The importance of DNA sequence for nucleosome positioning in the process of transcriptional regulation

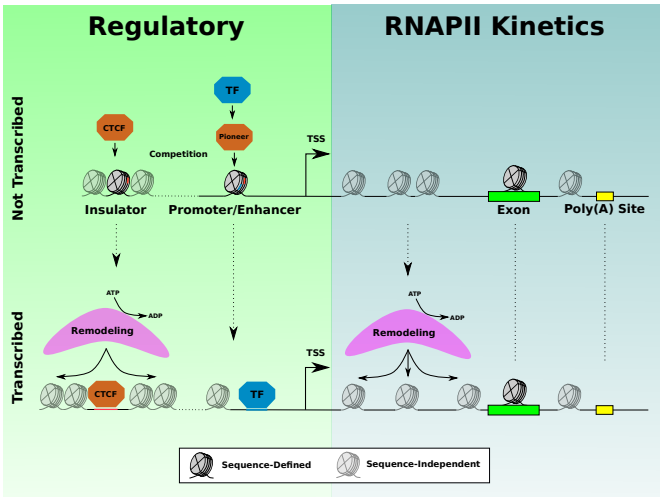
Malte Sahrhage [1*], Niels Benjamin Paul [1,2], Tim Beißbarth [1], Martin Haubrock [1*]

[1]Department of Medical Bioinformatics, University Medical Center, 37077 Göttingen, Germany and [2]Department of Cardiology and Pneumology, University Medical Center, 37075 Göttingen, Germany

ABSTRACT

Nucleosome positioning is a key factor in the regulation of transcription, as nucleosomes regulate the dynamic accessibility of chromatin for transcription factor binding and interact with the transcription machinery at every stage. Influences to steer nucleosome positioning are diverse, and the importance of the DNA sequence for this process in contrast to active chromatin remodeling has been the subject of a long discussion. In this study, we evaluate the functional role of DNA sequence for all major elements along the process of transcription. For this purpose, we developed a random forest classifier based on local DNA structure that assesses the sequence-intrinsic support for nucleosome positioning. On this basis, we created an easy-to-use data resource that can be applied in a genome-wide manner to the human genome. In our comprehensive analysis, we found a special role of DNA in mediating the competition of nucleosomes with cis-regulatory elements, in enabling steady transcription, for positioning of stable nucleosomes in exons and for nucleosome depletion during transcription termination. In contrast, we relate these findings to concurrent processes that generate strongly positioned nucleosomes in vivo that are not mediated by sequence, such as energy-dependent remodeling of chromatin.

GRAPHICAL ABSTRACT



INTRODUCTION

The nucleosome is a foundational protein complex in eukaryotic organisms comprised of four equal pairs of subunits (H2A, H2B, H3, H4) which wraps 147 bp of DNA (1, 2). The positioning of nucleosomes is very important for the regulation of transcription and the relevance of nucleosome positioning for the regulation of genes can be viewed from many perspectives. The broadest unit to span interaction potential are topologically associated domains (TADs). These are characterized by a loop compartment that brings together potentially interacting proteins at enhancers and promoters of genes close to each other (3). The boundaries of these logical units are set up at so called insulator regions which are defined by an interplay of CTCF binding (4) and a consequential strict positioning of a well-positioned nucleosome array (5). Whether the genes inside such units are transcribed is an intricately regulated process involving nucleosome positioning at all levels. The binding of transcription factors (TFs) to promoter and enhancer regions is crucial for the initiation of transcription. As a means of controlling access to transcription factor binding sites (TFBSs), nucleosomes occupy DNA and need active displacement through pioneer TFs or energy dependent remodeling to avoid random triggering of gene expression (6, 7). To achieve an orderly rate of transcription, the RNA polymerase II (RNAPII) is highly regulated by nucleosomes. In particular, the +1 nucleosome directly downstream of the transcription start site (TSS) and a sequential array of well-positioned nucleosomes is needed to reliably initiate transcription (8, 9). Along its way, RNAPII encounters more fuzzily positioned nucleosomes in intronic regions and well-positioned nucleosomes in exons. Since exonic DNA is the template for the subsequent translation of proteins, the positioning of nucleosomes supports the faultless transcription of these regions by serving as roadblocks to decrease the speed of RNAPII elongation, by marking alternative splice sites (10) and by triggering feedback signals for other concurrent

*To whom correspondence should be addressed. Email: malte.sahrhage@bioinf.med.uni-goettingen.de
Correspondence may also be addressed to Martin Haubrock Tel: +49-551-3961784; Email: martin.haubrock@bioinf.med.uni-goettingen.de

transcription initiation events on the same gene (11). The transcription ends at the transcription termination site (TTS), which sets up the signal where polyadenylation will be applied on the resulting pre-mRNA. These polyadenylation sites (Poly(A) sites) show an increased upstream nucleosome positioning, opposed to a mainly nucleosome depleted area directly at the DNA encoded prospective polyadenylation site (12). An extra layer of nucleosome functionality in all of these stages is added through the epigenetic modifications of histone proteins. Depending on the specific biochemical modulations of the histone sub-units, nucleosomes can carry marks for the active recruitment of further proteins and show a different flexibility in terms of the energy needed for displacement (13, 14). Since the differential increase or decrease of chromatin accessibility at all of these specific functional sites regulates transcription and thus cell identity, it is crucial to understand the principles that determine nucleosome positioning.

The influences on the positioning of nucleosomes have been debated for many years and can be grouped roughly into DNA sequence, trans-acting factors and active chromatin remodeling (15, 16, 17). The role of the DNA sequence is particularly controversial. It was postulated that there is a periodic repeat of A/T di-nucleotides that support the binding of nucleosomes (18, 19, 20). This pattern has been extended to be represented by local DNA structure (21) with the help of DNAshape (22). The benefits of the supposed patterns for nucleosome support are linked to the flexibility of the DNA strand to facilitate an easier attraction of nucleosomes (23, 24), which can be seen particularly well in extremely stiff poly(dA:dT) tracts that counteract nucleosome binding (25).

Some of these findings have been discovered with the help of a variety of computational tools to predict nucleosome positions, each with a different focus of training data, prediction target and methodology (26, 27, 28, 29). Among these, there are tools that employ HMMs (30, 31), SVMs (32) and recently more and more CNN architectures (33, 34, 35, 36). The difference in meaning between sequence-intrinsic signals and the actual *in vivo* positioning of nucleosomes is fundamental to understand. The sequence can only explain the potential of any given position to attract nucleosomes. Whether it is actually occupied by nucleosomes or opened up by regulatory mechanisms cannot be determined without the help of experimental chromatin accessibility data for the particular cell type. In other words, while experimental data can provide insights into the actual nucleosome positioning within cells, it does not directly elucidate the underlying reasons or mechanisms behind such positioning. On the other hand, machine learning on DNA sequence can give an intrinsic estimate of nucleosome attraction potential but cannot answer the question where nucleosomes are located *in vivo*. Therefore, we need to integrate both information to estimate the overlap of sequence-derived nucleosome support and experimental chromatin accessibility evaluation (37). For that purpose we use nucleosome occupancy data (MNase-seq) (38), chromatin accessibility data (DNase-seq) (39) and ChIP-seq data for both histone modifications and TFs, mainly in the two cell lines GM12878 and K562 available from ENCODE (40).

In this study, we used a sequence-based nucleosome support score to evaluate all elements involved in transcription for the human genome. For that purpose, we built a classification algorithm to apply in a genome-wide context and to provide an easy-to-use data resource which can be used without further computation for similar analyses in the future. Due to its relative robustness against overfitting we created a random forest (41) classifier based on the Guo2014 dataset (32) of nucleosomal sequences which is derived from MNase-seq in human $CD4^+$ T-cells (38). Our classifier uses a transformation from raw DNA sequences to a frequency spectrum of local DNAshape (Propeller Twist, Helix Twist, Minor Groove Width, Electrostatic Potential, obtained with the DNAshapeR package (42)). This binary classifier was applied in a sliding-window manner to the complete human genome in two different resolutions, yielding a genome-wide mapping of sequence-defined nucleosome support.

We aspire to give a comprehensive overview over the crucial parts of the transcriptional machinery and provide our genome-wide measure for sequence-intrinsic nucleosome support as a simple data resource. Our findings are embedded into a framework of previous studies that showed the relevance of DNA for nucleosome positioning for some specific elements and we add some new insights about the sequence-definition of nucleosome positioning in transcription. We demonstrate that there is a clear influence of DNA sequence for supporting nucleosome binding at positions of high competition with other DNA-binding proteins and where it is beneficial to accumulate nucleosomes by default, such as exons, or repel them in transcription termination sites. In contrast, we give examples of positioned, but sequence-independent nucleosome arrays in proximity to the TSS or around insulator sites and put our findings into the context of more unspecific, fuzzy nucleosomes.

METHODS

Datasets

The experimental chromatin accessibility datasets were obtained from ENCODE for the human genome in hg19 (40). They comprise of raw signal MNase-seq data (K562: ENCFF000VNN, GM12878: ENCFF000VME), DNase-seq data (K562: ENCFF352SET, GM12878: ENCFF901GZH), peak-called histone modification ChIP-seq data for K562 (H3K4me3: ENCFF779FUO, K3K27me3: ENCFF908KJV, H3K27ac: ENCFF252DWA, H3k9me3: ENCFF323KXF) and GM12878 (H3K4me3: ENCFF354MGT, K3K27me3: ENCFF734CUK, H3K27ac: ENCFF660QDF, H3k9me3: ENCFF713DDN) and raw signal for H3K4me3 (K562: ENCFF285EUA, GM12878: ENCFF154XCY). Besides, we used CTCF ChIP-seq peaks from ENCODE for our insulator analysis (K562: ENCFF002DDJ , GM12878: ENCFF710VEH)

Additionally, we used nucleosome occupancy data to evaluate the influence of the GATA3 TF in MDA-MB-23 cells with and without GATA3 expression. The according MNase-seq data was taken from (43). We calculated the mean of all three replicates for the condition of GATA3 expression and for the control cells.

Furthermore, we evaluated the evolutionary conservation at base pair-resolution with the PhyloP score (44), which was obtained using the UCSC table browser. (45)

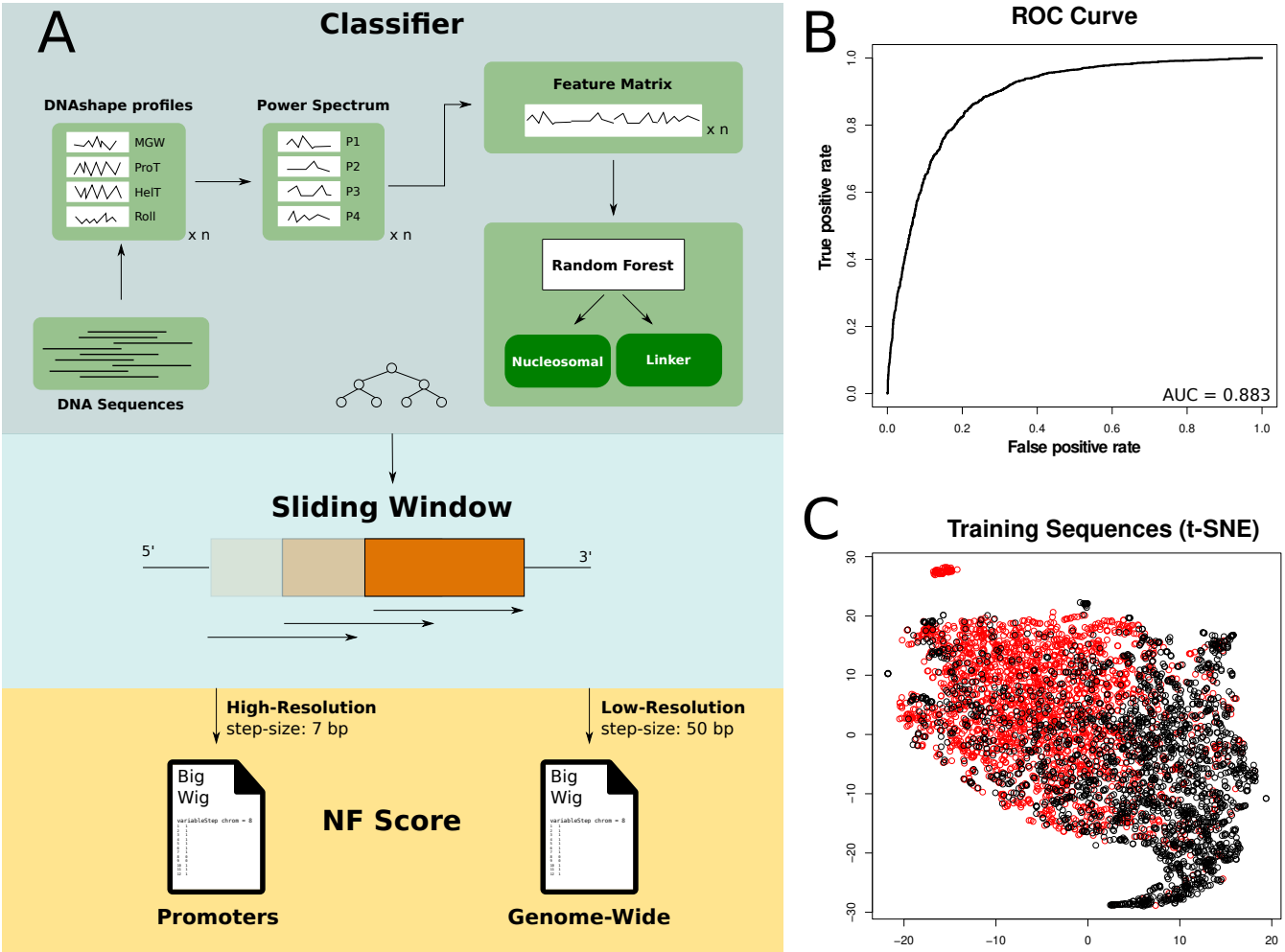


Figure 1. Processing and Classification of DNA Sequences. **(A)** Deriving the NF score. A random forest model is used to annotate DNA sequences as nucleosomal or linker. To build the feature matrix, each DNA fragment of length 147 bp is transformed into a frequency spectrum of local DNAshape. In order to create a pre-calculated mapping for the human genome, we applied the classifier genome-wide in a sliding-window approach with a step size of 50 bp and a detailed version for promoters with a step size of 7 bp. Both resources are provided as a downloadable BIGWIG file (see data availability section). **(B)** Receiver Operating Characteristic (ROC) curve and area under ROC curve (AUC). **(C)** t-SNE representation of training data based on the local DNAshape frequency spectrum to show how well the underlying data is separable in principle.

We compared all these data along a variety of different genome locations. The broad overview over different biological contexts is defined by chromHMM for K562 and GM12878 and defines 15 meaningful states originally derived from histone modification ChIP-seq data (46).

The gene mapping we use is based on RefSeq in hg19 (47) and was also obtained with the UCSC table browser. We exclude genes from the chromosome Y, since not all experimental chromatin accessibility data is available for it. In particular, we use the remaining 30142 unique TSS locations to define genes and promoters. Our promoter definition is always 1 kb upstream of any TSS. In case of multiple alternative TSS locations, all of these were considered as individual gene starts. Furthermore, we extracted all exons from the same mapping.

Moreover, we analyzed the elements at TTSs in the form of Poly(A) sites. The poly(A) site mapping with a separation into clusters of biological origin is taken from PolyASite 2.0 (48). To obtain single locations for the center of each site, we used the representative cluster location from the unique cluster ID column. The original assembly of the data is hg38. We used liftOver to convert the coordinates to hg19 via the rtracklayer R package (49).

Software

Apart form the individual software that is referenced in its due place, the analyses were performed using R (v. 4.3) (50) and we used the packages rtracklayer (v. 1.60.0) (49), ggplot2 (v. 3.4.2) (51), GenomicRanges (v. 1.52.0) (52), ROCR (v.1.0.11) (53) and randomForest (v. 4.7.1.1) (54).

Classifier

The classifier is based on the human Guo et al. 2014 dataset (32) which is in turn derived from an *in vivo* mapping of nucleosome positions in $CD4^+$ T-cells (38). It contains 4573 non-redundant FASTA sequences which are categorized as either nucleosomal or linker sequences. The FASTA files are transformed with the DNAshapeR tool (v. 1.28.0) (42),

resulting in a set of numerical vectors for Roll, HelT, ProT and MGW. Additionally, a power spectrum (Power Spectral Density, PSD) is applied to the DNAshape values in order to strengthen positional independence of the periodic nucleotide patterns, emphasizing the assumed underlying periodic sequence pattern and making it suitable for the use in a random forest classifier. In a next step, the spectral densities are combined to a feature matrix which is the input for a random forest classifier. The result of the classification is the categorization of a sequence being either *nucleosomal* or *linker* and serves as the basis for the two types of NF scores that will be described further in the next section.

The performance of the classifier was evaluated with an out-of-bag-error of 0.1865 and an AUC of 0.883. The ROC curve for the classifier can be seen in Figure 1 B. For a better understanding about how clearly the data can be expected to be separated, Figure 1 C shows a t-SNE representation of the data used to train the classifier. It can be seen that the data is not clearly separable, so it is not to be expected to get an almost perfect classifier.

Scores

NF score The classifier described above gives a binary prediction for a DNA segment of 147 bp in length. To allow for an easier handling of the estimates of nucleosome support based on nucleotide sequence, we pre-calculated two sets of scores for the human genome in hg19 assembly with different local resolutions. (1) The low-resolution score is a simple application of the classifier in a sliding window to all regions of the human genome with a fragment input length of 147 bp and a step size of 50 bp. The resulting NF score is an estimate of either 1 (nucleosomal) or 0 (linker) for the central 50 bp in any given 147 bp bin. (2) The high-resolution NF score provides a continuous score between 0 and 1 and is based on a much more local sliding window with a segment size 147 bp and a step-size of 7 bp. All base-pairs are thus part of multiple (147/7=21) predictions and the score represents the relative frequency of nucleosomal prediction results, this particular bp has been part of. To give an example, if an individual bp has been part of 21 sliding windows, of which 14 were classified as nucleosomal and 7 as linker, then the score at that base-pair is 14/21= 0.67.

Due to resource efficiency the high-resolution score was only pre-calculated for all human promoter regions with a radius of 3 kb around the TSS, since we want to give a good local estimate of the importance of the sequence for nucleosome positioning in specific individual promoter regions. If not specified otherwise, we always use the low-resolution score, since we normally show the mean of nucleosome support profiles over a large number of fragments and there is virtually no difference between averaging the low-resolution 0/1 predictions or doing the same with the 0-1 high-resolution score. Both resources are easily accessible as BIGWIG files and listed in the data availability section. Additionally, the code to classify nucleosomal vs. linker DNA can be found there.

DHS score In this study we analyze the influence of sequence on nucleosome positioning, However, the resulting accessibility in particular cell-lines can be very cell-type specific. Therefore, we created a score to generalize the accessibility over multiple cell-lines. The DHS (DNase Hypersensitive Sites) score is a measure to estimate the generalized accessibility of any given genome location. For that purpose we obtained peak-called DNase-seq data for all 403 primary cell lines from ENCODE in human. These we integrated as a simple count of how many primary cell lines show a DNase-seq peak at any given base-pair.

RESULTS

Sequence-defined nucleosome support characterizes functional genomic regions

We used a random forest classifier to derive a nucleosome formation (NF) score to estimate the nucleosome support based on the nucleotide sequence and applied it to a range of different genome locations. A broad range of entities can be found in the histone ChIP-seq derived mappings of chromHMM for the cell lines K562 and GM12878 (46). Figure 2 A shows the mean sequence-dependent nucleosome support (NF), the nucleosome occupancy (MNase-seq) and the chromatin accessibility (DNase-seq) for the 147 bp in the center of any given chromHMM fragment. In total there are 622257 fragments in K562 and 571339 in GM12878. A list of the individual fragment numbers for each subgroup can be found in Supplementary Table 1. We grouped all chromHMM states into the three categories: ”Regulatory” (Promoters/ Enhancers), ”Heterochromatin/Repetitive” and ”Transcription” (Transcriptional Transition/Elongation, etc. and Insulators).

The most intriguing feature is the relation between sequence-dependent nucleosome support and actual nucleosome occupancy. Active promoters show the highest DNA-intrinsic nucleosome support (NF: approx. 0.8). However, being crucial places for transcriptional regulation, they are also the most accessible regions (low MNase-seq of 0.4-0.55 and high DNase-seq of 0.65-1). The same nucleosome support can be observed in poised promoters, but with closed chromatin (low DNase-seq, high MNase-seq). This observation can be put in contrast to the highlighted group of heterochromatin fragments. Here, we observe rather low nucleosome support in both cell lines of roughly 0.4. At the same time the locations are very inaccessible with a nucleosome occupancy of 1 and very low DNase-seq signal of 0.03.

The last highlighted group in Figure 2 are insulators. In this group we can observe a relatively high nucleosome support of 0.6-0.66 and a medium accessibility with relatively high values for both MNase-seq of 1 and DNase-seq signal of 0.2-0.3.

In summary, these observations suggest that regions involved in active regulation of transcription have significantly higher levels of nucleosome support with well-defined sequence context. Whether these important locations are actually occupied by nucleosomes is dependent on whether they are actually used *in vivo* (Active vs. Poised Promoters). In comparison, constitutive heterochromatin has a much lower sequence-dependent nucleosome support while being mainly inaccessible.

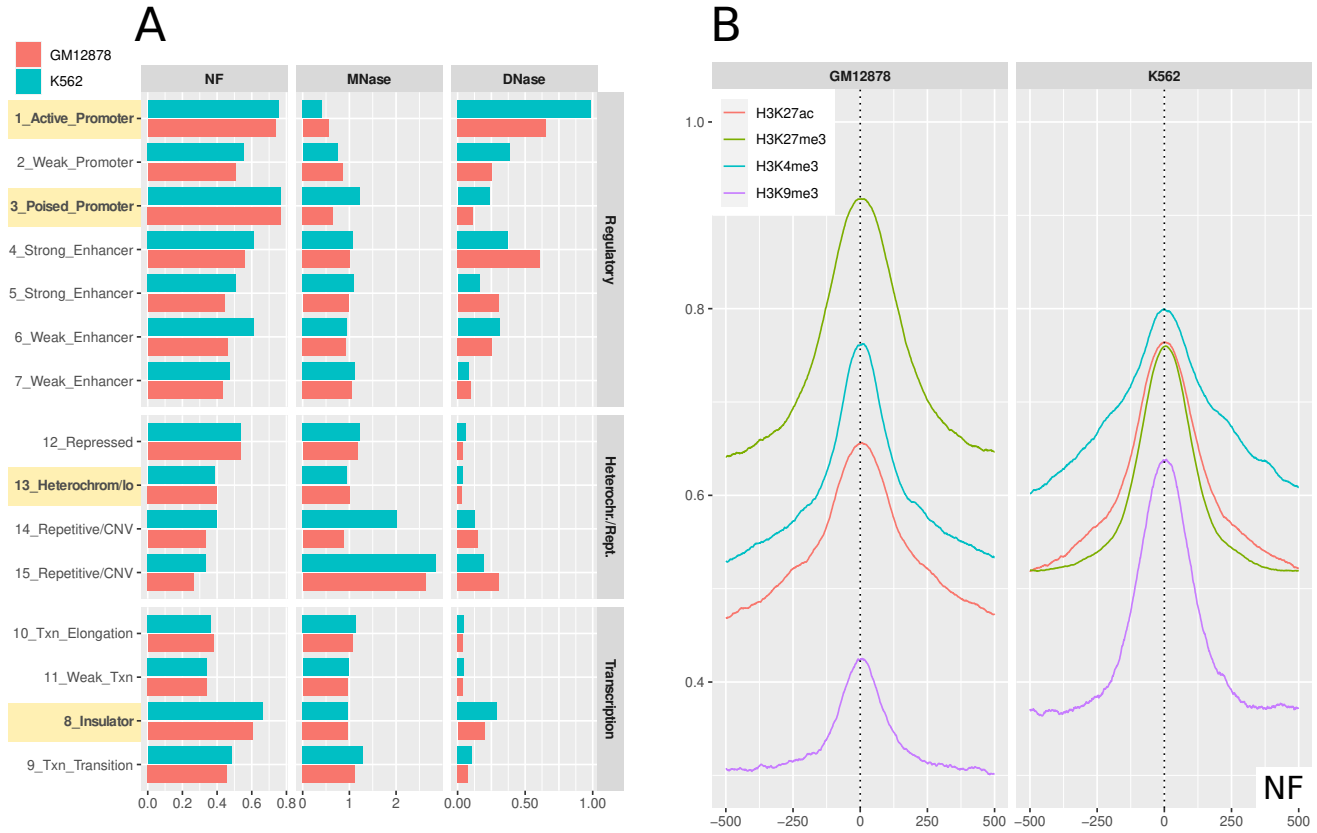


Figure 2. Nucleosome support in different regions of the genome and around histone modification ChIP-seq peaks. **(A)** chromHMM. Nucleosome Support (NF score), nucleosome occupancy (MNase-seq) and chromatin accessibility (DNase-seq) in all states of the chromHMM mappings. The highlighted regions exhibit a high NF score in active regulatory regions (active and poised promoters) as well as in insulators, opposed to low nucleosome support regions. Active promoters are rather accessible and poised promoters are closed off mostly. Heterochromatin exhibits pronounced inaccessibility, whereas insulators demonstrate a hybrid accessibility characterized by relatively elevated levels of both DNase-seq and MNase-seq signals. **(B)** Nucleosome support of histone modifications. A distinct peak is observed in proximity to each histone modification ChIP-seq peak, providing evidence of the influence of DNA sequence on the positioning of individual nucleosomes. This influence is higher in marks indicating a regulatory function like active promoters (H3K4me3), poised promoters (H3K27me3) and more generally accessible chromatin (H3K27ac) as compared to repressing histone marks like H3K9me3.

Histones can carry different kinds of modifications that determine their function and the flexibility of the surrounding chromatin accessibility (13, 14). Since chromHMM mapping is derived from ChIP-seq data of histone modifications, we gain more detailed insights into the local nucleosome support relative to the actual position of the histone modification. Therefore, we analyzed 518609 histone ChIP-seq peaks from ENCODE for the GM12878 and K562 cell lines. We chose the histone marks according to their role in relation to the derived chromHMM states. The H3K4me3 peaks are representative for active promoter histone marks (GM12878: 40325, K562: 32868), H3K27ac for enhancers (GM12878: 61462, K562: 52000), H3K27me3 represents poised promoters (GM12878: 40238, K562: 251901) and H3K9me3 is repressive chromatin as a general indication for constitutive heterochromatin (GM12878: 28968, K562: 10847).

Figure 2 B shows the average NF score around the set of different histone modification ChIP-seq peaks. It can be seen that in the specific experimental configurations, the previous findings of a higher NF score in regulatory important regions can be confirmed. For both cell lines it can be observed that the three activating marks are higher than the heterochromatin mark. Additionally, all histone modifications show a clear peak of nucleosome support around the central ChIP-seq peak. The active promoter mark H3K4me3 shows a peak at an NF score of 0.76-0.80. The poised promoter mark H3K27me3 is more variable between the two cell lines and can be found at levels of 0.76 and 0.92, respectively. The general accessibility mark in promoters and enhancers H3K27ac is at 0.66-0.76 and the peak of the heterochromatin mark H3K9me3 is located lowest at a level of 0.42-0.63. These results indicate a local sequence preference for nucleosome binding around the center of all histone modifications with a distinctly higher nucleosome support for regulatory elements.

Regulatory Nucleosomes: Competition between sequence-defined nucleosomes and transcription factors

Promoter regions The previous analyses suggest that DNA structure plays different roles in nucleosome binding in different biological environments. To get a better understanding of these differences, we zoom in on regulatory locations, in particular all human promoters, so that we obtain the chance to link the static information of sequence-defined regulatory potential and its effect on gene regulation. The promoters are defined as 1 kbp upstream of all 30142 RefSeq defined human genes (see methods). Alternative TSS locations

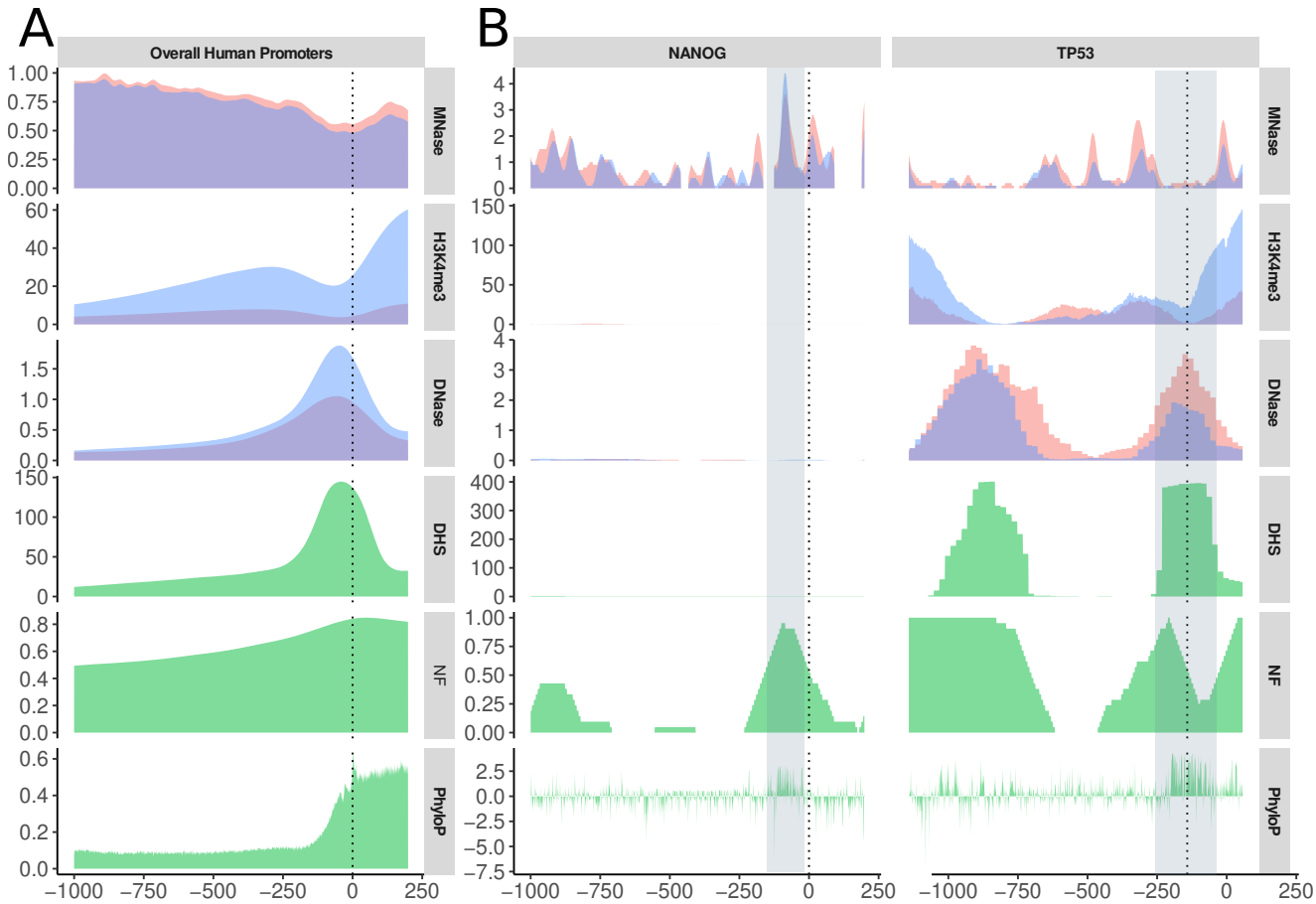


Figure 3. Profiles for experimental chromatin accessibility (MNase-seq, DNase-seq, H3K4me3 ChIP-seq) and static data (green: DHS score, NF score, PhyloP) as (A) the mean of all human RefSeq promoters and (B) the promoters of the individual genes NANOG and TP53 in K562 (blue) and GM12878 (red). The dotted line at position 0 marks the TSS. **(A)** Overall Promoter Regions. The chromatin accessibility increases, while the general nucleosome occupancy decreases. The sequence-dependent nucleosome support (NF score) increases towards the TSS along the increasing chromatin accessibility. Furthermore, the H3K4me3 ChIP-seq and MNase-seq signals exhibit an NDR upstream of the TSS and the evolutionary sequence conservation represented as PhyloP score, is generally decreased in the promoter region but increases steeply towards the TSS. **(B)** *NANOG Profiles.* In the context of the two cell lines, NANOG, due to its role as a developmental pioneer factor, does not exert any influence in K562 and GM12878. The marked area denotes an overlap of high sequence-defined nucleosome support as well as a high MNase-seq peak. That means the main site of nucleosome attraction potential is occupied by well-positioned nucleosomes. *TP53 Profiles.* TP53 is an important tumor suppressor in both cell lines. The NF score follows the peaks of high accessibility in DHS score and DNase-seq in location. In the annotated instance, the region characterized by an elevated nucleosome support exhibits a nucleosome-free state in vivo, indicating a displaced nucleosome position. Both gene promoters show an increased PhyloP score at the high NF score sites, indicating an evolutionary conservation of the nucleosome supporting sequence.

are counted as individual promoters and strand direction is aligned to 5'-3' direction.

We applied the high-resolution NF score (see methods) to all human promoters and compare this cell-type unspecific signal with the actual experimental evidence of chromatin accessibility in K562 and GM12878, measured by ENCODE MNase-seq, DNase-seq and H3K4me3 ChIP-seq data as well as a derived DHS score which integrates DNase-seq peaks over 403 primary cell lines (see methods). Additionally, we include PhyloP, a bp resolution evolutionary conservation score.

Figure 3 A shows the position-specific mean distribution of all available data in relation to the TSS. An extended version of this figure showing 1 kb up- and downstream of the TSS can be found in Supplementary Figure S1. The overall trend of the different data sources is very similar between both cell lines and confirms the general expectations of chromatin accessibility. While the general occupation with nucleosomes declines towards the TSS (MNase-seq from 0.9 to 0.5), the amount of promoter histone modification H3K4me3 increases towards the TSS (from 10 to 25 in K562, from 4 to 7 in GM12878) with a distinct valley about 100 bp from the TSS (DNase-seq and DHS score). This decline can be explained simultaneously with the gain of chromatin accessibility by putative proteins such as TFs, which is described by a peaked rise of DNase-seq (from 0.25/0.3 to 1.0/1.6) and DHS score (from 10 to 145) in the same area. The NF score increases towards the TSS with a more steady, less peaked slope from approx. 0.5 at -1 kb to 0.85 at the TSS. The evolutionary conservation (PhyloP) stays relatively stable at a level of 0.09 until it increases steeply to 0.45, starting at -200 bp and continuing high into the gene.

The average profiles of each signal in Figure 3 A are overall logical but mainly they show the clustering of DNA

accessibility near the TSS at the nucleosome depleted region (NDR). However, to understand the local interaction of nucleosomes and TFs, it is crucial to examine the specific positions within the entire promoter region. Figure 3 B shows examples of individual promoters and the relations between nucleosome binding, chromatin accessibility and the potential nucleosome support described by the NF score.

The transcription factor NANOG is a general developmental factor involved in the proliferation and renewal of embryonic stem cells (55) and thus it does not play a role in the differentiated cell lines. Figure 3 B (left) shows the integrated profiles for the gene of NANOG which should not be active. Therefore, the first notable thing is that there is neither a clear signal for DNase-seq, nor for H3K4me3 ChIP-seq. Additionally, the general accessibility in primary cell lines, described by the DHS score, is mainly zero overall, too. Albeit, the NF score is high near the TSS (highlighted region) with an NF score of 0.9-0.95. At the same location a maximum of nucleosome occupancy (MNase-seq of 4.4 in K562 and 3.6 in GM12878) can be observed. The evolutionary conservation at this point is much higher with up to 2.9 compared to the near-zero level in the rest of the promoter. This indicates a regulatory nucleosome position that is currently blocked by a well-positioned nucleosome, so that the activation of the NANOG gene requires an active displacement of this nucleosome.

In contrast, the transcription factor TP53 carries substantial importance in both the K562 and GM12878 cell line, as it functions as a critical tumor suppressor (56). Thus, its gene serves as an example of a promoter that is positively regulated. The profiles for this TF can be found in Figure 3 B (right). The accessible regions are separated into two distant peaks on the level of DNase-seq and DHS score, one of which is located around the TSS and the other one at 800 bp upstream. Both accessible regions seem to be important in almost all primary cell lines (DHS score of 401 at the upstream peak and 398 downstream). Both of these peaks furthermore exhibit an absence of nucleosomes in these areas, evident in an MNase-seq signal of 0-1 at the upstream peak and of 0-0.3 in the downstream peak. Interestingly, the downstream peak is flanked by a high nucleosome occupancy of up to 1.5-2.1 to either side, potentially marking shifted nucleosomes and/or the well-described +1 nucleosome downstream of the TSS. The sequence support for nucleosome binding in the form of NF score shows a broader plateau of 1 at the upstream portion of the accessible promoter and another maximum directly upstream of the TSS in the NDR with a sharper profile and a short decline before it peaks again closely downstream of the TSS.

Transcription Factors The pioneer transcription factor GATA3 has been described to bind well positioned nucleosomes (57). To check whether this positioning can be seen on DNA level as well, we analyzed 43504 GATA3 binding sites with our global NF score and compared them to published nucleosome occupancy (MNase-seq) in MDA-MB-23 cells (43). Figure 4 A shows the profiles of +-500 bp around GATA3 binding sites. It is evident that the area is supportive for nucleosome binding based on sequence as shown by an NF score that shows an increased peak of approx. 0.15 around the binding sites from 0.35 to 0.5. In the case of cells without GATA3 expression (control), there is a shift of increased nucleosome occupancy (0.63-0.7) around the sites, showing increased nucleosome occupancy on top of the closed binding sites. However, upon GATA3 expression, the peak observed in the data undergoes a split into a double peak at 0.675, corresponding to both sites, while exhibiting a distinct valley at 0.62 directly at the transcription factor binding sites (TFBSs). This behavior suggests the presence of a robust nucleosome that was initially positioned based on the DNA sequence but subsequently displaced due to the influence of pioneer transcription factors.

Insulators One of the regions with a rather high mean NF score of 0.66 in Figure 2 A were insulator regions. Insulators are DNA regulatory elements that play a crucial role in organizing and maintaining the spatial organization of the genome. The protein CTCF is a key factor in mediating insulator function by facilitating chromatin looping and regulating gene expression (58). The nucleosome occupancy and sequence-defined nucleosome support around 43247 ChIP-seq peaks for CTCF in K562 and 43631 peaks in GM12878 are compared in detail in Figure 4 B. The peaks are agnostic of strand orientation, since they are not mapped to a particular gene.

The nucleosome occupancy shows a very stable, phased array of nucleosomes around the CTCF peaks with an approx. 200 bp periodicity and an ongoing MNase-seq signal decrease per peak from 1.8 to 1.2 with increasing distance to the TSS. While the nucleosome occupancy is lowest directly at the CTCF peaks, the sequence-dependent nucleosome support is rather high specifically in these regions with an NF score of about 0.77 at the peak. The observed periodicity in nucleosome occupancy is completely absent on the sequence level, indicating a strong influence on nucleosome positioning at CTCF sites but active energy-driven chromatin remodeling once the site was cleared and the factor is bound to the DNA.

Sequence-defined nucleosome support in transcriptional regulation

In the previous part we focused mainly on the role of DNA for nucleosome positioning in promoters, in proximity to the TSS and in competition to TFs. The following sections will explore other functional elements of the genome in a similar way.

Active chromatin remodeling on transcription initiation It is known that there is a multitude of influences on the dynamic positioning of nucleosomes. One such phenomenon is the phasing of ordered nucleosome arrays downstream of the TSS upon transcription initiation (59). Figure 5 A shows the comparison of nucleosome occupancy (MNase-seq) and the DNA-intrinsic nucleosome support (NF) around the TSS of all human RefSeq genes. The proposed periodic downstream array of nucleosome positions is apparent, marked by arrows, stretching with a periodicity of 200 bp at MNase-seq levels between 0.75 and 0.9 (in GM12878) or 0.65 and 0.85 (in K562), respectively. At the same time, although the NF score is generally still relatively high close to the TSS and the +1 nucleosome with a value of 0.8-0.85, the periodicity is not mirrored on the sequence level and the general sequence-dependent nucleosome support decreases further

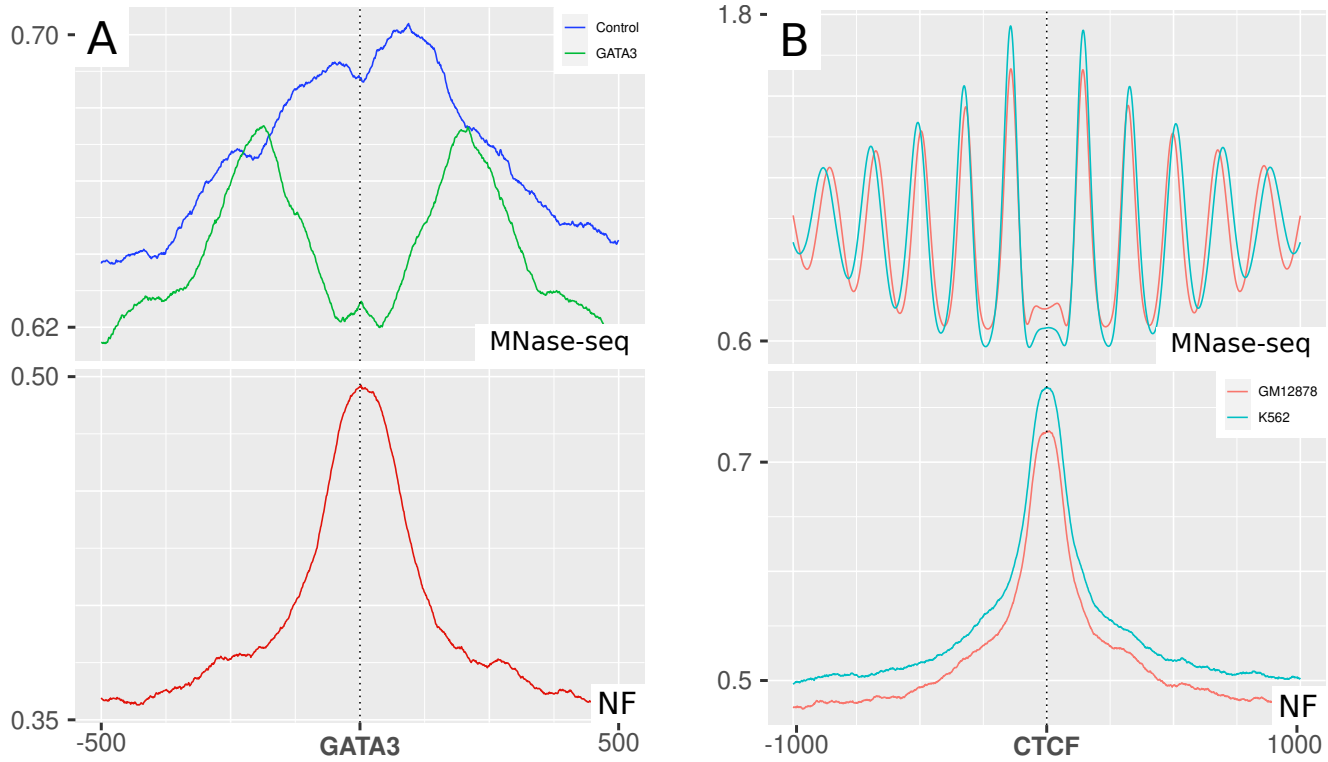


Figure 4. Nucleosome support (NF) and nucleosome occupancy (MNase-seq) surrounding the transcription factors GATA3 and CTCF. **(A)** GATA3. GATA3 binding sites of MDA-MB-23 cells with GATA3 expression (green: GATA3) and without (blue: Control). The nucleosome occupancy is higher towards the binding sites in the cells without GATA3 expression. This peak is depleted in the GATA3 expression cells and a double peak of nucleosome occupancy is left around the sites with a distance of approx. 200 bp. The nucleosome support displays a pronounced peak precisely overlapping the TFBSs, indicating the presence of an intrinsically well-positioned nucleosome that is actively depleted by GATA3. **(B)** CTCF. CTCF ChIP-seq peaks from ENCODE in GM12878 and K562. The nucleosome occupancy shows a very distinct periodicity around CTCF peaks. This periodicity of nucleosome positions is not apparent on the level of sequence-dependent nucleosome support, which exhibits a single peak directly around CTCF.

downstream. Therefore, it is reasonable to conclude that the dynamic re-localization of nucleosomes due to ATP-dependent remodeling can overwrite the sequence-intrinsic nucleosome guidance.

Sequence-mediated nucleosome occupation potential of exon regions While it was just shown that there is a sequence-independent creation of nucleosome arrays downstream of the TSS, that does not entirely describe the sequence-specific local nucleosome support for other regions along transcriptional regulation. It is described in literature that exons are targets of a particularly high nucleosome occupancy that is related to transcription speed of RNAPII and recognition of alternative splicing (10). To test whether this principle is also evident on the nucleotide sequence level, we analyzed the exons of all human genes by comparing the MNase-seq signal for K562 and GM12878 and the NF score again. Figure 5 B shows both average signals centered around the middle of 267864 exons with strand orientation taken into account. Both the nucleosome occupancy and the NF score show a distinct peak over the exon regions. The extent of the rise of the NF score towards the exon center is as high as approx. 20% from slightly above 50% to almost 70% and is distributed mainly between 100 bp upstream and 150 downstream of the exon midpoints. The nucleosome occupancy mirrors this pattern of a central peak of about 0.3-0.4 from 0.9 up to 1.25 in both cell lines around the exon. In addition, a notable observation indicates a relative depletion of MNase-seq signal directly upstream of the peak. For a more detailed analysis of these patterns, Supplementary Figure S2 provides an enhanced representation by separating the exons based on their length, revealing a higher level of nucleosome support in longer exons.

Sequence-defined nucleosome occupancy prevention around poly(A) sites in transcription termination Another important part of gene regulation and splicing is the nucleosome occupancy at poly(A) sites. These were shown to be explicitly free of nucleosomes, thus making them accessible for proteins marking the sites and interacting with RNAPII in transcriptional termination (12). We obtained the ENCODE MNase-seq signal for K562 and GM12878, as well as the NF score, for a list of 570740 poly(A) sites in the human genome (48). These are classified by the original authors into clusters of biological origin, of which we show sites in 21844 terminal exons (TE), 21844 in other exons (EX), 119071 in intergenic regions (IG) and 166393 in introns (IN). All sites are oriented by strand in 5' - 3' direction. Figure 5 shows the nucleosome occupancy and nucleosome support for these poly(A) sites. There appears to be a direct positive relation between the DNA-intrinsic nucleosome support and nucleosome occupancy. Additionally, a qualitative difference between the cluster of exons (EX) compared to all others can be observed. While both occupancy and nucleosome support

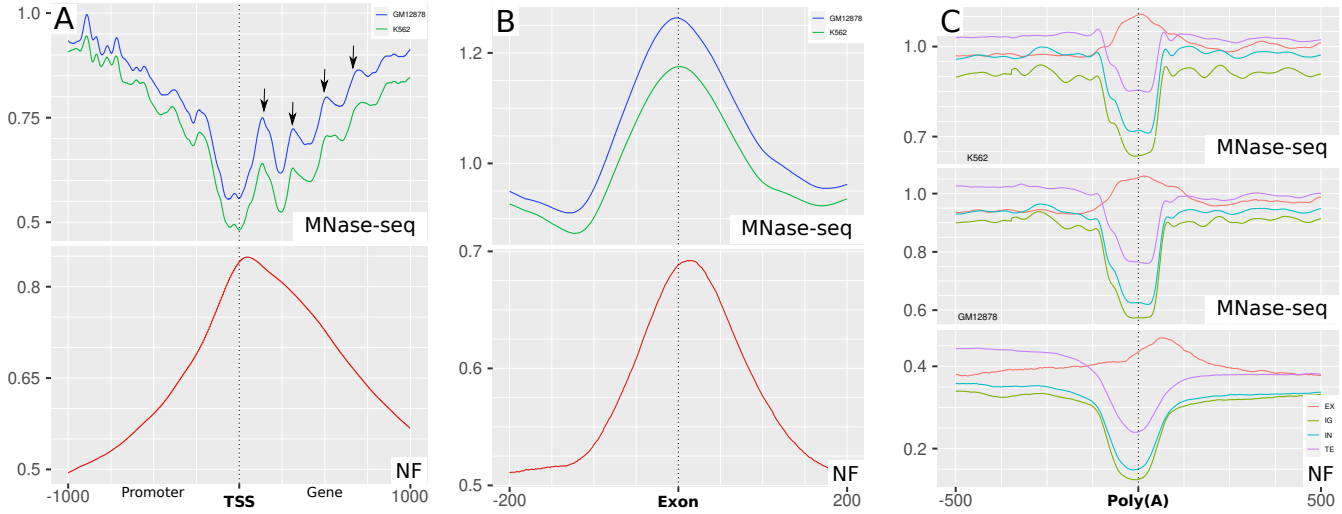


Figure 5. MNase-seq and NF score for different elements from transcription initiation to transcription termination. **(A)** Promoters. The profiles are consistently aligned to ensure that the promoter located at the 5' end is consistently positioned on the left side. The central vertical line represents the TSS location. The nucleosome occupancy declines around the TSS and shows an apparent periodicity to both directions. The NF score rises in anti-correlation towards the TSS and does not indicate the proposed periodicity. **(B)** Exons. The NF score shows a high sequence-defined nucleosome support for exon regions and experimentally a high nucleosome occupancy can be observed. This correlation strongly suggests the functional significance of the observed sequence-defined nucleosome. **(C)** Poly(A). Poly(A) sites, separated by cluster (TE=Terminal Exons, EX= Other Exons, IG = Intergenic, IN = Intron). All groups except for Exons show a distinct decline of both sequence-defined nucleosome support (NF score) and nucleosome occupancy (MNase-seq) around the site.

exhibit a decrease around the sites, the EX cluster shows an increase of 0.1 in MNase-seq in the central 250 bp and an increase of NF score of 7% at 170 bp downstream of the poly(A) site. At intronic (IN) and intergenic (IG) sites there is a 20% drop in nucleosome support relative to the already low level of 35% in the surroundings, regardless of reading direction. The same is apparent on nucleosome occupancy in both cell lines with a drop of 0.3 in MNase-seq signal. The group of poly(A) sites from the TE cluster shows a disbalance between an upstream nucleosome support of 45% upstream of the sites which recovers to 39% after a similar drop of 20% in NF score. The nucleosome occupancy in the TE group is slightly higher than the other clusters but does not exhibit a difference between up- and downstream of the poly(A) site. Supplementary Figure S3 gives an overview over all available clusters in the PolyASite 2.0 atlas, some of which were filtered out here for the sake of readability.

DISCUSSION

In this study, we analyzed the role of DNA sequence for the positioning of nucleosomes. For that purpose, we developed a random forest classifier to distinguish nucleosomal and linker DNA and used it to derived a score to estimate the sequence-defined nucleosome support of different functional elements in the human genome. The degree to which nucleosome positioning relies on DNA sequence has been a subject of extensive debate (15). These questions cannot be answered by analyzing either chromatin accessibility experiments or DNA sequence alone. On the one hand, the experimental verification of chromatin accessibility for any given cell-type will only describe the result of the multitude of counteracting influences on nucleosome positions. On the other hand, DNA can explain the importance of nucleotide-sequence on their potentially favoured binding locations, without reliably predicting the actual nucleosome occupancy in any given cell-type *in vivo*. Therefore, we compared the results of our sequence-defined nucleosome support in the form of the NF score with experimental data for chromatin accessibility and histone modifications and described the possible combinations of *in vivo* nucleosome positioning with its underlying DNA sequence support.

We could demonstrate that the DNA-intrinsic nucleosome support is not directly equivalent to the resulting nucleosome occupancy in a specific cell-line *in vivo*. Rather, there is a positive influence on nucleosome binding on the DNA of genomic regions which are involved in actively regulating transcriptional processes, such as promoters, enhancers, exons and in the center of insulators. Simultaneously, certain locations exhibit lower nucleosome support, which can serve functional purposes, such as well-positioned nucleosome occupancy arrays downstream of the transcription start site (TSS) and in the vicinity of insulators. Conversely, there are instances where lower nucleosome support indicates non-functionality, as observed in constitutive heterochromatin.

The increased reliance on DNA sequence in gene regulatory regions can be attributed to evolutionary pressures favoring the development of DNA sequences that facilitate optimal nucleosome binding, effectively regulating access to these regions. This effect has been described to be particularly important for organisms of a higher complexity (60). Especially these highly sequence-defined and evolutionary conserved positions can make a large difference in the regulation of gene activity and thus, the sequence can help to increase the resistance of accidentally triggering gene expression. The counteracting process to this state is an active eviction of the blocking nucleosome by a combination of (pioneer-) TFs, histone modification enzymes and active chromatin remodeling to gain accessibility for a particular position (61). Due to the inclusion of the relevant TFBSs

within the confined nucleosome positions, it can be inferred that there is a presumed co-evolution of proficient nucleosome organization and the existence of high-quality binding sites for activating TFs. Therefore, we interpret the crucial function of DNA in these ”regulatory” nucleosomes as a mechanism to regulate the competition between nucleosomes and TFs for spatial availability. This competition affects the energy required to activate biological processes, allowing for dynamic control and modulation. Further evidence for the systematic co-localization of regulatory proteins (high DHS score) with highly nucleosome-attracting regions (high NF score) can be observed in Supplementary Figure S4. That the accessibility of these sequence-defined nucleosomes plays a significant role in regulating gene expression is further supported by Supplementary Figure S5, in which gene expression values are compared between groups of genes based on a correlation or anti-correlation of NF score and chromatin accessibility data. The role of DNA sequence for different kinds of TFs under this assumption has been partially analyzed for different TF families (62, 63). Nevertheless, it would be worthwhile to research the differences in sequence-dependent nucleosome support in regard of pioneer functionality. The co-evolution to occupy highly important places would suggest a high nucleosome support at pioneer TFBSs. However, this presents a notable challenge as the binding sites of pioneer transcription factors are often located in close proximity to the subsequent TFs that bind in the following step. Furthermore, it is well-established that pioneer TFs can employ various modes of nucleosome eviction, which may not occur directly at the target site of chromatin accessibility (7, 64).

In contrast to these nucleosome positions determined by DNA sequence, constitutive heterochromatin, although commonly occupied by nucleosomes, does not demonstrate a comparable reliance on DNA sequence for their precise positioning. In analogy to the previous explanation of DNA evolution towards a targeted occlusion of TFBSs to prevent random gene activation, we propose a lower relevance of sequence since the higher-order packing of dense chromatin is a more unspecific process in which the accessibility of small individual positions on the genome is not expected to have such severe consequences as the revealing of TFBSs in cis-regulatory elements.

Apart from the evolution of nucleosome positioning advantage at highly competitive sites, we found other instances of locations that appear to have a high sequence-mediated nucleosome support. In particular, exons show a considerable nucleosome support relative to their surroundings. It is well described that the occupation of exon regions with nucleosomes is important in the pausing of RNAPII to ensure the correct transcription of a gene as well as the marking and spatial relocation of exons for the splicing machinery (10, 65). The principle of using the DNA to encode a favourable structure to guide nucleosomes to exon regions might again be an efficient means of increasing the intrinsic nucleosome support. In contrast to strengthening the nucleosome in competition with TFs, the nucleosome at exons might simply be guided to its place of effect without the need for further displacement by other processes. This provides another striking example of putative co-evolution that is directed to attract nucleosomes on the sequence which underlies a drive to reliably encode proteins at the same time.

The utilization of sequence to influence nucleosome positioning is evident not only in its ability to positively guide nucleosomes to specific locations. We showed the opposite effect for poly(A) sites, marking the end of transcription. The substantial contrast in sequence-dependent nucleosome support between poly(A) sites and the surrounding DNA clearly indicates the strategic prevention from nucleosome binding at these specific locations. The reason for this feature could be the attempt to keep chromatin at the TTS constantly accessible to easily allow the binding of proteins that form complexes together with the RNAPII to terminate transcription, cleave the mRNA and do the polyadenylation in the consecutive step. The nucleosome rejection fits well to the general observation of low nucleosome occupancy over Poly(A) sites. However, there are hints about the role of increased nucleosome occupancy around one end of the sites, regulating alternative polyadenylation (12). This could be the reason for a difference of nucleosome support towards both directions surrounding Poly(A) sites in terminal exons.

Opposed to the steering of functional nucleosomes by nucleotide sequence, we identified processes which exhibit strong nucleosome positioning that does not manifest on DNA sequence level. Notably, downstream of the TSS, a well-established pattern of regularly spaced nucleosome array formation is observed. At these locations the nucleosome support constantly decreases towards intronic locations without mirroring the periodicity of the nucleosome occupancy. This process is known to be mediated by active chromatin remodeling complexes which are organizing nucleosomes into an evenly spaced array by using energy from ATP (9). This process serves as an example for an important transcriptional mechanism which is not directly supported by local DNA structure. Overall, the nucleosome support around the +1 nucleosome stayed relatively high but the reason for that is most likely the proximity to the regulatory promoter elements which are well positioned by sequence.

In the end, it is notable that these sequence-dependent and -independent nucleosome positioning processes do not have to be mutual exclusive. This phenomenon is particularly pronounced at insulator sites, where the transcription factor CTCF plays a crucial role in delineating TAD boundaries. Here, it could be observed that on top of the actual binding location there is a very strong sequence-defined support for the nucleosome, suggesting the previously described mechanism of competition and co-evolution of TF-nucleosome interaction. Yet, as a result in the surroundings of CTCF bound locations, there is a clear well-positioned nucleosome array which is the result of active chromatin remodeling without the help of sequence support. While the binding of CTCF underlies the aforementioned competition with positioned nucleosomes, the spacing and symmetry of the surrounding array could be linked to active chromatin remodelers, such as Chd4 and Snf2h (5). This is just one example of the very close, local interplay of sequence support and energy-dependent remodeling of chromatin to decide the final *in vivo* positions of nucleosomes in transcriptional regulation. It is impossible to favour any of these principles in their importance for the collaborative regulation of transcription, from the formation of TAD boundaries, over the regulation of gene activity through the accessibility of chromatin for TFs or support mechanisms of the tasks

of RNAPII, to the clean termination of transcription, thus marking the last step in which the DNA is involved directly.

DATA AVAILABILITY

The code to execute predictions can be found at https://gitlab.gwdg.de/MedBioinf/generegulation/nfclassifier. The NF scores in BIGWIG file format can be downloaded at https://bioinformatics.umg.eu/resources/nfscore.

FUNDING

Research was funded by Deutsche Forschungsgemeinschaft (DFG) (KFO5002 to MS and SFB1002 to NP).

CONFLICT OF INTEREST DISCLOSURE

None declared.

ACKNOWLEDGEMENTS

We thank Argyris Papantonis and Alexander Ecker for fruitful discussions. MS is a member of the International Max Planck Research School for Genome Science (IMPRS-GS), part of the Göttingen Graduate Center for Neurosciences, Biophysics, and Molecular Biosciences (GGNB). NP is a member of Molecular Medicine PhD program, part of the Göttingen Graduate Center for Neurosciences, Biophysics, and Molecular Biosciences (GGNB). TB and MH are members of the Göttingen Campus Institute Data Science (CIDAS). Also, we thank the ENCODE Consortium and the ENCODE production laboratories generating the particular data sets.

REFERENCES

1. Roger D. Kornberg. Chromatin structure: A repeating unit of histones and DNA. *Science*, 184(4139):868–871, May 1974.
2. Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, September 1997.
3. Elphège P. Nora, Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L. van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385, April 2012.
4. Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, April 2012.
5. Christopher T Clarkson, Emma A Deeks, Ralph Samarista, Hulkar Mamayusupova, Victor B Zhurkin, and Vladimir B Teif. CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length. *Nucleic Acids Research*, 47(21):11181–11196, October 2019.
6. Kenneth S. Zaret and Jason S. Carroll. Pioneer transcription factors: establishing competence for gene expression. *Genes & Development*, 25(21):2227–2241, November 2011.
7. Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, 39(9):381–399, September 2014.
8. William K. M. Lai and B. Franklin Pugh. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nature Reviews Molecular Cell Biology*, 18(9):548–562, May 2017.
9. Elisa Oberbeckmann, Vanessa Niebauer, Shinya Watanabe, Lucas Farnung, Manuela Moldt, Andrea Schmid, Patrick Cramer, Craig L. Peterson, Sebastian Eustermann, Karl-Peter Hopfner, and Philipp Korber. Ruler elements in chromatin remodelers set nucleosome array spacing and phasing. *Nature Communications*, 12(1), May 2021.
10. Alberto R. Kornblihtt, Ignacio E. Schor, Mariano Alló, Gwendal Dujardin, Ezequiel Petrillo, and Manuel J. Muñoz. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Reviews Molecular Cell Biology*, 14(3):153–165, February 2013.
11. Nova Fong, Hyunmin Kim, Yu Zhou, Xiong Ji, Jinsong Qiu, Tassa Saldi, Katrina Diener, Ken Jones, Xiang-Dong Fu, and David L. Bentley. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes & Development*, 28(23):2663–2676, December 2014.
12. Huan Huang, Jiao Chen, Hongde Liu, and Xiao Sun. The nucleosome regulates the usage of polyadenylation sites in the human genome. *BMC Genomics*, 14(1), December 2013.
13. Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, May 2007.
14. Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, February 2011.
15. Kevin Struhl and Eran Segal. Determinants of nucleosome positioning. *Nature Structural & Molecular Biology*, 20(3):267–273, March 2013.
16. Yong Zhang, Zarmik Moqtaderi, Barbara P Rattner, Ghia Euskirchen, Michael Snyder, James T Kadonaga, X Shirley Liu, and Kevin Struhl. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nature Structural & Molecular Biology*, 16(8):847–852, July 2009.
17. Noam Kaplan, Irene K. Moore, Yvonne Fondufe-Mittendorf, Andrea J. Gossett, Desiree Tillo, Yair Field, Emily M. LeProust, Timothy R. Hughes, Jason D. Lieb, Jonathan Widom, and Eran Segal. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366, December 2008.
18. Sandra C. Satchwell, Horace R. Drew, and Andrew A. Travers. Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology*, 191(4):659–675, October 1986.
19. Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thåström, Yair Field, Irene K. Moore, Ji-Ping Z. Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, July 2006.
20. Daniel J. Gaffney, Graham McVicker, Athma A. Pai, Yvonne N. Fondufe-Mittendorf, Noah Lewellen, Katelyn Michelini, Jonathan Widom, Yoav Gilad, and Jonathan K. Pritchard. Controls of nucleosome positioning in the human genome. *PLoS Genetics*, 8(11):e1003036, November 2012.
21. Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of DNA shape in protein–DNA recognition. *Nature*, 461(7268):1248–1253, October 2009.
22. Tianyin Zhou, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Research*, 41(W1):W56–W62, May 2013.
23. Alexandre V. Morozov, Karissa Fortney, Daria A. Gaykalova, Vasily M. Studitsky, Jonathan Widom, and Eric D. Siggia. Using DNA mechanics to predict in vitro nucleosome positions and formation energies. *Nucleic Acids Research*, 37(14):4707–4722, June 2009.
24. Aakash Basu, Dmitriy G. Bobrovnikov, Basilio Cieza, Juan Pablo Arcon, Zan Qureshi, Modesto Orozco, and Taekjip Ha. Deciphering the mechanical code of the genome and epigenome. *Nature Structural & Molecular Biology*, 29(12):1178–1187, December 2022.
25. Eran Segal and Jonathan Widom. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Current Opinion in Structural Biology*, 19(1):65–71, February 2009.
26. Hui Liu, Ruichang Zhang, Wei Xiong, Jihong Guan, Ziheng Zhuang, and Shuigeng Zhou. A comparative evaluation on prediction methods of nucleosome positioning. *Briefings in Bioinformatics*, 15(6):1014–1027, September 2013.
27. Vladimir B. Teif. Nucleosome positioning: resources and tools online. *Briefings in Bioinformatics*, 17(5):745–757, September 2015.
28. Guo-Sheng Han, Qi Li, and Ying Li. Comparative analysis and prediction of nucleosome positioning using integrative feature representation and machine learning algorithms. *BMC Bioinformatics*, 22(S6), June 2021.
29. Mariya Shtumpf, Kristan V. Piroeva, Shivam P. Agrawal, Divya R. Jacob, and Vladimir B. Teif. NucPosDB: a database of nucleosome positioning

in vivo and nucleosomics of cell-free DNA. *Chromosoma*, 131(1-2):19–28, January 2022.
30. Liqun Xi, Yvonne Fondufe-Mittendorf, Lei Xia, Jared Flatow, Jonathan Widom, and Ji-Ping Wang. Predicting nucleosome positioning using a duration hidden markov model. *BMC Bioinformatics*, 11(1), June 2010.
31. Kun Fang, Tianbao Li, Yufei Huang, and Victor X. Jin. NucHMM: a method for quantitative modeling of nucleosome organization identifying functional nucleosome states distinctly associated with splicing potentiality. *Genome Biology*, 22(1), August 2021.
32. Shou-Hui Guo, En-Ze Deng, Li-Qin Xu, Hui Ding, Hao Lin, Wei Chen, and Kuo-Chen Chou. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*, 30(11):1522–1529, February 2014.
33. Mattia Di Gangi, Giosuè Lo Bosco, and Riccardo Rizzo. Deep learning architectures for prediction of nucleosome positioning from sequences data. *BMC Bioinformatics*, 19(S14), November 2018.
34. Juhua Zhang, Wenbo Peng, and Lei Wang. LeNup: learning nucleosome positioning from DNA sequences with improved convolutional neural networks. *Bioinformatics*, 34(10):1705–1712, January 2018.
35. Etienne Routhier, Edgard Pierre, Ghazaleh Khodabandelou, and Julien Mozziconacci. Genome-wide prediction of DNA mutation effect on nucleosome positions for yeast synthetic genomics. *Genome Research*, 31(2):317–326, December 2020.
36. Guo-Sheng Han, Qi Li, and Ying Li. Nucleosome positioning based on DNA sequence embedding and deep learning. *BMC Genomics*, 23(S1), April 2022.
37. Maria Tsompana and Michael J Buck. Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin*, 7(1), November 2014.
38. Dustin E. Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, March 2008.
39. Alan P. Boyle, Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, January 2008.
40. The Encode Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.
41. Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
42. Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. DNAshapeR: an r/bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, 32(8):1211–1213, December 2015.
43. Motoki Takaku, Sara A. Grimm, Takashi Shimbo, Lalith Perera, Roberta Menafra, Hendrik G. Stunnenberg, Trevor K. Archer, Shinichi Machida, Hitoshi Kurumizaka, and Paul A. Wade. GATA3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler. *Genome Biology*, 17(1), February 2016.
44. Katherine S. Pollard, Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, October 2009.
45. D. Karolchik. The UCSC table browser data retrieval tool. *Nucleic Acids Research*, 32(90001):493D–496, January 2004.
46. Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, February 2012.
47. Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, November 2015.
48. Christina J Herrmann, Ralf Schmidt, Alexander Kanitz, Panu Artimo, Andreas J Gruber, and Mihaela Zavolan. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Research*, October 2019.
49. Michael Lawrence, Robert Gentleman, and Vincent Carey. rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics*, 25(14):1841–1842, May 2009.
50. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.
51. Hadley Wickham. *Ggplot2: Elegant graphics for data analysis*. Use R! Springer International Publishing, Cham, Switzerland, 2 edition, June 2016.
52. Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9(8):e1003118, August 2013.
53. T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):7881, 2005.
54. Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.
55. Kyoung-Jae Choi, My Diem Quan, Chuangye Qi, Joo-Hyung Lee, Phoebe S. Tsoi, Mahla Zahabiyon, Aleksandar Bajic, Liya Hu, B. V. Venkataram Prasad, Shih-Chu Jeff Liao, Wenbo Li, Allan Chris M. Ferreon, and Josephine C. Ferreon. NANOG prion-like assembly mediates DNA bridging to facilitate chromatin reorganization and activation of pluripotency. *Nature Cell Biology*, 24(5):737–747, April 2022.
56. Brandon J. Aubrey, Andreas Strasser, and Gemma L. Kelly. Tumor-suppressor functions of the TP53 pathway. *Cold Spring Harbor Perspectives in Medicine*, 6(5):a026062, May 2016.
57. Hiroki Tanaka, Yoshimasa Takizawa, Motoki Takaku, Daiki Kato, Yusuke Kumagawa, Sara A. Grimm, Paul A. Wade, and Hitoshi Kurumizaka. Interaction of the pioneer transcription factor GATA3 with nucleosomes. *Nature Communications*, 11(1), August 2020.
58. Nick Owens, Thaleia Papadopoulou, Nicola Festuccia, Alexandra Tachtsidi, Inma Gonzalez, Agnes Dubois, Sandrine Vandormael-Pournin, Elphège P Nora, Benoit G Bruneau, Michel Cohen-Tannoudji, and Pablo Navarro. CTCF confers local nucleosome resiliency after DNA replication and during mitosis. *eLife*, 8, October 2019.
59. Lu Bai and Alexandre V. Morozov. Gene regulation by nucleosome positioning. *Trends in Genetics*, 26(11):476–483, November 2010.
60. Marco Tompitak, Cédric Vaillant, and Helmut Schiessel. Genomes of multicellular organisms have evolved to attract nucleosomes to promoter regions. *Biophysical Journal*, 112(3):505–511, February 2017.
61. Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, January 2019.
62. Desiree Tillo, Noam Kaplan, Irene K. Moore, Yvonne Fondufe-Mittendorf, Andrea J. Gossett, Yair Field, Jason D. Lieb, Jonathan Widom, Eran Segal, and Timothy R. Hughes. High nucleosome occupancy is encoded at human regulatory sequences. *PLoS ONE*, 5(2):e9129, February 2010.
63. Iros Barozzi, Marta Simonatto, Silvia Bonifacio, Lin Yang, Remo Rohs, Serena Ghisletti, and Gioacchino Natoli. Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Molecular Cell*, 54(5):844–857, June 2014.
64. Fangjie Zhu, Lucas Farnung, Eevi Kaasinen, Biswajyoti Sahu, Yimeng Yin, Bei Wei, Svetlana O. Dodonova, Kazuhiro R. Nitta, Ekaterina Morgunova, Minna Taipale, Patrick Cramer, and Jussi Taipale. The interaction landscape between transcription factors and the nucleosome. *Nature*, 562(7725):76–81, September 2018.
65. Robin Andersson, Stefan Enroth, Alvaro Rada-Iglesias, Claes Wadelius, and Jan Komorowski. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Research*, 19(10):1732–1741, August 2009.