scNOMe-seq detected footprints caused by binding of CTCF **(Fig. 2d and f)**. Of note, motifs associated with a footprint had significantly higher scores than motifs without a footprint suggesting that the motif score is a strong determinant of CTCF binding within these accessible regions **(Fig. 2e, g and Supplemental Fig. 17)**. The CTCF footprint could be observed at individual loci and comparing GpC methylation from multiple cell and suggests that scNOMe-seq also detected cell-to-cell variation in the footprint **(Figure 2h and Supplemental Fig. 18).** Footprint measurements such as this should be generally feasible for TFs whose motifs contain at least one GpC di-nucleotide and could be used to infer the activity of a wide range of transcription factors in single cells.

The pattern of GpC methylation adjacent to CTCF sites suggested that scNOMe-seq also detected the well-positioned nucleosomes flanking these regions (**Fig. 2a**)[3]. This observation was confirmed by the oscillatory distribution of the average GpC and CpG methylation around locations of well-positioned nucleosomes identified from MNase-seq data[16] (**Fig. 3a and Supplemental Fig. 14**). While nucleosome core particles are invariably associated with DNA fragments of 147 bp, nucleosomes are separated by linker DNA of varying lengths, resulting in different packaging densities between cell types[24] and between genomic regions within a cell[24,25]. To determine whether scNOMe-seq data can be used to measure the average linker length, average distances between nucleosome midpoints in single cells (phasing distances) were estimated by correlating the methylation status between pairs of cytosines in GpC di-nucleotides at offset distances from 3 bp to 400 bp **(Fig. 3c, d and Supplemental Fig. 19, 20).** The estimated phases fell between 187 bp and 196 bp (mean=196.7 bp) in GM12878 cells, and between 188 bp and 200 bp (mean =194.2 bp) in K562 cells (**Fig. 3e**). These estimates are in general agreement with phase estimates derived from MNase-seq data in human cells[24]. In addition, estimated phasing distances varied within individual cells depending on the chromatin context, similar to observation from bulk MNase-seq data[24] (**Fig. 3f**). These proof-of-principle experiments have been performed using commercial kits for bisulfite conversion and library amplification, additional optimization or alternative amplification approaches[8] are likely to increase the yield substantially. Ultimately, it should be possible to integrate the GpC MTase treatment into microfluidic workflows and combine this method with scRNA-seq, similar to recently published methods that combine scRNA-seq and methylome-sequencing[26]. scNOMe-seq will be particularly useful for studies that aim to simultaneously measure chromatin accessibility and DNA methylation, or that aim to measure activity of transcription factors in individual cells based on footprints in single cells. This approach could be applied to characterize the chromatin organization of single cells in heterogeneous mixtures of cells, for example to samples of primary cancer cells. scNOMe-seq could be applied to characterize the

chromatin organization of single cells in heterogeneous mixtures of cells, for example in samples of primary cancer cells.

Methods

Cell culture, nuclei isolation, and GpC methylase treatment

GM12878 and K562 cells were obtained from Coriell and ATCC, respectively. GM12878 were grown in RPMI medium 1640 (Gibco), supplemented with 2mM L-Glutamine (Gibco), and Penicilin and Streptavidin (Pen Strep, Gibco), and 15% fetal bovine serum (FBS, Gibco). K562 were grown in RPMI medium 1640 of the same composition but with 10% FBS. Cells were grown at 37 C and in 5% CO_2. NOMe-Seq procedure was performed based on protocols for CpG methyltransferase SSsi described in[18]and the GpC methyltransferase from *M.CviPl*[3], with some modification. Between 2x10^6 and 5x10^6 cells were harvested by centrifuging the cell suspension for 5 min at 500x g. Cells were washed once with 1x PBS, re-suspended in 1 ml lysis buffer (10mM Tris-HCl pH 7.4, 10mM NaCl, 3mM $MgCl_2$) and incubated for 10 min on ice. IGEPAL CA-630 (Sigma) was added to a final concentration of 0.025% and the cell suspension was transferred to a 2 ml Dounce homogenizer. Nuclei were released by 15 strokes with the pestle. Success of lysis was confirmed by inspection under a light microscope. Nuclei were collected by centrifuging the cell suspension for 5 min at 800x g at 4C and washed twice with cold lysis buffer without detergent. One million nuclei were resupended in reaction buffer to yield a suspension with a final concentration of 1x GpC MTase buffer (NEB), 0.32 mM S-Adenosylmethionine (SAM) (NEB), and 50 ul of GpC methyltransferase (4U/ul)) from *M.CviPl* (NEB). The suspension was carefully mixed before incubating for 8 min at 37 C after which another 25 ul of enzyme and 0.7 ul of 32 mM SAM were added for an additional 8 min incubation at 37C. To avoid disruption of nuclei incubation was stopped by adding 750 ul of 1x PBS and collecting the nuclei at 800 xg. Supernatant was removed

and nuclei were re-suspended in 500ul 1x PBS containing Hoechst 33342 DNA dye (NucBlue Live reagent, Hoechst). Nuclei were kept on ice until sorting.

Nuclei isolation using Fluorescence activated cell sorting (FACS), lysis, and DNA bisulfite conversion

Nuclei were sorted at the Flow Cytometry core at the University of Chicago on a BD FACSAria or BD FACSAria Fusio equipped with a 96-well-plate holder. To obtain individual and intact nuclei gates were set on forward and side scatter to exclude aggregates and debris. DAPI/PacBlue channel or Violet 450/500 channel were usedto excited the Hoechst 33342 DNA dye and to gate on cells with DNA content corresponding to cells in G1 phase of the cell cycle in order to maintain similar DNA content per cell and to remove potential heterogeneity attributable to cell cycle. Cells were sorted into individual wells pre-filled with 19 ul of 1x M-Digestion buffer (EZ DNA Methylation Direct Kit, Zymo Research) containing 1 mg/ml Proteinase K. Following collection, the plates were briefly spun to collect droplets that might formed during handling. Nuclei were lysed by incubating the samples at 50 C for 20 min in a PCR cycler. DNA was subjected to bisulfite conversion by adding 130 ul of freshly prepared CT Conversion reagent (EZ DNA Methylation Direct Kit, Zymo) to the lysed nuclei. Conversion was performed by denaturing the DNA at 98 C for 8 min followed by 3.5 hrs incubation at 65 C. DNA isolation was performed using the EZ DNA Methylation Direct Kit (Zymo Research) following the manufacturer's instruction with the modification that the DNA was eluted in only 8 ul of elution buffer.

Library preparation and sequencing

Libraries were prepared using the Pico Methyl-seq Library prep Kit (Zymo Research) following the manufacturer's instruction for low input samples. Specifically, the random primers were diluted 1:2 before the initial pre-amplification step and the first amplification was extended to a total of 10 amplification cycles. Libraries were amplified with barcoded primers allowing for multiplexing. The sequences can be found in **Supplemental Table 2**, primers were ordered from IDT. The purification of amplified libraries was performed using Agencourt AMPureXP beads (Beckmann

Coulter), using a 1:1 ratio of beads and libraries. Concentration and size distribution of the final libraries was assessed on an Bioanalyzer (Agilent). Libraries with average fragment size above 150 bp were pooled and sequenced. Libraries were sequenced on Illumina HiSeq 2500 in rapid mode (K562 cells) and HiSeq4000 (GM12878 cells).

Read processing and alignment

Sequences were obtained using 100 bp paired-end mode. For processing and alignment each read from a read pair was treated independently as this slightly improved the mapping efficiency. Before alignment, read sequences in fastq format were assessed for quality using fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were trimmed to remove low quality bases and 6 bp were clipped from the 5 prime end of each read to avoid mismatches introduced by amplification. In the case of GM12878 cells 6 bp were clipped from either end of the read. Only reads that remained longer than 20 bp were kept for further analyses. These processing steps were performed using trim_galore version 0.4.0 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with the following settings: *trim_galore --quality 30 --phred33 --illumina --stringency 1 -e 0.1 --clip_R1 6 --gzip --length 20 --output_dir outdir Sample.fastq.gz*. The trimmed fastq files were aligned using the bisulfite aligner bismarck version 0.15.0[27] which calls bowtie2[28] internally. Reads were aligned to the human genome (genome assembly hg38). Reads were aligned in single read mode using default settings. The amplification protocol used to generate the scNOMe-seq libraries yielded non-directional libraries and alignment was performed with the option —non_directional (*bismark --fastq --prefix SamplePrefix --output_dir output_dir --non_directional --phred33-quals --score_min L,0,-0.2 --bowtie2 genome_file trimmed.fastq.gz*). Some libraries contained small amounts of DNA from *C. elegans* as spike-ins, however these were not used during the analysis. Duplicates were removed using samtools version 0.1.19[29] on sorted output files from bismark (*samtools rmdup SamplePrefix.sorted.bam SampleAligned_rmdup.bam*).

Extraction of GpC and CpG methylation status

Coverage and methylation status of all cytosines was extracted using bismark_methylation_extractor[27] (*bismark_methylation_extractor -s --ignore 6 --output outdir --cytosine_report --CX --genome_folder path_to_genome_data SampleAligned_rmdup.bam*). The resulting coverage files were used to extract the methylation status of cytosines specifically in GpC and CpG di-nucleotides using the coverage2cytosine script which is part of Bismark[27]. The resulting coverage files contained cytosines in GCG context which are ambiguous given that they represent a cytosine both in GpC and CpG di-nucleotides. Coordinates of these ambiguous positions were identified using oligoMatch[30] and these positions were removed from the coverage files. The number of unconverted cytosines (estimated based on apparent methylation rates in non-GpC and non-CpG context) was low in all libraries (<1%). However, it was noted that unconverted cytosines were not randomly distributed but associated with entirely unconverted reads. Regions covered by a read with more than 3 unconverted cytosines in non-CpG and non-GpC context were removed from further analysis as well. The genotype was not taken into account as its effect on calling the methylation status incorrectly was deemed negligible for the analyses performed here.

Analysis of GpC and CpG methylation at genomic features in single cells

ScNOMe-seq data were compared to a number of genomic features in GM12878 and K562 cells collected by ENCODE[24] which were downloaded through the UCSC data repository[31]. These datasets are listed in **Supplemental Table 3**. While the scNOMe-seq data were aligned against human genome assembly hg38, some of the datasets were only available on genome assembly hg19 and the coordinates of these datasets were lifted from hg19 to hg38 using liftOver[30] (default re-mapping ratio 1). Nucleosome positions based on MNase-seq data in GM12878 were determined with DANPOS version 2.2.2[32] using default settings. Resulting intervals were lifted to hg38. After removing summit locations with occupancy values above 300, the top 5% (713361) of nucleosome positions based on their summit occupancy value were used.

GpC and CpG methylation density across intervals encompassing DNase hypersensitivity sites (DHSs), transcription factor binding sites (TFBS), and well positioned nucleosomes was calculated

across the 2 kb regions centered on the middle of these regions using the scoreMatrixBin function in the genomation package[33] in R[34]. Data were aggregated in 5 bp bins for each region and across all regions covered in a single cell. The average methylation level in pre-defined intervals (DHSs, TFBS) was determined by computing the average GpC or CpG methylation for each interval together with the number of GpC/CpGs covered in this interval using the map function from the bedtools[35] suite. If not specifically mentioned otherwise DHSs were considered 'covered' and used in analyses when at least 4 GpCs occurring within the predefined interval were covered by sequencing data in an individual cell. Covered DHSs that Because the frequency of CpG di-nucleotides is significantly lower, only 2 CpGs were required in order for a DHSs to be considered covered for analyses focused on endogenous DNA methylation. To estimate the number of cytosines within a given DHSs that could be covered only cytosines on the forward strand were counted. While each GpC dinucleotide can be measured on both strands and would therefore yield a count of two cytosines the data are sparse and each location will get at most a single read. This approach should therefore give a more conservative estimate of the possible GpC coverage. For analyses that used the scores of the peak regions, the peak scores reported the datasets from bulk samples were used[16].

For analyses that were centered on transcription factor binding motifs the PWMs were obtained from the JASPAR database (2014)[36] for the TFs CTCF (MA0139), EBF1 (MA0154), and PU.1(MA0080). Genome-wide scanning for locations of sequence matches to the PWMs was performed using matchPWM in the Biotstring package[37] in R with a threshold of 75% based on the human genome assembly hg38.

All plots were prepared using ggplot2[38], with the exception of heatmaps displaying the average methylation density around genomic features in individual cells which were prepared using heatmap.2 in gplots[39].

Comparison of chromatin accessibility between cells

Similarity in accessible chromatin between cells was calculated based on Jaccard similarity. Jaccard similarity index (eq. 1) was calculated between pairs of samples by first obtaining the intersection of DHSs covered in both samples of a pair with more than 4 GpCs. Each features was annotated as open or closed, depending on the methylation status (>= 40% methylation) and only pairs in which at least one of the members was open.

$$jac(A,B) = \frac{(A \cap B)}{(A \cup B)} \quad (1)$$

The similarity between samples from GM12878 and K562 cells was calculated based on the union set of DHSs from both cell lines. The similarity indexes of all pairwise comparisons were used to compute the distances between each cell. The resulting clustered data were displayed as a heat map.

CTCF footprints in single cells

CTCF footprints were measured by comparing the GpC methylation level in each motif to the methylation level in the 50bp flanking regions immediately upstream and downstream of the motif. Overlapping motifs were merged into a single interval before determining the coordinates for flanking regions. To ensure sufficient GpC coverage for each interval the resulting three interval was required to have at least one covered GpC and 4 GpCs covered in total among the three intervals. This analysis only included regions that were accessible based on the methylation status of the flanking regions (at least 50%). A CTCF footprint 'score' was determined by simply subtracting the average GpC methylation of the flanking regions from the GpC methylation of the motif.

scNOMe-seq data were displayed in the UCSC genome browser[30] by converting the GpC methylation coverage file into a bed file and using the methylation value as score. To facilitated the visualization of the data in the context of previous Encode data the methylation files were lifted to hg19. The tracks shown together with scNOMe-seq data are Open Chromatin by DNaseI HS from ENCODE/OpenChrom (Duke University) for DNaseI hypersensitivity, Nucleosome Signal from

ENCODE/Stanford/BYU, and CTCF ChIP-seq signal from Broad Histone Modification by ChIP-seq from ENCODE/Broad Institute. All data are from GM12878 cells.

Estimation of nucleosome phasing

Nucleosome phasing estimates were obtained by first calculating the correlation coefficients for the methylation status of pairs of GpCs ad different offset distances. These values were computed using a custom python script. Essentially, pairs of sequenced cytosines in GpC di-nucleotides were collected for each offset distance from 3bp to 400bp cytosine. At each offset distance the correlation of the methylation status was calculated across all pairs. Correlation coefficients were plotted against the offset distances revealing periodic changes in the correlation coefficient. The smoothened data were used to estimate the phasing distances by obtaining the offset distance corresponding to the local maximum found between 100 bp and 300 bp. To determine phase lengths of nucleosomes in different chromatin contexts the GpC coverage files were filtered for positions falling into categories defined by chromHMM[16,40] before obtaining the correlation coefficients.

Data access

Raw data and methylation coverage files are available at GEO (https://www.ncbi.nlm.nih.gov/geo/) under the accession number . Reviewers might use this link: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=glotcwqqjbqlvef&acc=GSE83882

Competing financial interest

The author declares no competing financial interests

Acknowledgements

I like to thank Yoav Gilad for support, and Greg Crawford and colleagues in the Department of Human Genetics for helpful suggestions and comments on the manuscript. Cell sorting was performed by M. Olson and D. Leclerc at the Flow Cytometry core of the University of Chicago. I am grateful to Jason Lieb for input and support at the beginning of this project.

Bibliography

1.Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498,** 236–240 (2013).

2.Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–375 (2014).

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

4.Kilgore, J. A., Hoose, S. A., Gustafson, T. L., Porter, W. & Kladde, M. P. Single-molecule and population probing of chromatin structure using DNA methyltransferases. *Methods (San Diego, Calif.)* **41,** 320–332 (2007).

5.Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348,** 910–914 (2015).

6.Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 1–15 (2015). doi:10.1038/nature14590

7.Jin, W. *et al.* Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* 1–17 (2015). doi:10.1038/nature15740

8.Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods* **11,** 817–820 (2014).

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

10.Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology* **33,** 1–11 (2015).

11.Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **109,** 21.29.1 (2015).

12.Maurano, M. T. & Stamatoyannopoulos, J. A. Taking Stock of Regulatory Variation. *Cell Systems* **1,** 18–21 (2015).

13.Bernstein, B. E. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* 1–5 (2013). doi:10.1038/nature12433

14.Taberlay, P. C., Statham, A. L., Kelly, T. K., Clark, S. J. & Jones, P. A. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Research* **24,** 1421–1432 (2014).

15.Small, E. C., Xi, L., Wang, J.-P., Widom, J. & Licht, J. D. Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America* **111,** NaN–NaN (2014).

16.ENCODE Project Consortium & The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

17.Schübeler, D. Function and information content of DNA methylation. *Nature* **517,** 321–326 (2015).

18.Miranda, T. B., Kelly, T. K., Bouazoune, K. & Jones, P. A. Methylation-Sensitive Single-Molecule Analysis of Chromatin Structure. 1–16 (2001). doi:10.1002/0471142727.mb2117s89

19.Krueger, F., Kreck, B., Franke, A. & Andrews, S. R. DNA methylome analysis using short bisulfite sequencing data. *Nature Methods* **9,** 145–151 (2012).

20.Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 1–7 (2011). doi:10.1038/nature10716

21.Ziller, M. J. *et al.* Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature* **518,** 355–359 (2015).

22.Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83–90 (2012).

23.Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343,** 776–779 (2014).

24.Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474,** 516–520 (2011).

25.Schones, D. E. *et al.* Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* **132,** 887–898 (2008).

26.Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods* 1–6 (2016). doi:10.1038/nmeth.3728

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

28.Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).

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

30.Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Research* **12,** 996–1006 (2002).

31.Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research* **42,** NaN–NaN (2014).

32.Chen, K. *et al.* DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Research* **23,** 341–351 (2013).

33.Akalin, A., Franke, V., Vlahoviček, K., Mason, C. E. & Schübeler, D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* **31,** 1127–1129 (2015).

34.R Core Team. R: A language and environment for statistical computing. (2015). at <https://www.R-project.org/>

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

36.Tan, G. JASPAR2014: Data package for JASPAR. at <http://jaspar.genereg.net/>

37.Pages, H., Aboyoun, P., Gentleman, R. C. & DebRoy, S. Biostrings: String objects representing biological sequences, and matching algorithms. (2016).

38.Wickham, H. ggplot2. 213 (2009). doi:10.1007/978-0-387-98141-3

39.Warnes, G. R., Bolker, B., Bonebakker, L. & Gentleman, R. gplots: Various R programming tools for plotting data. (2016). at <https://CRAN.R-project.org/package=gplots>

40.Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).

Figure Legends

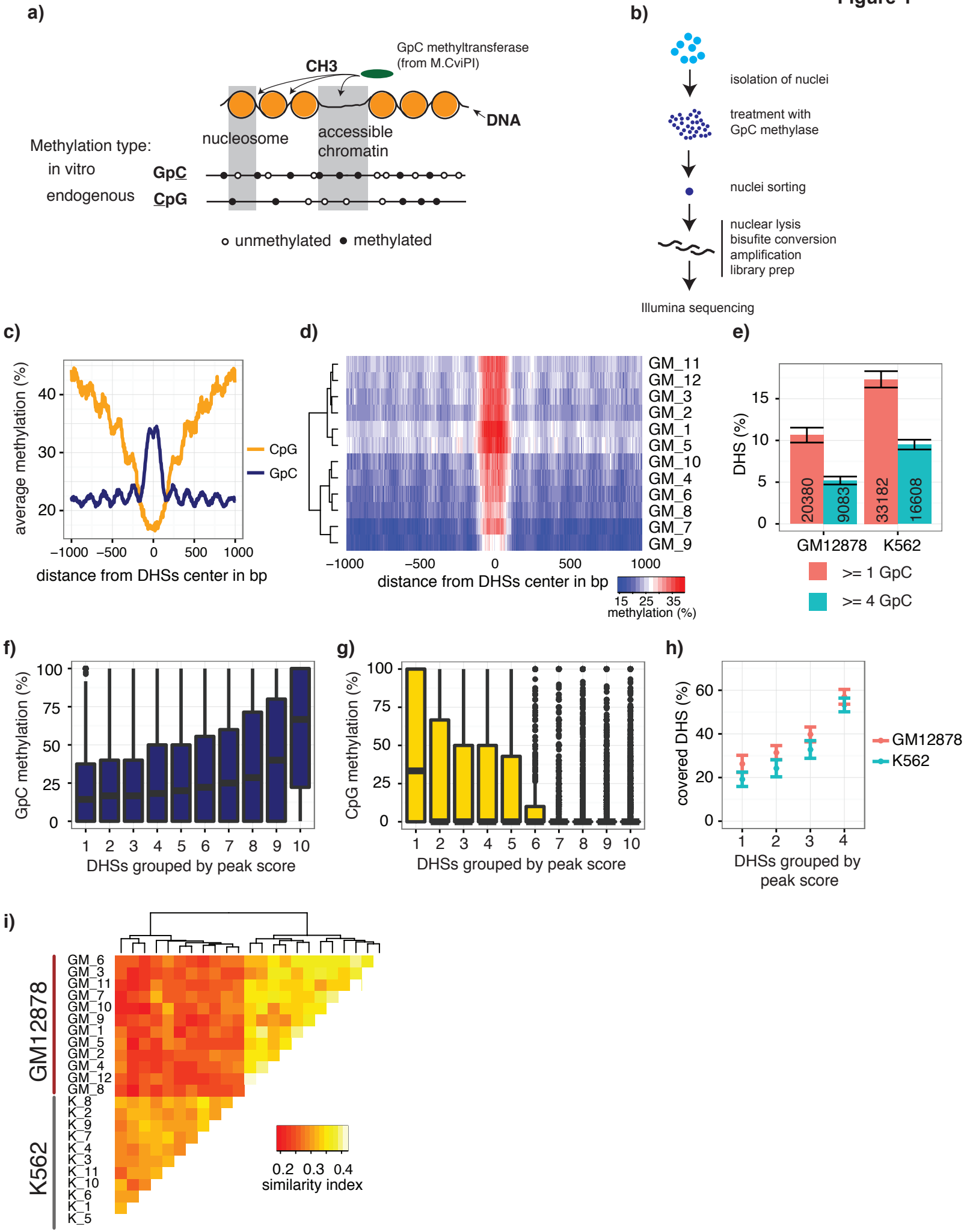
Figure 1. scNOMe-seq detected DNase hypersensitive sites in single cells. a) Schematic of GpC methyltransferase-based mapping of chromatin accessibility and simultaneous detection of endogenous DNA methylation. b) Schematic of scNOMe-seq procedure introduced in this study. c) Average GpC methylation level (blue) and CpG methylation level (orange) at DNase Hypersensitive sites (DHSs) in GM12878 cells. Regions are centered on the middle of DNase-seq peak locations. Shown is the average methylation across a 2 kb window of 12 GM12878 cells. d) Heatmap displaying the average GpC methylation level across the same regions as in c). Each row corresponds to an individual GM12878 cell. Cells were grouped by similarity. e) Proportion of DHSs covered by scNOMe-seq data in each cell. The proportion displayed corresponds to the fraction of DHSs covered by at least 1 or 4 GpCs in a given cell. Only DHSs with at least 1 or 4 GpCs, respectively, within their primary sequence were taken in consideration. Error bars represent the standard deviation. f) Average GpC methylation and g) endogenous CpG methylation at DHSs split into 10 groups based on associated DNase-seq peak scores from lowest to highest scores. Average methylation per DHSs was based on the data from 12 GM12878 cells. h) Fraction of accessible sites in individual GM12878 and K562 cells, respectively. DHSs were grouped by DNase-seq peak scores. DHSs was considered accessible if the average methylation for that locus was above 40%. Only DHSs with at least 4 covered GpCs were included. i) Heatmap shows similarity scores (pair-wise jaccard distances) between all GM12878 and K562 cells. Comparison was based on the union set of DHSs from GM12878 and K562 cells. Cells were grouped based on unsupervised hierarchical clustering.

Figure 2. **scNOMe-seq detected characteristic accessibility patterns at CTCF transcription factor binding sites and measured CTCF footprints at individual loci** a) Average GpC methylation level (blue) and CpG methylation level (orange) at CTCF binding sites in GM12878 cells. Regions are centered on motif locations. Shown is the average methylation across a 2 kb window of the pool of 12 GM12878 cells. b) Heatmap displaying the average GpC methylation across CTCF binding sites. Each row corresponds to an individual GM12878 cell and rows are grouped by similarity. c) Schematic outline the measurement of CTCF footprints in accessible regions. M denotes CTCF binding motifs within CTCF ChIP-seq regions and U and D indicate 50 bp upstream and downstream flanking regions. footprint score was determined by subtracting the average GpC methylation in the flanking regions from the GpC methylation at the motif. d) Heatmap displays GpC methylation in accessible regions found in a representative GM12878 cell (GM_1). Each row represents a single CTCF motif instance within a CTCF ChIP-seq region. Average methylation values for the motif and the 50 bp upstream and downstream regions are shown separately. Regions are sorted based on the footprint score. Displayed are only regions that had sufficient GpC coverage and that were considered accessible based on the methylation status of the flanking regions. e) Heatmap reporting the CTCF motif scores for the motif regions in d). Regions are sorted in the same order as in d). f) Average number of accessible regions at CTCF motifs and the average number of those with a detectable footprint per individual GM12878 cell. Error bars reflect standard deviation. g) Average CTCF motif scores in regions with and without CTCF footprint for all 12 GM12878 cells. Each line connects the two data points from an individual cell h) Combined display of scNOMe-seq data from this study and DNase hypersensitivity data, nucleosome occupancy, and CTCF ChIP-seq data from ENCODE. Upper panel shows a ~10 kb region containing a CTCF binding sites. DNaseI hypersensitivity data and nucleosome density show characteristic distribution around CTCF binding sites in GM12878 cells. Lower panel shows the GpC methylation data of 5 individual cells that had sequencing coverage in this region, 4 of the cells provide GpC data covering the CTCF motif located in the region. scNOMe-seq data tracks

show methylation status of individual GpCs. Each row corresponds to data from a single cell. These data indicate that binding of CTCF is detected in all 4 cells. Data are displayed as tracks in the UCSC genome browser (*http://genome.ucsc.edu)*.

Figure 3. Nucleosome phasing in single cells. a) Average GpC methylation level and b) CpG methylation level at well-positioned nucleosomes in GM12878 cells. Regions are centered on midpoints of top 5% of positioned nucleosomes. Shown is the average methylation across a 2 kb window of the pool of 12 GM12878 cells. c), d) Correlation coefficients for the comparison in methylation status between GpCs separated by different offset distances for GM12878 (c) and K562 (d) cells. Each line represents a single cell. Data are smoothened for better visualization. e) Distribution of estimated phase lengths for GM12878 and K562 cells. f) Nucleosome phasing in GM12878 in genomic regions associated with different chromatin states defined by chromHMM (ENCODE). Boxplot represents the distribution of estimated phase lengths from all 12 GM12878 cells and overlaid points indicate values of each individual cells.

Figure 1


a)
GpC methyltransferase
(from M.CviPI)
CH3
DNA
nucleosome
accessible
chromatin
Methylation type:
in vitro
endogenous
GpC
CpG
unmethylated
methylated
b)
isolation of nuclei
treatment with
GpC methylase
nuclei sorting
nuclear lysis
bisufite conversion
amplification
library prep
Illumina sequencing
c)
average methylation (%)
CpG
GpC
distance from DHSs center in bp
d)
GM_11
GM_12
GM_3
GM_2
GM_1
GM_5
GM_10
GM_4
GM_6
GM_8
GM_7
GM_9
distance from DHSs center in bp
methylation (%)
e)
DHS (%)
20380
9083
33182
16608
GM12878
K562
>= 1 GpC
>= 4 GpC
f)
GpC methylation (%)
DHSs grouped by peak score
g)
CpG methylation (%)
DHSs grouped by peak score
h)
covered DHS (%)
GM12878
K562
DHSs grouped by
peak score
i)
GM12878
K562
GM_6
GM_3
GM_11
GM_7
GM_10
GM_9
GM_1
GM_5
GM_2
GM_4
GM_12
GM_8
K_8
K_2
K_9
K_7
K_4
K_3
K_11
K_10
K_6
K_1
K_5
similarity index

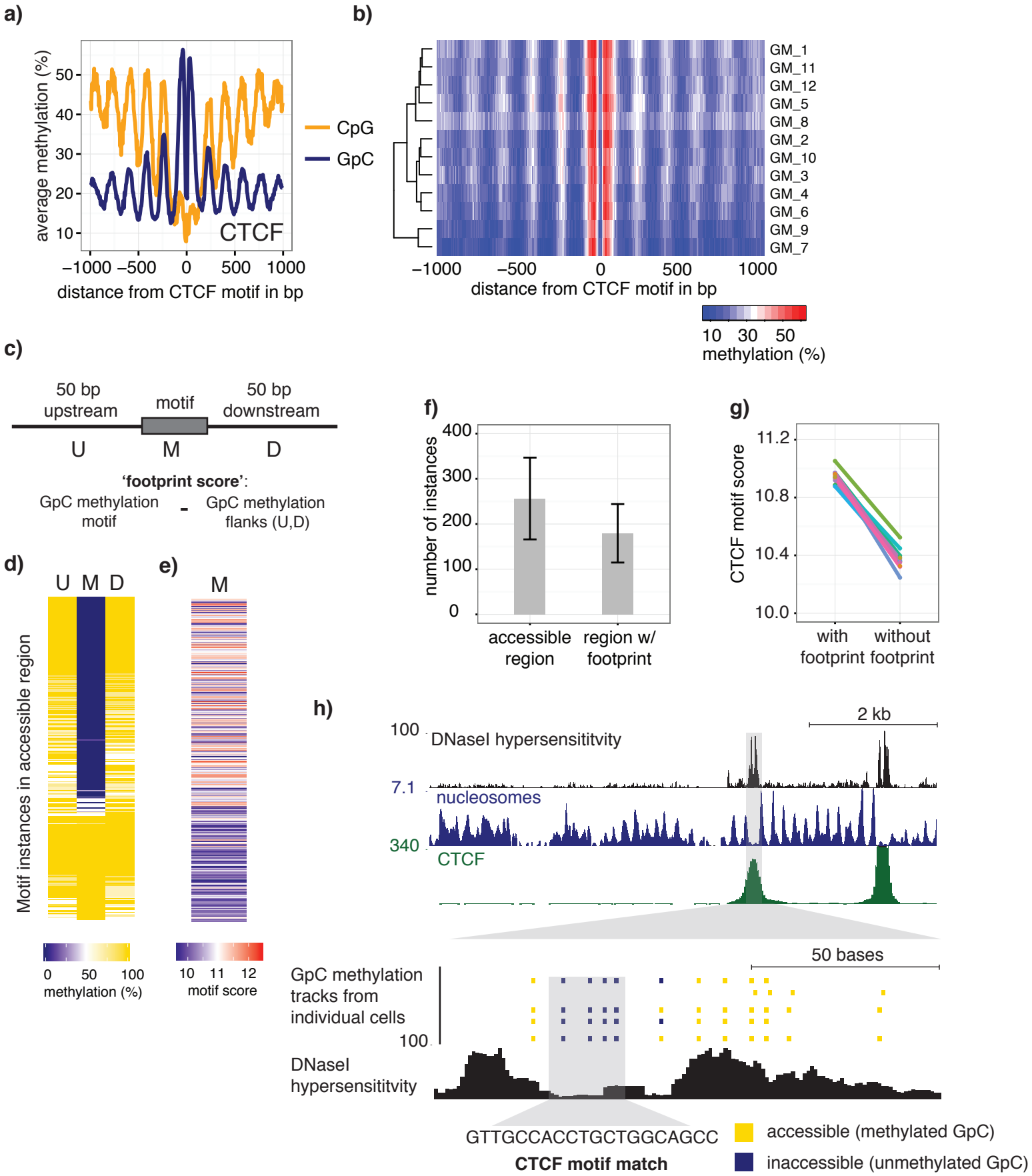
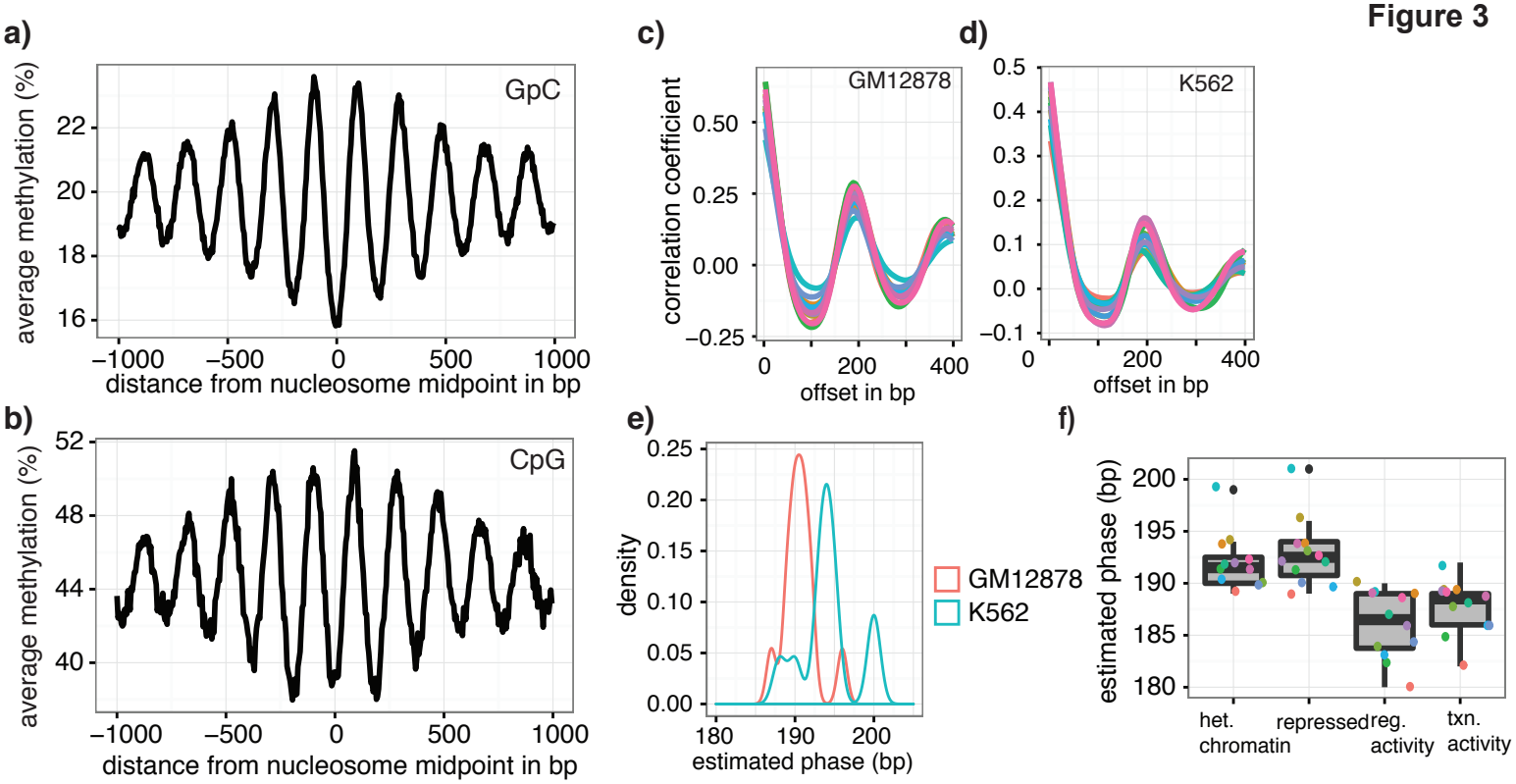

a)
average methylation (%)
50
40
30
20
10
CTCF
CpG
GpC
−1000 −500 0 500 1000
distance from CTCF motif in bp
b)
GM_1
GM_11
GM_12
GM_5
GM_8
GM_2
GM_10
GM_3
GM_4
GM_6
GM_9
GM_7
−1000 −500 0 500 1000
distance from CTCF motif in bp
10 30 50
methylation (%)
c)
50 bp upstream
motif
50 bp downstream
U
M
D
'footprint score':
GpC methylation motif
-
GpC methylation flanks (U,D)
d)
U M D
Motif instances in accessible region
0 50 100
methylation (%)
e)
M
10 11 12
motif score
f)
number of instances
400
300
200
100
0
accessible region
region w/ footprint
g)
CTCF motif score
11.2
10.8
10.4
10.0
with footprint
without footprint
h)
100
DNaseI hypersensititvity
2 kb
7.1
nucleosomes
340
CTCF
50 bases
GpC methylation tracks from individual cells
100
DNaseI hypersensititvity
GTTGCCACCTGCTGGCAGCC
CTCF motif match
accessible (methylated GpC)
inaccessible (unmethylated GpC)

Figure 3


a)
GpC
average methylation (%)
22
20
18
16
−1000
−500
0
500
1000
distance from nucleosome midpoint in bp
b)
CpG
average methylation (%)
52
48
44
40
−1000
−500
0
500
1000
distance from nucleosome midpoint in bp
c)
GM12878
correlation coefficient
0.50
0.25
0.00
−0.25
0
200
400
offset in bp
d)
K562
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
200
400
offset in bp
e)
density
0.25
0.20
0.15
0.10
0.05
0.00
180
190
200
estimated phase (bp)
GM12878
K562
f)
estimated phase (bp)
200
195
190
185
180
het. chromatin
repressed
reg. activity
txn. activity