Size-based expectation maximization for characterizing nucleosome positions and subtypes

Jianyu Yang[1], Kuangyu Yen[2,3]*, Shaun Mahony[1]*

[1] Center for Eukaryotic Gene Regulation, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA

[2] State Key Laboratory of Experimental Hematology, National Clinical Research Center for Blood Diseases, Haihe Laboratory of Cell Ecosystem, Institute of Hematology & Blood Diseases Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Tianjin 300020, China

[3] Department of Developmental Biology, School of Basic Medical Sciences, Southern Medical University, Guangzhou, China

* To whom correspondence should be addressed: kuangyuyen@ihcams.ac.cn; mahony@psu.edu

ABSTRACT

Genome-wide nucleosome profiles are predominantly characterized using MNase-seq, which involves extensive MNase digestion and size selection to enrich for mono-nucleosome-sized fragments. Most available MNase-seq analysis packages assume that nucleosomes uniformly protect 147bp DNA fragments. However, some nucleosomes with atypical histone or chemical compositions protect shorter lengths of DNA. The rigid assumptions imposed by current nucleosome analysis packages potentially prevent investigators from understanding the regulatory roles played by atypical nucleosomes.

To enable the characterization of different nucleosome types from MNase-seq data, we introduce the Size-based Expectation Maximization (SEM) nucleosome-calling package. SEM employs a hierarchical Gaussian mixture model to estimate nucleosome positions and subtypes. Nucleosome subtypes are automatically identified based on the distribution of protected DNA fragments. Benchmark analysis indicates that SEM is on par with existing packages in terms of standard nucleosome-calling accuracy metrics, while uniquely providing the ability to characterize nucleosome subtype identities.

Applying SEM to a low-dose MNase-H2B-ChIP-seq dataset from mouse embryonic stem cells, we identified three nucleosome types: short-fragment nucleosomes; canonical nucleosomes; and di-nucleosomes. Short-fragment nucleosomes can be divided further into two subtypes based on their chromatin accessibility. Interestingly, short-fragment nucleosomes in accessible regions exhibit high MNase sensitivity and are enriched at transcription start sites (TSSs) and CTCF peaks, similar to previously reported “fragile nucleosomes”. These SEM-defined accessible short-fragment nucleosomes are found not just in promoters, but also in distal regulatory regions. Additional analyses reveal their co-localization with the chromatin remodelers Chd6, Chd8, and Ep400.

In summary, SEM provides an effective platform for exploration of non-standard nucleosome subtypes.

INTRODUCTION

The nucleosome is the basic packaging unit of chromatin, typically comprising ~147bp DNA wrapped around the histone octamer (Luger et al. 1997). Nucleosomes participate in gene regulation by both physically impeding access to DNA and by serving as a substrate for interactions with regulatory proteins (Klemm et al. 2019). Chemical modifications on histone tails, including methylation, acetylation, and ubiquitination, can alter DNA affinity to the octamer and can be engaged by regulatory proteins (Kouzarides 2007; Torres and Fujimori 2015). Several histone tail chemical modifications are well correlated with transcriptional activity or repression (Kim et al. 2005; Papp and Müller 2006; Wang et al. 2008). Histone variants, such as H2A.Z and H3.3, have also been reported to play roles in many important biological events, such as DNA replication and enhancer activity (Zentner and Henikoff 2013; Chen et al. 2013b).

The most common technique for studying nucleosome landscapes across the genome is micrococcal nuclease sequencing (MNase-seq). MNase is an endo-exonuclease that preferentially digests accessible DNA between nucleosomes. After size selection, mono-nucleosome-sized DNA fragments are retained for high-throughput sequencing (Jiang and Pugh 2009; Mavrich et al. 2008). However, depending on the composition of the nucleosome and the factors engaging the nucleosome, not all nucleosomes protect the canonical 147bp of DNA. For example, nucleosomes engaged by Pol II lose one H2A-H2B dimer and transiently become hexamers (Ramachandran et al. 2017). Nucleosomes containing the histone variant H2A.Z can exhibit a distinct unwrapping state from the canonical nucleosome (Wen et al. 2020). Some studies employing alternative nucleosome mapping methods, including low-dose MNase-seq, methidiumpropyl-EDTA sequencing (MPE-seq) (Ishii et al. 2015), Cleavage Under Targets and Release Using Nuclease (CUT&RUN) (Brahma and Henikoff 2019), and chemical mapping (Voong et al. 2016), found MNase-sensitive nucleosome subtypes in yeast and mouse which protect shorter DNA fragments than canonical nucleosomes (Ishii et al. 2015; Voong et al. 2016; Brahma and Henikoff 2019). These studies have begun revealing variations in nucleosome composition across the genome.

Despite experimental results that have characterized a wider diversity of nucleosome composition, most nucleosome-calling software packages still assume that nucleosomes uniformly protect ~147bp of DNA (Chen et al. 2013a; Zhou et al. 2016; Becker et al. 2013). Current approaches use this rigid assumption when estimating the locations and occupancy properties of nucleosomes, making their performance sub-optimal when characterizing nucleosomes of non-canonical DNA length. Although some packages have begun incorporating the ability to detect nucleosome positioning dynamics (Zhou et al. 2016; Chen et al. 2013a), none are yet able to distinguish various nucleosome subtypes from MNase-seq data.

To resolve the lack of an effective method for characterizing nucleosome subtypes, we introduce a new nucleosome-calling package called Size-based Expectation Maximization (SEM). We evaluate the performance of SEM by comparing it to existing nucleosome-calling packages. We then apply SEM to analyze a low-dose MNase-ChIP-H2B dataset from mouse embryonic stem cells (mESCs) (Ishii et al. 2015), thereby demonstrating SEM's ability to characterize various nucleosome subtypes genome-wide.

RESULTS

A hierarchical Gaussian Mixture Model for characterizing nucleosome types

SEM is a hierarchical Gaussian Mixture Model (GMM), which probabilistically models the positions, occupancy, fuzziness, and subtype identities of nucleosomes from MNase-seq data (**Fig. 1**). The components of the mixture model represent individual nucleosomes; the properties of each nucleosome are modeled based on the mapped locations and lengths of MNase-seq fragments. Specifically, each nucleosome component is defined by its dyad location, occupancy, fuzziness, and the probability of belonging to each nucleosome subtype.

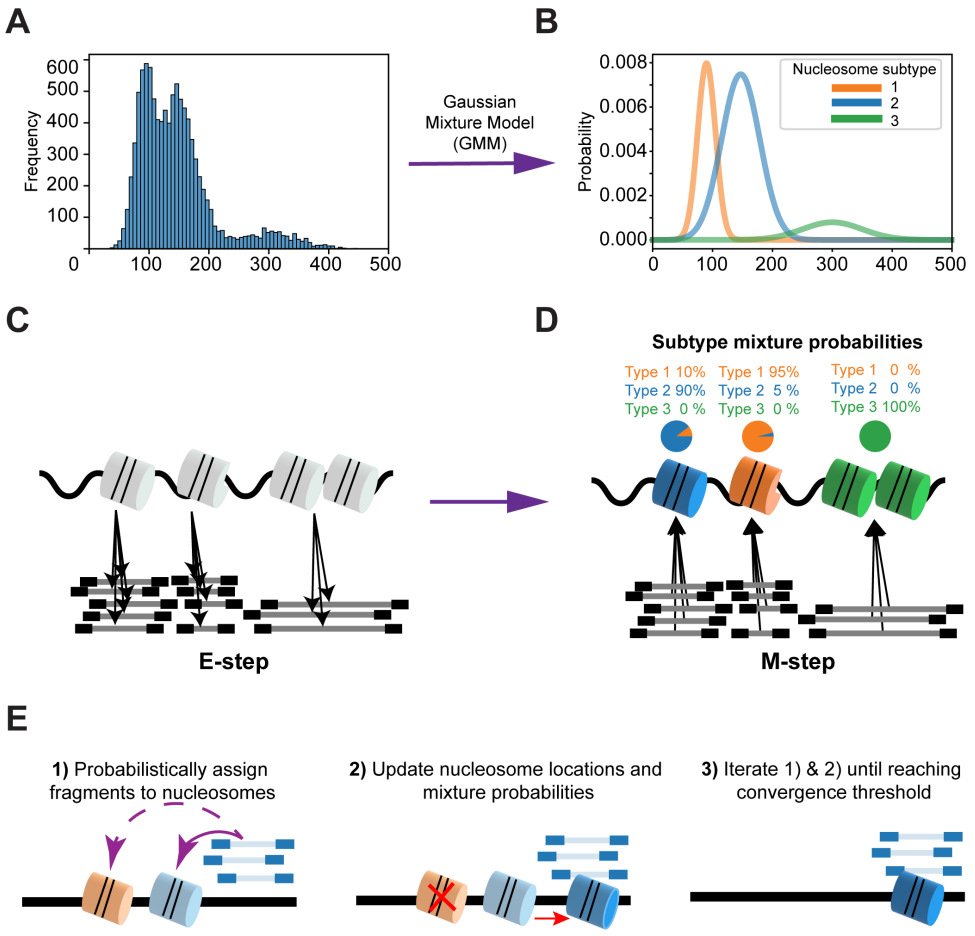


Figure 1: Overview of the SEM algorithm. **A)** SEM first calculates the global fragment size distribution from the MNase-seq data. **B)** A Gaussian Mixture Model is used to deconvolve the fragment size distribution into a set of nucleosome subtypes. **C)** In the Expectation step of the algorithm, each MNase-seq fragment is probabilistically assigned to nucleosome components according to the current locations, strengths, and subtype identities of the components. **D)** In the Maximization step of the algorithm, the various nucleosome properties are updated based on the current fragment assignments. **E)** Detailed illustration of how nucleosome properties are updated during EM iterations.

SEM runs in two major phases: nucleosome subtype discovery and nucleosome finding. During the first phase, SEM fits a GMM onto the fragment size distribution of the entire MNase-seq dataset to determine the Gaussian distribution parameters of each nucleosome subtype (**Fig. 1A, B**). The number of clusters, corresponding to the number of nucleosome subtypes, is an essential parameter of the model. It can be specified by the user according to prior knowledge of the biological sample, or SEM can automatically determine the best fit value using a Dirichlet Process Mixture Model (DPMM).

In the second phase, SEM uses a hierarchical GMM to compute the likelihood that each specific nucleosome is responsible for generating each MNase-seq fragment in the dataset. A Generalized Expectation Maximization (GEM) framework is used to calculate the latent assignment of MNase-seq fragments to nucleosomes and to estimate the various properties associated with each nucleosome (see **Methods**). Briefly, the Expectation step in the algorithm probabilistically assigns MNase-seq fragments to the nucleosome components that are most likely to have generated them based on their current locations and properties (**Fig. 1C**). Then the Maximization step updates the properties of the nucleosome components based on the best fit to the current fragment assignments (**Fig. 1D**). For example, a

nucleosome component's dyad location, occupancy, and fuzziness properties are re-estimated based on the mean midpoint location, weights, and variance of currently assigned MNase-seq fragments. Subtype identity is updated based on the size distribution of assigned fragments. The algorithm iterates through Expectation and Maximization steps until convergence (**Fig. 1E**).

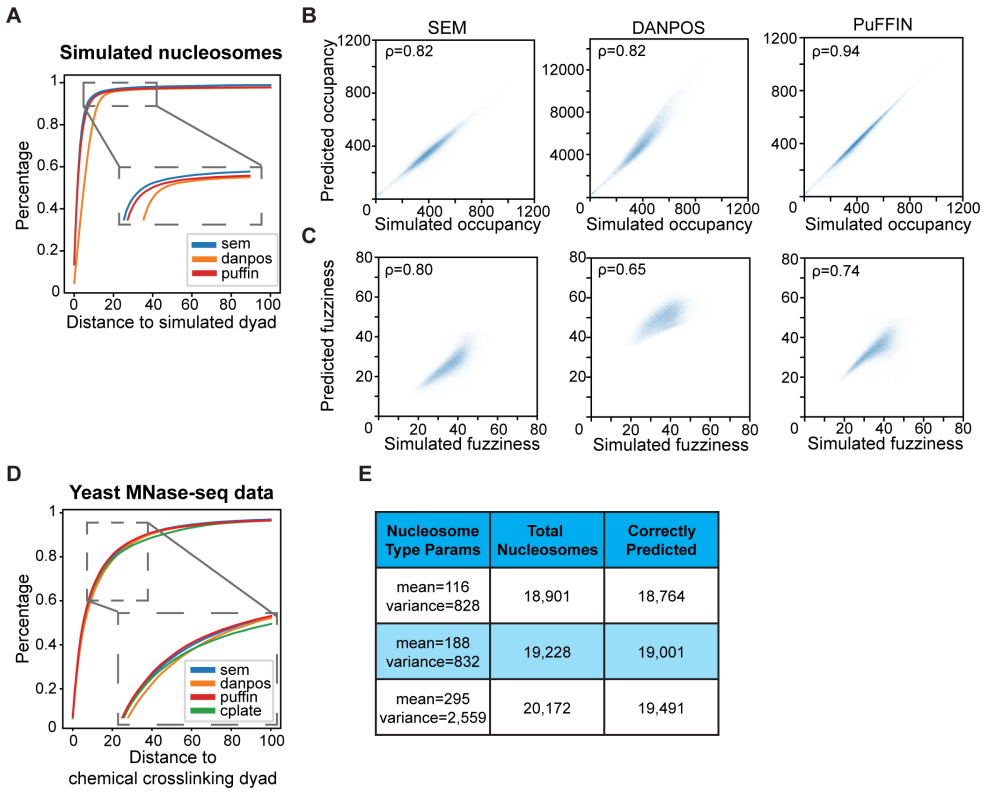
To better reflect the known biological properties of nucleosomes, several priors are integrated into the model. First, a sparse prior on nucleosome occupancy eliminates components that don't have enough fragments assigned, thereby encouraging solutions where each nucleosome is supported by MNase-seq data. Another sparse prior on subtype probability encourages each component to be a member of individual subtypes, which improves the interpretability of nucleosome subtype assignments. The physical exclusion of adjacent nucleosomes is also taken into consideration to ensure that neighboring nucleosomes do not overlap in the model.

SEM accurately predicts conventional nucleosome properties

We first evaluated the performance of SEM in predicting conventional nucleosome properties, including nucleosome dyad location, nucleosome occupancy, and fuzziness, on both simulated and real MNase-seq datasets. To generate the simulated dataset, we took reference nucleosome dyad locations from yeast (Jiang and Pugh 2009), based on a collection of nucleosome dyad locations compiled from six datasets. The occupancy and fuzziness of each nucleosome were then computed using a H4 MNase-ChIP-seq dataset against the reference nucleosome dyads (see **Methods**). Background "noise" reads following a Poisson distribution were added globally to the simulated dataset to test the robustness of each algorithm.

SEM's performance in this simulated dataset was evaluated against two other nucleosome-calling packages: DANPOS and PuFFIN (Chen et al. 2013a; Polishko et al. 2014). All three packages were executed using their default settings (see **Methods**). The numbers of nucleosomes predicted by each package was similar to the number in the reference set (i.e., ~60,000 nucleosomes). SEM and PuFFIN outperform DANPOS in predicting simulated nucleosome dyad locations (**Fig. 2A**). PuFFIN achieves the best estimates of simulated nucleosome occupancy, as evaluated by Pearson Correlation Coefficient (**Fig. 2B**). While SEM and DANPOS have similar correlations with the simulated occupancy, DANPOS has a general tendency to overestimate nucleosome occupancy (**Fig. 2B, middle panel**). SEM outperforms PuFFIN and DANPOS in estimating the simulated fuzziness levels, again evaluated by Pearson Correlation Coefficient (**Fig. 2C**).

We further evaluated the performance of SEM using real MNase-seq data (Zhou et al. 2016). Since no ground truth of nucleosome occupancy or fuzziness is available, our evaluation solely focused on predictions of nucleosome dyad locations. The nucleosome dyad locations obtained from the chemical crosslinking experiment were used as the "gold standard" due to its ability to specifically break DNA nucleotides near the dyad location. We included a fourth nucleosome-calling package, Cplate, in this comparison, as predicted nucleosome dyad locations in the same dataset were available in the original Cplate publication (Zhou et al. 2016). Notably, although SEM, DANPOS, and PuFFIN exhibited significant performance disparities in simulated datasets, all four software packages exhibited comparable performance when evaluated on the real MNase-seq dataset (**Fig. 2D**).



Nucleosome Type Params	Total Nucleosomes	Correctly Predicted
mean=116 variance=828	18,901	18,764
mean=188 variance=832	19,228	19,001
mean=295 variance=2,559	20,172	19,491

Figure 2: Comparison of SEM to existing nucleosome calling packages on common metrics. **A)** Cumulative percentage of distances between predicted nucleosome dyads and simulated nucleosome dyads in simulated MNase-seq data. **B)** Correlation between predicted occupancy and simulated occupancy. Each simulated nucleosome is assigned to the closest predicted nucleosome for the purposes of comparison. **C)** Correlation between predicted fuzziness and simulated fuzziness. Each simulated nucleosome is assigned to the closest predicted nucleosome for the purposes of comparison. **D)** Cumulative percentage of distances between predicted nucleosome dyads and chemical crosslinking dyads in real MNase-seq data (Voong et al., 2016). **E)** Numbers of total simulated nucleosomes and correctly predicted nucleosomes of each nucleosome subtype.

Additionally, we evaluated SEM's ability to distinguish various nucleosome subtypes. A simulated MNase-seq dataset, where reads are sampled from three distinct nucleosome subtypes, was generated. SEM achieved high accuracy in correctly identifying the subtypes from which each simulated nucleosome was sampled (**Fig. 2E**).

To summarize, our results indicate that SEM performs comparably with existing nucleosome-calling packages in predicting nucleosome dyad locations. Although PuFFIN displays greater accuracy in predicting nucleosome occupancy, SEM outperforms it in predicting nucleosome fuzziness. Overall, SEM yields comparable performance to other nucleosome-calling packages when it comes to conventional nucleosome metrics. In addition, SEM can precisely predict nucleosome subtypes in simulated data.

Genome-wide detection of nucleosome subtypes in mouse embryonic stem cells

To assess whether SEM can distinguish different types of nucleosomes from MNase-seq data, we applied it to a low-dose MNase-H2B-ChIP-seq dataset from mESCs (Ishii et al. 2015). The original study focused on characterizing the locations of so-called "fragile", or MNase-sensitive, nucleosomes, which protect shorter DNA fragments than canonical nucleosomes under low-dose MNase digestion (Chereji et al. 2017; Ishii et al. 2015). While the study found evidence for MNase-sensitive nucleosomes at TSSs and

CTCF binding sites, the overlapping fragment size distributions of canonical nucleosomes and fragile nucleosomes makes it difficult to distinguish between these nucleosome subtypes at individual sites. The original study applied hard fragment length cut-offs (50-100bp, 101-140bp, 140-190bp) to define different nucleosome subtypes. However, the application of these thresholds does not adequately distinguish between nucleosome subtypes at TSSs or CTCF binding sites, given the variations observed in the overall fragment size distribution (**Fig. 3A,B, Supplemental Fig. S1A**). Our goal here is to assess whether SEM can more clearly discriminate nucleosome subtypes and add additional insights into the nature of the MNase-sensitive nucleosomes characterized in the original study.

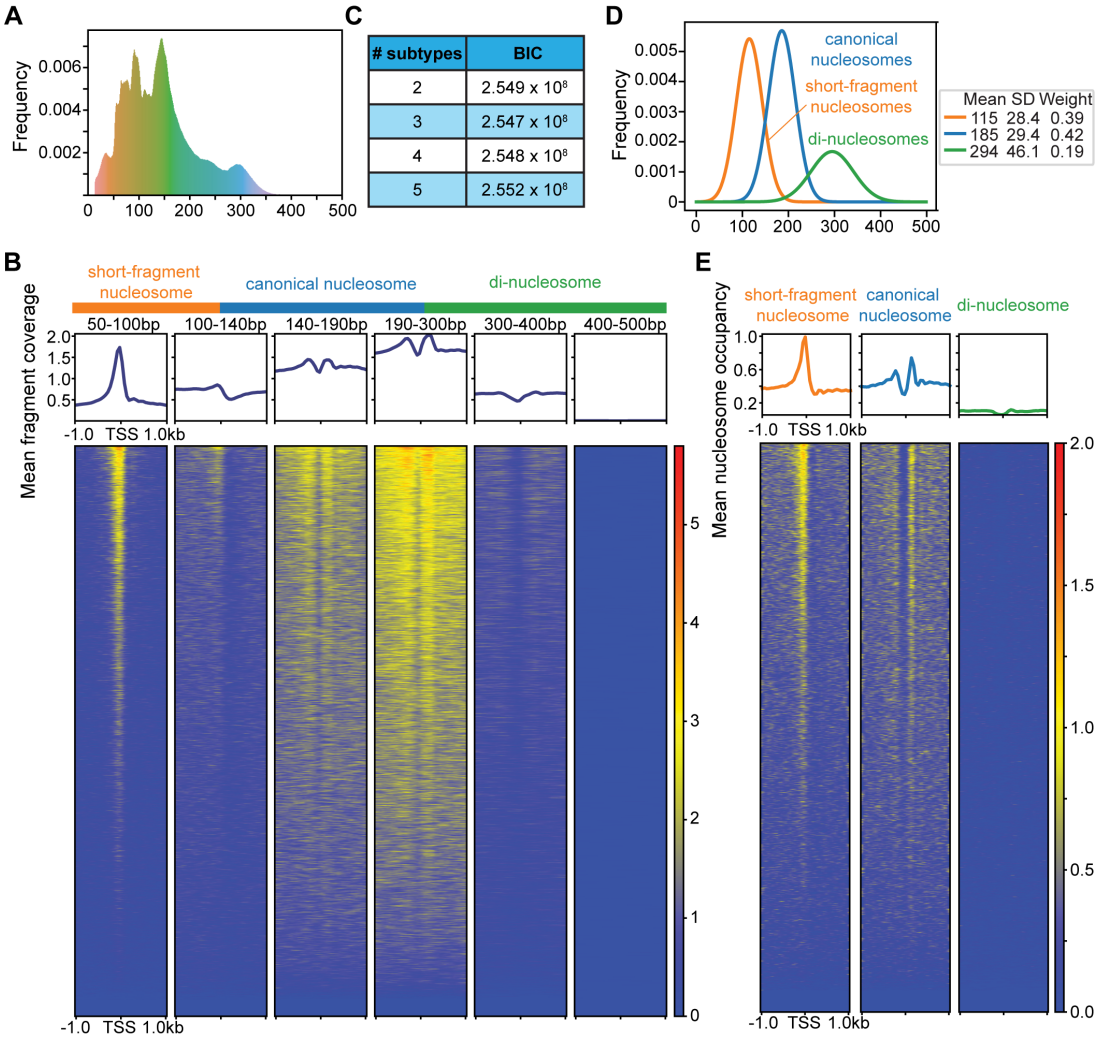


Figure 3: SEM characterizes three nucleosome subtypes in a low dose MNase-ChIP-seq dataset from mESCs. **A)** Fragment size distribution of the low dose MNase-ChIP-seq dataset. **B)** Heatmap and profile plot of MNase-seq fragments split by fragment size (50-100bp, 100-140bp, 140-190bp, 190-300bp, 300-400bp, 400-500bp) around TSSs. **C)** Bayesian Information Criterion (BIC) computed across varying numbers of nucleosome subtypes. **D)** Fragment size distribution of each nucleosome subtype as determined by SEM. **E)** Heatmap and profile plot of each SEM-defined nucleosome subtype around TSSs.

Visual inspection of the DNA fragment size distribution suggests that there are three main modes of MNase-protected fragments in the MNase-H2B-ChIP dataset, and thus three distinct nucleosome subtypes (**Fig. 3A**). We confirmed that including three nucleosome subtypes in the SEM model provided the best fit to the data by evaluating the Bayesian Information Criterion (BIC) metric over varying numbers

of subtypes (**Fig. 3C**). Based on their mean fragment sizes, we refer here to each SEM-defined nucleosome subtype as: short-fragment nucleosomes (~115bp mean size); canonical nucleosomes (~185bp mean size); and di-nucleosomes (~295bp mean size) (**Fig. 3A, D**). Here, the larger mean size of DNA fragments protected by the canonical nucleosome, compared to the typically reported 147bp, likely results from incompletely digested linker DNA at a subset of nucleosomes under the light MNase digestion conditions used in the MNase-H2B-ChIP experiment. SEM discovered a total of 6,200,234 nucleosomes along the mouse genome (**Supplemental Data D1**). More than half exhibit mixed nucleosome subtype identity, indicating the highly dynamic nature of nucleosomes across the whole genome (**Supplemental Fig. S1C, D**). In comparison with hard fragment length cut-offs, SEM's detection of nucleosome subtypes more clearly separates nucleosomes belonging to each subtype at TSSs and CTCF sites (**Fig. 3E, Supplemental Fig. S1B**). Specifically, short-fragment nucleosomes are enriched at the centers of TSSs and CTCF sites, while canonical nucleosomes are well-positioned in the flanking regions (**Fig. 3E, Supplemental Fig. S1B**).

The enrichment of SEM-defined short-fragment nucleosomes at TSSs and CTCF sites mirrors that of the MNase-sensitive nucleosomes defined in the original study. However, only a small proportion of all SEM-defined short-fragment nucleosomes overlap TSSs or CTCF sites (**Fig. 4A**). We therefore asked whether SEM's short-fragment nucleosome category consists of a homogenous class of MNase-sensitive nucleosomes or whether it encompasses multiple short-fragment nucleosome subtypes. To more clearly delineate the properties of each nucleosome subtype, we focus only on SEM-defined nucleosomes with occupancy greater than 5 reads and unambiguous subtype assignments (i.e., one of the nucleosome subtype probabilities >0.9).

We first subcategorized SEM's nucleosome subtypes according to their association with potential regulatory regions by overlapping with mESC ATAC-seq peaks (Ostapcuk et al. 2018). Two thirds of the SEM-defined accessible short-fragment nucleosomes overlap with TSSs or CTCF binding sites, while the non-accessible short-fragment nucleosomes are primarily located in distal regions (**Fig. 4A**). We then calculated the MNase sensitivity of each nucleosome subcategory by comparing the number of MNase-seq fragments from the low-dose MNase-H2B-ChIP-seq dataset with the number of fragments from a high-dose MNase-seq dataset (Ishii et al. 2015) in the +/-50bp region surrounding each nucleosome dyad (**Supplemental Fig. S2A**). Accessible short-fragment nucleosomes display high MNase sensitivity, consistent with the MNase-sensitive short-fragment nucleosomes found in the original study. However, the non-accessible short-fragment nucleosomes do not display significantly higher MNase sensitivity than non-accessible canonical nucleosomes (**Supplemental Fig. S2A**).

Since the non-accessible short-fragment nucleosomes do not display higher MNase sensitivity, we asked whether their shorter fragment lengths could be accounted for by MNase digestion biases towards *A/T*-rich sequences (Dingwall et al. 1981). We therefore compared the MNase sensitivity index for each nucleosome to the *A/T* content in the +/-50bp region surrounding each nucleosome dyad (**Fig. 4B**). While higher *A/T* content appears to correlate with higher MNase sensitivity at non-accessible nucleosome categories, the *A/T* content properties of non-accessible short-fragment nucleosomes are again similar to those of non-accessible canonical nucleosomes. However, non-accessible short-fragment nucleosomes display significantly higher *A/T* content at entry and exit sites compared with non-accessible canonical nucleosomes (*t*-test statistic=45.78, p-value < 2e-308 **Supplemental Fig. S2B**). This suggests that the MNase bias towards digesting *A/T*-rich sequences at nucleosome flanks could be responsible for the shorter fragment distributions at this subset of short-fragment nucleosomes. In other words, there may not be any substantial difference between the non-accessible short-fragment nucleosomes and canonical nucleosomes as defined by SEM, other than a greater susceptibility to MNase biases at entry and exit sites.

In contrast to the non-accessible short-fragment nucleosomes, the accessible short-fragment nucleosomes display high MNase sensitivity despite being generally *G/C* rich (**Fig. 4B**). We confirmed the levels of MNase sensitivity using the MNase accessibility (MACC) score (Mieczkowski et al. 2016). The MACC score is the slope of the fitted line between fragment counts and logarithmic scaled MNase concentration in titrated MNase-seq experiments, where a high MACC score represents a highly MNase sensitive region. The MACC score is substantially higher at accessible short-fragment nucleosomes compared with random nucleosome positions (**Supplemental Fig. S3A**). This indicates that SEM-defined

accessible short-fragment nucleosomes represent a subpopulation of *G/C*-rich, MNase-sensitive short-fragment nucleosomes that is distinctly separated from the broader pool of short-fragment nucleosomes.

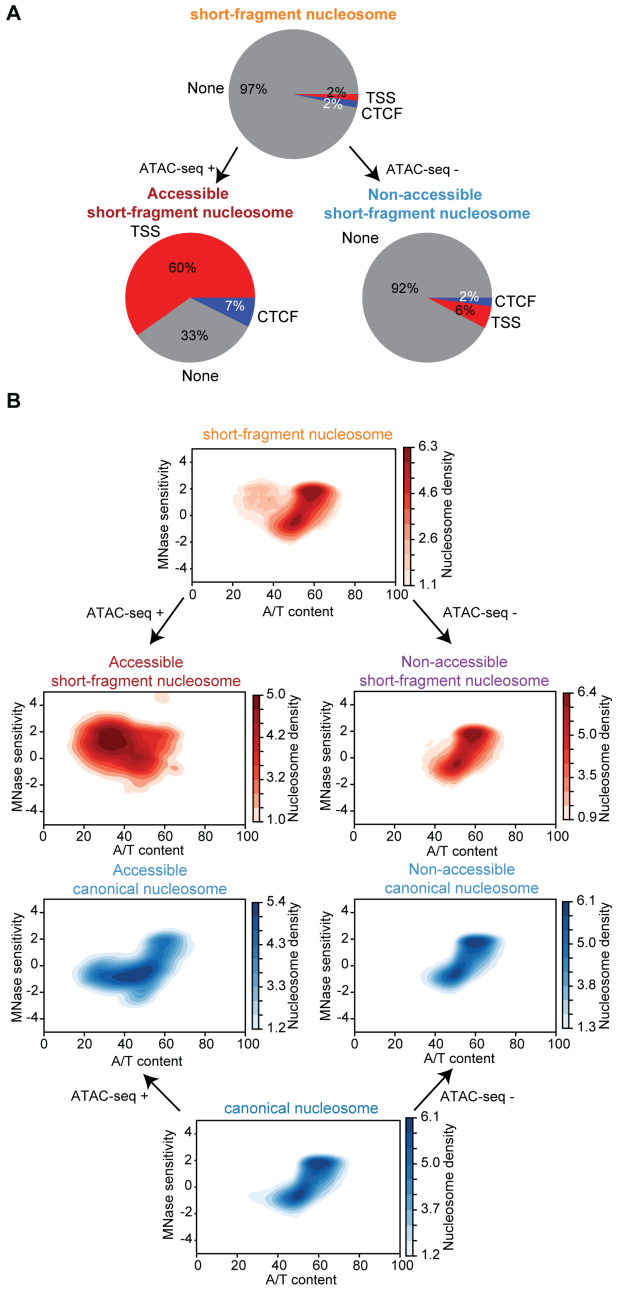


Figure 4: Subcategorization of SEM nucleosome subtypes reveals a population of accessible short-fragment nucleosomes. **A)** Percentages of short-fragment nucleosome subtype and subcategories located at TSS regions (+/-500bp) and CTCF binding sites. **B)** 2D density plots show the relationships between *A/T* content and MNase sensitivity of each nucleosome subtype and subcategory, color bars indicate the log density of nucleosomes. Nucleosome subcategories were obtained by examining if the nucleosome is fully overlapped by ATAC-seq peak (ATAC-seq +) or not (ATAC-seq -).

While the analyzed MNase-ChIP-seq experiment incorporated a H2B ChIP pulldown, it is possible that our detected short-fragment nucleosome subcategories result from MNase protection by non-nucleosomal regulatory proteins that artefactually crosslink with neighboring nucleosomes. To assess this potential concern, we examined nucleosome centering positioning (NCP) scores derived from a H4S47C-mediated chemical crosslinking approach (Voong et al. 2016). The accessible short-fragment nucleosomes display a clear peak in NCP scores, and nucleosome array phasing similar to canonical nucleosomes, providing orthogonal evidence for the presence of nucleosomes at the short-fragment nucleosome loci (**Supplemental Fig. S3B**).

In summary, SEM identified three distinct nucleosome subtypes in mESCs from the low-dose MNase H2B ChIP-seq dataset. By integrating chromatin accessibility data, the SEM-defined short-fragment nucleosomes were further subdivided into two subcategories. Accessible short-fragment nucleosomes display several properties corresponding to the previously proposed fragile nucleosomes, including high MNase sensitivity and enrichment at TSSs and CTCF binding sites.

Accessible short-fragment nucleosomes associate with distinct regulatory activities

Most previous studies of fragile or MNase-sensitive nucleosomes have focused on their occurrence at TSS and CTCF binding sites (Ishii et al. 2015; Voong et al. 2016). As shown above, SEM-defined accessible short-fragment nucleosomes are highly enriched at these regions (**Fig. 4A**). However, a substantial portion (33%) of SEM's accessible short-fragment nucleosomes do not overlap either TSSs or CTCF binding sites (**Fig. 4A**). To investigate the types of locations occupied by non-TSS/CTCF accessible short-fragment nucleosomes, we overlapped these nucleosomes with candidate cis-regulatory elements (cCREs) from the ENCODE SCREEN database (The ENCODE Project Consortium et al. 2020) (**Fig. 5A**). Notably, 17% of the accessible short-fragment nucleosomes overlap with either proximal or

distal enhancer-like signatures (pELSs, dELSs, respectively).

The association between accessible short-fragment nucleosomes and cis-regulatory elements suggests a possible role in transcriptional regulation. Therefore, we investigated whether accessible short-fragment nucleosomes overlap the occupancy of factors involved in nucleosome composition or histone modifications associated with regulatory activities in mES cells. A previous report in *Drosophila* suggested that H3.3/H2A.Z-containing nucleosomes are enriched at regulatory regions and are unstable under high salt conditions (Jin et al. 2009). Thus, we examined the enrichment of both histone variants around nucleosome subcategories. We found that H3.3 is highly enriched at regions flanking the accessible short-fragment nucleosomes, albeit locally depleted at dyad locations (**Fig. 5B**). H2A.Z also displays higher levels of enrichment flanking accessible short-fragment nucleosomes compared with non-accessible nucleosome subtypes (**Fig. 5B**). Active histone modifications, including H3K27ac, H3K4me3, H3K9ac, and H4ac, were enriched at sites adjacent to accessible short-fragment nucleosomes (**Supplemental Fig. S4**), while repressive marks, including H3K27me3 and H2AK119ub, do not show enrichment at accessible short-fragment nucleosome sites. Finally, several chromatin remodelers are enriched at accessible short-fragment nucleosome dyad locations, including Smarca4, Chd4, Chd6, Chd8, and Ep400. Ep400 is more prominently enriched at accessible short-fragment nucleosomes relative to other remodelers, and is centrally enriched at accessible short-fragment nucleosome dyad locations (**Supplemental Fig. S5**). These chromatin remodelers are known to influence the contact between DNA and the nucleosome (Clapier et al. 2017). Ep400 has also been reported to deposit H3.3 at promoter and enhancer regions (Pradhan et al. 2016), which may correspond to the H3.3 enrichment observed at accessible short-fragment nucleosome sites. Therefore, the enrichment of chromatin remodelers at accessible short-fragment nucleosomes suggests a mechanism for their destabilization.

Finally, we interrogated whether short-fragment nucleosomes are related to distinct transcription factor (TF) binding activities. Specifically, we performed enrichment analysis to investigate which transcription factors' binding sites (sourced through ChIP-Atlas (Zou et al. 2022; Oki et al. 2018)) significantly overlap with each nucleosome subcategory. The ChIP-seq peaks of many TFs are heavily enriched around accessible nucleosomes subtypes, including peaks for the pluripotency factors Oct4, Sox2, Nanog, and Klf4 (**Supplemental Table S1**). This observation is not necessarily surprising given the general enrichment of TF binding sites in accessible regions. TFs generally show lower fold-enrichment levels at non-accessible short-fragment nucleosomes, aligning with the expectation that nucleosomes located in less accessible regions would naturally attract fewer binding factors (**Supplemental Table S1**).

In summary, accessible short-fragment nucleosomes defined by SEM are not restricted to TSSs and CTCF binding sites and correspond to sites of specific chromatin remodeler enrichment. While the regulatory roles played by these atypical nucleosomes require further investigation, our analyses demonstrate the efficacy of SEM in characterizing diverse nucleosome types.

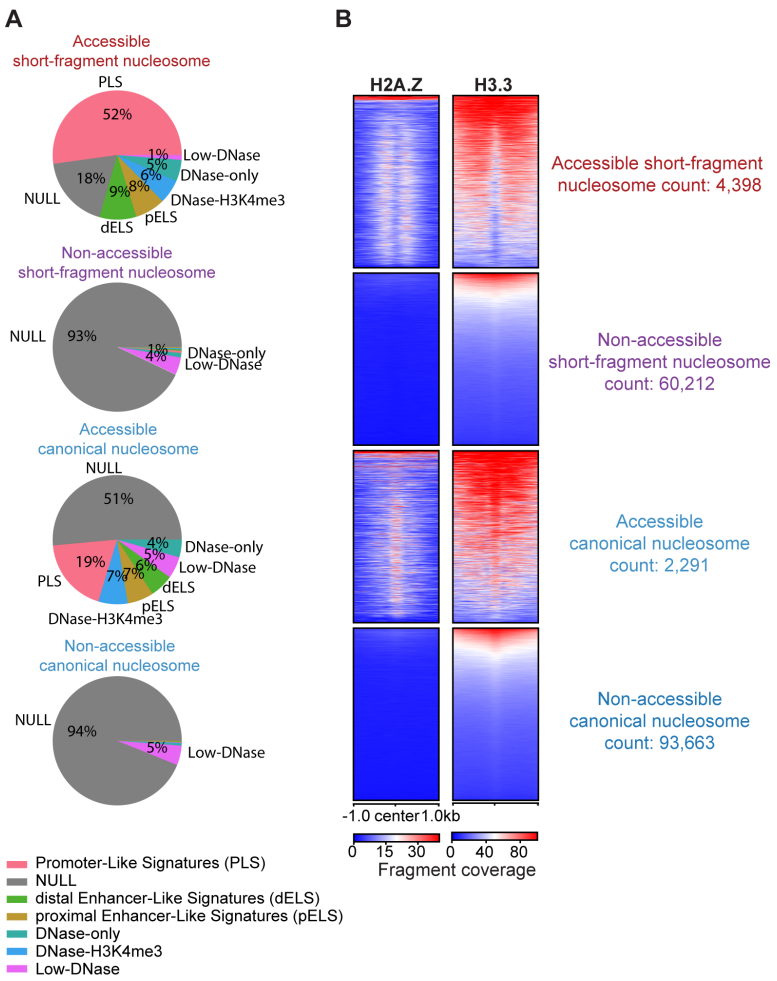


Figure 5: Accessible short-fragment nucleosomes are associated with regulatory elements and histone variants. **A)** Pie chart summarizing the overlap between SEM-defined nucleosome subcategories and ENCODE SCREEN cCRE types. **B)** Heatmap displaying the enrichment of histone variants H2A.Z and H3.3 around nucleosome dyad locations.

DISCUSSION

In recent years, there has been growing interest in nucleosome subtypes and compositional variants beyond canonical nucleosomes (Wen et al. 2020; Brahma and Henikoff 2019; Ramachandran et al. 2017; Hsieh et al. 2022). Here, we introduced SEM, a new nucleosome-calling framework capable of distinguishing various nucleosome subtypes from a single MNase-digested sequencing experiment. When compared with other nucleosome-calling packages, SEM exhibits comparable performance on conventional nucleosome-calling metrics while uniquely providing an automatic annotation of nucleosome subtype identity.

Using SEM on a low-dose MNase-H2B-ChIP-seq dataset from mESCs, we identified a nucleosome subtype that protects shorter DNA fragments than canonical nucleosomes. A further integration of ATAC-seq data classified the short-fragment nucleosomes into two subcategories, which we labeled accessible short-fragment nucleosomes and non-accessible short-fragment nucleosomes. Although SEM-defined non-accessible short-fragment nucleosomes may result from MNase digestion bias at entry and exit sites, it is worth noting that this nucleosome subcategory could be overlooked in traditional MNase-seq experiments conducting mono-nucleosome size selection. Thus, MNase-seq experiments that omit mono-nucleosome size selection, in combination with SEM analysis, could more accurately characterize nucleosome positions across the genome. Future versions of SEM could resolve the issues associated with MNase bias by incorporating the relationship between sequence content and DNA fragment size into the probabilistic model.

SEM-defined accessible short-fragment nucleosomes share a similar distribution to previously reported fragile nucleosomes at TSS and CTCF sites. However, SEM's ability to identify short-fragment nucleosomes genome-wide shows that accessible short-fragment nucleosomes are not restricted to TSS and CTCF sites but are also present in distal regulatory regions. Previous studies have suggested that histone variants H3.3 and H2A.Z regulate nucleosome stability or lead to nucleosome fragility (Wen et al. 2020; Jin et al. 2009; Jin and Felsenfeld 2007). We find an enrichment of H3.3, and to a lesser extent H2A.Z, at accessible short-fragment nucleosome sites. Accessible short-fragment nucleosomes are also associated with active histone marks but not repressive marks. Accessible short-fragment nucleosomes in mESCs are co-enriched with the chromatin remodelers Smarca4, Chd4, Chd6, Chd8, and Ep400,

suggesting that this nucleosome subcategory represent hotspots of chromatin remodeler activity. Assessing whether these chromatin remodelers play an active role in establishing short-fragment nucleosomes will require further experimental investigation (for example, via knock-down perturbations followed by MNase-seq experiments). Further work also remains to investigate whether SEM-defined accessible short-fragment nucleosomes play distinct regulatory roles in mESCs. However, our analyses demonstrate that SEM's genome-wide probabilistic approach to nucleosome subtype calling provides clear advantages over the typically employed fragment length threshold approaches.

METHODS

Nucleosome subtype characterization: To model different nucleosome subtypes present in an MNase-seq dataset, we first apply a Gaussian Mixture Model (GMM) to the overall fragment size distribution. This step can be processed with or without a user-defined number of subtypes. If the number of types has been defined, a finite GMM is used on the fragment size distribution to find the parameters of each type. GMM initialization is achieved using *k*-means clustering on the fragment sizes. Briefly, fragments are sorted by size and divided into percentile-based groups. The mean sizes of each group are used as initial centroids for *k*-means clustering. The final assignments from *k*-means serve as the starting point for GMM. Alternatively, if the number of subtypes has not been defined, a value can be automatically determined from the fragment size distribution by a Dirichlet Process Mixture Model (DPMM).

SEM probabilistic model: SEM is based on a hierarchical Gaussian Mixture Model that describes the likelihood of observing a set of MNase-seq fragments from a set of nucleosomes. Each nucleosome contributes a distribution of reads surrounding its genomic position to the overall mixture of reads. We assume that fragment locations are independently conditioned on the dyad location, fuzziness, and subtype mixture probabilities of their underlying nucleosomes.

SEM performs nucleosome discovery by finding the set of nucleosomes that maximizes the penalized likelihood of the observed MNase-seq fragments. First, we assume there are K nucleosome fragment size subtypes, where each subtype is a Gaussian distribution with mean $\Psi = \psi_1, \dots, \psi_K$, variance $\Phi = \phi_1, \dots, \phi_K$, and weight $\Omega = \omega_1, \dots, \omega_K$. Thus, the expected size distribution of fragments from each nucleosome follows a mixture of the subtype distributions, conditioned on their subtype probabilities. Throughout the whole genome, we consider N MNase-seq fragments that have been mapped to genomic locations $E = \epsilon_1, \dots, \epsilon_N$ with size $H = \eta_1, \dots, \eta_N$, and M potential nucleosomes at genomic locations $U = \mu_1, \dots, \mu_M$ with fuzziness $\Theta = \theta_1, \dots, \theta_M$ and subtype mixture probabilities $T = \tau_1, \dots, \tau_M = [\tau_{1,1}, \dots, \tau_{1,K},], \dots, [\tau_{M,1}, \dots, \tau_{M,K},]$. The locations of fragments from nucleosome m follow a Gaussian distribution, where the Gaussian mean equals the nucleosome dyad location μ_m and the variance equals the fuzziness parameter θ_m. We represent the latent assignment of fragments to nucleosomes that caused them as $Z = z_1, \dots, z_N$, where $z_i = j$ where j is the index of the nucleosome whose dyad is at position μ_j that generates fragment i.

The conditional probability of fragment r_i being generated from nucleosome j located in μ_j with mixture probability ρ_j and fuzziness θ_j is:

$$Pr(r_i|j,s) = \sum_{s=1}^{K} \frac{1}{\theta_j\sqrt{2\pi}} e^{-\frac{(\epsilon_i-\mu_j)^2}{2\theta_j^2}} * \tau_{j,s}\omega_s \frac{1}{\phi_s\sqrt{2\pi}} e^{-\frac{(\eta_i-\psi_s)^2}{2\phi_s^2}} \quad (1)$$

where $\tau_{j,s}$ represents the probability of nucleosome j belonging to fragment size type s.

The probability of a fragment is a convex combination of possible nucleosomes:

$$Pr(r_i|P,U,T,\Theta) = \sum_{j=1}^{M}\sum_{s=1}^{K} \rho_j \tau_{j,s} Pr(r_i|j,s) \quad (2)$$

where ρ represents the weighted occupancy of a nucleosome. The overall likelihood of the observed set of fragments is then:

$$Pr(E,H|P,U,T,\Theta)=\prod_{i=1}^{N}\sum_{j=1}^{M}\sum_{s=1}^{K}\rho_j\,\tau_{j,s}Pr(r_i|j,s) \qquad (3)$$

We incorporate several biological assumptions in the form of priors on the parameters. We place a sparseness-promoting negative Dirichlet prior α on the nucleosome weighted occupancy π, based on the assumption that each nucleosome should have sufficient numbers of assigned fragments to support its existence in the model: $Pr(\rho)\propto\prod_{j=1}^{M}(\rho_j)^{-\alpha},\alpha>0$ (4)

We also incorporate an additional prior to encourage nucleosomes to choose less ambiguous subtype identities. To do so, we place a Dirichlet prior β on the nucleosome type mixture probability τ.

$$Pr(\tau)\propto\prod_{j=1}^{M}\prod_{s=1}^{K}\tau_{j,s}^{-\beta},\beta>0 \qquad (5)$$

In summary, the complete-data log posterior is as follows:

$$\begin{aligned}logPr(U,P,T|E,H,\alpha,\beta)=&\sum_{i=1}^{N}\left[\sum_{j=1}^{M}\sum_{s=1}^{K}1\,(z_i=j)\left(log\rho_j+log\tau_{j,s}+logPr(r_i|j,s)\right)\right]\\&-\sum_{j=1}^{M}\alpha_j\,log\rho_j-\beta\sum_{j=1}^{M}\sum_{s=1}^{K}log\,\tau_{j,s}+C\end{aligned} \qquad (6)$$

Expectation Maximization (EM): We initialize mixing probabilities ρ with uniform probabilities, $\rho_j=1/M, j=1,\dots,M$. At the E step, we calculate the relative responsibility of each nucleosome subtype per nucleosome in generating each fragment as follows: $E(j,s)=\frac{\rho_j\tau_{j,s}Pr(r_i|j,s)}{\sum_{j'=1}^{M}\sum_{s'=1}^{K}Pr(r_i|j',s')}$ (7)

The maximum posterior probability (MAP) estimation of ρ,τ is as follows:

$$\hat{\rho}_j=\frac{Max(N_j-\alpha,0)}{\sum_{j'=1}^{M}Max\,(N_{j'}-\alpha,0)} \qquad (8)$$

$$\widehat{\tau_{j,s}}=\frac{Max(N_{j,s}-\beta N_j,0)}{\sum_{s'=1}^{K}M\,ax(N_{j,s}-\beta N_j,0)} \qquad (9)$$

where N_j represents the number of fragments assigned to nucleosome j, $N_{j,s}$ represents the number of fragments assigned to nucleosome j of type s. The value of α is the minimum number of MNase-seq fragments required to support a nucleosome surviving in this round. In the first five rounds of EM, sparse priors are set to their minimum value (1 for α and 0 for β) to avoid elimination of true nucleosomes due to poor initialization. Then sparse priors are gradually increased in the following five rounds until reaching the default or user-defined values. The default values of α and β are 1 and 0.05, respectively.

MAP values of $\mu_{j,s}$ are determined by finding the location with the maximized probability in +/-50bp flanking sites of the current nucleosome dyad. If the maximization step results in two components sharing the same weighted positions, they are combined in the next iteration of the algorithm.

Considering the efficiency of the algorithm, the fuzziness and the subtype mixture probabilities of each nucleosome are updated every two rounds. In the first several rounds, prior α and β is multiplied by an annealing factor according to the number of completed rounds to avoid too early elimination of nucleosomes. As EM proceeds, the log likelihood increases after each iteration, and convergence is defined when log likelihood increases less than 1% compared to previous iteration.

Exclusion zone: Each nucleosome protects DNA and should sterically exclude other nucleosomes at the same position. For example, canonical nucleosomes protect 147bp DNA, which means the minimum

distance between two canonical nucleosomes should be larger than 147bp. Thus, when there is no nucleosome eliminated during an iteration after sparse prior fully incorporated, an extra step will be taken to remove nucleosomes which are too close to adjacent nucleosomes. Each time the overlapped nucleosome with the lowest responsibility (i.e., occupancy) will be removed, until all nucleosomes have a large enough spacing to each other. The exclusion zone is currently set to 127bp to avoid false removal of nucleosomes due to the inaccurate prediction of nucleosome dyads.

MNase-seq data simulation: To simulate the MNase-seq dataset on the sacCer3 reference genome, we first took the nucleosome dyad location maps from (Jiang and Pugh 2009) as a reference. MNase H4 ChIP-seq datasets from SRR3649286, SRR3649291 (Chereji et al. 2017) were used to infer the nucleosome occupancy and fuzziness for simulation. Specifically, we used the MNase-seq fragments within +/-73bp around each nucleosome dyad to compute the occupancy and fuzziness. Simulated MNase-seq fragments were then generated given the computed nucleosome metrics, distributed following a Gaussian with mean (nucleosome dyad), variance (nucleosome fuzziness), and weight (nucleosome occupancy) parameters set from the above data. Background noise fragments following a Poisson Distribution were added to the simulated dataset to test the robustness of each algorithm. We used background noise ratio 0.05 in this study. The Poisson mean λ is determined by the ratio of background noise among the whole dataset using the formula:

$$\lambda = (\#\ total\ fragments) * ratio_{background}/l_{genome}$$

Nucleosome-calling on MNase-seq datasets: On both simulated and real datasets, SEM, DANPOS, and PuFFIN were run under default setting. Cplate predictions on real MNase-seq dataset were taken from reference (Zhou et al. 2016).

Determining nucleosome subcategories in mESCs: SEM was run on a low-dose MNase H2B ChIP-seq dataset (SRR2034510, SRR2034511) from (Ishii et al. 2015) under default settings with the number of nucleosome subtypes set to 3. Nucleosomes are then filtered such that occupancy is greater than 5 reads and one of the nucleosome subtype probabilities > 0.9. The nucleosome subcategories were determined by overlapping the filtered nucleosomes with ATAC-seq peaks. To ensure the whole nucleosome is located within an accessible region, we required the dyad location of accessible nucleosome has to be at least 100bp away from the boundary of the ATAC-seq peak. The remaining nucleosomes are deemed non-accessible.

Data processing: All FASTQ files were first trimmed by trimmomatic (Bolger et al. 2014) with the options "LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 {adapter_file}:2:30:10" using the corresponding adapters, then trimmed FASTQ files were mapped to the mm10 reference genome by Bowtie2 (Langmead and Salzberg 2012). Mapped BAM files were filtered by the criteria that MAPQ of both reads in a pair >=10. ATAC-seq peak calling was performed by using Genrich (https://github.com/jsh58/Genrich) with ATAC-seq mode enabled by "-j" option.

ChIP-Atlas enrichment analysis: ChIP-seq peak enrichment analysis was performed through the ChIP-Atlas website. Specifically, each nucleosome subcategory was overlapped with peaks from ChIP experiments targeting "TFs and Others" in "Pluripotent stem cell". The threshold for significance was set to 50.

Bayesian Information Criterion (BIC): BIC scores were computed on GMM models fitted with numbers of components set to 2, 3, 4, and 5. The formula is: $BIC = k * ln(n) - 2ln(\hat{L})$, where $\hat{L}$ is the likelihood of the fitted model, k is the number of the parameters, n is the number of data points.

Statistical test on *A/T* content: The ratio of *A/T* nucleotides at nucleosome entry and exit sites was computed by calculating *A/T* nucleotide percentages in the regions bounded by [-100bp, -50bp] and [+50bp, +100bp] relative to the nucleosome dyad. Then an independent *t*-test was performed to compare the percentages at non-accessible short-fragment nucleosomes to those at non-accessible canonical nucleosomes.

Software availability: The SEM software package is available on GitHub (https://github.com/YenLab/SEM) and Bioconda (https://anaconda.org/bioconda/sem). The version of the SEM code used in this manuscript is archived as **Supplemental Code**.

ACKNOWLEDGEMENTS

This work was supported by NIH R35-GM144135 and National Science Foundation DBI CAREER 2045500 (to S.M.). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders. This work was initiated by J.Y. as a Masters student in K.Y.'s lab at the Department of Developmental Biology, School of Basic Medical Sciences, Southern Medical University, Guangzhou, China. Work by the Yen lab is supported by the National Natural Science Foundation of China (grants 31522031 & 31571526). The authors thank the members of the Center for Eukaryotic Gene Regulation at Penn State for helpful feedback and discussions.

REFERENCES

Becker J, Yau C, Hancock JM, Holmes CC. 2013. NucleoFinder: a statistical approach for the detection of nucleosome positions. *Bioinformatics* **29**: 711–716.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Brahma S, Henikoff S. 2019. RSC-Associated Subnucleosomes Define MNase-Sensitive Promoters in Yeast. *Mol Cell* **73**: 238-249.e3.

Chen K, Xi Y, Pan X, Li Z, Kaestner K, Tyler J, Dent S, He X, Li W. 2013a. DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res* **23**: 341–351.

Chen P, Zhao J, Wang Y, Wang M, Long H, Liang D, Huang L, Wen Z, Li W, Li X, et al. 2013b. H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. *Genes and Development* **27**: 2109–2124.

Chereji RV, Ocampo J, Clark DJ. 2017. MNase-Sensitive Complexes in Yeast: Nucleosomes and Non-histone Barriers. *Mol Cell* **65**: 565-577.e3.

Clapier CR, Iwasa J, Cairns BR, Peterson CL. 2017. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nat Rev Mol Cell Biol* **18**: 407–422.

Dingwall C, Lomonossoff GP, Laskey RA. 1981. High sequence specificity of micrococcal nuclease. *Nucleic Acids Res* **9**: 2659–2673.

Hsieh LJ, Gourdet MA, Moore CM, Muñoz EN, Gamarra N, Ramani V, Narlikar GJ. 2022. A hexasome is the preferred substrate for the INO80 chromatin remodeling complex, allowing versatility of function. *Mol Cell*. http://dx.doi.org/10.1016/j.molcel.2022.04.026.

Ishii H, Kadonaga JT, Ren B. 2015. MPE-seq, a new method for the genome-wide analysis of chromatin structure. *Proc Natl Acad Sci U S A* **112**: E3457–E3465.

Jiang C, Pugh BF. 2009. A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome. *Genome Biol* **10**: 1–11.

Jin C, Felsenfeld G. 2007. Nucleosome stability mediated by histone variants H3.3 and H2A.Z. *Genes and Development* **21**: 1519–1529.

Jin C, Zang C, Wei G, Cui K, Peng W, Zhao K, Felsenfeld G. 2009. H3.3/H2A.Z double variant-containing nucleosomes mark “nucleosome-free regions” of active promoters and other regulatory regions. *Nat Genet* **41**: 941–945.

Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. 2005. A high-resolution map of active promoters in the human genome. *Nature* **436**: 876–880.

Klemm SL, Shipony Z, Greenleaf WJ. 2019. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**: 207–220.

Kouzarides T. 2007. Chromatin modifications and their function. *Cell* **128**: 693–705.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**: 251–260.

Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, et al. 2008. Nucleosome organization in the Drosophila genome. *Nature* **453**: 358–362.

Mieczkowski J, Cook A, Bowman SK, Mueller B, Alver BH, Kundu S, Deaton AM, Urban JA, Larschan E, Park PJ, et al. 2016. MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat Commun* **7**: 1–11.

Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, Kawaji H, Nakaki R, Sese J, Meno C. 2018. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep* **19**. http://dx.doi.org/10.15252/embr.201846255.

Ostapcuk V, Mohn F, Carl SH, Basters A, Hess D, Iesmantavicius V, Lampersberger L, Flemr M, Pandey A, Thomä NH, et al. 2018. Activity-dependent neuroprotective protein recruits HP1 and CHD4 to control lineage-specifying genes. *Nature* **557**: 739–743.

Papp B, Müller J. 2006. Histone trimethylation and the maintenance of transcriptional ON and OFF states by trxG and PcG proteins. *Genes Dev* **20**: 2041–2054.

Polishko A, Bunnik EM, Le Roch KG, Lonardi S. 2014. PuFFIN - a parameter-free method to build nucleosome maps from paired-end reads. *BMC Bioinformatics* **15**: 1–10.

Pradhan SK, Su T, Yen L, Jacquet K, Huang C, Côté J, Kurdistani SK, Carey MF. 2016. EP400 Deposits H3.3 into Promoters and Enhancers during Gene Activation. *Mol Cell* **61**: 27–38.

Ramachandran S, Ahmad K, Henikoff S. 2017. Transcription and Remodeling Produce Asymmetrically Unwrapped Nucleosomal Intermediates. *Mol Cell* **68**: 1038-1053.e4.

The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**: 699–710.

Torres IO, Fujimori DG. 2015. Functional coupling between writers, erasers and readers of histone and DNA methylation. *Curr Opin Struct Biol* **35**: 68–75.

Voong LN, Xi L, Sebeson AC, Xiong B, Wang J-P, Wang X. 2016. Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. *Cell* **167**: 1555-1570.e15.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh T-Y, Peng W, Zhang MQ, et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40**: 897–903.

Wen Z, Zhang L, Ruan H, Li G. 2020. Histone variant H2A.Z regulates nucleosome unwrapping and CTCF binding in mouse ES cells. *Nucleic Acids Res* **48**: 5939–5952.

Zentner GE, Henikoff S. 2013. Regulation of nucleosome dynamics by histone modifications. *Nat Struct Mol Biol* **20**: 259–266.

Zhou X, Blocker AW, Airoldi EM, O'Shea EK. 2016. A computational approach to map nucleosome positions and alternative chromatin states with base pair resolution. *Elife* **5**: 1–28.

Zou Z, Ohta T, Miura F, Oki S. 2022. ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res* **50**: W175–W182.