nature

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41586-018-0567-3

In the format provided by the authors and unedited.

Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing

Binbin Lai[1], Weiwu Gao[1,2], Kairong Cui[1], Wanli Xie[1,3], Qingsong Tang[1], Wenfei Jin[4], Gangqing Hu[1], Bing Ni[2] & Keji Zhao[1]*

[1]Laboratory of Epigenome Biology, Systems Biology Center, National Heart, Lung and Blood Institute, NIH, Bethesda, MD, USA. [2]Department of Pathophysiology and High Altitude Pathology, Third Military Medical University, Chongqing, China. [3]Institute of Anesthesiology and Critical Care Medicine, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. [4]Department of Biology, South University of Science and Technology of China, Shenzhen, China. *e-mail: zhaok@nhlbi.nih.gov

© 2018 Springer Nature Limited. All rights reserved.

Methods

Cell lines

NIH3T3 cells and mESCs were obtained from ATCC (ATCC® CRL-1658™ and ATCC® SCRC-1036™). In addition to the authentic commercial source, our own data (RNA-seq and MNase-seq) validate the identity of these cells.

Generation of scMNase-seq libraries

NIH3T3 cells were maintained in culture medium DMEM (Invitrogen cat# 10566-016) supplemented with 10% FBS (Sigma cat# F4135-500ML) and 100U/ml Penicillin-Streptomycin (Invitrogen cat# 15140-122). mESCs were cultured as described previously[30]. Mouse naïve CD4 T cells were isolated from B6 mice as described[31]. Single live cells were sorted by flow cytometer and deposited directly into each tubes of a PCR strip-tube, which containing 32 μl of cell lysis buffer (10mM Tris-HCl, pH 7.5, 10mM NaCl, 3mM $CaCl_2$, 0.1% Triton X-100). For growing AN3-12 haploid ES cells[32], we use cell culture-treated dishes coated with gelatin. Cells were cultured in ES cell media containing LIF (ESGRO, Millipore ESG1107) + 1.5μM GSK3i (CHIR9901, Stemgent Inc. S1263). Haploid cells were FACS sorted using Hoechst 33342 (Sigma B2261).

For Micrococcal Nuclease (MNase) digestion, 8 μl diluted MNase (1:3000 in lysis buffer) was added to the sorted single cells and incubated at 37°C for 5 min. The reaction was stopped by addition of 80 μl of Stop Buffer (10mM Tris-HCl, pH 7.5, 10mM NaCl, 0.15% SDS, 10mM EGTA) and 1 μl of 20mg/ml Proteinase K. Following incubation at 55°C for 1 hour, DNA was purified by Phenol-chloroform extraction and precipitation with ethanol in the presence of glycogen. Universal adaptors were then ligated to the DNA ends as described[33]. The libraries were amplified using indexed primers with the PCR condition: 98°C, 10"; 67°C, 30"; 72°C, 30" for 23 cycles. The fragments between 160bp to 300bp were isolated on E-gel and subjected to paired-end sequencing on Illumina HiSeq2500. The sequencing reads were mapped to the mouse genome mm9.

Datasets

We obtained sequences for 48 NIH3T3, 198 mouse ESC, and 278 naïve CD4 T cell scMNase-seq libraries, which are summarized in Supplementary Table 1. In addition, we generated three groups of NIH3T3 scMNase-seq libraries (10 single cells for each group) with different MNase concentrations for titration purpose, and 12 haploid mouse ESC scMNase-seq

libraries. We also generated bulk-cell MNase-seq for mouse ESCs and naïve CD4 T cells. DNase-seq, ChIP-seq, RNA-seq, scDNase-seq and Drop-seq datasets used in this study were downloaded from public databases, which are summarized in Supplementary Table 2.

Peak calling for ChIP-seq and DNase-seq and gene expression measurements

Reads of ChIP-seq and RNA-seq were mapped to genome (mm9) using Bowtie 2[34]. Reads with MAPQ ≤ 10 or redundant reads that mapped to the same location and orientation were removed. Peaks for histone modifications were identified using MACS2[35] with parameter "--broad". Peaks for CTCF and p300 ChIP-seq were identified using MACS2 with default parameters. DHSs peaks were downloaded directly from mouse ENCODE project database, which were called using ENCODE protocol. Gene expression level was measured by calculating reads per kilobase per million mapped reads (RPKM) for each gene. Active genes were defined as genes with RPKM ≥1.

Nucleosome and Subnucleosome-sized particle position identification

Paired-end sequencing reads for scMNase-seq and bulk-cell MNase-seq were mapped to mouse genome (mm9) using Bowtie 2. Discordantly aligned pairs, reads with mapping quality (MAPQ) ≤10, and redundant reads were removed from further analysis. We selected the fragments with range of 140-180 bp as canonical nucleosomes and the short fragments (≤80 bp) as subnucleosome-sized particles for further analysis. The midpoints of nucleosomes and subnucleosomes-sized particles were defined as their positions. The position distributions of defined nucleosomes and subnucleosome-sized particles relative to TSSs and CTCF binding sites are consistent between pooled scMNase-seq and bulk-cell MNase-seq, and are also consistent with reported before[5,36].

Nucleosome and Subnucleosome-sized particle occupancy relative to TSSs, CTCF binding sites, and DHSs from pooled scMNase-seq

We calculated tag density for nucleosome and subnucleosome-sized particles from pooled scMNase-seq and bulk-cell MNase-seq in 10 bp genomic bins around TSSs, CTCF binding sites, and DHSs by normalizing tag counts to total mapped tag counts (tag counts per kilobase per million mapped tags) and drew average profiles around TSSs and CTCF binding sites. For each DHS, we calculated the fraction of nucleosomes occupied within 200 bp window around DHS

from all nucleosomes within 400 bp window around DHS to measure nucleosome occupancy or depletion level at DHS for pooled scMNase-seq or bulk-cell MNase-seq. We calculated genome-wide correlation for the fraction of nucleosomes occupied at DHS centers between pooled scMNase-seq and bulk-cell MNase-seq. For each cell type, we selected top 10,000 DHSs with highest DNase I tag density for this analysis. DHSs with low nucleosome coverage (<10 nucleosomes within 400 bp window around DHS) were further excluded for calculation. We also calculated the tag density for subnucleosome-sized particles within 2 kb window centered around DHSs for pooled scMNase-seq and bulk-cell MNase-seq, and calculated genome-wide correlation for the subnucleosome-sized particle density at all identified DHSs between pooled scMNase-seq and bulk-cell MNase-seq.

Nucleosome positioning stringency and reference nucleosome position

For NIH3T3 cells, we quantified the consistency of nucleosome positioning by calculating positioning scores for pooled scMNase-seq or bulk-cell MNase-seq datasets using a similar approach by Gaffney et al.[7] We calculated the positioning score at a genomic site k as the fraction of midpoints in a 201 bp window surrounding k that are within 15 bp of k: $\left(\sum_{j=k-15}^{k+15} x_j\right)/\left(\sum_{j=k-100}^{k+100} x_j\right)$, where x_j is the number of nucleosome midpoints at site j. We defined reference nucleosome position as the site with maximum positioning score within the ±100 bp region. We identified reference nucleosome positions along the genome by iteratively identifying the site with maximum positioning score and removing other sites within ±100 bp surrounding the selected site. Reference positions with less than 10 cells that have nucleosomes within ±100 bp region were excluded from further analysis. If a nucleosome position is within 15 bp of a reference position, we called the nucleosome is positioned. We defined -1 and +1 nucleosome position (-1 N, +1 N) relative to TSS as the reference positions located within corresponding peaks in nucleosome occupancy profile relative to TSS.

Aggregated nucleosome profiles from pooled scMNase-seq and bulk-cell MNase-seq

We used center-weighted occupancy score to plot aggregated nucleosome profiles at single base resolution. Specifically, we defined the nucleosome center positioning (NCP) score S_k at position k, as the normalized nucleosome count with center at position k. We defined center-weighted occupancy score at position k as $\sum_{j=-73}^{73} S_{k+j} w_j$, where w_j is the Gaussian weight equal to $e^{-(j/20)^2/2}$. We defined similarity score between aggregated nucleosome profiles from pooled

scMNase-seq and bulk-cell MNase-seq within a 400 bp region centered around TSSs and DHSs as Pearson correlation coefficient between two vectors of center-weighted occupancy score within the region.

Mathematics model for quantifying nucleosome spacing uniformness

Suppose we have n cells with ploidy m and we consider distance from any two nucleosomes from the n cells. We draw histogram in density scale (smoothed density plot) of nucleosome-to-nucleosome distance within a range (d_1, d_2) with a bin size equal to 1 bp. The shape of the histogram depends on the probability of observing a distance d between two nucleosomes, denoted as $p(d)$. We have $p(d) = p_1 * A(d) + (1 - p_1) * B(d)$, where p_1 is the probability of assigning two nucleosomes from the same allele, $A(d)$ is the probability of observing a distance d between two nucleosomes from the same allele, and $B(d)$ represents the probability of observing a distance d between two nucleosomes that are not from the same allele. Note that p_1 can be estimated as $1/(n * m)$. Therefore, we have $p(d) = \frac{A(d)}{n*m} + \left(1 - \frac{1}{n*m}\right) * B(d)$.

We first consider a fixed step of nucleosome distance, e.g. adjacent nucleosomes (step=1) or nucleosomes spanning 1 (step = 2), 2 (step = 3), …, *k* nucleosomes. The probability $A(d)$ follows Gaussian distribution: $A(d) \sim N(\mu, \sigma_A)$, where μ corresponds to the most common distance. On the other hand, $B(d) = \sum_j A(d - j)C(j)$, where j is the relative position of assigned nucleosome from other alleles/cells with respect to the nucleosome from the same allele, and $C(j)$ is the probability of observing position difference j by assigning a nucleosome from a different allele/cell relative to the nucleosome from the same allele. $C(j)$ should also follows Gaussian distribution: $C(j) \sim N(0, \sigma_C)$. σ_C indicates the standard deviation of nucleosomes differences across alleles/cells. If nucleosomes have high positioning stringency, σ_C is low. For example, $\sigma_C = 0$ indicates nucleosomes are perfectly positioned across cells. If nucleosomes have low positioning stringency, σ_C is high. As $B(d)$ is the convolution of two Gaussian function $A(d)$ and $C(j)$, $B(d)$ also follows Gaussian distribution and have parameters: $B(d) \sim N(\mu + 0, \sqrt{{\sigma_A}^2 + {\sigma_C}^2})$. From $p(d) = \frac{A(d)}{m*n} + \left(1 - \frac{1}{m*n}\right) * B(d)$, we have that $p(d)$ also has a Gaussian curve with mean = μ and standard deviation ${\sigma}^2 = \left(\frac{1}{mn}\right)^2 * {\sigma_A}^2 + \left(1 - \frac{1}{mn}\right)^2 * ({\sigma_A}^2 + {\sigma_C}^2) = \left(\left(\frac{1}{mn}\right)^2 + \left(1 - \frac{1}{mn}\right)^2\right) * {\sigma_A}^2 + \left(1 - \frac{1}{mn}\right)^2 * {\sigma_C}^2$.

Now we consider all nucleosome pairs with distance <2000 bp. Given that a normal adjacent nucleosome distance is ~185 bp for the studied cells, we get a mixture of Gaussian distributions that describe distances between adjacent nucleosomes, nucleosomes spanning 1, 2, …, and 9 nucleosomes. The histogram displays a periodic Gaussian curve with phase step corresponding to nucleosome step. Moreover, we would expect the more the step, the larger σ_A is. Therefore, we observe the declined peak shape as the distance increases.

Our goal is to measure the uniformness of the nucleosome space within specific regions. This can be directly indicated by standard deviation σ_A. However, suppose we don't use haploid single cell, we cannot directly get Gaussian curve with σ_A. Instead, what we observe is Gaussian curve with standard deviation

$$\sigma^2 = \left(\left(\frac{1}{mn}\right)^2 + \left(1 - \frac{1}{mn}\right)^2\right) * {\sigma_A}^2 + \left(1 - \frac{1}{mn}\right)^2 * {\sigma_C}^2$$

for each peak.

For simplicity, for each step i from a periodic Gaussian curve, we introduce a relative peak height $h(i)$ to measure the variation for step i. Relative peak height is defined as $h = c - (l + r)/2$, where c is the density value at the peak summit in the density profile, l and r represent the values at the left and right troughs for the peak respectively. Intuitively, high relative peak height corresponds to sharp peak which has a small standard deviation, while low relative peak height corresponds to gentle peak which has a large standard deviation.

We describe distance uniformity in a comparative manner between active chromatin regions (e.g. active promoters, DHS regions) and silent chromatin regions (e.g. silent promoters, non-DHS regions) with the same dataset. Different datasets display different histogram for the same regions, and cell number affects the distribution as we discuss below.

We discuss the effect of cell number on distribution of nucleosome-to-nucleosome distance. As Gaussian curve shape is determined by standard deviation, we discuss the standard deviation σ^2 as the function of cell number n. We have $\sigma^2 = \left(\left(\frac{1}{mn}\right)^2 + \left(1 - \frac{1}{mn}\right)^2\right) * {\sigma_A}^2 + \left(1 - \frac{1}{mn}\right)^2 * {\sigma_C}^2$. Because we want to use σ^2 to estimate ${\sigma_A}^2$, the term $\left(1 - \frac{1}{mn}\right)^2 * {\sigma_C}^2$ is an interference term, which we denote as I. From the formula, the interference term gets larger when n gets larger. When $m = 1, n = 1$, (i.e. haploid single cell), we have $I = 0$. When $m = 2, n = 1$, (diploid single cell), we have $I = 3/4 * {\sigma_C}^2$. When n get large enough, $I \approx {\sigma_C}^2$.

Next, we consider the comparison of distance uniformity between active chromatin region and silent chromatin region. We have variation of distance for silent chromatin as

${\sigma^2}_{silent} = \left(\left(\frac{1}{mn}\right)^2 + \left(1-\frac{1}{mn}\right)^2\right) * {\sigma_A}^2{}_{silent} + \left(1-\frac{1}{mn}\right)^2 * {\sigma_C}^2{}_{silent}$, and variation of distance for active chromatin as ${\sigma^2}_{active} = \left(\left(\frac{1}{mn}\right)^2 + \left(1-\frac{1}{mn}\right)^2\right) * {\sigma_A}^2{}_{active} + \left(1-\frac{1}{mn}\right)^2 * {\sigma_C}^2{}_{active}$.

Then we have

$$ {\sigma^2}_{silent} - {\sigma^2}_{active} = \left(\left(\frac{1}{mn}\right)^2 + \left(1-\frac{1}{mn}\right)^2\right) * \left({\sigma_A}^2{}_{silent} - {\sigma_A}^2{}_{active}\right) + \left(1-\frac{1}{mn}\right)^2 * \left({\sigma_C}^2{}_{silent} - {\sigma_C}^2{}_{active}\right) $$

and

$$ {\sigma_A}^2{}_{silent} - {\sigma_A}^2{}_{active} = \frac{1}{\left(\frac{1}{mn}\right)^2 + \left(1-\frac{1}{mn}\right)^2}\left({\sigma^2}_{silent} - {\sigma^2}_{active}\right) - \frac{\left(1-\frac{1}{mn}\right)^2}{\left(\frac{1}{mn}\right)^2 + \left(1-\frac{1}{mn}\right)^2}\left({\sigma_C}^2{}_{silent} - {\sigma_C}^2{}_{active}\right) \quad (1) $$

Empirically, nucleosome variation in silent chromatin is larger than in active chromatin, which has also been supported by the data in this study. Thus, we have:

$$ {\sigma_C}^2{}_{silent} > {\sigma_C}^2{}_{active} \quad (2) $$

Using single cell datasets, we have observed that ${\sigma^2}_{silent} < {\sigma^2}_{active}$, by comparing relative peak height. Together with (1), (2), we have:

$$ {\sigma_A}^2{}_{silent} - {\sigma_A}^2{}_{active} < 0, $$

i.e., variation of nucleosome space from single allele for silent chromatin regions is smaller than active chromatin regions.

However, if ${\sigma^2}_{silent} > {\sigma^2}_{active}$, as what we observed from bulk-cell datasets, we cannot easily determine whether ${\sigma_A}^2{}_{silent} < {\sigma_A}^2{}_{active}$ unless we can get a precise value of ${\sigma^2}_{silent}$, ${\sigma^2}_{active}$, ${\sigma_C}^2{}_{silent}$, ${\sigma_C}^2{}_{active}$, which cannot be obtained from relative peak height.

Next, we discuss the condition in respect to cell number under which we can observe ${\sigma^2}_{silent} < {\sigma^2}_{active}$.

For simplicity, we suppose cells are diploid, as is the case for most cells in our study, we have m = 2. We suppose: ${\sigma^2}_{silent} - {\sigma^2}_{active} < 0$.

Then, we have

$$ \left(\left(\frac{1}{2n}\right)^2 + \left(1-\frac{1}{2n}\right)^2\right) * \left({\sigma_A}^2{}_{silent} - {\sigma_A}^2{}_{active}\right) $$

$$+\left(1-\frac{1}{2n}\right)^2 * \left({\sigma_C}^2{}_{silent} - {\sigma_C}^2{}_{active}\right) < 0$$

Then, we have

$$\frac{1}{(2n-1)^2} > \frac{{\sigma_C}^2{}_{silent} - {\sigma_C}^2{}_{active}}{{\sigma_A}^2{}_{active} - {\sigma_A}^2{}_{silent}} - 1 \quad (3)$$

If ${\sigma_C}^2{}_{silent} - {\sigma_C}^2{}_{active} \leq {\sigma_A}^2{}_{active} - {\sigma_A}^2{}_{silent}$, then (4) is true for any n >0. However, this might not true for nucleosomes positioning. We know nucleosome spacing variance reflects the variance of linker DNA, which is constrained by chromatin structure, while the nucleosome positions across different cells may be quite random in silent chromatin. Thus, we hypothesize that ${\sigma_C}^2{}_{silent} - {\sigma_C}^2{}_{active} > {\sigma_A}^2{}_{active} - {\sigma_A}^2{}_{silent}$. Then, we further have:

$$n < \frac{1}{2}\left(1 + \sqrt{\frac{{\sigma_A}^2{}_{active} - {\sigma_A}^2{}_{silent}}{({\sigma_C}^2{}_{silent} - {\sigma_C}^2{}_{active}) - ({\sigma_A}^2{}_{active} - {\sigma_A}^2{}_{silent})}}\right)$$

We denote $\frac{1}{2}\left(1 + \sqrt{\frac{{\sigma_A}^2{}_{active} - {\sigma_A}^2{}_{silent}}{({\sigma_C}^2{}_{silent} - {\sigma_C}^2{}_{active}) - ({\sigma_A}^2{}_{active} - {\sigma_A}^2{}_{silent})}}\right)$ as *s,* which is a finite value. When $n < s$, we can observe ${\sigma^2}_{silent} < {\sigma^2}_{active}$, when n> s, we will see ${\sigma^2}_{silent} > {\sigma^2}_{active}$. As long as $s > 1$, we can observe higher relative peak height in silent chromatin regions than active chromatin regions using single cells, which cannot be observed using bulk cells due to $n > s$.

Next, we describe the way how we select nucleosome pairs within a region for drawing the density plot and whether it affects the distance distribution. Suppose the region size is L_R, and the distance range is (d_1, d_2). We suppose nucleosomes can be located at any positions with the same probability along the genome. When selecting nucleosome pairs, we require at least one nucleosome within the region. The probability of observing the distance d between any selected nucleosome pairs would be the same, which we denote as $P_1(d) = c$, where $c = \frac{1}{d_2 - d_1}$.

Previously, we have described the probability of nucleosome-to-nucleosome distance d based on the intrinsic of nucleosome spacing distribution and nucleosome positioning variance across cells or alleles, which we denote as $P_2(d)$. Thus, we have probability of observing the nucleosome-to-nucleosome distance d by selecting any two nucleosomes from regions follows: $p(d) = P_1(d) * P_2(d) = c * P_2(d)$, which is independent on region size L_R, suggesting the region length does not affect the density plot.

To determine the summits and troughs of peaks for density plot from a relatively low sample size (nucleosome pairs), plot line was smoothed using *smooth.spline* function in R

package if needed. For fair comparison, the same smoothing parameter was used for two conditions in comparison.

Quantifying nucleosome spacing uniformity in various datasets and chromatin regions

We compared the nucleosome spacing profiles and relative peak heights from 1 million randomly selected 1,000-bp genomic regions based on one NIH3T3 single cell or pooled 2, 4, 8, 16 and 48 single cells, respectively. We observed that the space phasing and peak heights gradually decline as we mixed more cells (**Extended Data Fig. 3b**). We down-sampled NIH3T3 scMNase-seq libraries to 1/2, 1/3 and 1/4 to increase the data sparsity level, and compared the nucleosome space phasing and relative peak heights from 1 million randomly selected 1,000-bp genomic regions based on full, 1/2, 1/3, and 1/4 sized libraries. The result shows that reducing the library size does not change the nucleosome spacing pattern and relative peak heights (**Extended Data Fig. 3c**). We measured space uniformity for regions containing tightly positioned (positioning score, PS $\geq$ 6.5), moderately positioned (4.5 $\leq$ PS $<$ 6.5), or loosely positioned (PS < 4.5) nucleosomes based on NIH3T3 single cells, pooled single cells and bulk cells. We found that nucleosome space phasing dramatically decreases for the pooled single cells and bulk cells in regions containing loosely positioned nucleosomes (**Extended Data Fig. 3d**). We compared space uniformity between active and silent promoter regions, and DHS and non-DHS regions based on single cells for all three cell types (**Fig. 2a-d, Extended Data Fig. 4a-b**). Promoter regions were defined as $\pm$2kb from TSSs. DHS regions were defined as $\pm$2kb from DHS centers. Non-DHS regions were defined as genomic regions not covered by DHS peak and at least 2kb away from any DHS centers. We separated NIH3T3 cells into two groups based on their genomic coverage and examined relative peak heights in DHS and non-DHS regions for both low and high coverage cells respectively. Both low and high coverage cells reveal the same conclusion that relative peak height in non-DHS regions is higher than that in DHS regions (**Extended Data Fig.4c**). We examined nucleosome spacing uniformity in DHS and non-DHS regions for haploid mESCs and in haploid chromosome X from diploid mESCs. The results are consistent with that obtained from diploid single cells: nucleosome uniformity is higher in non-DHS regions than DHS regions (**Extended Data Fig. 4e-g**). We obtained additional scMNase-seq libraries for NIH3T3 cells with two different MNase concentrations (0.6 unit and 2.4 unit per million cells) and examined relative peak heights in DHS and non-DHS regions for both two conditions. Both conditions reveal the same conclusion that relative peak height in non-DHS regions is higher than in DHS regions (**Extended Data Fig. 4l-m**). We also measured space uniformity in genomic regions marked by histone modifications based on

NIH3T3 single cells cells (**Extended Data Fig. 4n-o**). Not marked regions were defined as intervals of merged histone modification peaks.

Nucleosome variance

Nucleosome variance is defined as distance between two overlapping nucleosomes within a cell (across alleles) or across cells. We measure nucleosome variance within genomic regions by averaging all the variances within the range of 3–82 bp between all nucleosome pairs from those regions. For fair comparison between within-cell and across-cell variances, we didn't include zeros or small distances (0–2 bp) because nucleosomes with identical positions in each single-cell dataset had been removed. In addition, we didn't include distances larger than half of the nucleosome-to-nucleosome distance. We set an upper bound as 82 bp for convenience because most nucleosome-to-nucleosome distances are larger than 165 bp. Excluding small distances does not affect the comparison of variance between different regions. For example, using a range of 0-82 bp revealed the same trend of variance change from DHS center to distal regions with the use of range 3-82 bp (**Extended Data Fig. 5e**).

Synchronized shift score

We measured consistency of two adjacent nucleosomes' shifts from reference positions described as above by introducing synchronized shift score. For a given pair of adjacent nucleosome reference positions, we collected all non-overlapping nucleosome pairs from the same single cells that meet following criteria as valid nucleosome pairs: two nucleosomes are within ±100 bp from each of reference positions separately; at least one nucleosome is not within ±15 bp from reference positions. Suppose d_1 and d_2 are the distances of two nucleosomes from their corresponding reference positions, two nucleosomes are defined as synchronized-shift pair if $|d_1 - d_2| \leq 30$ and both shifts have the same orientation. We defined nucleosome shift score for a group of reference nucleosome pairs that share a specific feature, e.g. within histone modification peaks (**Extended Data Fig. 6b**) or having the same space (**Extended Data Fig. 6c**). Then, we calculated the fraction of valid nucleosome pairs that are synchronized shifted, denoted as S. Meanwhile, we randomly collected 10 million reference pairs that are at least 1 Mb away from each other and calculated the same value as a background, denoted as S'. Synchronized shift score for the feature was defined as the ratio of

observed fraction to the background as S/S'. The analysis was performed using NIH3T3 cell datasets.

Nucleosome spacing flanking DHSs

We calculated nucleosome space flanking DHS for strong DHSs (top 25% with highest DNase I density) and weak DHSs (bottom 25%) by calculating distances (from 100 to 2000 bp) of nucleosomes from the +1 nucleosomes within the first nucleosome-occupancy peak (155 to 200 bp from DHS) on the same side within single cells. Gaussian smoothed density plot was used to determine summits and troughs for each peak and average space between two peaks (**Extended Data Fig. 6e-f**). For comparison, we randomly selected the same number of genomic sites in non-DHS regions as pseudo-DHSs and used the same method to calculate nucleosome space. The analysis was performed using NIH3T3 cell datasets.

Bimodal nucleosome spacing across DHS and its links to cell-to-cell variation in nucleosome positioning and chromatin accessibility

We profiled distribution of distances (from 100 to 500 bp) between nucleosome pairs across DHS center with one nucleosome located within the first nucleosome-occupancy peak (-200 to -155 bp from DHS) on the opposite side. The distance distribution reveals two distinct peaks corresponding to narrow (summit at ~190 bp) or wide (summit at ~300 bp) space. We defined two ranges of nucleosome space based on peaks shape and summit position; narrow space ranges from 150 to 230 bp and wide space ranges from 260 to 340 bp, and assign each nucleosome pair as having either wide or narrow space for each cell and each DHS. We separated DHSs into 5 subgroups based on the fraction of wide space. The DHSs with total space number <5 were not included in this analysis. We calculated DNase-seq tag density, pooled scMNase-seq subnucleosome-sized particle density, cell-to-cell nucleosome variation and accessibility variation for each subgroup of DHSs. Mann-Whitney rank test was used to determine the significance of value differences between the first subgroup and the last subgroup. Tag density was calculated by normalizing read count (RPKM) in each DHS region. Accessibility variation was measured by coefficient of variation of subnucleosome-sized particle density in each 2,000 bp genomic window centered by DHS. We tested whether the ratio between subnucleosome-sized fragments and nucleosomal fragments in different cells is correlated with the ratio between the DHSs with wide space and the DHSs with narrow space.

We plotted fragment size ratio versus space ratio for each cell and calculated Pearson correlation (**Extended Data Fig. 7d**). The result shows that the nucleosome spacing at DHSs is not correlated with the fragment size ratio in different cells. This analysis was performed using NIH3T3 cell datasets.

Cell-to-cell DHS variation, gene expression variation and their correlations with nucleosome variance across cells

DHS variation was defined as coefficient of variation of tag density from 5 scDNase-seq libraries for each 2000 bp genomic window centered by DHS. All DHSs were sorted by DHS variation. Average nucleosome variance across cells and average DHS variation in bins of 500 sorted DHSs were calculated and used to draw dot plot and to calculate the Pearson correlation (**Fig. 3f**). Next, we separated DHSs into 6 subgroups according to the number of cells with detected DHS and calculated the fraction of cells that have positioned nucleosomes at -1 or +1 reference position surrounding DHS for each DHS subgroup (**Extended Data Fig. 7e**). We used Mann-Whitney rank test to determine the significance of value differences between the first and the last subgroups. We downloaded available mouse NIH3T3 single-cell gene digital expression data from Drop-seq paper[9] and calculated gene expression level as $\log_2$(TPM+1). Gene expression variation was defined as coefficient of variation of expression level from single cells. We defined top 25% of genes with smallest expression variation as low expression variation genes and top 25% of genes with largest expression variation as high variation genes. We compared nucleosome variance across cells at +1 nucleosome (+1 N) relative to TSS between two gene groups (**Fig. 3g**). Mann-Whitney rank test was used to determine the significance of value differences between two groups (P value = 1.2e-3). We further calculated the fraction of single cells that were detected as expressed (digital expression level $\geq$ 1) for each gene. We collected two gene groups: one has top 25% of genes with most cells detected as expressed and the other has top 25% of genes with least cells. We compared fraction of cells with positioned nucleosomes at +1 nucleosomes relative to TSS between two gene groups (**Fig. 3f**) and used Mann-Whitney rank test to determine the significance of the difference (P value = 2.2e-2). Top 1,000 genes with smallest variance across cells at +1 nucleosomes relative to TSS were collected for gene ontology analysis using DAVID (v6.7) online resources[37]. This analysis was performed using NIH3T3 cell datasets.

Clustering NIH3T3 cells, T cells and mouse ESCs based on cell-to-cell nucleosome positioning dissimilarity score matrix

We got the union set of DHSs for NIH3T3 cells, T cells and mouse ESCs as described above. We used mean values of nucleosome variance within 2000 bp window centered by DHSs to measure nucleosome dissimilarity score between two cells. Because some cells have low coverage, the mean variance represents only a small number of DHSs and may not reflect true similarity between two cells. Therefore, we considered the mean variance as valid only if there are more than 100 DHSs represented in the calculation. Further, the cells that have valid mean variance with less than 100 other cells were excluded. Then, we used the nucleosome dissimilarity score as the distance between cells to cluster the cells by applying the R function *hclust*. The symmetric nucleosome dissimilarity score matrix with clusters were drawn in heat map and the cell type, the experiment time, and fragment size ratio for each sample were indicated (**Extended Data Fig. 7k**).

DNA sequence preference and MNase cleavage bias analysis

We used NIH3T3 cells to perform this analysis. Di-nucleotide frequency was calculated and normalized by dividing the frequency by its genome-wide expectation at each site within ±200 bp from nucleosome midpoints. We randomly selected 10 million nucleosomes and calculated fraction of AA/TT/AT/TA in all di-nucleotides within flanking region (defined as 75-150 bp away from midpoint) for each nucleosome. We sorted nucleosomes by AA/TT/AT/TA fractions and calculated average nucleosome variance and average AA/TT/AT/TA fractions in bins of 10K nucleosomes for dot plot. To check whether DNA sequence preference was due to MNase cleavage bias, we examined sequence preference at the cleavage sites, which correspond to both fragment-ends positions in our data. Consistent with previous report[14], MNase tends to cleave at A or T nucleotides. The cleavage bias is weak and does not extend beyond 1-2 base pairs, which cannot explain the sequence preference in the linker or nucleosome core regions which extend up to 150 base pairs. Moreover, different nucleosome variance is not associated with a change in the sequence at cleavage sites, suggesting the analysis of the correlation between DNA sequence preference and nucleosome variance is robust to the introduction of MNase cleavage biases.

Single nucleotide variance analysis

We examined the association of the cell-to-cell single nucleotide variance with nucleosome positioning variance in NIH3T3 single cells. To identify cell-to-cell single base variance, we used annotated mouse SNP dataset downloaded from UCSC genome browser (mm9, snp128). If both reference and alternative bases are observed at a genomic site in different cells (NIH3T3 cells), we say there is a SNP at that site. We calculated SNP frequency (counts per 1K nucleosomes) at each position relative to nucleosome midpoints within a ±200 bp window for nucleosome subgroups with different nucleosome variance across cells. To analyze SNPs within TF motif match, we first used PIQ[38] to identify TF motif matches within DHSs by combining DNase-seq dataset and PWM motifs for TFs downloaded from JASPAR database[39]. We next identified SNPs that were likely to reduce TF binding affinities by searching SNPs that were within motif matches and met following criteira: $S_{ref} - S_{alt} \geq 0.3$, where S_{ref} is the probability of reference base at the motif PWM and S_{alt} is the probability of alternative base. We examined whether a CTCF motif match occupied by nucleosomes had more alternative bases. For each sequence that covers a CTCF motif match with a SNP, we determined whether it has a reference base or an alternative base. Bases with low sequencing quality score (Q<20) were not included. Then we summed up all the reference bases and alternative bases from either nucleosome, subnucleosome-sized particle, DNase-seq tag, and CTCF ChIP-seq tag, and calculated the ratio of the number of alternative bases to reference bases for each category. Significance were determined by Fisher Exact Test. We calculated the frequency of identified SNPs as total counts divided by total lengths of motif matches for DHS subgroups separated by nucleosome variance. To determine the significance of observed SNP frequency difference between the first and last subgroups, we randomly shuffled DHSs among four subgroups and calculated the same value for 10^5 times. We calculated P value as the fraction of the values obtained by random that was equal to or smaller than the observed value. We further focused on DHSs within promoter region (±1K bp from TSS) and separated genes into four subgroups by expression variation. We used the same approach to calculate the frequency of identified SNPs for each subgroup and calculated the P value of the difference between the first and the last subgroups. We used NIH3T3 cell datasets to perform this analysis. It is of note that measurement of nucleosome variance over genomic regions with SNPs can also be achieved by other methods such as methylation footprinting[40,41].

Identification of de novo enhancers in EB, T_H1 and T_H2 cells

We identified newly formed enhancers in EB cells after differentiation from mouse ESCs, and newly formed enhancers in T_H1 or T_H2 cells after differentiation from naïve T cells. We used p300 ChIP-seq to identify enhancers in EB cells and selected those that do not overlap with p300 and H3K27ac peaks in mouss ESCs as de novo enhancers. We used p300 ChIP-seq to identify enhancers in T_H1 and T_H1 cells separately. We selected T_H1 enhancers that do not overlap with T_H1 enhancers and not overlap with H3K27ac peaks in naïve T cells as T_H1 specific de novo enhancers. Similarly, we identified T_H1 specific de novo enhancers. Furthermore, the common T_H1 and T_H1 enhancers were excluded in this analysis.

Nucleosome occupancy score at de novo enhancers in undifferentiated cells

We calculated fraction of nucleosomes at the enhancer center: $f = c_{center}/(c_{flanking} + c_{center})$, where c_{center} represents the sum of the nucleosomes located within ±100 bp window centered by enhancer centers (p300 peak summits) and $c_{flanking}$ represents the sum of the nucleosomes located at the regions with distance of 100–200 bp from enhancer centers. Next, we calculated this score in a background DHS pool. We collected DNase-seq datasets in a broad range of cell types downloaded from mouse ENOCDE projects[26] (total 33 different cell types and total ~6.5 million DHSs, see datasets reference in **Supplementary Table 2**) and pooled all the DHSs from different cell types. This dataset represents potential DHSs that are inaccessible in majority cell types but may be open in specific cell condition. We calculated fraction of nucleosomes at these DHS centers as the background score f'. Then, the nucleosome occupancy score is calculated as: $s = \log_2(f/f')$. We used hypergeometric distribution to determine whether the fraction of nucleosomes at de novo enhancer centers is significantly smaller than a score calculated from random draws from the background potential DHSs. We plot nucleosome occupancy score and P value for each single cell and used cutoff s <-0.1 and P value <0.05 to identify primed undifferentiated cells.

We calculated nucleosome occupancy difference between two groups of cells, i.e., significantly nucleosome-depleted cells described above and the remaining cells, for each de novo enhancer, as follows: $d = (c_{center,depleted} - c_{flanking,depleted}) - (c_{center,non\_depleted} - c_{flanking,non\_depleted})$, where c represents normalized nucleosome count within the enhancer center (*center*) or in the flanking region (*flanking*) defined above for pooled nucleosome-depleted cells (*depleted*) or the remaining cells (*non_depleted*). Enhancers with less than 2 nucleosome calls in pooled nucleosome-depleted cells were not included in this analysis. Enhancers were ranked in ascending order. Top 1,000 enhancers were used for motif analysis

by SeqPos program[42] and/or gene ontology (GO) analysis by GREAT program [43] with default settings.

Code availability

Custom codes for nucleosome spacing uniformness quantification and nucleosome occupancy score calculation are available at https://github.com/binbinlai2012/scMNase.

Data Availability

The scMNase-seq data sets have been deposited in the Gene Expression Omnibus database with accession number GSE96688.

Methods-only References

30 Hu, G. *et al.* H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* **12**, 180-192 (2013).

31 Hu, G. *et al.* Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nat. Immunol.* **14**, 1190-1198 (2013).

32 Elling, U. *et al.* Forward and reverse genetics through derivation of haploid mouse embryonic stem cells. *Cell Stem Cell* **9**, 563-574 (2011).

33 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).

34 Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359 (2012).

35 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

36 Voong, L. N. *et al.* Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. *Cell* **167**, 1555-1570.e1515 (2016).

37 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic. Acids Res.* **37**, 1-13 (2009).

38 Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* **32**, 171-178 (2014).

39 Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic. Acids Res.* **32**, D91-94 (2004).

40 Bell, O. *et al.* Accessibility of the Drosophila genome discriminates PcG repression, H4K16 acetylation and replication timing. *Nat. Struct. Mol. Biol.* **17**, 894-900 (2010).

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

42 Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, R83 (2011).
43 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495-501 (2010).