Single-molecule mapping of chromatin accessiility using NOMe-seq/dSMF

Michaela Hinks[1,#], **Georgi K. Marinov**[1], **Anshul Kundaje**[1,2], **Lacramioara Bintu**[3], and **William J. Greenleaf**[1,4,5,6,#]

[1] *Department of Genetics, Stanford University, Stanford, CA 94305, USA*
[2] *Department of Computer Science, Stanford University, Stanford, CA 94305, USA*
[3] *Department of Bioengineering, Stanford University, Stanford, CA 94305, USA*
[4] *Center for Personal Dynamic Regulomes, Stanford University, Stanford, California 94305, USA*
[5] *Department of Applied Physics, Stanford University, Stanford, California 94305, USA*
[6] *Chan Zuckerberg Biohub, San Francisco, California, USA*
[#] *Corresponding authors*

Abstract

The bulk of gene expression regulation in most organisms is accomplished through the action of transcription factors (TFs) on *cis*-regulatory elements (CREs). In eukaryotes, these CREs are generally characterized by nucleosomal depletion and thus higher physical accessibility of DNA. Many methods exploit this property to map regions of high average accessibility, and thus putative active CREs, in bulk. However, these techniques do not provide information about coordinated patterns of accessibility along the same DNA molecule, nor do they map the absolute levels of occupancy/accessibility. SMF (**S**ingle **M**olecule **F**ootprinting) fills these gaps by leveraging recombinant DNA cytosine methyltransferases (MTase) to mark accessible locations on individual DNA molecules. In this chapter we discuss current methods and important considerations for performing SMF experiments.

Key words: Enhancers, Promoters, Chromatin accessibility, SMF, High-throughput sequencing

1 Introduction

The development of assays such as ChIP-seq [1, 2] enabled the direct mapping of genome-wide TF binding while methods such as ATAC-seq [3], DNase-seq [4, 5], and MNase-seq [6] have provided unbiased global mapping of accessible DNA and nucleosome positioning, with open chromatin generally being a proxy indicator of TF occupancy. These methods have enabled identification of CREs and the profiling the average occupancy of TFs across the genome. While powerful, identifying genome-wide TF binding in bulk across tens of thousands of cells is insufficient to fully understand mechanisms of TF action. In contrast, single molecule methods such as NOMe-seq [7] (Nucleosome Occupancy and Methylome sequencing) and SMF [8] (single-molecule footprinting) enable profiling of accessible DNA and TF occupancy within individual molecules, thus potentially providing invaluable information about

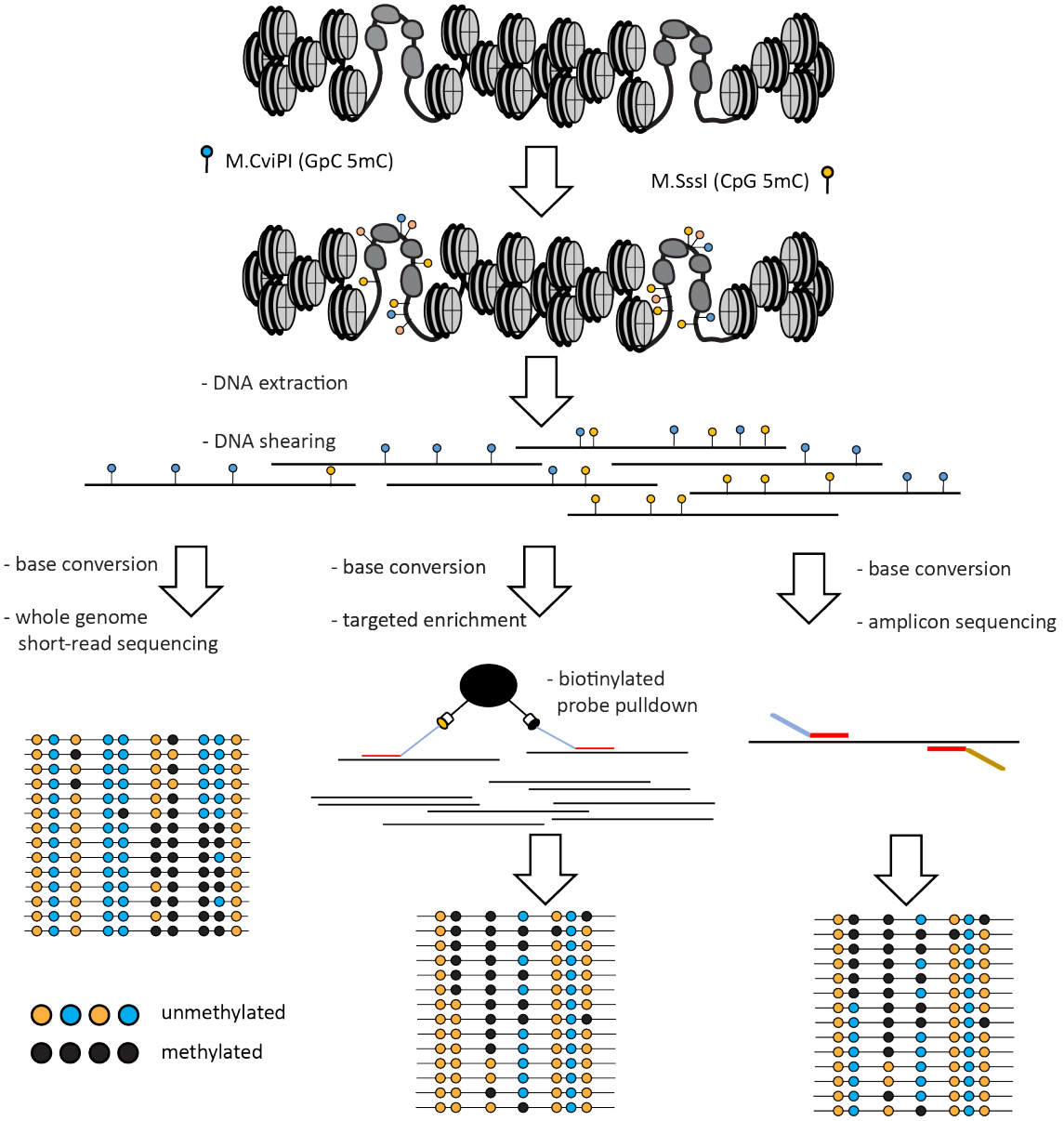

Fig. 1: Outline of the NOMe-seq/dSMF assay. As a first step, nuclei are isolated from cells and chromatin is incubated with the M.CviPI (GpC) and/or M.SssI (CpG) DNA methyltransferases (CpG can usually only be used in biological contexts in which there is no endogenous CpG DNA methylation). DNA is methylated where it is accessible, i.e., where it is not protected by nucleosomes and bound transcription factors. DNA is then purified and fragmented, and chemical or enzymatic conversion is carried out. Three different readout strategies can be applied subsequently – unbiased whole-genome sequencing (left), targeted enrichment using probe-hybridization pulldown, or amplicon sequencing (see the text for more details). After sequencing, single-molecule accessibility maps are generated based on the methylation status of informative positions along DNA.

binding cooperativity and dependencies between individual accessibility states. The core principle underlying all SMF assays is the use of DNA methyltransferases to deposit methyl groups on accessible DNA, followed by detection of the methylation on individual molecules of interest.

Several different versions of the SMF assays can be carried out, based on which DNA MTase, DNA methylation detection method, sequencing modality, and sequence enrichment strategy are used. In this chapter, we provide important considerations for performing SMF experiments intended for sequencing on Illumina instruments, either in an unbiased genome-wide or targeted manner.

2 Materials

Prepare a master stock of the ATAC-RSB buffer without detergents in a large volume (e.g. 50 mL) and store it 4 °C.

2.1 *Methylation Buffers and Reagents*

Prepare the RSB-Lysis and RSB-Wash buffers immediately before use by adding the necessary detergents; keep on ice.

1. IGEPAL CA-630 detergent (Sigma Cat# 11332465001; supplied as a 10% solution).
2. Tween-20 detergent (Sigma Cat# 11332465001, supplied as a 10% solution; store at 4 °C).
3. Digitonin detergent (Promega Cat# G9441, supplied as a 2% solution in DMSO; store at -20 °C)).
4. RSB buffer (master stock)
   10 mM Tris-HCl pH 7.4
   10 mM NaCl
   3 mM $MgCl_2$.
5. RSB-Lysis buffer
   10 mM Tris-HCl pH 7.4
   10 mM NaCl
   3 mM $MgCl_2$
   0.1% IGEPAL CA-630
   0.1% Tween-20
   0.01% Digitonin.
6. Lysis Wash Buffer (RSB-wash)
   10 mM Tris-HCl pH 7.4
   10 mM NaCl
   3 mM $MgCl_2$
   0.1% Tween-20.
7. GpC Methyltransferase (M.CviPI) Reaction Buffer (NEB Cat # B0227SVIAL). This bugger is supplied with GpC Methyltransferase Cat # M0227S as a 10× stock without the S-adenosylmethionine). It final composition (1×) is as follows:
   50 mM NaCl

50 mM Tris-HCl (pH 8.5)
10 mM DTT
32 mM S-adenosylmethionine (SAM) (NEB Cat # B9003SVIAL, supplied with all NEB DNA methyltransferase enzymes). SAM is to be added immediately prior to use. Avoid repeated freeze–thawing of SAM as it is an unstable reagent.

8. GpC MTase (M.CviPI) (NEB Cat # M0227S, supplied at 4,000 units/mL).
9. CpG MTase (M.SssI) (NEB Cat # M0226S, supplied at 4,000 units/mL).
10. $MgCl_2$ (Thermo Fisher Scientific Cat # AM9530G).
11. 2 M Sucrose solution (Sigma Aldrich Cat # S0389).

2.2 *Library building, sequencing and quality evaluation*

1. Monarch Genomic DNA Purification Kit (NEB, Cat # T3010L) or equivalent.
2. NEBNext Enzymatic Methyl-seq Kit (EM-seq, NEB, Cat # E7120L) and associated reagents or EZ-DNA Methylation-Gold Kit (Zymo Research Cat# D5005 (or equivalent), depending on the exact type of SMF experiment being performed (see more details below).
3. *Optional, required if doing probe hybridization enrichment of genomic locations*: SureSelectXT Methyl-Seq Library Preparation kit (Agilent, Cat# G9651A) and associated reagents.
4. *Optional, required if doing probe hybridization enrichment of genomic locations*: SureSelectXT Mouse Methyl-Seq target enrichment panel and associated reagents (Agilent, Cat# 931052) (or equivalent).
5. *Optional, required if doing probe hybridization enrichment of genomic locations* - Dynabeads MyOne Streptavidin T1 (Thermo Fisher Scientific Cat# 65601)
6. Agencourt AMPure XP Kit (Beckman Coulter Genomics Cat# A63880).
7. 10 M NaOH, molecular biology grade (Sigma Cat# 72068).
8. 100% Ethanol, molecular biology grade (Sigma-Aldrich Cat# E7023).
9. 1× Low TE Buffer (10 mM Tris-HCl, pH 8.0, 0.1 mM EDTA) (Thermo Fisher Scientifi Cat# 12090015).
10. 200-μL PCR tubes.
11. Sequencing primers/adapters.
12. NEBNext High-Fidelity 2× PCR Master Mix (NEB, Cat# M0541S).
13. Qubit fluorometer or equivalent.
14. QuBit tubes.
15. QuBit dsDNA HS Assay Kit (Thermo Fisher Scientific Cat# Q328500).
16. TapeStation (Agilent) or equivalent, e.g. BioAnalyzer (Agilent).

17. TapeStation D1000 tape and reagents (Agilent).

2.3 *General materials and equipment*

1. 1.5-mL microcentrifuge tubes, preferably low protein and DNA binding.
2. 2-mL, 15-mL and 50-mL tubes.
3. Incubator (37 °C), or a ThermoMixer.
4. Tabletop centrifuge.
5. Thermal cycler.
6. MinElute PCR Purification Kit (Qiagen Cat# 28004/28006), Zymo DNA Clean and Concentrator Kit (Zymo Cat# D4013/D4014), or equivalent.
7. Nuclease-free H_2O.
8. 1× PBS buffer solution.
9. qPCR machine (StepOne or equivalent).
10. Covaris E220 or equivalent method for shearing genomic DNA (gDNA).

2.4 *Software packages*

1. UCSC Genome Browser [9, 10] utilities: http://hgdownload.cse.ucsc.edu/admin/exe/.
2. R: https://www.r-project.org/.
3. Python (version 2.7 or higher) https://www.python.org/.
4. TGL Kmeans: https://github.com/tanaylab/tglkmeans.
5. SciPy: https://www.scipy.org/.
6. Matplotlib: https://matplotlib.org/.
7. Trimmomatic [11]: http://www.usadellab.org/cms/?page=trimmomatic.
8. Cutadapt [12]: https://cutadapt.readthedocs.io/en/stable/.
9. TrimGalore: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
10. `bwa-meth` [13]: https://github.com/brentp/bwa-meth.
11. `samtools` [14]: http://www.htslib.org/
12. `PicardTools`: https://broadinstitute.github.io/picard/
13. MethylDackel: https://github.com/dpryan79/MethylDackel.
14. Additional scripts: https://github.com/georgimarinov/GeorgiScripts. Contains python scripts used in the examples shown below; some of the scripts depend on having `pysam` (https://pysam.readthedocs.io/en/latest/index.html) and `pyBigWig` (https://github.com/deeptools/pyBigWig) installed.

3 Methods

The general outline of the dSMF assay is shown in Figure 1. Nuclei are first isolated from cells, then chromatin is methylated using a 5mC methyltransferase and genomic DNA is purified. Next, base conversion of unmethylated cytosines are converted to uracils is carried out and sequencing libraries are prepared. In most cases a GpC methyltransferase is used, e.g. M.CviPI, which methylates cytosines in a GpC dinucleotide context. This is because the genomes of mammals, plants and many other species contain endogenous methylation in CpG context. However, if endogenous CpG methylation is not present in the samples being analyzed (e.g. yeast, *Drosophila*, specially engineered mammalian cells that lack endogenous methylation [15], and others), an additional CpG methyltransferase can be used, e.g. M.SssI. This improves the resolution of the assay as the number of informative positions can be increased by a factor of two. Historically, the difference between NOMe-seq [7] and dSMF [8] (dual-enzyme SMF) has been that the latter uses both enzymes.

There are several ways to create a dSMF sequencing library, including via hybridization-based probe enrichment of genomic regions [15], targeted PCR amplification of specific loci, or by unbiased whole-genome sequencing of methylated DNA. Here we describe a generalized protocol for creating dSMF libraries following these approaches using commercially available kits.

We also note that it is possible to carry out SMF on crosslinked material, but we advise that the exact parameters of any such protocol be individually optimized depending on the specifics of the experiment. The protocol described here is for native chromatin.

3.1 *Preparation of nuclei*

The first step of the SMF procedure is to prepare nuclei for methylation. The nuclei lysis delineated here is different from most previously published SMF protocols and identical to the Omni-ATAC cell lysis procedure [16] as we have found that optimal and consistent results are obtained that way. It will work well for most mammalian and insect cell lines. Note that tissues and eukaryotic cells with cell walls (e.g. yeast and plant cells) will require different lysis and nuclei isolation procedures.

1. Count 1×10^6 live cells (if working with a mammalian-sized genome) and aliquot into a microcentrifuge tube (*see* **Note 1**).
2. Centrifuge cells at 500 *g* for 5 min at 4 °C
3. Carefully aspirate the supernatant avoiding the pellet.
4. Wash cells by resuspending in 200 μL ice cold 1× PBS.
5. Centrifuge cells at 500 *g* for 5 min at 4 °C.
6. Aspirate supernatant.
7. Add 200 μL of cold RSB-Lysis Buffer and pipette up and down several times.
8. Incubate on ice for 3 minutes

9. Add 1.2 mL cold RSB-Wash Buffer and invert several times to mix well.
10. Centrifuge at 500 g for 10 min at 4 °C.
11. Carefully aspirate the supernatant as fully as possible while avoiding the pellet. Proceed immediately to methylation.

3.2 *Methylation treatment*

Carry out methylation as follows:

1. Without resuspending, add 100 μL of CviPI Reaction Buffer to cells.
2. To each sample add 50 μL of GpC MTase M.CviPI (4 U/μL). Pipette gently ∼6× to mix (*see* **Note 2**).
3. Incubate at 37 °C for 7.5 min in a Thermomixer at 1,000 rpm.
4. Add more GpC MTase. Add additional 25 μL of low concentration M.CviPI (4 U/μL) and 2.4 μL more 32 mM SAM to the same tube, pipette 3× to mix, and return to 37 °C with shaking for another 7.5 minutes.
5. (*Optional*, *see* **Note 3** and discussion above). Add CpG MTase. Add 3 μL of high concentration (20 U/μL) M.SssI and 2.4 μL more SAM to the same tube, pipette 3× to mix, and return to 37 °C with shaking for another 7.5 minutes.
6. At this point, proceed immediately to DNA purification or freeze cells in methylation solution at −20 °C.

3.3 *DNA purification*

Quench MTase by adding 175 μL of the lysis buffer from the NEB Monarch gDNA extraction kit, along with 3μL RNase A and 1μL Proteinase K (supplied with Monarch gDNA kit). (*see* **Note 4**).

Purify gDNA following the NEB Monarch gDNA extraction kit instructions.

Following elution, quantify gDNA using Qubit.

3.4 *Library preparation – whole genome SMF*

If carrying out whole-genome SMF, libraries are to be generated from this point using standard methods for carrying out whole genome cytosine methylation profiling, in which unmethylated cytosines are converted into uracils and final sequencing libraries are generated from the converted DNA. Two main approaches exist – bisulfite conversion or enzymatic conversion.

For bisulfite conversion, we recommend the EZ-DNA Methylation-Gold Kit. However, bisulfite conversion leads to fragmentation of DNA, often to shorter fragments than what is desired for SMF experiments, where a key objective is to obtain molecules as long as can be sequenced on a short-read platform and thus maximize the information contained within each single molecule. Bisulfite conversion generally leads to fragments shorter than 200 bp, often considerably shorter, which has historically necessitated careful size selection of the subsequently generated libraries in order to maximize the coverage of long fragments [8].

Enzymatic conversion using the NEB EM-seq kit offers an attractive alternative as it does not degrade DNA and fragment size can be carefully controlled. As a first step before entering the EM-seq procedure, DNA needs to be sheared to the desired size. The Covaris E220 instrument allows a convenient solution for controlling fragment length, but other methods for shearing can be used too.

Note that the EM-seq kit contains important positive and negative controls – pUC19 and Lambda DNA, that are respectively fully methylated and unmethylated, and are invaluable for monitoring the efficiency of methylation conversion. Either add those to your samples as a ∼1% spike-in before shearing, or maintain a stock of pre-sheared pUC19 and Lambda to be mixed with sonicated samples prior to conversion. If using bisulfite conversion, use unsheared controls.

Depending on the exact kit used, follow the manufacturer's instructions for final sequencing library generation.

3.5 *Library preparation – probe hybridization enrichment*

A significant practical challenge to the application of single-molecule footprinting approaches to mammalian genomes is the very high sequencing depth that needs to be achieved in order to fully take advantage of the wealth of information contained in single molecules. These reads need to be as long as possible too (see further discussion below). Consequently, sequencing costs quickly become a major consideration when working with large genomes.

However, given that most of the genome is inaccessible and active CREs comprise only a small portion of it, it is possible to greatly reduce costs by using hybridization capture to enrich for a desired subset of the genome. As an example, this approach has been previously successfully used to apply dSMF to many thousands of promoters and enhancers in the mouse genome [15], using the SureSelectXT Mouse Methyl-Seq target enrichment panel from Agilent. Other probe sets and enrichment protocols are likely to work as well.

The exact details of the protocol will vary depending on the specifics of the kit used. A general outline of the procedure is as follows:

For a probe hybridization enrichment dSMF experiment, footprinted DNA is sheared using a Covaris sonicator, end-repaired and methylated adapters are ligated, creating a pre-capture library. Adapters need to be methylated in order to block their conversion during the subsequent steps and enable PCR amplification. The pre-capture library is then hybridized with a biotinylated set of target probes, and purified using streptavidin bead capture. The captured molecules are subjected to bisulfite conversion, and then PCR-amplified.

3.6 *Library preparation – amplicon-targeted SMF*

Even greater levels of enrichment and depth of coverage can be obtained by selectively amplifying individual loci. This approach works best together with the EM-seq conversion kit because, as discussed above, it provides better preserved DNA compared to bisulfite treatment. Footprinted whole-genome DNA is used as input and carried through the

EM-seq procedure up to the last, final library amplification step. Then PCR primers specific for a locus (or loci) of interest are used to make the final targeted library. The challenge when using this approach is that PCR primers need to be selected and/or designed in such a way that they work on converted DNA; the exact specifics of that selection will vary depending on the particulars of the experiment carried out.

3.7 *Library quantification and evaluation of library quality*

Before libraries can be sequenced, they need to be properly quantified, and their quality evaluated. There are two components to this process – first, evaluation of the insert distribution, and second, quantification.

1. Examination of library size distribution. This step can be carried out using a variety of instruments that are now available for this purpose, such as a TapeStation or a BioAnalyzer. In our practice we prefer to use a TapeStation (with the D1000 or HS D1000 kits) due to its ease of use, flexibility and rapid turnaround time.
2. Quantification of library concentration. For most high-throughput sequencing applications, where fragment size is unimodal, this step can be carried out with a sufficient degree of accuracy using a Qubit fluorometer. Typically, dSMF falls in that category. For libraries with complex fragment distributions, as well as for higher accuracy of quantification, qPCR can be used. Commercial kits, such as the NEBNext Library Quant Kit for Illumina or the KAPA Library Quantification Kits, exist for that purposes, and custom in-house quantification methods can also be used (see the first chapter in this book on ATAC-seq for details).

3.8 *Sequencing*

The protocol described here generates libraries designed to be sequenced on Illumina sequencers. Because every molecule in an SMF library contains information about its unique accessibility state throughout the sequence, it is advisable to perform longer read-length sequencing than is necessary to simply align the fragments. It is best to sequence all molecules completely (e.g. a 300-bp insert would be sequenced with 150 cycles in Read 1 and 150 cycles in Read 2). Paired-end sequencing is preferable to single-end sequencing to improve quality, though single-end will also work provided the reads are sufficiently long. It is recommended to sequence SMF libraries to high depth, i.e. $\sim$1000$\times$ coverage of the size of the probed portion of the genome. This leads to, on average, 1000 unique molecules per genomic locus that are each read once. Sequencing depth can be adjusted based on the user probe set and the frequency of accessibility states observed.

Due to the relatively high cost of longer-read Illumina sequencing, users may wish to perform quality control checks on the library prior to full sequencing. A useful way to verify that the library is complex and captures chromatin accessibility is to sequence it at a fraction of the optimal depth using shorter read-length sequencing (e.g. as 2$\times$36mers). This way, the user can check that methylation is detected at GpC locations

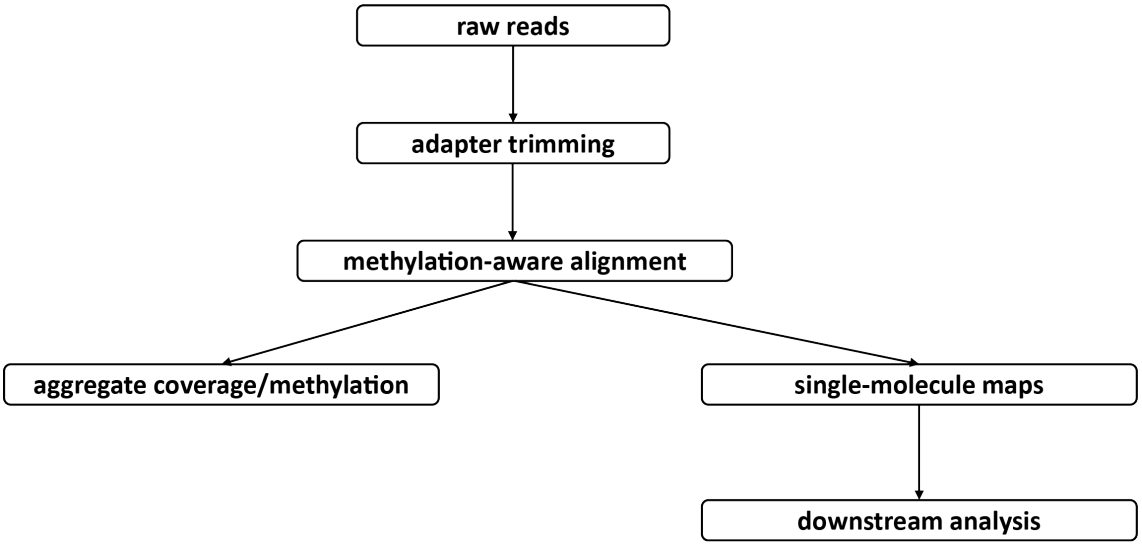

Fig. 2: Outline of NOMe-seq/dSMF computational processing. Raw sequencing reads are first trimmed of adapters (note that it is important to do this properly depending on the type of conversion protocol used for making the libraries). They are then aligned against the target genome or amplicons in a methylation-aware manner. Subsequently, alignments are used to make aggregate methylation tracks (if data is to be used to evaluate bulk accessibility) and single-molecule plots (for actual footprinting).

and ensure that there is a diversity of probed regions represented.

3.9 *Computational analysis*

The overall outline of NOMe-seq/dSMF data processing is shown in Figure 2. Briefly, reads are trimmed of adapters and then aligned against the genome or a set of target amplicons. These alignments are then used to evaluate bulk-level accessibility and to carry out analysis at the level of individual single molecules.

3.9.1 *Adapter trimming*

If working with EM-seq datasets, Trimmomatic can be used to trim adapters as follows:

```
java -jar trimmomatic-0.36.jar PE
EM-seq.read1.fastq.gz EM-seq.read1.fastq.gz
EM-seq.read1.paired.fastq.gz
EM-seq.read1.unpaired.fastq.gz
EM-seq.read2.paired.fastq.gz
EM-seq.read2.unpaired.fastq.gz
ILLUMINACLIP:Trimmomatic-0.36/adapters/adapters.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

For bisulfite data, we recommend clipping the first 9 bases off reads due to their usually lower quality in addition to adapter remova, using `trim_galore`, as follows:

```
trim_galore
```

```
--path_to_cutadapt ./cutadapt
--clip_R1 9 --clip_R2 9
--three_prime_clip_r1 6
--three_prime_clip_r2 6
--paired EM-seq.read1.fastq.gz
         EM-seq.read1.fastq.gz
```

3.9.2 *Read mapping and alignment filtering*

We use BWAmeth for alignment of base-converted datasets.

1. The first step is to prepare a reference, as follows:

   ```
   python bwameth.py index bwameth-indexes/genome.fa
   ```

2. Next, reads are mapped against the reference (while filtering out low quality alignments and unmapped reads):

   ```
   python bwameth.py --reference bwameth-indexes/genome.fa
   EM-seq.read1.paired.fastq.gz EM-seq.read2.paired.fastq.gz
   | samtools view -F 1804 -q 30 -bT
   bwameth-indexes/genome.fa - | samtools sort -
   EM-seq.bwameth
   ```

3. The next step is to remove potential PCR duplicates. Note that this step generally applies only to whole-genome and probe-capture libraries where there is diversity of fragment coordinates.

   ```
   java -Xmx4G -jar picard-tools-1.99/MarkDuplicates.jar
   INPUT=EM-seq.bwameth.bam OUTPUT=EM-seq.bwameth.dedup.bam
   METRICS_FILE=EM-seq.bwameth.dedup.metric
   VALIDATION_STRINGENCY=LENIENT ASSUME_SORTED=true
   REMOVE_DUPLICATES=true
   ```

4. Use `samtools` to index the final BAM file:

   ```
   samtools index EM-seq.bwameth.dedup.bam
   ```

3.9.3 *Methylation conversion assessment*

In order to evaluate the efficiency of methylation conversion, use the same procedure as described above to map reads against the Lambda and pUC19 genomes.

Then use the custom `MethylationPercentageContext.py` script to calculate the average methylation levels in GpC and CpG contexts, e.g. for pUC19:

```
python MethylationPercentageContext.py
EM-seq.pUC19.dedup.bam pCU19.fa CG,GC
EM-seq.pUC19.dedup.CG-GC-meth-perc
```

The pUC19 plasmid is the methylated positive control and should show very high (90%+) levels of specifically CpG methylation, while Lambda DNA is the unmethylated negative control and should exhibit minimal levels of methylation.

3.9.4 *Methylation calling*

The next step is to extract methylation calls, using `MethylDackel`:

```
MethylDackel extract --CHG --CHH genome.fa
EM-seq.bwameth.dedup.bam
```

Note the parameters used so that both CpG and GpC contexts are included in the output. However, further filtering is needed in order to specifically obtain GpC positions, described further below.

3.9.5 *Bulk accessibility or methylation profile generation*

For the purpose of generating bulk accessibility profiles (this is often useful for genome browser visualization of results), execute the following steps:

1. Compress the `MethylDackel` output:

   ```
   gzip EM-seq.bwameth.dedup_CHG.bedGraph
   gzip EM-seq.bwameth.dedup_CHH.bedGraph
   gzip EM-seq.bwameth.dedup_CpG.bedGraph
   ```

2. Extract GpC positions from the `MethylDackel` output, for GpC positions, using the `MethylationPercentageSmooth-dSMF.py` custom script:

   ```
   python MethylationPercentageSmooth-dSMF.py
   EM-seq.bwameth.dedup_CHH.bedGraph.gz
   genome.fa GpC 10 -MethylDackel -minCov 10 >
   EM-seq.bwameth.dedup_CHH.GpC-only.minCov10.wig
   ```

3. Do the same for CpG positions:

   ```
   python MethylationPercentageSmooth-dSMF.py
   EM-seq.bwameth.dedup_CpG.bedGraph.gz
   genome.fa CpG 10 -MethylDackel -minCov 10 >
   EM-seq.bwameth.dedup_CHH.CpG-only.minCov10.wig
   ```

   Note that in this case we also apply a minimal coverage cutoff of 10 reads. This can be adjusted as needed.

4. These steps create bedGraph files from which bigWig files to be used on a genome browser can be generated:

   ```
   UCSC-utils/wigToBigWig
   EM-seq.bwameth.dedup_CHH.GpC-only.minCov10.wig
   genome.chrom.sizes
   EM-seq.bwameth.hg38.dedup_CHH.GpC-only.minCov10.bigWig
   ```

   Note that for this step a `chrom.sizes` file is needed. This file can be created using the `makeChromSizesFromFasta.py` custom script

5. It is also often useful to know what the raw read coverage is along the genome. This can be generated using many different existing tools, in this case we use the custom `makewigglefromBAM-NH.py` script:

   ```
   python makewigglefromBAM-NH.py title
   EMs-eq.dedup.bam genome.chrom.sizes
   EMs-eq.dedup.coverage.wig -uniqueBAM
   ```

   Convert into a bigWig file as above.

3.9.6 *Metaprofile evaluation*

It is often useful to generate metaplots over a defined set of genomic features, for quality evaluation (e.g. assessing how strong the methylation levels are around active promoters) and for other analysis tasks (e.g. measuring average footprinting by TFs at their occupancy sites).

1. As a first step, extract the wanted sequence contexts from the `MethylDackel` output using the `BismarckSequenceContextFilter.py` custome script, e.g. as follows for GpC:

   ```
   python BismarckSequenceContextFilter.py
   EM-seq.bwameth.hg38.dedup_CHH.bedGraph.gz GC
   genome.fa | gzip >
   EM-seq.bwameth.hg38.dedup_CHH.GpC-only.bedGraph.gz
   ```

2. Then generate the metaprofile using the `signalAroundPeaks-nano.py` custom script. This script can be run with a wide variety of inputs and window lengths around the desired viewpoints. In this example, we use it to generate a metaprofile around annotated transcription start sites:

   ```
   python signalAroundPeaks-nano.py
   annotation.TSS-0bp.bed 0 1 3 1000 10
   EM-seq.bwameth.hg38.dedup_CHH.GpC-only.bedGraph.gz
   EM-seq.bwameth.hg38.dedup_CHH.GpC-only.TSS_profile
   -bismark.cov
   ```

   The `annotation.TSS-0bp.bed` can be generated from a GTF files using the `TSS_bed_FromGTF.py` custom script.

3.9.7 *Generating single-molecule maps*

Finally, we illustrate the generation of single molecule maps. This is done using the `dSMF-footprints.py` script, which has as a dependency the `heatmap.py` custom script, and has a wide variety of options for color adjustment, minimal read coverage filtering, and others.

It takes as input a BAM file, a BED file with the windows over which single molecules are to be plotted, the genome sequence, and the sequence context(s) (`GC`, `CG`, or `both`).

```
python dSMF-footprints.py EM-seq.bwameth.hg38.dedup.bam
genome.fa GC region.bed 0 1 2 3 EM-seq
-heatmap heatmap.py 10 10 binary 10,100 -minCov 0.9
```

In this case we filter out all alignments that do not cover at least 90% of the input regions, and plot the single molecules using the `binary` Matplotlib colormap, meaning that methylated positions will be shown as light while protected unmethylated positions will be shown in black.

4 Expected results

Figures 3A shows bulk accessibility, CpG methylation maps and raw read coverage tracks for previously published [15] probe capture dataset in mouse.

Figures 3B and 3C show typical CpG (endogenous methylation) and GpC (accessibility) metaprofiles around transcription start sites and around occupancy sites of the CTCF transcription factors (which is well known to be a strong driver of nucleosomal occupancy in the vicinity of its occupancy sites [17]) for the same dataset.

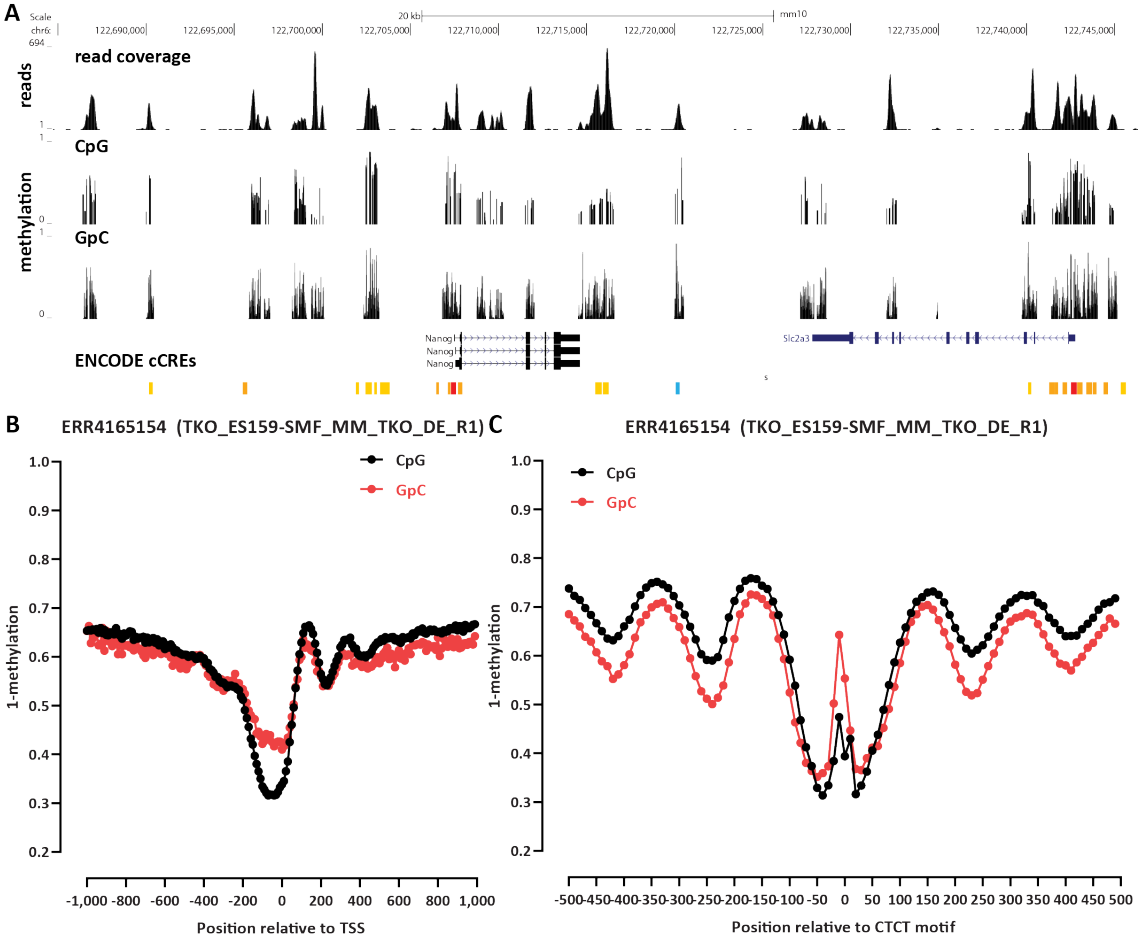

Fig. 3: Aggregate accessibility analysis of NOMe-seq/dSMF datasets. (A) Read coverage and aggregate CpG/GpC methylation genome browser tracks show accessibility and/or endogenous methylation levels around the genome. In this case, reduced representation probe-capture dSMF dataserts (obtained from ArrayExpress accessions E-MTAB-9033 and E-MTAB-9123) is show, thus the uneven coverage; (B) Metaplot showing average accessibility levels around TSSs in the mouse genome; (C) Metaplot showing average footprinting levels at occupied CTCF motifs.

Examples of single-molecule maps showing footprint protection around binding sites for the CTCF transcription factor are shown in Figure 4.

Fig. 4: Examples of single-molecule accessibility measurements. Shown are dSMF single-molecule maps (obtained from ArrayExpress accessions E-MTAB-9033 and E-MTAB-9123 [15]). (A) High levels of occupancy by the CTCF transcription factor (middle). (B) CTCF (middle) and possible nucleosome (left) footprints.

5 Notes

1. Note that the efficiency of the methylation reaction is potentially dependent on the ratio between the amount of enzyme present and the quantity of input material. Therefore one should be careful to avoid using too many cells as this could lead to suboptimal level of methylation in open chromatin. The input cell number should be scaled according to genome size and ploidy, i.e. a fission yeast cell (a ∼12-Mbp haploid genome) contains ∼500× less chromatin than a typical human cell (a ∼3-Gbp diploid genome).
2. We have found that the high concentration of glycerol in the final methylation reaction is important for maintaining cell solubility and has little or no adverse effect on methyltransferase function. As a result, we recommended to use the low concentration of methyltransferase supplied by NEB at the time of writing. If using higher concentrations of enzymes from another source, adding extra glycerol to the methylation reaction may help to prevent cells from clumping together. Low concentrations of non-ionic detergents such as Tween-20 may also prevent cell clumping, but further optimization would be required.
3. It is possible to do single molecule footprinting for intended use with Illumina sequencing with only one or both of GpC or CpG methyltransferases. The particular application will determine which option is advisable. When working in organisms such as *Drosophila* that contain no endogenous DNA methylation, using both enzymes is recommended for maximal footprinting resolution. CpG methylation exists endogenously in mammalian cells, so users may opt to only use GpC methyltransferase in order to distinguish between natural and synthetic DNA methylation.
4. In our experience, using the Monarch Genomic DNA Purification Kit is the easiest way to obtain high quality, purified genomic DNA after methylation. However, other methods, such as phenol-chloroform extraction, have been demonstrated to work. It is likely other kits for genomic DNA extraction also perform well. If using a different column-based gDNA purification kit, care should be taken to increase the amount of DNA binding and cell lysis buffers to ensure the ratio of kit buffer volume to sample buffer volume remains the same as in the manufacturer's instructions. Inappropriate buffer volume ratios may lead to poor DNA binding to the column and subsequent low yield.

Acknowledgements

The authors thank members of the Greenleaf, Bintu and Kundaje labs for many helpful discussions. This work was supported by NIH grants (P50HG007735, RO1 HG008140, U19AI057266 and UM1HG009442 to W.J.G., 1UM1HG009436 to W.J.G. and A.K., 1DP2OD022870-01 and 1U01HG009431 to A.K., and HG006827 to C.H.), the Rita Allen Foundation (to W.J.G.), the Baxter Foundation Faculty Scholar Grant, and

the Human Frontiers Science Program grant RGY006S (to W.J.G). W.J.G is a Chan Zuckerberg Biohub investigator and acknowledges grants 2017-174468 and 2018-182817 from the Chan Zuckerberg Initiative. Fellowship support provided by the Stanford School of Medicine Dean's Fellowship (G.K.M.).

References

1. Johnson DS, Mortazavi A, Myers RM, Wold B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830):1497–1502.
2. Mikkelsen TS, Ku M, Jaffe DB, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**(7153):553–560.
3. Buenrostro JD, Giresi PG, Zaba LC, et al. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218.
4. Crawford GE, Holt IE, Whittle J, et al. (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**, 123–131.
5. Boyle AP, Davis S, Shulha HP, et al. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322.
6. Schones DE, Cui K, Cuddapah S, et al. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**(5):887–898.
7. Kelly TK, Liu Y, Lay FD, et al. (2012) Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* **22**, 2497–2506.
8. Krebs AR, Imanci D, Hoerner L, Gaidatzis D, et al. (2017) Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol Cell* **67**, 411–422.e4.
9. Kuhn RM, Haussler D, Kent WJ (2013) The UCSC genome browser and associated tools. *Brief Bioinform* **14**, 144–161.
10. Kent WJ, Zweig AS, Barber G, et al. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207.
11. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15):2114–2120.
12. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**(1):10–12.
13. Pedersen BS, Eyring K, De S, et al. (2014) Fast and accurate alignment of long bisulfite-seq reads. *arXiv* 1401.1129
14. Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.
15. SÖnmezer C, Kleinendorst R, Imanci D, Barzaghi G, Villacorta L, Schübeler D, Benes V, Molina N, Krebs AR. 2021. Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo. *Mol Cell* **81**(2):255–267.e6.
16. Corces MR, Trevino AE, Hamilton EG, et al. (2017) An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**, 959–962.
17. Fu Y, Sinha M, Peterson CL, Weng Z. 2008. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* **4**(7):e1000138