Detection of new pioneer transcription factors as cell-type specific nucleosome binders

Yunhui Peng[1,2*], Wei Song[2], Vladimir B. Teif [3], Ivan Ovcharenko[2], David Landsman[2] and Anna R. Panchenko[4,5,6,7*]

[1] current address: Institute of Biophysics and Department of Physics, Central China Normal University, Wuhan 430079, China

[2] National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

[3] School of Life Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK

[4] Department of Pathology and Molecular Medicine, Queen's University, ON, Canada

[5] Department of Biology and Molecular Sciences, Queen's University, ON, Canada

[6] School of Computing, Queen's University, ON, Canada

[7] Ontario Institute of Cancer Research, Toronto, ON, Canada

*Correspondence: Yunhui Peng yunhuipeng@ccnu.edu.cn

Anna Panchenko anna.panchenko@queensu.ca

Abstract

Wrapping of DNA into nucleosomes restricts accessibility to the DNA and may affect the recognition of binding motifs by transcription factors. A certain class of transcription factors, the pioneer transcription factors, can specifically recognize their DNA binding sites on nucleosomes , initiate local chromatin opening and facilitate the binding of co-factors in a cell-type-specific manner. For the majority of human pioneer transcription factors, the locations of their binding sites, mechanisms of binding and regulation remain unknown. We have developed a computational method to predict the cell-type-specific ability of transcription factors to bind nucleosomes by integrating ChIP-seq, MNase-seq and DNase-seq data with details of nucleosome structure. We have demonstrated the ability of enrichment scores in discriminating pioneer from canonical transcription factors and predicted new potential pioneer transcription factors in H1, K562, HepG2 and HeLa cell lines. Lastly, we systemically analyzed the interaction modes between various pioneer transcription factors and detected several clusters of distinctive binding sites on nucleosomal DNA.

Introduction

In eukaryotic cell, DNA is packaged in the form of chromatin yet, it should be dynamically accessed during transcription and replication processes at high spatiotemporal precision[1]. Open chromatin is thought to comprise actively transcribed genes, while compact chromatin contains repressed genes. However, many recent observations point to a limited association between DNA accessibility, chromatin compaction and gene transcription at the global genomic scale. Indeed, rapid transcription activation may occur by relatively small changes of DNA solvent exposure, and localized nucleosomal array structures can be dynamically exposed without the large-scale chromatin rearrangements[2-5]. Nucleosomes represent the basic subunits of chromatin structure and function. They comprise a histone octamer of four types of core histones of two copies each, and ~147 base pairs of DNA wrapped around them[6]. Wrapping of DNA into nucleosomes inherently restricts DNA accessibility and the recognition of binding motifs by transcription factors. Intrinsically disordered histone tails flank histone core domains and may also modulate DNA accessibility by forming transient interactions with the nucleosomal and linker DNA[7]. The control of the DNA accessibility at nucleosomal and subnucleosomal scales is of major importance in understanding of how certain transcription factors can target compact chromatin to induce transcription activation or repression[8-10].

The differentiation of cells into different lineages occurs through chromatin reprogramming, involving the cooperative behavior of various transcription factors[11,12]. Although nucleosomes generally hinder the binding of transcription factors (TFs), a certain class, so-called pioneer transcription factors (PTF), can specifically recognize their binding sites on nucleosomal DNA, initiating local chromatin opening and facilitating subsequent binding of other co-factors in a cell-type specific manner[2,12]. Several studies have revealed the critical roles of pioneer transcription factors in mediating the cell-type specific gene expression and establishment of cell lineage reprogramming[11-13].

Significant efforts have been made to characterize the interaction landscape between various TFs and nucleosomes[14-18]. Using the NCAP–SELEX approach, a recent study has characterized the interaction modes between nucleosomes and 220 transcription factors[14]. Another high-throughput protein microarray study of 593 human TFs systematically identified the structural features of TFs binding with nucleosomes [19]. It has been revealed that the vast majority of TFs preferably bind naked DNA instead of nucleosomal DNA at

physiological concentrations, whereas certain transcription factors specifically target nucleosomes at different locations and orientations[14-16]. Moreover, several structures of pioneer transacription factors in complex with nucleosomes have recently been resolved[20-23]. Despite significant advances in recent experimental approaches, the interaction modes of most transcription factors with nucleosomes remain obscure and their cell-type-specific pioneer activities are largely unknown. The development of new computational approaches can improve our understanding of binding properties of various TFs with nucleosomes, thereby helping to identify novel pioneer transcription factors.

Through the advances in high-throughput sequencing techniques, a large volume of data has been generated (*e.g.* ChIP–seq, ATAC-seq, Dnase-seq and MNase-seq data[24]), which allows us to gain insights into the chromatin structure and details of epigenetic binding events. The rapid growth of such data sets has further stimulated the development of computational methods to characterize cell-type specific transcription factors [25-28]. Several machine learning models have been proposed to predict transcription factor binding sites and identify sequence context features critical for TF binding[25,27-29]. In addition, gene regulatory network-based approaches have helped identify the key transcription factors in cell fate determination[30,31]. Despite the success of computational methods in genome-wide prediction of binding sites of canonical transcription factors, the locations of pioneer transcription factors' binding sites, mechanisms of binding and regulation have not been systematically explored.

Integrating data on nucleosome positioning and DNA accessibility (MNase-seq, ATAC-seq and DNase-seq) with the data on DNA binding events, available through ChIP-seq and other methods, can reveal the interplay between transcription factor binding and nucleosome positioning, providing insights into the mechanisms of pioneer transcription factor interactions with nucleosomes and their ability to modulate chromatin accessibility[32-35].We have developed a computational method to study the ability of transcription factors to bind nucleosomes by using ChIP-seq, MNaseq-seq and DNase-seq data sets from five different cell lines and (Figure 1). Our results point to the capability of our method to discriminate between pioneer and canonical transcription factors using experimental benchmarks. Additionally, we have predicted several transcription factors as potentially new cell-type specific pioneer transcription factors in H1,K562, HepG2 and HeLa cell lines and performed multiple validations. Lastly, we systemically analyzed the interaction modes between various pioneer transcription factors and nucleosomal DNA and detected several clusters of distinctive binding sites on nucleosomal DNA.

Methods

Genome-wide mapping of nucleosome dyads and footprint regions

The overall workflow of our computational framework is shown in Figure 1. High-coverage micrococcal nuclease sequencing data (MNase-seq) of five human cell lines (H1, HepG2, MCF-7, K562, and HeLa) from paired-end sequencing was used for nucleosome mapping (Supplementary Table 1). The raw MNase-seq data was downloaded from the NCBI Sequence Read Archive (SRA) and converted into the fastq format using SRA Toolkit[36]. Then, the downloaded fastq files were processed using the mnaseseq pipeline from nf-core (https://github.com/nf-core/mnaseseq)[37], a recently developed bioinformatics pipeline for MNase seq data analysis. Adapter trimming of sequencing reads was performed with Trim Galore. Then, adapter-trimmed reads were mapped to the reference genome using Burrows-Wheeler Aligner (BWA)[38]. Human genome GRCh37 was used as a reference genome for reads mapping and the maximum number of mismatches in alignment was set to 4. The minimum and maximum insert sizes for filtering of mono-nucleosome paired-end reads were set to 120 and 180 base pairs. Duplicate reads were marked using Picard MarkDuplicates command (http://broadinstitute.github.io/picard/) and discarded from the analysis to avoid PCR duplication artifacts (less than 10% of reads were duplicated). Read libraries of replicates from the same experiment condition were merged into the analysis. Next, the BAM sequence alignment files were converted into BED format using bedtools v2.30.0[39]. The aligned reads with fragment sizes from 146 to 148 base pairs were selected for mapping of the representative dyad positions (center of the nucleosomal DNA).

To determine the representative dyad position of nucleosomes, we implemented and modified a previously developed nucleosome mapping protocol[40] and dyad positions were determined as midpoints of mapped MNase-seq reads[41]. Namely, we used a triweight kernel K as a weighting function (Equation 1) and the kernel-smoothed dyad counts D at each genomic coordinate i is calculated as [40] .

$$K(u) = \begin{cases} (1 - (\frac{u}{h})^2)^3 & |u| \leq h \\ 0 & |u| > h \end{cases} \quad (1)$$

$$D(i) = \frac{1}{Nh} \sum_{j=1}^{N} K(i - j) \cdot d(j) \quad (2)$$

Here N is the length of a chromosome, $d(j)$ is the dyad count at genomic coordinate j and $D(i)$ is the smoothed dyad count at genomic coordinate i. Small values of bandwidth h lead to less smoothing but more accurate estimates of dyad positions. In our case, we chose a relatively small bandwidth value h=15 to improve the accuracy of the mapped dyad positions.

Next, we identified the genomic locations with the local maximum values of the smoothed dyad counts using bwtool[42] and the minimum distance between the neighboring local maxima was set to 150 base pair with "find local-extrema -maxima -min-sep=150". Then, for every 60-bp window centered at each local maxima, one representative dyad location was determined as the dyad location with the highest number of dyad counts in this interval. Other dyad positions within the same 60-bp interval were discarded. If two or more dyad positions had the same dyad counts within the same interval, the dyad position located closest to the local maximum of the smoothed counts was selected as the representative dyad. Last, nucleosome regions (NRs) were determined as genomic regions centered at the representative dyad positions and flanked by 73 bp segments on each side.

Genome-wide mapping of active enhancers, open chromatin and nucleosome-depleted regions

To map the open chromatin and enhancer regions, we used DNase-seq, H3K27ac ChIP-seq and H3K4me1 ChIP-seq data from the following five human cell lines (H1, HepG2, MCF7, K562, and HeLa-S3, Supplementary Table 1 and 2). The narrowPeak files were downloaded from the ENCODE and Roadmap project[43,44]. The open chromatin regions were identified as genomic regions centered at narrow peaks and flanked by 1000 bp segments on each side. The active enhancer regions were defined as the open chromatin regions overlapped with both H3K27ac and H3K4me1 ChIP-seq narrow peaks. The nucleosome-depleted regions (NDRs) represent genomic regions free of nucleosomes and were identified as genomic regions not covered by any mono-nucleosome fragments (120 to 180 base pairs length) in MNase-seq data in all replicate experiments.

Using open chromatin regions from the DNase-seq data, we identified *differentially* and *conserved open chromatin* regions using the "intersect" command from the BEDTools suite[39]. Conserved open chromatin regions represent open chromatin regions that are more than 80% shared between embryonic H1 and at least one other differentiated cell line used in this

study[45] (Supplementary Figure 1). Chromatin regions with differential accessibility ("*differentially open chromatin regions*") are defined as those that have less than 20% overlap between open regions in H1 embryonic cell line and any other differentiated cell line[45] (Supplementary Figure 1). For motif enrichment analysis of transcription factors, we selected two sets of NRs and NDRs: 1) NDRs located in the open chromatin regions and all identified NRs using the MNase-seq data; 2) NDRs located in conserved open chromatin regions and NRs located in the *differentially* open regions so that their accessibility may be associated with TF binding.

Analysis of dinucleotide patterns of nucleosomal DNA

Using identified representative nucleosome dyad positions, we examined dinucleotide patterns of nucleosomal DNA and mapped genomic coordinates of WW/SS (where W is A or T, and S is G or C) and YY/RR (R = A or G, and Y = C or T) dinucleotides on nucleosome regions (NRs). Then, we aligned NRs by superimposing their dyad positions and computed the frequency of observed dinucleotides at each location of nucleosomal DNA (as a function of distance in base pairs from the nucleosome dyad). As a result, we observed pronounced dinucleotide patterns at specific nucleosomal DNA positions for all cell lines which are indicative of the high qulity of the data and nucleosome mapping procedure (Supplementary Figure 2).

Genome-wide mapping of transcription factor binding sites

To map the genome-wide locations of binding sites of various transcription factors, we matched the ChIP-seq data from the ENCODE project[43] with the MNase-seq data for the same human cell lines (H1, HepG2, MCF-7, HeLa-S3 and K562) (Supplementary Table 3). In case of HeLA cell line, we used HeLa-S3 ChiP-seq data, which is a clonal derivative of HeLA. In total, ChIP-seq data for 225 transcription factors could be matched with the corresponding MNase-seq data from the same cell type. All available narrow peak files of ChIP-seq of these transcription factors were downloaded, and files corresponding to the same transcription factor were merged for further analysis. To map binding sites from ChIP-seq narrow peaks, referred to as "ChiP-seq TF motif", we downloaded position frequency matrices (PFMs) for each TF from the JASPAR CORE database [46]. Then we applied FIMO program [47] to scan the DNA sequences within ChIP-seq narrow peaks using each PFM and identified motifs with P-values less than 10^{-4} for further analysis.

Next, we calculated the transcription factor binding motif profiles or briefly *"motif profiles"*. To do that, we first mapped ChIP-seq TF motifs for a given TF to the closest nucleosome dyad position and then aligned all dyad positions and corresponding ChIP-seq TF motifs. For each TF binding motif, we considered sequences in both strands of DNA. Then, we counted the number of base pairs from different TF binding motifs that mapped at each location of nucleosomal DNA and flanking DNA regions (+/- 1000 bp) around the dyad.

Transcription factor motif enrichment analysis

Pioneering transcription factors (PTFs) can engage nucleosomal DNA, while truly canonical TFs cannot bind to nucleosomes. To predict pioneer transcription factors, we calculated the binding motif enrichment of different TFs on nucleosomal regions (NRs) compared to nucleosome-depleted regions (NDRs). Namely, we counted a number of base pairs of ChIP-seq TF motifs overlapped with the NRs and NDRs regions and constructed the following contingency table (Supplementary Table 4). We then calculated the enrichment score for each transcription factor (Equation 4) and applied a Fisher's exact test to evaluate the significance of TF motif enrichment on NRs compared with NDRs.

$$\textit{Enrichment score (Odds ratio)} = (a/c)/(b/d) \qquad (4)$$

Here a and c are the number of base pairs overlapped with ChIP-seq TF motifs on NR and NDR respectively. Counts b and d correspond to the number of base pairs outside of ChIP-seq TF motifs on NR and NDR respectively. We further collected the data on expression levels of various TFs from the NIH roadmap epigenomics program (https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/) and identified TFs which are highly expressed in each cell line (RPKM value >= 10).

Receiver operating curves (ROC) and precision-recall (PR) curve analyses were further performed to evaluate the power of enrichment scores in discriminating pioneer transcription factors from other transcription factors. 32 known pioneer transcription factors from the literature (which were included in our list of 225 TFs) were used as positives (Supplementary Table 5) and other transcription factors were considered as negatives for this test.

Clustering of transcription factor binding motif profiles

To identify the prevalent interaction modes of various TFs with nucleosomes, we have performed clustering of TF binding motif profiles. Prior to clustering, we filtered out potentially low-quality binding profiles using the following criteria. First, we removed under-represented TFs with the total genome-wide cumulative sum of ChiP-seq TF motifs on nucleosome regions of less than 500 base pairs. Second, due to the two-fold symmetry of DNA in nucleosome structures, ChiP-seq TF motifs on DNA complementary plus and minus strands should be structurally superimposed if the nucleosome structure is rotated by 180 degrees. If we are analyzing relative large number of binding motifs on both DNA strands, binding motif profiles should be symmetrical with respect to the nucleosome dyad. Therefore, we calculated the Pearson correlation coefficient (PCC) of motif profiles between two symmetrical nucleosomal halves (positive and negative superhelical locations of nucleosomal DNA) for each TF and removed those with Pearson correlation coefficient values less than 0.4.

Next, we applied t-distributed stochastic neighbor embedding (t-SNE) and projected all profiles onto two dimensions using the Rtsne function from the R package[48]. Then, the projected data were subjected to k-medoids clustering using the pam function from the R package with the optimal number of clusters equal to six (Supplementary Figure 9). The silhouette width is an estimate of the goodness of clusters, its values close to 1 correspond to a cluster where most objects are much closer to other objects in the same cluster than to other clusters. For each cluster, members with silhouette width <= 0.25 were considered as outliers and 33 outliers removed.

Since Micrococcal Nuclease has a sequence bias and cleaves DNA upstream of A or T more efficiently than of G or C nucleotides, a certain nucleotide preference exists around the ends of nucleosomal DNA reads produced by the MNase-seq experiments. It may potentially bias the TF binding profiles near the nucleosomal DNA ends. Therefore, in our analysis, we excluded regions near the nucleosomal DNA ends. To identify binding modes between TFs and nucleosomes, we clustered binding motif profiles of different TFs centered at the nucleosomal dyad locations (+/- 60 base pair from dyad) using k-medoids.

Results

Transcription factor binding motif enrichment analysis can identify pioneer transcription factors

Nucleosomes are generally considered an impediment to the binding of TFs to DNA and thus binding sites of TFs are typically depleted in DNA regions with high nucleosome occupancy. However, pioneer TFs can recognize their binding motifs on nucleosomal DNA and trigger the opening of chromatin to recruit other TFs in a cell-type-specific fashion [2,49]. Therefore, we hypothesized that DNA binding sites of pioneer transcription factors should not be depleted and, in some cases, should be enriched on nucleosome regions (NRs). At the same time, those canonical transcription factors that can preferentially bind to the naked DNA should exhibit the depletion of binding sites on nucleosome regions and enrichment in nucleosome-free regions. To quantitively assess these trends, we have performed binding motif enrichment analysis for each of the 225 transcription factors and calculated the motif enrichment scores (Supplementary Table 6-9). We have further analyzed the enrichment of TF binding sites on nucleosome regions (NRs) compared to the nucleosome depleted regions (NDRs)(see Methods for definition).

To validate our results, from the literature, we have compiled three sets of known pioneer transcription factors (which could also be found in our list of 225 TFs) as positives (Supplementary Table 5 and 10) whereas other transcription factors were considered as negatives for this test. Test set 1 includes 32 known pioneer transcription factors from our data set. Some known PTFs were not included as they were not present in the original data sets. Test set 1 comprises Test set 2 and 3 and other pioneer transcription factors. Test set 2 includes 11 known pioneer transcription factors with roles in cell differentiation. Test set 3 includes six known pioneer transcription factors critical for the maintenance of embryonic stem cells or the reprogramming of somatic cells into induced pluripotent stem cells (Supplementary Table 10). As we show in the next section, the negative set may also contain pioneer transcription factors, therefore the classification accuracy values provided below can be considered as a lower bound estimate. Then, we performed the enrichment anlysis by calculating the binding motif enrichment of different TFs on nucleosomal regions compared to nucleosome-depleted regions.

The enrichment score of the 32 known PTFs from Test set 1 was found to be significantly higher than for other factors (p-value = $2.12*10^{-7}$ for all TFs and p-value = $4.33*10^{-5}$ for highly expressed TFs, Mann Whitney U test, Supplementary Figure 3 and Supplementary

Table 11). The results also show the ability of enrichment scores in the classification of pioneer transcription factor activity (Supplementary Table 11, ROC AUC = 0.69, PR AUC = 0.33 and maximal MCC= 0.31 for all TFs and ROC AUC = 0.71, PR AUC = 0.39 and maximal MCC= 0.31 for highly expressed TFs).

Validation pertaining to the ability of pioneer transcription factors to open the closed chromatin was assessed. We hypothesized that the enrichment score calculated based on NRs located in differentially open chromatin regions and NDRs located in conserved open chromatin regions would perform best in the classification of pioneer transcription factors essential for cell differentiation (Test set 2). As can be seen in Figure 2 and Supplementary Table 11, it is indeed the case and the classification accuracy increases from ROC AUC = 0.69 to 0.89 (from 0.71 to 92 for expressed TFs) upon the inclusion of differentially open regions compared to open regions (maximal Maximal Matthews correlation coefficient increased from 0.31 to 0.42). We found that known pioneer transcription factors that acted as key regulators of cell differentiation had the highest enrichment scores in our ranking (Figure 2c). These cases (Test set 2) mainly included pioneer transcription factors from the FOXA, GATA, and CEBP families [50-52] with GATA1 and GATA2 showing the highest enrichment scores.

Interestingly, pioneer transcription factors in Test set 3 (responsible for the maintenance of embryonic stem cell or reprogramming of somatic cells) showed significantly lower enrichment scores compared with other TFs (Figure 2c and Supplementary Figure 4). Yamanaka pioneer transcription factors (POU5F1/OCT4 and KLF4)[53] were strongly depleted at nucleosomes (Figure 2c). It has been previously shown that Yamanaka pioneer transcription factors might target partial sequence motifs on nucleosomal DNA and require other factors for their binding to nucleosomes[16] and therefore their enrichment score might not be expected to be high. We also found known pioneer transcription factors with relatively low enrichment scores including NFYA, NFYB, NFYC and ESRRB. These, transcription factors regulate stem cell proliferation and maintenance of stem cell identity[54-56] (Supplementary Figure 4). If we repeat our anaysis by redefining differentially open regions as those closed in differentiated cell lines and open in H1 embryonic cell line, then ESSRB and Yamanaka pioneer transcription factor POU5F1 (OCT4) showed significantly higher enrichment scores (Supplementary Figure 5). This could be explained by the roles of

Yamanaka factors in cellular reprogramming – they reprogram somatic differentiated cells into induced pluripotent stem cells.

Since PTFs often target enhancers, we have repeated the enrichment anlysis using NRs located in differentially active enhancer regions and NDRs located in the conserved active enhancer regions and the performance in the PTF classification was slightly worse (Supplementary Figure 6). Here, differential enhancer regions refer to the active enhancer regions in an embryonic cell line (H1) that have less than 20% overlap with any active enhancer regions in differentiated cell lines. Conserved active enhancer regions represent active enhancer regions that are more than 80% shared between embryonic H1 and at least one other differentiated cell line used in this study. We also explored different thresholds in defining *differentially open* and *conserved open chromatin regions* but the performance in classifications of PTF was not significantly affected by the threshold choice (Supplementary Figure 7).

Finally, we performed an additional validation using recent data from the high-throughput protein microarray and electromobility shift assays (EMSA) experiments on human TFs which systematically assessed transcription factor binding preferences to nucleosomal DNA versus naked DNA[19]. These authors classified all studied transcription factors with respect to their strengths of nucleosome binding into three clusters: strong binders, which bind strongly to both naked DNA and nucleosomal DNA (cluster 1) and weak binders, which bind weakly to both naked DNA and nucleosomal DNA (cluster 2). Cluster 3 consists of strong binders which bind strongly to naked DNA but weakly to nucleosomal DNA. We found that TFs from their cluster 3 had the lowest enrichment scores although this trend was not statistically significant because of a lack of the data in this cluster. As to TFs in cluster 1 and 2, half of them had binding sites enriched on nucleosome regions (including known pioneer transcription factors FOXA1, GATA4 and CEBPA) which correspond to strong nucleosome binders, and another half had binding sites enriched on nucleosome depleted regions (Figure 2e).

Prediction of novel transcription factors as potential pioneer transcription factors

Using gene expression data from the Roadmap Epigenomics Program, we have identified 35 TFs in H1, K562, HepG2 and HeLa cell lines (no expression data was available for MCF-7) that ranked highly in the enrichment analysis and were also highly expressed in the

corresponding cell lines (Supplementary Table 7). Among these 35 TFs, 15 TFs were well-characterized PTFs such as GATA, FOXA and CEBP factors, ESRRB, NEUROD1, SPI1 and subunits of the AP-1 complex. To validate the remaining 20 PTF predictions, we performed literature searches and found that RFX5 had the ability to displace nucleosomes[14]. In addition, many of the predicted PTFs such as ZKSCAN1, USF1,USF2 and SRF, were confirmed by a recent study[57].

Next, using the enrichment score calculated based on NRs located in differentially open chromatin regions and NDRs located in conserved open chromatin regions, we identified 40 transcription factors that could act as pioneer transcription factor with essential roles in cell differentiation in K562, HepG2 and HeLa cell lines. These identified TFs had their DNA binding sites significantly enriched on nucleosomal DNA of differentially open chromatin regions and were highly expressed in the corresponding cell lines (Supplementary Table 9). Among these 40 TFs, 15 were well-characterized PTFs including GATA, FOXA and CEBP factors and subunits of AP-1 complex. For the remaining 25 PTF predictions, seven transcription factors were annotated in the literature as potential pioneer transcription factors and/or potential nucleosome binders (Supplementary Table 12). For instance, HNF4A was annotated as a potential pioneer transcription factor active in chromatin remodeling in the liver[58], LEF1 was identified as a regulatory high mobility group (HMG) box protein that could bind to nucleosomes [59] and CUX1 could specifically interact with its recognition motif in a nucleosomal context [60].

Association between binding motif profiles and nucleosome occupancy for pioneer transcription factors

The enrichment analysis described above is the first step to estimate the propensity of transcription factor binding sites to be located on nucleosomal footprints. The enrichment analysis might have high specificity but a compromised sensititivity as it can misclassify those pioneer transcription factors that can bind to both naked DNA and nucleosomes at similar concentrations. Therefore, the next step would be to evaluate the actual locations of TF binding sites with respect to the nucleosomal dyad. To this end, we tested if there is a significant association between the binding motif profiles (see Methods) and nucleosome occupancy values +/-400 bp around the dyad (Figure 3a). Our results showed that 87% of all 225 transcription factors had negative Pearson correlation coefficients (PCC) between binding motif profile and nucleosome occupancy values (Figure 3b) which is consistent with

the fact that nucleosomes generally restrict the access of TFs to their binding sites on DNA molecules.

We have also identified 37 transcription factors that had a statistically significant positive correlation coefficients between binding motif profile and nucleosome occupancy and could be classified as potential nucleosome binders and pioneer transcription factors (Supplementary Table 13 and Figure 3). Among these predictions there were six known PTFs, such as POU5F1(OCT4), GATA3, CEBPB, ATF2, NFYA and NFYB (Supplementary Table 13). As we mentioned previously, NFYB, NFYA and POU5F1 had low enrichment scores, but were identified as nucleosome binders by this analysis. It could be explained by their dual nature: these factors can bind to nucleosomes using certain binding arrangements, as evident from their postive correlation coefficients, but at the same time, they can interact with the naked DNA, as evident from their low enrichment scores. Supplementary Table 13 shows 14 predictions that overalpped with the predicted PTFs from the enrichment analysis (ATF2, BACH1, CEBPB, ESRRA, HMBOX1, NFATC3, NFYB, SREBF2, USF1, USF2, ZNF24, ZNF274, ZNF282, ZKSCAN1). Among those with a high PCC > 0.5, were ZKSCAN1, ESR1, NFATC3, ZBTB7B, MAX, and TBX2. Half of them were also predicted to be pioneer by a recent study[57]. With the exception of a few cases, none of transcription factors was identified as being pioneer in all cell types because pioneer activity is often cell-type specific. For example, CTCF was identified as having a significant correlation (although low) between binding motif profile and nucleosome occupancy for the human embryonic stem cell line (PCC = 0.1, p-values < 0.05), but not for other somatic cell lines (Figure 3d). Indeed, previous studies have indicated that CTCF proteins could access the binding sites in nucleosomes and may function as pioneer transcription factors in embryonic stem cells and to a lesser extent in differentiated cells[61-63].

Deciphering the interaction modes between transcription factors and nucleosomes

Using nucleosome consecutive affinity purification–systematic evolution of ligands by exponential enrichment (NCAP–SELEX) approach, a recent study characterized the interaction landscape between pioneer transcription factors and nucleosomes. It revealed different binding modes of pioneer transcription factors: DNA end binding, dyad binding, gyre binding and periodic binding[14]. The NCAP–SELEX approach was based on the analysis of enrichment of specific sequences from the DNA libraries. These sequences were reconstituted into nucleosomes, and incubated with TFs[14]. Then, the dissociated nucleosomal

DNA was separated from the intact nucleosomes and the analysis of the enrichment of sequences allowed to identify TF binding specificities and binding site locations on nucleosomal DNA[14]. To compare our TF motif profiles with NCAP–SELEX data, we filtered out low-quality motif profiles using the criteria described in the Methods section and then calculated motif enrichment scores for TFs with different binding modes identified by NCAP–SELEX approach. Due to the limited number of transcription factors observed in both our dataset and NCAP–SELEX study (24 transcription factors), we mainly focused on the DNA end binding super helical locations (SHL) from +/- 5.5 to +/- 7 and the dyad binding modes with SHL from 0 to +/- 1.5. To estimate the preferential binding of TFs to the ends of nucleosomal DNA compared to the nucleosomal dyad, we calculated the end/dyad binding ratio ($R_{end/dyad}$) as the number of binding motifs at the DNA ends (SHLs from +/-5.5 to +/-7) divided by the number of binding motifs near dyad regions (SHLs from 0 to +/-1.5).

In NCAP-SELEX experimental analyses, to quantify the preference of PTF binding to nucleosomes[14], binding signals were compared by calculating the mutual information (MI) content between 3-mer distributions at two non-overlapping positions of the ligand, aimed at finding if SELEX ligand may contact these positions at the same time. Since nucleosomes can form on most DNA sequences, whereas TFs bind to only a few specific sequences, the NCAP-SELEX study calculated the enriched MI (E-MI) score to separate the TF signals from nucleosome signals by limiting the MI measure to the top ten most enriched 3-mer pairs. E-MI penetration score corresponds to E-MI drop by half compared to the E-MI maximum and larger values pointed to the favorable binding to the dyad regions[14]. As our $R_{end/dyad}$ and experimental EMI penetration values should be anti-correlated, we indeed observed a statistically significant negative correlation between $R_{end/dyad}$ and EMI penetration values (PCC = -0.45, p-value = 0.025). This shows that our computational analysis is able to capture the detailed features of TF binding motifs on nucleosomes. For the EMI intensity, although a negative linear dependence trend was evident, the correlation was not significant (Supplementary Figure 8).

Clustering reveals several groups of pioneer transcription factor binding sites

Previous studies indicated that while nucleosome positioning acts as a barrier to TF binding, there is a number of different signatures of mutual TF/nucleosome positioning[64]. We attempt to characterize the details of PTF binding motifs at near-single nucleotide resolution on

nucleosomal DNA and performed K-medoids clustering of binding motif profiles for all 225 transcription factors (Figure 4 and Supplementary Figure 9). Transcription factors in the first and second cluster showed that motif density increased with the distance from the dyad. These clusters include cases of canonical factors or those PTF that preferentially interact with the ends of nucleosomal DNA from super helical locations (SHL) +/- 5.5 to SHL +/- 6. Such pioneer transcription factors include CEBPB and GATA3, consistent with the NCAP-SELEX data. In addition, it has been recently shown that GATA3 factor targets nucleosomal DNA around the SHL +/-5.5 position[23]. In clusters 4 and 6, TFs preferentially interact with the nucleosomal DNA around the dyad regions and SHL from +/- 3 to SHL +/- 5, whereas transcription factors from clusters 3 and 5 preferentially occupy SHL +/- 1 to SHL +/- 2.5 respectively.

We found that many transcription factors from the same family and even from different cell types were assigned to the same cluster. For example, all known pioneer transcription factors from FOXA families belong to cluster 4 and indeed it has been shown that FOXA1 can target its binding motifs to nucleosomal DNA near the dyad and linker regions[65]. However, there are a few exceptions, CEPBP, FOSL, CREB1, CTCF, MEF2A, NFYA, NFYB and YY1, where the same transcription factor from different cell lines has been assigned to different clusters. Possible reasons may include a non-linear binding motifs on DNA which can be close in space but occupy different regions on DNA. In addition, the cell-type-specific epigenetic marks recognized by pioneer transcription factors or cooperativity binding between factors are differentially expressed in various cell types.

Discussion

The modulation of chromatin accessibility with high spatiotemporal precision is a subject of ongoing debate as it is a crucial factor in regulation of transcription, replication and DNA repair processes. The question of how chromatin accessibility is precisely modulated involves multiple aspects, including epigenetic modifications, PTF infiltration, ATP-dependent remodeling and spontaneous chromatin dynamics. There are about two thousand transcription factors in the human genome and a few hundred pioneer transcription factors among them. Yet, for the vast majority of human PTFs, the locations of their binding sites and mechanisms of binding and regulation remain unknown[19,66]. There are many different experimental assays that can provide information on transcription factor binding sites but these assays suffer from multiple drawbacks and often require prior knowledge of the TFs being tested. Moreover,

pioneer transcription factors are very dynamic, may target partial binding sites and work cooperatively with other factors. For example, two recent cryo-EM structures of the pioneer transcription factor SOX2-nucleosome complexes showed different mechanisms of binding[21,22]. All this complicates the identification and characterization of human pioneer transcription factors pointing to the pressing need to develop predictors and classifiers [49,67-69] .

The goal of this study has been to gain functional insights into the mechanisms of binding and infiltration of PTFs into chromatin at the level of the nucleosome. To achieve this goal, we used the ChIP-seq, MNaseq-seq and DNase-seq data from the five different cell lines for 225 human transcription factors. A computational framework has been to systemically investigate the ability of transcription factors to bind to nucleosomes. As a result, we found that using the information on differentially open chromatin regions (open in one cell line, closed in another) leads to the highest classification accuracy in discriminating pioneer from canonical transcription factors. This finding supports the view that TF binding to nucleosomes leads to DNA and chromatin opening and might correlate with the reprogramming potential[19]. Our study has also verified the known and predicted several dozens of new pioneer transcription factors as nucleosome binders (Supplementary 7, 9 and 13). These predicted cases include transcription factors without previously known pioneer activity which could be subject to the future experimental validation. Finally, we have identified six distinctive clusters of TF binding profiles with the nucleosomal DNA. These clusters point to the diversity of binding motifs where transcription factors belonging to the same cluster may exhibit potential competitive binding.

We should mention that our classification method and the data used in this study have certain limitations. First, interactions of PTFs with nucleosomes may depend on binding of other factors. Second, PTF binding relies on specific recognition of DNA binding sites and can be governed by nucleosome dynamics. Indeed, initial stages of binding can happen via the DNA site exposure through thermal fluctuations, leading to DNA unwrapping and the formation of DNA looping or twist defects[70-72]. Pioneer transcription factors can exploit DNA unwrapping and trap nucleosomes in a partially unwrapped states[73]. Transcription factor binding motifs close to the nucleosome entry-exit sites may have increased exposure but would be error-prone to capture using MNase-seq data as a nucleotide bias exists around the ends of nucleosomal DNA reads produced by the MNase-seq experiments. Moreover, as we showed recently the degree of spontaneous DNA unwrapping from nucleosomes depends on the

incorporation of histone variants, like H2A.Z[74], which is, in turn, might be needed for pioneer transcription factor recruitment to promote embryonic stem cell differentiation[75]. The third important group of limitations pertains to the nucleosome positioning in a genome, as binding dissociation constants for PTFs depend on where nucleosomes are located. Indeed, nucleosome positioning and stability is not uniform throughout the genome and depends on local nucleosomal DNA sequence, histone variant deposition and epigenetic chromatin modifications. Finally, pioneer transcription factors may exhibit multivalent binding recognizing not only DNA binding sites but also some parts of the histone core or histone tail regions. Deducing such dependencies from the experimental essays utilized in this study is very challenging, if not impossible.

Nucleosomes represent hub points in epigenetic signaling pathways and identifying complex epigenetic relationships at the level of single nucleosomes may yield functional insights into the mechanisms of binding and infiltration of pioneer transcription factors into chromatin during differentiation and reprogramming. Transcription factors regulate a large number of signaling pathways and their dysregulation contributes to a plethora of human diseases, including diabetes, cardiovascular diseases and many cancers. Conventional transcription factors have been used for a long time as biomarkers and drug targets, however, the targeted therapeutic potential of pioneer transcription factors is lagging. To fill this gap, integrative approaches using large-scale low- or medium-resolution data with precise molecular modeling or protein-protein docking may provide the required detailed characterization of many predicted PTFs in the future.

Acknowledgments

ARP was supported by the Department of Pathology and Molecular Medicine, Queen's University, Canada. ARP is the recipient of a Senior Canada Research Chair in Computational Biology and Biophysics and a Senior Investigator Award from the Ontario Institute of Cancer Research, Canada. ARP acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) (No. RGPIN/02972-2021 ARP). ARP would like to thank David Clark for helpful discussions. YP was supported by the National Natural Science Foundation of China (No.12205112) and Fundamental Research Funds for Central China Normal University. YP, WS, IO and DL were supported by the Intramural Research Program of the National Library of Medicine, NIH. VBT acknowledges

support by Cancer Research UK (grants EDDPMA-Nov21\100044 and SEBPCTA-2022/100001).

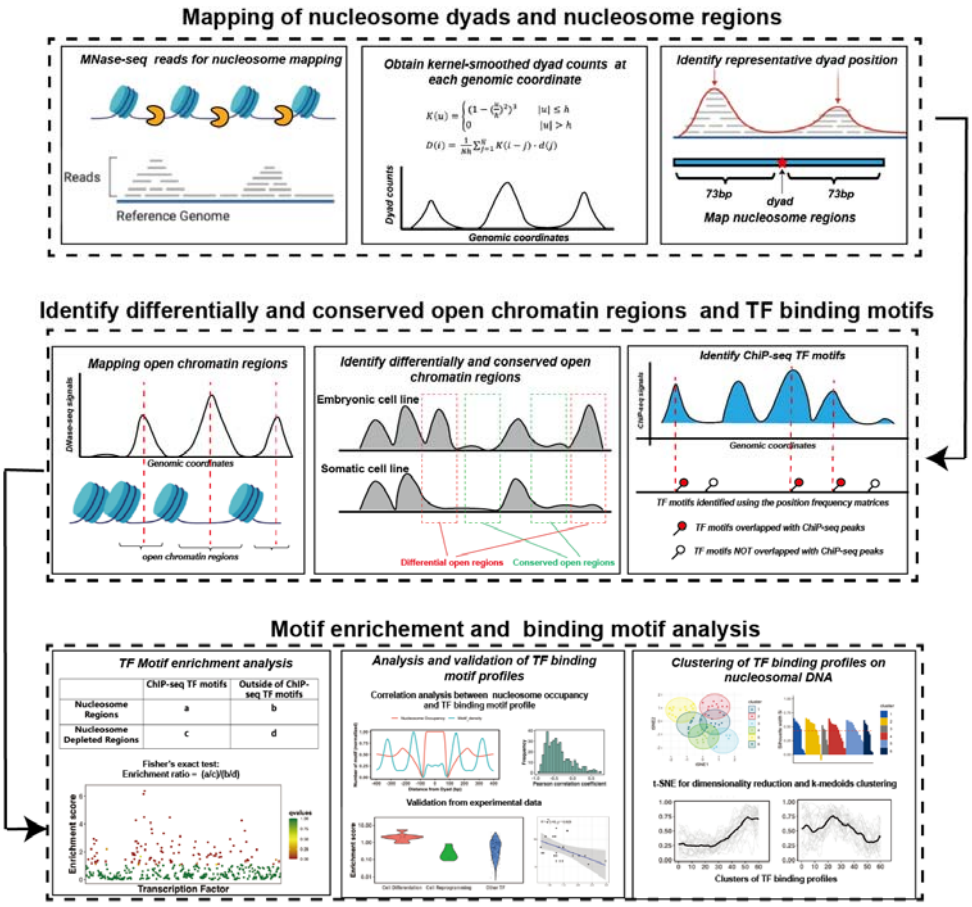

Figure 1. The computational framework to analyze the ability of transcription factors to bind to nucleosomes by integrating ChIP-seq, MNaseq-seq and DNase-seq data for motif enrichment and binding motif analysis.

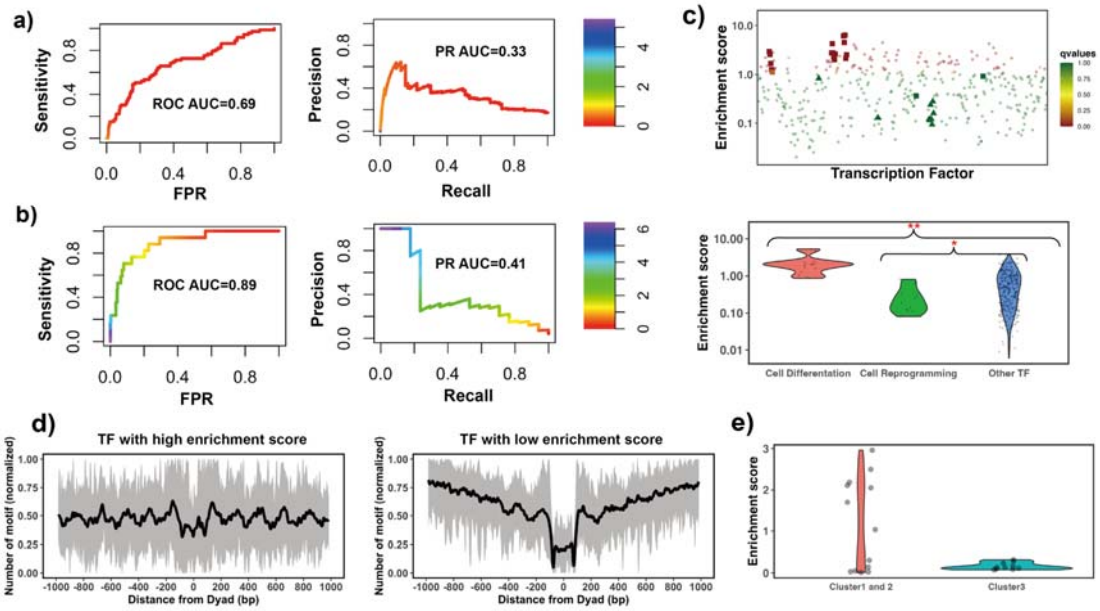

Figure 2. Identifying pioneer transcription factors using motif enrichment analysis. a) TF motif enrichment score is used to distinguish 32 known PTFs (Test set 1) from other other TFs. Receiver operating curve (ROC) and precision-recall (PR) curve analysis of motif enrichment scores are performed. Here nucleosome regions (NRs) were determined as genomic regions (147 bp long) centered at the representative dyad positions and the nucleosome-depleted regions (NDRs) represent genomic regions free of nucleosomes and are located in open chromatin regions.**b)** TF motif enrichment score is used to distinguish 11 known PTFs with essential roles in cell differentiation (Test set 2) from other other TFs. Receiver operating curve (ROC) and precision-recall (PR) curve analyses of motif enrichment scores are performed. Here NRs in differentially open and NDRs in conserved open chromatin regions are used in enrichment analysis. **c)** Classification of pioneer transcription factors by binding motif enrichment scores. Known pioneer transcription factors from Test set 2 and Test set 3 are indicated by squares and triangles, while other TFs are shown as circles. Colors corresponds to false discovery rate (FDR) q-values. Mann–Whitney U tests are performed under the null hypothesis that PTF's mean values of enrichment scores are equal to canonical TFs. * - p-value < 0.05; ** - p-value < 0.005. **d)** Binding motif profile of TFs with the highest and lowest motif enrichment scores (ranked at the top or bottom 20% among all TFs). The number of motifs for each TF is normalized within the range between 0 and 1 as follows: $X(i)_{normalized}) = (X(i) - X_{min})/(X_{max} - X_{min})$, X(i) is the number of sequences which have TF binding sites at the i_{th} base pair from the nucleosomal dyad position; X_{max} and X_{min} represent the maximal and minimal counts of sequence fragments. **e)** Comparison of the enrichment score of TFs in different clusters identified from recent EMSA experiments[19]. Only 13 transcription factors could be found in EMSA and our data set. Cluster 1 and 2 include strong binders to both naked DNA and nucleosomal DNA andweak binders to both naked DNA and nucleosomal DNA (only one TF). Cluster 3: strong binders to naked DNA but weak binders to nucleosomal DNA.

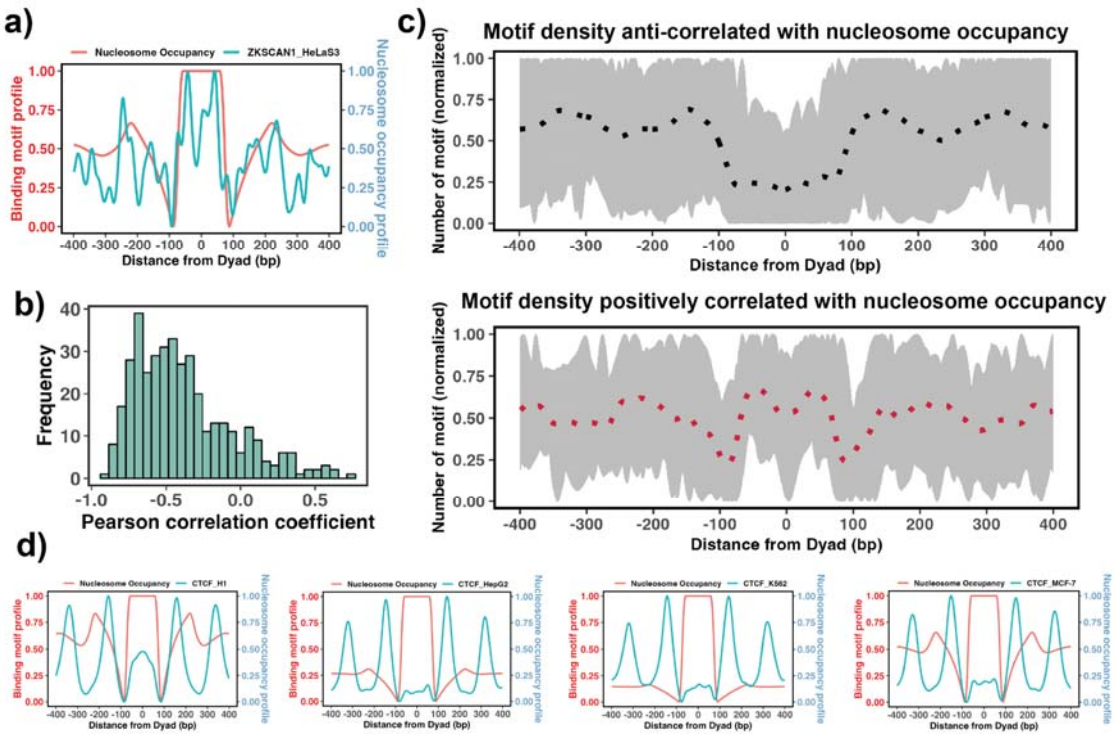

Figure 3. Association between the binding motif profiles and nucleosome occupancy. **a)** Motif profiles for ZKSCAN1 in HeLa-S3 cell line is shown as an example. **b)** Pearson correlation coefficients between motif profiles and nucleosome occupancy values for each TF (n=225) (the median value of correlation coefficient = -0.46). **c)** Binding motif profiles of TFs with positive (red, correlation coefficient >= 0.2 and p-value < 0.05) or negative correlation coefficients (black, correlation coefficient <= -0.4 and p-value < 0.05) between motif profile and nucleosome occupancy. Dashed lines correspond to the average of binding motif profiles. **d)** Comparison of binding motif profiles of CTCF between H1 embryonic stem cell line and somatic cell lines.

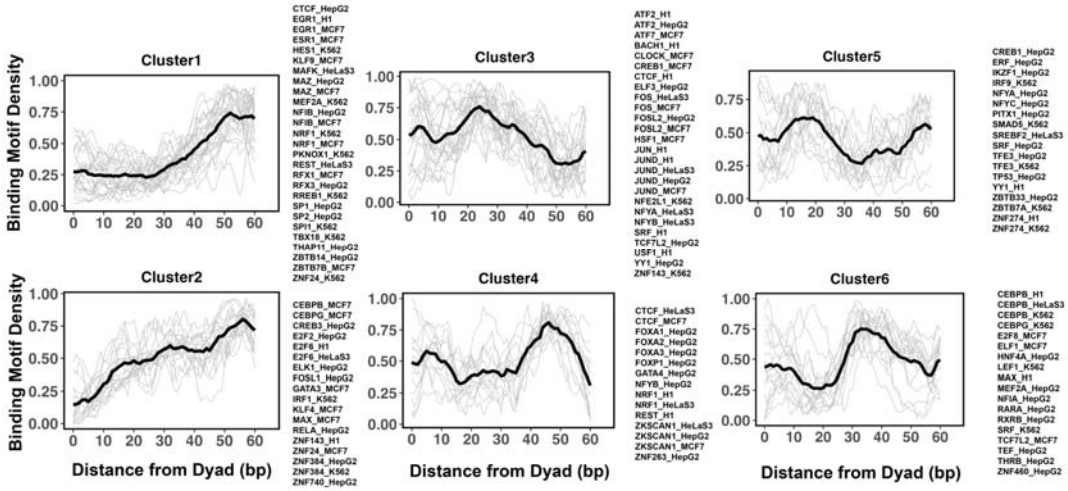

Figure 4. Clusters of TF binding motif profiles on nucleosomal DNA. Binding motif profiles centered at nucleosomal dyad locations (+/- 60 base pair from dyad) are clustered using k-medoids clustering with k=6. The entry/exit regions of nucleosomal DNA were excluded as a certain nucleotide bias exists around the ends of nucleosomal DNA reads produced by the MNase-seq experiments. Binding motif profiles between two symmetrical nucleosomal halves are combined for each TF. The black line represents the averaged profiles of all TFs in the same cluster. Cluster members with silhouette width <= 0.25 were considered as outliers and removed.

Reference

1 Isbel, L., Grand, R. S. & Schubeler, D. Generating specificity in genome regulation through transcription factor sensitivity to chromatin. *Nat Rev Genet*, doi:10.1038/s41576-022-00512-6 (2022).

2 Zaret, K. S. Pioneer Transcription Factors Initiating Gene Network Changes. *Annu Rev Genet* **54**, 367-385, doi:10.1146/annurev-genet-030220-015007 (2020).

3 Iwafuchi-Doi, M. The mechanistic basis for chromatin regulation by pioneer transcription factors. *Wiley Interdiscip Rev Syst Biol Med* **11**, e1427, doi:10.1002/wsbm.1427 (2019).

4 Kono, H. & Ishida, H. Nucleosome unwrapping and unstacking. *Curr Opin Struct Biol* **64**, 119-125, doi:10.1016/j.sbi.2020.06.020 (2020).

5 Chereji, R. V., Eriksson, P. R., Ocampo, J., Prajapati, H. K. & Clark, D. J. Accessibility of promoter DNA is not the primary determinant of chromatin-mediated gene regulation. *Genome Res* **29**, 1985-1995, doi:10.1101/gr.249326.119 (2019).

6 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260, doi:10.1038/38444 (1997).

7 Peng, Y., Li, S., Landsman, D. & Panchenko, A. R. Histone tails as signaling antennas of chromatin. *Curr Opin Struct Biol* **67**, 153-160, doi:10.1016/j.sbi.2020.10.018 (2021).

8 Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**, 207-220, doi:10.1038/s41576-018-0089-8 (2019).

9 Kornberg, R. D. & Lorch, Y. Primary Role of the Nucleosome. *Mol Cell* **79**, 371-375, doi:10.1016/j.molcel.2020.07.020 (2020).

10 Peng, Y., Li, S., Onufriev, A., Landsman, D. & Panchenko, A. R. Binding of regulatory proteins to nucleosomes is modulated by dynamic histone tails. *Nat Commun* **12**, 5280, doi:10.1038/s41467-021-25568-6 (2021).

11 Penalosa-Ruiz, G., Bright, A. R., Mulder, K. W. & Veenstra, G. J. C. The interplay of chromatin and transcription factors during cell fate transitions in development and reprogramming. *Biochim Biophys Acta Gene Regul Mech* **1862**, 194407, doi:10.1016/j.bbagrm.2019.194407 (2019).

12 Balsalobre, A. & Drouin, J. Pioneer factors as master regulators of the epigenome and cell fate. *Nat Rev Mol Cell Biol* **23**, 449-464, doi:10.1038/s41580-022-00464-z (2022).

13 Jaenisch, R. & Young, R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell* **132**, 567-582, doi:10.1016/j.cell.2008.01.015 (2008).

14 Zhu, F. *et al.* The interaction landscape between transcription factors and the nucleosome. *Nature* **562**, 76-81, doi:10.1038/s41586-018-0549-5 (2018).

15 Meers, M. P., Janssens, D. H. & Henikoff, S. Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded. *Mol Cell* **75**, 562-575 e565, doi:10.1016/j.molcel.2019.05.025 (2019).

16 Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555-568, doi:10.1016/j.cell.2015.03.017 (2015).

17 Luzete-Monteiro, E. & Zaret, K. S. Structures and consequences of pioneer factor binding to nucleosomes. *Current Opinion in Structural Biology* **75**, doi:10.1016/j.sbi.2022.102425 (2022).
18 Tan, C. & Takada, S. Nucleosome allostery in pioneer transcription factor binding. *Proc Natl Acad Sci U S A* **117**, 20586-20596, doi:10.1073/pnas.2005500117 (2020).
19 Fernandez Garcia, M. *et al.* Structural Features of Transcription Factors Associating with Nucleosome Binding. *Mol Cell* **75**, 921-932 e926, doi:10.1016/j.molcel.2019.06.009 (2019).
20 Min, J. & Liu, K. Structures of chromatin modulators in complex with nucleosome. *Curr Opin Chem Biol* **63**, 105-114, doi:10.1016/j.cbpa.2021.02.018 (2021).
21 Michael, A. K. *et al.* Mechanisms of OCT4-SOX2 motif readout on nucleosomes. *Science* **368**, 1460-1465, doi:10.1126/science.abb0074 (2020).
22 Dodonova, S. O., Zhu, F., Dienemann, C., Taipale, J. & Cramer, P. Nucleosome-bound SOX2 and SOX11 structures elucidate pioneer factor function. *Nature* **580**, 669-672, doi:10.1038/s41586-020-2195-y (2020).
23 Tanaka, H. *et al.* Interaction of the pioneer transcription factor GATA3 with nucleosomes. *Nat Commun* **11**, 4136, doi:10.1038/s41467-020-17959-y (2020).
24 Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research* **46**, D794-D801, doi:10.1093/nar/gkx1081 (2018).
25 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, doi:10.1038/nbt.2798 (2014).
26 Jankowski, A., Tiuryn, J. & Prabhakar, S. Romulus: robust multi-state identification of transcription factor binding sites from DNase-seq data. *Bioinformatics* **32**, 2419-2426, doi:10.1093/bioinformatics/btw209 (2016).
27 Zheng, A. *et al.* Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat Mach Intell* **3**, 172-180, doi:10.1038/s42256-020-00282-y (2021).
28 Avsec, Z. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* **53**, 354-366, doi:10.1038/s41588-021-00782-6 (2021).
29 Kishan, K. C., Subramanya, S. K., Li, R. & Cui, F. Machine learning predicts nucleosome binding modes of transcription factors. *BMC Bioinformatics* **22**, 166, doi:10.1186/s12859-021-04093-9 (2021).
30 Xu, Q. *et al.* ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination. *Nucleic Acids Res* **49**, 7966-7985, doi:10.1093/nar/gkab598 (2021).
31 Heuts, B. M. H. *et al.* Identification of transcription factors dictating blood cell development using a bidirectional transcription network-based computational framework. *Sci Rep* **12**, 18656, doi:10.1038/s41598-022-21148-w (2022).
32 Teif, V. B. *et al.* Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Research* **24**, 1285-1295, doi:10.1101/gr.164418.113 (2014).
33 Chen, D. *et al.* Nonlinear relationship between chromatin accessibility and estradiol-regulated gene expression. *Oncogene* **40**, 1332-1346, doi:10.1038/s41388-020-01607-2 (2021).
34 Yu, X., Singh, P. K., Tabrejee, S., Sinha, S. & Buck, M. J. DeltaNp63 is a pioneer factor that binds inaccessible chromatin and elicits chromatin remodeling. *Epigenetics Chromatin* **14**, 20, doi:10.1186/s13072-021-00394-8 (2021).

35 Gong, W. *et al.* ETV2 functions as a pioneer factor to regulate and reprogram the endothelial lineage. *Nat Cell Biol* **24**, 672-684, doi:10.1038/s41556-022-00901-3 (2022).
36 Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database, C. The sequence read archive. *Nucleic Acids Res* **39**, D19-21, doi:10.1093/nar/gkq1019 (2011).
37 Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* **38**, 276-278, doi:10.1038/s41587-020-0439-x (2020).
38 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
39 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
40 Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516-520, doi:10.1038/nature10002 (2011).
41 Gaffney, D. J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS Genet* **8**, e1003036, doi:10.1371/journal.pgen.1003036 (2012).
42 Pohl, A. & Beato, M. bwtool: a tool for bigWig files. *Bioinformatics* **30**, 1618-1619, doi:10.1093/bioinformatics/btu056 (2014).
43 Consortium, E. P. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710, doi:10.1038/s41586-020-2493-4 (2020).
44 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).
45 Piroeva, K. V. *et al.*, doi:10.1101/2022.12.20.518743 (2022).
46 Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**, D91-94, doi:10.1093/nar/gkh012 (2004).
47 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).
48 Maaten, L. v. d. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*.
49 Zaret, K. S. & Mango, S. E. Pioneer transcription factors, chromatin dynamics, and cell fate control. *Curr Opin Genet Dev* **37**, 76-81, doi:10.1016/j.gde.2015.12.003 (2016).
50 Costa, R. H., Kalinichenko, V. V., Holterman, A. X. & Wang, X. Transcription factors in liver development, differentiation, and regeneration. *Hepatology* **38**, 1331-1347, doi:10.1016/j.hep.2003.09.034 (2003).
51 Smale, S. T. Pioneer factors in embryonic stem cells and differentiation. *Curr Opin Genet Dev* **20**, 519-526, doi:10.1016/j.gde.2010.06.010 (2010).
52 Lee, C. S., Friedman, J. R., Fulmer, J. T. & Kaestner, K. H. The initiation of liver development is dependent on Foxa transcription factors. *Nature* **435**, 944-947, doi:10.1038/nature03649 (2005).
53 Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676, doi:10.1016/j.cell.2006.07.024 (2006).
54 Bungartz, G., Land, H., Scadden, D. T. & Emerson, S. G. NF-Y is necessary for hematopoietic stem cell proliferation and survival. *Blood* **119**, 1380-1389, doi:10.1182/blood-2011-06-359406 (2012).

55 Gao, H. *et al.* Esrrb plays important roles in maintaining self-renewal of trophoblast stem cells (TSCs) and reprogramming somatic cells to induced TSCs. *J Mol Cell Biol* **11**, 463-473, doi:10.1093/jmcb/mjy054 (2019).
56 Iyer, H., Collins, J. J., 3rd & Newmark, P. A. NF-YB Regulates Spermatogonial Stem Cell Self-Renewal and Proliferation in the Planarian Schmidtea mediterranea. *PLoS Genet* **12**, e1006109, doi:10.1371/journal.pgen.1006109 (2016).
57 Pop, R. T. *et al.* Identification of mammalian transcription factors that bind to inaccessible chromatin. *Nucleic Acids Res*, doi:10.1093/nar/gkad614 (2023).
58 Qu, M., Qu, H., Jia, Z. & Kay, S. A. HNF4A defines tissue-specific circadian rhythms by beaconing BMAL1::CLOCK chromatin binding and shaping the rhythmic chromatin landscape. *Nat Commun* **12**, 6350, doi:10.1038/s41467-021-26567-3 (2021).
59 Steger, D. J. & Workman, J. L. Stable co-occupancy of transcription factors and histones at the HIV-1 enhancer. *EMBO J* **16**, 2463-2472, doi:10.1093/emboj/16.9.2463 (1997).
60 Last, T. J., van Wijnen, A. J., de Ridder, M. C., Stein, G. S. & Stein, J. L. The homeodomain transcription factor CDP/cut interacts with the cell cycle regulatory element of histone H4 genes packaged into nucleosomes. *Mol Biol Rep* **26**, 185-194, doi:10.1023/a:1007058123699 (1999).
61 Teif, V. B. *et al.* Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol* **19**, 1185-1192, doi:10.1038/nsmb.2419 (2012).
62 Teif, V. B. *et al.* Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Res* **24**, 1285-1295, doi:10.1101/gr.164418.113 (2014).
63 Voong, L. N. *et al.* Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. *Cell* **167**, 1555-1570 e1515, doi:10.1016/j.cell.2016.10.049 (2016).
64 Kundaje, A. *et al.* Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res* **22**, 1735-1747, doi:10.1101/gr.136366.111 (2012).
65 Iwafuchi-Doi, M. *et al.* The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Mol Cell* **62**, 79-91, doi:10.1016/j.molcel.2016.03.001 (2016).
66 Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665, doi:10.1016/j.cell.2018.01.029 (2018).
67 MacCarthy, C. M. *et al.* OCT4 interprets and enhances nucleosome flexibility. *Nucleic Acids Res* **50**, 10311-10327, doi:10.1093/nar/gkac755 (2022).
68 Zaret, K. S., Lerner, J. & Iwafuchi-Doi, M. Chromatin Scanning by Dynamic Binding of Pioneer Factors. *Mol Cell* **62**, 665-667, doi:10.1016/j.molcel.2016.05.024 (2016).
69 Nagaich, A. K. *et al.* p53-induced DNA bending and twisting: p53 tetramer binds on the outer side of a DNA loop and increases DNA twisting. *Proc Natl Acad Sci U S A* **96**, 1875-1880, doi:10.1073/pnas.96.5.1875 (1999).
70 Poirier, M. G., Oh, E., Tims, H. S. & Widom, J. Dynamics and function of compact nucleosome arrays. *Nat Struct Mol Biol* **16**, 938-944, doi:10.1038/nsmb.1650 (2009).
71 Li, S., Peng, Y., Landsman, D. & Panchenko, A. R. DNA methylation cues in nucleosome geometry, stability and unwrapping. *Nucleic Acids Res* **50**, 1864-1874, doi:10.1093/nar/gkac097 (2022).
72 Brandani, G. B., Niina, T., Tan, C. & Takada, S. DNA sliding in nucleosomes via twist defect propagation revealed by molecular simulations. *Nucleic Acids Res* **46**, 2788-2801, doi:10.1093/nar/gky158 (2018).

73 Donovan, B. T., Chen, H., Jipa, C., Bai, L. & Poirier, M. G. Dissociation rate compensation mechanism for budding yeast pioneer transcription factors. *Elife* **8**, doi:10.7554/eLife.43008 (2019).
74 Li, S., Wei, T. & Panchenko, A. R. Histone variant H2A.Z modulates nucleosome dynamics to promote DNA accessibility. *Nat Commun* **14**, 769, doi:10.1038/s41467-023-36465-5 (2023).
75 Li, Z. *et al.* Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell* **151**, 1608-1616, doi:10.1016/j.cell.2012.11.018 (2012).