Chapter 15

Single-Molecule Multikilobase-Scale Profiling of Chromatin Accessibility Using m6A-SMAC-Seq and m6A-CpG-GpC-SMAC-Seq

Georgi K. Marinov, Zohar Shipony, Anshul Kundaje, and William J. Greenleaf

Abstract

A hallmark feature of active cis-regulatory elements (CREs) in eukaryotes is their nucleosomal depletion and, accordingly, higher accessibility to enzymatic treatment. This property has been the basis of a number of sequencing-based assays for genome-wide identification and tracking the activity of CREs across different biological conditions, such as DNAse-seq, ATAC-seq, NOMeseq, and others. However, the fragmentation of DNA inherent to many of these assays and the limited read length of short-read sequencing platforms have so far not allowed the simultaneous measurement of the chromatin accessibility state of CREs located distally from each other. The combination of labeling accessible DNA with DNA modifications and nanopore sequencing has made it possible to develop such assays. Here, we provide a detailed protocol for carrying out the SMAC-seq assay (Single-Molecule long-read Accessible Chromatin mapping sequencing), in its m6A-SMAC-seq and m6A-CpG-GpC-SMAC-seq variants, together with methods for data processing and analysis, and discuss key experimental and analytical considerations for working with SMAC-seq datasets.

Key words Chromatin accessibility, SMAC-seq, Nanopore sequencing, DNA modifications, m6A, EcoGII

1 Introduction

Chromatin accessibility is a key feature of the regulation of gene expression and many other aspects of chromatin biology in eukaryotes. Nearly all eukaryote genomes are packaged by nucleosomes, with each nucleosome being a dimer of two tetramers composed of the four core nucleosomal histones H3, H4, H2A and H2B. Packaging by nucleosomes has a generally inhibitory effect on RNA polymerase activity and to the occupancy of DNA by regulatory

Georgi K. Marinov and Zohar Shipony contributed equally with all other contributors.

Julia Horsfield and Judith Marsman (eds.), *Chromatin: Methods and Protocols*,
Methods in Molecular Biology, vol. 2458, https://doi.org/10.1007/978-1-0716-2140-0_15,

proteins. Accordingly, active regulatory regions in the genome are characterized by depleted nucleosomal occupancy and increased chromatin accessibility. This has turned out to be a highly useful property enabling the identification of candidate cis-regulatory elements and the tracking of their activity across cell types and conditions.

Mapping accessible chromatin relies on the preferential enzymatic action of various reagents whose access to DNA is occluded by the presence of nucleosomes. Four decades ago it was initially recognized that active cREs are hypersensitive to cleavage by DNase enzymes [1–3]. DNase hypersensitivity remained the primary approach for mapping cREs well into the genomic era, being first coupled to microarrays [4–6], and eventually high-throughput massively parallel sequencing [7–9].

The advent of high-throughput sequencing enabled the development of numerous novel strategies for mapping active CREs. ATAC-seq [10], which relies on the preferential insertion of the Tn5 transposase enzyme into open chromatin, has emerged as the most convenient, versatile, and widely used method for studying the chromatin state of the eukaryotic cell, including down to single cell level [11, 12].

Other methods have also been developed, using restriction enzymes [13], nicking enzymes [14], small molecules [15], viral integration [16], and others.

All of these methods share two common features—they involve fragmentation of DNA and they enrich for accessible DNA during sequencing library generation. Consequently, it is first, not possible to enumerate accessibility states within the cellular population, that is, how often is a given CRE accessible, and second, there is no way to study the relationship between the chromatin states of distant regulatory elements, as the linkage between them is lost during fragmentation.

An alternative strategy to cleavage-based methods is to label accessible DNA with methyltransferase enzymes, then read out methylation states using high-throughput sequencing. This is the basic idea behind the NOMe-seq assay [17] and its later dSMF extension [18]. NOMe-seq uses the GpC methyltransferase M. CviPI to label accessible DNA at GpC positions. Genomic DNA is then subjected to bisulfite readout, providing single-molecule and fractional methylation (and thus accessibility) maps genome-wide. Only M.CviPI can be used in mammalian genomes due to the presence of endogenous CpG methylation, and only the m5C modification can be utilized as this is what can be read out with base pair resolution using short-read sequencing. This presents a limitation, as GpC nucleotides are only found once every ~25 bp in a mammalian genome. In organisms such as Drosophila that do not have endogenous methylation, both a GpC and a CpG methyltransferase (M.SssI) can be used, increasing resolution to ~10 bp

on average, in the form of the dSMF assay. This has allowed the enumeration of protein occupancy states at unprecedented resolution at a single-molecule level [18]. Yet short-read approaches of this kind are still quite limited in their capabilities.

First, these resolution values are averages. In reality genomes contains some quite large stretches with no informative positions (Fig. 1), and not much can be done to address that limitation as long as m5C in GpC/CpG contexts is the only available modification.

Second, it is only possible to analyze fragments no longer than 600 bp due to read-length limitations of short-read sequencers. Even this has been very difficult to achieve, as DNA methylation has traditionally been mapped using bisulfite sequencing, and bisulfite treatment severely degrades DNA to lengths considerably shorter than 600 bp. The introduction of the EM-seq method [19] as an alternative to bisulfite conversion has largely eliminated the degradation issue, but short reads are still short reads, making it impossible to study chromatin states on the scale of many kilobases along the chromatin fiber.

With the advent of long read sequencing technologies, and especially nanopore sequencing, these limitations have been overcome. Nanopore sequencing is capable of reading out arbitrary DNA modifications [20, 21], and of doing so along the length of DNA molecules tens of kilobases long, allowing for the simultaneous capture of the chromatin states of CREs located far apart. This has enabled the development of a qualitatively new class of functional genomic assays [22–24].

The MeSMLR-seq [23] and nanoNOMe [24] assays have adapted the NOMe-seq approach to nanopore sequencing, using a GpC methyltransferase to label accessible DNA, then reading it out using nanopore sequencing. However, while this approach preserves long-range contiguity, it still suffers from the limitations imposed by the density of informative modification positions in the genome (Fig. 1).

In contrast, SMAC-seq [22] uses dense modifications, found once every few nucleotides in the genome. Accessible DNA is enzymatically labeled using a methyltransferase enzyme (or multiple such enzymes), high molecular weight (HMW) DNA is isolated, then subjected to nanopore sequencing, which allows for the direct detection of DNA modifications and thus the assembly of an accessibility map at the single molecule level and on multikilobase scales (Fig. 2). In addition, the dense modifications that SMAC-seq is based on also provide information about nucleosome occupancy/positioning [25] and even transcription factor footprints [26, 27]. Finally, the long reads provided by nanopore sequencing allow chromatin accessibility and nucleosome positioning to be profiled within repetitive regions of the genome that are otherwise not uniquely mappable using short reads.

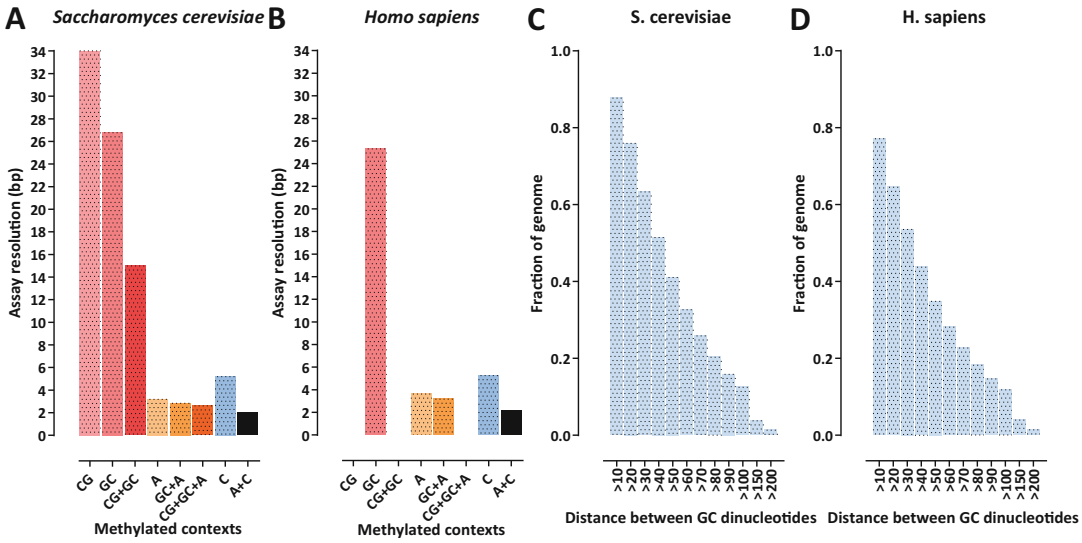

Fig. 1 Overview of the SMAC-seq experimental protocol. Chromatin is treated with m6A and optionally also with CpG and GpC 5mC methyltransferases, which preferentially methylate DNA bases within accessible chromatin. HMW DNA is then isolated and subjected to nanopore sequencing. After read mapping and identifying modified bases, the accessibility state within individual chromatin fibers can be reconstructed

Here, we describe an m6A-SMAC-seq protocol based on the m6A (N6Methyladenosine) methyltransferase EcoGII [28], which labels A bases nonspecifically in all contexts (*see* **Note 1**) as well as a m6A-CpGGpC-SMAC-seq protocol, which uses multiple modifications (m6A and m5C modifications in CpG and GpC contexts) and which can be used in organisms without endogenous DNA methylation. We also describe basic data processing and analysis procedures for working with SMAC-seq datasets.

2 Materials

SMAC-seq uses standard laboratory reagents with the exception of the m6A methyltransferase in the m6A version of the assay (*see* **Note 2**). Other versions of the assay involving different modifications may also require custom reagents.

2.1 SMAC-Seq Buffers and Reagents

1. Nuclei Lysis Buffer: 10 mM Tris–HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA, 0.5% NP-40.
2. Nuclei Wash Buffer. This is the same as the Lysis Buffer except for the absence of NP-40: 10 mM Tris pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA.
3. M.CviPI Reaction Buffer: 50 mM Tris-HCl pH 8.5, 50 mM NaCl, 10 mM DTT,

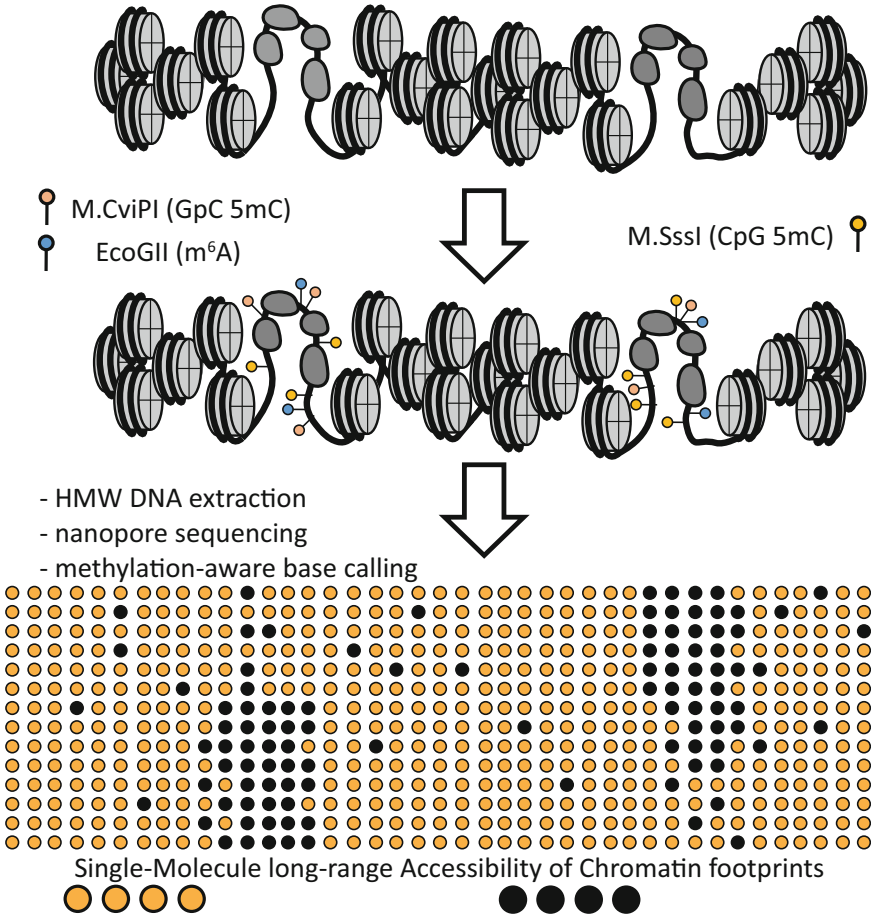

Fig. 2 Impact of the use of dense modifications on the theoretical resolution of methylation-based chromatin accessibility assays. (**a** and **b**) Theoretical average resolution of the SMAC-seq assay for different modification sequence contexts in the *S. cerevisiae* and *H. sapiens* genomes. (**c** and **d**) Limitations of using GpC m5C modifications alone due to the nonuniform distribution of GpC dinucleotides in the genome, which results in many large regions without any informative positions

4. CutSmart Reaction buffer: 1× CutSmart buffer, 0.3 M sucrose.
5. Stop Buffer: 20 mM Tris pH 8.5, 600 mM NaCl, 1% SDS, 10 mM EDTA.
6. Sorbitol Buffer: 1.4 M Sorbitol, 40 mM HEPES-KOH pH 7.5, 0.5 mM $MgCl_2$.
7. 100 T Zymolase
8. M.CviPI methyltransferase.
9. EcoGII methyltransferase (*see* **Note 2**).
10. M.SssI methyltransferase.

11. 32 mM S-adenosylmethionine (SAM)
12. Sucrose solution (prepare a highly concentrated solution, e.g., 2 M).
13. Molecular biology-grade 1 M $MgCl_2$ solution.

2.2 DNA Isolation and Size Selection

1. HMW DNA. We have most often used the MagAttract HMW DNA Kit (Qiagen), but other approaches for isolating HMW can also be applied, such as the NEB Monarch Genomic DNA Purification Kit, the Nanobind CBB Big DNA Kit, and others.
2. Size selection. Several solutions now also exist for HMW size selection that eliminates shorter fragments. We have used the Short Read Eliminator Kit (Circulomics) with fairly consistent levels of success, but equivalent approaches are also applicable.

2.3 Nanopore Sequencing Flow Cells and Reagents

Nanopore and SMAC-seq data can be generated using any of the Oxford Nanopore Technologies (ONT) platforms (Flongle, MinION, GridION, or PromethION). Which one to use is a decision to be made on the basis of the desired output, which in turn is determined by the needed coverage based on genome size, the properties of the genome studied, and so on (*see* **Notes 4** and **5**). ONT offers a variety of library preparation options, the two main ones relevant to SMAC-seq being the following.

1. The Ligation Sequencing Kits are to be used if maximum read length is desired. These require ~1000 ng of input HMW DNA.
2. The Rapid Barcoding Kit uses a transposases to simultaneously fragment DNA and attach adapters to the resulting pieces. Thus it will yield shorter molecules ($\leq$ ~10 kbp) but it allows the pooling of multiple samples in the same run (which is useful if, for example, working with an organism with a small genome and on a PromethION) and works with smaller amount of input HMW DNA (~400 ng).

Which kit is to be used depends on what the optimal choice is with respect to the particular research question and experimental system.

2.4 General Materials and Equipment

1. 1.5-mL microcentrifuge tubes, preferably low protein and DNA binding (*see* **Note 6**).
2. 2 mL, 15 mL and 50 mL tubes.
3. Magnetic stands for 1.5 mL and 2 mL tubes.
4. Thermomixer.
5. Molecular biology-grade 200 proof EtOH.
6. Tabletop centrifuge.
7. Nuclease-free H_2O.

8. 1× PBS.
9. AMPure beads.
10. QuBit fluorometer (ThermoFisher Scientific) or equivalent.
11. QuBit dsDNA HS kit (ThermoFisher Scientific).
12. TapeStation (Agilent) or equivalent.
13. TapeStation Genomic DNA Reagents (Agilent).
14. TapeStation Genomic DNA Screentape (Agilent).

2.5 Computational Resources

The computational analyses described are designed to run on standard Linux systems through the UNIX command line. The maximal memory usage depends on the size of the datasets but is usually less than ~50 GB. However, note that nanopore sequencing datasets can occupy very large amounts of disk space (i.e., many terabytes), thus it is advisable to use a computing system with ample storage (*see* **Note 7**).

2.6 Genomic Sequence and Annotation Files

1. A FASTA file containing the GRCh38 version of the human genome can be downloaded from the UCSC Genome Browser at http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz. Genome files can also be obtained from ENSEMBL (http://ensemblgenomes.org/) and from the NCBI website (http://www.ncbi.nlm.nih.gov/assembly/). However, it has to be noted that in the case of the human genome, reference FASTA files available in public repositories contain alternative haplotype contigs, that is, alternative versions of sequences already present in the assembly. These alternative haplotypes should be removed from reference files before use. The ENCODE Project [29] provides such filtered files from its portal at https://www.encodeproject.org/data-standards/reference-sequences/. The sacCer3 version of the Saccharomyces cerevisiae genome can be obtained from http://hgdownload.cse.ucsc.edu/goldenPath/sacCer3/bigZips/sacCer3.fa.gz
2. Genome annotations in GTF format can be obtained from UCSC, ENSEMBL, NCBI or ENCODE.

2.7 Software Packages

1. UCSC Genome Browser [30, 31] 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. Minimap2 [32] (version 2.17) https://github.com/lh3/minimap2.
8. Tombo [33] (version 1.5) https://nanoporetech.github.io/tombo/.
9. Albacore https://nanoporetech.com/.
10. Megalodon https://github.com/nanoporetech/megalodon.
11. Guppy https://nanoporetech.com/.
12. Rerio https://github.com/nanoporetech/rerio.
13. tabix: http://www.htslib.org/doc/tabix.html (*see* **Note 8**).
14. Additional scripts: https://github.com/georgimarinov/SMAC-seq-scripts. Contains python scripts for processing and post-processing of SMACseq data used in the examples shown below.

3 Methods

The principle behind the assay and the typical SMAC-seq experimental procedure are outlined in Fig. 2. SMAC-seq consists of the following basic steps:

1. Nuclei isolation.
2. Enzymatic treatment of chromatin.
3. HMW DNA extraction.
4. Nanopore sequencing.
5. Read mapping and calling modified basis.
6. Aggregate and single-molecule accessibility analysis.

We provide several slightly different protocols for working with yeast (*see* **Note 12**) as well as with mammalian and fly cells.

3.1 Nuclei Isolation (Budding Yeast)

Start with 2.5×10^8 yeast cells (the equivalent to 1×10^6 human cells).

1. Spin cells for 1 min at 13000 rpm. Remove supernatant.
2. Wash cells with 100 μL Sorbitol Buffer.
3. Spin cells 1 min at 13000 rpm. Remove supernatant.
4. Resuspend pellet in 200 μL Sorbitol Buffer + 10 mM DTT + 0.5 mg/mL 100T Zymolase.
5. Incubate for 5 min at 30 °C, shaking 300 rpm.
6. Centrifuge for 2 min at 5000 rpm. Remove supernatant.
7. Add 100 μL SB buffer (no DTT) and resuspend gently.

8. Centrifuge for 2 min at 5000 rpm. Remove supernatant.
9. Add 100 μL ice-cold lysis buffer.
10. Incubate on ice for 10 min.
11. Spin down at 5000 rpm for 5 min at 4 °C.
12. Wash with 100 μL cold wash buffer.
13. Spin at 5000 rpm for 5 min at 4 °C.
14. Resuspend in M.CviPI Reaction Buffer (100 μL).

3.2 Enzymatic Treatment of Chromatin for m6A-GpC-CpGSMAC-Seq (Budding Yeast)

1. Add 200 U of M.CviPI and 200 U of EcoGII.
2. Add SAM to a final concentration of 0.6 mM, and sucrose to a final concentration of 300 mM.
3. Incubate at 30 °C for 7.5 min.
4. Add 128 pmol SAM (= 4 μL 32 mM solution) and another 100 U of both enzymes.
5. Incubate at 30 °C for 7.5 min.
6. Add 60 U of M.SssI.
7. Add 128 pmol SAM (= 4 μL 32 mM solution) (*see* **Note 3**).
8. Add $MgCl_2$ to a final concentration of 10 mM.
9. Incubate at 30 °C for 7.5 min.
10. Stop reaction by adding an equal volume of Stop Buffer.

3.3 Nuclei Isolation for Human, Drosophila, and Other Cells Without Cell Walls

Start with 1×10^6 diploid human cells. Scale accordingly according to genome size, variations in cell ploidy, the aimed-for amount of sequencing (*see* **Note 9**).

1. Wash cells with 1× PBS.
2. Centrifuge for 5 min at 500 × *g* at 4 °C. Remove supernatant.
3. Resuspend cells in 200 μL ice-cold Nuclei Lysis Buffer.
4. Incubate on ice for 10 min.
5. Centrifuge for 5 min at 500 × *g* at 4 °C. Remove supernatant.
6. Resuspend nuclei in 200 μL cold Nuclei Wash Buffer.
7. Centrifuge for 5 min at 500 × *g* at 4 °C. Remove supernatant.
8. Resuspend nuclei in 200 μL CutSmart Reaction buffer.

3.4 Enzymatic Treatment of Chromatin for m6A-SMAC-Seq (Human Cells)

1. Add 200 U of EcoGII.
2. Add SAM at 0.6 mM and sucrose at 300 mM.
3. Incubate at 37 °C for 10 min.
4. Stop reaction by adding SDS to a concentration of 0.2%.

3.5 Enzymatic Treatment of Chromatin for m6A-GpC-CpGSMAC-Seq (Drosophila Cells)

1. Add 200 U of M.CviPI and 200 U of EcoGII.
2. Add SAM at a final concentration of 0.6 mM and sucrose at a final concentration of 300 mM.
3. Incubate at 30 °C for 7.5 min.
4. Add 128 pmol SAM (= 4 μL 32 mM solution) and another 100 U of both enzymes.
5. Incubate at 30 °C for 7.5 min.
6. Add 60 U of M.SssI.
7. Add 128 pmol SAM.
8. Add $MgCl_2$ at 10 mM.
9. Incubate at 30 °C for 7.5 min.
10. Stop reaction by adding SDS to a concentration of 0.2%.

3.6 HMW DNA Isolation

Here we describe HMW DNA using the Qiagen MagAttract HMW DNA Kit. Many other kits/protocols can also be used with similar success.

1. Add 20 μL Proteinase K into a 2 mL tube.
2. Add 200 μL of sample.
3. Add 4 μL RNase A solution and 150 μL Buffer AL. Mix by vortexing.
4. Incubate at room temperature for 30 min.
5. Add 15 μL MagAttract Suspension G beads.
6. Add 280 μL Buffer MB and incubate at room temperature for 3 min at 1400 rpm in a Thermomixer.
7. Separate the beads on a magnetic stand, carefully and completely remove the supernatant.
8. Add 700 μL Buffer MW1 and incubate at room temperature for 1 min at 1400 rpm in a Thermomixer.
9. Separate the beads on a magnetic stand, carefully and completely remove the supernatant.
10. Add 700 μL Buffer MW1 and incubate at room temperature for 1 min at 1400 rpm in a Thermomixer.
11. Separate the beads on a magnetic stand, carefully and completely remove the supernatant.
12. Add 700 μL Buffer PE and incubate at room temperature for 1 min at 1400 rpm in a Thermomixer.
13. Separate the beads on a magnetic stand, carefully and completely remove the supernatant.
14. Add 700 μL Buffer PE and incubate at room temperature for 1 min at 1400 rpm in a Thermomixer.

15. Separate the beads on a magnetic stand, carefully and completely remove the supernatant.
16. Add 700 µL nuclear-free H_2O by slowly pipetting on the side of the tube opposite to the beads while on the magnetic stand. Do not disturb the pellet, otherwise DNA loss can ensue.
17. Remove H_2O, and repeat the H_2O wash step.
18. Add an appropriate volume of Buffer AE, that is, 100–200 µL (*see* **Note 10**).
19. Incubate at room temperature for 3 min at 1400 rpm in a Thermomixer.
20. Separate the beads on a magnetic stand, carefully transfer the supernatant to a new DNA lo-bind tube using a wide bore tip.
21. Measure DNA concentration using a Qubit dsDNA HS assay.
22. Evaluate the DNA size distribution profile on the TapeStation using the gDNA screen tape and reagents.
23. Store the DNA at 4 °C (*see* **Note 11**).

3.7 DNA Size Selection

Selection of very HMW DNA using the Circulomics Short Read Eliminator Kit is described here. Use either the SRE or the SRE XL version depending on the properties of the genome studied and the input DNA size distribution. The SRE XL version will remove fragments $\leq$40 kbp while the SRE one will eliminate fragments $\leq$25 kbp.

1. Start with a total volume of 60 µL at DNA concentration between 50 and 150 ng/µL in a 1.5 mL DNA lo-bind tube.
2. Add 60 µL of Buffer SRE or Buffer SRE XL. Mix by tapping.
3. Centrifuge at 10,000 × *g* for 30 min at room temperature.
4. Carefully remove the supernatant without disturbing the DNA pellet (note that the pellet is not visible; always place the tube with the hinge facing outward to ensure reliable positioning of the pellet at the bottom of the tube).
5. Add 200 µL of 70% EtOH (make fresh immediately before use). Do not tap or mix. Centrifuge at 10,000 × *g* for 2 min at room temperature.
6. Carefully remove the supernatant without disturbing the DNA pellet.
7. Repeat the 70% EtOH wash and centrifugation step.
8. Add at least 50 µL Buffer EB and incubate at room temperature for 20 min (*see* **Note 10**).
9. Resuspend well by tapping.
10. Measure DNA concentration using a Qubit dsDNA HS assay.

11. Evaluate the DNA size distribution profile on the TapeStation using the gDNA screen tape and reagents.
12. Store the DNA at 4 °C (*see* **Note 11**).

Example TapeStation results for poor-quality, high-quality and post-size selection HMW DNA are shown in Fig. 3.

3.8 Nanopore Library Construction and Sequencing

Carry out nanopore library construction and sequencing according to the manufacturer's instructions depending on the particular kit and flow cell/sequencer being used.

3.9 Computational Analysis

The basic processing of SMAC-seq data described here consists of the following steps:

1. Initial base calling.
2. Read mapping.
3. Generating modification calls.
4. Compilation of basic data statistics.
5. Generation of aggregate modification scores.
6. Generation of averaged coverage tracks.

Analysis at the single molecule level can be subsequently carried out.

The overall workflow is summarized in Fig. 4.

3.9.1 Read Mapping and Modification Calling

There are two different ways to extract modifications. Historically, SMAC-seq per-read modification calls were extracted using Tombo, which is a non–model-based DNA modification caller for any context. It is no longer updated and requires base calling using the older and less accurate base calling software Albacore. The state-of-art way to call modifications is Megalodon. Megalodon is a command-line tool that combines base calling using Guppy with modified base calling based on pretrained modification calling models in the Rerio package, in which all-context m6A and 5mC models are available. Because of the higher accuracy of these models and the ease of use of Megalodon, this is at present the preferred method for calling modifications.

Calling modifications with Tombo involves the following steps (run these commands for each individual fast5 file in parallel to speed up the process):

1. Base calling using Albacore. Tombo requires that reads are first base-called using Albacore. Running Albacore requires the user to specify the exact type of flow cell and the kit used to build the library, as follows:

```
read_fast5_basecaller.py --flowcell {FLOW_CELL}
--kit {RUN_KIT} -i {FAST5_DIR}
-t {NUMBER_OF_THREADS} -s {OUTPUT_DIR}
-o fastq,fast5 --disable_filtering
```

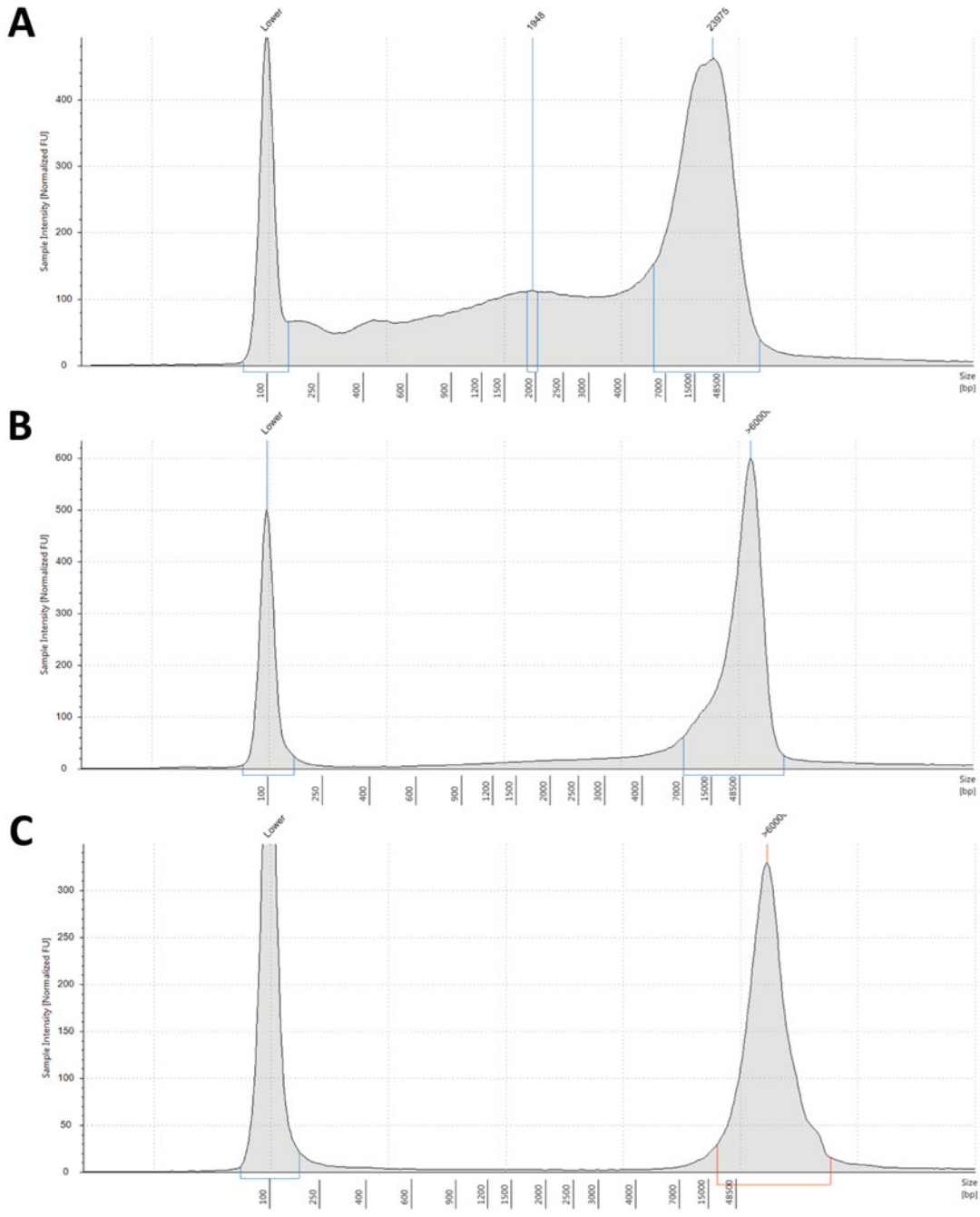

Fig. 3 HMW DNA isolation and size selection for long-read sequencing. It is of critical importance for the success of SMAC-seq experiments (and many other long read-based assays) to use high quality HMW DNA as input to sequencing. Numerous protocols exist for isolating HMW DNA and HMW DNA size selection. Shown are TapeStation gDNA profiles for a DNA sample with poor size distribution (**a**), a DNA sample with good size distribution (**b**), and a DNA sample after size selection using the Circulomics Short Read Eliminator Kit (**c**)

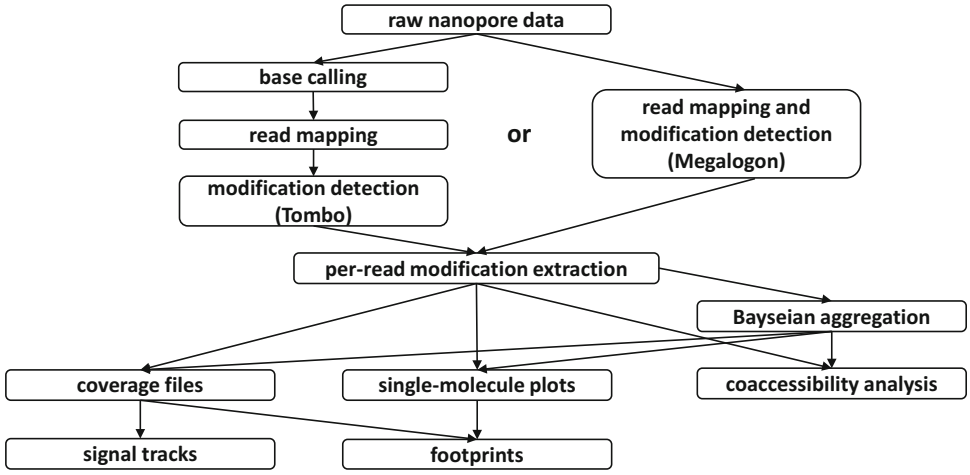

Fig. 4 Summary of the SMAC-seq analysis workflow. For Tombo processing, raw nanopore read traces are first subjected to base calling, mapped to the reference genome, and modified bases are then identified after "resquiggling" of the reads. The newer Megalodon-based processing combines these steps in one. Per-read modification calls are then extracted, and converted into a common file format that allows for downstream tasks to be carried out

2. Read preprocessing. Following base calling at the read level using Albacore, Tombo maps every read to its corresponding fast5 signal track, as follows:

```
tombo preprocess annotate_raw_with_fastqs
--processes {NUMBER_OF_THREADS} --overwrite
--fast5-basedir {FAST5_DIR}
--fastq-filenames {ALBACORE_PRODUCED_FAST5}
```

3. Tombo resquiggling. Next, the reads are mapped and nanopore signal is "resquiggled" against the reference genome as follows (note that Tombo uses minimap2 to carry out the mapping):

```
tombo resquiggle --ignore-read-locks
--processes {NUMBER_OF_THREADS} --overwrite
{FAST5_DIR} {REFERENCE_GENOME}
```

4. Tombo de novo modification calling. To call m6A and 5mC modifications in all contexts we use the de novo mode of Tombo as follows:

```
tombo detect_modifications de_novo
--statistics-file-basename {STATS_FILE_NAME}
--per-read-statistics-basename {MODS_FILE_NAME}
```

```
 --processes {NUMBER_OF_THREADS}
 --multiprocess-region-size 2000000
 --fast5-basedirs {FAST5_DIR}
```

3.9.2 Tombo Extraction

Default Tombo outputs do not include information about modification at the basepair single-molecule level. These need to be extracted using the Tombo Python API using custom-written scripts. Run the TomboSingleReadsExtract-tombo_de_novo-1.5.py script in order to convert Tombo per_read_stats files into text files. The script has multiple options for different sequence contexts, excluding certain sequence contexts, etc.:

```
 python TomboSingleReadsExtract-tombo_de_novo-1.5.py tombo.
per_read_stats genome.fa outfile_prefix
 [-m5C-only] [-m6A-only] [-CG-only] [-CG-CG-only]
 [-GC-only] [-m6A-CG-only] [-m6A-GC-only]
 [-m6A-GC-CG-only] [-doT] [-T-only]
 [-generic bases(comma-separated)]
 [-excludeContext string(...,stringN) radius]
 [-excludeChr chr1[...,chrN]]
 [-chrPrefix string]
```

Example for A positions:

```
python TomboSingleReadsExtract-tombo_de_novo.py
0.tombo.per_read_stats genome.fa 0.tombo.m6A-only
-m6A-only
```

Example for A, CpG and GpC positions:

```
python TomboSingleReadsExtract-tombo_de_novo.py
0.tombo.per_read_stats genome.fa 0.tombo.m6A-GC-CG-only
-m6A-GC-CG-only
```

Run the script for each individual `tombo.per_read_stats` file.

3.9.3 Read Mapping and Modification Calling Using Megalodon

Megalodon is run in one step as follows:

```
megalodon {LOCATION_OF_FAST5_FILES}
--guppy-params "-d {PATH_TO_RERIO_MODELS}"
 --guppy-config res_dna_r941_min_modbases-all-context_v001.
cfg
--outputs basecalls,mod_basecalls,per_read_mods
--reference {REFRENCE_GENOME}
--write-mods-text --output-directory {OUTPUT_DIR}
--guppy-server-path {LOCATION_OF_GUPPY_BIN}
```

For the purposes of downstream single-molecule analysis the `--outputs basecalls,mod_basecalls,per_read_mods` and `--write-mods-text` options need to be specified. These will result in output of per-read modifications in a text format.

3.9.4 Megalodon Per-Read Modification Extraction

To extract the per-read modification, we run the following script:

```
python megalodon-to-single_line.py *.per_read_modified_base_calls.txt
*megalodon.reads.tsv
```

Run the script for each individual Megalodon file.

3.9.5 Merging and Indexing

Merge the converted files into a single file, and sort by coordinates in the same step:

```
cat *.reads.tsv | sort -k1,1 -k2,2n -k3,3n
| bgzip > merged.reads.tsv.bgz
```

Then `tabix`-index the file:

```
tabix -s 1 -b 2 -e 3 merged.reads.tsv.bgz
```

This will create a `tabix`-index —.bgz— file in the following format, with one entry for each read:

1. Column 1: chromosome.
2. Column 2: left-most modified/informative position within the read.
3. Column 3: right-most modified/informative position within the read.
4. Column 4: . character (for legacy reasons).
5. Column 5: nanopore read ID.
6. Column 6: nan (for legacy reasons).
7. Column 7: comma-separated list of modified/informative positions.
8. Column 8: comma-separated list of Tombo probabilities, matching the order of the positions in Column 7.

3.9.6 Calculate Mapping Statistics

Calculate read mapping statistics as follows:

```
python NanoporeTSVMappingStats.py
merged.reads.tsv.bgz
NanoporeTSVMappingStats-merged
```

This will produce a short report with the total number of mapped reads, the total number of mapped bases, the mean mapped read length and the median read length.

3.9.7 Create Coverage File

While the true strength of SMAC-seq lies in the single-molecule analysis, SMAC-seq data can also be highly informative at an aggregate level, which allows for CREs and positioned nucleosomes to be discerned by visualization of average SMAC-seq profiles on a genome browser. For the purpose of such analyses, a coverage file in the style of the output from the popular bisulfite sequencing analysis tool Bismark [34] is created, using the `methylation-reads-tsv-to_coverage.py` script:

```
python methylation_reads_all.tsv threshold outfile
[-stranded +|-] [-minAbsLogLike float]
[-minAbsPValue float]
[-BayesianIntegration window(bp) step alpha beta pseudosam-
plesize] [-N6mAweight pseudosamplesize genome.fa]
[-saveNewSingleMoleculeFile filename]
```

Nanopore DNA modification data is not binary, instead it is recorded as probabilities. It thus has to be binarized at some threshold. We have found, through exploration of the parameter space and comparison to known biological truths, that the most intuitive threshold of 0.5 works optimally [22]. Example:

```
python methylation-reads-tsv-to_coverage.py
merged.reads.tsv.bgz 0.5 merged.cutoff_0.5.coverage
```

Convert the resulting plain text file to a `.bgz` file:

```
cat merged.cutoff_0.5.coverage |
bgzip > merged.cutoff_0.5.coverage.bgz
```

Then `tabix`-index it:

```
tabix -s 1 -b 2 -e 3 merged.cutoff_0.5.coverage.bgz
```

The format of the `coverage` file is as follows:

1. Column 1: chromosome.
2. Column 2: left-most position of the modified/informative sequence context.
3. Column 3: right-most position of the modified/informative sequence context.
4. Column 4: number reads in which the sequence context is methylated.

5. Column 5: number reads in which the sequence context is unmethylated.
6. Column 6: total number of reads.

3.9.8 Bayesian Integration

Even when using m6A, SMAC-seq still does not cover every single nucleotide in the genome, and coverage varies substantially between different locations depending on local sequence content differences. In addition, base calling for ONT data is still far from perfectly accurate (*see* **Note 13**), and detecting modifications is particularly challenging. On the other hand, the biologically meaningful length scale for DNA accessibility is not necessary the individual basepair, but somewhat larger sequence contexts.

For these reasons we often use aggregate accessibility scores over fixed-length windows, which combine information over all available informative positions in the window, thus providing more reliable, even if coarser-grained, views of accessibility patterns. This is done using a simple Bayesian procedure, as follows.

For a given window of width w, specified by coordinates $c,i,i+w$ (where c is the chromosome, and i is the leftmost coordinate of the window), and for all reads $r \in R_{c,i,i+w}$ fully spanning the window, we obtain all Tombo probabilities $p_{r,(c,j)}$ such that $j \in [i, i+w)$ for the assayed sequence contexts on the corresponding genomic strand (*see* **Note 14**). We usually use a Beta prior $B(\alpha,\beta)$, with $\alpha = \beta = 10$, which is updated based on each probability $p_{r,(c,j)}$ for all $j \in [i,i+w)$ (but the prior can be easily changed if necessary, *see* below), in order to obtain a final accessibility score $p_{r,(c,i,i+w)}$ for read r and window $c,i,i+w$.

This Bayesian integration calculation is also carried out using the same `methylation-reads-tsv-to_coverage.py` script. For efficiency of calculation, compute it in parallel on the individual converted tombo files, as follows (for a 10-bp context and (10,10) prior):

```
python methylation-reads-tsv-to_coverage.py
0.tombo 0.5
0.tombo.all0.cutoff_0.5.coverage.BI_w10_a10_b10
-minAbsPValue 0.4 -BayesianIntegration 10 1 10 10 50
-saveNewSingleMoleculeFile
0.tombo.BI_w10_a10_b10.reads.tsv
```

Merge the Bayesian integration files:

```
cat *tombo.BI_w10_a10_b10.reads.tsv
| sort -k1,1 -k2,2n -k3,3n
| bgzip > merged.BI_w10_a10_b10.reads.tsv.bgz
```

Then `tabix`-index the resulting .bgz file:

```
tabix -s 1 -b 2 -e 3 merged.BI_w10_a10_b10.reads.tsv.bgz
```

3.9.9 Filtering Fully Methylated Reads

On occasions, we observe a population of reads that appear as fully methylated across their whole length or over large segments of it. They are most likely derived from dead cells or represent some other undesired artefact. In order to remove such potentially artefactual reads, we obtain a "filtered" read set by removing all reads containing a $\geq$1-kbp stretch that is $\geq$75% methylated (while also filtering out reads shorter than 1 kb).

This operation can be carried out using the `filterFullyMethylatedReads.py` script as follows:

```
 python filterFullyMethylatedReads.py methylation_reads_all.
tsv WindowSize minFraction
 [-keepShort] [-missingBasesFilter genome.fa basecontexts(com-
ma-separated) minFraction
 [-doMBFSet]]
```

3.9.10 Create Genome Browser Tracks

In order to create average-methylation (and thus accessibility) tracks that can be visualized on a genome browser such the UCSC or the WashU ones, use the following script:

```
python coverage_to_wig.py coverage.bgz window step chrField
MfieldID UfieldID chrom.sizes outprefix [-minCov N_reads]
```

where the M and the U fields indicate the column IDs of the numbers of methylated and unmethylated reads, respectively, and the window and step parameters specify the width and the stride used for averaging the signal (i.e., window of 50 and step of 5 means that the average methylation level over 50 bp windows tiling the genome every 5 bp will be outputted).

This script will output two `bedGraph` files—a `coverage.wig` one (which contains the number of reads covering a position) and a `meth.wig` one (which contains the fraction of methylated reads). These can then be converted into `bigWig` files that can in turn be displayed on a genome browser using the `wigToBigWig` program from the UCSC utilities:

```
wigToBigWig meth.wig chrom.sizes meth.wig
```

where the chrom.sizes files contains one line per chromosomes including the chromosome name and its length in bp (tab-separated).

An example of an average SMAC-seq profile is shown in Fig. 5.

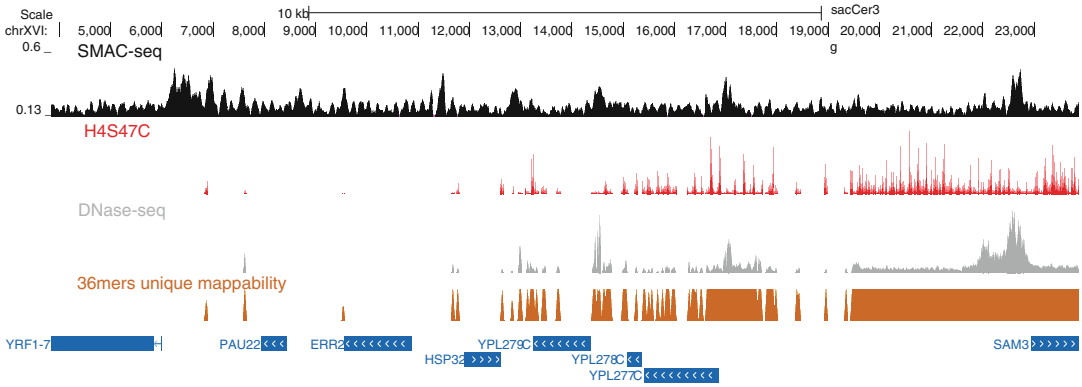

Fig. 5 Examples of average m6A-CpG-GpC-SMAC-seq profiles visualized on the UCSC Genome Browser. Shown is a subtelomeric regions on chrXVI. SMAC-seq signal provides information both about accessible open chromatin measures (peaks in DNAse-seq data) and positioned nucleosomes. The latter are shown here in the form of H4S47C chemical nucleosome mapping [35], which maps the positions of dyads (SMAC-seq signal is enriched on nucleosome linkers, thus the inverse relationship between the two). SMAC-seq, being a long-read assay, also provides information about repetitive regions of the genome (in this case, telomeres, which are not uniquely mappable with short reads as shown by the 36-mer unique mappability track)

3.9.11 Making Metaplots Around a Position

A common analysis task is to generate a metaplot around a given set of genomic features (such as TSSs, positioned nucleosomes, TF binding motifs, and others). The `coverage.bgz` can be used to make such metaplots, as follows, with a variety of parameters (window size, minimal coverage per position, different input file formats, stranded or unstranded, and others):

```
python signalAroundPeaks-nano.py inputfilename chrFieldID
posField strandField radius window coverage.bgz outputfile-
name [-bismark.cov] [-bed] [-minCov N]
[-unstranded] [-narrowPeak]
```

Examples of such plots around yeast transcription starts sites and human occupied CTCF motifs are shown in Fig. 6.

3.9.12 Making Single Molecule Plots

One of the two key strengths of SMAC-seq is the ability to analyze accessibility at the single molecule level. There are many ways to do that, due to the nonbinary nature of raw nanopore data and of the long length of nanopore reads, which allows for/requires analysis at different resolution levels. Single molecule maps can be generated using the continuous modification probability values or they can be binarized.

The `SMAC-footprints-from-methylation-reads-tsv-tabix.py` and `SMAC-footprints-from-methylation-reads-tsv-tabix-kmeans.py` scripts can be used to generate such plots. The first script will apply hierarchical clustering while the second one will use *k*-means (in our experience, we obtain

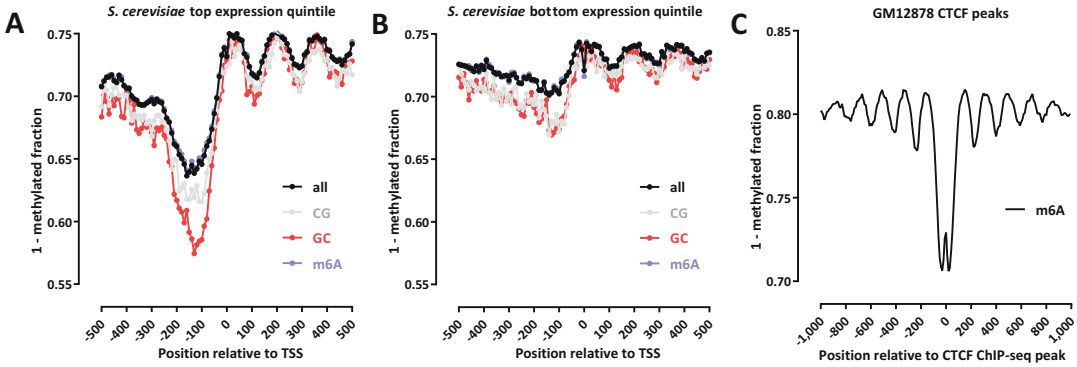

Fig. 6 Examples of average SMAC-seq metaprofiles over predefined genomic features. (**a** and **b**) Average m6-CpG-GpC SMAC-seq profiles for the top 20% and bottom 20% of genes (ranked by expression levels) in *S. cerevisiae*. Profiles are split by modification channel. (**c**) Average m6-SMAC-seq profile around CTCF ChIP-seq peaks in the human GM12878 cell line. CTCF is known to strongly position nucleosomes in the vicinity of its occupancy sites [36]. ChIP-seq peaks were obtained from the ENCODE Project Consortium [29]

decidedly better results using the *k*-means approach). The commands are otherwise the same. There is a wide variety of options regarding the input list of region (which can be in any format), the display (averaging over arbitrary number of basepairs), subsampling of reads, color schemes, binarization or continuous display, and others:

```
python methylation_reads_all.tsv peak_list chrFieldID
leftFieldID rightFieldID strandFieldID tabix_path outfile_-
prefix [-resize factor] [-subset N] [-label fieldID] [-minCov
fraction]
[-minPassingBases fraction] [-minReads N]
[-unstranded] [-minAbsLogLike float]
[-scatterPlot colorscheme minScore maxScore color|none]
[-window bp] [-readStrand +|-]
[-printMatrix] [-deleteMatrix] [-binarize threshold]'
```

The following command will generate binarized single molecule maps retaining only reads that completely span the input set of regions, averaging over 10 bp windows:

```
python SMAC-footprints-from-methylation-reads-tsv-tabix-
kmeans.py
SMAC-seq.reads.tsv.bgz regions.bed 0 1 2 3 tabix
SMAC-seq.regions.binary-0.5-gist_heat.10bp
-window 10 -minCov 1 -binarize 0.5
-scatterPlot gist_heat 0 1.1 w -unstranded
```

An example of such a single-molecule level visualization for yeast m6ACpG-GpC-SMAC-seq data is shown in Fig. 7a.

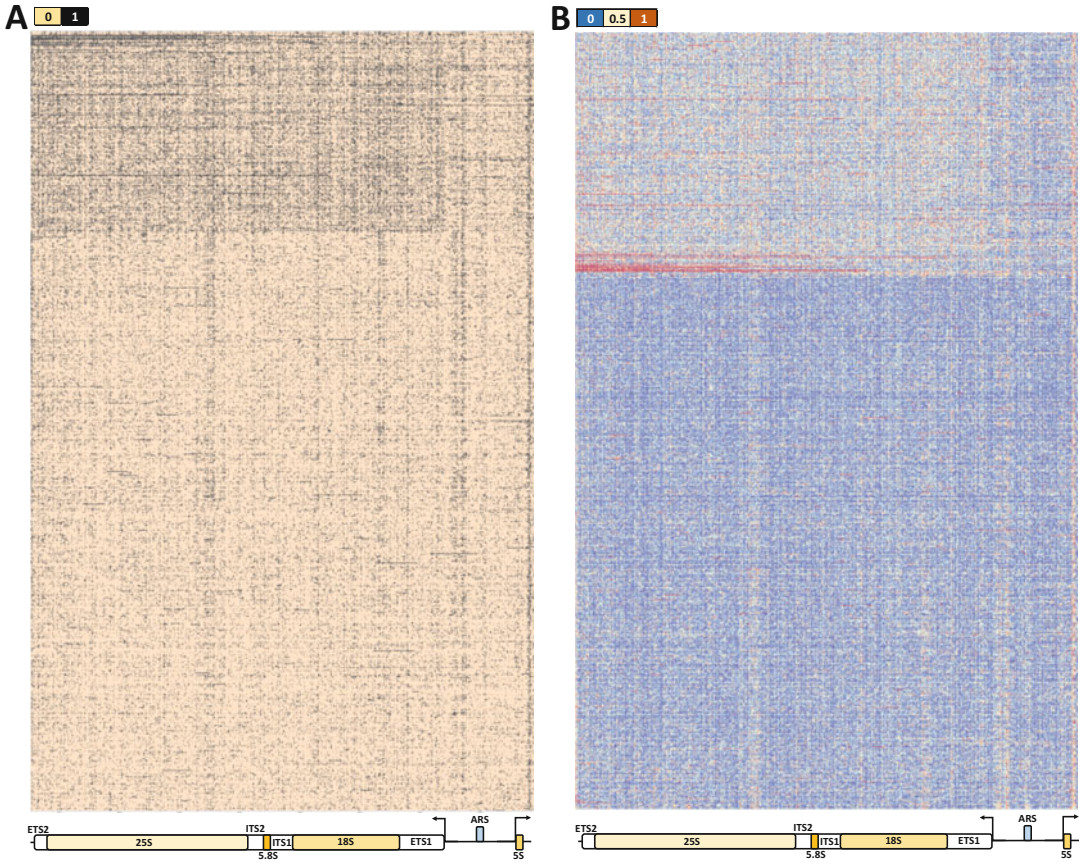

Fig. 7 Examples of single molecule m6A-CpG-GpC-SMAC-seq maps in *S. cerevisiae*. Shown is the yeast rDNA locus, binarized (**a**) and as a continuous display (**b**). Yeast rDNA is organized into multicopy (~150) arrays, consisting of ~9.1 kb units, each containing a copy of the 35S precursor pre-rRNA, transcribed by Pol I, a 5S RNA, transcribed by Pol III, and a replication origin ARS element, located in nontranscribed (NTS) regions of the array. The rDNA chromatin structure adopts two distinct conformations [37, 38]—an inactive nucleosomal state and an extremely highly transcriptionally active, largely devoid of nucleosomes (and thus highly accessible) state. Note that 1000 reads were sampled at random for each plot, and that different samplings are shown in (**a**) and (**b**)

The following command will generate continuous-signal single molecule maps retaining only reads that completely span the input set of regions, averaging over 10 bp windows:

```
 python SMAC-footprints-from-methylation-reads-tsv-tabix-
kmeans.py
 SMAC-seq.reads.tsv.bgz regions.bed 0 1 2 3 tabix
 SMAC-seq.regions.binary-0.5-RdYlBu.10bp
 -window 10 -minCov 1
 -scatterPlot RdYlBu 0 1 w -unstranded
```

An example of such a single-molecule level visualization for yeast m6ACpG-GpC-SMAC-seq data is shown in Fig. 7b.

3.9.13 Calculating NMI Matrices

Finally, another common analysis task when working with SMAC-seq data is to estimate the degree of single-molecule coaccessibility along the chromatin fiber.

To this end, we apply a Normalized Mutual Information as follows. Each chromosome c is split into windows of size w. For each such window $(c,i,i+w)$, the maximum range to the right of it, $(c,j,j+w)$ such that the span $(c,i,j+w)$ is covered by $\geq M$ reads, is identified. All reads spanning $(c,j,j+w)$ are then extracted and subsampled down to M reads (usually $M = 100$). Accessibility scores are then aggregated and binarized for all windows located in the span $(c,j,j+w)$, and for all M reads fully spanning it, resulting in a local coaccessibility matrix LCM of size $M \times (j+w-i)/w$. A Normalized Mutual Information (NMI) score for each pair of columns LCM_k and LCM_l is then calculated as follows:

$$\begin{aligned}\mathrm{MI}(\mathrm{LCM}_k,\mathrm{LCM}_l) = {} & p(0,0)\log_2\left(\frac{p(0,0)}{p_k(0)\,p_l(0)}\right)\\ & +p(1,1)\log_2\left(\frac{p(1,1)}{p_k(1)\,p_l(1)}\right)\\ & +p(0,1)\log_2\left(\frac{p(0,1)}{p_k(0)\,p_l(1)}\right)\\ & +p(1,0)\log_2\left(\frac{p(1,0)}{p_k(1)\,p_l(0)}\right)\end{aligned} \tag{1}$$

While, in principle, mutual information cannot be negative, NMI scores are normalized and rescaled in the interval $(-1,1)$ so that anticorrelated regions are given negative scores (this is done for visualization and interpretation purposes):

$$\mathrm{NMI}(\mathrm{LCM}_k,\mathrm{LCM}_l) = \begin{cases} \dfrac{\mathrm{MI}(\mathrm{LCM}_k,\mathrm{LCM}_l)}{\sqrt{H(\mathrm{LCM}_k)H(\mathrm{LCM}_l)}} & \text{for } p(0,0)+p(1,1) \geq 0.5 \\ -\dfrac{\mathrm{MI}(\mathrm{LCM}_k,\mathrm{LCM}_l)}{\sqrt{H(\mathrm{LCM}_k)H(\mathrm{LCM}_l)}} & \text{for } p(0,0)+p(1,1) < 0.5 \end{cases} \tag{2}$$

where H refers to the entropy of each individual distribution.

To calculate NMI matrices, the `SingleMoleculeCorrelation-NMI-matrix.py` script can be used:

```
python SingleMoleculeCorrelation-NMI-matrix.py
 SMAC-seq.reads.tsv.bgz regions.bed chrFieldID leftField
rightFieldID minCoverage windowsize stepsize tabix_location
outfileprefix
 [-subsample N] [-expectedMaxDist bp] [-label fieldID]
```

Example:

```
python SingleMoleculeCorrelation-NMI-matrix.py
SMAC-seq.reads.tsv.bgz regions.bed 0 1 2 50 1 1200 tabix
NMI.min50cov.1bp.regions.SMAC-seq -expectedMaxDist 1500
```

If running genome-wide, split the genome into overlapping bins for parallelization efficiency, for example, 50 kb in size with a 10 kb stride, and calculate a separate matrix for each, then take the average NMI values for each pair of coordinates for downstream analyses.

An example of the results of NMI analysis for yeast m6A-CpG-GpCSMAC-seq data is shown in Fig. 8.

4 Notes

1. EcoGII deposits m6A modifications without a sequence preference, but it does not do so with perfect efficiency. It is not conclusively established why, that is, whether the presence of neighboring already modified bases prevents further methylation or whether perhaps the enzyme is highly nonprocessive and stays bound to DNA for a prolonged period after completion of the reaction, thus occluding neighboring bases from further enzymatic action. The methylation efficiency reported by NEB is ~50%; however, this is based on a relatively short treatment (5 min). On the other hand, based on the original more detailed study describing EcoGII [28], methylation efficiency seems to be closer to 80% for a prolonged treatment of about an hour. The incubation times during a SMAC-seq experiment would place the expected efficiency somewhere in between these values. Unfortunately, the most straightforward imaginable experiment that would properly establish EcoGII methylation efficiencies in the context of a nanopore-based experiment—nanopore sequencing of naked DNA subjected to EcoGII treatment—is not in fact possible because of the strong bias of the Oxford Nanopore platform against fully methylated templates, which simply do not sequence well and are mostly discarded.
2. EcoGII is commercially available as a solution at relatively low concentration, and if sufficiently many units of it are to be used, the volume needed becomes too large and could interfere with the labeling reaction. For these reasons, we are using a custom-made highly concentrated EcoGII from NEB.
3. SAM is unstable. This is one of the reasons why it is added twice to the reaction, and it is also why it should be handled carefully, avoiding repeated freeze-thaw cycles.

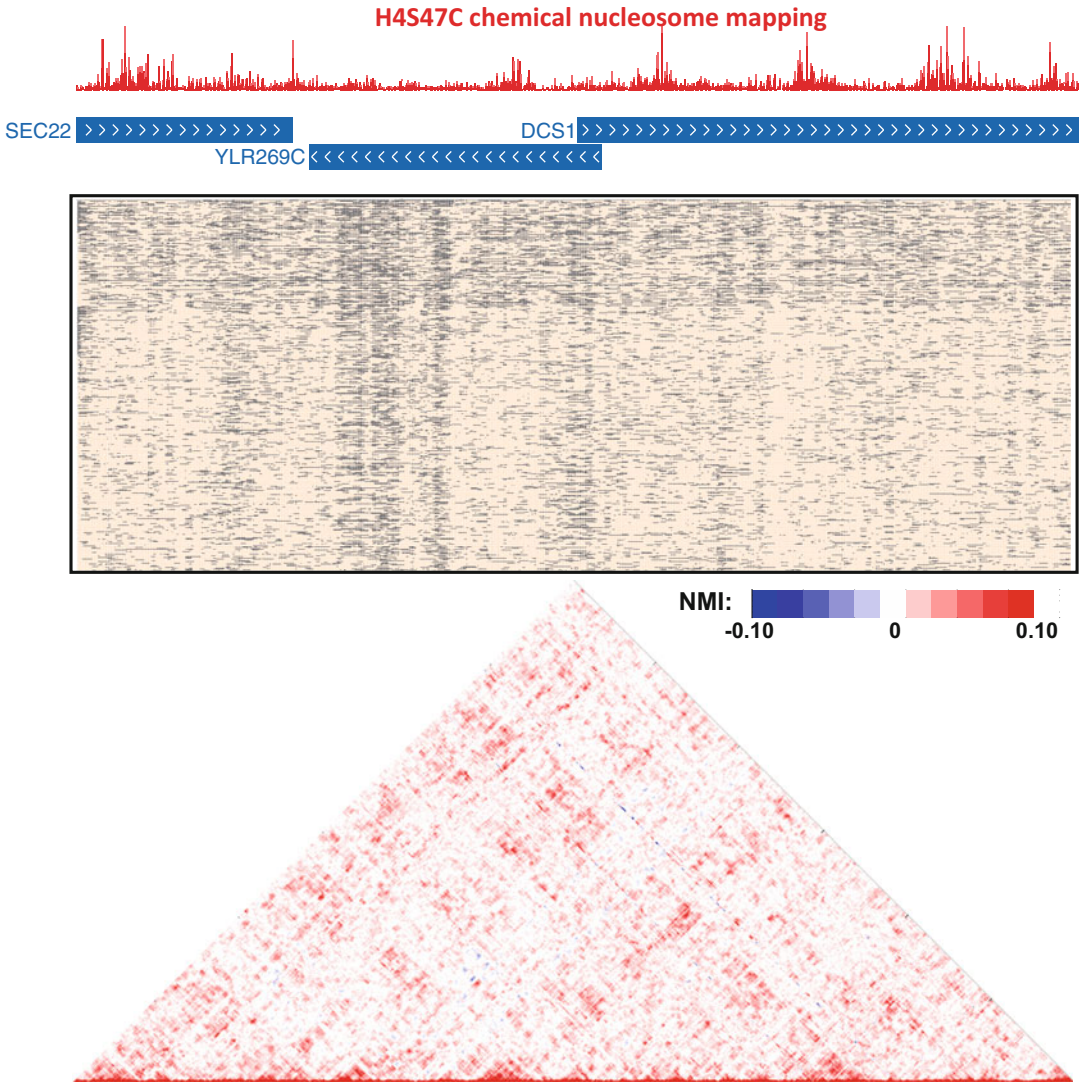

Fig. 8 Example of m6A-CpG-GpC-SMAC-seq coaccessibility maps (measured in terms of NMI) in *S. cerevisiae*. Shown is the promoter region of the DCS1 gene, together with chemical mapping nucleosome positioning data (top), the single-molecule SMAC-seq map (middle), and the coaccessibility map (bottom)

4. ONT offers multiple sequencing platforms, and it is advisable to be familiar with the properties of each and the tradeoffs between them. The MinION is the main ONT sequencing platform, typically generating ~10 Gbp data (although higher throughput runs have also been observed in practice, up to 20 Gbp) and several million reads. In the context of SMAC-seq, a single MinION is often sufficient to generate adequate coverage over a small genome such as that of yeasts.

The Flongle is a miniaturized flow cell that also runs on the MinION instrument, typically generating ~100,000 reads. It is not sufficient for production-scale runs, but as it is priced at ~1/9 of the cost of a MinION flow cell, it is ideal for testing protocol, carrying out QC runs, etc.

The GridION can use either MinION or Flongle flow cells and run five of them in parallel.

The PromethION is the high-throughput ONT sequencer. It uses different flow cells, each of which can generate up to ~100 Gbp of data (and more than ten million reads), and can run 48 such flow cells in parallel at the same time. Each such flow cell is priced at more than twice the cost of a MinION flow cell. To study larger and more complex eukaryotic genomes using SMAC-seq, the throughput of the PromethION becomes necessary, and often multiple such flow cells are needed.

5. It is important to note that "coverage" means very different things in the contexts of genome sequencing and SMAC-seq. Usually, "coverage" refers to how many reads cover a given position in the genome on average. However, the more relevant metric for SMACseq is instead "coverage at length L," that is, how many reads cover two position spread apart at a given instance. One of the main goals of SMAC-seq is to capture the coordinated behavior of distal CREs and this is only possible when sufficiently many single molecules containing both CREs have been sequenced. Across eukaryotes a clear trend is observed—as genome size increases, CREs become spread apart more and more. Thus, while 2 yeast promoters are on average 1.5–2 kb apart, the distance between an enhancer and its cognate promoter in a mammalian genome is often tens of kilobases. Thus, the required sequencing throughput to achieve the same effective "coverage at length *L*" does not scale linearly once the fact that even with careful size selection there are still many more shorter nanopore reads than very long ones is taken into account.
6. Low-binding tubes are preferable in order to minimize DNA loss.
7. The sheer volume of nanopore sequencing data presents a different level of challenge in terms of computational infrastructure compared to short-read sequencing. A single PromethION flow cell can produce 100 Gbp of data within 48 h, and a PromethION instrument can in principle run 48 such flow cells in parallel.

   However, base calls are far from the only information that needs to be stored. For analysis of SMAC-seq datasets (and of DNA modifications in general), the nanopore current signal itself is what is most important, as it is used during the

resquiggling and DNA modification detection steps. Thus, the actual disk space footprint of such a flow cell is between one and two orders of magnitude higher than storing the base calls alone.

In addition, a separate challenge has historically been posed by the number of files. This has changed with more recent versions of the ONT processing software, but historically ONT data has been stored in a large number of individual small files, which could be so large that it reached the limit on the number of files per use that many shared computational clusters have in place, necessitating sequential processing of datasets in batches and cleaning of files in between each.

8. The files containing single-molecule SMAC-seq information can be huge in size, surpassing 1 TB on occasions. Random access is critical for downstream analysis to be practical. The workflows described here achieve this by using `tabix` indexing of coordinate-sorted files.
9. Nanopore sequencing involves no amplification of DNA while having strict constraints on the minimum amount of DNA that is to be used as input to each sequencing run. A typical PromethION run uses at least 1 μg of DNA, but if size selection is to be applied prior to it, this corresponds to several times more input DNA per run. A typical diploid human cell contains ~6 pg of DNA, thus 1×10^6 cells contain ~6 μg of DNA. Multiple PromethION runs are required to obtain good coverage for a mammalian-sized genome, thus tens of micrograms of DNA are needed as input to size selection and then sequencing. Scale up reactions accordingly based on the specifics of the experiment with these considerations in mind.
10. Elution volumes are important for nanopore sequencing. All ONT sequencing kits have a minimum requirement for the amount of input DNA but also a maximum limit to the volume in which it is contained. Concentrating DNA using beads will result in significant losses while doing so by evaporation leads to its degradation. Thus, it is best to have a large amount of DNA in a small volume. However, there is a trade-off between the elution volume and the efficiency of elution—larger elution volumes lead to better overall yields. Thus, the optimal elution volume is to be decided based on the number of cells used for the SMAC-seq reaction and the exact ONT kits that are to be used for sequencing.
11. HMW DNA is stable for a long time at 4 °C, but it is strongly recommended not to freeze it at −20 °C or −80 °C as this will likely result in fragmentation. Also, highly concentrated HWM DNA can sometimes precipitate out of solution after prolonged storage so make sure to inspect tubes before use.

Resuspend by tapping the tubes with your fingers, do not pipette up and down as this is also thought to lead to HMW DNA degradation. In addition, always transfer HMW DNA using wide bore tips to prevent shearing.

12. Yeast (and fungal cells in general) have thick cell walls comprised of polysaccharides, lipids and chitin in various proportions. They present a barrier to the access of most enzymes to the nucleus, thus protocols tailored to such cells involve treatment with zymolyase or chitinase enzymes [39], with the exact details varying depending on the species studied.
13. Nanopore sequencing is a powerful tool for detecting DNA modifications, but discerning modified bases from raw nanopore signal is not yet a fully resolved problem, especially for methylation modifications, which do not provide a huge shift in current signal relative to the unmodified base. Detection of m6A is more challenging that detection of m5C, possibly because a single methyl group changes the overall properties of a purine base to a lesser extent than it does for a pyrimidine. In addition, it should be noted that current implementations of nanopore sequencing do not actually read out a single bases at a time. Instead, they read several bases at a time and the problem of base calling and modification detection is solved not in the small space of bases but in the much larger space of k-mers of size 5 or 6. Base calling errors are therefore at present an unavoidable part of the reality of dealing with nanopore datasets.

    In our experience, the error rate for calling m6A at the level of a single base within a single molecule in the context of SMAC-seq is in the 20–25% range, while that for m5C is somewhere around 15%. However, we expect the performance to improve significantly in the future through a combination of computational and experimental approaches.
14. Unlike CpG and GpC sequence contexts, which are symmetric, and therefore bases that are to be modified are present at the same position on both strand, m6A provides different information on the forward and reverse strand, as it is not a symmetric sequence context. This is a partial limitation of m6A-SMAC-seq, because different profiles can be generated from the two strands in some situations.

Acknowledgments

The authors thank members of the Greenleaf and Kundaje labs for many helpful discussions. This work was supported by NIH grants UM1HG009436 and P50HG007735 (to W.J.G.). W.J.G. is a Chan Zuckerberg investigator. Z.S. is supported by EMBO

Long-Term Fellowship EMBO ALTF 1119-2016 and by Human Frontier Science Program LongTerm Fellowship HFSP LT 000835/2017-L. G.K.M. was supported by the Stanford School of Medicine Dean's Fellowship.

References

1. Wu C (1980) The 50 ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature 286(5776):854–860
2. Keene MA, Corces V, Lowenhaupt K et al (1981) DNase I hypersensitive sites in Drosophila chromatin occur at the 50 ends of regions of transcription. Proc Natl Acad Sci U S A 78:143–146
3. McGhee JD, Wood WI, Dolan M et al (1981) A 200 base pair region at the 50 end of the chicken adult β-globin gene is accessible to nuclease digestion. Cell 27:45–55
4. Dorschner MO, Hawrylycz M, Humbert R et al (2004) High-throughput localization of functional elements by quantitative chromatin profiling. Nat Methods 1:219–225
5. Sabo PJ, Humbert R, Hawrylycz M et al (2004) Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. Proc Natl Acad Sci U S A 101:4537–4542
6. Sabo PJ, Kuehn MS, Thurman R et al (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods 3:511–518
7. 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
8. Boyle AP, Davis S, Shulha HP et al (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132(2):311–322
9. Thurman RE, Rynes E, Humbert R et al (2012) The accessible chromatin landscape of the human genome. Nature 489(7414):75–82
10. 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
11. Buenrostro JD, Wu B, Litzenburger UM et al (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523(7561):486–490
12. Cusanovich DA, Daza R, Adey A et al (2015) Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 348(6237):910–914
13. Chereji RV, Eriksson PR, Ocampo J, Clark DJ (2019) DNA accessibility is not the primary determinant of chromatin-mediated gene regulation. bioRxiv:639971
14. Ponnaluri VKC, Zhang G, Estève PO et al (2017) NicE-seq: high resolution open chromatin profiling. Genome Biol 18(1):122
15. Umeyama T, Ito T (2017) DMS-seq for in vivo genome-wide mapping of protein-DNA interactions and nucleosome centers. Cell Rep 21(1):289–300
16. Timms RT, Tchasovnikarova IA, Lehner PJ (2019) Differential viral accessibility (DIVA) identifies alterations in chromatin architecture through large-scale mapping of lentiviral integration sites. Nat Protoc 14(1):153–170
17. 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(12):2497–2506
18. 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(3): 411–422.e4
19. Vaisvila R, Ponnaluri VKC, Sun Z et al (2019) EM-seq: detection of DNA methylation at single base resolution from picograms of DNA. bioRxiv:2019.12.20.884692
20. Simpson JT, Workman RE, Zuzarte PC et al (2017) Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods 14: 407–410
21. Rand AC, Jain M, Eizenga JM et al (2017) Mapping DNA methylation with highthroughput nanopore sequencing. Nat Methods 14: 411–413
22. Shipony Z, Marinov GK, Swaffer MP et al (2020) Long-range single-molecule mapping of chromatin accessibility in eukaryotes. Nat Methods 17:319–327
23. Wang Y, Wang A, Liu Z et al (2019) Singlemolecule long-read sequencing reveals the chromatin basis of gene expression. Genome Res 29:1329–1342

24. Aughey GN, Estacio Gomez A, Thomson J et al (2018) CATaDa reveals global remodelling of chromatin accessibility during stem cell differentiation in vivo. eLife 7:e32341
25. Schones DE, Cui K, Cuddapah S et al (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132:887–898
26. Hesselberth JR, Chen X, Zhang Z et al (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods 6(4):283–289
27. Neph S, Vierstra J, Stergachis AB et al (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489:83–90
28. Murray IA, Morgan RD, Luyten Y et al (2018) The non-specific adenine DNA methyltransferase M.EcoGII. Nucleic Acids Res 46:840–848
29. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74
30. Kuhn RM, Haussler D, Kent WJ (2013) The UCSC genome browser and associated tools. Brief Bioinform 14:144–161
31. Kent WJ, Zweig AS, Barber G et al (2010) BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 26: 2204–2207
32. Li H (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32(14):2103–2110
33. Stoiber MH, Quick J, Egan R, Lee JE, Celniker SE, Neely R, Loman N, Pennacchio L, Brown JB (2017) De novo identification of DNA modifications enabled by genomeguided nanopore signal processing. bioRxiv:094672
34. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27(11): 1571–1572
35. Brogaard K, Xi L, Wang JP, Widom J (2012) A map of nucleosome positions in yeast at basepair resolution. Nature 486(7404):496–501
36. 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
37. Conconi A, Widmer RM, Koller T, Sogo JM (1989) Two different chromatin structures coexist in ribosomal RNA genes throughout the cell cycle. Cell 57(5):753–761
38. Goetze H, Wittner M, Hamperl S, Hondele M, Merz K, Stoeckl U, Griesenbeck J (2010) Alternative chromatin structures of the 35S rRNA genes in Saccharomyces cerevisiae provide a molecular basis for the selective recruitment of RNA polymerases I and II. Mol Cell Biol 30(8):2028–2045
39. Schep AN, Buenrostro JD, Denny SK et al (2015) Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome Res 25:1757–1770