



Single-molecule long-read sequencing reveals the chromatin basis of gene expression

Yunhao Wang, Anqi Wang, Zujun Liu, et al.

Genome Res. published online June 14, 2019

Access the most recent version at doi:[10.1101/gr.251116.119](https://doi.org/10.1101/gr.251116.119)

P<P	Published online June 14, 2019 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .

Comprehensive immune receptor profiling.
Discover the **DriverMap™ AIR Assay** difference.



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

1 **Single-molecule long-read sequencing reveals the chromatin**
2 **basis of gene expression**

3 Yunhao Wang^{a,b,1}, Anqi Wang^{a,b,1}, Zujun Liu^{a,b}, Andrew L. Thurman^b, Linda S. Powers^b,
4 Meng Zou^b, Yue Zhao^{a,b}, Adam Hefel^b, Yunyi Li^b, Joseph Zabner^b, Kin Fai Au^{a,b,c,2}

5 ^a Department of Biomedical Informatics, The Ohio State University, OH 43210, USA

6 ^b Department of Internal Medicine, University of Iowa, Iowa City, IA 52242, USA

7 ^c Department of Biostatistics, University of Iowa, Iowa City, IA 52242, USA

8 ¹ Y. W. and A. W. contributed equally to this work.

9 ² To whom correspondence should be addressed. E-mail: kinfai.au@osumc.edu

10

11 **Key words**

12 Single-molecule long-read sequencing, nucleosome positioning, chromatin
13 accessibility, Oxford Nanopore Technologies

1 ABSTRACT

2 Genome-wide chromatin accessibility and nucleosome occupancy profiles have been
 3 widely investigated, while the long-range dynamics remains poorly studied at the
 4 single-cell level. Here we present a new experimental approach MeSMLR-seq
 5 (methyltransferase treatment followed by single-molecule long-read sequencing) for
 6 long-range mapping of nucleosomes and chromatin accessibility at single DNA
 7 molecules, and thus achieve comprehensive-coverage characterization of the
 8 corresponding heterogeneity. MeSMLR-seq offers direct measurements of both
 9 nucleosome-occupied and nucleosome-evicted regions on a single DNA molecule,
 10 which is challenging for many existing methods. We applied MeSMLR-seq to haploid
 11 yeast, where single DNA molecules represent single cells, and thus we could
 12 investigate the combinatorics of many (up to 356) nucleosomes at long range in single
 13 cells. We illustrated the differential organization principles of nucleosomes
 14 surrounding the transcription start site for silently- and actively-transcribed genes, at
 15 the single-cell level and in the long-range scale. The heterogeneous patterns of
 16 chromatin statuses spanning multiple genes were phased. Together with single-cell
 17 RNA-seq data, we quantitatively revealed how chromatin accessibility correlated with
 18 gene transcription positively in a highly-heterogeneous scenario. Moreover, we
 19 quantified the openness of promoters and investigated the coupled chromatin
 20 changes of adjacent genes at single DNA molecules during transcription
 21 reprogramming. Especially, we revealed the coupled changes of chromatin
 22 accessibility for two neighboring glucose transporter genes in response to the change
 23 of glucose concentration.

24 INTRODUCTION

25 In eukaryotic organisms, cells are faced with genetic information storage and
 26 packaging problems. As the carrier of genetic information, instead of folding into a
 27 disorganized yarn ball, DNA strands wrap around thousands of protein cores like
 28 "beads on a string". As the fundamental unit of chromatin, a nucleosome consists of
 29 ~147 bp DNA wrapping around a histone octamer composed of four core histones
 30 (H2A, H2B, H3 and H4) (Luger et al. 1997). Nucleosomes are connected by stretches
 31 of "linker DNA". Dynamic packaging of nucleosomes results in two different chromatin
 32 accessibility statuses: open (accessible and active genomic regions with sparse
 33 nucleosome occupancy) and closed (inaccessible and inactive genomic regions with
 34 dense nucleosome occupancy). Positioning of nucleosomes and dynamic changes of
 35 chromatin status play important regulatory roles in DNA-templated processes such as
 36 transcription, DNA replication and repair (Bell et al. 2011).

37 Current genome-wide methods of nucleosome occupancy and/or chromatin
 38 accessibility mapping are mainly based on three types of assays followed by
 39 short-read sequencing technologies: 1) nucleosome's protection of nucleosomal DNA
 40 sequences from endogenous and exogenous enzymes (e.g., MNase-seq, DNase-seq,

1 ATAC-seq, NOME-seq and MPE-seq) (Schones et al. 2008; Song and Crawford 2010;
 2 Cui and Zhao 2012; Kelly et al. 2012; Buenrostro et al. 2015a; Ishii et al. 2015); 2)
 3 chromatin immunoprecipitation using a specific histone antibody (e.g., ChIP-seq with
 4 H3) (Wal and Pugh 2012); and 3) solubility differences between nucleosomal DNA
 5 and naked linker DNA (e.g., FAIRE-seq) (Bianco et al. 2015). In particular, NOME-seq
 6 treats target sample with exogenous methyltransferase to detect nucleosome
 7 occupancy and chromatin accessibility: the nucleosome protects nucleosomal DNA
 8 from being methylated by exogenous methyltransferase, while cytosines in naked
 9 linker DNA sequences are methylated to 5-methylcytosine (5mC) (Kelly et al. 2012).
 10 The following bisulfite sequencing identifies this methylation profile as bisulfite can
 11 convert unmethylated cytosine to uracil, which discriminates 5mC from unmethylated
 12 cytosine.

13 These methods can map averaged patterns of nucleosome occupancy and chromatin
 14 accessibility in a population of cells, failing in precise identification at the single-cell
 15 level. Although the single-cell versions of these methods have been recently
 16 developed (Small et al. 2014; Buenrostro et al. 2015b; Jin et al. 2015; Pott 2017; Clark
 17 et al. 2018; Lai et al. 2018; Li et al. 2018), the corresponding sparse sequencing
 18 coverage and short read length lack information for addressing complex long-range
 19 chromatin status and nucleosome occupancy. Therefore, the heterogeneity of
 20 nucleosome occupancy and chromatin accessibility is rarely studied. Moreover, it is
 21 even more challenging to define nucleosome occupancy patterns and dynamics and
 22 chromatin accessibility at single DNA molecules, so it is hard to detect subtle but
 23 meaningful differences between seemingly identical cells. This is a critical gap of
 24 understanding the mechanism of how nucleosomes assemble, disassemble and slide.
 25 In addition, in contrast to the well-solved phasing problems (e.g., exon-splicing and
 26 allele-specific DNA methylation), the phased status of nucleosomes and chromatin
 27 accessibility at single DNA molecules remains incomplete (Kuleshov et al. 2014;
 28 Tilgner et al. 2018).

29 The emerging single-molecule long-read sequencing technology (i.e., Oxford
 30 Nanopore Technologies, ONT) provides unique data features that are capable of
 31 filling the gap: 1) 5mC can be directly detected at the single-base resolution at the
 32 single-molecule level based on ONT electrolytic current signal dynamics without
 33 bisulfite conversion (Rand et al. 2017; Simpson et al. 2017); 2) unlike the other
 34 sequencing platforms (such as Sanger sequencing and Second Generation
 35 Sequencing (SGS, e.g., Illumina)), PCR amplification is not required for ONT
 36 sequencing, so each ONT read can reveal the genomic events at the single-molecule
 37 level; 3) ONT reads are ultra-long (up to 2.3 Mb) (Payne et al. 2018) so that they can
 38 cover combinatorics of many nucleosomes and different chromatin statuses spanning
 39 multiple genomic elements. Leveraging the informative ONT sequencing technology,
 40 we developed an experimental approach MeSMLR-seq (methyltransferase treatment
 41 followed by ONT single-molecule long-read sequencing) and the corresponding
 42 bioinformatics method, to investigate heterogeneous and dynamic insight of
 43 long-range chromatin status and nucleosomes. Instead of bisulfite conversion (with

PCR amplification) and short-read sequencing, the footprint of exogenous 5mCs from GpC-specific methyltransferase treatment is detected at single DNA molecules (without any PCR amplification) by ONT sequencing in the MeSMLR-seq protocol, and is next used to detect nucleosome occupancy and chromatin accessibility computationally.

We applied MeSMLR-seq to haploid *Saccharomyces cerevisiae* cells, where single DNA molecules represent single cells, so it allows the “one-to-one” link between sequencing read (i.e., sequencing molecule) and haploid cell. Thus, each single MeSMLR-seq read can be used to mimic a single cell in a given genomic region and the heterogeneity can be investigated without single-cell sequencing. With the unique output of MeSMLR-seq, we revealed the chromatin basis of gene transcription.

RESULTS

Overview of MeSMLR-seq

In brief, the experimental approach MeSMLR-seq (**M**ethyltransferase treatment followed by **S**ingle-**M**olecule **L**ong-**R**ead sequencing) contains two main steps: 1) methyltransferase (M.CviPI) treatment to convert cytosine to 5mC at GpC sites at naked linker DNA and open chromatin; and 2) ONT sequencing to detect 5mC profile that is subsequently used to identify nucleosome occupancy and chromatin accessibility (Fig. 1). The first step has been shown feasible at both the bulk-cell and single-cell level by NOMe-seq and the other previous studies (Small et al. 2014; Pott 2017; Clark et al. 2018; Li et al. 2018). In addition, ONT has been reported to detect 5mC at CpG sites (Rand et al. 2017; Simpson et al. 2017), based on which an in-house tool (named NP-SMLR, see Supplemental Code) was developed to map 5mC profile at GpC sites for MeSMLR-seq data.

In the proof-of-concept application of MeSMLR-seq to haploid *Saccharomyces cerevisiae* (BY4741 strain), an additional step was applied to digest the cell wall that serves as a barrier against methyltransferase treatment to genomic DNA: yeast cells were treated with Zymolyase to generate spheroplasts (Fig. 1). After the subsequent methyltransferase treatment, extracted genomic DNA without any PCR amplification was directly submitted to library preparation and ONT sequencing. The genomic DNA that underwent *in vivo* spheroplast methylation was referred as target sample of MeSMLR-seq. In addition, we prepared negative control and positive control samples as training data for 5mC detection (see the below section “Detection and phasing of nucleosome occupancy at single DNA molecules”). Native genomic DNA extracted from yeast without M.CviPI treatment was used as negative control (all cytosines at GpC sites were unmethylated) since there is no endogenous 5mC on yeast genome as previously reported (Capuano et al. 2014). Genomic DNA treated by M.CviPI (without spheroplast methylation) was used as positive control (all cytosines at GpC sites were converted to 5mCs).

As the efficiency of M.CviPI methylation served a critical role in the whole protocol, it was evaluated at selected genomic regions by bisulfite sequencing as previously described (Small et al. 2014). The methylation efficiency of positive control sample was 99.37% and 13 single colonies of the selected region from target sample were all successfully-methylated, indicating high methylation efficiency.

Using ONT GridION platform with R9.4.1 chemistry, we sequenced one flow cell per sample and generated 0.9 million (positive control), 1.2 million (negative control) and 1.3 million (on average for six target samples) reads (i.e., sequencing molecules), separately, which were uniquely aligned to yeast genome (Supplemental Table S1). The longest sequencing molecule was 63.1 kb. In particular, from the target sample where yeast was grown in rich media (1% yeast extract, 2% peptone and 2% glucose) we generated 1.4 million sequencing molecules with the median length of 7.2 kb, covering 821× of the yeast genome.

Detection and phasing of nucleosome occupancy at single DNA molecules

We first identified 5mC methylation status for every GpC site on each DNA molecule based on the ONT sequencing current signal (referred as event level). Since the previous studies (Rand et al. 2017; Simpson et al. 2017) showed the event level depended on the context sequence (e.g., 6-mer), our positive and negative control data were used to train signal distributions for each 6-mer containing target GpC dinucleotide under the occasions of methylation and unmethylation. The event levels of a given 6-mer from the target sample were compared with the corresponding trained distributions to obtain a posterior of methylation for every GpC site on each molecule, which we denoted as the methylation score (Supplemental Fig. S1A). There was no obvious bias of 5mC methylation calling between the molecules that were aligned to forward and reverse strands, and the areas under the receiver operating characteristic curve (AUC) were both 0.86 (Fig. 2A). Correlation analysis of methylation status of paired GpC sites at single molecules showed a pattern with periodic distance of 170-180 bp, which was the same as the length of nucleosomal DNA (147 bp) plus regular linker DNA (20-30 bp) (Fig. 2B). Therefore, we can identify nucleosome occupancy at single molecules from the methylation profiles by developing the bioinformatics method **NP-SMLR** (Nucleosome Positioning detection by Single-Molecule Long-Read sequencing) as below (see Supplemental Code).

Let $X_1X_2 \cdots X_l$ be a molecule, where X_i is the i -th base. Denote s_i as the methylation score of X_i , if X_i is the cytosine of the GpC dinucleotide. Suppose that the methylation scores of all GpC sites are independent. Nucleosome occupancy detection refers to finding a path $\pi = \pi_1\pi_2 \cdots \pi_l$ that maximizes the likelihood of signals:

$$\pi^* = \operatorname{argmax}_{\pi} \prod_{t=1}^n \Pr(s_{i_t} | \pi_{i_t}).$$

1 π_i takes the value from $\{L, N_1, N_2, \dots, N_{147}\}$. L represents the linker region; N_m
 2 represents the m -th base within a nucleosome; $i_1, i_2, \dots, i_n$ are the positions of
 3 cytosines that belong to GpC dinucleotides. The elements of path π are restricted
 4 such that: 1) N_m is followed by N_{m+1} ($1 \leq m \leq 146$); 2) N_{147} is followed by L ; and 3)
 5 L is followed by L or N_1 . The problem is essentially an alignment between a
 6 sequence of nucleotides and a sequence of nucleosomal statuses. NP-SMLR adopts
 7 dynamic programming algorithm (Needleman and Wunsch 1970) for solution: a matrix
 8 regarding the nucleotide sequence and nucleosomal statuses is made, entries are
 9 updated iteratively, and the optimal path is obtained through backtracking (Fig. 2C
 10 and Supplemental Fig. S1B).

11 Due to the lack of more advanced experimental technology to generate a gold
 12 standard, we evaluated the accuracy of nucleosome occupancy detection at the
 13 single-molecule level by simulation tests. The tests were performed under different
 14 settings of nucleosome coverage (proportion of bases covered by nucleosomes,
 15 ranges from 30-90%) and GpC frequency (ranges from 1-10%) (Fig. 2D). The
 16 accuracy increased with GpC frequency, while the effect of nucleosome coverage
 17 was mild. In case of yeast genome with 3.75% density of GpC sites, NP-SMLR was
 18 very robust to reach the accuracy of 80% regardless of different nucleosome
 19 coverages, which represented different scenarios of chromatin status (Fig. 2D). These
 20 results highlight the accuracy and robustness of MeSMLR-seq on single-molecule
 21 long-range mapping of nucleosomes.

22 **Performance of nucleosome occupancy detection at the bulk-cell level**

23 In terms of nucleosome occupancy at the bulk-cell level, MeSMLR-seq provided
 24 consistent and comparable results with the widely-used method MNase-seq
 25 (Supplemental Methods) (Hughes and Rando 2014; Weiner et al. 2015). The
 26 averaged Pearson's correlation coefficient between three MeSMLR-seq data
 27 (forwardly, reversely aligned molecules and their combination) and three MNase-seq
 28 replicates was 0.75 (Fig. 3A). The 77% of nucleosomes called by MeSMLR-seq were
 29 also detected by MNase-seq (Fig. 3C). For example of the *DAL* (degradation of
 30 allantoin) gene cluster, the nucleosome peaks called by MeSMLR-seq and
 31 MNase-seq were generally well aligned (Fig. 3B). In the long-range scale, single
 32 MeSMLR-seq reads can phase a number of nucleosomes (median number was 37
 33 and maximal number was 356 in our data), so that it captures the dynamics and
 34 heterogeneity of nucleosome occupancy among DNA molecules (Fig. 3D and
 35 Supplemental Table S2). For instance, 35 to 61 nucleosomes (median number 58)
 36 were phased at the single molecules covering the *DAL* gene cluster across a 10 kb
 37 genomic region (Fig. 3E), which illustrated large-range variation as well as local subtle
 38 difference of nucleosome occupancy.

39 **Direct long-range evidence of differential nucleosome organization**

40 A few single-cell epigenome sequencing approaches have revealed the heterogeneity
 41 of chromatin status and nucleosome positioning within a cell population (Small et al.

2014; Buenrostro et al. 2015b; Jin et al. 2015; Pott 2017; Clark et al. 2018; Lai et al. 2018; Li et al. 2018). Recently, Lai *et al.* reported the differential nucleosome organization principles for silent and active genes using single-cell MNase-seq (Lai et al. 2018) (Fig. 4A). However, these studies lacked a long-scale nucleosome occupancy scene at the single-cell resolution due to short sequencing length and sparse data coverage within single cells. As shown above, MeSMLR-seq can determine the heterogeneous long-range phasing of nucleosomes, so we can investigate nucleosome organization logic in a comprehensive way (Fig. 3E).

We focused on the nucleosome organization surrounding TSS, which plays an important role in transcription regulation (Voss and Hager 2014). For each gene, we measured the heterogeneity of nucleosome positioning by the standard deviation of the distances between +1 nucleosome and TSS over all single cells. Compared to active genes, silent genes showed larger heterogeneity of nucleosome positioning among different cells (Fig. 4B and Supplemental Fig. S2A). Next, we evaluated the uniformity of nucleosome spacing within single cells by the variation of the distance between adjacent nucleosomes. In contrast to active genes, the nucleosomes surrounding TSS of silent genes were more uniformly spaced (Fig. 4C and Supplemental Fig. S2B). For instance, at the bulk-cell level, nucleosomes surrounding TSS of the lowly-expressed gene *AUA1* (FPKM=0) were poorly positioned (Fig. 4D), while there were well-positioned nucleosomes (including -1, +1, +2, +3 and +4 nucleosomes) surrounding TSS of the active gene *EMW1* (FPKM=77) and a pronounced nucleosome-depletion region (NDR) upstream of the TSS (Fig. 4E). At the single-cell level, the positioning of +1 nucleosome of *AUA1* had a continuous shift pattern across different cells, whereas it was relatively steady for *EMW1* (Fig. 4D, E). Compared with *EMW1*, the distances between +1 nucleosomes and TSS for *AUA1* were more approximate to a uniform distribution (Supplemental Fig. S3A, B), which represented the ideal occasion for continuous shift pattern. In addition, the spacing of nucleosomes surrounding TSS of *AUA1* was relatively uniform within single cells (Fig. 4D and Supplemental Fig. S3C), while there was a pronounced NDR in the upstream of TSS of *EMW1*, which disrupted the uniformity of nucleosome spacing (Fig. 4E and Supplemental Fig. S3D). MeSMLR-seq resolves these differential nucleosome organization principles with direct and convincing evidence at a long-range scale from single molecules/cells that are hard to be obtained by the bulk-cell and short-read sequencing approaches.

Single-molecule long-range measurement of chromatin accessibility

Based on the methylation profiles of MeSMLR-seq data, we also mapped the chromatin accessibility of yeast genome at both bulk-cell level and single-molecule level. To assess the performance on the bulk-cell chromatin accessibility mapping, we compared MeSMLR-seq with two widely-used methods, ATAC-seq (Schep et al. 2015) and DNase-seq (Zhong et al. 2016) (Supplemental Methods). Genome-wide chromatin accessibility profile revealed by MeSMLR-seq data was highly consistent with ATAC-seq (averaged Pearson's $r=0.80$) and DNase-seq (averaged Pearson's

r=0.82) (Fig. 5A, B and Supplemental Fig. S4). In addition, >83% (1,615/1,934) significantly-accessible regions called by MeSMLR-seq were also supported by either ATAC-seq or DNase-seq (Fig. 5C). These results indicate that MeSMLR-seq provides comparable results with the existing methods on the bulk-cell level chromatin accessibility mapping.

At the single-molecule level, a MeSMLR-seq read can fully cover multiple adjacent genes (median number was 4 and maximal number was 40 in our data), therefore we could examine the long-range chromatin accessibility at the single-molecule/-cell level (Fig. 5D and Supplemental Table S3). For example, 34 MeSMLR-seq molecules fully covered the 9 kb genomic region ChrII:370000-379000 that encompasses four genes (*NRG2*, *TIP2*, *BAP2* and *TAT1*). Based on the 5mC footprint, we identified the chromatin status ("open" or "closed") of the promoters for four genes on each molecule and thus defined and quantified the coupled chromatin status patterns. In total, these molecules detected 13 out of 16 (4^2 , four genes with binary status "open" or "closed") possible combinatorial patterns of the coupled chromatin statuses of four gene promoters (Fig. 5E). For instance, four genes in Pattern 1 (supported by 2 molecules) all had "open" promoters, whereas the promoters of four genes were all closed in Pattern 6 (supported by 14 molecules). Therefore, MeSMLR-seq is able to analyze the coupled chromatin statuses of adjacent genes and to investigate the heterogeneity of chromatin status within a cell population, which is challenging for the existing methods.

Heterogeneous openness of gene promoter

Leveraging the single-molecule and long-range information of MeSMLR-seq data, we can discover and measure different levels of promoter openness instead of binary status. In the promoter region (ChrXVI:66400-67550) of the cell cycle regulation gene *CLN2*, the bulk-level chromatin accessibility profiles generated by the existing methods and MeSMLR-seq all showed a significant openness (Fig. 6, upper panel), while it was not clear if the promoters of *CLN2* among all cells were open, or if the open regions were similar in size. Based on the single-molecule nucleosome occupancy profiles in the promoter region, 304 molecules that fully covered this region were clustered into three groups with different levels of promoter openness: closed (Cluster 1 with 176 molecules), narrowly-open (Cluster 2 with 75 molecules) and widely-open (Cluster 3 with 53 molecules) (Fig. 6, lower-right panel). The 5mC profiles at the molecules from three clusters also showed the difference of the widths of openness (Fig. 6, lower-left panel). This unique output of MeSMLR-seq is bringing new opportunities to perform quantitative analysis of the heterogeneous and dynamic promoter status.

In addition, the MeSMLR-seq data revealed consistent nucleosome occupancy and chromatin accessibility profiles with the previous studies based on bulk-cell short-read data (Hughes and Rando 2014) in the upstream of TSS, gene body region, and binding region of several important transcriptional regulators (Supplemental Note 1 and Supplemental Fig. S5, S6), as well as the dynamics of chromatin status during

transcription changes in response to different nutrition conditions (Supplemental Note 2 and Supplemental Fig. S7).

Quantitative relationship between gene expression and chromatin accessibility in cell populations

Though the analyses above showed that the promoters of the highly-expressed genes over a cell population were generally more accessible than the lowly-expressed genes (Supplemental Fig. S5), the quantitative relationship between promoter openness and gene transcription in a cell population remained unclear. Based on unique MeSMLR-seq data, we were able to calculate the fraction of cell subpopulation with open promoter of a given gene. With single-cell RNA-seq data for 2,812 yeast cells generated in this study (Supplemental Methods), we also calculated the fraction of cells with expression (read count ≥ 1) of a given gene (referred as expression frequency). The expression frequency within a cell population was positively correlated with the fraction of cells with open promoter (Fig. 7A). For example, the genes with open promoter in $\geq 40\%$ cells had significantly larger expression frequency than the ones with open promoter in $< 10\%$ cells (P -value $< 2.2 \times 10^{-16}$, Wilcoxon rank sum test, Fig. 7A).

When grouping the genes based on expression frequency, we observed similar positive correlation (Fig. 7B). In addition, considering the bulk-cell expression, the highly-expressed genes had relatively large fractions of cell subpopulation with open promoter in comparison to the lowly-expressed ones (P -value $< 2.2 \times 10^{-16}$, Wilcoxon rank sum test, Fig. 7C). These results suggest that chromatin accessibility of promoter at the single-molecule/-cell level detected by MeSMLR-seq data can contribute to the prediction of gene expression level and frequency in a cell population.

Coupled chromatin accessibility changes of adjacent genes during transcription reprogramming

Making full use of the single-molecule and long-range advantages of MeSMLR-seq data, we explored the coupled chromatin status changes of two adjacent glucose transporter genes, *HXT3* and *HXT6* during transcription reprogramming. The transport of glucose across the plasma membrane is the first step of glucose metabolism, and the glucose (also called hexose) transporter genes play essential regulatory roles in glucose sensing, signaling and utilization in a yeast cell (Ozcan and Johnston 1999). Hxt3 and Hxt6 have different affinities to glucose (low-affinity for Hxt3 and high-affinity for Hxt6) and thus respond differently to the change of glucose concentration. With the decrease of glucose concentration, the expression of *HXT3* decreased whereas *HXT6* increased, which corresponded to their low- and high-affinity of glucose (Fig. 8A).

For each glucose concentration (2%, 1%, 0.5% and 0.125%), we counted MeSMLR-seq molecules to estimate the fractions of cell subpopulations with two opposite coupled chromatin accessibility patterns: "Open-*HXT3* and Closed-*HXT6*"

and “Closed-*HXT3* and Open-*HXT6*”. The fraction of cell subpopulation with the coupled pattern “Open-*HXT3* and Closed-*HXT6*” decreased along with the reduction of glucose concentration, whereas “Closed-*HXT3* and Open-*HXT6*” increased (Fig. 8B). The changes of two coupled patterns significantly matched the expression dynamics of two genes in response to glucose concentration change (P -value $< 3.9 \times 10^{-8}$, Chi-square test, Fig. 8). These proof-of-concept results highlight the promising utility of MeSMLR-seq on studying complex epigenetic changes during transcription reprogramming.

DISCUSSION

In this study, we showed the consistent bulk-level nucleosome occupancy and chromatin accessibility profiles generated by MeSMLR-seq with existing methods. With the unique output of MeSMLR-seq, we investigated the organization principles of nucleosomes surrounding TSS and studied the heterogeneity of combinatorial chromatin statuses over multiple genomic regions. Together with single-cell RNA-seq data, the relationship between chromatin accessibility and gene transcription was investigated quantitatively. Finally, we revealed the coupled chromatin changes of adjacent genes during transcription reprogramming.

A large number of studies have demonstrated key regulatory roles for nucleosome positioning and chromatin accessibility in eukaryotic gene expression (Li et al. 2005; Li et al. 2007; Petesch and Lis 2008; Jiang and Pugh 2009) as well as DNA repair, recombination and other DNA-dependent processes (Lipford and Bell 2001; Dalal et al. 2007; Schwartz et al. 2009; Tilgner et al. 2009; Cole et al. 2011; Lai and Pugh 2017). The relationship between nucleosome positioning, chromatin accessibility and gene expression has been studied most extensively (Rando and Winston 2012). However, unlike the well-studied heterogeneity of gene expression based on single-cell analyses, the heterogeneity of nucleosome positioning and chromatin accessibility is poorly studied due to limitations in experimental and sequencing techniques. Previous bulk-cell studies based on the well-developed experimental techniques established the fundamental knowledge base, while their corresponding versions at the single-cell platforms have not yet lead to more details. This is largely due to the sparse sequencing coverage and short read length. MeSMLR-seq provides an alternative way to address this bottleneck: long read length guarantees the full length of genomic region of interest (e.g., whole gene body together with the flanking neighborhood) can be covered by many single reads (that is, single DNA molecules). In the application to haploid organisms, MeSMLR-seq read population represents cell population, so the heterogeneity at the cell level can be investigated. In this study, MeSMLR-seq provides a long-range chromatin status landscape and nucleosome occupancy detection at the single-molecule/-cell level. The investigation of coupled chromatin changes and differential nucleosome organization principles in response to nutrition changes underline the unique MeSMLR-seq output on exploring these complex epigenetic events.

However, it should be noted that the molecule-cell link does not hold in diploid or polyploid organisms, as the molecule populations are a mix of allele-specific and cell-specific events. It leads to challenges yet opportunities in the further development of new experimental (e.g., single-cell barcoding) and statistical (e.g., data deconvolution) approaches. Once cell subpopulations can be reconstructed from a molecule population, we could distinguish the allele-specific epigenome precisely from different cell subpopulations and achieve more accurate investigation of how epigenetics events behave differently at different alleles. Regardless of the wide interest on the cell-level study, the characterization of nucleosome positioning and chromatin status at single DNA molecules by MeSMLR-seq will also bring very unique and informative data to reveal the dynamic nucleosome positioning mechanism, such as assembly, disassembly, and sliding.

Besides the single-molecule information, the long length of MeSMLR-seq reads, which allows correlation analysis of exogenous and endogenous methylation statuses over different positions, could be informative for some research topics: 1) correlation of exogenous 5mC events has shown the nucleosome occupancy pattern in this study (Fig. 2B), and thus DNA loops or other larger spatial chromatin domain that affects exogenous methylation could be also identified, which would require specific library preparation to generate even longer ONT reads; 2) As endogenous 5mC can be also detected, MeSMLR-seq can be applied to higher organisms (e.g., human) to study how methylation status at different genomic region coordinates, but it could also provide direct evidence to address the controversial topics about how methylation status and nucleosome positioning and chromatin openness correlates. For example, in human cells, the endogenous 5mC mainly occurs at CpG sites and can be distinguished from exogenous GpC-specific 5mC detection using ONT data. Thus, the DNA methome, nucleosome occupancy and chromatin accessibility can be simultaneously measured on a single DNA molecule in human. Further, the ultra-long length of ONT reads (up to Mbps) enables the analysis of the coupled dynamics of DNA methylation and chromatin status of adjacent genes, since the median distance between adjacent genes is 36 kb in human.

On a technical view, there is relatively few application of ONT data in epigenetics research, as the corresponding experimental approaches or bioinformatics methods are rarely developed, although numerous applications of ONT data have been published rapidly with improved data quality and cost efficiency. In addition to the previously reported studies of identifying methylation and three-dimensional spatial organization of chromatin (<https://nanoporetech.com/resource-centre/pore-c-using-nanopore-reads-delineate-long-range-interactions-between-genomic-loci>), MeSMLR-seq contributes a new technique in the toolkit of single-molecule long-read sequencing to obtain the first-hand details of epigenetics at single DNA molecules. More innovative studies with single-molecule long-read sequencing should be explored and expected to advance our studies to discover novel and complex biological insights.

METHODS

Yeast strain and growth

Saccharomyces cerevisiae BY4741 strain was used in this study. Yeast cells were separately grown at 30°C in the media including 1% yeast extract, 2% peptone and different carbon sources. Yeast cells were collected in the mid-log phase (OD₆₀₀ of 0.3-0.6) and subjected to MeSMLR-seq, bulk-cell RNA-seq and single-cell RNA-seq experiments (Supplemental Table S4 and Supplemental Methods).

MeSMLR-seq experiment

Preparation and methylation of yeast spheroplasts were performed as previously described (Small et al. 2014) (Fig. 1). Briefly, yeast cells were treated with Zymolyase (amsbio, final conc. = 0.25 mg/mL) in 1 M sorbitol and 50 mM Tris (pH7.4) and 10 mM β-mercaptoethanol. Spheroplasts were washed using 1 M sorbitol twice before methyltransferase treatment. GpC-specific methyltransferase M.CviPI (NEB) supplemented with 160 μM SAM S-adenosylmethionine was used to methylate spheroplasts at 37°C for 45 min. Genomic DNA was extracted using PCI (Phenol:chloroform:isoamyl alcohol, 25:24:1) and purified by Genomic DNA Clean & Concentrator™-10 Kit (Zymo Research).

We denote the above mentioned genomic DNA that undergoes *in vivo* spheroplast methylation as target sample of MeSMLR-seq. Native genomic DNA extracted from yeast without M.CviPI treatment was used as negative control (all cytosines at GpC sites are unmethylated). There is no endogenous 5mC in yeast genome as reported in previous study (Capuano et al. 2014). Genomic DNA treated by M.CviPI (without spheroplast methylation) was used as positive control (all cytosines at GpC sites are 5mCs).

The efficiency of M.CviPI methylation was evaluated using bisulfite sequencing as previously described (Small et al. 2014). Firstly, bisulfite conversion was performed using EZ DNA Methylation Lighting Kit (Zymo Research). Secondly, PCR amplification targeted to specific genomic regions was performed by ZymoTaq PreMix (Zymo Research). *CHA1* gene region (ChrIII:15713-16074), *CYS3* gene region (ChrI:130966-131117), *GAL10* gene region (ChrII:278464-278738) and *PHO5* gene region (ChrII:430248-430388) were amplified for evaluating the methylation efficiency of positive control. The *PHO5* gene region (ChrII:430843-431498), which was shown in the Figure 1 of the previous study (Small et al. 2014), was used to estimate the efficiency of spheroplast methylation (i.e., target sample of MeSMLR-seq). Thirdly, TA cloning was performed by TOPO™ TA Cloning Kit (Life Technologies). Single colonies were picked and plasmids were extracted by QIAprep Spin Miniprep Kit (QIAGEN). Finally, plasmids were sequenced by Sanger sequencing. For positive control, we estimated the efficiency of methylation as the percentage of 5mC over all GpC sites (totally 53 GpC sites for four target gene regions). Three single colonies

were sequenced per gene region; and the methylation efficiency of positive control was $((53 \times 3) - 1) / (53 \times 3) = 99.37\%$. For target sample of MeSMLR-seq, we considered it as successfully-methylated if the single colony included at least one 5mC. In total, 10 colonies were sequenced; and the methylation success rate of target sample was up to 100% (10/10). The Sanger sequences for templates and colonies were provided in Supplemental Table S5.

Native genomic DNA (negative control), methylated genomic DNA (positive control) and extracted genomic DNA after spheroplast methylation (target sample) were directly submitted to ONT sequencing. In brief, the genomic DNA was fragmented (size = 8 kb) using Megaruptor. Sequencing library was prepared using the 1D Ligation Sequencing Kit (SQK-LSK108). ONT sequencing was performed using GridION platform with R9.4.1 flow cells.

MeSMLR-seq data preprocessing

The software Albacore developed by ONT was used to perform basecalling for ONT raw signals. The base-called ONT sequencing data were aligned to *sacCer3* reference genome using BWA software (version 0.7.17-r1188) (Li and Durbin 2010) with the “mem” mode and the “-x ont2d” parameter. Nanopolish (version 0.8.5) (Simpson et al. 2017) with the “eventalign” mode and the “--scale-events” parameter was used to generate the alignments between event levels and 6-mers for each sequencing molecule, which were utilized for the following GpC-specific 5mC detection.

Since we used ONT 1D sequencing strategy in this study, a DNA molecule from yeast cell might be sequenced twice (i.e., forward and reverse strands). Thus, to achieve the “one-to-one” link between ONT sequencing molecule and haploid yeast cell, we classified all molecules into two groups based on their aligned genomic strands: forward and reverse.

The information of MeSMLR-seq data (including data size, read length, error rate, alignment rate, and genome coverage) was summarized in Supplemental Table S1.

GpC-specific 5mC detection at the single-molecule level and single-base resolution by MeSMLR-seq

For each of the 6-mer pattern that includes cytosine of GpC dinucleotide, represented here by k , we trained two models for the event level; one was for unmethylated cytosine, and the other for methylated cytosine. We modeled the event level for unmethylated cytosine by a Gaussian distribution $\mathcal{N}(\mu_0, \sigma_0^2)$. From the negative control data preprocessed by “nanopolish eventalign”, we pooled the event levels that were aligned to k , and estimated μ_0 and σ_0^2 by the sample mean and variance, respectively. Considering the fact that the efficiency of exogenous methylation was not always 100% (99.37% in our experiment) (Supplemental Fig. S1A, right panel), we modeled the event level for methylated cytosine by a Gaussian mixture model $\rho \mathcal{N}(\mu_1, \sigma_1^2) + (1 - \rho) \mathcal{N}(\mu_2, \sigma_2^2)$, $0 < \rho < 1$. The parameters μ_1 , μ_2 , σ_1^2 , σ_2^2 , and ρ

were estimated based on the event levels from positive control data by EM algorithm. Denote the probability density functions of the two models as $f_N(x; k)$ (unmethylated cytosine) and $f_P(x; k)$ (methylated cytosine), respectively. Namely,

$$f_N(x; k) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}},$$

$$f_P(x; k) = \rho \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + (1-\rho) \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}.$$

The x in the above formulas represents the value of event level. The area of the overlapped region under the two probability density functions $f_N(x; k)$ and $f_P(x; k)$ is calculated. The discrimination of the 6-mer k is defined as 1 - the area of overlap.

Given a sequencing molecule from target sample, we detected 5mC for all GpC sites. For each of the GpC site, we listed all the 6-mers on the reference genome that cover the cytosine at the dinucleotide (Supplemental Fig. S1A, left panel). The 6-mer with >1 GpC site included or >10 aligned event levels from the molecule was excluded for 5mC detection. Among the remaining 6-mers, the one with the maximal discrimination was chosen for the calculation of methylation score. Denote k' as the selected 6-mer, and x' as the event level that was aligned to k' . The event level x' was filtered out if one of $\log f_P(x'; k')$ or $\log f_N(x'; k')$ was <-10; otherwise, the methylation score of the GpC site was calculated as

$$s = \frac{f_P(x'; k')}{f_P(x'; k') + f_N(x'; k')}.$$

The score s was essentially the posterior probability of methylation given a non-informative prior. If multiple event levels were aligned to k' , then $f_P(x'; k')$ and $f_N(x'; k')$ were replaced by the product of the multiple likelihood.

As a cross-validation, we randomly split each of the negative control and positive control data into two half. One of the half was used to train the models. Using the trained models, we detected 5mC on the other half. With the real methylation status of the test data being known, we were able to evaluate the detection results. It turned out that the area under the ROC curve (AUC) was 0.86 (Fig. 2A).

Nucleosome occupancy detection at the single-molecule level by MeSMLR-seq

We developed a bioinformatics method, named **NP-SMLR** (Nucleosome Positioning detection by **Single-Molecule Long-Read** sequencing), to detect and phase nucleosomes at the single-molecule level (Fig. 2C; Supplemental Code).

Let $X_1 X_2 \cdots X_l$ be a molecule, where X_i is the i -th base. Denote s_i as the methylation score of X_i , if X_i is the cytosine of the GpC dinucleotide. Suppose that

the event levels of all GpC sites are independent. Nucleosome occupancy detection refers to finding a path $\pi = \pi_1 \pi_2 \cdots \pi_l$ that maximizes the likelihood of signals:

$$\pi^* = \operatorname{argmax}_{\pi} \prod_{t=1}^n \Pr(s_{i_t} | \pi_{i_t}).$$

π_i takes the value from $\{L, N_1, N_2, \dots, N_{147}\}$. L represents the linker region; N_m represents the m -th base within a nucleosome; $i_1, i_2, \dots, i_n$ are the positions of cytosines that belong to GpC dinucleotides. The elements of path π are restricted that: 1) N_m is followed by N_{m+1} ($1 \leq m \leq 146$); 2) N_{147} is followed by L ; and 3) L is followed by L or N_1 .

Based on the methylation scores of all GpC sites from all molecules in negative and positive control training data, we can fit two density curves using the “density” command in R (version 3.3.0), respectively. The two density functions are denoted as $q_N(\cdot)$ and $q_P(\cdot)$, respectively (Supplemental Fig. S1B). A dummy methylation score $s_i = -1$ is added for X_i if it is not a cytosine of GpC dinucleotide. Define

$$p_i(\pi_i) \triangleq 1_{\{s_i = -1\}} + 1_{\{s_i \neq -1\}} \cdot \Pr(s_{i_t} | \pi_{i_t}) = 1_{\{s_i = -1\}} + 1_{\{s_i \neq -1\}} \cdot q_P(s_i)^{1_{\{\pi_i = L\}}} \cdot q_N(s_i)^{1_{\{\pi_i \neq L\}}}.$$

Let $a_{\pi_i, \pi_{i+1}}$ be the compatibility indicator of two adjacent states such that

$$a_{\pi_i, \pi_{i+1}} = 1_{\{\pi_i = N_m, \pi_{i+1} = N_{m+1}, 1 \leq m \leq 146\}} + 1_{\{\pi_i = L, \pi_{i+1} = N_1\}} + 1_{\{\pi_i = L, \pi_{i+1} = L\}}.$$

The objection function can therefore be expressed as

$$\mathcal{L} = p_1(\pi_1) \prod_{i=2}^n p_i(\pi_i) a_{\pi_{i-1}, \pi_i}.$$

Define

$$\ell_{k, \zeta} = \max_{\pi_1 \dots \pi_k, \pi_k = \zeta} p_1(\pi_1) \prod_{i=2}^k p_i(\pi_i) a_{\pi_{i-1}, \pi_i}.$$

Then the maximum of objection function can be obtained by iteration:

$$\ell_{k+1, \xi} = \max_{\zeta} \ell_{k, \zeta} \cdot p_{k+1}(\xi) \cdot a_{\zeta, \xi},$$

$$\max_{\pi} \mathcal{L} = \max_{\xi} \ell_{n, \xi}.$$

Accordingly, π^* can be obtained through dynamic programming (Fig. 2C). We start by building an $l \times 148$ matrix V . Line i corresponds to X_i , the i -th base of the molecule. Column 1 corresponds to the linker; and the other columns (from Column 2 to Column 148) correspond to $N_1, N_2, \dots, N_{147}$, separately. Initialize $V[1, 1] = p_1(L)$, and $V[1, j] = p_1(N_{j-1})$, $2 \leq j \leq 148$. Elements in Line i ($2 \leq i \leq n$) are then calculated iteratively. For Column 1, the element $V[i, 1]$ is set as $\max\{V[i-1, 1], V[i-1, 147]\} q_P(s_i)$ if X_i is cytosine of GpC, or $\max\{V[i-1, 1], V[i-1, 147]\}$ otherwise. For Column j ($2 \leq j \leq 148$), $V[i, j]$ is set as $V[i-1, j-1] q_N(s_i)$, or $V[i-1, j-1]$ otherwise. When updating an element, we record the position of the previous element that leads to the maximal value, and store the position as a pointer. After updating all elements, the maximal element in the last line is found (elements

that equal to one are not considered), and the nucleosome occupancy detection is completed through the backtracking of pointers. All calculations are performed in log scale to avoid rounding error.

We evaluated the accuracy of nucleosome occupancy detection (NP-SMLR) through simulation tests under different nucleosome coverage and GpC frequency (Fig. 2D). In detail, DNA sequence (3-kb length) was simulated with randomly assigned GpC sites at given frequency. Lengths of linkers between nucleosomes were sampled independently and sequentially. At each time, the linker length was sampled from the normal distribution $N(v_1, \gamma_1^2)$ with probability τ , and $N(v_2, \gamma_2^2)$ with probability $1 - \tau$, corresponding to regular nucleosome array and open region with specific biological functions, respectively. We set $v_2 > v_1$ and $\gamma_2^2 > \gamma_1^2$. Nucleosomes were then placed on the DNA sequence, with their distance being set as the above simulated linker length. Methylation scores for GpC sites occupied by nucleosomes were generated based on the score distribution of negative control data, whose density function was $q_N(\cdot)$. For GpC sites within linkers, $q_P(\cdot)$ was used instead. NP-SMLR was applied on the simulated sequence. Denote $\hat{Z}_i$ and Z_i as the predicted and real indicators of whether the i -th base locates in nucleosome or not, respectively. The accuracy was defined as

$$A = \frac{1}{l} \sum_{i=1}^l 1_{\{\hat{Z}_i = Z_i\}},$$

where l is the length of the simulated DNA sequence. In simulation tests, we set $v_1 = 15$, $\gamma_1 = 5$, $\gamma_2 = 10$, $\tau = 0.1$. We set v_2 as 15, 50, 100, 200, 300, 400, 500, 600, respectively to achieve different nucleosome coverage (defined as the proportion of bases covered by nucleosomes). For each parameter setting, the above simulation was carried out for 1,000 times.

Bulk-cell level nucleosome occupancy analyses based on MeSMLR-seq data

The genomic coordinates of all nucleosomes predicted by NP-SMLR at the single-molecule level were pooled and subjected to iNPS software (version 1.2.2) (Chen et al. 2014) with default parameters to generate bulk-cell level nucleosome occupancy profile and to call nucleosome peaks.

The nucleosome occupancy profiles were used to generate Fig. 3A, B; Fig. 4D, E (upper panel); Supplemental Fig. S5A, B; and Supplemental Fig. S6D, E. The nucleosome peaks called by iNPS were used for the comparison with MNase-seq (Fig. 3C).

Measurement of nucleosome positioning heterogeneity

The heterogeneity of nucleosome positioning was measured by the variation of the +1 nucleosome positioning relative to TSS across different cells (Fig. 4B and Supplemental Fig. S2A). For each molecule/cell, we first defined the nucleosome

whose center was located in the downstream of TSS and closest to TSS as the +1 nucleosome. Next, we sorted the distances between +1 nucleosome and TSS, and removed the upper 10% values for robustness. The standard variance of the remaining values was used to represent the heterogeneity of nucleosome positioning for each gene.

Measurement of nucleosome spacing uniformity

The uniformity of nucleosome spacing was measured by the variation of the distance between adjacent nucleosomes (i.e., the length of linker region) (Fig. 4C and Supplemental Fig. S2B). For each gene, the molecules that fully covered the region (from upstream 500 bp to downstream 100 bp of TSS) were chosen. For each molecule, we calculated the lengths of all linker regions that were located in the region “-500, +100”. Then, we calculated the absolute deviation of linker length pair-wisely. The sum of the deviation values was divided by the number of linker pairs. The obtained value, which described the variation of nucleosome distance, was namely the nucleosome spacing uniformity.

Chromatin accessibility mapping at the single-molecule level based on MeSMLR-seq data

Based on the methylation scores of all GpC sites per molecule, we detected accessible chromatin regions along the molecule. Given a single molecule $X_1X_2 \cdots X_L$, where X_i is the i -th base, we defined the interval from X_i to X_j as an accessible region if: 1) X_i and X_j were adjacent GpC sites; 2) the corresponding methylation scores s_i and s_j were >0.5 ; and 3) the distance between X_i and X_j was <100 bp. The continuous accessible regions were merged. Given an accessible region, the chromatin accessibility score was defined as the median methylation score among all GpC sites within this region.

In this study, we only considered the accessible regions with the length ≥ 100 bp for each molecule. Genome-wide chromatin accessibility profile was generated through merging accessible regions of all molecules. The chromatin accessibility profile was used to generate the Fig. 5A, B; Fig. 6 (upper panel); Supplemental Fig. S4A, B; Supplemental Fig. S5C; Supplemental Fig. S6B, C; and Supplemental Fig. S7A.

Chromatin accessibility peak calling at the bulk-molecule/-cell level based on MeSMLR-seq data

We defined significantly-accessible genomic regions as described in the previous study (Grunberg et al. 2016). Let G_i be the i -th base of the genome. Denote $X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(M)}$ as the bases from M sequencing molecules that covered G_i , and

$s_i^{(1)}, s_i^{(2)}, \dots, s_i^{(M)}$ as the corresponding methylation scores if G_i is a GpC site. Define

$r_i = \frac{1}{M} \sum_{j=1}^M 1_{\{s_i^{(j)} > 0.5\}}$, which is the ratio of methylated bases (methylation score >0.5),

and denote $\bar{r}$ as the average of ratios of all GpC sites. We defined the interval between G_i and G_j as a significantly-accessible region if: 1) G_i and G_j were adjacent GpC sites; 2) $r_i > 1.5\bar{r}$, and $r_j > 1.5\bar{r}$; and 3) the distance between G_i to G_j was <100 bp. The continuous accessible regions were merged to generate a longer accessible genomic region (referred as "chromatin accessibility peak").

In this study, we only considered the peaks with the length ≥ 100 bp. For sequencing molecules aligned to forward and reverse genomic strands, we defined chromatin accessibility peaks, separately. The overlapped peaks between the forward and reverse strands were used for the comparison with two existing methods (i.e., ATAC-seq and DNase-seq) (Fig. 5C).

Definition of gene promoter region and measurement of gene accessibility

To quantitatively measure the accessibility of genes, we first defined the promoter region for each gene. Briefly, chromatin accessibility peaks (including both forward and reverse strands) were called using MeSMLR-seq data for each biological sample. For each biological sample, the overlapped peaks between forward and reverse strands for MeSMLR-seq were merged together. Next, we combined the merged peaks of MeSMLR-seq from all biological samples and the overlapped peaks between two biological replicates of DNase-seq. For each gene, 1) if there was only one peak that was located within the upstream 500 bp and downstream 100 bp of TSS (named "-500, +100" region), the peak was defined as the promoter region; or 2) if there were multiple peaks that were located in the "-500, +100" region, the peak that had the longest overlap was defined as the promoter region; or 3) if there was no peak locating in the region "-500, +100", the region "-500, +100" was defined as the promoter region.

At the single-molecule level, the accessibility score of a gene was calculated as the median methylation score among all GpC sites within the promoter region. For all molecules covering the promoter of a given gene, we categorized them into two chromatin statuses: "open" if the accessibility score was >0.5 ; "closed" otherwise. The defined promoter region and the corresponding accessibility score were used to generate the Fig. 5E; Fig. 6 (upper panel); Fig. 7; Fig. 8C, D; Supplemental Fig. S6A; and Supplemental Fig. S7D.

Analyses of dynamic gene expression and chromatin accessibility among three carbon sources

Differentially-expressed genes were identified using Cuffdiff (version 2.2.1) (Trapnell et al. 2010) (q -value <0.01) between glucose (Glu) and other two carbon sources, galactose (Gal) and raffinose (Raf). Overall, there were 700 up-regulated and 682 down-regulated genes in Gal (Glu vs. Gal); 605 up-regulated and 727 down-regulated genes in Raf (Glu vs. Raf). These differentially-expressed genes were used to

generate the Supplemental Fig. S7B-D. Gene enrichment analyses in Supplemental Fig. S7C was performed using DAVID (version 6.8) (Huang et al. 2009).

For the differential chromatin accessibility analyses, we first calculated the bulk-cell-level chromatin accessibility as the ratio of those with “open” status among the molecules that fully covered the gene promoter. For each gene, the differential chromatin accessibility score was computed as the difference of bulk-cell-level chromatin accessibility between two carbon sources (Glu minus Gal for “Glu vs. Gal”; Glu minus Raf for “Glu vs. Raf”).

DATA ACCESS

The MeSMLR-seq data generated in this study have been submitted to the NCBI BioProject database (<http://www.ncbi.nlm.nih.gov/bioproject/>) under accession number PRJNA510813. The bulk-cell RNA-seq and the single-cell RNA-seq data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE131702. The Sanger sequences data generated in this study are summarized in the Supplemental Table S5. The public sequencing data used in this study are summarized in the Supplemental Table S6. The source code of NP-SMLR is in the Supplemental Material (Supplemental_Code.zip) and is also available at <https://github.com/Au-Lab/NP-SMLR>.

ACKNOWLEDGEMENTS

We would like to thank Dr. Charles Brenner (Department of Biochemistry, University of Iowa) for providing yeast BY4741 strain. We are grateful to Drs. David Stoltz and Massimo Attanasio (Department of Internal Medicine, University of Iowa) for experimental support. We thank Audrey Bollas for critical reading of the manuscript. This work was supported by an institutional fund of Department of Internal Medicine, University of Iowa and an institutional fund of Department of Biomedical Informatics, The Ohio State University (to K.F.A., Y.W., A.W., and Z.L.), the National Institutes of Health (R01HG008759 to K.F.A., Y.W., A.W., and Z.L.), and the Multidisciplinary Lung Research Career Development Program (T32HL007638 to A.T.).

Author contributions: K.F.A. and Y.W. devised the experiments. Y.W., L.S.P. and A.H. performed the experiments. K.F.A., A.W., A.L.T. and M.Z. devised the bioinformatics method. K.F.A., Y.W. and A.W. devised the data analyses. K.F.A., Y.W., A.W., Z.L., A.L.T. Y.Z. and Y.L. performed the data analyses. All of the authors discussed and interpreted the results throughout the project. K.F.A., Y.W. and A.W. wrote the paper with inputs from all of the other authors.

DISCLOSURE DECLARATION

1 The authors declare that they have no competing interests.

2 **FIGURE LEGENDS**

3 **Fig. 1 Overview of MeSMLR-seq.** Experimental approach (methyltransferase
4 treatment plus ONT sequencing) in yeast and the corresponding bioinformatics
5 analyses (5mC detection, chromatin accessibility mapping and nucleosome phasing).

6 **Fig. 2 5mC detection and nucleosome occupancy detection by MeSMLR-seq data.**
7 **(A)** ROC curve of 5mC detection on GpC sites. The molecules that were aligned to
8 forward (fwd) and reverse (rev) genomic strands were analyzed separately. **(B)**
9 Correlation coefficients between methylation scores of mutually paired GpC sites from
10 the same molecules with respect to their corresponding distances. **(C)** Dynamic
11 programming algorithm for nucleosome occupancy detection (NP-SMLR). A matrix
12 regarding the nucleotide sequence (row) and nucleosomal statuses (column) is made,
13 followed by initialization, iterative update for entries, and backtrack search for optimal
14 path (see Methods for details). **(D)** Accuracy of nucleosome occupancy detection
15 under different nucleosome coverage and GpC frequencies.

16 **Fig. 3 Performance evaluation of MeSMLR-seq on bulk-level nucleosome**
17 **occupancy, and single-molecule long-range phasing of nucleosomes.** **(A)**
18 Correlation of nucleosome occupancy profiles generated by MeSMLR-seq and
19 MNase-seq. For MeSMLR-seq, the molecules that were aligned to forward (fwd) and
20 reverse (rev) genomic strands were analyzed separately. **(B)** Nucleosome occupancy
21 profiles at the bulk-cell level provided by MeSMLR-seq and MNase-seq. **(C)** Overlap
22 of nucleosomes detected by MeSMLR-seq and MNase-seq at the bulk-cell level. **(D)**
23 Number of nucleosomes phased at single sequencing molecules of MeSMLR-seq
24 data under 2% glucose condition. **(E)** Detection and phasing of nucleosomes at the
25 single-molecule level by NP-SMLR. Each grey line represents a molecule. Green oval
26 represents nucleosome.

27 **Fig. 4 Differential nucleosome organization principles for silent and active**
28 **genes.** **(A)** Previous studies revealed nucleosome organization patterns surrounding
29 TSS of silent (left) and active (right) genes (Lai et al. 2018). Nucleosome positioning in
30 promoter regions of silent genes showed large variation among cells but was highly
31 uniformly spaced within each cell. In contrast, nucleosome positioning surrounding
32 TSS of active genes showed little variation among cells but relatively non-uniformly
33 spacing within each cell. **(B)** Heterogeneity of nucleosome positioning for silent
34 (FPKM=0) and active (FPKM>50) genes. The heterogeneity of nucleosome
35 positioning was measured by the standard deviation (SD) of the distances between
36 +1 nucleosomes and TSS. The *P*-value was calculated by Wilcoxon rank sum test. **(C)**
37 Uniformity of nucleosome spacing for silent (FPKM=0) and active (FPKM>50) genes.
38 See Methods for the definition of uniformity. The *P*-value was calculated by Wilcoxon
39 rank sum test. **(D)** Long-range nucleosome positioning patterns for the
40 silently-transcribed gene *AUA1* across different cells. Each row represents a cell and

nucleosome is labeled as blue bar. **(E)** Long-range nucleosome positioning patterns for the actively-transcribed gene *EWM1* across different cells.

Fig. 5 Performance evaluation of MeSMLR-seq on bulk-level chromatin accessibility mapping, and single-molecule long-range mapping of chromatin accessibility. **(A)** Correlation of chromatin accessibility profiles generated by MeSMLR-seq, ATAC-seq and DNase-seq. **(B)** Chromatin accessibility profiles at the bulk-cell level provided by MeSMLR-seq, ATAC-seq and DNase-seq. **(C)** Overlap of the significantly-accessible regions (peaks) called by MeSMLR-seq, ATAC-seq and DNase-seq. **(D)** Number of genes covered by single sequencing molecules of MeSMLR-seq data under 2% glucose condition. **(E)** Single-molecule long-range mapping of chromatin accessibility by MeSMLR-seq. Each line represents a molecule. GpC site is labeled as rainbow-color dot, with methylation score from 0 (blue) to 1.0 (red). Thirteen combinatorial patterns of the promoter status of four genes are shown with different numbers of supporting sequencing molecules/cells. A promoter was defined as “open” (highlighted by red box) if the methylation scores of the including GpC sites had a median value greater than 0.5, and “closed” (highlighted by blue box) otherwise.

Fig. 6 Heterogeneous promoter openness of *CLN2* in a cell population revealed by MeSMLR-seq. The bulk-level chromatin accessibility profiles (upper panel) were provided by ATAC-seq, DNase-seq, and MeSMLR-seq. MeSMLR-seq molecules were clustered into three groups with different promoter openness (by *k*-means clustering of the nucleosome occupancy profiles, bottom right panel): closed, narrow open and wide open. Each row represents a molecule (i.e., a cell) and nucleosome is labeled as blue bar. The corresponding methylation profiles at GpC sites on each molecule are shown on the bottom left panel. Each line represents a molecule (i.e., a cell). GpC site is labeled as rainbow-color dot, with methylation score from 0 (blue) to 1.0 (red).

Fig. 7 Quantitative relationship between chromatin accessibility and gene expression. **(A, B)** Quantitative relationship between chromatin accessibility and gene expression in a cell population. The former **(A)** was measured by the fraction of cells with open promoter, and the latter **(B)** by the fraction of cells with expression (based on single-cell RNA-seq data). Genes were binned by one of the indices and the distribution of the other is shown. The gene was considered as “expressed” in a cell if the corresponding UMI (unique molecular identifier) count was ≥ 1 . **(C)** Quantitative relationship between the bulk-cell gene expression and the cell population ratio of open promoter. Genes were binned based on the bulk-cell gene expression level (RNA-seq data).

Fig. 8 Relationship between chromatin accessibility and co-expression of *HXT3* and *HXT6*. **(A)** Expression levels of *HXT3* and *HXT6* in response to glucose concentration change. FPKM from bulk-cell RNA-seq data was taken as the expression level. **(B)** Change of the coupled chromatin statuses of *HXT3* and *HXT6* in response to different glucose concentration. Chromatin accessibility in promoters of

HXT3 and *HXT6* at the single-cell level is shown. Each line represents a molecule (i.e., cell). GpC site is labeled as rainbow-color dot, with methylation score from 0 (blue) to 1.0 (red). A promoter was defined as “open” (highlighted by red box) if the methylation scores of the including GpC sites had a median value greater than 0.5, and “closed” (highlighted by blue box) otherwise. Cells are shown in four groups that corresponded to four glucose concentrations. The cell fractions are also shown on the bar charts. The *P*-value was calculated by Chi-square test under the null hypothesis that alternative openness status of the two genes was independent with glucose concentration.

REFERENCES

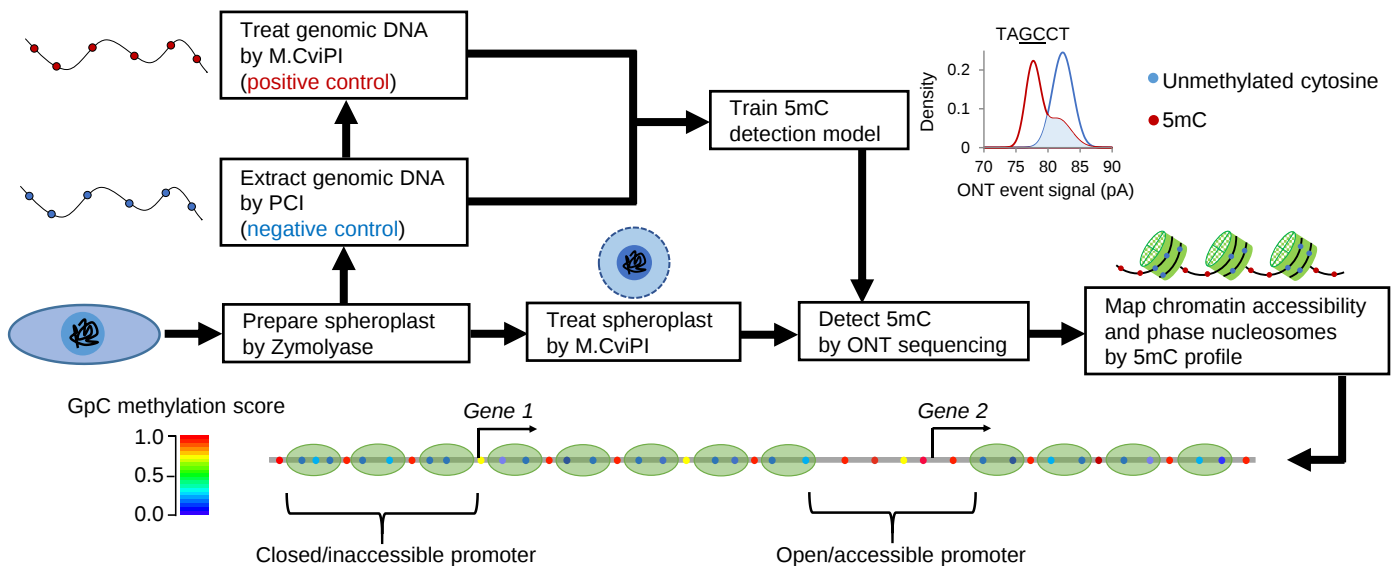
- Bell O, Tiwari VK, Thoma NH, Schubeler D. 2011. Determinants and dynamics of genome accessibility. *Nat Rev Genet* **12**: 554-564.
- Bianco S, Rodrigue S, Murphy BD, Gevry N. 2015. Global Mapping of Open Chromatin Regulatory Elements by Formaldehyde-Assisted Isolation of Regulatory Elements Followed by Sequencing (FAIRE-seq). *Methods Mol Biol* **1334**: 261-272.
- Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015a. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**: 21.29 21-29.
- Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. 2015b. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**: 486-490.
- Capuano F, Mulleder M, Kok R, Blom HJ, Ralser M. 2014. Cytosine DNA methylation is found in *Drosophila melanogaster* but absent in *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and other yeast species. *Anal Chem* **86**: 3697-3702.
- Chen W, Liu Y, Zhu S, Green CD, Wei G, Han JD. 2014. Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. *Nat Commun* **5**: 4909.
- Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC et al. 2018. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* **9**: 781.
- Cole HA, Howard BH, Clark DJ. 2011. The centromeric nucleosome of budding yeast is perfectly positioned and covers the entire centromere. *Proc Natl Acad Sci U S A* **108**: 12687-12692.
- Cui K, Zhao K. 2012. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. *Methods Mol Biol* **833**: 413-419.
- Dalal Y, Furuyama T, Vermaak D, Henikoff S. 2007. Structure, dynamics, and evolution of centromeric nucleosomes. *Proc Natl Acad Sci U S A* **104**: 15974-15981.

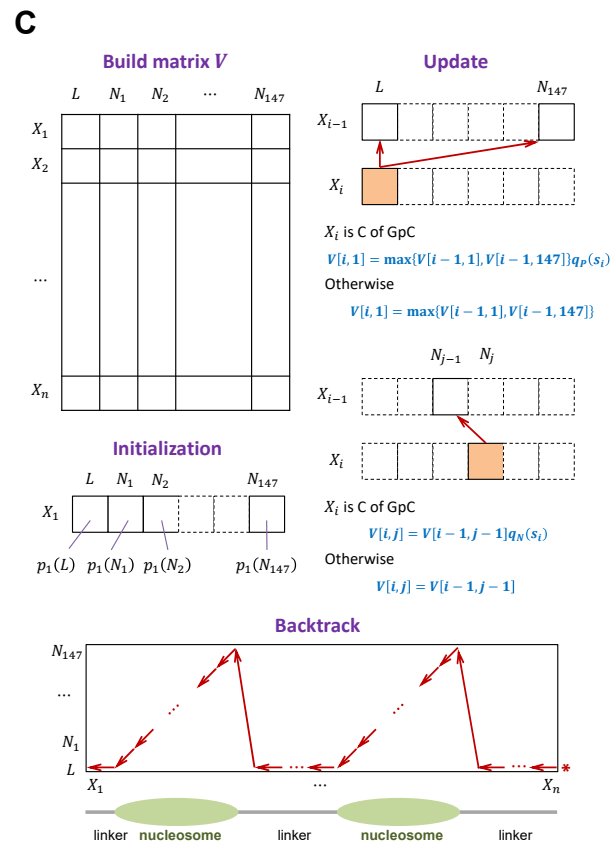
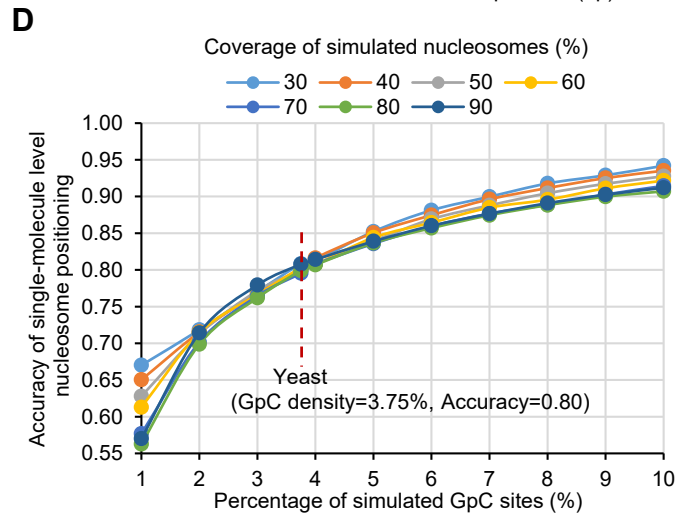
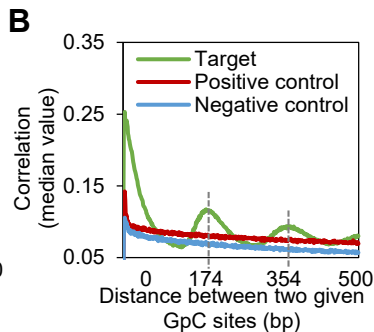
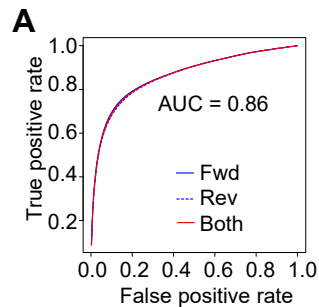
- 1 Grunberg S, Henikoff S, Hahn S, Zentner GE. 2016. Mediator binding to UASs is
2 broadly uncoupled from transcription and cooperative with TFIID recruitment
3 to promoters. *EMBO J* **35**: 2435-2446.
- 4 Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of
5 large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44-57.
- 6 Hughes AL, Rando OJ. 2014. Mechanisms underlying nucleosome positioning in vivo.
7 *Annu Rev Biophys* **43**: 41-63.
- 8 Ishii H, Kadonaga JT, Ren B. 2015. MPE-seq, a new method for the genome-wide
9 analysis of chromatin structure. *Proc Natl Acad Sci U S A* **112**: E3457-3465.
- 10 Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances
11 through genomics. *Nat Rev Genet* **10**: 161-172.
- 12 Jin W, Tang Q, Wan M, Cui K, Zhang Y, Ren G, Ni B, Sklar J, Przytycka TM, Childs R
13 et al. 2015. Genome-wide detection of DNase I hypersensitive sites in single
14 cells and FFPE tissue samples. *Nature* **528**: 142-146.
- 15 Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. 2012. Genome-wide
16 mapping of nucleosome positioning and DNA methylation within individual
17 DNA molecules. *Genome Res* **22**: 2497-2506.
- 18 Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, Kertesz M, Snyder M.
19 2014. Whole-genome haplotyping using long reads and statistical methods.
20 *Nat Biotechnol* **32**: 261-266.
- 21 Lai B, Gao W, Cui K, Xie W, Tang Q, Jin W, Hu G, Ni B, Zhao K. 2018. Principles of
22 nucleosome organization revealed by single-cell micrococcal nuclease
23 sequencing. *Nature* **562**: 281-285.
- 24 Lai WKM, Pugh BF. 2017. Understanding nucleosome dynamics and their links to
25 gene expression and DNA replication. *Nat Rev Mol Cell Biol* **18**: 548-562.
- 26 Li B, Carey M, Workman JL. 2007. The role of chromatin during transcription. *Cell* **128**:
27 707-719.
- 28 Li G, Levitus M, Bustamante C, Widom J. 2005. Rapid spontaneous accessibility of
29 nucleosomal DNA. *Nat Struct Mol Biol* **12**: 46-53.
- 30 Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler
31 transform. *Bioinformatics* **26**: 589-595.
- 32 Li L, Guo F, Gao Y, Ren Y, Yuan P, Yan L, Li R, Lian Y, Li J, Hu B et al. 2018.
33 Single-cell multi-omics sequencing of human early embryos. *Nat Cell Biol* **20**:
34 847-858.
- 35 Lipford JR, Bell SP. 2001. Nucleosomes positioned by ORC facilitate the initiation of
36 DNA replication. *Mol Cell* **7**: 21-30.
- 37 Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. 1997. Crystal
38 structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**:
39 251-260.
- 40 Needleman SB, Wunsch CD. 1970. A general method applicable to the search for
41 similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443-453.
- 42 Ozcan S, Johnston M. 1999. Function and regulation of yeast hexose transporters.
43 *Microbiol Mol Biol Rev* **63**: 554-569.

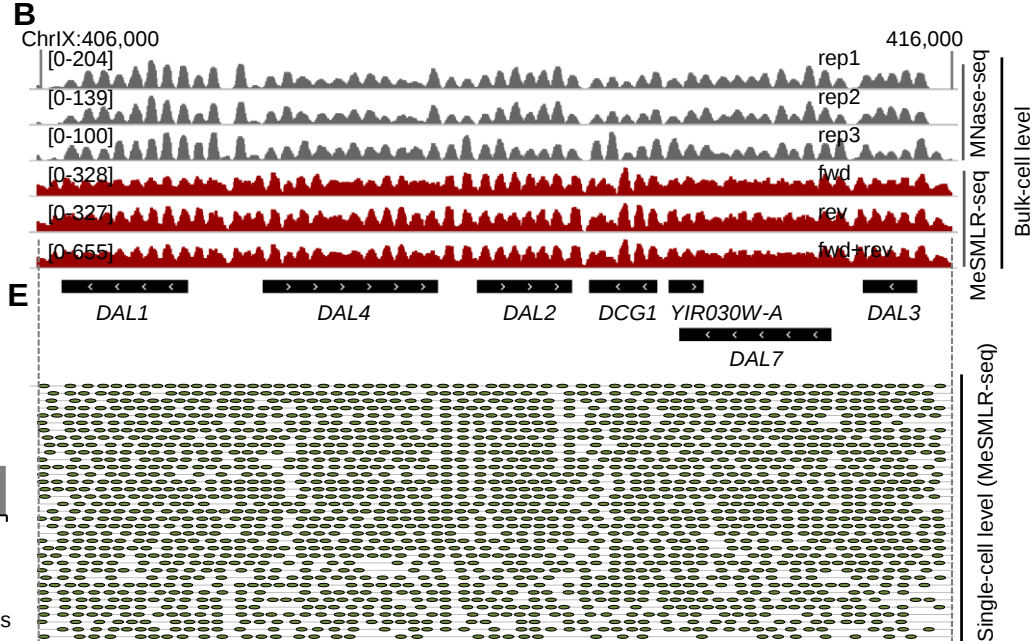
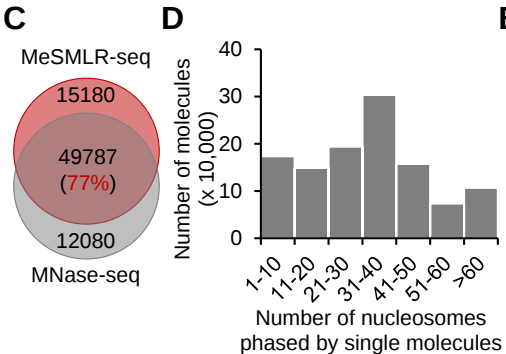
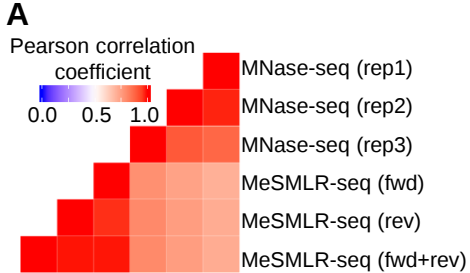
- 1 Payne A, Holmes N, Rakyan V, Loose M. 2018. BulkVis: a graphical viewer for Oxford
2 nanopore bulk FAST5 files. *Bioinformatics* doi:10.1093/bioinformatics/bty841.
- 3 Petesch SJ, Lis JT. 2008. Rapid, transcription-independent loss of nucleosomes over
4 a large chromatin domain at Hsp70 loci. *Cell* **134**: 74-84.
- 5 Pott S. 2017. Simultaneous measurement of chromatin accessibility, DNA methylation,
6 and nucleosome phasing in single cells. *Elife* **6**.
- 7 Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B.
8 2017. Mapping DNA methylation with high-throughput nanopore sequencing.
9 *Nat Methods* **14**: 411-413.
- 10 Rando OJ, Winston F. 2012. Chromatin and transcription in yeast. *Genetics* **190**:
11 351-387.
- 12 Schep AN, Buenrostro JD, Denny SK, Schwartz K, Sherlock G, Greenleaf WJ. 2015.
13 Structured nucleosome fingerprints enable high-resolution mapping of
14 chromatin architecture within regulatory regions. *Genome Res* **25**: 1757-1770.
- 15 Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. 2008.
16 Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**:
17 887-898.
- 18 Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon-intron
19 structure. *Nat Struct Mol Biol* **16**: 990-995.
- 20 Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting
21 DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**:
22 407-410.
- 23 Small EC, Xi L, Wang JP, Widom J, Licht JD. 2014. Single-cell nucleosome mapping
24 reveals the molecular basis of gene expression heterogeneity. *Proc Natl Acad*
25 *Sci U S A* **111**: E2462-2471.
- 26 Song L, Crawford GE. 2010. DNase-seq: a high-resolution technique for mapping
27 active gene regulatory elements across the genome from mammalian cells.
28 *Cold Spring Harb Protoc* **2010**: pdb prot5384.
- 29 Tilgner H, Jahanbani F, Gupta I, Collier P, Wei E, Rasmussen M, Snyder M. 2018.
30 Microfluidic isoform sequencing shows widespread splicing coordination in the
31 human transcriptome. *Genome Res* **28**: 231-242.
- 32 Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R.
33 2009. Nucleosome positioning as a determinant of exon recognition. *Nat*
34 *Struct Mol Biol* **16**: 996-1001.
- 35 Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL,
36 Wold BJ, Pachter L. 2010. Transcript assembly and quantification by
37 RNA-Seq reveals unannotated transcripts and isoform switching during cell
38 differentiation. *Nat Biotechnol* **28**: 511-515.
- 39 Voss TC, Hager GL. 2014. Dynamic regulation of transcriptional states by chromatin
40 and transcription factors. *Nat Rev Genet* **15**: 69-81.
- 41 Wal M, Pugh BF. 2012. Genome-wide mapping of nucleosome positions in yeast
42 using high-resolution MNase ChIP-Seq. *Methods Enzymol* **513**: 233-250.

- 1 Weiner A, Hsieh TH, Appleboim A, Chen HV, Rahat A, Amit I, Rando OJ, Friedman N.
- 2 2015. High-resolution chromatin dynamics during a yeast stress response.
- 3 *Mol Cell* **58**: 371-386.
- 4 Zhong J, Luo K, Winter PS, Crawford GE, Iversen ES, Hartemink AJ. 2016. Mapping
- 5 nucleosome positions using DNase-seq. *Genome Res* **26**: 351-364.

6

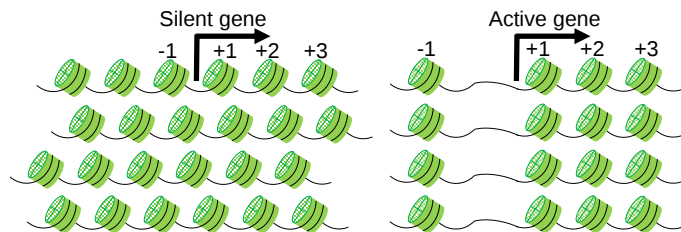
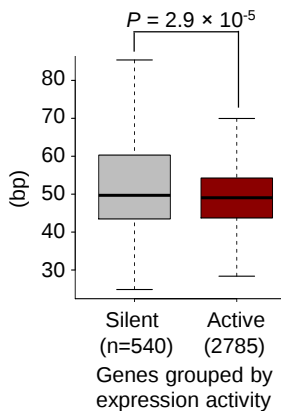
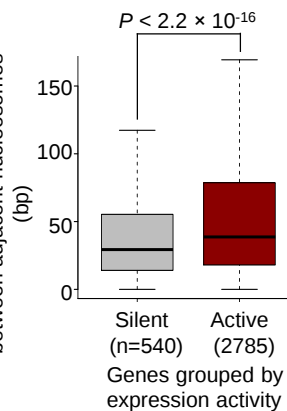




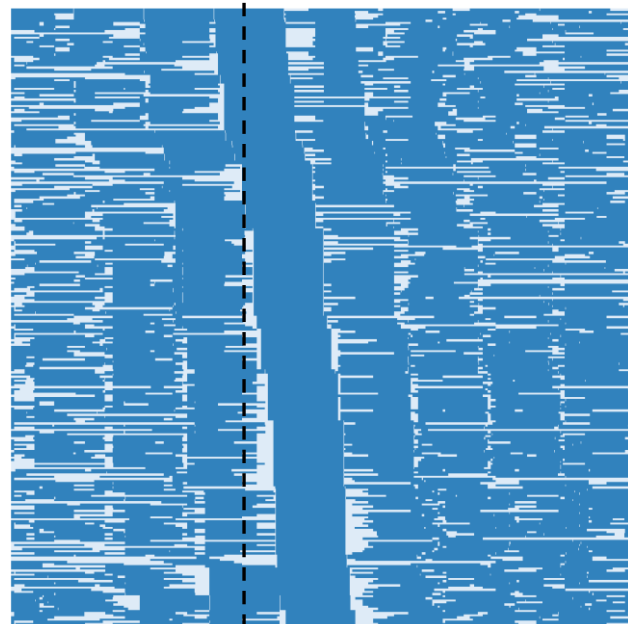


A

Nucleosome organization pattern surrounding TSS

**B**SD of the distance between
+1 nucleosome and TSS
(bp)**C**Variation of the distance
between adjacent nucleosomes
(bp)**D**TSS → *AUA1* (silent gene, FPKM = 0)

Poorly-positioned



High heterogeneity of +1 nucleosome positioning
High uniformity of nucleosome spacing

Bulk-cell level

Nucleosome positioning at the single-cell level
(sorted by the distance of +1 nucleosome to TSS)**E**TSS ← *EMW1* (active gene, FPKM = 77)

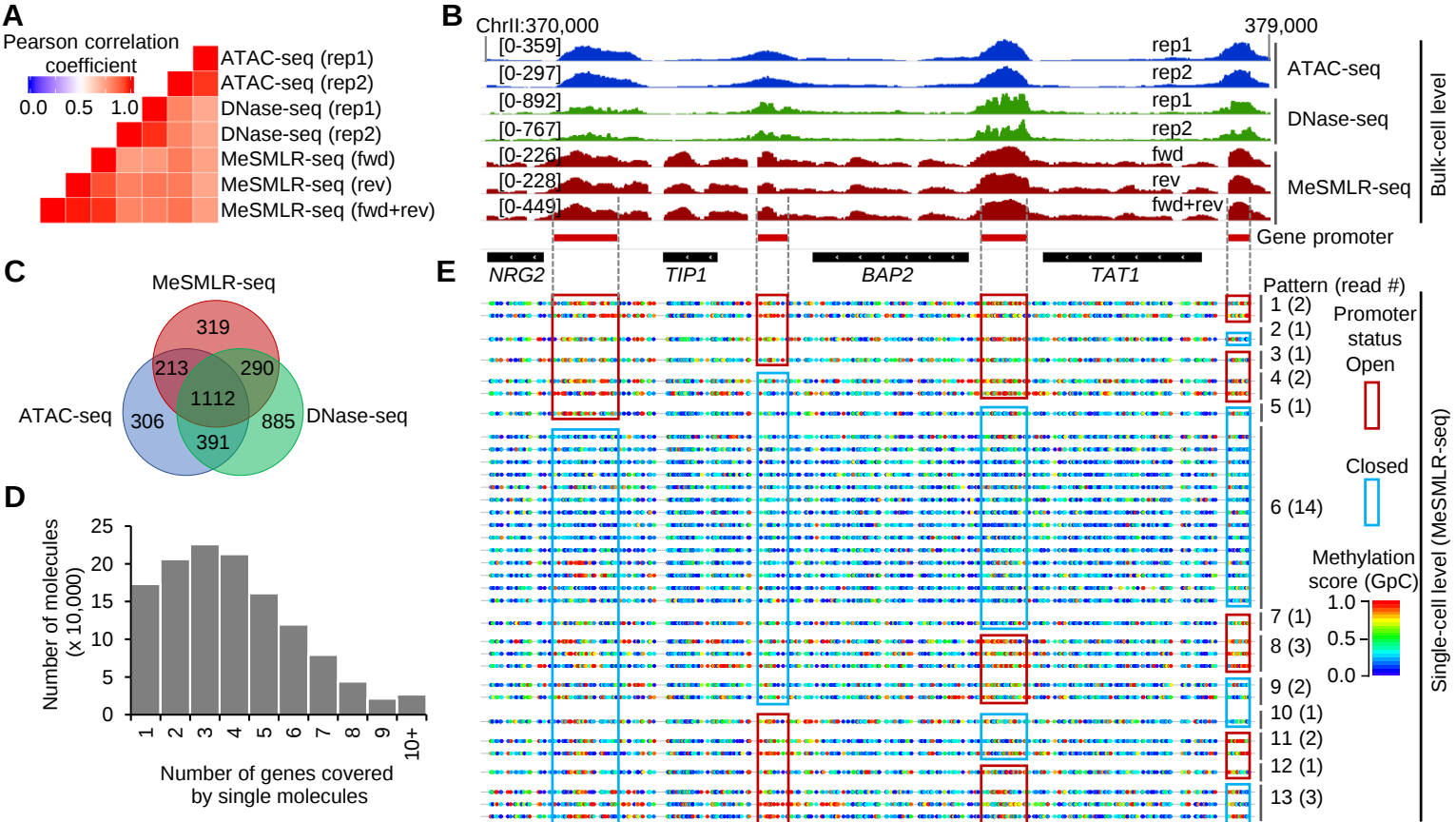
Well-positioned



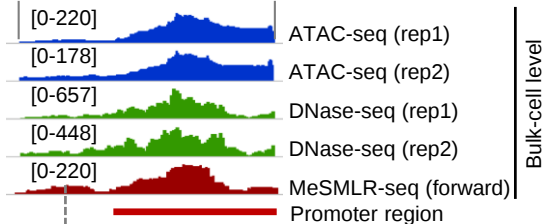
Low heterogeneity of +1 nucleosome positioning
Low uniformity of nucleosome spacing

Bulk-cell level

Nucleosome positioning at the single-cell level
(sorted by the distance of +1 nucleosome to TSS)



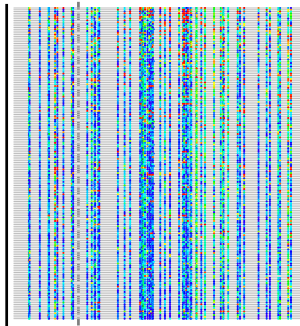
ChrXVI:66,400 67,550



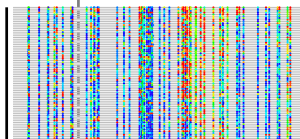
Methylation score (GpC)



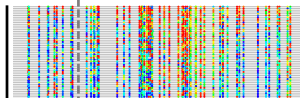
Cluster 1
(n=176)



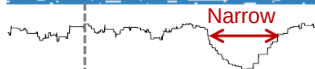
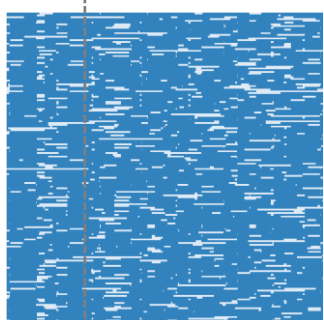
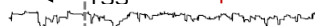
Cluster 2
(n=75)



Cluster 3
(n=53)



Chromatin accessibility

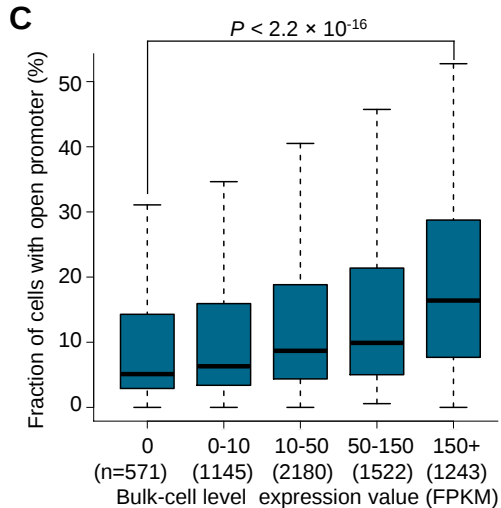
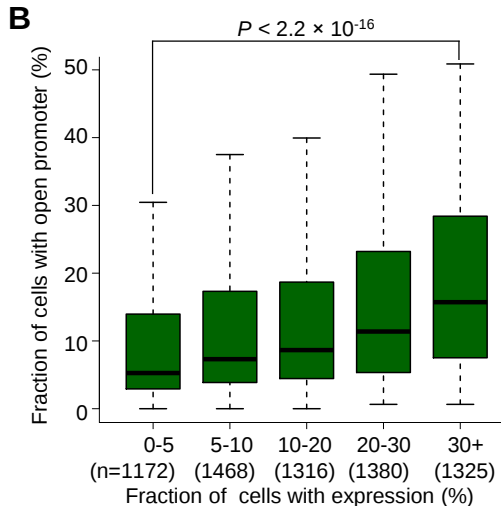
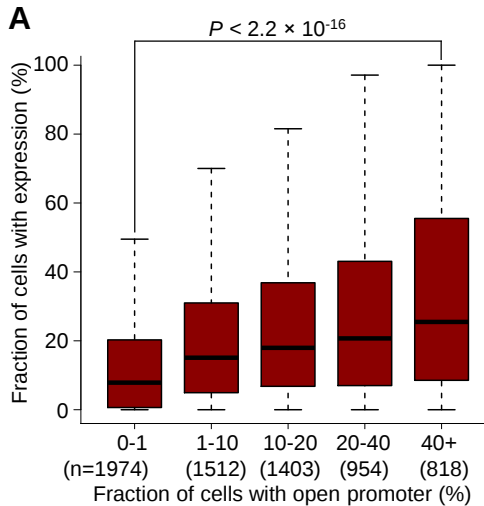


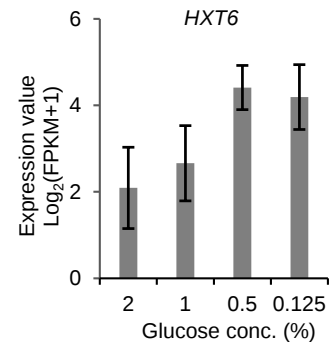
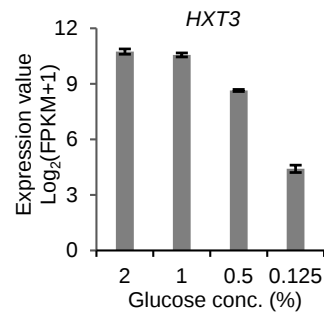
Nucleosome positioning

Single-cell level

Single-cell level

Single-cell level



A**B**