SORBONNE UNIVERSITÉ

École Doctorale Sciences de la Nature et de l'Homme – ED 227

THÈSE DE DOCTORAT

Spécialité : **Biologie moléculaire**

Présentée et soutenue publiquement par

Maxime Christophe

le **05 décembre 2025**

Décrypter le positionnement des nucléosomes à l'aide de réseaux de neurones profonds interprétables

Une approche fondée sur la séquence pour explorer la structure de la chromatine dans les cellules souches embryonnaires murines

Sous la direction de :

M. Julien Mozziconacci, Professeur d'université, MNHN

M. Pablo Navarro Gil, Directeur de recherche, Institut Pasteur

Jury :

Mme Federica Battistini	Lecturer Professor, UB, Barcelone – Rapportrice
Mme Émilie Elvira-Matelot	Chargée de recherche, INSERM – Rapportrice
Mme Élodie Laine	Professeure d'université, SU, Paris – Examinatrice
M. Julien Mozziconacci	MNHN, Paris – Directeur de thèse
M. Pablo Navarro Gil	Institut Pasteur, Paris – Co-directeur de thèse
M. Vladimir Teif	Associate Professor, Université d'Essex – Rapporteur

Présidé par :

Mme Élodie Laine Professeure d'université, SU, Paris

Avant-propos

Cette thèse a été réalisée au sein du Muséum National d'Histoire Naturelle, Département Adaptation du Vivant, au sein de l'unité *STRucture et INstabilité des Génomes* (CNRS UMR7196, INSERM U1154, MNHN, Sorbonne Université), dans l'équipe *ADN Répété Chromatine et Évolution*.

Le présent manuscrit a été rédigé et soutenu en langue anglaise sous le titre :

Deciphering Nucleosome Positioning Using Interpretable Deep Neural Networks

A Sequence-Based Approach to Explore Chromatin Structure in Mouse Embryonic Stem Cells

Remerciements

Je remercie les membres de mon jury pour avoir accepté et pris le temps de lecture, d'évaluation et de discussion de mon travail de recherche.

Julien, merci pour ton accompagnement, ta vision et ton enthousiasme scientifique. Nos échanges furent précieux, y compris ceux qui n'ont jamais pris la forme de mots.
Pablo, merci pour ta codirection et ton expertise.

Je remercie l'ensemble du laboratoire Structure et Instabilité des Génomes qui m'a accueilli et permis de passer ces trois années en son sein. Je remercie tout particulièrement Jean-Baptiste pour ses valeurs personelles et scientifiques. Lauréline pour sa simplicité et son naturel qui ont été un repère précieux. Bernadette pour son authenticité. Thomas pour son regard extérieur. Alexandra pour les discussions scientifiques et celles -toutes aussi importantes- qui l'étaient moins.
Je remercie et je souhaite du courage à mes collègues doctorants et doctorantes avec qui nous avons partagé une partie de ce voyage, notamment Alex avec qui j'ai pu le plus partager et particulièrement Valentin qui est devenu un ami.

Je remercie ma mère, qui m'a appris à faire de mon mieux dans ce que je sais faire mais aussi ce que je ne pense pas savoir faire.
Je remercie mon père, le seul à m'avoir expliqué *a priori* ce qu'était vraiment une thèse.

Je remercie et félicite Clara avec qui je partage tant depuis longtemps et pour plus longtemps encore, ainsi que tous les amis et les amies sympathiques qui m'entourent. Je remercie Roxane pour m'avoir fait contractuellement rire.

Je remercie les amis parisiens et les amies parisiennes, qui partagent mon quotidien.

Ya3ychek, n7ebbek barcha ya Bel7a

Et merci à toi aussi, qui te reconnaîtras.

Pour Georges et Sydney

Résumé/Abstract

Résumé

Les nucléosomes, unités fondamentales de la chromatine, enroulent l'ADN et participent à la fois à l'architecture du génome et à la régulation génique. Si la séquence d'ADN influence leur positionnement in vitro, son impact in vivo reste limité [1]. Le modèle de barrière [2] propose une organisation autour d'un nucléosome bien positionné. Nous introduisons un modèle de « génome ponctué », dans lequel des régions de positionnement (Nucleosome Positioning Regions, NPRs) induisent la formation de nucléosomes phasés ou de régions dépourvues de nucléosomes (Nucleosome Depleted Regions, NDRs). Ces NPRs incluent facteurs de transcription, répétitions dispersées et microsatellites.

En suivant l'approche développée chez la levure (Routhier et al., 2021), nous avons entraîné un modèle de deep learning interprétable sur le génome murin, afin d'extraire des règles prédictives adaptées aux génomes de grande taille. Les réseaux neuronaux convolutifs (CNN) sont bien adaptés pour capturer des caractéristiques multi-échelles, des k-mers aux motifs et à leurs combinaisons. Nous avons utilisé des cartes de positionnement nucléosomique en cellules souches embryonnaires murines (mESC), issues de deux méthodes complémentaires :

(1) MNase-seq, basé sur la digestion de l'ADN linker mais limité par des biais enzymatiques [3] ;

(2) Clivage chimique, qui modifie les histones pour cliver l'ADN via un système cuivre/H_2O_2, réduisant ces biais [4].

L'apprentissage sur un génome riche en séquences répétées (∼60 %) impose d'éviter le surapprentissage tout en conservant leur information. Nous avons filtré les régions à faible couverture/cartographiabilité puis évalué les prédictions par mutagenèse in silico, générant une carte de scores à la paire de bases. Ces scores ont été analysés avec XSTREME [5] et complétés par des expériences de génomique synthétique pour tester directement l'effet des éléments identifiés.

Le modèle retrouve des déterminants connus, comme CTCF, puissant inducteur de phasage nucléosomique [3, 6]. Sa capacité à phaser dépend de la conservation du motif et d'un pré-motif en amont, et notre approche distingue précisément les motifs effectivement liés, validés par ChIP-seq. De façon inattendue, nous identifions la famille SP/KLF comme nouveaux contributeurs au phasage. Nous montrons aussi que certains éléments répétitifs dispersés participent activement à l'organisation : certains SINEs, porteurs de motifs CTCF, agissent comme points de phasage [7],

tandis que les LINEs et LTR contribuent également, suggérant un rôle structurel plus large des répétitions dans la chromatine des mammifères. Enfin, les microsatellites révèlent des signatures de positionnement dépendant de leur séquence, renforçant leur potentiel régulateur.

Ainsi, notre approche de deep learning interprétable met en évidence une logique combinatoire où facteurs de transcription, répétitions dispersées et microsatellites structurent l'organisation nucléosomique dans les génomes complexes.

Abstract

Nucleosomes, the fundamental units of chromatin, wrap DNA and contribute both to genome architecture and gene regulation. While DNA sequence influences nucleosome positioning in vitro, its impact in vivo remains limited [1]. The barrier model [2] proposes an organization centered around a well-positioned nucleosome. Here, we introduce a "punctuated genome" model, in which Nucleosome Positioning Regions (NPRs) promote the formation of phased nucleosome arrays or nucleosome-depleted regions (NDRs). These NPRs include transcription factors, dispersed repeats, and microsatellites.

Following the approach developed in yeast [8], we trained an interpretable deep learning model on the mouse genome to extract predictive rules applicable to large genomes. Convolutional Neural Network (CNN) are well suited to capture multi-scale genomic features, from short k-mers to motifs and their combinations. We used nucleosome maps from mouse embryonic stem cells (mESCs) obtained with two complementary methods:

(1) MNase-seq, based on linker DNA digestion but limited by enzymatic bias [3];

(2) Chemical cleavage, which modifies histones to cleave DNA via a copper/H_2O_2 system, thereby reducing such biases [4].

Training on a genome rich in repetitive sequences ($\sim$60 %) required strategies to avoid overfitting while retaining informative content. We filtered out regions of low coverage/mappability and evaluated predictions using in silico mutagenesis, generating base-pair resolution score maps. These scores were analyzed with XSTREME [5] and complemented with synthetic genomics experiments to directly test the effect of identified elements.

Our model recovers known determinants such as CTCF, a strong inducer of nucleosome phasing [3, 6] . Its phasing capacity depends on motif conservation and the presence of an upstream pre-motif, and our approach accurately distinguishes motifs bound in vivo, as confirmed by ChIP-seq. Unexpectedly, we identify the SP/KLF family as novel contributors to nucleosome phasing. We also show that certain dispersed repeats actively participate in chromatin organization: some SINEs carrying CTCF motifs act as phasing points [7], while LINEs and LTRs also contribute, suggesting a broader structural role of repeats in mammalian chromatin. Finally, microsatellites display sequence-dependent nucleosome positioning patterns, reinforcing their regulatory potential.

Thus, our interpretable deep learning approach highlights a combinatorial logic in which transcription factors, dispersed repeats, and microsatellites collectively shape nucleosome organization in complex genomes.

Mots-clefs/Key-words

Mots-clefs

positionnement des nucléosomes ; apprentissage profond ; réseaux de neurones convolutifs ; intelligence artificielle interprétable ; séquence d'ADN ; architecture de la chromatine ; MNase-seq ; clivage chimique ; éléments régulateurs cis ; éléments répétitifs ; génome de souris ; facteurs de transcriptions ; CTCF ; génomique synthétique

Key-words

nucleosome positioning; deep learning; convolutional neural networks; interpretable AI; DNA sequence; chromatin architecture; MNase-seq; chemical cleavage; cis-regulatory elements; repetitive elements; mouse genome; transcription factors; CTCF; synthetic genomics

Contents

List of Abbreviations

cfDNA circulating cell-free DNA

ChIP-seq Chromatin ImmunopreciPitation with sequencing

CNN Convolutional Neural Network

DNA DesoxyriboNucleic Acid

ERV Endogenous RetroVirus

G4 G-quadruplex

GPU Graphical Processing Unit

IAP Intracisternal A-particle

ISM *In Silico* Mutagenesis

LINE Long Interspersed Nuclear Element

LTR Long Terminal Repeat

mESC mouse embryonic stem cells

NDR Nucleosome Depleted Regions

NFR Nucleosome Free Region

NLP Natural Language Processing

NN Neural Network

NOMe-seq Nucleosome Occupancy and Methylome Sequencing

NPR Nucleosome Positioning Region

NRL Nucleosome Repeat Length

PIC Pre-Initiation Complex

PWM Position Weight Matrix

RNA RiboNucleic Acid

SINE Short Interspersed Nuclear Element

TAD Topological Associated domain

TE Transposable Element

TF Transcription Factor

TFBS Transcription Factor Binding Site

TR Tandem Repeat

TSS Transcription Start Site

ZF Zinc Finger

Chapter 1

The Genome: Structure, Function, and Organization

The genome is the complete set of information that encodes the structure and function of cells, tissues, and the organism as a whole. DesoxyriboNucleic Acid (DNA) is a long, double-stranded molecule that stores the information for cell phenotype and more generally, whole organism for multicellular. in the form of a chemical code. Each strand is composed of a sugar-phosphate backbone, to which nitrogenous bases (also called nucleotides) adenine (A), thymine (T), cytosine (C), and guanine (G) are attached. The two strands wind around each other to form a double helix, a structure famously revealed by Rosalind Franklin [9]. DNA is ubiquitous in all organisms and serves as the universal medium for storing genetic information, each cell carry at least a copy of this genome which can be free of organize in a nucleus. This information is transcribed into RiboNucleic Acid (RNA), which differs from DNA by its chemical composition, as its nucleotides contain a ribose sugar and uracil instead of thymine, and is mostly found in single-stranded form. This RNA is then translated into chains of amino acids that fold into proteins. These proteins perform a wide range of structural and functional roles at every level of biological organization. Remarkably, the genetic code that governs this flow of transcription and translation is nearly identical across all forms of life [10].

1.1 Multicellular genomes are gene-sparse

The central characteristic of living organisms is reproduction, which ensures the continuity of life across generations. A fundamental challenge of reproduction is the faithful duplication of genetic material. In unicellular organisms, cell division by mitosis allows both the preservation and the propagation of the organism, as each division produces a new, independent cell. In multicellular organisms, by contrast, reproduction usually involves meiosis, which generates gametes with half the genetic content, later restored at fertilization. In both contexts, replication in its broadest sense (the accurate copying of genetic material) remains essential, as it underpins both cellular continuity and the

transmission of hereditary information.

In unicellular organisms, each cell functions independently and will replicate again in full autonomy. In contrast, multicellular organisms are composed of cells that assemble into tissues, organs, and ultimately the whole organism. These cells are specialized and no longer operate autonomously. This raises a fundamental question: how can different cell types, all carrying the same genetic material, acquire distinct phenotypes and cooperate to sustain the organism's global biological functions ? In multicellular organisms, specialization arises from stem cells which could basically adopt any phenotype. This process, also known as cell fate determination, is a finely tuned mechanism that programs each cell with a specific set of expressed proteins, thereby defining its phenotype. Genome regulation is both qualitative and quantitative: certain proteins are absent from specific cell types, while others are ubiquitous but expressed at different levels depending on the cell type.

Genomes can be broadly characterized by the number of genes they contain and their overall length. Intuitively, one might expect that genome length scales proportionally with gene number across species. However, in multicellular organisms, this correlation breaks down: genome size increases dramatically without a corresponding rise in gene count.

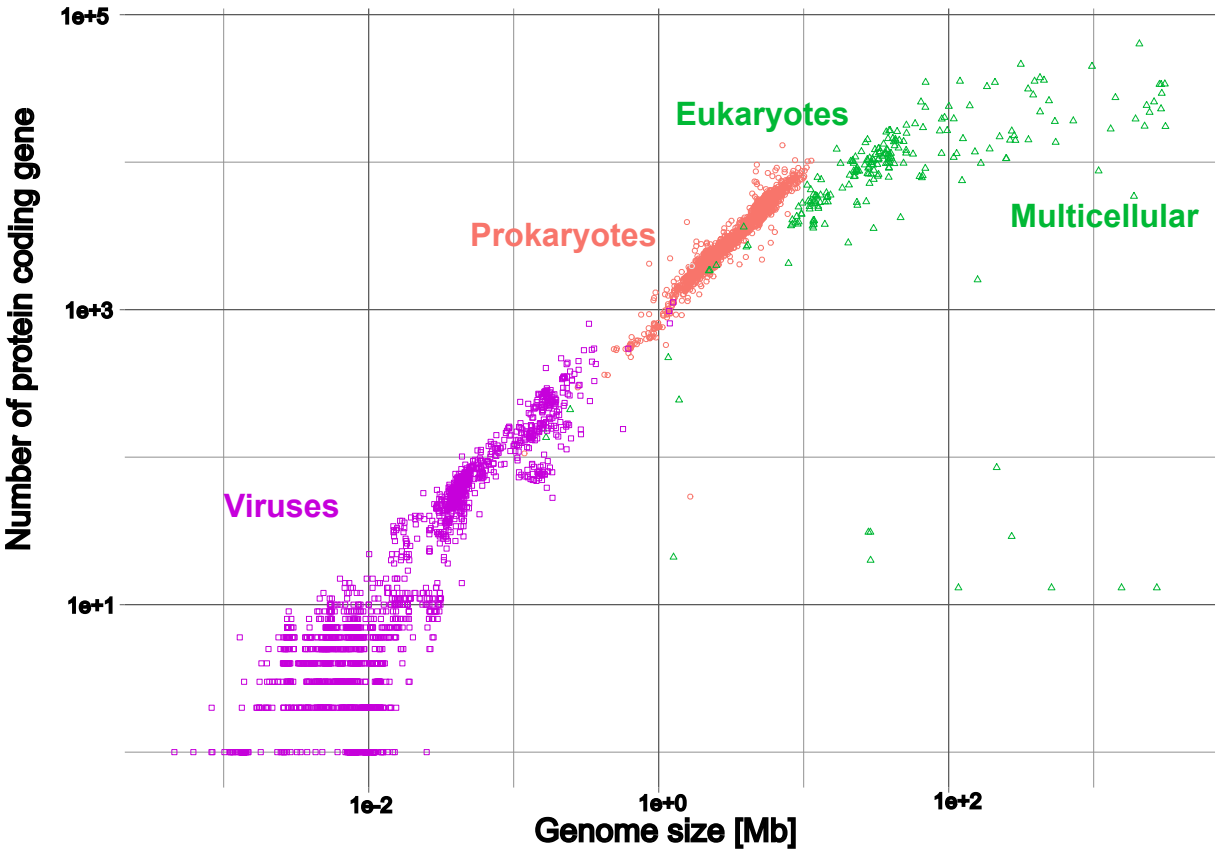


Figure 1.1: **Relationship between genome size and number of genes.**Log-log plot of the total number of annotated proteins in genomes submitted to GenBank as a function of genome size. Based on data from NCBI genome reports, styled after Koonin (2011). Modified from a figure by Estevezj, available on Wikipedia, licensed under CC BY-SA 3.0.

This observation allows us to distinguish between:

(1) Compact genomes, in which a large proportion of the DNA sequence codes for genes such as in Saccharomyces cerevisiae, where coding sequences represent approximately 70% of the genome [11] and up to 97% for some bacterial genomes [12].

(2) Gene-sparse genomes, typical of multicellular organisms, where intergenic and intronic regions dominate, with 1.2% of coding regions for the human genome [13].
The gene-sparsity of complex genomes is thought to reflect the need for sophisticated and flexible gene regulation such as enhancers, silencers, and insulators, enabling cellular differentiation and the emergence of diverse cell types [14].

Indeed, the expansion of non-coding regions may have been a prerequisite for the evolution of multicellularity [15]. Two major factors have been proposed to explain the expansion of genomes in multicellular organisms: whole-genome duplications and the proliferation of repetitive elements. Whole-genome duplications provide raw material for evolutionary innovation by creating gene redundancy, which can lead to subfunctionalization or neofunctionalization. This process facilitates the emergence of new regulatory and developmental functions, potentially supporting the evolution of complex multicellularity [16]. Phylogenetic studies on early metazoans show that critical transcription factors required for development are highly conserved and predate the divergence of modern animal lineages. However, subsequent genome expansions, together with the retention of multiple gene copies, enabled parallel evolution and the co-option of duplicated genes for new functions [17] such as the well-studied POU [18] and Sox families [19]. This increase in the repertoire of transcription factors allowed a refinement of gene regulation and the emergence of highly specialized cell phenotypes, suggesting that genome expansion was one of the key drivers of multicellularity.

Repetitive elements, particularly transposable elements, represent a substantial fraction of many eukaryotic genomes. Beyond their contribution to genome size, these sequences can be co-opted as regulatory elements, such as enhancers or insulators [20], and shape chromatin architecture, allowing long range regulation [21]. Their expansion is therefore thought to play a critical role in increasing genome size [22], therefore potential regulatory complexity, which is essential for cell differentiation and tissue specialization. We will return later to their contribution to higher-order chromatin structures such as domains and compartments.

In both cases, genome expansion is ultimately constrained by nuclear volume and can only occur in parallel with efficient chromatin compaction, thereby placing nucleosome positioning as a central determinant of genome expansion, evolutionary innovation, and genomic complexity.

1.2 Anatomy of the chromatin

DNA in the nucleus is not found in a naked form, but rather is packaged with histone-proteins to form a complex structure known as chromatin. Chromatin plays a key role in regulating gene accessibility and genome stability. Its fundamental unit is the nucleosome, which consists of 147 base pairs of DNA wrapped around a histone octamer. The dynamic organization of chromatin allows the genome to be compacted to fit within the nucleus while still enabling precise control over gene expression, replication, and repair.

Persistence length

The persistence length is a basic mechanical property quantifying the bending stiffness of a polymer. While it is rigorously defined using the polymer's diameter, Young's modulus, and bending stiffness, it can be simply understood as the typical length scale over which the polymer's deflection is negligible, and appears straight. Below this length, the polymer resists bending and behaves like a rigid rod. Beyond it, thermal fluctuations become strong enough to introduce noticeable bends, so the polymer starts to look flexible and can eventually fold back or cross over itself. A famous example of this property is the extreme (and theoretically estimated) 10^{18}m persistence length of uncooked spaghetti [23], meaning that spaghetti will always break before bending, even at very large scales. It is often insightful to compare a polymer's persistence length to its actual contour length, as this ratio determines whether the polymer behaves more like a flexible thread or a stiff rod. In the same spirit, raw DNA has a large persistence length ($\sim$500 Å, depending on the sequence) compared to its diameter (20 Å) [24, 25], making it effectively a rigid polymer at its scale.

1.2.1 Nucleosomes handle chromatin compaction

Each cell of our organism carries a copy of the genome, which is stored in the cell nucleus. This constitutes a real physical challenge, as a stretched copy of the entire genome in mammals can reach several meters in length, while the cell nucleus is only about 10 µm in diameter. On top of that, DNA has a relatively high persistence length among natural polymers (see text box).

Living organisms have developed different strategies for DNA compaction. In this process, eukaryotes differ from viruses or bacteria by relying on histone proteins (H2A, H2B, H3, and H4) to assist with DNA compaction [26]. These proteins assemble into an octamer to form the so-called nucleosome. DNA wraps around the nucleosome core particle in 1.65 turns of a 147-bp left-handed helix. On top of this wrapping, successive

nucleosomes can assemble into higher-order structures, stabilized by the linker histone H1. H1 binds near the nucleosome dyad (the central base pair where the DNA entry and exit points are symmetric) and simultaneously contacts the two linker DNAs, forming a stem-like structure that reduces their flexibility and promotes a regular trajectory of nucleosomes in the fiber [27]. This higher-order structure achieves a compaction ratio between 30 and 40, meaning that one micrometer of chromatin fiber contains 30–40 micrometers of naked DNA [28].

The Nucleosome Repeat Length (NRL) is defined as the average distance between the centers of two adjacent nucleosomes, *i.e.*, the sum of the 147 bp of DNA wrapped around the histone octamer and the linker DNA connecting consecutive nucleosomes. It provides a simple metric to describe the large-scale organization of chromatin and reflects how tightly nucleosomes are packaged along the genome. The NRL typically ranges between 165 and 220 bp. Variations in NRL have been linked to differences in cell type, developmental state, and histone H1 occupancy, and thus it is widely used as a quantitative descriptor of chromatin compaction and accessibility [29–31].

1.2.2 The spatial and temporal organization of the genome

The genome is not a static linear molecule, but a dynamic and spatially organized entity. The concept of the 4D genome refers to the integration of three-dimensional genome architecture with its temporal dynamics across the cell cycle, development, and environmental responses.

First, chromatin is not uniformly packed. Spatial studies of the nucleus have revealed two major chromatin states, organized into megabase-scale compartments. Compartment A is mostly accessible, enriched in transcriptionally active regions, and is generally associated with euchromatin. In contrast, compartment B is densely packed, enriched in silent genomic regions with tight DNA-DNA interactions, and is generally associated with heterochromatin [32, 33]. Compartments can be subdivided on a finer scale. Topological Associated domain (TAD) are contiguous regions with a median size of $\sim$880 kbp that are enriched in internal contacts [34]. Unlike compartments, TADs are conserved across cell types and can switch between compartments [35]. This spatial organization plays a critical role in gene regulation by facilitating or restricting interactions between enhancers, promoters, and other regulatory elements. Advances in techniques such as Hi-C or Micro-C are now revealing how the genome folds, unfolds, and rearranges to coordinate complex gene regulatory programs in space.

Over time, cells progress through different phases known as the cell cycle. For most of it, the cell is in an interphase state, during which transcription and genome regulation are active. During interphase (G1, S, and G2 phases), the genome is actively transcribed, replicated, and regulated, while maintaining a relatively open chromatin organization [36]. In contrast, during mitosis chromatin undergoes maximal condensation and extensive rearrangement into visible chromosomes. This global reorganization leads to a loss of

accessibility and transcriptional silencing, as most regulatory factors are displaced. Nevertheless, some histone modifications and transcription factors remain bound, a process known as mitotic bookmarking, which facilitates the rapid reactivation of transcriptional programs in daughter cells [3].

A multitude of molecular machineries govern multidimensional genome organization, including transcription factors and their complexes, chromatin remodelers, and the replication machinery. While these factors interact with nucleosomes and influence their positioning, in this thesis we focus specifically on sequence determinants, acknowledging that they act in concert with the broader epigenetic and regulatory landscape.

1.2.3 Nucleosomes in genome regulation

Genome regulation refers to the control of gene activity. Genes can be active, inactive, or expressed at intermediate levels, depending on the action of protein complexes that enhance or repress their expression. Transcription requires the recruitment of RNA polymerase II to the Transcription Start Site (TSS). To achieve this, transcription factors bind to enhancer regions and recruit chromatin remodelers that shape the chromatin to make it accessible to other transcription factors and to the Pre-Initiation Complex (PIC), including RNA polymerase II. Enhancer regions can be located proximally (within hundreds of base pairs) or at long distances, sometimes over megabases [37]. Spatial studies discussed previously showed that these interaction (even the distant one) occurs in TADs, which loop-shaped region of chromatin, extruded through cohesin-CTCF complex [32]. The initiation of the transcription requires the effective binding of transcription factors to the DNA, therefore chromatin accessibility. Nucleosome perturbation at the TSS can both repress or activate the transcription by influencing the PIC formation on particular genes [38, 39]. This represents the most intuitive regulatory function of nucleosomes: by physically restricting access to DNA, they can modulate the binding of transcription factors and the activity of the transcriptional machinery.

Beyond their structural role, nucleosomes also carry epigenetic modifications such as histone acetylation, methylation, or ubiquitination, that act as regulatory signals. These histone marks can either activate or repress gene expression by altering chromatin accessibility or recruiting chromatin-associated complexes [40, 41]. In this way, nucleosomes not only organize the genome spatially but also encode regulatory information that contributes to gene expression programs. Several well-characterized histone modifications that illustrate how nucleosomes integrate structural and regulatory functions are listed in Table 1.1.

Altogether, these mechanisms highlight that nucleosomes act not only as packaging units but also as dynamic regulators of genome function, bridging DNA sequence, chromatin organization, and transcriptional control

Histone mark	Location	Functional association
H3K4me3	Promoters (TSS)	Active transcription initiation [40, 41]
H3K27ac	Enhancers, promoters	Active enhancers, open chromatin [42]
H3K36me3	Gene bodies	Transcription elongation, splicing [40, 41]
H3K27me3	Polycomb regions	Transcriptional repression [43]
H3K9me3	Pericentromeric heterochromatin	Constitutive heterochromatin, silencing [44]
H4K20me3	Heterochromatin, telomeres	Genome stability, DNA repair, repression [45]
H3/H4 acetylation	Promoters, enhancers	Open chromatin, transcription activation [40]

Table 1.1: **Examples of histone modifications and their regulatory functions**

1.3 Repeated elements in sparse genomes

This study broadly addresses repeated elements, which constitute over 60% of the mouse genome [46]. Once dismissed as genomic "junk", these sequences are now increasingly recognized for their functional roles in transcriptional regulation [47] and 3D chromatin architecture [21]. Inherited from ancient mobile elements, some of these repeats have proliferated and contributed to regulatory innovation [7, 48]. Repeats can be broadly classified into two categories: transposable elements and satellite DNA.

1.3.1 From parasitism to partnership: the dual nature of transposable elements

Transposable Elements (TEs) constitute a major fraction of the repetitive genome and play a dual role as both drivers of genetic innovation and sources of genomic instability. First dicovered in maize by Barbara Mc Clintock [49], they propagate through two distinct mechanisms. Retrotransposons (class I) follow a "copy-and-paste" mechanism: they are transcribed into RNA, reverse-transcribed into cDNA, and reintegrated into the genome. DNA transposons (class II), in contrast, move through a "cut-and-paste" process, excising themselves from one genomic locus and inserting into another [50]. The main transposable elements families are described in Table 1.2.

Historically dismissed as genomic parasites due to their self-replicating nature, TEs are now recognized as key regulators of genome function, particularly in development and cell identity [20]. Their relationship with the host genome is shaped by an evolutionary conflict: while unchecked TE activity can threaten genome integrity, their insertional diversity also provides a rich substrate for regulatory innovation.

TEs can profoundly influence genome function and organismal fitness through a cascade of effects ranging from local gene regulation to broader impacts on cellular phenotype [20]. Upon integration, a TE may disrupt coding sequences, alter regulatory landscapes, or introduce new cis-regulatory elements such as promoters, enhancers, or insulators. While many TE insertions are neutral and ultimately lost through random mutation, deleterious insertions often impair cellular function and are eliminated through negative selection. In some cases, however, harmful insertions can persist if they are

TE class/type	Canonical structure	Most common form in host genomes
LTR retrotransposons (e.g., ERVs)	LTR – *gag* – *pol* – *env* – LTR	Solitary LTRs or fragments with loss of coding regions; still potentially regulatory
LINEs	5'UTR – ORF1 – ORF2 – 3'UTR	5'-truncated insertions, often lacking promoter and ORF1
SINEs	Short non-coding sequence with internal Pol III promoter	Largely intact due to their small size; retain promoter
DNA transposons	TIR – *Transposase* – TIR	MITEs (short elements with TIRs, lacking coding regions)

Table 1.2: **Canonical vs. truncated structures of major TE types.** LTR: Long Terminal Repeat, TIR: Terminal Inverted Repeat

epigenetically silenced by the host. Finally, a subset of insertions may acquire beneficial regulatory activity and be retained by positive selection. This dynamic interplay between loss, repression, and co-option contributes to the recurrent evolutionary turnover of regulatory sequences.

To mitigate the potential threat of uncontrolled TE activity, host genomes have evolved multiple, hierarchically layered mechanisms of epigenetic repression. In mouse embryonic stem cells (mESC), the chromatin landscape of transposable elements follows their evolutionary age and regulatory potential. Young and recently mobilized elements, which remain capable of transcription or enhancer-like activity, are the most stringently controlled. They are strongly enriched for H3K9me3 and SETDB1-dependent repression, and frequently co-occupied by poised or active marks such as H3K27ac or H3K4me1, resulting in bivalent chromatin states [51]. This tight control reflects the host's need to silence potentially active TE copies while preserving the possibility of regulated usage. In contrast, older and degenerated elements—largely transcriptionally inert—tend to carry constitutive but weaker repressive marking. Over evolutionary time, many such elements become stably integrated into host regulatory networks through molecular domestication.

Recent work shows that specific families of zinc-finger proteins (ZFP), including ZFP with SCAN domain, derived from a co-opted retroviral capsid domain, bind distinct TE subfamilies and actively anchor nucleosomes over these TE-derived regulatory sequences. Loss of these factors results in nucleosome redistribution and, in some cases, acquisition of enhancer-like chromatin states [52]. These findings illustrate how domesticated TEs can form the substrate for rapidly evolving regulatory modules whose function relies on precise chromatin positioning. In this continuum, epigenetic repression is therefore strongest at evolutionarily young elements where the risk of uncontrolled activity is highest, and progressively relaxes as elements decay, become inert, or are co-opted into host gene-regulatory architecture.

A prominent example of this repression occurs with the young murine L1Md Long Interspersed Nuclear Element (LINE) family. The promoters found on 5' present tandem repeated monomers, their number and sequence composition vary between subfamilies [53], but they often harbor multiple short motifs capable of recruiting transcription factors (e.g. YY1, RUNX3, SOX) and influencing transcription initiation [54]. Two observations highlight why repression is necessary yet must be timed. In somatic contexts, loss of H3K9me3-based repression at intronic L1Md perturbs nearby gene programs (observed during or ionization) [55]. Conversely, a brief, coding-independent burst of LINE-1 activity after fertilization promotes global chromatin accessibility, whereas premature silencing or prolonged activation are deleterious [56]. This stage-dependent logic helps explain how TE sequences can later be co-opted as *bona fide* regulatory modules.

In some cases, the regulatory properties of TE have been co-opted by the host genome. TE-derived sequences can be stably integrated into gene regulatory networks, especially during development, where they serve as alternative enhancers or promoters, and may even contribute to species-specific traits. This interplay between repression, innovation, and selection makes TEs both a threat and a source of evolutionary novelty. Endogenous RetroViruss (ERVs) stand out for their regulatory potential in mESCs. While most are transcriptionally silenced, a subset remains accessible and carries active histone marks. Their Long Terminal Repeat (LTR), including solo LTRs, are frequently bound by core pluripotency factors such as OCT4, SOX2, and NANOG, and can act as alternative promoters or enhancers [48]. These regulatory activities, though initially incidental, have been co-opted by the host genome to shape transcriptional programs during early development [20].

One well-studied example of such co-option involves the SINE B2 family of retrotransposons. These elements can harbor CTCF motifs and thereby act as boundary elements that limit the expansion of active chromatin domains [57]. Their regulatory influence, however, is strongly modulated by the epigenetic environment: enrichment in repressive histone marks such as H3K9me3 can silence B2 motifs, preventing their activity as chromatin borders and thus safeguarding the genome against uncontrolled regulatory interference [58]. On an evolutionary scale, TE expansions have contributed to the rewiring of genome architecture and regulatory networks. In mammals, lineage-specific insertions can create novel transcription factor binding sites and alter chromatin organization. In the mouse genome, Schmidt et al. demonstrated that SINE B2 elements are particularly enriched for newly acquired CTCF binding sites [7]. These insertions have introduced functional CTCF motifs that reshape chromatin loops, illustrating how waves of TE activity may remodel the three-dimensional genome organization in rodents.

In summary, TEs represent a dynamic and versatile component of the genome. Once viewed as selfish DNA, they are now increasingly seen as partners in gene regulation, whose activity is modulated by a finely tuned balance of conflict, repression, and evolutionary co-option.

1.3.2 Functional Implications of Satellite DNA and Microsatellites in Genome Architecture

Satellite DNA refers to tandemly repeated sequences, predominantley located in heterochromatic regions such as centromeres and telomeres. Based on the size of the repeat unit, they are classified into satellites, minisatellites, and microsatellites. Though once dismissed as inert DNA, satellite repeats are now known to play key roles in chromatin compaction, centromere identity, and genome integrity [59]. Because the mm10 genome assembly does not provide a telomere-to-telomere reconstruction, centromeric and telomeric regions remain poorly annotated, precluding detailed analysis of satellite DNA at these loci. Nevertheless, other families of Tandem Repeat (TR) are present in the assembly; a prominent example is the LINE-1 TE-derived tandem repeat formed by part of the ORF2 and the 3'LTR [60].

Microsatellites, also known as simple sequence repeats (SSRs), are highly polymorphic and prone to replication slippage, making them valuable as genetic markers. Recent studies have highlighted their multifaceted roles in genome biology, including modulation of transcription factor binding, spacing of regulatory elements, cytosine methylation, alternative splicing, mRNA stability, transcriptional initiation and termination, unusual secondary structures, nucleosome positioning and modification, higher-order chromatin organization, non-coding RNA production, and meiotic recombination hotspots [61].

Chapter 2

State of the art: in search of nucleosome positioning determinants

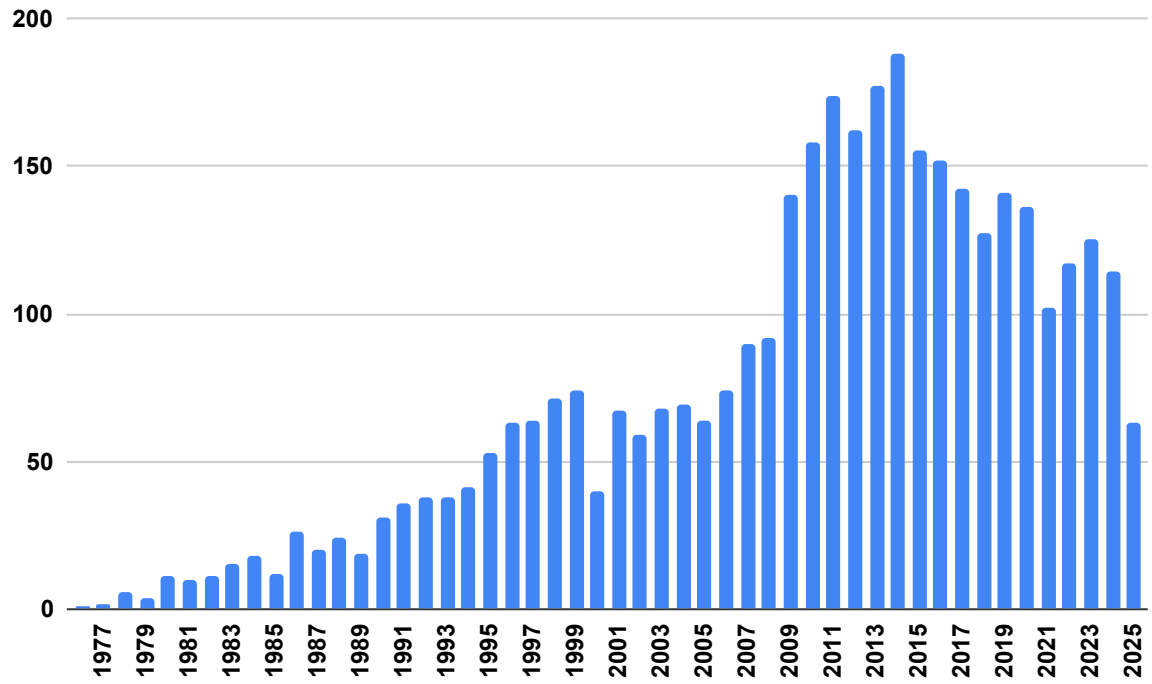


Figure 2.1: **Number of publications per year retrieved from PubMed using the query "nucleosome positioning"** (accessed on June 23, 2025).

2.1 Experimental methods for nucleosome positioning

Nucleosome positioning plays a crucial role in understanding how chromatin structure influences gene regulation and genome folding. Since the first biochemical and electron microscopy evidence of nucleosome arrangement in chromatin by Compton et al. in 1976 [62], the study of nucleosome positioning has become a central topic in molecular biology. Over the decades, a wide range of experimental and computational methods

have been developed to characterize how nucleosomes are distributed along the genome. The increasing interest in this topic is illustrated by the growing number of scientific publications over the years (Figure 2.1).

2.1.1 Microscopy

The first evidence of nucleosome particles on DNA was observed using electron microscopy [63], and their regular arrangement was later confirmed through similar techniques [62]. This gave rise to the "beads-on-a-string" model, suggesting that the genome was "remarkably uniform and simple" [64]. Since then, observing nucleosomes has remained a technical challenge, but microscopy techniques have continuously improved, offering increasingly higher resolution. Over the past decade, many long-held assumptions about chromatin structure have been overturned. Super-resolution microscopy revealed that nucleosomes are not evenly positioned across the genome, but rather form small, heterogeneous clusters termed “nucleosome clutches”, interspersed throughout the chromatin [65]. Furthermore, regular spacing or higher-order folding, such as the classical 30 nm fiber, has been shown to occur only in vitro under specific conditions, and no evidence of this kind of chromatin structure has been found in vivo [65–67].

2.1.2 MNase-sequencing

Several molecular biology techniques have been developed to map nucleosome positioning genome-wide. Among the most widely used is MNase-seq (Micrococcal Nuclease digestion followed by sequencing), which takes advantage of the nucleosome's ability to protect the underlying DNA from enzymatic digestion. The resulting protected DNA fragments, typically around 147 base pairs in length, are then sequenced and mapped to infer nucleosome positions with high resolution [68, 69]. Figure 2.2 provides a simplified illustration of this principle. In actual experiments, remaining fragments are filtered to retain nucleosome-sized fragments, typically between 130 and 200 base pairs.

This technique presents certain limitations, as micrococcal nuclease exhibits sequence preferences, particularly favoring AT-rich regions and underrepresenting GC-rich DNA [70]. Furthermore, chromatin compaction is highly heterogeneous, especially between heterochromatin (tightly packed) and euchromatin (loosely organized), making it difficult to capture nucleosome positions with uniform accuracy. In compact regions, limited enzymatic accessibility may lead to underrepresentation, while "fragile" nucleosomes, which are highly sensitive to digestion, may be lost under standard conditions. To address these issues and improve genome-wide consistency, some authors have implemented digestion gradients, varying enzyme concentration or exposure time to better capture nucleosome occupancy across chromatin states [71]. The work presented in this manuscript is based on the deep MNase-seq data published by Festuccia et al and processed by Luis Altamirano from Pablo Navarro Gil's team at Institut Pasteur [3] which provided both an experimental signal uniquely-mapped and multi-mapped (see box).

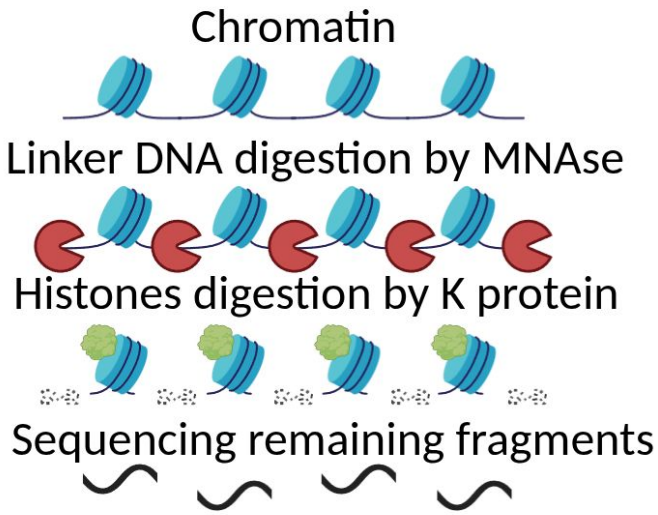


Figure 2.2: **Graphical abstract of MNase-seq assay**. DNA-linkers are digested by MNase while the nucleosomal fragments remains intact and are then sequenced. Created with Biorender

Genome coverage and mappability

Mappability quantifies how uniquely a k-mer can be mapped within a given genome. It is a key metric used to identify genomic regions where reads may align ambiguously. At each genomic position p, the mappability score is defined as the inverse of the number of occurrences of the k-mer starting at p in the genome; that is,

$$\text{Mappability}(p) = (\#\,\{\text{genomic matches of sequence}[p : p + k]\})^{-1}$$

A major challenge when studying mammal DNA is the difficulty of mapping small reads in some parts of the genome due to the highly represented repeated elements (low mappability). Using a signal resulting from reads that map uniquely to the genome will result in non covered region. One strategy to overcome this limitation is to allow ambiguous mapping by randomly attributing the read to one of the best match.
In our context, as the MNase-seq data were generated using Illumina technology, mappability was computed for 150-mers across the genome (k=150).

2.1.3 Chemical Cleavage

Chemical cleavage mapping relies on targeted hydroxyl radical cleavage near the DNA backbone at sites that flank the nucleosome center symmetrically, through copper-mediated reactions. It enables high-resolution nucleosome positioning by directly probing DNA–histone contacts [72]. This technique was developed to overcome some of the limitations of MNase-seq, particularly its sequence bias and sensitivity to digestion conditions. Notably, Voong et al. (2017) demonstrated that chemical cleavage can recover fragile nucleosomes that are typically lost with MNase digestion [73]. However, the method requires genetically modifying cells to express the H4S47C histone variant (substituting serine 47 with cysteine on histone H4 to allow copper tagging) and involves substantial computational processing to accurately reconstruct nucleosome maps. The experimental data published by Voong et al. have been used to train the chemical cleavage model.

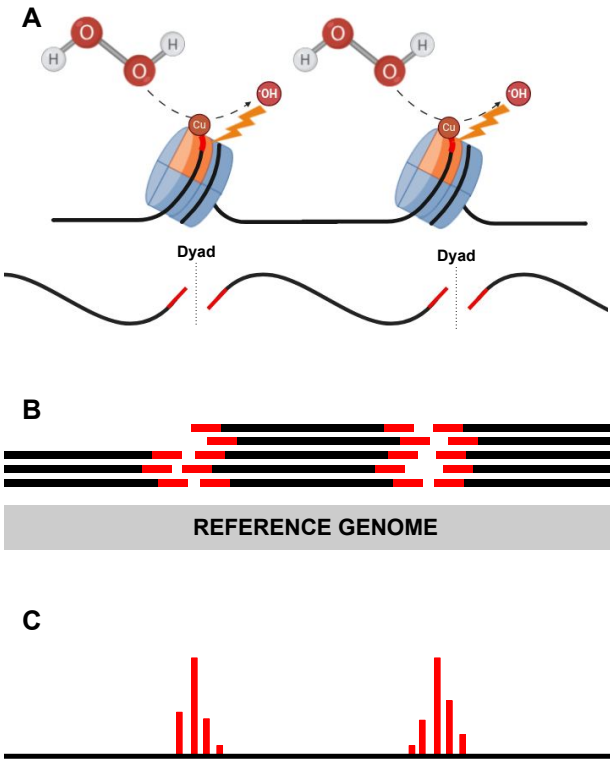


Figure 2.3: **Graphical abstract of chemical cleavage assay.**
A. Histone H4 serine 47 is substituted with cysteine (H4S47C) to covalently attach a sulfhydryl-reactive, copper-chelating label, N-(1,10-phenanthrolin-5-yl)iodoacetamide. Upon addition of hydrogen peroxide, a Fenton reaction occurs, generating localized hydroxyl radicals that cleave DNA near the nucleosomal dyad.
B. The resulting DNA fragments, typically ranging from 140 bp to 200 bp, are aligned to the reference genome.
C. Signal deconvolution enables reconstruction of nucleosome dyad positions genome-wide.
Created with bio-render, adapted from Voong *et al.* [4]

2.1.4 Histone ChIP-seq

Chromatin ImmunopreciPitation with sequencing (ChIP-seq) uses antibodies to target specific protein and enable to retrieve their positiong using sequencing. Leveraging the structure of nucleosome which is an octamer of histones, it is possible to use ChIP-seq on histone. This approach is way less resolutive than MNase-seq or Chemical cleavage but it is interesting to flag active and inactive region if the ChIP-seq targets epigenetics modification on histones.

2.1.5 Transposase-based method

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a fast and sensitive technique to map open chromatin regions using Tn5 transposase. While primarily used to identify accessible regulatory regions, fragment size distribution can indirectly inform nucleosome positioning: short fragments (<100bp) correspond to nucleosome-free regions, whereas mononucleosome-sized fragments (around 150bp) indicate protected DNA. Although less precise than MNase-seq or chemical cleavage for nucleosome mapping, ATAC-seq remains widely used due to its simplicity and suitability for low-input and single-cell applications [74].

CUT&RUN (Cleavage Under Targets and Release Using Nuclease) and CUT&Tag (Cleavage Under Targets and Tagmentation) are antibody-directed methods enabling high-resolution mapping of protein–DNA interactions. In CUT&RUN, a protein A–micrococcal nuclease fusion is recruited to chromatin by a specific antibody and selectively cleaves DNA adjacent to the bound protein, releasing fragments into solution for sequencing [75]. CUT&Tag follows a similar principle but employs a protein A–Tn5 fusion that simultaneously cleaves and inserts sequencing adapters at antibody-bound sites, streamlining library preparation and reducing background [76]. Both methods require fewer cells and produce higher signal-to-noise ratios than ChIP-seq, making them valuable tools for profiling transcription factors, histone modifications, and chromatin-associated proteins with minimal input.

2.1.6 Single molecule nucleosome occupancy

Nucleosome Occupancy and Methylome Sequencing (NOMe-seq) uses a CpG methyltransferase to probe chromatin accessibility before bisulfite conversion, providing a simultaneous readout of nucleosome occupancy and DNA methylation at single-molecule resolution [77]. This method relies on the principle that accessible regions of DNA (those not occupied by nucleosomes or bound proteins) are methylated by the exogenous enzyme. After bisulfite conversion and sequencing, unmethylated CpG sites indicate protected regions, whereas methylated sites correspond to accessible DNA. NOMe-seq is particularly powerful because it captures both nucleosome positioning and endogenous CpG methylation in the same molecule, thus linking chromatin structure with epigenetic state. It has

been applied in various contexts, including cancer epigenomics and early development, where chromatin accessibility and DNA methylation play crucial regulatory roles.

PCP (Proximity Copy-Paste) assigns unique DNA barcodes to DNA segments that are in 3D proximity on the same long molecule, enabling single-molecule readout of nucleosome positions, spacing, and long-range connectivity [78]. By proximity-tagging and copying barcode sequences across contacting regions on an intact fiber, PCP preserves both the linear order of nucleosomes and their higher-order associations on a single read. In practice, it resolves regularly spaced arrays that can be positioned or delocalized, detects multi-way long-range contacts consistent with loop clustering, and reveals non-canonical particles such as overlapping di-nucleosomes, all within one experiment.

Both approaches resolve chromatin at the single-molecule (long-read) scale: PCP captures physical connectivity and fiber architecture, whereas NOMe-seq captures accessibility and epigenetic state. Importantly, PCP does not sequence native ultra-long reads directly; instead, it uses proximity barcoding to reconstruct continuous single-molecule fibers at long-read scale, which still entails technically demanding library preparation, stringent controls, and dedicated reconstruction and analysis pipelines.

2.1.7 Adaptations to single-cell resolution

Recent advances have enabled the adaptation of nucleosome mapping techniques to single-cell resolution, revealing chromatin heterogeneity across individual cells. Techniques such as single-cell MNase-seq, scATAC-seq, and single-cell NOMe-seq have been developed to capture nucleosome positioning and chromatin accessibility in individual nuclei. These approaches offer insights into cell-to-cell variability in chromatin structure and gene regulation, which are especially relevant in developmental biology, cancer, and tissue complexity. However, single-cell methods often suffer from sparse coverage and increased technical noise, requiring careful computational processing and aggregation to extract meaningful patterns. Despite these limitations, they represent a critical step toward understanding genome organization at the cellular level.

2.1.8 Medical perspective of nucleosome studies

This emergence of experimental methods show the interest of the community for nucleosome positioning studies (Figure 2.1) which is a central and fundamental question to better grasp the genome regulation and refine the global knowledge of life sciences. However the study of nucleosome is also of clinical interest. Deep sequencing of circulating cell-free DNA (cfDNA) found in plasma has been used to infer nucleosome footprints. Snyder et al. (2016) demonstrated that it is possible to retrieve the cell type of origin from sequenced cfDNA and to generate a genome-wide nucleosome density map [79]. The idea is that cfDNA originates predominantly from leukocytes undergoing apoptosis or other forms of cell death. As a consequence, the nucleosome landscape inferred from cfDNA fragmentation patterns largely reflects the chromatin organization of circulating immune

cells in healthy individuals. However, many pathologies alter the composition, activation state, or turnover rate of specific blood-cell populations. Such changes modify the relative contributions of different cell types to the cfDNA pool. The resulting nucleosome landscape becomes a weighted mixture of cell-type–specific chromatin signatures, shifted away from the profile characteristic of healthy leukocytes. Quantifying the magnitude and direction of this shift provides information about the underlying condition and can serve as a sensitive marker of disease state or immune dysregulation. This technique is minimally invasive and holds great potential for the diagnosis of cancer-related disorders [80, 81] or sepsis [82, 83].

2.2 Computational methods for nucleosome positioning

In addition to the diverse experimental techniques developed to map nucleosome positions genome-wide, a growing number of computational tools have been introduced to predict nucleosome occupancy directly from the DNA sequence. In 2016, Vladimir Teif cataloged 19 online tools dedicated to the theoretical prediction of nucleosome positioning, reflecting the growing interest in sequence-based modeling approaches [84]. These tools span a wide range of methodologies, including statistical mechanics, DNA bendability models, and machine learning algorithms. Since then, the number of available resources has continued to grow, with more than forty computational tools now accessible for either analyzing experimental data or predicting intrinsic nucleosome formation probabilities from sequence alone. This proliferation of tools highlights the complexity of nucleosome positioning and the need for integrative approaches that combine biophysical principles with high-throughput data analysis. The Teif Lab maintains a curated and regularly updated list of these resources is maintained by the Gene Regulation Teif Lab, offering a valuable reference for researchers navigating this evolving landscape [85].

2.2.1 Sequence-based models

In the late 1980s and early 1990s, early computational efforts aimed to identify consensus high-affinity sequences for nucleosomes by aligning nucleosome-bound DNA [86]. However, these alignment-based approaches failed to uncover a clear canonical motif, highlighting the degenerate and context-dependent nature of nucleosomal DNA, but they revealed sequence composition preferences and periodic features, such as dinucleotide periodicity. Later improvements in sequence alignment algorithms confirmed the absence of a motif and the presence of dinucleotide preferences [87].

This limitation prompted a shift toward statistical modeling approaches, which aimed to capture the complex and subtle sequence features associated with nucleosome affinity. Rather than searching for a strict consensus, these models evaluated how local sequence composition, dinucleotide frequencies, and periodic patterns contribute to nucleosome formation. A key advance came with the work of Segal et al. (2006), who developed a probabilistic model trained on *in vitro* reconstituted nucleosomes, incorporating 5-mer statistics and periodic filters to predict nucleosome occupancy across the genome [88]. This approach was later extended by Kaplan et al. (2009), who used genome-wide *in vitro* reconstitution data to refine sequence-based predictions and systematically compare them to *in vivo* nucleosome maps, thus highlighting the interplay between DNA-encoded signals and trans-acting regulatory mechanisms [69]. This statistical framework provided key insights into the sequence determinants of nucleosome occupancy. Yet, because it does not explicitly model the mechanics of DNA deformation, it leaves unanswered questions

about the physical feasibility of nucleosome formation at predicted sites. To address this, a complementary line of research grounded in biophysical principles has emerged.

2.2.2 Statistical positioning models

Building upon early theoretical work such as the Tonks gas model [89], which states that gas particles along a one-dimensional lattice will exhibit density oscillations around an obstacle, Kornberg et al. proposed the idea of nucleosome statistical positioning [90]. In this view, the genome harbors barriers against which nucleosomes stack and position according to Tonks' principle. This simple model was applied and confirmed in yeast MNase-seq data [2, 91]. However, later studies showed that the barrier model alone, as an interaction-free framework, was not sufficient to fully explain nucleosome positioning. Nucleosome–nucleosome interactions were demonstrated to be necessary to fit experimental data, and refinements such as a two-body potential [92] or an explicit attractive force between nucleosomes [93] were proposed.

The statistical positioning model is now challenged because it relies on assumptions about the location of chromatin barriers and is not sufficient to explain nucleosome positioning without incorporating nucleosome–nucleosome interactions. Nevertheless, it still remains a good average model, especially in yeast, where gene-dense genomes present regularly positioned TSS, explaining the majority of nucleosome positioning.

2.2.3 Biophysical models

Biophysical models aim to predict nucleosome positioning by simulating the physical and chemical constraints that govern the interaction between DNA and histone proteins. These models typically rely on principles from polymer physics, electrostatics, and structural biology to estimate the energetic cost of wrapping DNA around the histone octamer at specific loci.

One class of models, based on DNA elasticity, treats the double helix as a deformable polymer and quantifies the bending energy required to accommodate the superhelical path imposed by the nucleosome. Building upon the idea that dinucleotides position nucleosomes because of the intrinsic physical properties of DNA, Miele developed a nucleosome positioning model using only sequence-dependent DNA flexibility and intrinsic curvature [94]. The addition of histone–DNA interactions allowed DNABEND to refine sequence-specific energy landscapes [95]. These elasticity-based models are complemented by structure-based methods, which derive atomistic interaction potentials from high-resolution nucleosome crystal structures [96]. Other approaches incorporate chromatin packing constraints, such as NRL or steric hindrance, into energetic models of nucleosome arrays [97, 98]. Monte Carlo simulations have also been used to explore nucleosome arrangement at the population level by simulating nucleosome exclusion and spacing constraints in the presence of experimental occupancy data [99]. Probabilistic extensions of these models have emerged more recently. For instance, HiddenFoot integrates

single-molecule footprinting data with a biophysical formalism (using thermodynamical principles) to infer nucleosome and transcription factor occupancy at base-pair resolution [100].

Despite their mechanistic interpretability, biophysical models face important limitations. Many models rely on equilibrium thermodynamics and neglect the inherently dynamic nature of chromatin, including the activity of ATP-dependent remodelers and transcriptional processes [47]. Energy models trained on specific sequence or structural features may not generalize across different cell types, species, or chromatin contexts. Incorporating genome-wide experimental data (e.g., MNase-seq) into these frameworks often entails high computational cost and susceptibility to experimental noise. The influence of transcription factors, histone modifications, and chromatin boundaries remains difficult to integrate explicitly and requires additional biophysical assumptions. In response to these challenges, recent efforts have turned to deep learning approaches. Neural network–based models bypass the need for explicit feature engineering by learning complex, nonlinear relationships directly from DNA sequence. These models can capture a broader spectrum of regulatory signals and outperform traditional models in predicting nucleosome occupancy, particularly in complex, cell-type-specific chromatin environments.

2.2.4 Machine learning approach

Machine learning refers to a broad set of computational techniques that enable mathematical models to learn from data. Instead of defining explicit rules, these methods infer patterns and structures automatically by optimizing model parameters. Common approaches include decision trees, which split data based on informative features; k-means clustering, which groups similar data points without labels; and linear or logistic regression, which model relationships between variables. These models are used for tasks such as classification, prediction, clustering, and dimensionality reduction. Their simplicity and interpretability make them widely applicable across domains, from biology to economics. Among these techniques, neural networks have enabled remarkable advances across disciplines. They now power applications ranging from medical image analysis [101] and natural language translation [102] to autonomous driving [103] or protein structure prediction [104]. Such versatility stems from their ability to capture highly nonlinear and hierarchical relationships within complex datasets.

Neural Networks principles

A simple way to understand neural networks is to imagine a machine that can convert numbers into some other numbers. Since the machine has parameters, it is possible to adjust them until the numbers produced by the device match the expected ones. Mathematically speaking the machine will just perform series of matrix product and thresholding. This technology has its roots in the work of

Gauss and Legendre on regression at the end of the XVIIIe century, long before the computer existed. Today, neural networks are ubiquitous and have evolved to accomplish various tasks from predicting prices, diagnosing diseases or even talking almost like a human.

2.2.4.1 Neural Networks for genomics

Advancements in sequencing and high-throughput technologies (new generation sequencing) since the close of the previous century have resulted in a substantial increase in data collection within the field of genomics. Consequently, the field of computer science has emerged as a pivotal component of genomics research. Originally conceptualized in the 1960s, neural networks were limited in their practical applications due to computational constraints. The emergence of efficient hardware like Graphical Processing Unit (GPU) has made it feasible to train deep neural networks by performing millions of calculations per second [105]. Neural network have proven its ability to decipher complex rules in genome regulation such as syntax of motif [106]. It can also serve as a sandbox to experiment synthetic sequence or mutation effects.

2.2.4.2 Convolutional Neural Networks for genomics

The principle of Convolutional Neural Network (CNN) is based on emulating the visual process occurring in animal brains, with the idea that object recognition arises from detecting and combining patterns at different scales [107, 108]. This principle has been adapted to computation and is illustrated in Figure 2.4. Starting from the detection of basic shapes through convolution, a CNN progressively reduces the dimensionality of the input while preserving the most relevant features (via pooling). This process can be repeated multiple times at different scales (by varying convolution filter size and pooling operations), and the extracted features are finally combined through fully connected layers to recognize more sophisticated patterns at higher levels of abstraction. A significant advantage of the CNN architecture is the strong reduction in receptive field size and input dimensionality before the final fully connected layer, which substantially decreases the computational cost compared to architectures relying solely on multiple fully connected layers.

Largely adopted for image recognition with the rise of standard datasets and a worldwide race for improved performance [109, 110], CNNs have also proven highly effective for studying the long sequences of letters that constitute DNA. These models learn to detect patterns at different scales, such as short DNA motifs, and to capture the syntax of these motifs. In doing so, CNNs can uncover intricate relationships within DNA sequences, making them powerful tools in fields such as genomics and bioinformatics. The first deep-learning models for biological sequences demonstrate large improvement compared to traditionnal machine learning models for different tasks such as predicting

DNA/RNA-protein binding [111], annotating genomes [112, 113] or predicting chromatin structure [114].

Several studies have also addressed nucleosome positioning using deep learning architectures trained directly on raw sequence [115–118]. These models, which range from CNNs to more complex multi-layer networks, consistently achieve high predictive accuracy across species, confirming that nucleosome occupancy is strongly encoded in the DNA sequence itself. However, their contribution has remained mostly at the predictive level.

Efforts to interpret these models have been scarce and largely limited to descriptive analyses such as k-mer enrichment at favored nucleosome positions [119, 120]. Such approaches highlight biases like AA/TT/TA periodicity or GC-rich preferences, but provide little mechanistic insight into how these features integrate with transcription factor binding or repeat-derived elements to shape chromatin organization. Moreover, most previous works did not systematically test whether the patterns learned by the networks correspond to experimentally validated regulatory determinants.

A particularly compelling advance is the design of DNA sequences for *in vivo* nucleosome positioning [121]. The authors combined deep mutational scanning with kinetic Monte Carlo optimization to design synthetic yeast DNA sequences that, when assembled in tandem arrays, produced nucleosomal arrays with NRL larger than the natural 165 bp. Importantly, these arrays were validated *in vivo*, demonstrating that computationally designed sequences can directly impose nucleosome phasing rules. This work illustrates the feasibility of functional chromatin design and identifies sequence rules with real consequences for nucleosome organization, at least in gene-dense genomes.

2.2.4.3 Language Models for Genomics

Inspired by advances in Natural Language Processing (NLP), recent years have seen the rise of transformer-based models in genomics. Originally introduced by Vaswani et al. (2017) [122], the transformer architecture relies entirely on attention mechanisms to capture long-range dependencies within sequential data. Unlike convolutional neural networks, which operate on fixed receptive fields, transformers can attend to all positions in a sequence simultaneously, enabling the modeling of complex, non-local interactions, a key feature of genomic regulation.

Genomic sequences, much like language, exhibit hierarchical and context-dependent patterns. This analogy has led to the development of language models for DNA, trained on large corpora of genomic sequences in a self-supervised manner. One of the earliest examples is DNABERT, which adapted the BERT architecture to k-mer tokenized DNA and demonstrated transfer learning capabilities across multiple genomic tasks [123]. More recently, large-scale pretraining efforts such as the Nucleotide Transformer [124] have scaled up these models to hundreds of millions of parameters, trained on entire genomes across species.

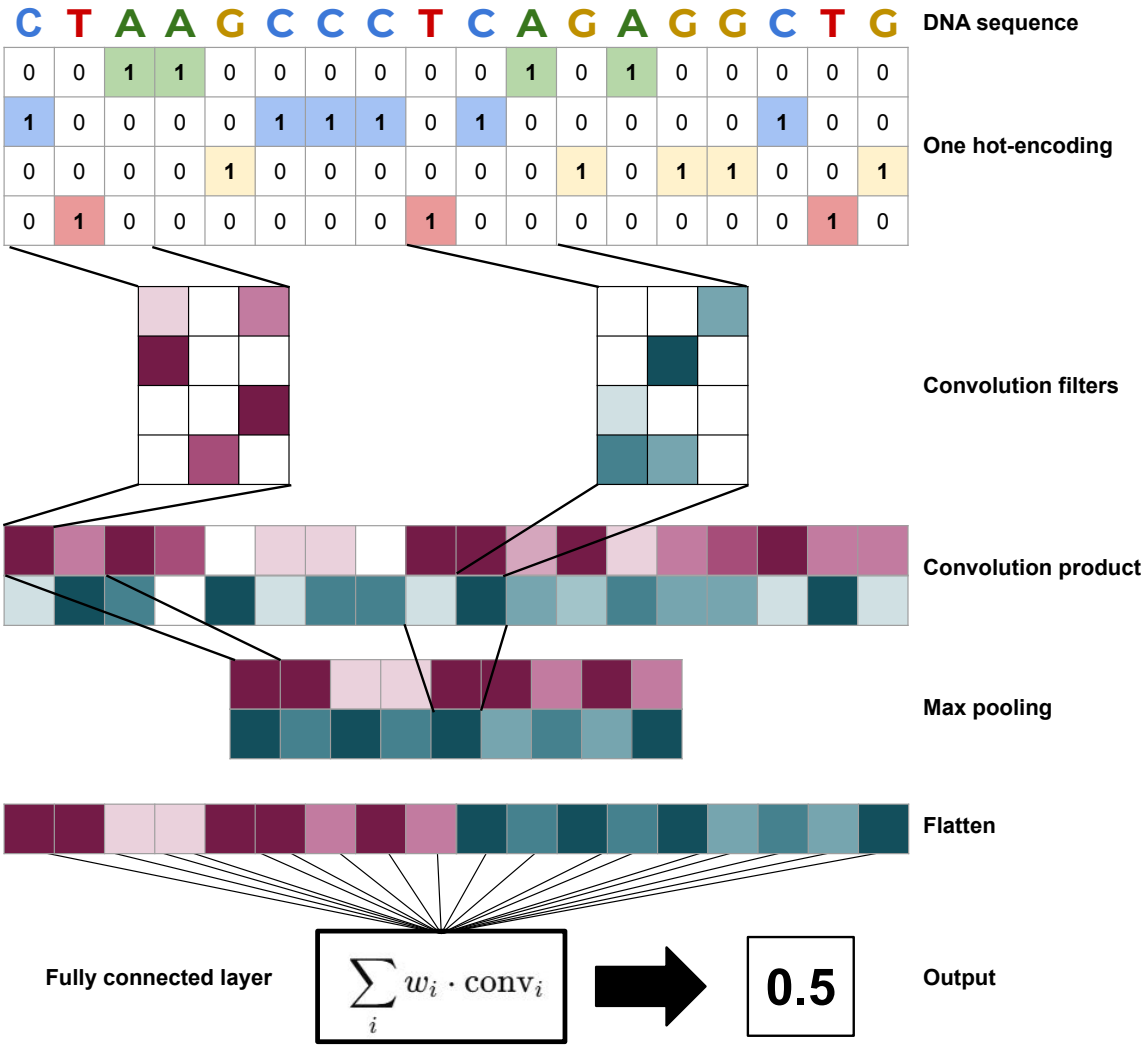


Figure 2.4: **Graphical overview of a convolutional neural network.** A DNA sequence is encoded as a binary matrix. Convolutional filters scan the sequence to detect local patterns, while pooling operations reduce dimensionality and retain the most relevant features. Multiple convolution–pooling layers allow the network to capture information at different scales. A final dense layer integrates the extracted features to produce a prediction.

In the context of nucleosome positioning, transformer-based models are still emerging but show great potential. A notable example is Borzoi, a nucleotide-resolution transformer pretrained on the human genome and fine-tuned to predict chromatin features such as nucleosome occupancy, DNA accessibility, and transcription factor binding [125]. Borzoi achieves high predictive accuracy while offering interpretable attention maps, making it a promising tool for deciphering the latent regulatory code of chromatin.

Beyond prediction, language models also serve as exploratory tools: their attention weights and learned embeddings can reveal biologically meaningful patterns, such as sequence periodicity, motif co-occurrence, or genome compartmentalization. As these models continue to grow in size and resolution, they are expected to play an increasingly central role in sequence-to-function modeling in genomics.

Despite their impressive performance, transformer-based language models come with notable limitations: they are computationally expensive to train and deploy, often requiring specialized hardware and large datasets, and their complex internal representations make biological interpretation challenging.

2.3 The determinants of nucleosome positioning

2.3.1 Sequence and chromatin: a dual influence

In their seminal 1988 paper, Kornberg and Stryer proposed that the DNA sequence alone may not possess sufficient intrinsic affinity for stable histone association, suggesting instead that regular nucleosome positioning might arise from boundary effects within chromatin rather than from sequence-dictated placement [90]. This early insight hinted at a more nuanced interplay between DNA and higher-order chromatin structure - an idea that would gain further traction as subsequent studies began to examine how both local sequence features and contextual cues shape nucleosome landscapes.

Early models of nucleosome positioning proposed that the DNA sequence alone encodes sufficient information to guide nucleosome placement. A landmark study by Segal et al. (2006) introduced the idea of a "genomic code" for nucleosome organization. They suggested that specific sequence motifs like AA/TA dinucleotides (with a ~10.5bp periodic enrichment) facilitate DNA bending and rotational positioning on the histone octamer. By combining *in vitro* selection experiments with probabilistic modeling, Segal and colleagues showed that around 50% of *in vivo* nucleosome could be pretricted from the sequence features [88]. However, this model was challenged and refined by Kaplan et al. (2009), who performed high-throughput *in vitro* nucleosome reconstitution on entire yeast and human genomes. Their data confirmed that sequence preferences do drive nucleosome positioning *in vitro*, and that models trained on such data can predict occupancy with high accuracy under controlled conditions. Yet, by comparing these *in vitro* profiles to *in vivo* nucleosome maps, Kaplan et al. demonstrated that many sequence-predicted nucleosomes are absent or displaced in living cells. This discrepancy highlighted the role of trans-acting factors, such as chromatin remodelers, transcription factors, and replication-associated processes, which can override or reprogram sequence-encoded preferences [69]. Extending this analysis to human cells, Valouev et al. (2011) mapped nucleosome positioning *in vivo* in multiple human cell types using MNase-seq and found that the influence of DNA sequence diminishes further in complex genomes. While certain DNA features (e.g., CpG content, poly(dA:dT) tracts) still contribute locally to nucleosome positioning, Valouev et al. showed that cell-type-specific chromatin landscapes dominate the observed patterns. Notably, they found a weak correlation between predicted *in vitro* occupancy and actual *in vivo* nucleosome maps in humans, emphasizing the dominant role of regulatory architecture, including transcriptional state, histone modifications, and boundary elements like CTCF [1]. This view is further supported by studies of the Widom 601 sequence, which was selected for its exceptionally strong nucleosome positioning ability *in vitro* [126] and has become a widely used model for nucleosome positioning studies. Despite its high affinity in biochemical assays *in vitro*, integration of large tandem repeats of the 601 sequence into the yeast genome failed to reproduce the expected nucleosome positioning pattern. This result highlights how the chromatin environment in vivo can

override even the strongest intrinsic sequence signals [127].

Together, these studies delineate a clear trajectory: from an early emphasis on intrinsic sequence-driven positioning toward a more nuanced view in which DNA sequence acts as a baseline framework, modulated or even overridden by cellular context and chromatin dynamics. The periodic dinucleotide patterns remain important for understanding nucleosome formation potential, but real genomic nucleosome organization reflects a complex interplay between sequence, chromatin remodelers, transcriptional activity, and higher-order genome architecture. Rather than being mutually exclusive, these mechanisms likely cooperate. For instance, a DNA region may intrinsically favor nucleosome formation, but the presence of a transcription factor could displace the nucleosome, effectively acting as a barrier. This concept gave rise and confirmed the Kornberg's insight of a barrier model, in which a nucleosome-depleted region created by regulatory proteins establishes a boundary that organizes surrounding nucleosomes into a phased, periodic array [2]. The most recurrent case of study is the TSS, which typically displays a Nucleosome Free Region (NFR) followed by a well-positioned +1 nucleosome. This arrangement is highly conserved across eukaryotes and has been consistently observed in genome-wide maps of nucleosome occupancy [2, 128, 129]. The NFR acts as a strong barrier, largely maintained by the binding of general transcription factors and chromatin remodelers, which prevents nucleosome assembly directly at the promoter. The +1 nucleosome downstream of the NFR serves as a reference point, from which statistical positioning and phasing propagate into the gene body. Thus, transcription start sites exemplify how sequence context, DNA-binding proteins, and remodeling activities cooperate to establish barrier-dependent nucleosome arrays. It is, however, noteworthy that despite the presence of a similarly regular nucleosome array upstream of the TSS, chemical mapping data reveal nucleosome occupancy within promoters [73], in contrast to the NFR typically described. While transcription start sites illustrate barrier formation through the combined action of DNA sequence context, transcription factors, and remodelers, other sequence elements may also generate barriers intrinsically. For instance, sequence elements like microsatellites may contribute to barrier formation themselves. Repetitive motifs such as GAA have been associated with nucleosome exclusion zones, as a repetitive element can affect locally the DNA stiffness [130], suggesting that certain sequence patterns both attract and repel nucleosomes depending on context.

Summing up these findings, a plausible explanation for the genome-wide dispersion of nucleosome patches is that the genome itself is punctuated by sequence elements that recruit or act as chromatin barriers. These include transcription factor binding motifs, transcription start sites that attract RNA polymerase, and microsatellite regions that may intrinsically disrupt nucleosome formation. From this view, the determinants of nucleosome positioning are effectively encoded within the DNA sequence, not as direct placement instructions, but as proxies that influence the recruitment of regulatory machinery and the establishment of chromatin boundaries.

2.3.2 Chromatin remodelers are actively driving nucleosome position

Chromatin remodelers are protein complexes that can shape chromatin. These enzymes are highly conserved in metazoans. Their control over nucleosomes is exerted both through histone post-translational modifications and nucleosome displacement. Their ATPase domain allows them to leverage sufficient energy to unwrap DNA from around the nucleosome, eject nucleosomes, or relocate them [131].

Although chromatin remodelers do not bind DNA in a sequence-specific manner, Rippe et al. (2007) showed *in vitro* that they can read DNA sequence features and position nucleosomes accordingly [132]. Other *in vivo* studies in yeast have shown that DNA sequence can indirectly influence the action of chromatin remodelers by serving as a binding platform for transcription factors that recruit them [133, 134].

Going further, by adding chromatin remodelers and a few Transcription Factor (TF), Oberbeckmann and colleagues (2024) showed that chromatin domains formed *in vitro* correlate with those observed *in vivo*. Despite some differences (such as additional NFR), nucleosome positioning also tends to be similar [135]. These remodelers are grouped into conserved families, each with distinct interaction partners and functional roles. Table 2.1 summarizes the main families, their partners, and associated functions. These factors are recognized determinants of nucleosome positioning, but since they are not completely independent of the DNA sequence, their effects on nucleosome positioning might still be captured from sequence alone.

Table 2.1: Function of chromatin remodelers, adapted from Tyagi et al. (2016)

Remodeler Family / Subtype	**Interacting Partner(s)**	**Function**
SWI/SNF (Switching Defective)	-	Transcriptional activation and repression
ISWI (Imitation Switch)	-	Nucleosome spacing, DNA damage repair, transcriptional repression
CHD (Chromodomain-Helicase-DNA binding)	-	ATPase activity, chromatin remodeling, HDAC activity
Subfamily 1 (Chd1/Chd2)	SSRP1, ACT-rich DNA	Nucleosome relocation; helicase activity
Subfamily 2 (Chd3/4/5)	H3K36, HDAC1/2, ATR, TRIM27, TRIM28, unmodified histones, H3K27me3	HDAC activity; ATP-dependent chromatin remodeling; epigenetic repression; neuronal chromatin regulation
Subfamily 3 (Chd6–9)	RNA Pol II, NRF2, NQO1, CTCF, Duplin, PPAR1a, CBFA1, osteocalcin, myosin	Transcriptional activation; redox homeostasis; developmental regulation; nuclear receptor activation; osteogenic differentiation

Remodeler Family / Subtype	Interacting Partner(s)	Function
INO80 Family	-	DNA helicase activity, replication and repair
INO80	YY1, Rvb1/2, NFRB, Arp4, Arp5, Arp8	Chromatin remodeling; DNA repair and replication
SWR1	Arp4, Arp6, Swc2, Rvb1/2	H2A.Z–H2B histone exchange

2.3.3 Pioneers factors engage nucleosome rearrangement

Pioneer TF are defined by their ability to bind closed chromatin and convert it to an open state. Upon engaging nucleosomal DNA, they recruit chromatin remodelers and/or non-pioneer TF, promoting nucleosome eviction or repositioning and thereby increasing accessibility. Pioneer factors are key drivers of cell-fate programming because they activate enhancers and promoters that are initially embedded in facultative heterochromatin [136]. In mouse, the pluripotency factors Sox2, Klf4, and Oct4 target silenced, nucleosome-enriched regions, recognize partially occluded Transcription Factor Binding Site (TFBS) wrapped on nucleosomes, and remodel local chromatin architecture [137, 138]. Similarly, NeuroD1 can trigger neuronal differentiation by binding epigenetically repressed neuronal genes [139]. In all cases, pioneer-factor binding reshapes chromatin topology and alters transcription by activating previously silent genes [138].

2.3.4 CTCF as a nucleosome positioning anchor

Among the many factors influencing nucleosome positioning, the transcription factor CTCF (CCCTC-binding factor) plays a particularly prominent role. Initially identified as an insulator protein [34], CTCF is now recognized as a central architectural regulator of the genome, coordinating both chromatin looping and local chromatin accessibility. One of its most striking features is its ability to generate phased nucleosome arrays flanking its binding sites [29, 140]. Unlike pioneer factors, which typically recruit remodeling complexes to rearrange nucleosomes, CTCF can directly bind to DNA even when it is wrapped around a nucleosome and independently induce nucleosome repositioning [141]. Moreover, CTCF has been shown to preserve the nucleosomal arrangement surrounding its sites throughout the cell cycle [140]. CTCF-induced phasing is not a passive consequence of binding, but rather depends on an active interplay with cohesin and other chromatin-associated factors. The presence of additional transcription factor binding motifs nearby modulates the extent of nucleosome repositioning, showing that sequence context (both within and around the CTCF motif) acts as a combinatorial signal for local chromatin remodeling [141, 142]. Furthermore, the directionality of the motif, driven by its asymmetry, plays a critical role in the downstream organization of nucleosomes and in the formation of chromatin loops. The asymmetry leads to polarized phasing, with

stronger periodicity observed downstream of the motif, consistent with its orientation-dependent function in loop extrusion and domain insulation [34, 143, 144]. These findings support a model in which CTCF binding sites encode chromatin architecture through their sequence, not only by recruiting architectural proteins but also by directly imposing a nucleosomal landscape. Thus, CTCF exemplifies how DNA sequence alone can instruct precise and reproducible chromatin patterns. Concerning CTCF, its motif is described as a triptych of an upper motif, a 15bp core motif and a downstream motif. While binding can occur efficiently through the core motif alone, primarily engaging Zinc Finger (ZF) 3-7, the upstream and downstream segments-interacting with ZF 9-11 and 1 respectively contribute significantly to the overall binding stability and specificity [145]. Recent work has shown that these submotifs are not equivalent in their contribution to binding affinity and chromatin organization. The core motif is essential for CTCF recognition and DNA binding, while the upstream and downstream flanking sequences modulate the strength, orientation, and functional output of the site. Do et al. demonstrated that variations within the flanking motifs can alter nucleosome phasing patterns, even when the core binding is maintained [146]. This suggests that the extended CTCF motif functions as a modular signal, with different segments encoding different layers of regulatory information - from binding stability to chromatin remodeling potential. This intricate encoding highlights the dual nature of CTCF sites as both sequence-specific transcription factor binding sites and architectural nucleosome-positioning elements.

2.4 Stakes and purposes of my study

Upscaling the study done by Routhier et al. from yeast to mouse presents significant challenges. Yeast have compact genomes, with narrow intergenic regions and a dense organization of genes and regulatory elements, often arranged consecutively. In contrast, mammalian genomes are much larger, and genes account for only about 3% of their total length. The key question, therefore, was whether neural networks could maintain their predictive power in such a sparse genomic landscape.

CNN have to ability to capture and exctract information from huge, noisy signals. Previous studies provide insight of certain genomics features that locally position nucleosome on the genome [115–118], achieving accurate predictions across species. Yet, little is known about the biological features these models can reveal. Most interpretability efforts have focused on k-mer enrichment at favored nucleosome positions [119, 120], providing limited insight into the regulatory logic encoded in the sequence. Using neural network aims to generate an exhaustive and comprehensive map of Nucleosome Positioning Region (NPR). On top of that, it constitutes a sandbox to generate synthetic sequences or study the effect of *in silico* mutated sequences. Given the complex interplay between DNA sequence, chromatin folding, and genome regulation, we ask whether a neural network could achieve the *tour de force* of reconstructing a meaningful portrait of regulatory logic

from nucleosome positioning alone.
We hypothesize that CNNs, when properly trained, can extract relevant regulatory signals even in challenging genomic contexts such as low-mappability regions, repetitive elements, or gene-sparse loci. Furthermore, we posit that once such models produce accurate and biologically consistent predictions, they can be leveraged to explore *in silico* synthetic sequences. This opens the possibility to simulate and interrogate nucleosome behavior in controlled sequence contexts, offering a new framework for studying the sequence determinants of chromatin organization.

Given these considerations, our study addresses three main questions:

1. Can nucleosome positioning be accurately predicted from DNA sequence alone in the large, gene-sparse genome of a mammal, as previously shown in the compact genome of yeast?

2. If so, can the neural network be interpreted to reveal biologically relevant determinants of nucleosome positioning, beyond simple k-mer enrichment?

3. What is the role of repetitive elements—known to carry regulatory potential—in shaping nucleosome organization, and can functional motifs embedded within them be detected by the model?

To address these questions, the next chapter presents a comprehensive analysis of model performance, interpretability, and biological insight, leveraging both *In Silico* Mutagenesis (ISM) and motif discovery. We aimed to go further than simple k-mer analysis by leveraging synthetic genomics to explore the impact of controlled sequence modifications, thereby testing specific hypotheses about the sequence determinants of nucleosome positioning.

Chapter 3

Overview of datasets and modelling framework

In next chapters, we ask whether a CNN can predict *in vivo* nucleosome positioning from DNA sequence alone, and what this reveals about the underlying sequence determinants. For clarity, we briefly summarize here the datasets, models and interpretation strategy; full methodological details are provided in Appendices A–B.

3.1 Experimental nucleosome maps

We used two independent genome-wide nucleosome maps in mouse embryonic stem cells (mESC). First, MNase-seq data were processed from paired-end fragment midpoints to obtain a continuous nucleosome occupancy signal along the genome. Only the uniquely mapping reads were keep to conserve the signal. The raw coverage was smoothed by Gaussian convolution, clipped at the 99th percentile to remove extreme values, and normalized to the $[0, 1]$ range. Second, chemical-cleavage data were lifted over the mm10 assembly and converted into nucleosome occupancy by aggregating nucleosome core particle scores with a center-weighted Gaussian scheme. These two assays provide complementary *in vivo* measurements that differ in biochemical principle and serve as independent targets for sequence-based prediction.

3.2 Sequence-based neural network models

All predictions are made from 2 kb one-hot encoded genomic windows. The models output nucleosome occupancy at five positions distributed across the central part of the window, which provides a local profile while stabilizing the interpretation of the network (multi-head architecture). Two CNN architectures were implemented: (i) a lightweight model used for MNase-seq data and (ii) a higher-capacity variant optimized for chemical-cleavage profiles. Both share the same overall structure: several convolutional layers with non-linear activation and pooling, followed by a small dense layer and a 5-unit sigmoid

output layer.

Training is performed by minimizing a custom loss that combines mean absolute error and the complement of the Pearson correlation coefficient between predicted and experimental nucleosome occupancy. This objective encourages the network to reproduce both the amplitude and the shape of the experimental profiles. To mitigate the strong imbalance in target values (most genomic positions having low occupancy), we apply a bin-wise reweighting of the loss so that all occupancy ranges in $[0, 1]$ contribute more uniformly to learning.

3.3 Interpretation strategy

To move beyond predictive performance and extract sequence determinants of nucleosome positioning, we use ISM as our primary interpretation tool. For each input sequence, all possible single-nucleotide substitutions are introduced in silico, and the resulting changes in predicted occupancy are aggregated into a mutascore per genomic position. High mutascore loci define compact regions where the model is most sensitive to sequence variation.

We then identify Nucleosome Positioning Regions (NPRs) as short genomic intervals where ISM signals from the MNase-seq and chemical-cleavage models are jointly elevated. These NPRs seqlets serve as the basis for downstream analyses: motif discovery and enrichment using the MEME suite (XSTREME and SEA) and enrichment tests with RepeatMasker annotations to quantify the contribution of repetitive-element families. Additional details on ISM, NPRs, motif analyses and repeat enrichment procedures are provided in Appendix B.

Chapter 4

Neural networks can predict *in vivo* nucleosome positioning

4.1 Nucleosome are positioned in island along the genome.

To investigate nucleosome organization, assays such as MNase-seq and chemical cleavage aim to sequence the DNA fragments protected by histones. As these experiments are typically performed on a population of thousands to millions of cells (bulk experiment), the resulting signal reflects an average nucleosome occupancy landscape along the genome. Regions consistently occupied by nucleosomes across most cells appear as well-defined peaks. When nucleosomes are regularly spaced and aligned in phase across the population, a periodic pattern emerges in the signal: these are referred to as phased nucleosomes. In contrast, regions where nucleosome positions vary from cell to cell result in broader or attenuated signals, indicative of fuzzy nucleosomes whose positions are not strictly conserved as illustrated in Figure 4.1.

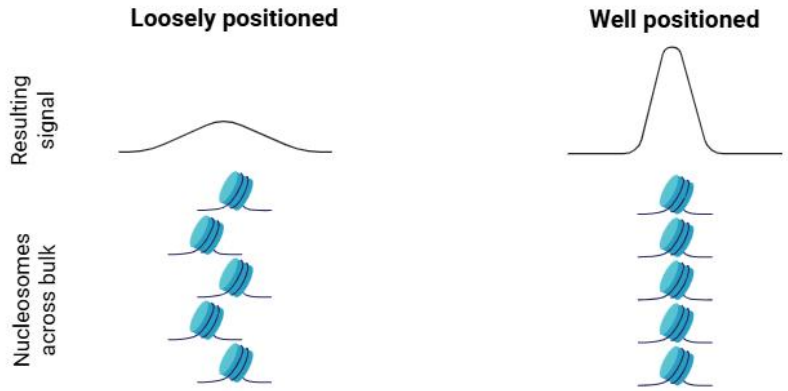


Figure 4.1: **Schematic representation of nucleosome fragment aggregation in bulk-cell data.** Created with biorender

In the same vein, arrays of regularly positioned nucleosomes won't clearly appear in the signal, depending whether or not these arrays are phased in most cells in the bulk. The

resulting signal will vary from stochastic noise to a flat line with no defined nucleosomes as schematized in Figure 4.2.

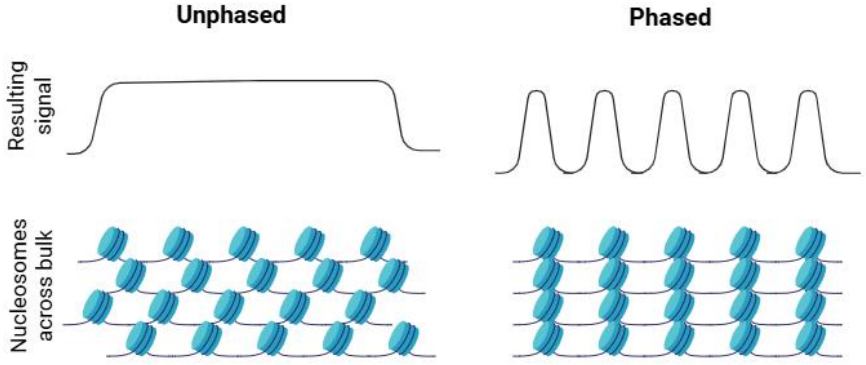


Figure 4.2: **Schematic representation of nucleosome phased and unphased arrays in bulk-cell data.** Created with biorender

The implication of this retrieved average nucleosome landscape is that the determinants we can identify are necessarily those recurrent across the majority of cells in the bulk population. In contrast, biological processes such as transcription or replication often occur heterogeneously in an asynchronous bulk, meaning that cell-to-cell variability in nucleosome positioning linked to these processes is likely to be masked [129].

Although the concept of NRL remains debated at the single-cell level [65], bulk experiments allow estimating an average nucleosome spacing across the genome. To this end, we applied an autocorrelation analysis on a 10-kb sliding window to recover the distance between neighboring nucleosomes. While this approach differs from previous methods [1], our estimates are consistent with other reports in mouse [29]. We obtained an average NRL of 187bp, corresponding to 147bp of wrapped DNA plus a 40bp linker region (Figure 4.3).

The characteristic nucleosome pattern can be detected using a wavelet transform, which is designed to identify local frequencies in a signal (see Appendix B), in contrast to Fourrier transform that will completely loose the space resolution. When examining nucleosome positioning across the genome (Figure 4.4), we observe that nucleosomes are broadly distributed, but regularly phased arrays occur only in restricted regions. Using wavelet decomposition, we can detect periodic signals in the MNase-seq coverage at multiple scales, which allows us to distinguish between background nucleosome occupancy and genuine phased domains. We refer to these regions of strong periodicity as *nucleosome islands*, reflecting local patches of ordered arrays embedded within a largely irregular nucleosome landscape.

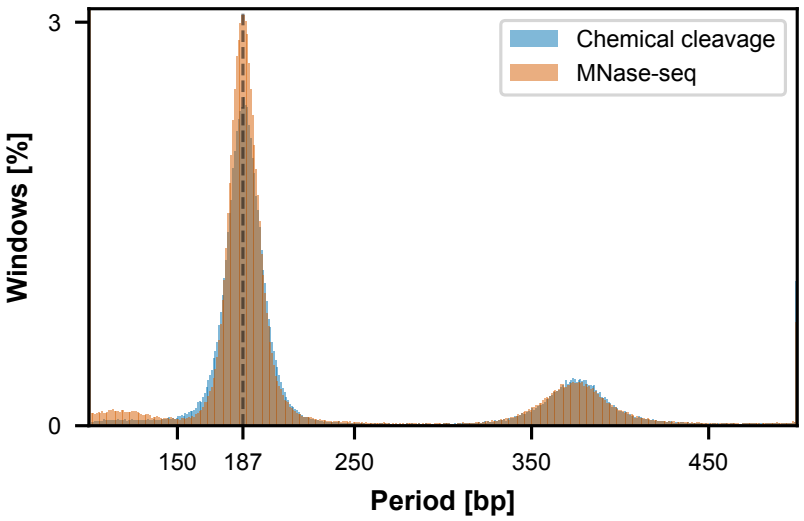


Figure 4.3: **Autocorrelation of experimental signals computed on sliding windows.** Histograms of window-based autocorrelation values for MNase-seq (orange) and chemical cleavage (blue) data reveal a dominant periodic component at approximately 187 bp.

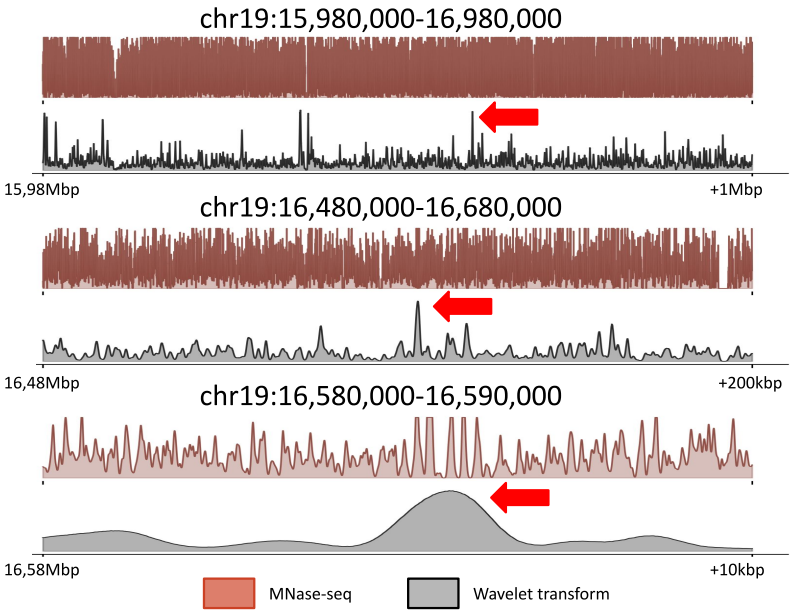


Figure 4.4: **Representative MNase-seq signal and its wavelet transform at different scales.** Wavelet peaks coincide with phased nucleosome arrays.

Figure 4.5 shows the distribution of nucleosome island sizes, as estimated from the width of the wavelet transform peaks. Genome-wide, nucleosome islands contain up to 15 phased-nucleosomes, with an average of about 5 nucleosomes per array (based on MNase-seq data).

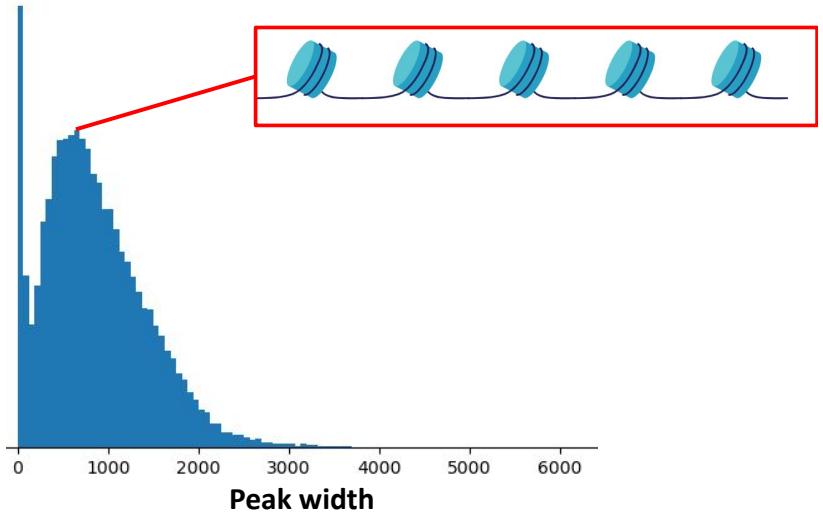


Figure 4.5: **Wavelet peaks width distribution.** A mean width of ∼1 kb indicates arrays of roughly five phased nucleosomes.

These results suggest that nucleosomes are organized in well-phased arrays, but only around specific genomic hot-spots. The genome would be therefore punctuated by sequence elements that strongly position nucleosomes in their vicinity.

We asked whether a CNN, given only the DNA sequence, could learn the determinants underlying both the genome-wide nucleosome landscape and the phased arrays that emerge locally. If successful, such a model would not only capture the basal rules of nucleosome positioning but also highlight strong sequence features acting as focal points for chromatin organization.

4.2 Overall performances

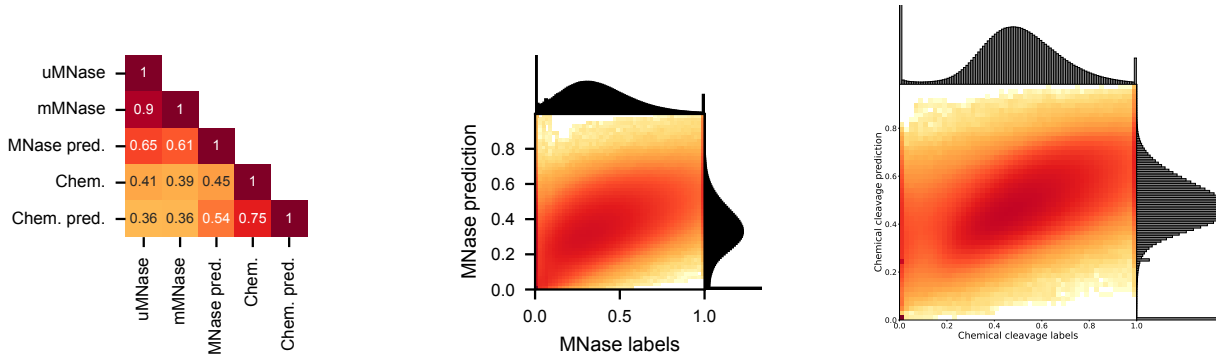
What is a good prediction ?

Assessing the performance of a machine learning model in biology is not straightforward. Most experimental data suffer from biases or limitations, either due to the technology used or the intrinsic variability of biological systems. As such, a prediction that perfectly fits experimental data may simply replicate those limitations.
Moreover, multiple criteria can be considered when comparing biological signals. In the case of nucleosomes, the most relevant aspect is the predicted position of nucleosomes. Pearson correlation is often used to assess the similarity in shape between two signals, allowing for the comparison of predicted and experimental occupancy patterns-focusing on relative positioning, rather than the absolute values of the signal.
Beyond purely quantitative agreement, a truly meaningful prediction must also capture biologically relevant features. In other words, it should not only fit the data; it should make sense in light of established biological knowledge. For example, one of our model trained on Hi-C data (not shown here) achieved excellent global correlation and signal similarity, yet completely ignored charachteristic nucleosome phasing around CTCF binding sites or TSS, which are well-known nucleosome organizers. Such a model, despite its metrics, fails to capture an essential aspect of chromatin structure.

To assess the performance of our network, we computed the Pearson correlation between the experimental and predicted nucleosome positioning profiles. Chromosome 19, which was excluded from the training and validation process, served as the test set, ensuring that the Neural Network (NN) had not encountered any of its sequences during learning. The resulting correlations are shown in Figure 4.6a.

As expected, the two MNase-seq mapping strategies (Box 2.1.2) yield very similar predictions. Our CNN models exhibit strong predictive performance, achieving Pearson correlations of 0.75 and 0.65 for chemical cleavage and MNase-seq data, respectively. Surprisingly, the predicted profiles are more strongly correlated with each other than the corresponding experimental datasets, suggesting that the two experimental signals may share a common underlying informational structure captured by the model.

Our model uses an input sequence of 2,001 base pairs (see Appendix A), which defines the maximum span of information it can exploit to predict nucleosome occupancy. Despite this local window constraint, the model achieves excellent predictive performance, with Pearson correlations reaching up to 0.75 on held-out chromosomes. This demonstrates that a substantial portion of nucleosome positioning signals is encoded in the



(a) Pearson correlation matrix (b) MNase correlation histogram (c) Chemical cleavage correlation histogram

Figure 4.6: **Pearson correlation between experimental and predicted nucleosome profiles.** (a) Pairwise correlations between all signals. (b) Histogram of correlation between predicted and experimental MNase-seq signals across test windows. (c) Same as (b), but for chemical cleavage data. uMNase: uniquely mapped MNase; mMNase: multimapped MNase; Chem.: chemical cleavage; pred.: prediction.

immediate sequence context. Nevertheless, such a finite receptive field inherently limits the model's ability to integrate broader genomic cues, such as distal regulatory elements, long-range chromatin interactions, or positioning signals emerging from higher-order nucleosome phasing. These long-range dependencies may be critical in specific contexts, particularly around insulators, enhancers, or within TADs. Future efforts could address this by incorporating architectures capable of capturing extended dependencies, such as dilated convolutions, recurrent layers, or transformer-based models. This would allow the model to account for both local sequence determinants and global chromatin context, potentially leading to even more biologically faithful predictions.

Figure 4.7: **Nucleosome occupancy around CTCF sites** Aggregated plot of mnase-seq (left) and chemical clevage (right) on 63506 CTCF sites genome wide, sites were retrieved by JASPAR with mapping quality over 900. blue : prediction, orange : experimental data

In addition to achieving high predictive performance, our model captures fine-grained biological features of nucleosome organization, which can be visualized through aggregated

nucleosome occupancy profiles centered on CTCF binding sites (Figure 4.7). These plots, based on predictions from held-out test regions, display the canonical phased nucleosome arrays flanking CTCF. Strikingly, the model recapitulates this pattern for both MNase-seq and chemical cleavage datasets, even though no CTCF annotation was used during training. This suggests that the model has learned to recognize intrinsic sequence features that encode chromatin architecture around such regulatory elements. Furthermore, our predicted profiles mirror subtle experimental differences between the two datasets. In particular, predictions trained on the chemical cleavage signal exhibit a central dip in nucleosome occupancy, corresponding to the so-called fragile nucleosome directly over the CTCF motif. This feature, well described by Voong et al. (2017), is underrepresented or absent in MNase-seq data due to the MNase enzyme's preferential digestion of loosely bound or partially unwrapped nucleosomes [73]. Its presence in the chemical cleavage-based predictions and absence in the MNase-based ones demonstrates that the model has internalized not just the sequence rules, but also the technical biases and sensitivities of the input data. This ability to reproduce both shared and method-specific chromatin features further supports the biological and experimental interpretability of the model's predictions. It also validates the idea that deep learning can serve as a proxy for chromatin profiling, while enabling new in silico experiments. Building on these results, we next asked whether the model could generalize to more challenging regions of the genome; in particular, those typically excluded from experimental analyses due to low mappability.

4.3 Trained model accurately predicts the nucleosome density over non-mappable regions.

As we used the uniquely-mapped signal (see Box 2.1.2) to train MNase-seq model, we excluded repeated elements that are were sufficiently covered (see Appendix A). At the genome-wide scale, the MNase-seq model achieves a Pearson correlation of 0.65 with the experimental data. To assess the model's performance in poorly mappable regions, we compared its predictions to the alternative multi-mapped signal: we selected genomic segments with length between 50 and 1000bp, with a mappability strictly lower than 1. Regions showing high agreement between the uniquely-mapped and multi-mapped signals were discarded, yielding approximately 200,000 regions for analysis.

Figure 4.8 shows the correlation between the model's predictions and both types of mapping strategies. We observe that, in these low-mappability regions, around 60% of the correlations are higher with multi-mapped signal, suggesting that the model trained on uniquely-mapped regions is able to generalize from its training on mappable regions and infer the nucleosome occupancy on the poorly covered regions.

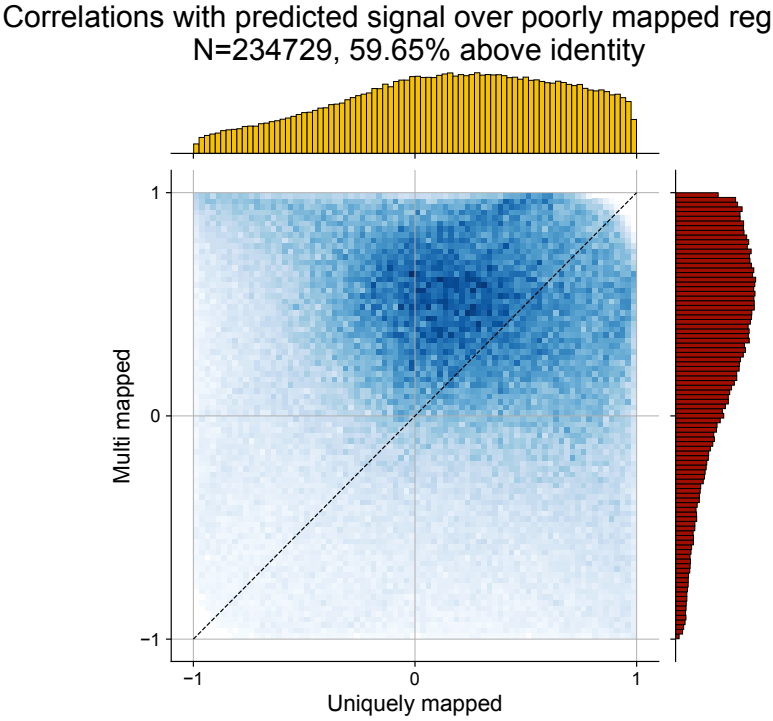


Figure 4.8: **Comparison of prediction–experiment correlations under different MNase-seq mapping strategies in low-mappability regions.** Each point represents a poorly mappable genomic window, with the Pearson correlation between the predicted MNase-seq profile and the uniquely mapped experimental signal shown on the x-axis, and the correlation with the multimapped signal on the y-axis. The diagonal denotes identity. A majority of windows (∼60%) fall above identity, indicating higher agreement between model predictions and the multimapped representation of MNase-seq in these regions.

To illustrate this, Figure 4.9 shows an example of a poorly mappable region. While the uniquely mapped signal is extremely sparse, the multi-mapped signal reveals a clear nucleosomal pattern. The model prediction aligns with the multi-mapped signal, demonstrating that the network captures meaningful features even in regions traditionally considered unreliable.

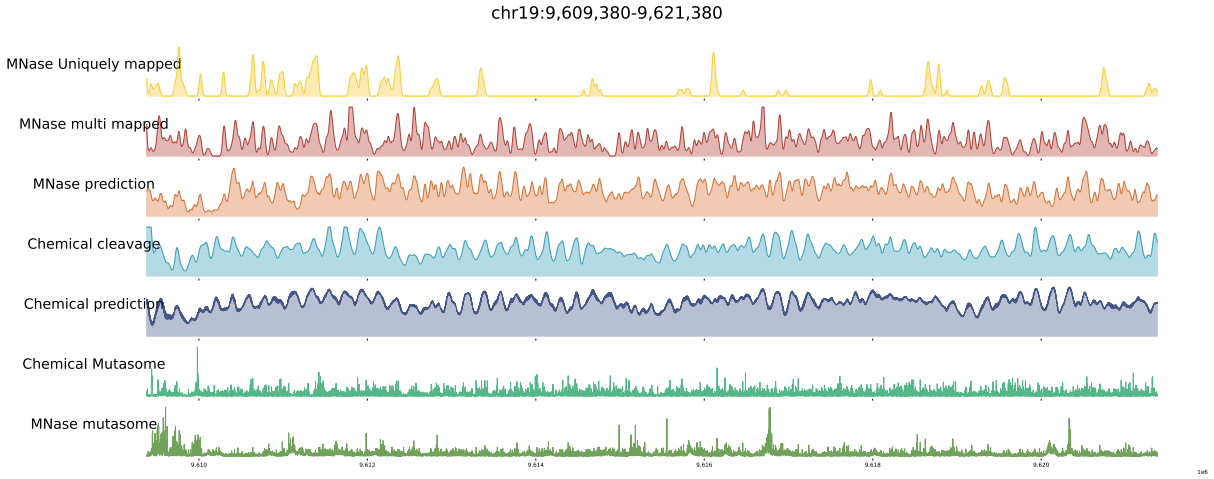


Figure 4.9: **Nucleosome occupancy on poorly mappable region.** yellow:uniquely mapped mnase-seq data, red: multimapped mnase-seq data, orange: predictions of mnase-seq model.

4.4 Internal validation and relation to previous work

Direct comparison with other nucleosome positioning models was not performed here for two main reasons. First, most existing approaches are classification-based (linker vs nucleosome) or tailored to *in vitro* short-reads datasets, whereas our task involves predicting continuous occupancy profiles *in vivo* from sequence alone. Second, although our methodology closely follows the interpretable deep learning framework introduced by Routhier *et al.* for the compact yeast genome, the substantial genomic differences between yeast and mouse (particularly in genome size, repeat content, and mappability) mean that a direct benchmark is not strictly equivalent.

Nevertheless, a qualitative comparison with Routhier's results is informative. In their study, correlations between predicted and experimental nucleosome profiles reached 0.68 on the yeast genome [8], which is gene-dense and largely well-mappable. In our case, we obtain correlations up to 0.75 (chemical cleavage) and 0.65 (MNase-seq) on the much larger and more repetitive mouse genome, where a substantial fraction of nucleosomes lie in regions of low mappability. Achieving performance in this range, despite the greater genomic complexity, suggests that the model retains predictive power in the face of genomic complexity. The robustness of our model can be evaluated internally through several complementary lines of evidence.

Further internal validation comes from multiple complementary lines of evidence. First, high correlations obtained on a completely held-out chromosome (19) indicate that

the network generalizes beyond sequences seen during training. Second, the convergence of predictions between MNase-seq and chemical cleavage data, while still reflecting their specific differences (for example, the presence or absence of the fragile nucleosome at the center of CTCF sites), suggests that the network captures genuine biological determinants while also incorporating the technical biases specific to each method. Finally, the ability to predict relevant profiles in low-mappability regions—where experimental signal is notoriously degraded—supports the use of the model as a tool for "completion" or exploration of genomic regions traditionally ignored.

Altogether, these observations confirm that the information contained within the local 2kbp sequence context is sufficient to reconstruct biologically interpretable nucleosome positioning signals, paving the way for a systematic analysis of NPRs discussed in the following chapter.

Chapter summary

- CNN with sequence input alone can reproduce *in vivo* nucleosome occupancy profiles with high accuracy in large genomes.
- Comparing to the training experimental MNase-seq and chemical cleavage data, the models achieved correlations up to 0.75 on held-out genomic regions and captured both global occupancy patterns and fine-scale features such as phased arrays around CTCF sites.
- The predictions generalize beyond mappable regions, providing meaningful nucleosome profiles even in repeat-rich domains where experimental signals are degraded.

Chapter 5

Opening the black box: *In Silico Mutagenesis* extracts nucleosome positioning rules

The previous section established that our convolutional neural network achieves accurate predictions of nucleosome positioning from DNA sequence alone, even under challenging conditions such as low-mappability regions. We now turn to the question of *why the model makes these predictions and specifically, which sequence features it relies upon ?*.

Neural networks are often criticized for their "black box" nature: while inputs and outputs are accessible, the internal mechanisms driving predictions remain largely opaque. Several strategies have been developed to address this limitation, broadly divided into two families: perturbation-based methods and gradient-based methods.

Gradient-based approaches, such as saliency maps, Integrated Gradients, or DeepLIFT, estimate the importance of each input feature by computing the gradient of the output with respect to the input. They are computationally efficient but rely on backpropagated signals through a highly non-linear model, which can make them less intuitive to interpret in a biological context. Perturbation-based approaches, by contrast, systematically modify parts of the input—such as masking or mutating specific bases—and observe how the output changes. This makes them model-agnostic and conceptually straightforward, at the cost of higher computational demands.

In this work, we adopt ISM a perturbation-based approach in which each nucleotide in a sequence is mutated in turn, and the resulting change in the model's prediction is recorded. The principle of this method illustrated in Figure 5.1 is to measure the difference in prediction when a mutation occurs in any sequence: the network prediction will noticeably change only if the locus where the mutation occurred matter for the neural network to accurately predict the nucleosome positioning in its vicinity.

ISM directly links the importance of each base to an observable change in prediction, making it easier to relate the network's behaviour to underlying sequence features. Although more computationally demanding than gradient-based alternatives, its conceptual

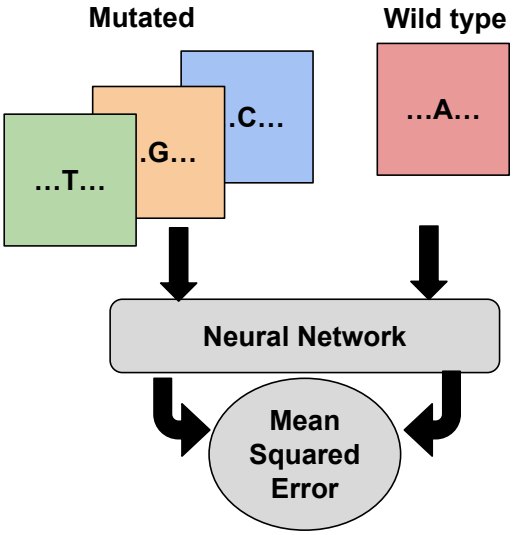


Figure 5.1: **Graphical abstract of In Silico Mutagenesis**. Sequences are muted in silico in order to observe change in the CNN prediction

simplicity and minimal assumptions make it well suited to our goal: extracting biologically interpretable rules of nucleosome positioning from the model.

To illustrate the output of ISM in a genomic context, we selected a representative locus displaying phased nucleosome arrays. Figure 5.2 shows, for both MNase-seq and chemical cleavage models, the experimental signal, the corresponding model predictions, and the ISM score track.ISM yields a generally flat signal, punctuated by sharp peaks that cluster in small genomic regions. In this example, ISM high-scoring regions align with phased nucleosomes in the experimental data, highlighting that the model's attributions coincide with well-established chromatin organization patterns. For completeness, we also computed a saliency map track for the same locus, which yields a qualitatively similar pattern to ISM. Although we do not analyse saliency maps further in this work, their concordance with ISM supports the robustness of the attribution signal.

To assess robustness of the method, we compared ISM profiles from the two independently trained networks (MNase-seq vs. chemical cleavage). Although the raw experimental signals are only moderately correlated (Figure 4.6a), their ISM profiles show high agreement (correlation of 0.59), indicating that both networks share similar sequence determinants of nucleosome positioning (Figure 5.3).

We further validated ISM against orthogonal data: DNase-seq and ATAC-seq (chromatin accessibility assays) and a wavelet-based computational analysis. As shown in Figure 5.4, ISM peaks show strong concordance with these orthogonal measures of chromatin structure. Differences reflect assay properties: wavelets produce broader regions; DNase-seq highlights the most accessible sites (often the highest ISM peaks); ATAC-seq displays wider signal around the ISM peaks. ISM consistently pinpoints compact, high-

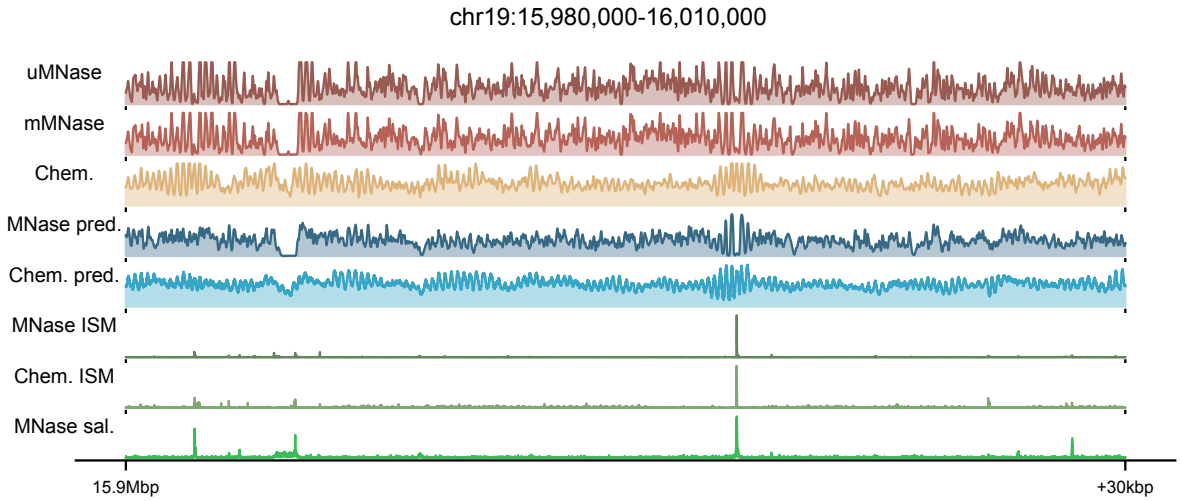


Figure 5.2: **Overview of nucleosome occupancy signals and network interpretation.** Genome browser view showing experimental signals, model predictions, and ISM signal, whose peaks coincide with phased nucleosome arrays in the experimental data. sal.: saliency

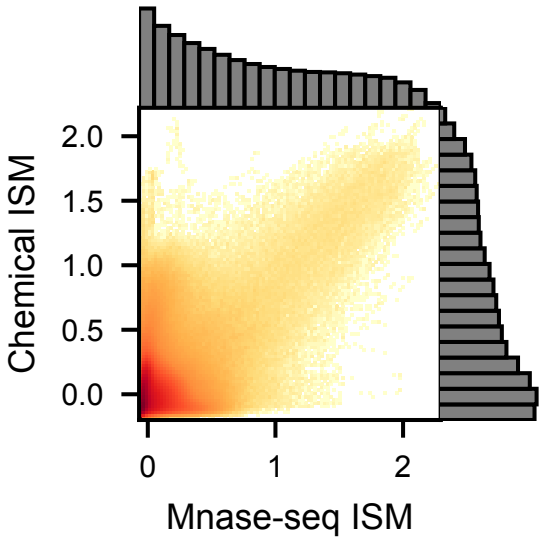


Figure 5.3: **Correlation of ISM scores** between the chemical-cleavage–trained model (y-axis) and the MNase-seq–trained model (x-axis) PCC =0.59. Signals are z-scored and axis log10-scaled.

specificity sites with minimal noise.

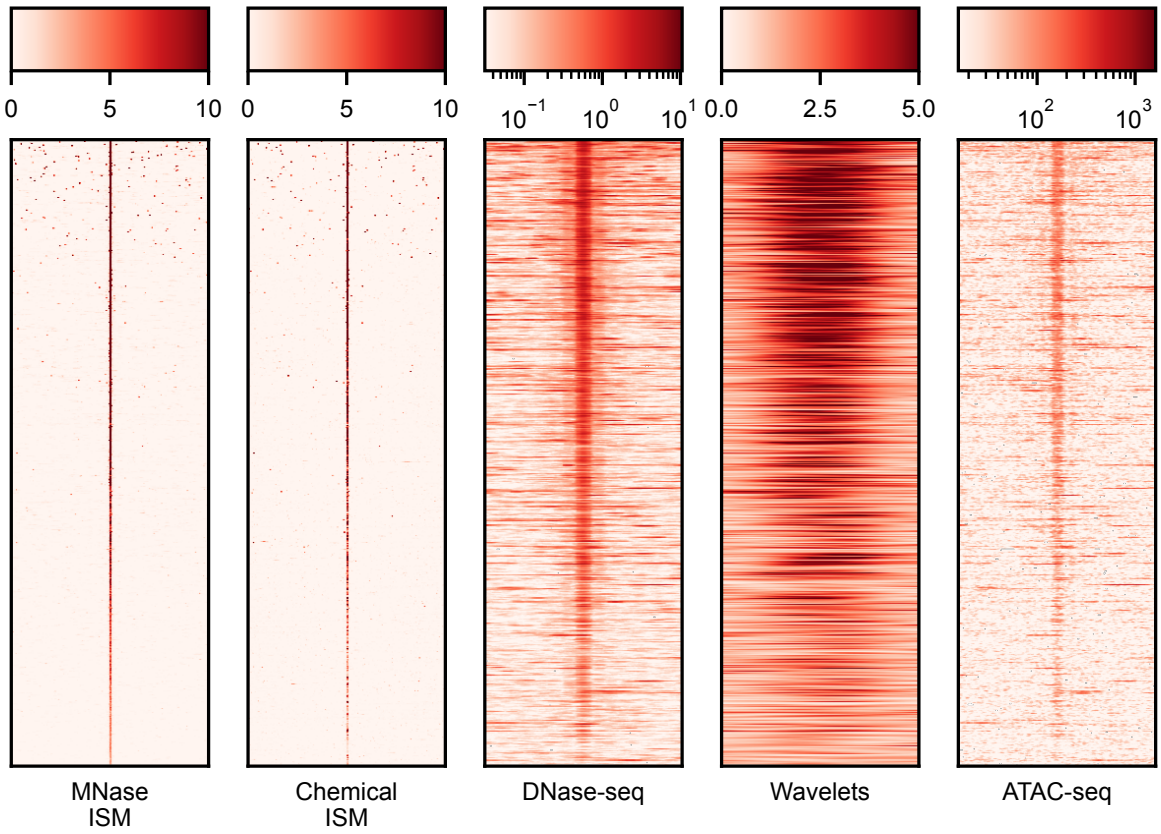


Figure 5.4: **Concordance of ISM with orthogonal datasets on chromosme 19. N = 1614** Overlap between ISM peaks, DNase-seq, ATAC-seq and wavelet analysis. The sequences are sorted in descending order of MNase-seq ISM

Zooming in on a single locus illustrates the resolution advantage of ISM: it highlights sequence determinants on the scale of TFBS (around 20bp), whereas other signals extend over hundreds to thousands of base pairs (Figure 5.5).

Together, these results indicate that ISM flags *bona fide* biological determinants of nucleosome organization rather than assay-specific artifacts. We next examine the sequence content of these Nucleosome Positioning Regions (NPR) and the classes of elements that drive them.

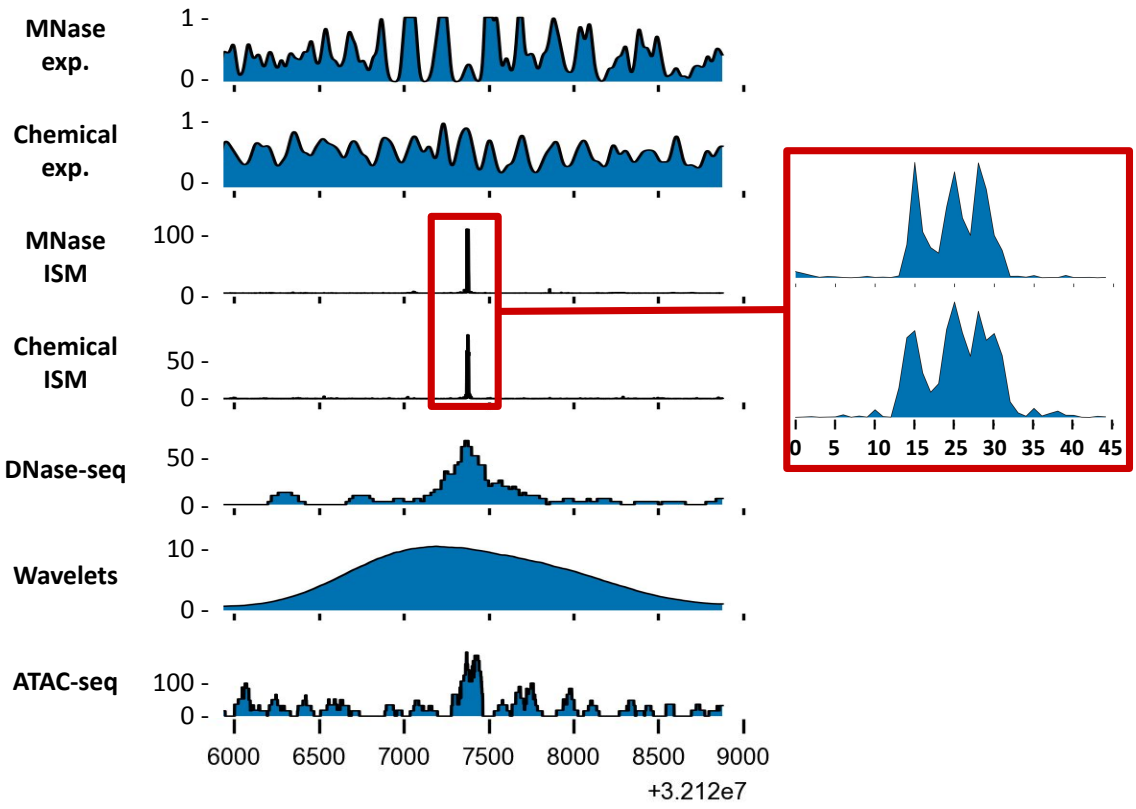


Figure 5.5: **Comparaison of ISM with orthogonal signals and focus on ISM peak.** Signal at coordinates chr19:32,125,877-32,128,876

5.1 In silico Mutagenesis highlights Nucleosome Positioning Regions

Having established that ISM highlights compact, biologically relevant regions (Nucleosome Positioning Regions) that correlates with nucleosome islands, we next examined whether these regions are enriched for motifs that could point to sequence-specific factors that directly or indirectly shape nucleosome organization. The methodology used to call NPRs is described in Appendix B and yielded a total of 613,545 regions of 20bp, which were concordant between MNase-seq and chemical cleavage ISM signals.

Their distribution along the genome is shown in Figure 5.6. The average distance is 3,916 bp, which does not differ from the expected average distance obtained by an equidistant positioning of the same number of NPRs genome-wide. However, the cumulative distribution of inter-NPR distances reveals distinct regimes. First, ∼9.4% of intervals fall below 20bp, reflecting overlapping sequences introduced by the procedure. Second, only a small fraction of distances occur between 20 and 500 bp, as indicated by the shallow slope of the curve, suggesting that closely spaced NPRs are relatively rare. Third, the majority of distances accumulate between 500 bp and 10kb, where the curve rises

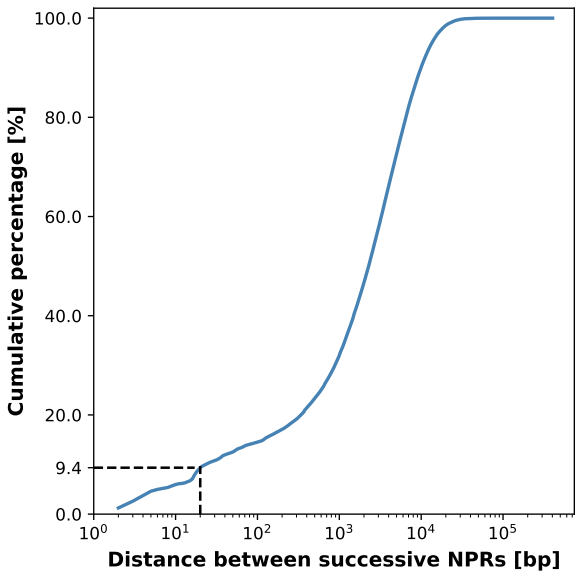


Figure 5.6: **Cumulative distribution of NPR distances.** About 9.4% of intervals are below 20 bp, reflecting overlapping extracted sequences of NPR. Distances between 500 bp and 10 kb account for the majority of cases, while very large gaps (>10 kb) are rare, leading to the plateau.

steeply, defining the typical range of inter-NPR spacing. Beyond 10kb, the distribution reaches a plateau, indicating that very large gaps between NPRs are infrequent and likely correspond to extended regions depleted of positioned nucleosomes.

To further investigate the dispersion of NPRs, we projected experimental and predicted nucleosome occupancy signals onto compartments defined from Micro-C data. Compartments were called using the first principal component of a normalized contact matrix [147], which partitions the genome into two chromatin states. The orientation of PC1 was subsequently assigned by correlation with transcriptional activity, defining A- and B-like compartments (Figure 5.7): RNA-seq coverage clearly distinguishes active regions consistent with A compartment, from transcriptionally silent regions consistent with the B compartments. Experimental nucleosome occupancy profiles show distinct behaviors: MNase-seq exhibits strong differences in average signal between compartments with much lower value in the B compartment, whereas the chemical cleavage signal is more uniform but tends to be higher in B compartments, suggesting opposite biases between the two assays. Predicted signals correlate with their respective experimental data but display smoother profiles across compartments. Finally, NPRs are detected genome-wide with pronounced local enrichments and moderate increase in average value in A compartments.

At the level of higher-order chromatin, it is noteworthy that both prediction models show little global variation between compartments, resembling the experimental signal obtained by chemical cleavage. Nevertheless, NPRs are not homogeneously distributed across the genome: although present in both compartments, they are enriched in A, the transcriptionally active compartment.

To further investigate the link with transcription, we assessed model predictions around

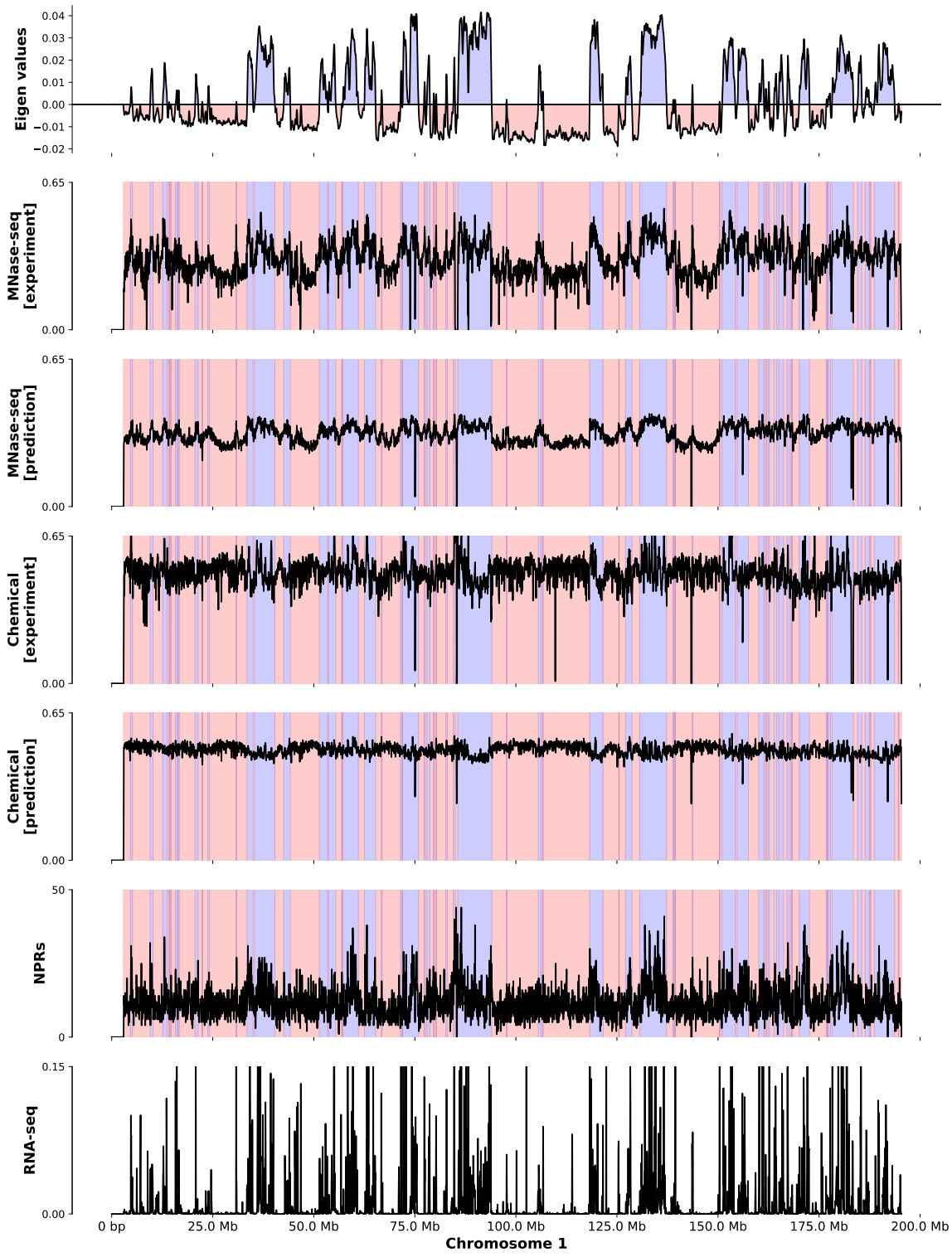


Figure 5.7: **Comparison of nucleosome positioning signals with compartmentalization on chromosome 1.** The top panel shows the first principal component of the normalized Micro-C contact matrix, used to define two compartments (highlighted in red and blue). Subsequent panels display experimental and predicted nucleosome occupancy profiles from MNase-seq and chemical cleavage assays, followed by the number of nucleosome positioning regions (NPRs) and RNA-seq coverage.

TSS selected for their nucleosome-positioning effect (see Methods and references). As shown in Figure 5.8, our models correctly capture the presence of a nucleosome-depleted region (NDR). However, they fail to reproduce the precise positioning of the +1 nucleosome and the phased downstream arrays observed experimentally, despite the increased ISM signal at TSS.

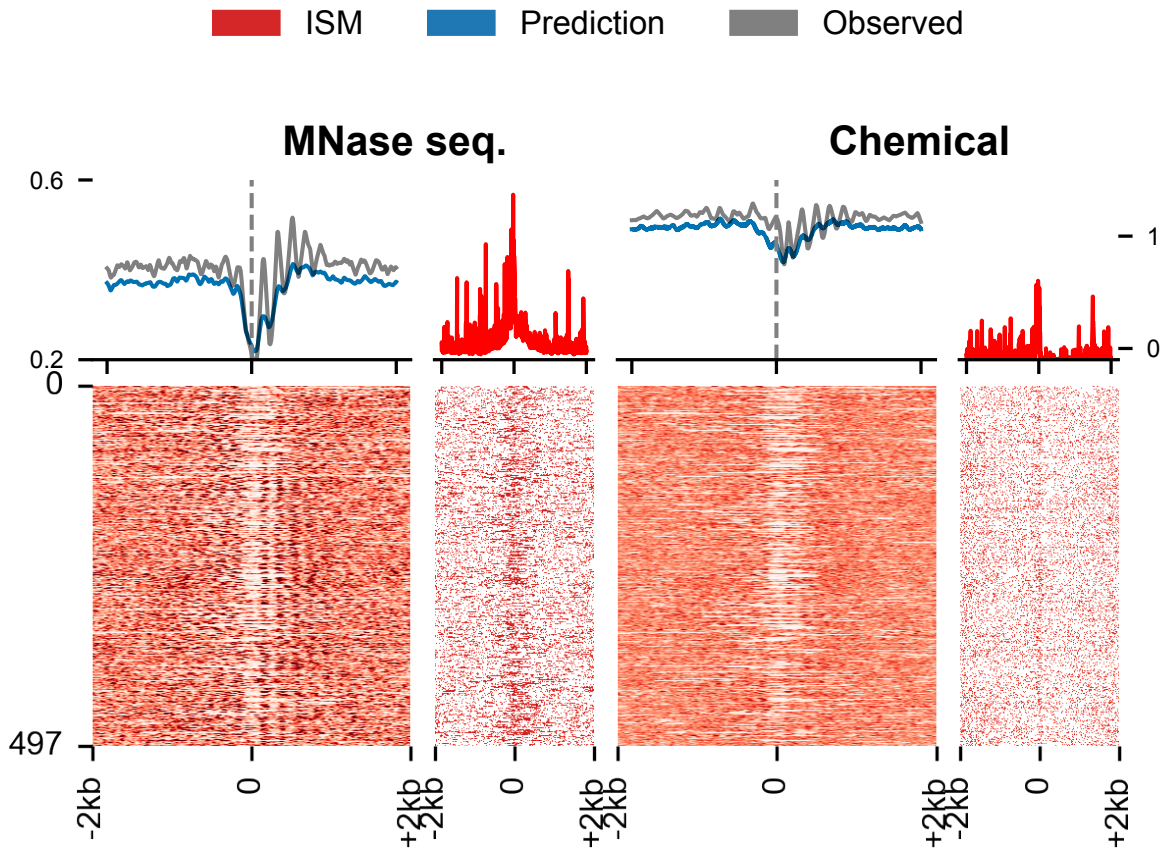


Figure 5.8: **Nucleosome positioning and *in silico* mutagenesis (ISM) around selected transcription start sites (TSS).** Top panels: aggregate nucleosome occupancy observed experimentally (grey), predicted by the model (blue), and ISM signal (red). Bottom panels: heatmaps showing the observed nucleosome occupancy and ISM across individual loci, aligned on the TSS (dashed line).

These results show that NPRs punctuate the genome in a non-random manner and point to a connection with transcriptional activity. However, the precise contribution of sequence features remains unclear, motivating a closer examination of the sequence content of these regions to identify potential transcription factor binding motifs.

5.2 Transcription factors binding sites motifs

To focus on motifs with potential functional relevance in our system, we retained only those corresponding to transcription factors expressed above 1 transcript per million (TPM) in mouse embryonic stem cells, according to ENCODE RNA-seq data .

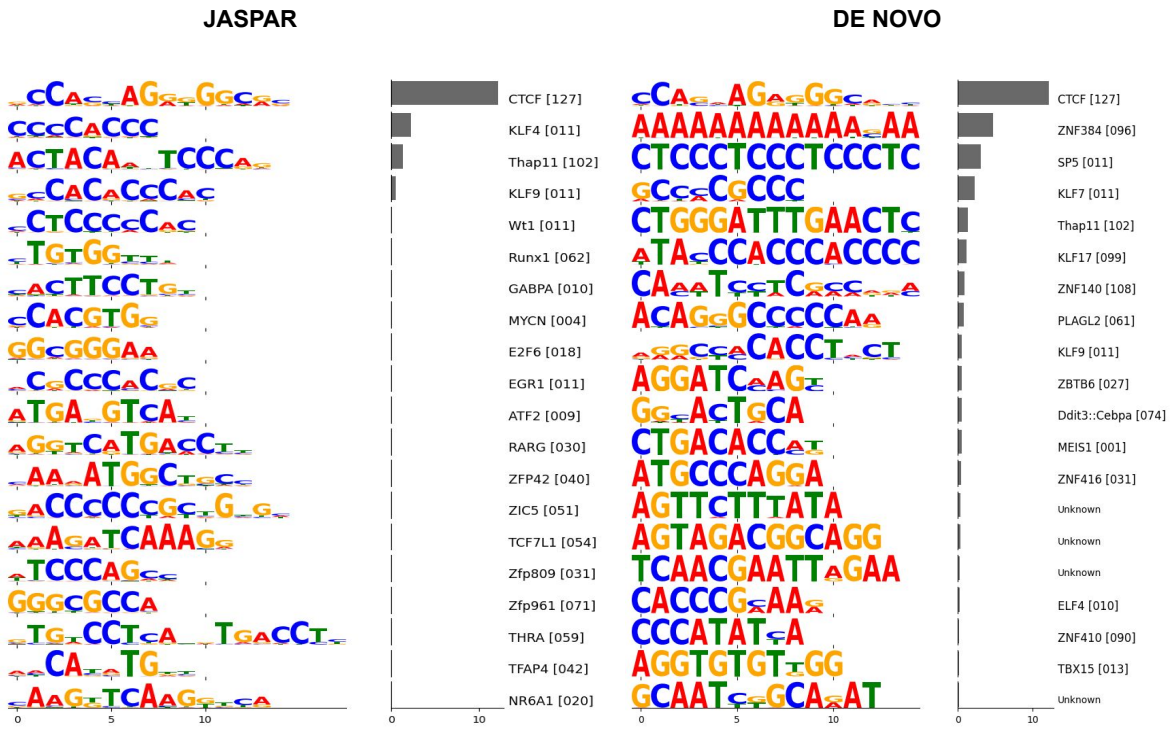


Figure 5.9: **Motif retrived from sequence analysis** issued from jaspar database (left) or *de novo* with STREME (right). The bar plot gives the percentage of NPR carrying the motif. The name is the batch match using TOMTOM (regarding E-value) and the [JASPAR cluster]

Figure 5.9 shows the top 20 motif found from JASPAR database and *de novo* with STREME (s. De novo motif discovery (STREME) revealed enriched sequence patterns specific to our dataset, some of which deviate from canonical motifs. When comparing these results to known transcription factor binding sites from JASPAR using XSTREME, we observed partial overlap but also notable differences. This discrepancy can be explained by the different aims of the two approaches: de novo algorithms identify the most enriched signals in the sequences at hand, potentially uncovering context-specific or degenerate variants, whereas database scanning relies on predefined consensus motifs that represent averaged binding profiles across diverse conditions. Thus, the combination of both approaches provides complementary insights, highlighting both the canonical binding determinants and the dataset-specific variants that may drive nucleosome organization in mESC.

JASPAR provides matrix clusters, which group transcription factor binding motifs based on similarity. This hierarchical clustering uses correlation between position weight matrices to align and organize motifs, and each cluster is annotated with the correspond-

ing structural class of the transcription factor. Such clustering simplifies the exploration of large motif collections by reducing redundancy and highlighting families of related motifs. In contrast, clusters produced by STREME (or other de novo approaches) arise directly from the enrichment of sequence patterns in the input data, and therefore reflect dataset-specific overrepresented motifs. JASPAR clusters represent a curated, knowledge-driven organization of known motifs, whereas STREME clusters represent an unsupervised, data-driven discovery of enriched patterns. Together, they provide complementary perspectives: JASPAR clusters offer a biological framework for interpretation, while de novo clusters capture the motifs most relevant in the studied context [148, 149].

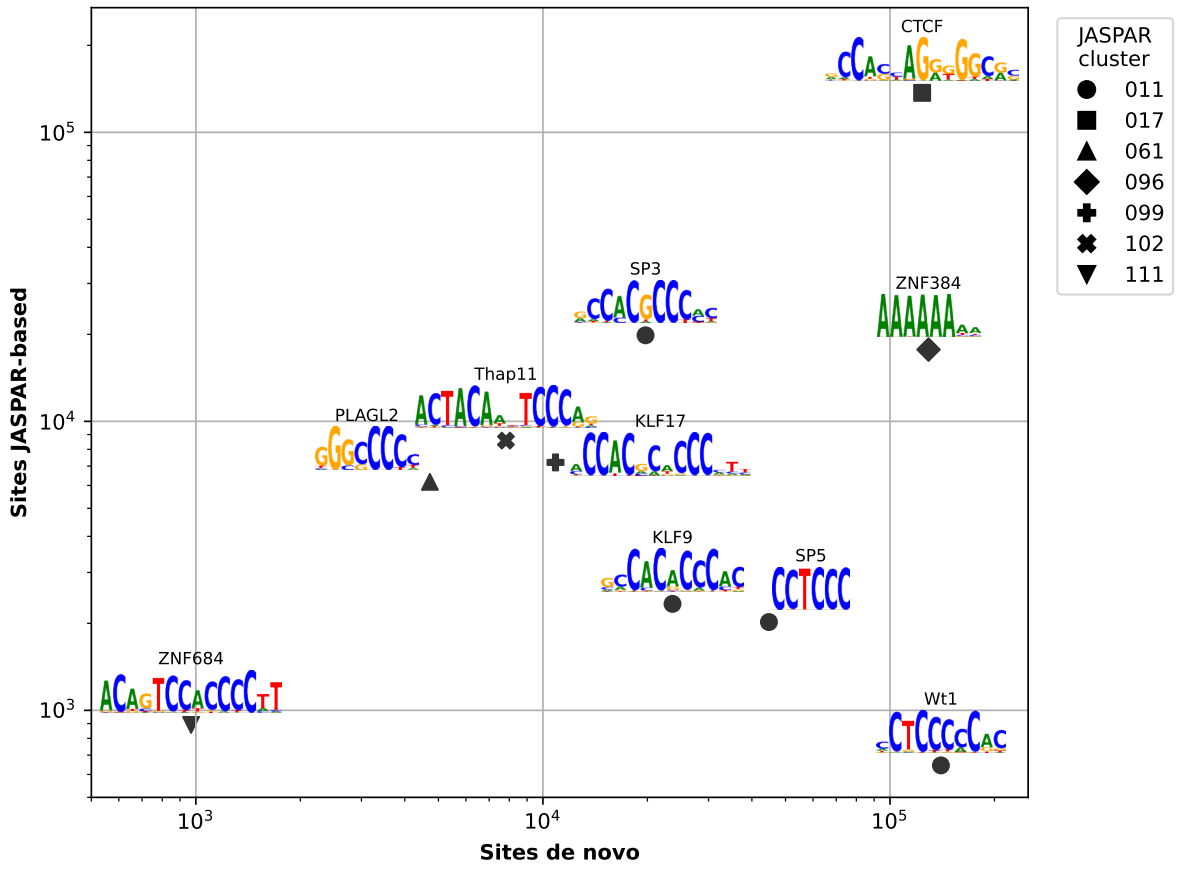


Figure 5.10: **Number of sites retrieved from XSTREME motif discovery** The scatter plots shows the number of motifs from JASPAR (y-axis) and *de novo* (x-axis) discovered.

Among the XSTREME motif discovery, ten motifs are retrieved in both approach (clustered together by XSTREME) and clustered in seven hierarchical clusters according to JASPAR (Figure 5.10). The JASPAR cluster 11 is composed of representative of the KLF/SP TFBS familly and appear four times in this selection.

Among the motifs enriched in NPRs, CTCF stands out as the single most frequent and biologically interpretable hit. Given its well-established role in chromatin architecture and nucleosome phasing, we examined in detail how the model captures and responds to CTCF binding sites.

5.2.1 CTCF as the conductor of nucleosome positioning

the largest fraction of intersecting NPR carry CTCF binding cite. To better understand the sequence features driving the model's predictions at CTCF sites, we compared the in silico mutagenesis (ISM) scores to the information content of the CTCF motif (Figure 5.11). The information content of each position in the Position Weight Matrix (PWM) reflects the sequence specificity of CTCF binding, with higher values indicating more conserved and functionally constrained bases.

In the left panel of Figure 5.11, we show the CTCF consensus motif (as discovered in our data) alongside the average ISM score at each base within the motif. Strikingly, the ISM signal peaks over the same positions that exhibit high information content, particularly in the central core region known to engage zinc fingers 4 to 8. This alignment suggests that the model has learned to associate these conserved bases with their functional impact on nucleosome positioning.

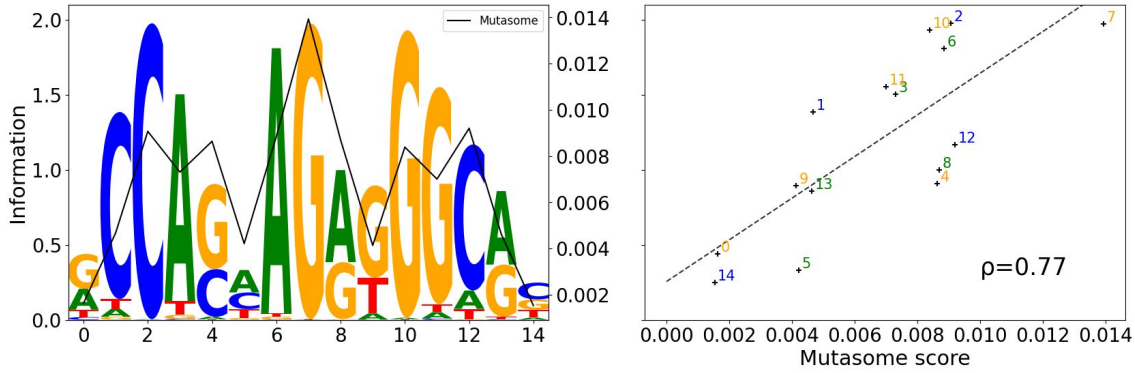


Figure 5.11: ISM score correlates with PWM information on CTCF core binding site. Discovered CTCF motif and ISM score (left panel), Scatter plot of the information (y-axis) and the ISM score (x-axis) (right panel)

The right panel confirms this relationship quantitatively: a positive correlation is observed between ISM scores and PWM information content across motif positions. This indicates that the model is not only sensitive to the presence of the motif but also to its internal structure and variability, further supporting the biological interpretability of the network's learned features.

5.2.2 Among the large amout of CTCF sites, network can discriminate the positioning ones

While CTCF motifs are widespread across the genome, only a subset are bound in vivo and associated with phased nucleosome arrays. To test whether the model distinguishes between functional and inert motif instances, we stratified CTCF sites by ChIP-seq binding status [3] and examined predicted nucleosome landscapes around each group.

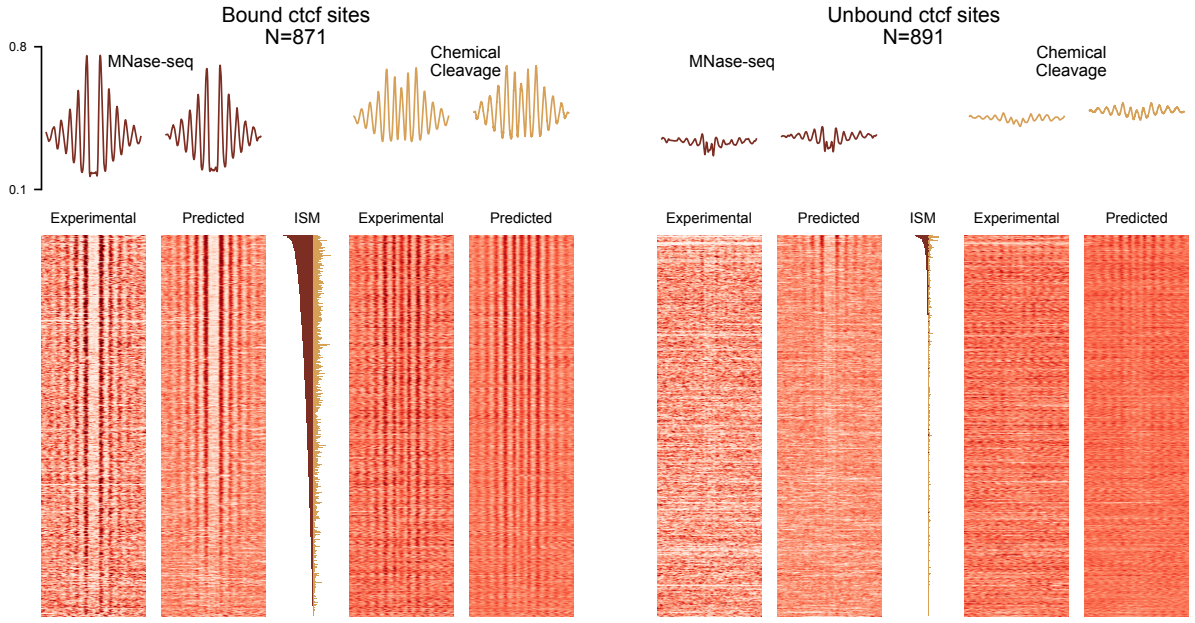


Figure 5.12: **Nucleosome occupancy and ISM around CTCF sites**.

Remarkably, Figure 5.12 shows that the model recovers the phased nucleosome arrays specifically around bound CTCF sites, while unbound motifs lack such structure in the predictions. The figure shows CTCF sites on chromosome 19 as they haven't took part of the training. This result confirms that the model captures subtle sequence determinants and local context features associated with functional CTCF occupancy, highlighting its ability to generalize beyond average profiles and to discriminate biologically relevant binding events.

To further confirm that the model has internalized sequence features specifically associated with CTCF-mediated nucleosome organization, we computed the average ISM signal on CTCF sites, stratified by binding status.

As shown in Figure 5.13, bound CTCF sites exhibit significantly higher ISM scores compared to unbound sites, for both the MNase-seq and chemical cleavage-trained models. This strong signal suggests that the model has learned to associate specific sequence features (such as the CTCF motif and its surrounding context) with nucleosome positioning activity. In contrast, unbound motif instances elicit little to no ISM response, indicating a lack of influence on the predicted nucleosome occupancy. These findings provide further support for the model's capacity to distinguish regulatory from inert motif instances purely based on sequence, and to assign functional relevance through its internal representation. Interestingly, several sites classified as unbound by ChIP-seq still display phased arrays of well-positioned nucleosomes. These sites correspond to those with the highest ISM scores among the unbound set. They also show poorly mappable nucleosome occupancy, which may explain why they appear as false negatives.

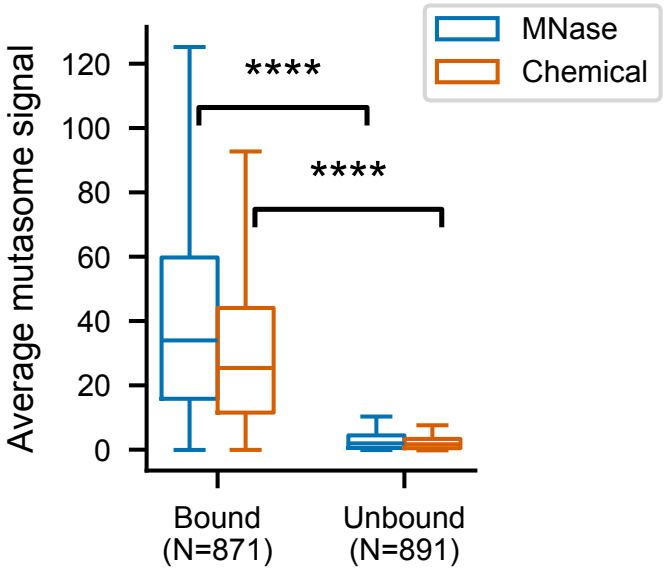


Figure 5.13: **Average ISM signal on CTCF sites** Significativity computed using Mann-Withney test ($p <= 1e-4$)

5.2.3 SP/KLF family position nucleosomes

From all the motifs matching with the JASPAR-11 cluster (KLF/SP), identified in our analysis (Figure 5.10), STREME-2 motif [GCCCCGCCC] most closely matches the JASPAR motif for KLF7 (Fig. 5.14). However, members of the KLF/SP transcription factor family share highly similar G-box motif, making unambiguous assignment to a single factor difficult from sequence alone. Given this similarity, functional insights from well-characterized family members can inform interpretation.

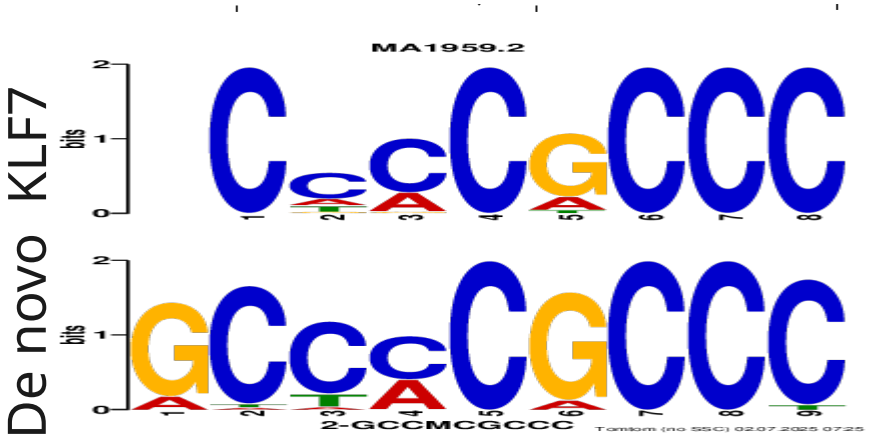


Figure 5.14: **Logoplot of the *de novo* retrieved STREME-2 motif and its best match (KLF7) in the JASPAR database.**

The chromatin landscapcape around this motif is shown in Figure 5.15 where nucleosome-depleted region (NDR) flanked by regularly phased nucleosomes is displayed with MNase-seq data, whereas the chemical-cleavage data show a phased-nucleosome array but no cen-

tral depletion. This effect can be observed in both unique and repeated sequences, except for the LINE family. It is also noteworthy that experimental data and model predictions display similar nucleosome occupancy. Notably, some loci lack experimental coverage yet still yield predictions. The motif is mostly retrieved in unique genomic sequences (∼65%) and in LTR elements (∼20%). STREME-2 is not the only motif attributable to the SP/KLF TFBS family, and further representatives will be discussed in a following section.

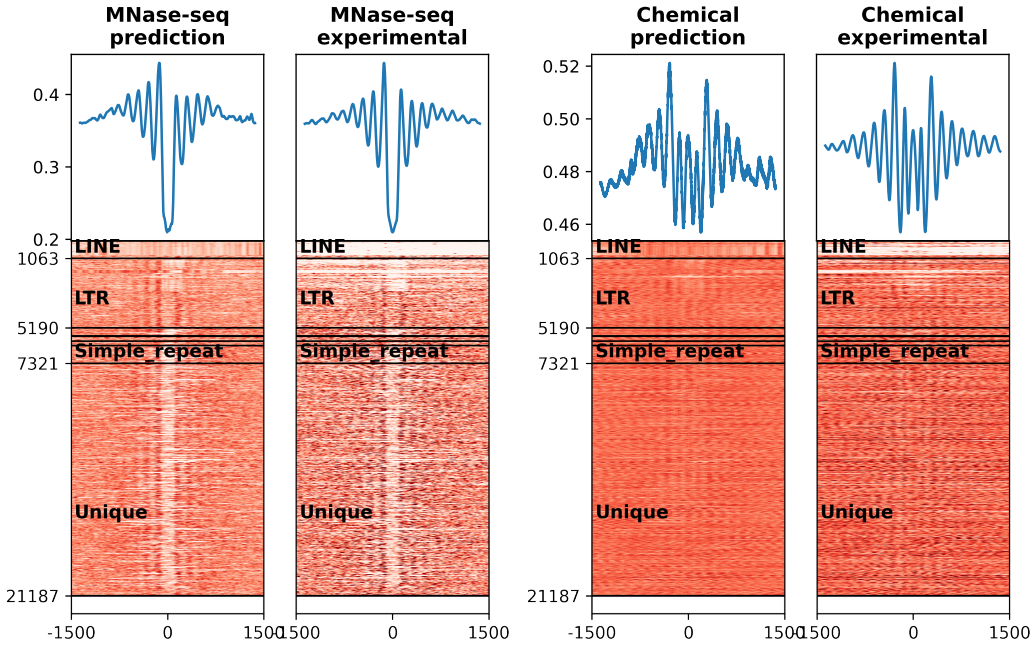


Figure 5.15: **Predicted nucleosome occupancy and genomic distribution of the STREME-2 motif from the xstreme analysis.**

Together, these findings underscore the role of SP/KLF-like motifs as important sequence determinants of nucleosome positioning. The consistent association of the G-box with phased nucleosomal arrays, and its enrichment in both unique genomic regions and LTR elements, suggest that members of the SP/KLF family contribute broadly to chromatin organization in mouse embryonic stem cells.

5.2.4 Pluripotency factors are retrieved in Nucleosome Positioning Regions

THAP11 (Ronin) has been identified as an essential factor for embryogenesis and the pluripotency of mouse embryonic stem cells [150]. Retrieved both *de novo* ($E = 6.81e-208$) and from JASPAR database($E = 2.36e-1314$), it is surprisingly only retrieved in B2_MM1 Short Interspersed Nuclear Element (SINE) subfamily. This motif is associated with NPR and phased-array of nucleosome in MNase-seq data, Chemical-cleavage data display really similar profile with a putative fragile-nucleosome (Figure 5.16).

Several several other TFBS with established roles in mouse embryonic stem cell reg-

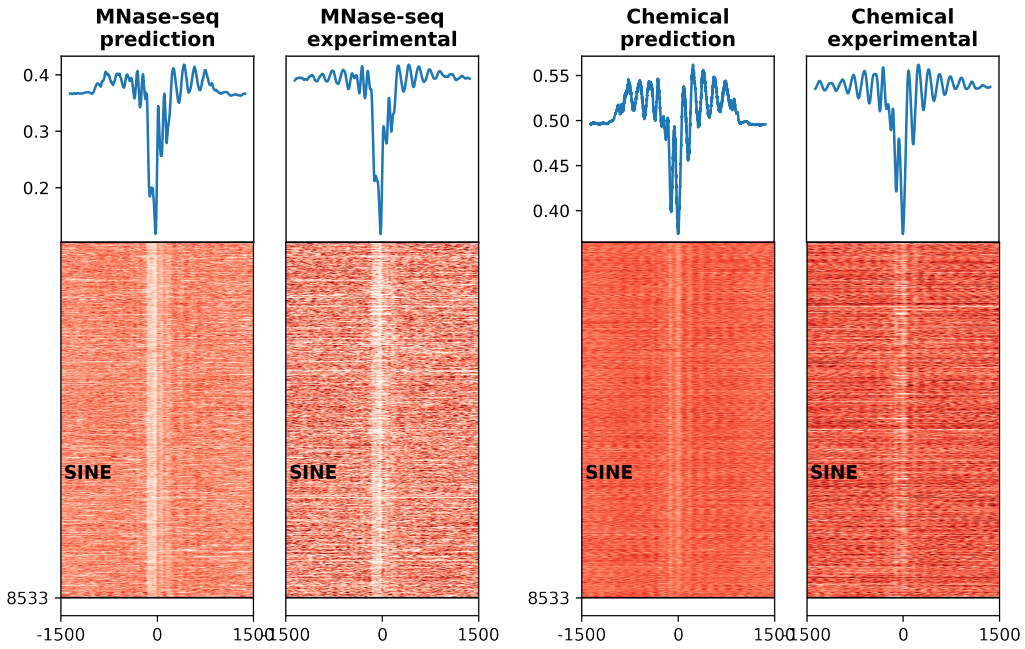


Figure 5.16: **Predicted nucleosome occupancy and genomic distribution of the STREME-3 motif from the xstreme analysis.**

ulation were recovered in the JASPAR-based search. YY1 ($E = 6.03e - 1921$), an architectural protein with sequence-specific DNA binding activity, has been implicated in promoter–enhancer looping, nucleosome organization at regulatory elements and chromatin remodeler partner [131]. Nanog ($E = 1.47e - 66$),Oct4 ($E = 1.96e - 150$), Sox family ($E \in [5.63e - 169, 4.84e - 0.17]$) and Oct4::SOX2 heterodimer ($E = 1.268e - 112$) core pluripotency factors, are known for shaping the chromatin landscape of ESCs [3].

Although these motifs were not recovered *de novo* in our STREME analysis, their enrichment in NPRs is consistent with their established roles in regulating nucleosome positioning and accessibility in stem cell chromatin. These additional factors, together with the strong signals from CTCF and SP/KLF, reinforce the view that NPRs identified by ISM correspond to sequence-encoded regulatory hotspots in the mESC genome. The complete list of discovered TFBS and E-values can be found in Appendix E.

Looking at the retrieved *de novo* motifs shown in Figure 5.9, an interesting fact is that beyond canonical regulatory sites such as CTCF or KLFs, the PWM gives a single nucleotide per position, indicating high-conservation in the motif and most likely a repetitive origin. This observation suggests that transposable elements and other genomic repeats may act as reservoirs for nucleosome-positioning motifs. In summary, ISM reveals that the neural network relies on compact, interpretable, and biologically meaningful sequence features to predict nucleosome positioning. These features correspond not only to known transcription factor binding sites like CTCF and SP/KLF, but also to *de novo* motifs enriched in repeats, hinting at a broader role for repetitive elements in chromatin

organization. The next chapter will explore this connection in depth.

Chapter summary

- ISM is an intuitive approach to extract rules of nucleosome positioning from the trained model
- ISM highlights compact NPRs that coincide with phased nucleosomes, replicate across models, and align with orthogonal assays.
- Motif analyses show that NPRs are enriched for canonical regulators such as CTCF and SP/KLF, as well as pluripotency factors and repeat-derived motifs.

  Together, these results demonstrate that the network relies on interpretable, biologically meaningful sequences to organize nucleosomes, bridging known TFBSs and opening on the analysis of repetitive elements

Chapter 6

From repeats to regulation: the repeated genome's role in nucleosome organization

6.1 Transposable elements are actively involved in nucleosome positioning

Figure 6.1 shows the enrichment of RepeatMasker subfamilies in NPRs, the detailled method of computation can be found in Appendix B. Among the different *repClass* categories, SINE, LINE, LTR, and satellite-related sequences (including *low_complexity* and *Simple_repeat*) stand out due to their strong association with NPRs. In particular, the B2 SINE family, which is already one of the most abundant classes of repeats in the mouse genome, shows a marked enrichment, suggesting that these elements are not only passively incorporated into the chromatin landscape but actively contribute to the establishment of phased nucleosome arrays. Similarly, LINE-1 (L1Md) and LTR elements also display significant associations, consistent with their potential to provide sequence features and potential genomic regulation that strongly position nucleosomes [48, 55]. Altogether, this analysis indicates that repetitive elements, beyond their role in genome expansion, serve as key sequence determinants that punctuate the genome with hotspots of nucleosome positioning.

To better understand their specific contribution to chromatin organization, we next examined in greater detail the main repeat families enriched in nucleosome positioning regions, namely SINEs, LINEs, LTRs, and microsatellites elements.

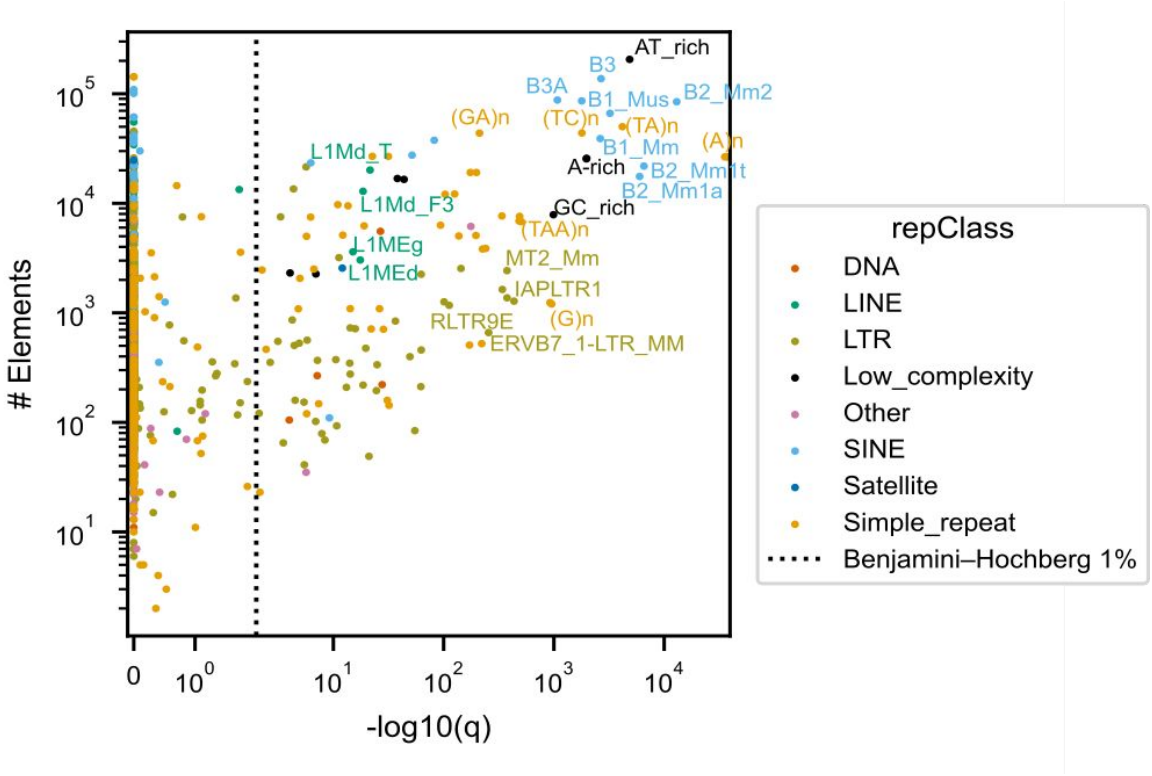


Figure 6.1: **Enrichissment of repeat subfamilies (repName from RepeatMasker) in Nucleosome Positioning Regions**. The y-axis indicates total count of elment on the genome, while the x-axis the p-values corrected by Benjamini-Hochberg method. Points are colored by repeat class (repClass from RepeatMasker).

6.1.1 B2 SINEs as carriers of functional CTCF sites

As discussed in the introduction, SINEs, and in particular the B2 family (comprising B3A, B2_Mm2, B2_Mm1t, B3, B2_Mm1a), are known to harbor CTCF binding sites. Such insertions can introduce functional insulator elements that shape local chromatin structure and nucleosome organization [7]. Figure 6.2 shows that a substantial fraction (∼80%) of positioning CTCF motifs is embedded within B2-family SINE repeats.

Interestingly, the strongest nucleosome-positioning CTCF sites are found in unique sequences, this is shown in Figure 6.3 that depicts the mean amplitude of the signal for each cluster, both model agree on a higher amplitude of B3s over B2s and the maximum is reach over unique sequences. The effect is visible on the heatmaps in Figure 6.2.

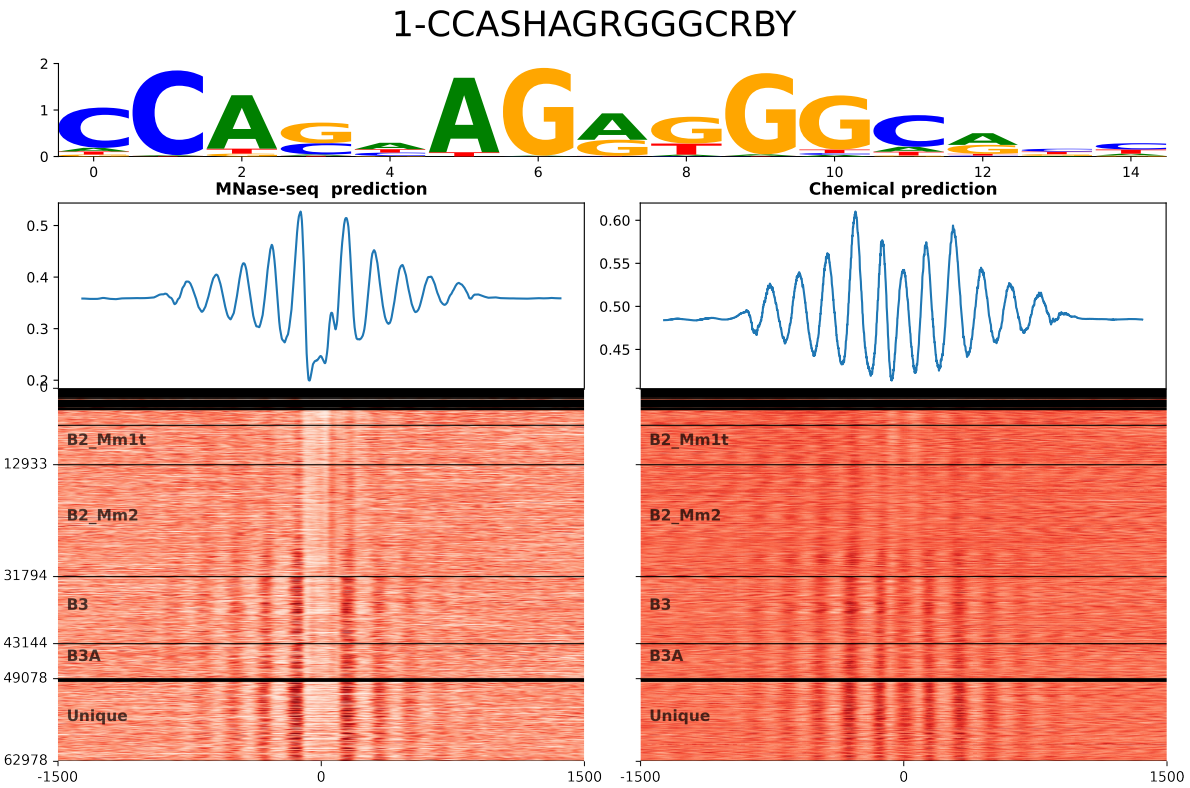


Figure 6.2: **Nucleosome positioning around NPRs harboring CTCF motifs and their genomic distribution.** Heatmaps are sorted by genomic localization and ascending ISM score. All sequences are oriented with the CTCF motif in the same direction.

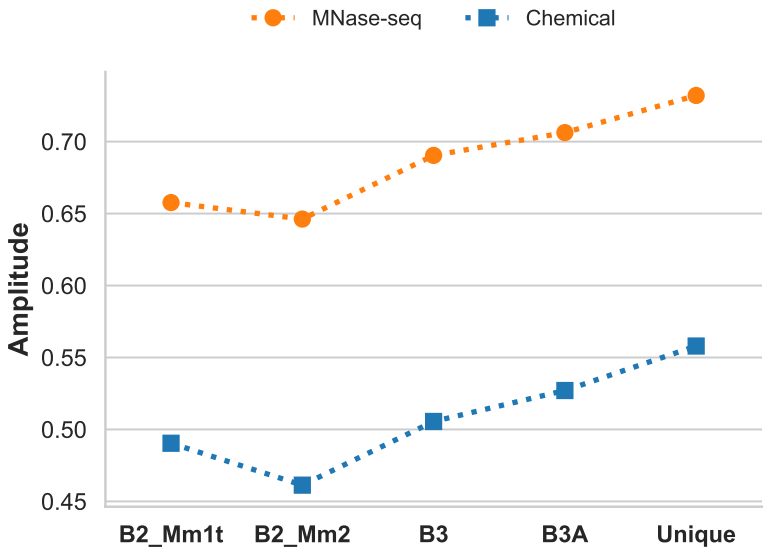
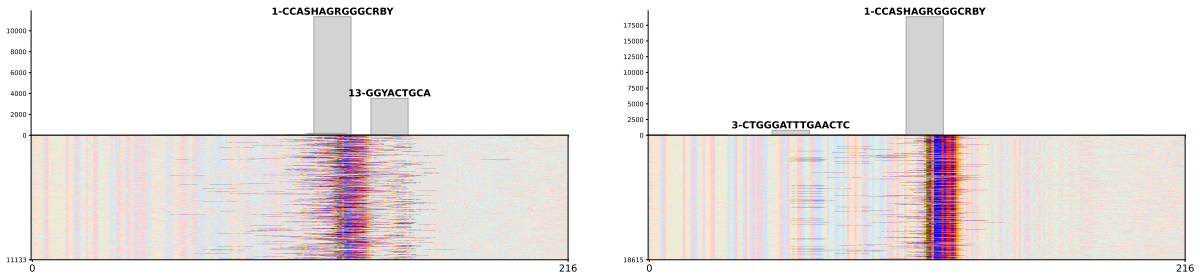


Figure 6.3: **Amplitude of nucleosome signal by repeat family**

When comparing B3 and B2_Mm2, it is clear that B3 elements carry a supplementary motif (STREME-13) absent in B2_Mm2 repeat (Figure 6.4). This observation is further confirmed by Figure 6.5 which show that motif-13 is found in B3, B3A and unique sequences. This motif match with Ddit3::Cebpa ($E = 4.9e-1$) motif in JASPAR database, however considering its location in the genome and its strong similarity with the the well-described upstream CTCF U-motif [7, 142, 151] we can reasonnably considering the motif 13 as such.



(a) **B3** (b) **B2_Mm2**

Figure 6.4: **Motifs found NPRs carried by B3 and B2_Mm2 SINEs.** Both subfamilies share motif 1 (CTCF), B3 also carry motif 13.

Our model recapitulates this property: SINE elements overlapping conserved CTCF motifs exhibit clear nucleosome phasing patterns, consistent with the regulatory activity of their embedded CTCF sites. Phasing was even stronger for full CTCF motifs (upstream motif + core motif). This result highlights the model's ability to distinguish functional motif instances even within repetitive contexts, demonstrates the high resolution of the ISM approach, and confirms its capacity to predict nucleosome positioning in line with known biology.

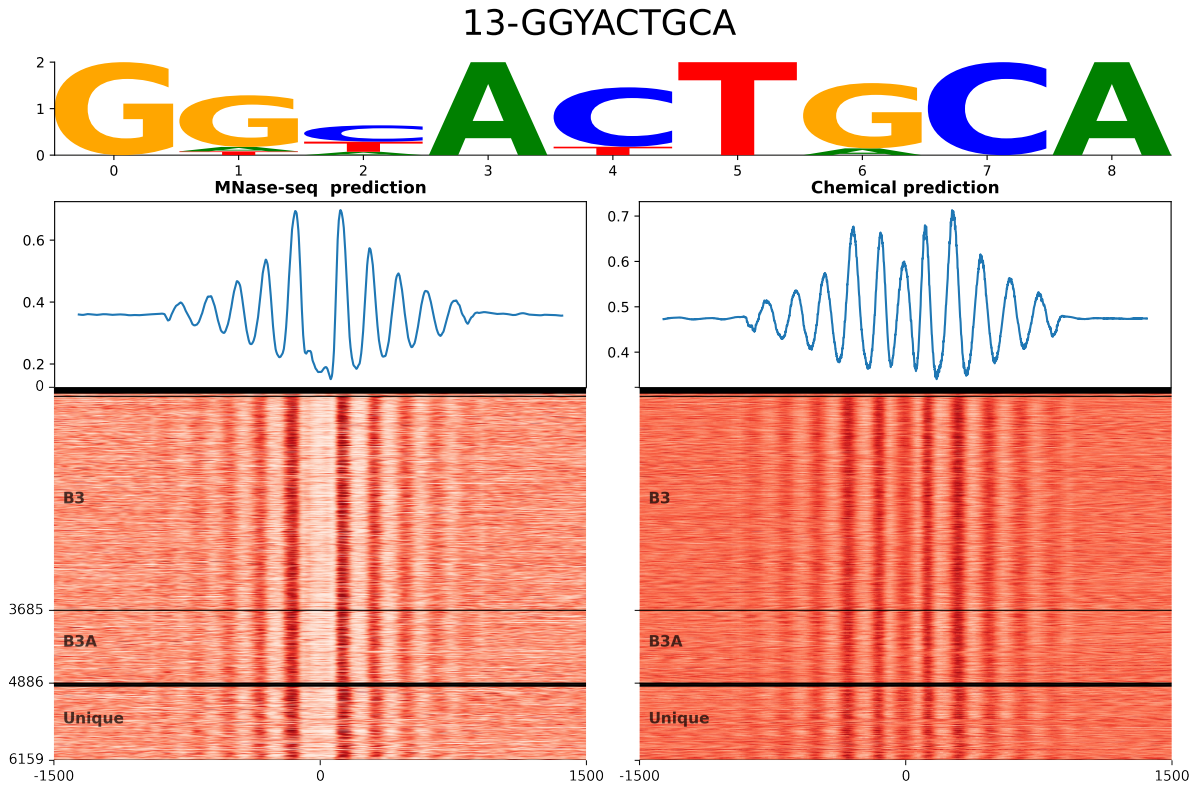


Figure 6.5: **Nucleosome positioning around U-motif containing CTCF sites.** Putative CTCF's U-motifs are found in unique sequences and in the B3 subfamily.

6.1.2 Long Terminal Repeats

Beyond their classification as transposable elements, many LTR sequences originate from Endogenous RetroViruses (ERVs), ancient viral insertions that became fixed in the genome over evolutionary time. A typical ERV includes coding regions flanked by two LTRs, which harbor regulatory sequences such as promoters and enhancers. Even in the absence of intact viral genes, these LTRs often persist as so-called solo LTRs and retain their regulatory potential. This is particularly relevant in embryonic stem cells, where ERV-derived LTRs can function as divergently transcribed regulatory elements, actively shaping the chromatin landscape. These sequences are frequently co-opted to drive transcription and contribute to enhancer activity [48]. The observed enrichment of NPR on LTRs in our study (Figure 6.1) may therefore reflect the structural and regulatory influence of ERV-derived elements on nucleosome positioning. MT2_MM is a representative solo LTR element of this type [152]. Figure 6.6 shows that several motifs are consistently identified as NPRs on this repeat.

The two main motifs (STREME 47 and 40) retrieved on MT2_Mm are also found on other repetitive and unique sequences. Interestingly, for both motifs, predictions indicate only a NPR on MT2_Mm sequences, whereas in unique sequences they predict regular and phased nucleosome positioning (Figure 6.7). Both motifs exhibit very high information content, meaning they are highly consistent in nucleotide composition. Motif

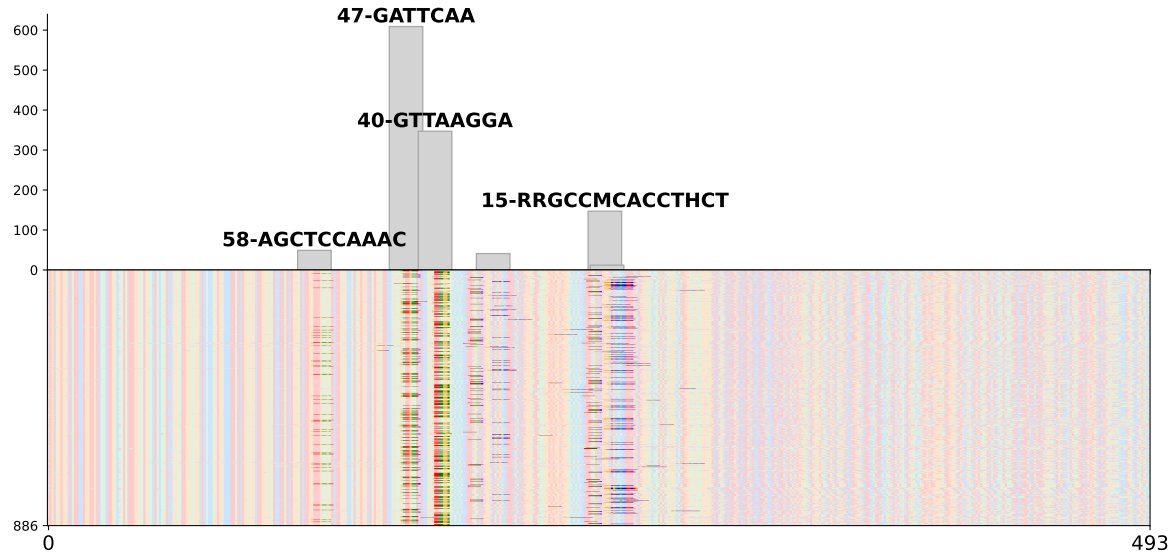
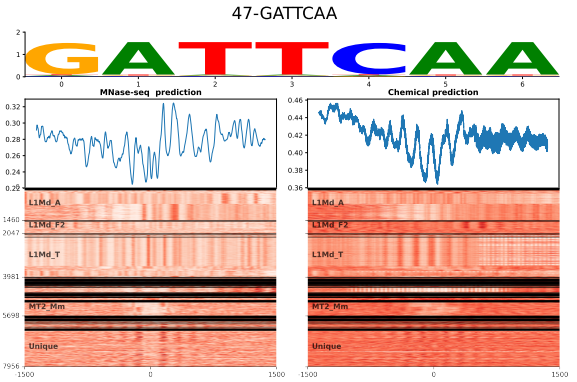


Figure 6.6: **MT2_Mm elements carry nucleosome-positioning–relevant regions (NPRs).** Bar plots indicate the number of motifs retrieved within MT2_Mm elements. Repeats are aligned relative to their consensus sequence, and the underlying colored heatmap represents the nucleotide composition across aligned elements.

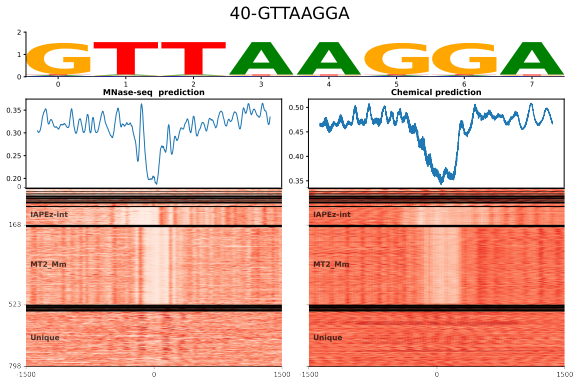
47 (Figure 6.7a) significantly ($E = 2.2e-01$) matches the Dux (Figure 6.7a) TFBS. Dux is a recognized driver of totipotency in mESC [153]. Motif 40 (Figure 6.7b) has no match in the JASPAR database (Appendix E).

Motif 15 significantly matches KLF9 ($E = 3.35e-02$), a transcription factor of the KLF/SP family discussed earlier. In addition to MT2_Mm, it is mostly found in ORR families, another LTR-derived group. This motif also occurs in several LINE sequences and in unique sequences. Figure 6.8 shows that the MNase-seq model depicts the characteristic pattern of a NPR with phased nucleosomes on each side, with a particularly strong effect in ORR subfamilies and a moderate effect in unique sequences. However, this effect is inverted in the chemical cleavage model: phasing in ORR is barely detectable, whereas it appears in unique sequences.

Among LTRs, Intracisternal A-particle (IAP) elements form a family of rodent-specific endogenous retroviruses. They remain among the most transcriptionally active retrotransposons in mouse embryonic stem cells [154], and their regulatory potential has been linked to both physiological and pathological processes. Many IAPs insertions contain promoter or enhancer sequences capable of driving transcription in a cell type–specific manner, often influencing nearby gene expression. Importantly, IAPs can act as alterna-



(a) **Motif 47** Found in MT2_Mm, LINEs, and unique sequences



(b) **Motif 40** Found in MT2_Mm, IAPEz-int, and unique sequences

Figure 6.7: **The two main motifs retrieved on MT2_Mm NPRs**

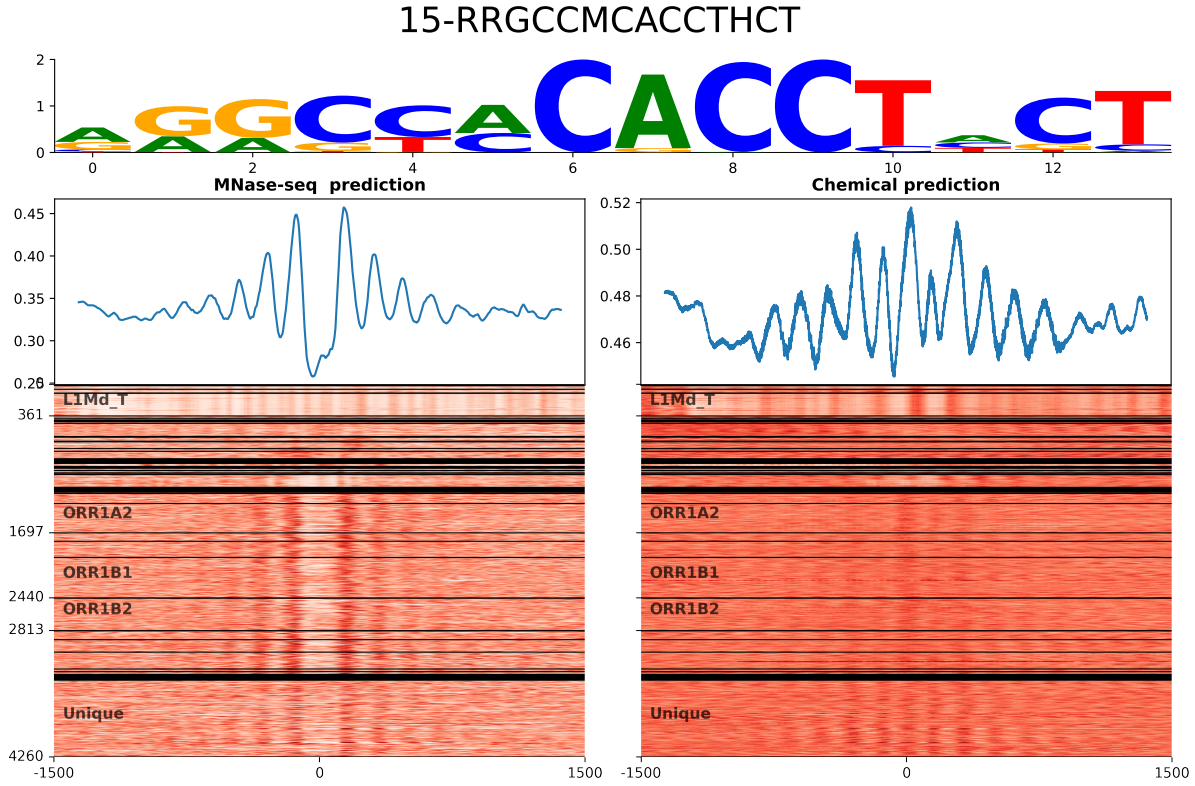


Figure 6.8: **Nucleosome occupancy and distribution across the genome of motif 15**

tive promoters for host genes or generate novel transcripts, thereby altering chromatin accessibility and local nucleosome organization [54,155]. The enrichment of NPRs at IAP loci in our dataset suggests that these elements may contribute to nucleosome positioning both through their intrinsic sequence composition and by recruiting transcription factors to their LTRs.

IAPEz-int (derived from the Repbase consensus IAPEZI) [156] is a representative of this rodent-specific LTR family. Its size is remarkable (∼7 kb), considering it and the high similarity to the consensus (>88% [156]) of the repeats, it is expected to highly similar prediction from the model. NPRs can be observed at both ends of the sequence (Figure 6.9).

Three of these motifs (in addition to motif 40 described in Figure 6.7b) are shown in Figure 6.10, motif 44 and 46 are almost exclusively found in IAPEz sequences. Motif 46 show a predicted NPR on the motif locus for both model whereas sequences carrying motif 44 are predicted unconcordantly for mnase-seq and chemical-cleavage model at the motif site; however the upstream NPR which correspond to the KLF/SP motif (streme 2 motif, Figure 5.15).

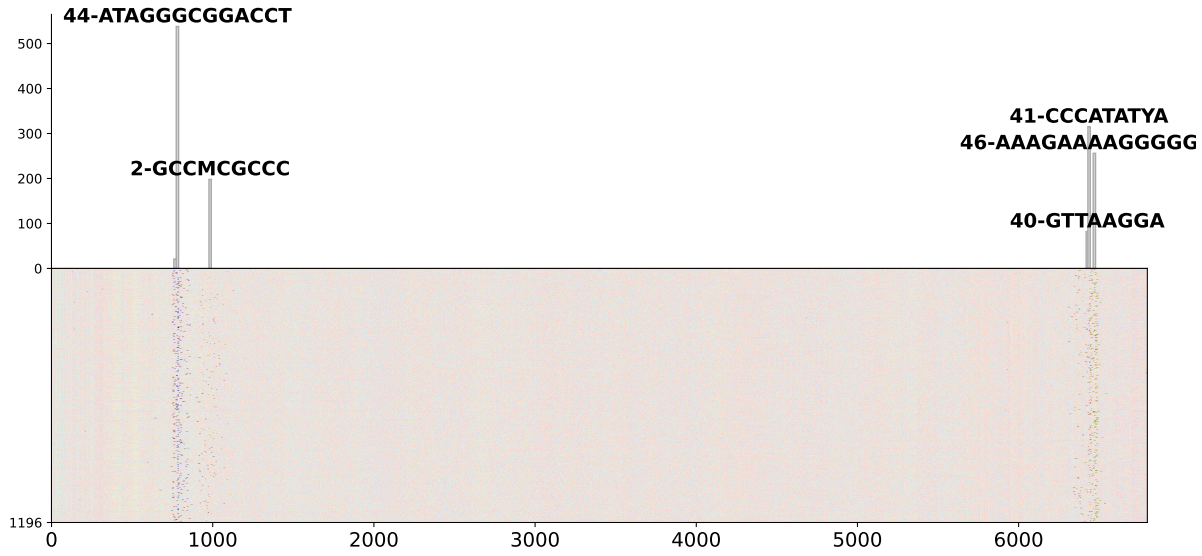
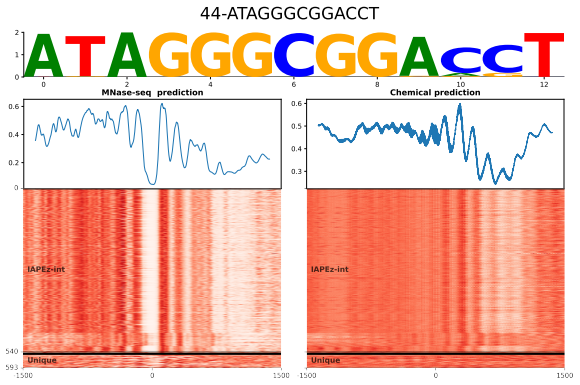
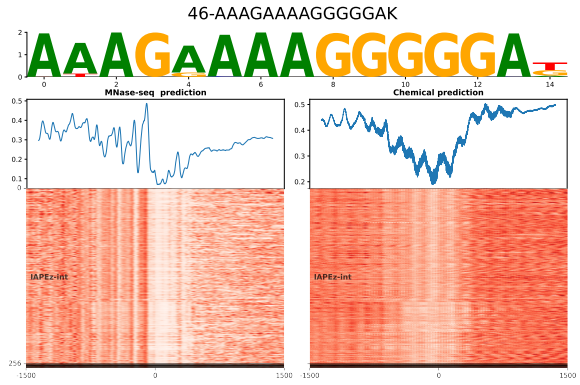


Figure 6.9: **IAPEz-int** Bar plots indicate the number of motifs retrieved within elements. Repeats are aligned relative to their consensus sequence, and the underlying colored heatmap represents the nucleotide composition across aligned elements.

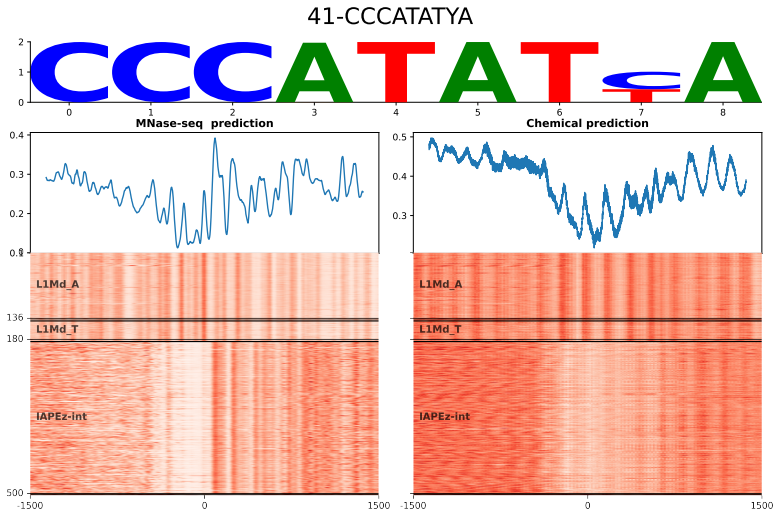
These results highlight LTRs as strong candidates for sequence determinants of nucleosome positioning. By combining known TFBS with constitutive motifs consistently recognized by our model as nucleosome positioning regions, they provide a dual regulatory logic that can shape local chromatin architecture. This suggests that LTRs are not only passive remnants of retroviral insertions but active contributors to genome organization, with potential implications for gene regulation in both developmental and evolutionary contexts.



(a) **Motif 44** Mostly found in IAPEz, with few motifs in unique sequences



(b) **Motif 46** Almost exclusively retrieved in IAPEz



(c) **Motif 41** Mostly found in IAPEz and L1Md repeats

Figure 6.10: **Nucleosome occupancy and distribution across the genome for STREME motifs 44, 46 (top), and 41 (bottom).**

6.1.3 ISM reveals constitutive conserved motifs of transposable elements

The study of the LINE repeats revealed a large variety of motifs on L1 sequences. A characteristic of LINEs is their composition in tandem repeats. Figure 6.11 shows the density of motifs discovered along the L1Md_A elements, as expected motifs appear regularly repeated a period of approximately 200b, reflecting the underlying structure.

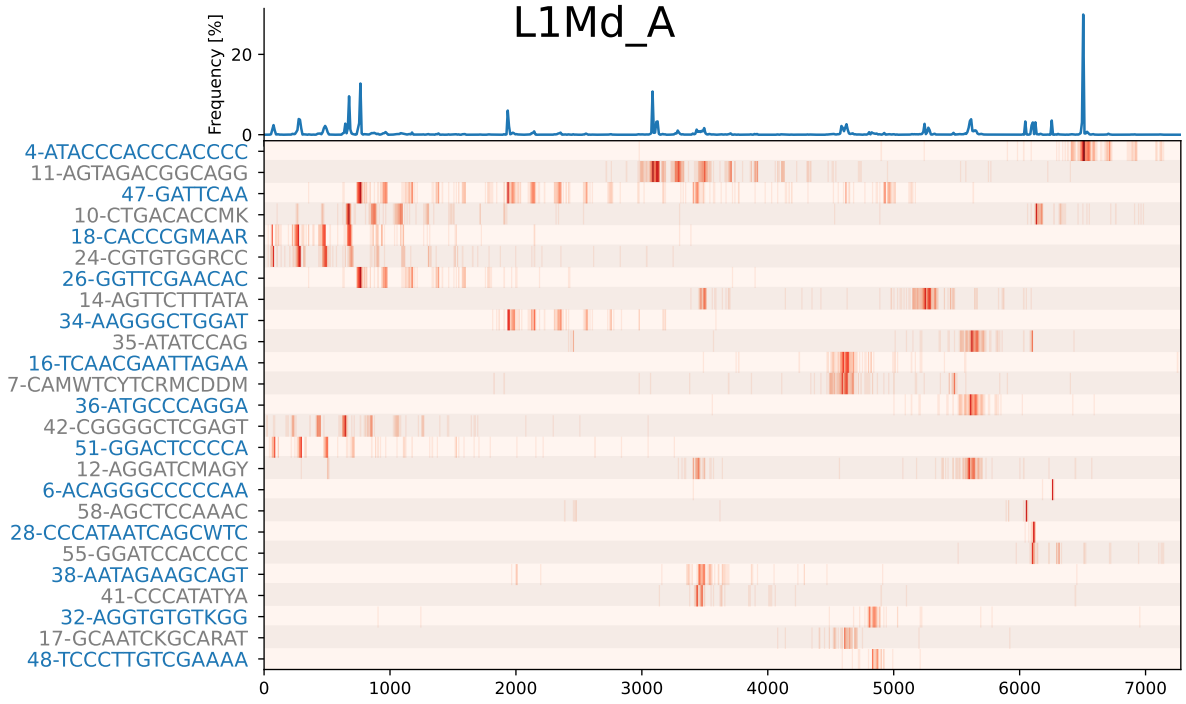


Figure 6.11: **NPRs on L1Md_A**. Each line in the heatmaps depict the density of the motif, regarding their position in elements. Motifs are sorted by occurrence.

Among the recovered motifs, STREME 4 was the most abundant in L1Md_A. This G-box-like motif was associated with the KLF/SP family. The *de novo* STREME 4 motif matched weakly with KLF17 ($E = 9.38e-02$) and KLF4 ($E = 6.24e-01$); however, KLF17 was also identified directly in the XSTREME scan ($E = 4.68e-605$). Predicted nucleosome occupancy around STREME 4 is consistent across both models, showing an asymmetric pattern: a regularly phased nucleosome array followed by an NPR and a weaker signal toward ORF–3' (Figure 6.12).

Several additional motifs were retrieved (Figure 6.13). Among them, only motif 52 was clustered by XSTREME with motif 15 (described in the LTR section), which is related to the KLF/SP family of binding sites. By contrast, motifs 11, 18, 24, 42, and 51 did not yield significant matches in the JASPAR database. The corresponding position weight matrices show little to no variation across occurrences (Figure 6.13), and their genomic distribution indicates that they are restricted to LINEs, consistent with strong conservation.

The associated nucleosome landscapes typically show regularly phased nucleosomes, sometimes accompanied by a local decrease in baseline signal. Notably, the predicted

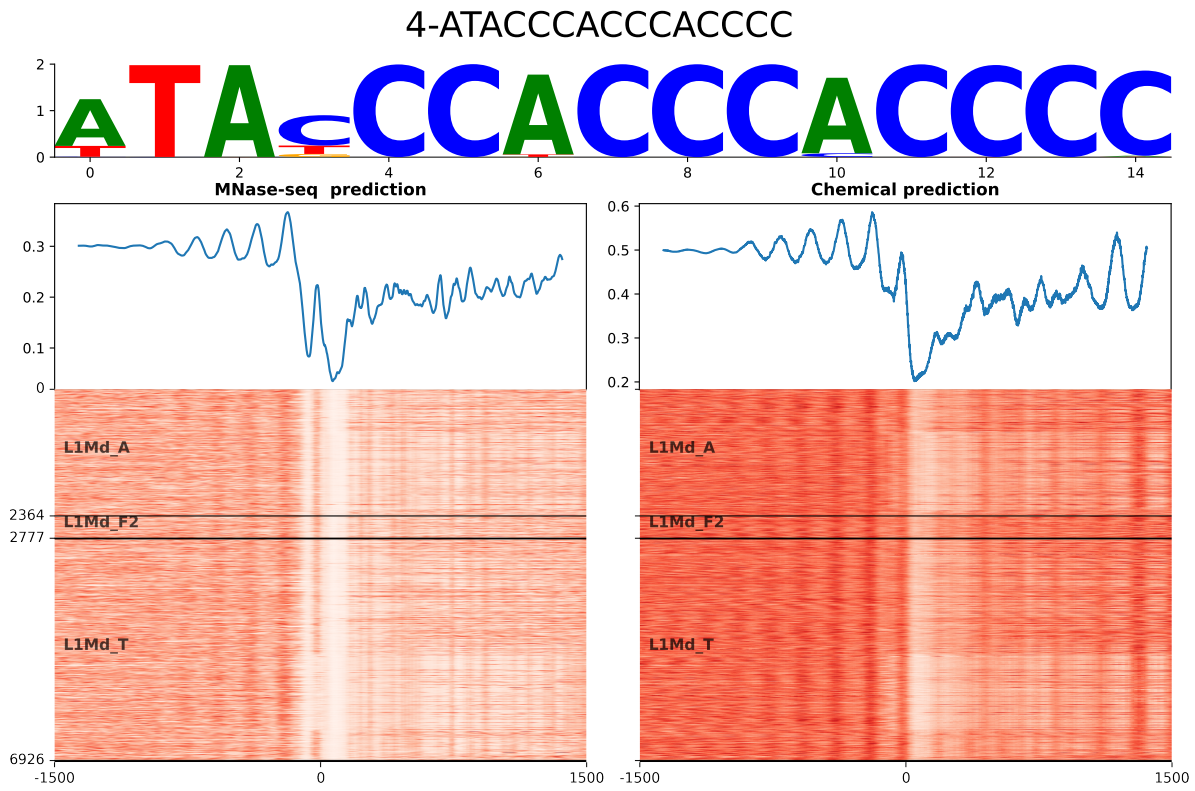


Figure 6.12: **STREME-4 Motif**. 3'-5' oriented

landscapes are robust across different LINE 1 subfamilies: for example, motifs 18, 34, and 11 display similar patterns in both L1Md_A and L1Md_T. However, motifs 10 and 24 exhibit divergent predictions depending on the subfamily in which they occur.

Among the motifs retrieved by ISM, several do not correspond to any annotated transcription factor binding site but instead derive from internal regions of transposable elements. Their recovery by ISM suggests that the model is sensitive to these conserved repeat-derived patterns, regardless of whether they act through direct protein binding or other mechanisms.

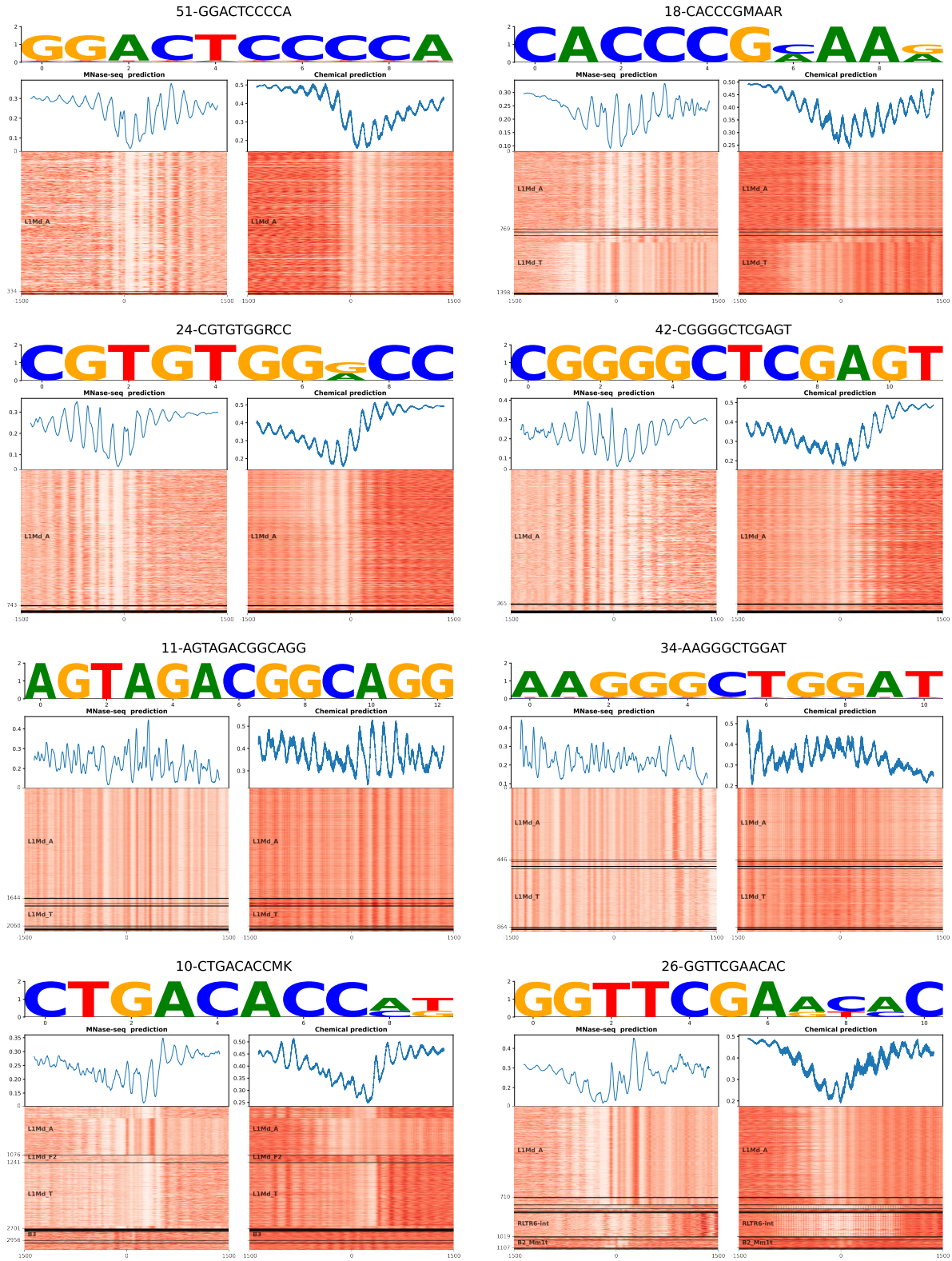


Figure 6.13: **STREME motifs identified on L1Md_A sequences**

A prominent example of repeat-associated sequence features was observed in LINE-1 monomers. RepeatMasker annotated three L1Md_A elements; however, closer inspection revealed that this region contains 35 tandem monomers of 208 bp, corresponding to repeated promoter A units (Appendix D, Figure 6.14). Both prediction models identified phased nucleosomes aligned with the monomeric structure, despite low amplitude in the chemical cleavage model predictions. Nucleosome dyads coincide with the monomer repeats, suggesting that sequence periodicity alone can drive nucleosome phasing independently of transcription factor binding.

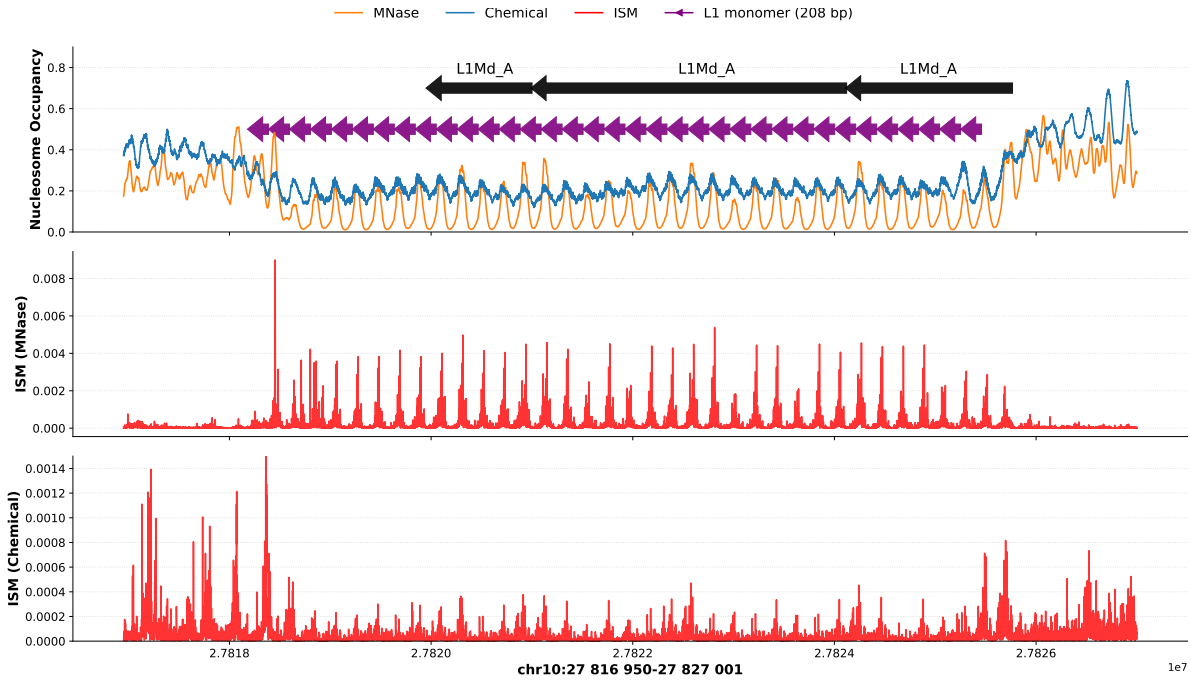


Figure 6.14: **Nucleosome occupancy and ISM score on a L1 element with phased nucleosomes** Top plot represent the nucleosomic density predicted by both models with annotation of repeatMasker (element-level) and Tandem Repeat Finder (monomer-level). The two plots underneath show ISM response of each model.

Although this structure is unlikely to function as a canonical promoter, the periodic organization strongly influences nucleosome phasing: predicted nucleosome dyads align with the monomer repeats, suggesting that specific conserved sequence periodicity, rather than transcription factor binding, can also shape local nucleosome organization. In the context of nucleosome positioning, their regular size and GC-rich sequence (63%) composition may facilitate the establishment of phased nucleosomes, either by providing sequence-specific anchoring sites or by shaping local DNA physical properties. The presence of these motifs in high-scoring ISM regions may therefore reflect the model's detection of repeat-derived sequence features that can serve as organizational landmarks in the chromatin landscape.

Together, these findings underscore the ability of LINE-1 elements to shape chromatin architecture not only via transcription factor recruitment but also through intrinsic sequence periodicity and structural DNA features. The ISM approach highlights that repeat-derived motifs, even in the absence of canonical regulatory function, can act

as sequence-encoded nucleosome-positioning signals, suggesting a dual role for LINEs as both reservoir of TFBSs and chromatin scaffolds.

6.2 Local nucleotide enrichment and intrinsic DNA features shape nucleosome organization

Beyond transposable elements, our genome-wide analysis highlights low-complexity sequence features as recurrent nucleosome-positioning cues. Satellites and simple repeats are over-represented among NPRs (Figure 6.1), and the XSTREME summary (Fig. 5.9) nominates two prototypical motifs—poly(dA:dT) and the $(CTCC)_n$ motif as strong shapers of local chromatin architecture. Because these elements are short, abundant, and fully sequence-encoded, they provide a clean testbed to disentangle intrinsic DNA mechanics from protein-mediated effects. In what follows, we focus on these low-complexity tracts to show how they create barrier-like NDRs and phased arrays, and when their effects coincide with specific factor occupancy, thereby linking sequence-encoded cues to nucleosome organization.

6.2.1 GC-content is a strong determinant of nucleosome positioning

Consistent with work in yeast [157], nucleosome occupancy is highest at intermediate GC content. Specifically, for our data the MNase-seq signal peaks between 30–60% GC, whereas the chemical-cleavage signal shows a broader plateau between 25–70% GC (Figure 6.15). Voong *et al.* attribute the attenuation at low GC to MNase bias, which preferentially digests A/T-rich sequences [73]. In the same vein, the high-GC shift could be explained by preferential MNase digestion of DNA wrapped around "fragile" nucleosomes, which are frequent near GC-rich TFBS, as retrieved by our XSTREME analysis (Figure 5.10). Both assays display a strong decay at the extreme GC values. Conversely, the ISM signal shows two local maxima at the edges of the intermediate-GC range, suggesting specific properties associated with both low- and high-GC sequences.

6.2.2 Adenine arrays have a pivotal role in nucleosome positioning

Among NPR, poly-A tandem repeats stand out in the XSTREME analysis (Figure 5.9). Their nucleosomal profile shows a pronounced Nucleosome Depleted Regions (NDR) and the presence of a putative fragile nucleosome for the chemical-model, flanked by phased nucleosome arrays (Figure 6.16). This pattern is characteristic of the presence of a chromatin barrier, as described by Mavrich et al. [2]. Both MNase- and Chemical-cleavage–based models consistently show the presence of this barrier, in agreement with previous studies highlighting (A/T)-rich sequences as nucleosome repellent, possibly acting as a chromatin barrier [2, 88, 130].

TOMTOM analysis further matched the poly-A motif with ZNF384 binding sites.

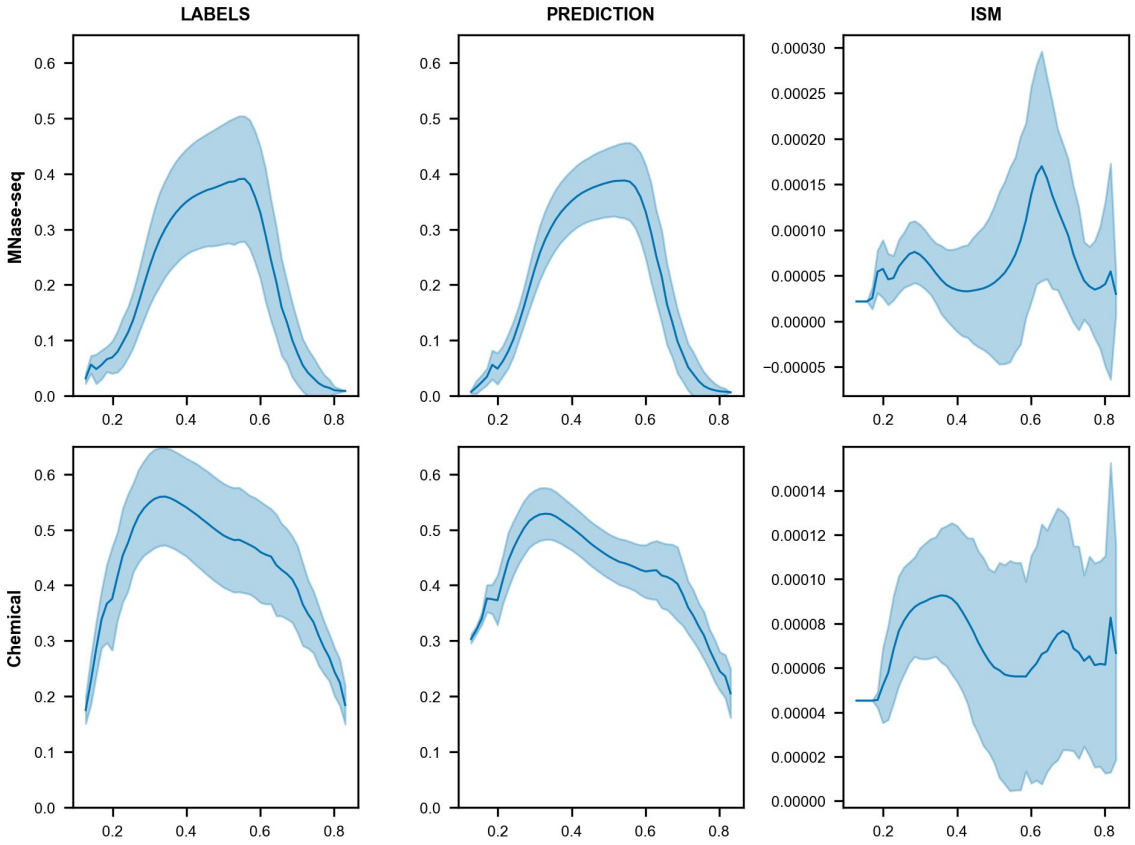


Figure 6.15: **Influence of GC on nucleosome landscape and ISM**. Mean and standard deviation of signal using 500bp rolling mean. Signal was filtered to discard non mappable regions

While ZNF384 has been mostly studied in the context of leukemia, recent work highlighted its role in chromatin organization, showing enrichment at TAD borders and preferential interactions with SINE elements [158]. This makes ZNF384 a compelling candidate as a nucleosome positioning determinant.

When aligning ZNF384 ChIP-seq peaks on the sequences, the nucleosome landscape (Figure 6.17) essentially recapitulates the pattern seen on the general set of poly-A repeats. Importantly, the ZNF384 metaplot does not significantly deviate from the STREME-5 ($(A)_n$) landscape, suggesting that only a subset of high-ISM poly(dA:dT) tracts are bound by ZNF384, while the positioning property itself is intrinsic to the poly-A DNA sequence.

Together, these results suggest a dual role of poly-A microsatellites in mESC, as effective an TFBS for ZNF384, as chromatin barriers due to their intrinsic physical properties.

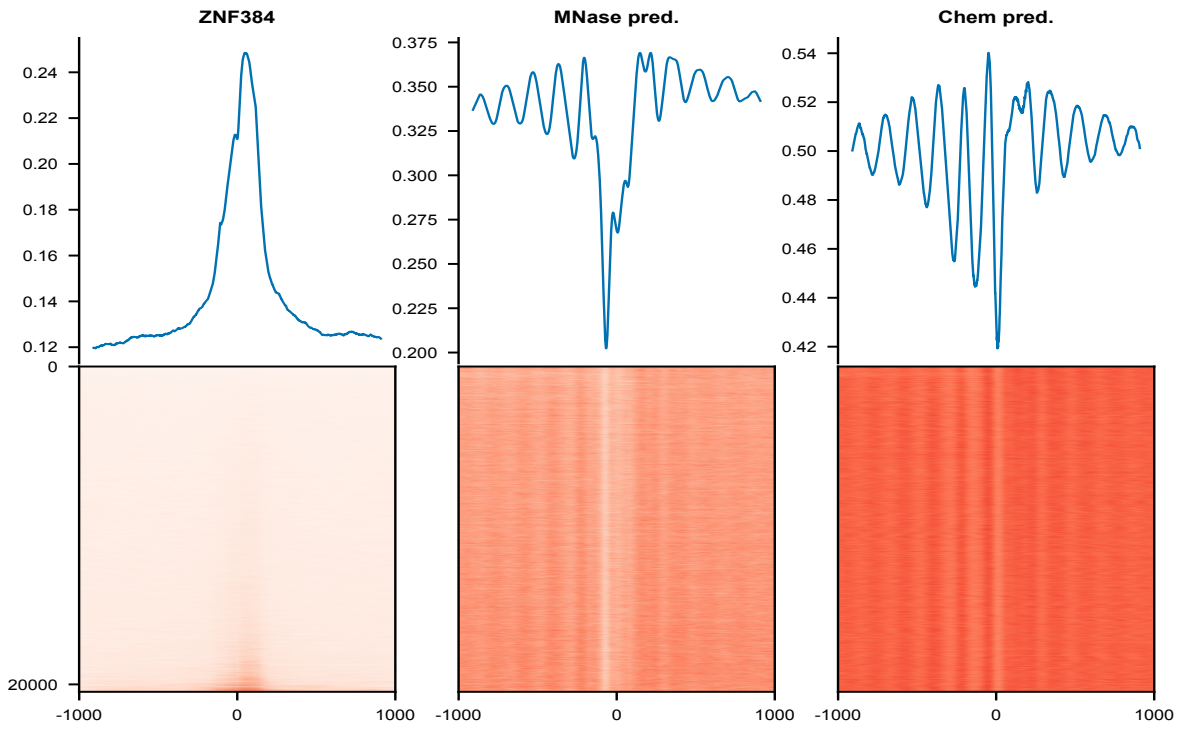


Figure 6.16: **Aggregated plot of predicted nucleosome occupancy around poly-A microsatellites**. ZNF384 Chip-seq is log10 scaled.

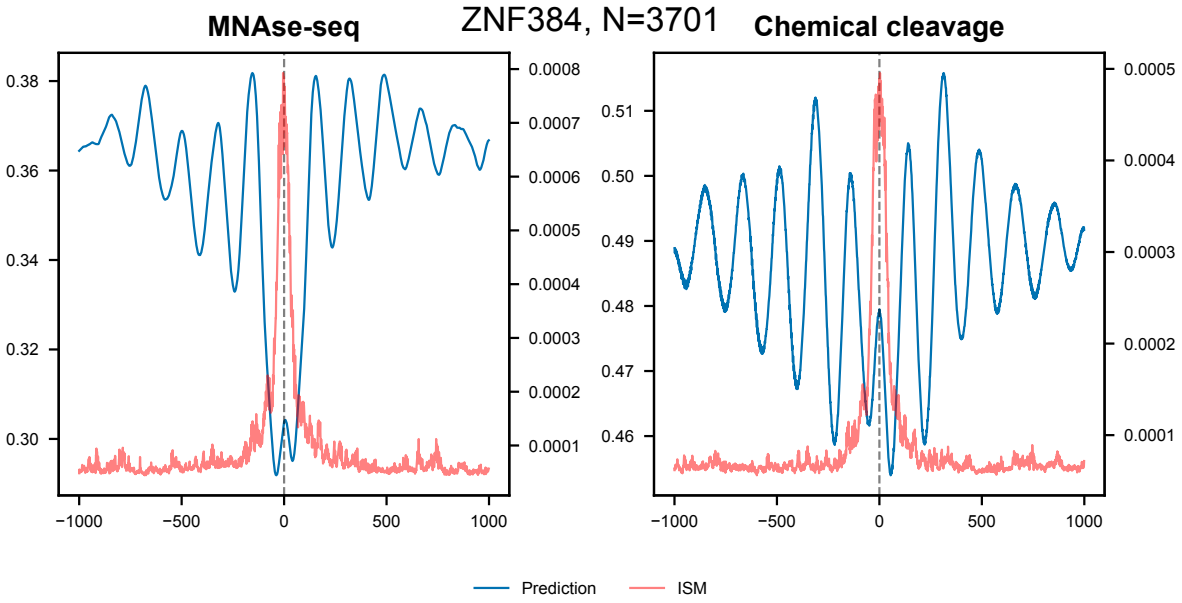


Figure 6.17: **Aggregated plot of nucleosome occupancy around ZNF384 ChIP-seq peaks**.

6.2.3 G-rich sequences

In contrast to the poly(A)-rich motifs described in the previous section, which are generally associated with nucleosome-depleted regions, we also investigated poly(dG:dC) sequences. These G-rich sequences are known to adopt non-canonical DNA structures, such as G-quadruplexes [159], and have been proposed to promote nucleosome formation in certain contexts [119, 120, 160]. Among the *de novo* motifs retrieved by XSTREME, one corresponded to a highly repetitive poly(dG:dC) tract and also matched entries in the JASPAR database. This MEME-3 motif ($E = 3.20e-154$) corresponds to a poly(dG:dC) tract of 15 bp (Figure 6.18.B). XSTREME clustered it with motifs retrieved directly from the JASPAR scanning: ZNF148 ($E = 1.71e-974$), VEZF1 ($E = 9.61e-198$), ZNF740 ($E = 3.88e-129$), RREB1 ($E = 1.27e-030$), and PRDM9 ($E = 3.46e-022$) (cluster 136 in Appendix E) for which we didn't find any ChIP-seq analysis. The TOMTOM analysis matched MEME-3 with numerous motifs. We report only the signficant one ($E <= 5e-2$) in Table 6.1. As expected from XSTREME clustering, we retrieved motifs VEZF1, ZNF740 and ZNF148 with signifant E-value, however RREB1 and PRDM9 are not significant in the TOMTOM analysis. SP1 and SP4 from the SP/KLF family were also identified, consistent with their known G-box binding motifs.

Name	**E-value**
ZNF740 (MA0753.3)	$3.06e-03$
ZNF281 (MA1630.3)	$3.33e-03$
VEZF1 (MA1578.2)	$1.04e-02$
PATZ1 (MA1961.2)	$1.41e-02$
ZNF148 (MA1653.2)	$2.25e-02$
SP1 (MA0079.5)	$2.95e-02$
SP4 (MA0685.2)	$4.71e-02$

Table 6.1: E-values of motifs associated MEME-3 match in TOMTOM analysis.

The effect of this motif on nucleosome landscape is shown on Figure 6.18. The MNase-seq model does not predict a NPR around the TFBS, whereas the chemical cleavage model shows a decrease in signal in the vicinity of the poly-G motif. In both models, however, phased nucleosomes flanking the motif can be observe

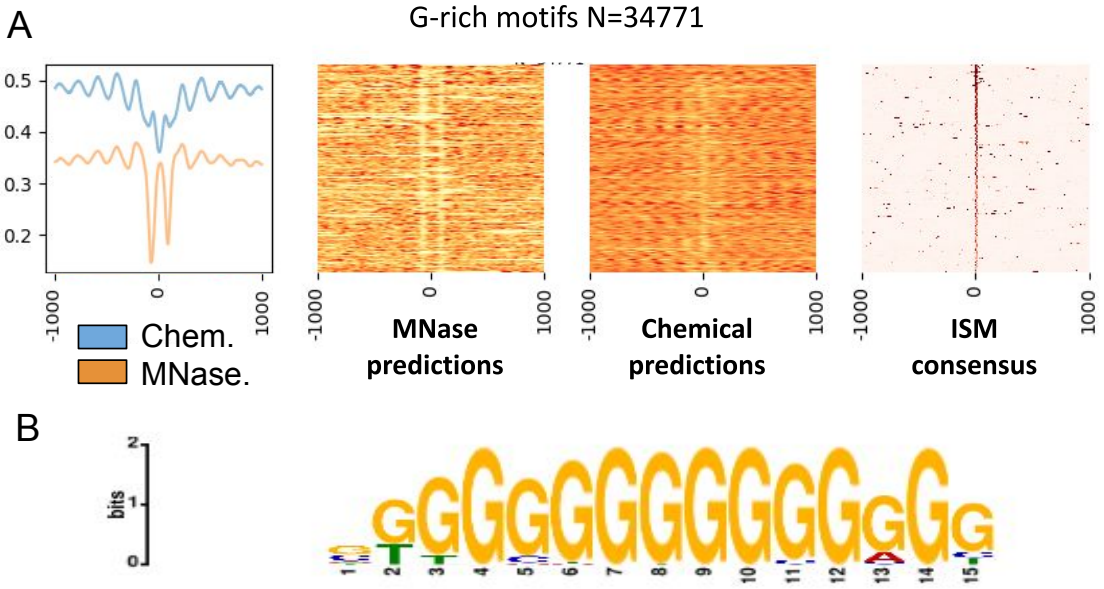


Figure 6.18: **A. Aggregated plot of predicted nucleosome occupancy and consensus ISM over G-rich motifs.** Motif has been highlighted by the Xstreme analysis, from the extracted Nucleosome Positioning Sequences. **B. Logo of *de novo* motif discovered by MEME**

The third most recurrent *de novo* motif retrieved in our analysis is also a G-rich motif. STREME-9 is a microsatellite-like motif CTCCCTCCCTCCCTC. This motif clustered with the SP5 TFBS from the JASPAR database ($E = 6.30e - 45$) and also matched SP5 directly with a significant E-value of $5.39e - 4$ in TOMTOM analysis. It represents another G-box–like motif but stands out for its peculiar $(CTCC)_n/(GGAG)_n$ composition. Figure 6.19 shows that, in both models, this motif is associated with an asymmetrical NPR flanked by phased nucleosome arrays. As expected, the motif is enriched in simple repeats such as microsatellites and G/GA/CT-rich sequences.

These results highlight that simple repeats are not mere genomic noise, but can shape distinct nucleosome landscapes and potentially provide binding sites for transcription factors.

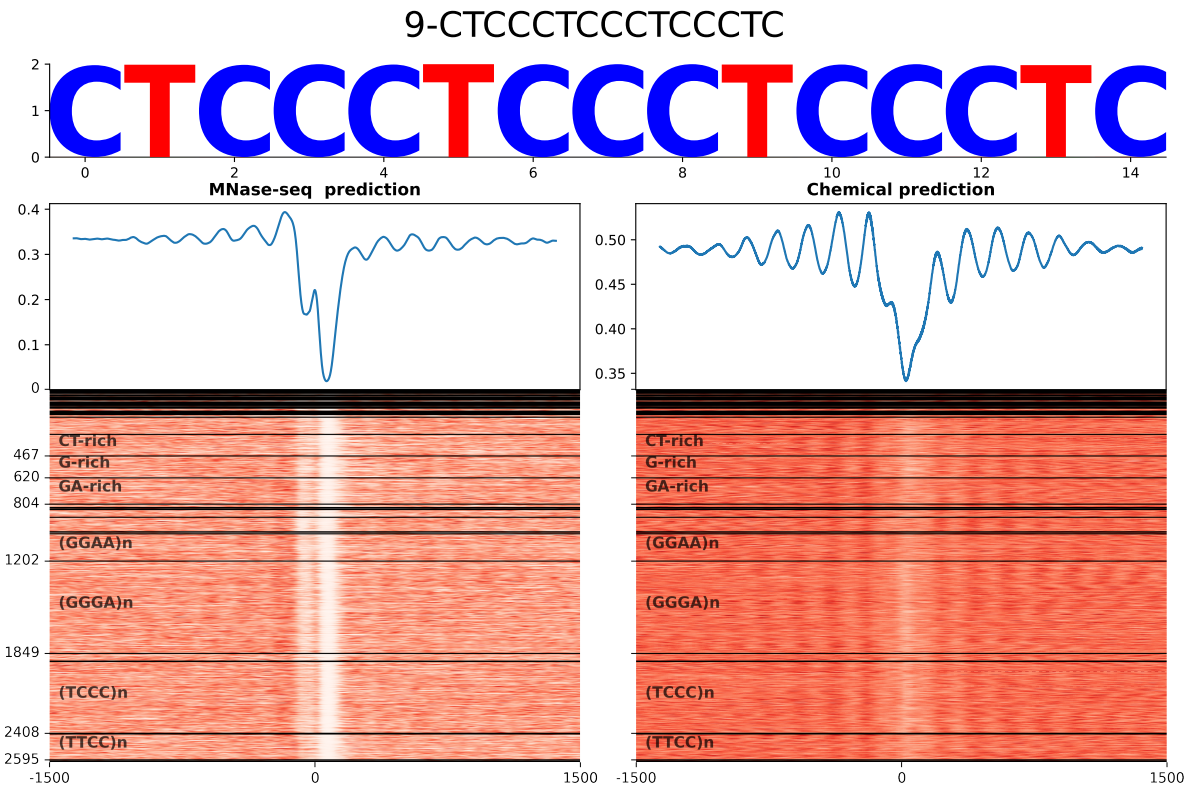


Figure 6.19: **STREME 9 motif**.

6.2.4 Perspective on G4-like motifs

G-quadruplex (G4) are non-canonical nucleic acid structures formed in guanine-rich regions of DNA or RNA. They consist of stacks of guanine tetrads—planar arrangements of four guanine bases connected through Hoogsteen hydrogen bonds—stabilized by monovalent cations such as potassium [161]. The canonical sequence motif capable of forming a G-quadruplex is

$$G_{\geq 3}N_{1-7}G_{\geq 3}N_{1-7}G_{\geq 3}N_{1-7}G_{\geq 3}$$

where $G_{\geq 3}$ represents a stretch of at least three consecutive guanines and N_{1-7} denotes a loop of one to seven arbitrary nucleotides. G-quadruplexes can form in promoter regions, telomeres, and untranslated regions, where they are thought to play regulatory roles in processes such as transcription, replication, and translation.

Interestingly, Figure 6.20 shows that G4-like motifs are associated with alterations in nucleosome occupancy. Both experimental and predicted MNase-seq profiles exhibit a NDR centered on the G4-like motif, consistent with the idea that these structures act as physical or functional barriers to nucleosome positioning. However, no phased nucleosomes can be observed around it.

In addition, the ISM signal peaks sharply at the center of the motif, indicating that the model has learned to associate these sequences with a high-impact disruption of the

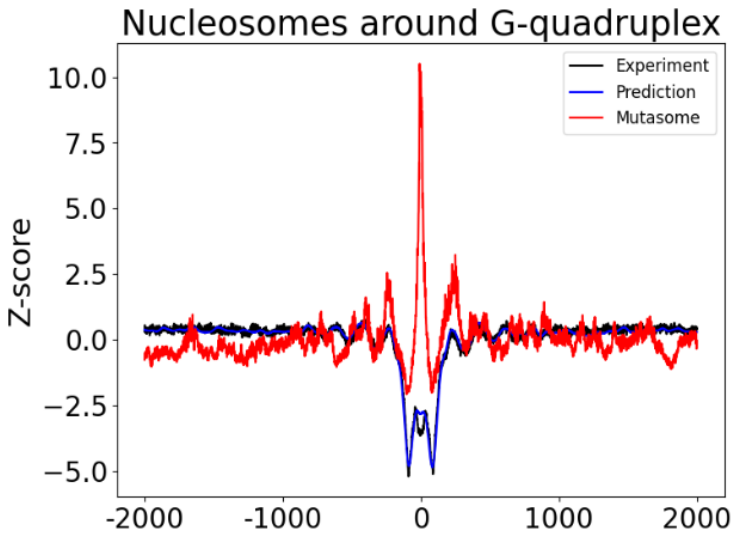


Figure 6.20: **Nucleosome occupancy and ISM around G4-like motifs.**Prediction has been made with MNase-seq model on retrieved G4 motif genome wide

predicted nucleosome landscape. The strong concordance between experimental data, model predictions, and ISM analysis supports the biological relevance of G4-like motifs as local chromatin organizers.

Chapter summary

This chapter reveals that repetitive and low-complexity DNA sequences are not passive genomic passengers, but active participants in chromatin organization. Using ISM approach, we identified SINEs, LINEs, LTRs, and microsatellites as nucleosome positioning hotspots, each contributing distinct architectural logic.

- SINEs (e.g., B2, B3) embed CTCF sites.
- LTRs and ERVs encode high-information motifs influencing nucleosome spacing and accessibility.
- LINEs exhibit internal sequence periodicity matching nucleosome arrays and provide a reservoir of TFBS.
- Microsattellites act as chromatin barrier either by their exotic intrinsic physical properties yet still carry potential TFBSs.

Together, these data suggest a model where genomic repeats double as organizational elements, with functional implications for chromatin state, transcription factor binding, and genome evolution.

In this chapter, we demonstrated that convolutional neural networks trained on DNA sequence alone can reproduce *in vivo* nucleosome occupancy profiles with high accuracy. Using MNase-seq and chemical cleavage data, the models achieved correlations up to 0.75 on held-out genomic regions and captured both global occupancy patterns and fine-scale features such as phased arrays around CTCF sites. Importantly, the predictions generalize beyond mappable regions, providing meaningful nucleosome profiles even in repeat-rich domains where experimental signals are degraded.

Beyond predictive performance, the models internalized both biological determinants of nucleosome organization and the specific biases of the underlying experimental assays. Together, these results show that local sequence context is sufficient to encode much of the nucleosome positioning landscape, while also highlighting limitations linked to long-range chromatin interactions that fall outside the receptive field of our architecture. This chapter therefore establishes the model as a reliable tool for exploring nucleosome positioning genome-wide and provides the foundation for the subsequent motif- and repeat-focused analyses.

Chapter 7

Leveraging neural network potential: synthetic genomics

Nucleosome Positioning score (NPscore)

To better understand how the model behaves regarding the nature of the sequence, we defined a nucleosome positioning score defined as follow:

$$\text{NPscore} = \sum_{i=1}^{L} |S(i) - B(i)| \,.$$

where $S(i)$ denotes the signal of modified background at position i, $B(i)$ the background signal at position i, and L the length of the region under consideration.

7.1 Deciphering CTCF motif

7.1.1 ISM highlights robust CTCF motifs

While CTCF motif is highly represented in the genome, not all its TFBS are positioning nucleosomes. Interestingly, the ISM can discriminate positioning motifs from the CTCF sites without phased nucleosome (Figure 5.12). This also correlates with the binding of the protein (Figure 5.13). As the model can differentiate the behaviour around each site, we can suppose that the binding determinant of CTCF is encrypted in the DNA sequence. To further investigate the behaviour of the model regarding CTCF motif, we ran *in silico* experiments consisting of mutating CTCF motifs.

The first experiment aimed to probe the model's focus and better understand how it identifies CTCF motifs. Figure 7.1 shows the effect of CTCF motifs, selected from the genome, when inserted into different random backgrounds. Each motif was paired with 500 randomly generated background sequences that respect the global GC-content of the mouse genome. CTCF motifs were splitted in two groups using the ISM consensus: CTCF motifs flagged as NPR (ISM +) and motifs not considered as NPR (ISM -)

(see Appendix B). The heatmaps reveal a clear nucleosome positioning across most backgrounds and motifs for ISM-positive group, in both models. In contrast, positioning fades for ISM-negative motifs, independently of the background. This demonstrates that the model does not only rely on the surrounding sequence context to recognize CTCF, but rather captures a specific sequence-encoded signature intrinsic to the motif itself. On top of that, one can notice that the logoplots differ slightly from each other. For instance, the C at position 19 is more frequent in ISM-tagged motifs, whereas the U-motif at position 1,3,4 are less variable in ISM-low motifs (respectively T, C, A).

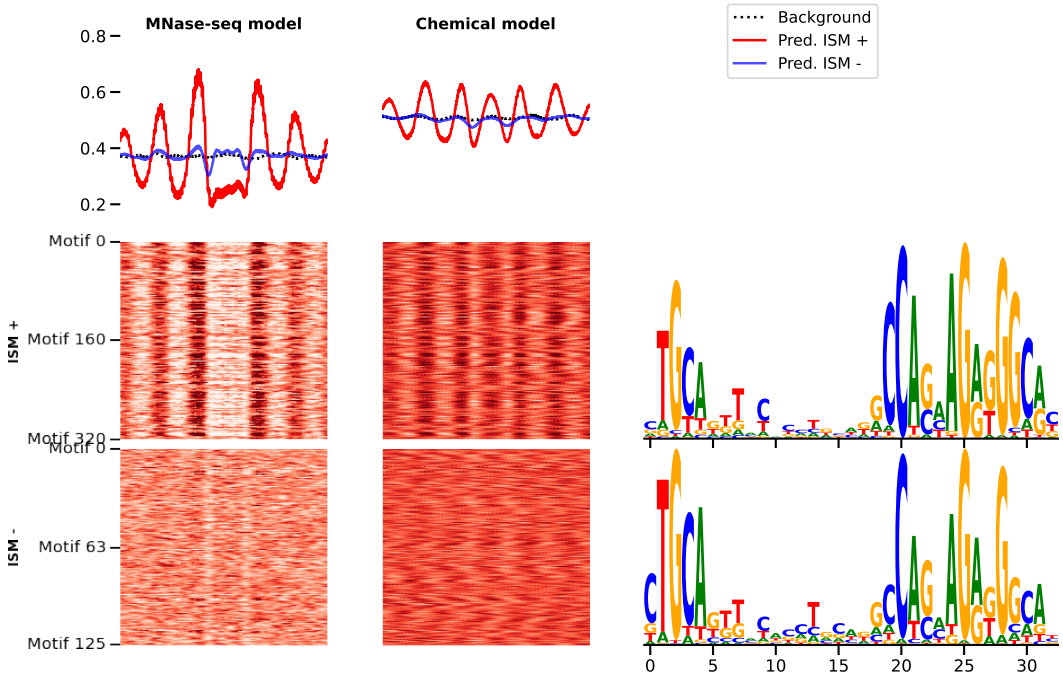
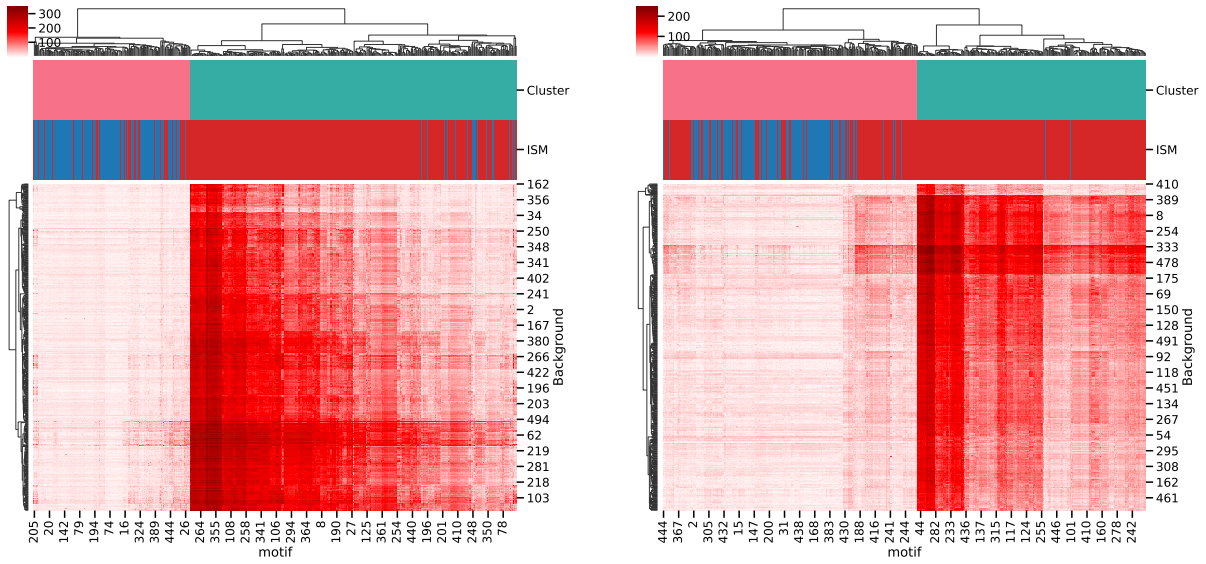


Figure 7.1: **Effect of full motif insertion on random backgrounds with respect to the ISM score.** CTCF motifs identified within NPRs were inserted into random background sequences, forming the *ISM+* category. CTCF motifs located outside NPRs were similarly inserted, forming the *ISM–* category. Both sets of motifs were tested on the same 500 randomly generated background sequences.

To assess whether the positioning ability of CTCF motifs depends on the surrounding sequence context, we clustered the NPscore values obtained for each motif inserted into multiple random background sequences (Figure 7.2). The clustering was performed using Bray–Curtis distance and the UPGMA algorithm. Both the MNase-seq model (Figure 7.2a) and the chemical-cleavage model (Figure 7.2b) reveal two distinct clusters of non-positioning CTCF motifs across backgrounds: one with consistently low NPscore on all backgrounds (cluster 1), and another with variable NPscore depending on the background (cluster 2). Interestingly, some CTCF motifs position robustly on any background, but most motifs in cluster 2 show positioning only on specific backgrounds, highlighting

that the nucleosome positioning potential of CTCF motifs is highly context-dependent.



(a) MNase-seq model

(b) Chemical cleavage model

Figure 7.2: **Clustering of NPscore values for CTCF motifs across random backgrounds.** Each cell represents the NPscore predicted when inserting a given CTCF motif into a specific random background sequence. Rows correspond to background sequences and columns to CTCF motifs. Both axes were ordered using hierarchical clustering based on Bray–Curtis distance and the UPGMA algorithm, grouping motifs and backgrounds with similar NPscore profiles. Two main motif clusters were defined from the first bifurcation of the hierarchical tree: one cluster (cyan) shows consistently low NPscore values across all backgrounds, indicating motifs unable to position nucleosomes robustly; the second cluster (pink) displays variable NPscore depending on the background, suggesting that nucleosome positioning by these motifs is context-dependent. In both models — (a) MNase-seq and (b) chemical cleavage — this separation between low and variable NPscore clusters is conserved, although the distribution and magnitude of NPscore differ between the two assays.

To determine whether the two clusters identified previously reflect an intrinsic ability to position nucleosomes, we compared them with the ISM categories defined from motif insertion experiments (Figure 7.3). ISM+ motifs correspond to CTCF motifs originally located within NPRs, whereas ISM− motifs correspond to those outside NPRs. A strong association is observed between cluster identity and ISM category, especially for the MNase-seq model. For MNase-seq, cluster 1 is largely depleted of ISM+ motifs (35 ISM+ vs 286 ISM−), while cluster 2 shows the opposite pattern (110 ISM+ vs 16 ISM−), yielding a strong correlation (MCC = 0.73, Fisher's exact test $p = 6.4 \times 10^{-55}$). For the chemical-cleavage model, the same trend is observed but with a weaker effect (112 ISM+ vs 209 ISM− in cluster 1, and 124 ISM+ vs 2 ISM− in cluster 2; MCC = 0.57, $p = 8.4 \times 10^{-41}$).

These results indicate that CTCF motifs found within NPRs are predominantly associated with cluster 2, which corresponds to motifs capable of positioning nucleosomes in a context-dependent manner, whereas motifs outside NPRs mostly belong to cluster 1

and fail to position nucleosomes regardless of the surrounding sequence. This correspondence shows that the ISM analysis effectively discriminates between motifs that can promote nucleosome positioning and those that cannot, capturing their intrinsic and context-dependent potential. In mouse embryonic stem cells, the MNase-seq model displays this relationship most strongly, highlighting that ISM-derived scores provide a reliable indicator of nucleosome positioning ability.

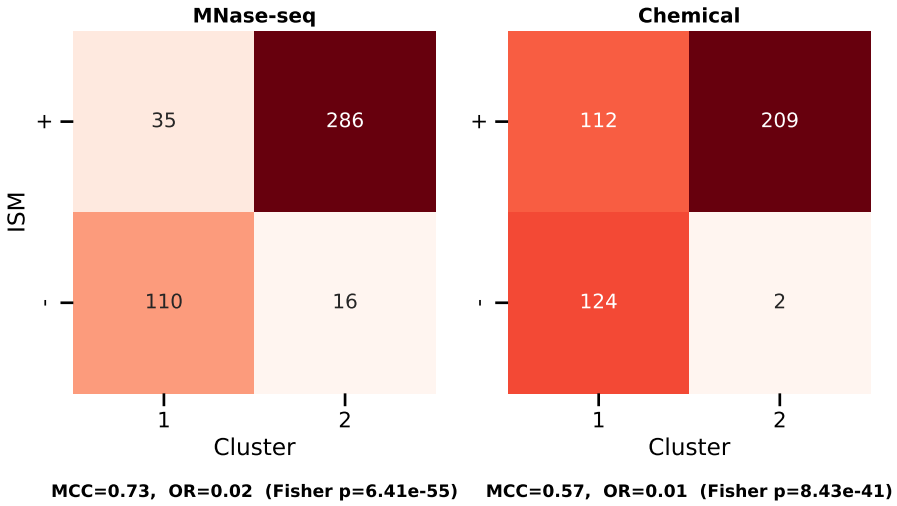


Figure 7.3: **Association between ISM category and NPscore-based clusters of CTCF motifs.** Contingency matrices comparing the cluster assignment (from NPscore-based hierarchical clustering) with the ISM category (defined by whether the original motif overlaps a nucleosome positioning region, NPR). (a) For the MNase-seq model, cluster 1 is mostly composed of ISM− motifs, whereas cluster 2 contains mostly ISM+ motifs, indicating a strong correlation (Matthews correlation coefficient, MCC = 0.73; odds ratio, OR = 0.02; Fisher's exact test $p = 6.4 \times 10^{-55}$). (b) For the chemical-cleavage model, the same association is observed but with a lower strength (MCC = 0.57; OR = 0.01; $p = 8.4 \times 10^{-41}$). The MCC quantifies the overall agreement between ISM category and cluster label (+1 = perfect correlation, 0 = random), while the odds ratio (OR) measures the enrichment of ISM+ motifs within cluster 2 relative to cluster 1. These results show that motifs forming the high-NPscore cluster are generally those originally located within NPRs.

7.1.2 The CTCF upstream motif reinforce CTCF nucleosome phasing ability

A second experiment was designed to evaluate the impact of the CTCF core motif and its upstream (U) motif on nucleosome organization over a single randomized background. As shown previously and in Figure 7.4, there appear to be two populations of CTCF motifs: one that strongly positions nucleosomes regardless of background, and another whose positioning ability depends on the surrounding sequence. For some motifs, insertion of the core motif alone is sufficient to induce phased nucleosome arrays around the insertion site. Closer inspection of the scatterplots (Figure 7.4B,C) reveals that, on this background,

the U-motif has a particularly strong impact on moderately positioning CTCF sites, while it exerts little to no influence on motifs with either very strong or very weak positioning ability.

These results are consistent with the known stabilizing role of the U-motif, which enables CTCF to engage two additional zinc fingers, thereby enhancing its DNA-binding affinity and chromatin residence time [162], and potentially reinforcing its barrier effect on nucleosome positioning.

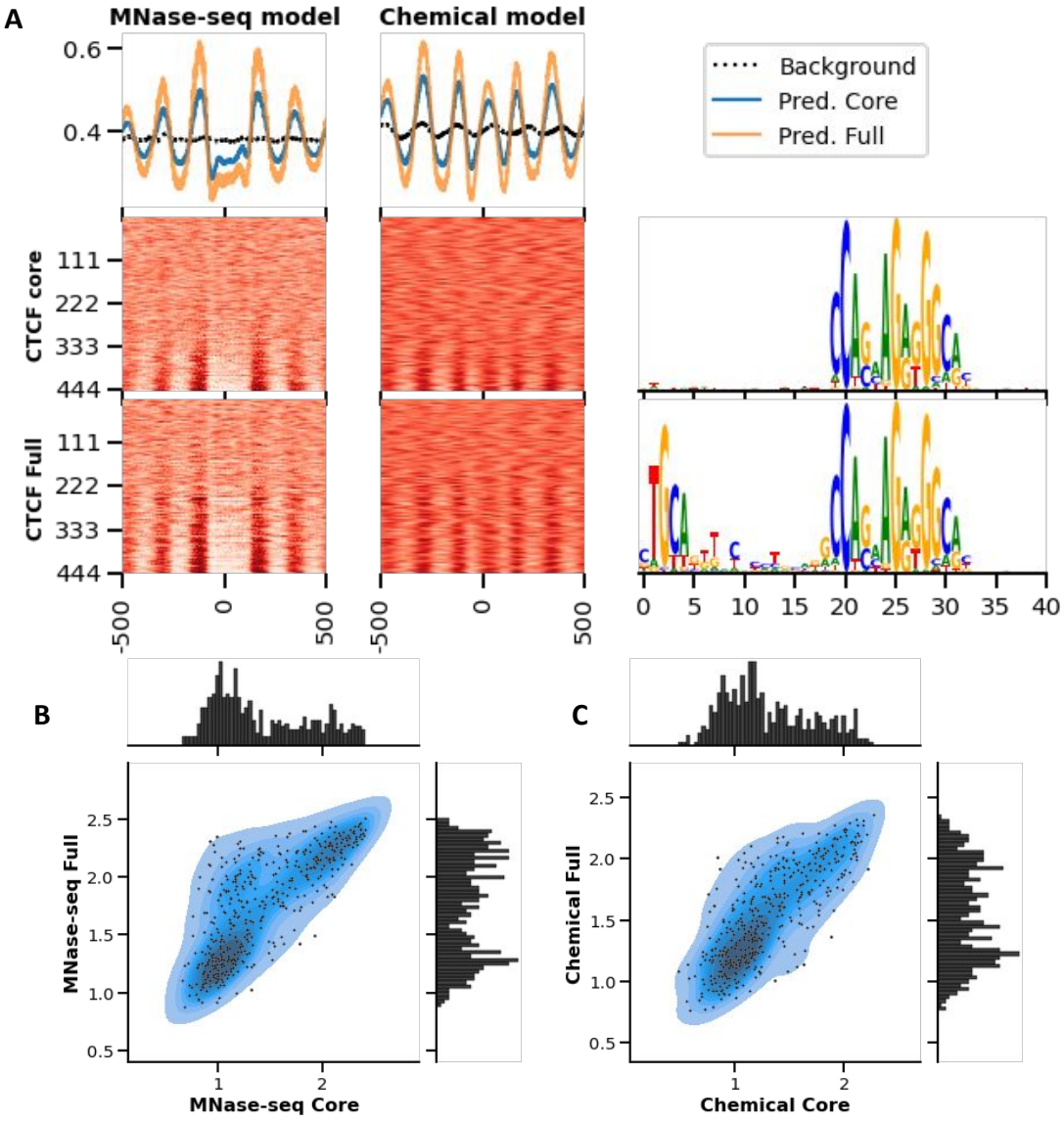
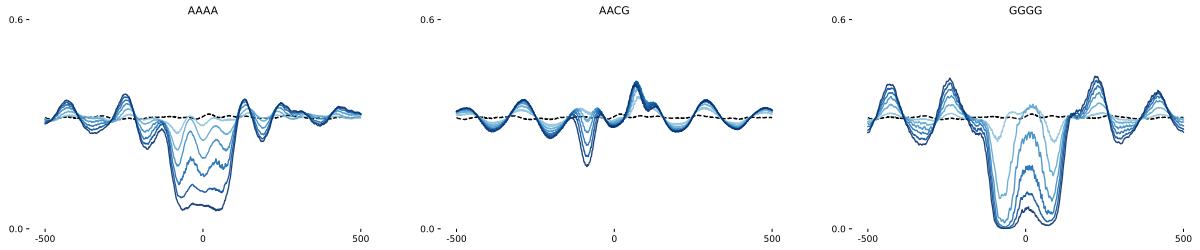


Figure 7.4: **A. Effect of core motif and full motif inserted on single random background.** CTCF position nucleosome on more background when the core motif is preceeded by the U-motif. **B,C. Scatterplot of NPscores for MNase-seq and Chemical model.** the scatter plots indicate the variation of the NPscore regarding the core and full condition

7.2 Specific microsatellites act as chromatin barrier



(a) AAAA (b) AACG (c) GGGG

Figure 7.5: **In silico insertion of tandem repeats on random sequence**. Aggregated plot of 1000 random sequence with *in silico* insertion of 4-mers repeated [2, 4, 6, 8, 10, 12] times in shades of blue. Background control without insertion is represented in black dotted line.

To better understand how microsatellites can influence nucleosome positioning, we generated synthetic microsatellites as repeated 4-mers (repetitions range from 2 to 14), that we inserted on random backgrounds. The network predictions are consistent across reverse-complementary and cyclically equivalent 4-mers, which are strictly identical when repeated in tandem. Figure 7.5 shows three representative examples from this experiment. The aggregated signal over background sequences displays no nucleosome positioning pattern. In contrast, the insertion of a single 4-mer in tandem gradually induces a positioning signal, which becomes stronger as the number of repeated monomers increases. However all 4-mers do not have a clear effect on the nucleosome positioning as shown in Figure 7.5b.

Figure 7.6 shows that the NPscore is higher for the extreme values of the GC content of the repeated 4-mers. The result is concordant for both model with few outliers for chemical cleavage model at 50% GC-content. This result confirm what we observed both in experimental data and prediction regarding GC-content: sequences with extreme GC-content values act as chromatin barrier and are flagged as NPR by both models (Figure 6.15).

For tandem repeats composed entirely of longer monomers, the relationship with GC content differed, as described in the following section.

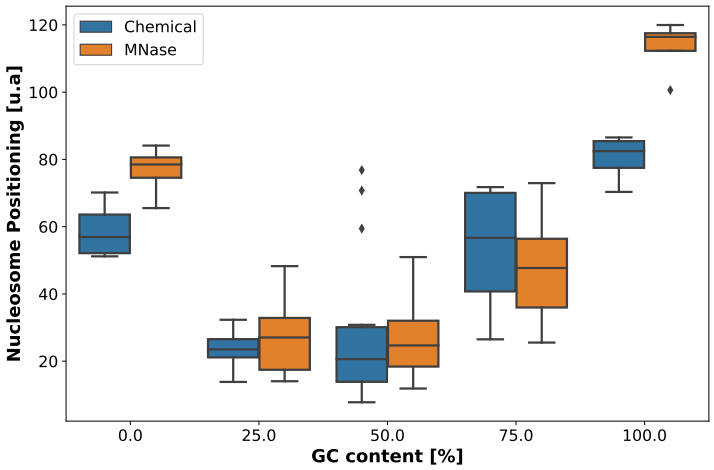


Figure 7.6: **Nucleosome positioning score of synthetic 4-mers repeats regarding their GC content** from MNase-seq model (orange) and Chemical cleavage model (blue)

7.3 Tandem repeats

Tandem repeats are predominantly located in centromeric and pericentromeric regions [60]. Figure 7.7 shows the predicted signal amplitude for synthetic tandem repeats of varying monomer sizes and GC contents. The distribution is largely unaffected by monomer size, but strongly dependent on GC content. At extreme values of GC, predictions flatten to low amplitude, particularly in the MNase-seq model. For this model, signals are essentially flat for GC contents above 0.7 and below 0.3. The chemical-cleavage model shows a more moderate decline, with flat profiles at 100% GC and plateaus for GC contents below 0.2.

Within the intermediate GC range [0.3–0.6], the distributions are broad: some repeats show strong nucleosome positioning, while others lack any clear positioning signature. Predictions for natural tandem repeats such as minor satellites (MinSat), major satellites (MajSat), and the L1 monomer (Figure 6.14) fall toward the lower end of this distribution.

In summary, although extreme GC contents lead to very restricted amplitude distributions, tandem repeats of varying lengths can nevertheless support a wide range of nucleosome positioning patterns.

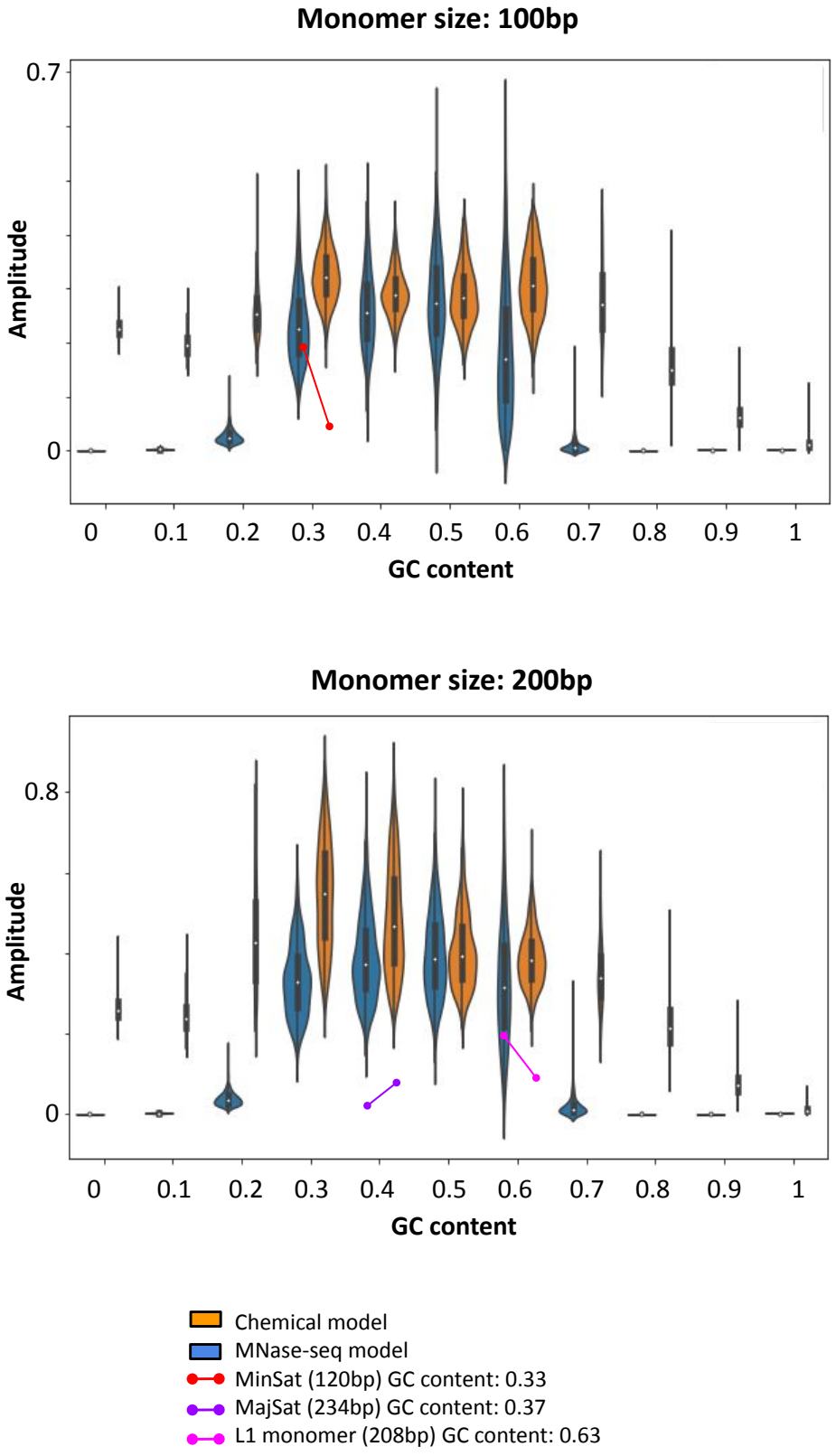


Figure 7.7: **Amplitude distribution of predicted signal on synthetic and natural tandem repeats**

Chapter summary

In this chapter, we leveraged the predictive power of our neural network in synthetic settings to probe sequence determinants of nucleosome positioning.

- ISM confirmed that the model discriminates between positioning and non-positioning CTCF motifs, and highlighted the stabilizing role of the upstream U-motif in reinforcing nucleosome phasing.
- Synthetic microsatellites revealed that extreme GC content generates strong chromatin barrier effects.
- Synthetic tandem repeats with longer monomers support diverse nucleosome patterns at intermediate GC content, whereas natural tandem repeats display poor nucleosome positioning.

Together, these experiments demonstrate that the model captures not only canonical transcription factor binding signals but also intrinsic sequence properties of repetitive DNA. They also illustrate how CNNs can be used as an *in silico* sandbox for biological exploration.

Chapter 8

Discussion

8.1 Convolutional Neural Networks capture complex interplay inside gene-sparse genomes

We showed that CNN can predict nucleosome positioning on large genomes with a fidelity comparable to that observed in gene-dense regions [8]. At a finer scale, the models reproduced well-established biological features, such as the characteristic nucleosome pattern around acknowledged nucleosome-positioning TFBS like CTCF.

Although predicted nucleosome occupancy is less well defined at TSS, the models were still able to predict a NDR. In addition, NPR were enriched in compartment A, indicating that the models integrate (at least partially) the transcriptional landscape of the cell. While the presence of an NDR is required for the transcriptional machinery to initiate transcription, the positioning and dynamics of nucleosomes at TSS remain unresolved [163–165].

Notably, these studies have shown that average nucleosome occupancy profiles around TSS can obscure substantial gene-to-gene variability. This heterogeneity likely reflects the asynchronous transcriptional states across individual cells in bulk populations, as well as dynamic remodeling processes that are not captured in static measurements. As a result, the apparent "fuzziness" in predicted occupancy around TSS may not necessarily reflect model failure, but rather the biological variability embedded in the training data. Furthermore, the determinants of nucleosome positioning at these loci may involve trans-factors or chromatin remodelers acting at distances beyond the receptive field of our model (2001 bp), or may even be independent of the DNA sequence itself.

8.2 Large genomes are punctuated with Nucleosome Positioning Regions

The ISM method demonstrated strong robustness, with high correlation between the two models. This approach yielded a sufficient number of precise nucleosome determinants (∼600k), regularly spaced across the genome with an average distance of 4 kb, to account

for nucleosome positioning at a genome-wide scale, consistent with the barrier model in large genomes such as mouse [2]. The study of these NPR revealed both known TFBS and previously uncharacterized motifs, each with the ability to position nucleosomes in their vicinity.

8.3 ISM confirms previously identified simple features of nucleosome positioning

Some di-, tri-, and even tetra-nucleotides were predicted as favoring or disfavoring nucleosome formation [119, 166, 167]. Here, we demonstrate that locally highly repetitive sequences can act as chromatin barriers and influence nucleosome distribution in their vicinity through their intrinsic nature, particularly regarding their GC content.

We further confirmed previous models indicating poly-A/T arrays as nucleosome-excluding sequences [119, 167], and demonstrated that these sequences act as chromatin barriers—positioning phased arrays of nucleosomes around a NPR—and that this effect is not linked to the binding of ZNF384. Leveraging synthetic genomics, we generalized this rule by showing that any 4-mer with extreme GC content (0% or 100%) acts as a chromatin barrier independently of the surrounding sequence. However, it is important to note that GC-rich sequences correlate with biological features that may influence nucleosome positioning beyond the raw mechanical properties of the DNA itself. Among these are G-quadruplexes, CpG islands, TSSs, and TFBSs such as the G-box of SP/KLF motifs, and in those case we exhibited NDR whereas the litterature suggested affinity in GC-rich sequences.

8.4 ISM precisely identifies transcription factor binding sites

Beyond these simple features, we identified canonical TFBS as nucleosome positioning factors. The KLF/SP family appears as a novel nucleosome-positioning factor *in vivo*. While the precise transcription factor bound at each occurrence cannot be unambiguously inferred from sequence alone, the strong similarity to well-characterized factors such as KLF4 supports the idea that pioneer-like properties underlie this positioning activity. These observations place SP/KLF-like motifs among the recurrent sequence features that shape nucleosome landscapes. Notably, this nucleosome organization mirrors that observed at KLF4 binding sites in mouse embryonic stem cells, where KLF4 binding is compatible with nucleosome occupancy at the motif itself and leads to phasing of adjacent nucleosomes. The recovery of this motif by Soufi et al. in *in vitro* pioneer factor binding assays [137] further supports the notion that G-box KLF/SP sites can serve as anchoring points for nucleosome positioning. Although the KLF/SP family is widely associated with GC-rich binding motifs, its direct contribution to nucleosome phasing *in vivo* has remained poorly documented. Our results provide indirect yet robust evidence that

these motifs actively contribute to phased arrays of nucleosomes, thus extending their role beyond canonical transcriptional regulation.

Beyond KLF/SP, our analyses retrieved additional motifs. Both *de novo* discovery and JASPAR-guided searches consistently recovered Thap11 (Ronin) and Dux, two factor that are still poorly described in the context of nucleosome positioning but acknowledged as essential mouse development factors, involved in pluripotency [150, 153], alongside a collection of well-established pluripotency regulators, including YY1, Nanog, Oct4, Sox family members, and the Oct4:Sox2 heterodimer [3, 131]. These observations underscore how deeply the interplay between nucleosome positioning and DNA sequence is embedded within the regulatory network of embryonic stem cells, to the point that the architecture alone is able to capture essential determinants of pluripotency.

Taken together, the influence of local sequences and transcription factors further confirms a duality in nucleosome-positioning rules. The sequence itself positions nucleosomes, but this influence can be overcome by external factors such as DNA-binding proteins, higher-order chromatin folding, or chromatin remodelers.

8.4.1 CTCF has a unique dynamic yet encoded in its motif

In line with numerous reports of its ability to shape chromatin and displace nucleosomes [3,31,143,168,169], CTCF was strongly and recurrently recovered in our sequence analyses. Strikingly, the models could discriminate DNA-bound CTCF sites with high precision, suggesting that NNs may help resolve such challenging loci and extend predictive scope beyond nucleosome positioning to broader protein–DNA interactions.

Unlike Romero *et al.*, who inferred binding preferences by training directly on TF-specific ChIP-seq data [170], our approach identifies such determinants indirectly from nucleosome landscapes. This demonstrates that interpretable models trained on chromatin organization can recover generalizable regulatory logic, rather than being restricted to single datasets or factors.

Synthetic genomics reinforced these findings by revealing the importance of the Umotif in B3 SINE–derived CTCF TFBSs, which is absent from B2_Mm2 SINEs despite both carrying a CTCF core motif [7]. Schmidt *et al.* linked this compositional difference to evolutionary age, with B3 elements predating B2 [7, 171]. Our results are consistent with a loss of the U-motif in murine B2 while it has been retained in B3. Since both core motifs are represented by the same PWM, this highlights intra-motif dependencies invisible to conventional motif models and points to a scenario of parallel evolution between B3 and B2_Mm2.

A key determinant of CTCF function is its residence time, which depends on the protein's ability to engage its eleven zinc fingers [151]. The upstream auxiliary motif enables binding of additional fingers, thereby extending residence time [145]. This may in turn relax selective pressure on the core motif, providing an evolutionary explanation for the divergence observed between B2_Mm2 and B3 SINEs while preserving overall CTCF

occupancy.

This hypothesis of coexisting CTCF configurations is also consistent with the observation that the consensus motif alone poorly predicts TADs borders. Because CTCF competes with other transcription factors that recognize highly similar motifs [172, 173], even subtle motif variations or intra-motif nucleotide dependencies may be functionally relevant—an aspect that our network appears to capture.

Finally, CTCF motifs found in unique genomic regions coexist as both solo-core and U+core variants. Whether these motifs ultimately derive from ancient SINEs remains an open question. Nonetheless, our networks prove capable of detecting repeated-element–derived motifs with high sensitivity and thus hold promise for addressing broader evolutionary questions on the origin and diversification of regulatory sequences. In this sense, interpretable neural networks may open a new path beyond consensus-based annotations, offering a dynamic way to trace the evolutionary logic of regulatory sequences

8.4.2 Deep investigation of transposable elements reveals nucleosome positioning regions

The ability of deep learning to overcome mappability limitations allowed us to extend our analysis to repeated elements. Beyond the well-described CTCF-harboring SINEs, we highlighted the contribution of LTRs, which are known regulators and, as shown here, also act as nucleosome-positioning sequences. LINEs elements not only carry established TFBSs, but also harbor conserved motifs without clear database matches, which nevertheless colocalize with regularly phased nucleosome arrays.

These observations align with the recognized role of repeats in shaping chromatin structure at large genomic scales [21], while also underscoring their regulatory impact when co-opted as cis-regulatory elements [20]. For transposable elements that are normally repressed, nucleosomes provide a first line of defense by bearing repressive epigenetic marks and preventing their activation. Yet, their numerous embedded TFBSs and promoters remain poised to enter regulatory networks if silencing is relaxed, for instance through the loss of repressive methylation [55]. In this sense, nucleosome positioning emerges both as a mechanism of genome stability and as a substrate for regulatory innovation.

Whether recruited into regulatory circuits or silenced as genomic parasites, our results show that transposable elements consistently act as nucleosome-positioning factors. They should therefore be regarded not merely as passive passengers, but as active architectural players in the organization of the chromatin landscape.

8.4.3 Simple repeats but active chromatin shaping actors

In the absence of external factors, nucleosomes tend to position according to sequence preferences. Our synthetic microsatellite experiments confirmed this principle: short tandem repeats with extreme GC content acted as strong chromatin barriers, robustly inducing phased nucleosome patterns. In contrast, natural tandem repeats in the genome showed only weak phasing compared to synthetic constructs. This discrepancy highlights the complexity of genomic repeats, which may balance opposing sequence signals within their monomers, be shaped by higher-order chromatin contexts absent from synthetic settings, or be blurred by technical challenges such as mapping ambiguities in highly repetitive regions. Together, these observations suggest that synthetic assays expose the "pure" sequence rules of nucleosome positioning, whereas natural repeats reveal how such rules are modulated in vivo. Further work will be required to disentangle these possibilities.

It is still important to remind that we found putative TFBSs on microsatellites and even cross poly-As track with ZNF380 chip-seq, showing that they can also engage in interaction with proteins. Moreover some studies suggest that microsatellites would be transcribed [174], which would match the nucleosome patterns observed in the vicinity of simple repeats. Thus, microsatellites should not be viewed merely as passive sequence oddities: their exotic base composition shapes nucleosome organization, while their ability to recruit proteins or even undergo transcription highlights their potential as active regulatory actors within the chromatin landscape.

8.5 Where models agree, biology emerges

Strikingly, independently trained neural networks converge toward highly concordant predictions, exhibiting stronger agreement with each other than the experimental assays themselves. This is observed even though each model faithfully learns and reflects the specific biases of its training assay, whether MNase-seq or chemical cleavage, which were generated in different laboratories and rely on distinct biochemical principles to probe nucleosome positioning.

This convergence is not restricted to canonical genomic regions but is also observed over natural L1 tandem repeats, where nucleosomes are consistently phased across predictions. These repetitive regions are known to exacerbate protocol-specific biases, making the agreement between models particularly remarkable.

Importantly, the models do not ignore experimental artifacts; rather, they internalize them, as illustrated by their ability to capture assay-specific features such as fragile nucleosomes. While model predictions retain method-dependent characteristics, they nonetheless show higher mutual concordance than the assays themselves.

In silico mutagenesis further reveals that this convergence is even more pronounced at the level of underlying sequence determinants. Models trained on distinct assays rely

on highly similar sets of nucleosome-positioning–relevant regions (NPRs), indicating that they internalize a shared, sequence-encoded logic of nucleosome organization.

Consequently, deep learning models can generate assay-specific predictions while being grounded in common biological determinants. This dissociation between prediction space and explanatory space highlights their ability to disentangle reproducible biological signals from method-dependent bias and noise.

More broadly, these results indicate that increasing dataset size alone is insufficient to resolve chromatin organization. Progress instead relies on the integration of complementary experimental modalities, whose convergence of viewpoints enables predictive models to anchor their representations in robust and reproducible biology.

8.6 Limits and perspectives

Although our study was trained on experimental data from different laboratories, most conclusions rely on statistical computation or *in silico* experiments. A first limitation is that the two datasets were used separately for training. A unified model trained jointly on both would likely converge toward a shared set of weights, capturing generalizable features while minimizing dataset-specific biases. For example, fragile nucleosomes are robustly captured in chemical cleavage but erased in MNase-seq, and our models reproduce this difference. A unified model trained jointly on both assays may help reconcile such discrepancies, disentangling genuine biology from assay-specific biases.

A second limitation lies in the 2,001-bp receptive field of our networks. While some long-range effects may be partially encoded in local sequence features (e.g., repeated elements or CTCF motifs acting as TAD borders), higher-order chromatin interactions or distal regulatory mechanisms remain inaccessible to the current model. Addressing this will require extending input lengths, integrating external contact maps and multi-omics data, or developing new representations of chromatin beyond the primary DNA sequence.

A third limitation concerns motif discovery. Tools such as STREME, MEME, and the broader XSTREME suite depend on statistical enrichment relative to background models, making them sensitive to the choice of controls and compositional biases. Moreover, their reliance on PWMs assumes positional independence and cannot capture higher-order dependencies or DNA structural features. As emphasized by the authors of these methods, enrichment-based motif discovery is intrinsically constrained by data resolution and by the presence of repetitive or low-complexity regions, which may inflate significance or obscure meaningful motifs. Reported motifs should therefore be interpreted cautiously and ideally validated with orthogonal experimental approaches [175].

A fourth limitation is the cell-type specificity of our training data. Because the models were trained exclusively on mouse embryonic stem cells, some extracted features may reflect stem-cell–specific determinants of chromatin organization rather than universal nucleosome-positioning principles. Predictions on synthetic constructs, in particular, may not directly transfer across cell types with different remodeler activity or regulatory programs. Future studies should evaluate model generalizability across diverse contexts and species.

Finally, new nucleosome mapping techniques—especially long-read and single-cell approaches—offer promising avenues to overcome current limitations. By resolving dynamic aspects of chromatin organization and mitigating short-read mapping biases, these emerging datasets could be integrated with deep learning models to provide a more complete and mechanistic understanding of nucleosome positioning.

8.7 Conclusion

As biology becomes as computational as it is experimental, progress will increasingly depend on the integration of expertise across disciplines, as recent studies demonstrate [125, 176]. Machine learning is not a replacement for *in vivo* experimentation, but rather an *in silico* sandbox in which hypotheses can be generated, prioritized, and explored at low marginal cost. By recovering biological signal from noisy data, such approaches complement experimental work. This thesis illustrates how interpretable deep learning can reproduce established principles of chromatin organization while uncovering novel determinants of nucleosome positioning, providing both mechanistic insight and experimentally testable predictions.

Bibliography

[1] Anton Valouev, Steven M. Johnson, Scott D. Boyd, Cheryl L. Smith, Andrew Z. Fire, and Arend Sidow. Determinants of nucleosome organization in primary human cells. *Nature*, 474(7352):516–520, 2011. Number: 7352 Publisher: Nature Publishing Group.

[2] Travis N. Mavrich, Ilya P. Ioshikhes, Bryan J. Venters, Cizhong Jiang, Lynn P. Tomsho, Ji Qi, Stephan C. Schuster, Istvan Albert, and B. Franklin Pugh. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research*, 18(7):1073–1083, 2008.

[3] Nicola Festuccia, Nick Owens, Thaleia Papadopoulou, Inma Gonzalez, Alexandra Tachtsidi, Sandrine Vandoermel-Pournin, Elena Gallego, Nancy Gutierrez, Agnès Dubois, Michel Cohen-Tannoudji, and Pablo Navarro. Transcription factor activity and nucleosome organization in mitosis. *Genome Research*, 29(2):250–260, 2019. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[4] Lilien N. Voong, Liqun Xi, Amy C. Sebeson, Bin Xiong, Ji-Ping Wang, and Xiaozhong Wang. Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. *Cell*, 167(6):1555–1570.e15, 2016.

[5] Charles E. Grant and Timothy L. Bailey. XSTREME: Comprehensive motif analysis of biological sequence datasets. Pages: 2021.09.02.458722 Section: New Results.

[6] Binbin Lai, Weiwu Gao, Kairong Cui, Wanli Xie, Qingsong Tang, Wenfei Jin, Gangqing Hu, Bing Ni, and Keji Zhao. Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature*, 562(7726):281–285, 2018.

[7] Dominic Schmidt, Petra C. Schwalie, Michael D. Wilson, Benoit Ballester, Angela Gonçalves, Claudia Kutter, Gordon D. Brown, Aileen Marshall, Paul Flicek, and Duncan T. Odom. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, 148(1):335–348, 2012.

[8] 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, 2023. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[9] Rosalind E. Franklin and R. G. Gosling. Evidence for 2-Chain Helix in Crystalline Structure of Sodium Deoxyribonucleate. *Nature*, 172(4369):156–157, July 1953. Publisher: Nature Publishing Group.

[10] Eugene V. Koonin and Artem S. Novozhilov. Origin and evolution of the genetic code: the universal enigma. *Iubmb Life*, 61(2):99–111, February 2009.

[11] Roger P. Alexander, Gang Fang, Joel Rozowsky, Michael Snyder, and Mark B. Gerstein. Annotating non-coding regions of the genome. *Nature Reviews Genetics*, 11(8):559–571, August 2010. Publisher: Nature Publishing Group.

[12] Miriam Land, Loren Hauser, Se-Ran Jun, Intawat Nookaew, Michael R. Leuze, Tae-Hyuk Ahn, Tatiana Karpinets, Ole Lund, Guruprased Kora, Trudy Wassenaar, Suresh Poudel, and David W. Ussery. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, 15(2):141–161, March 2015.

[13] Gilbert S. Omenn. Reflections on the HUPO Human Proteome Project, the Flagship Project of the Human Proteome Organization, at 10 Years. *Molecular & Cellular Proteomics : MCP*, 20:100062, February 2021.

[14] Arnau Sebé-Pedrós, Elad Chomsky, Kevin Pang, David Lara-Astiaso, Federico Gaiti, Zohar Mukamel, Ido Amit, Andreas Hejnol, Bernard M. Degnan, and Amos Tanay. Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nature Ecology & Evolution*, 2(7):1176–1188, July 2018. Publisher: Nature Publishing Group.

[15] Emma P. Bingham and William C. Ratcliff. A nonadaptive explanation for macroevolutionary patterns in the evolution of complex multicellularity. *Proceedings of the National Academy of Sciences*, 121(7):e2319840121, February 2024. Publisher: Proceedings of the National Academy of Sciences.

[16] Yves Van de Peer, Steven Maere, and Axel Meyer. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, 10(10):725–732, October 2009. Publisher: Nature Publishing Group.

[17] Claire Larroux, Graham N. Luke, Peter Koopman, Daniel S. Rokhsar, Sebastian M. Shimeld, and Bernard M. Degnan. Genesis and Expansion of Metazoan Transcription Factor Gene Classes. *Molecular Biology and Evolution*, 25(5):980–996, May 2008.

[18] Kathryn Phillips and Ben Luisi. The virtuoso of versatility: POU proteins that flex to fit. *Journal of Molecular Biology*, 302(5):1023–1039, October 2000.

[19] Megan Wilson and Peter Koopman. Matching SOX: partner proteins and co-factors of the SOX family of transcriptional regulators. *Current Opinion in Genetics & Development*, 12(4):441–446, August 2002.

[20] Edward B. Chuong, Nels C. Elde, and Cédric Feschotte. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics*, 18(2):71–86, 2017. Publisher: Nature Publishing Group.

[21] Axel Cournac, Romain Koszul, and Julien Mozziconacci. The 3d folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Research*, 44(1):245–255, 2023.

[22] Manfred Schartl, Joost M. Woltering, Iker Irisarri, Kang Du, Susanne Kneitz, Martin Pippel, Thomas Brown, Paolo Franchini, Jing Li, Ming Li, Mateus Adolfi, Sylke Winkler, Josane de Freitas Sousa, Zhuoxin Chen, Sandra Jacinto, Evgeny Z. Kvon, Luis Rogério Correa de Oliveira, Erika Monteiro, Danielson Baia Amaral, Thorsten Burmester, Domitille Chalopin, Alexander Suh, Eugene Myers, Oleg Simakov, Igor Schneider, and Axel Meyer. The genomes of all lungfish inform

on genome expansion and tetrapod evolution. *Nature*, 634(8032):96–103, October 2024. Publisher: Nature Publishing Group.

[23] G. V Guinea, F. J Rojo, and M Elices. Brittle failure of dry spaghetti. *Engineering Failure Analysis*, 11(5):705–714, October 2004.

[24] Peter Gross, Niels Laurens, Lene B. Oddershede, Ulrich Bockelmann, Erwin J. G. Peterman, and Gijs J. L. Wuite. Quantifying how DNA stretches, melts and changes twist under tension. *Nature Physics*, 7(9):731–736, September 2011.

[25] Guohui Zheng, Luke Czapla, A. R. Srinivasan, and Wilma K. Olson. How stiff is DNA? *Physical Chemistry Chemical Physics*, 12(6):1399–1406, January 2010. Publisher: The Royal Society of Chemistry.

[26] Vladimir B. Teif and Klemen Bohinc. Condensed DNA: Condensing the concepts. *Progress in Biophysics and Molecular Biology*, 105(3):208–222, May 2011.

[27] Jan Bednar, Isabel Garcia-Saez, Ramachandran Boopathi, Amber R. Cutter, Gabor Papai, Anna Reymer, Sajad H. Syed, Imtiaz Nisar Lone, Ognyan Tonchev, Corinne Crucifix, Hervé Menoni, Christophe Papin, Dimitrios A. Skoufias, Hitoshi Kurumizaka, Richard Lavery, Ali Hamiche, Jeffrey J. Hayes, Patrick Schultz, Dimitar Angelov, Carlo Petosa, and Stefan Dimitrov. Structure and Dynamics of a 197 bp Nucleosome in Complex with Linker Histone H1. *Molecular Cell*, 66(3):384–397.e8, May 2017.

[28] 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. Publisher: Nature Publishing Group.

[29] Vladimir B. Teif, Yevhen Vainshtein, Maïwen Caudron-Herger, Jan-Philipp Mallm, Caroline Marth, Thomas Höfer, and Karsten Rippe. Genome-wide nucleosome positioning during embryonic stem cell development. *Nature Structural & Molecular Biology*, 19(11):1185–1192, November 2012. Publisher: Nature Publishing Group.

[30] Daria A. Beshnova, Andrey G. Cherstvy, Yevhen Vainshtein, and Vladimir B. Teif. Regulation of the nucleosome repeat length in vivo by the DNA sequence, protein concentrations and long-range interactions. *PLOS Computational Biology*, 10(7):e1003698, 2014. Publisher: Public Library of Science.

[31] 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, 2019.

[32] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950):289–293, October 2009.

[33] Boyan Bonev and Giacomo Cavalli. Organization and function of the 3D genome. *Nature Reviews Genetics*, 17(11):661–678, November 2016. Publisher: Nature Publishing Group.

[34] Suhas S. P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and

Erez Lieberman Aiden. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680, December 2014. Publisher: Elsevier.

[35] Jesse R. Dixon, David U. Gorkin, and Bing Ren. Chromatin Domains: the Unit of Chromosome Organization. *Molecular cell*, 62(5):668–680, June 2016.

[36] Aline V. Probst, Elaine Dunleavy, and Geneviève Almouzni. Epigenetic inheritance during the cell cycle. *Nature Reviews Molecular Cell Biology*, 10(3):192–206, March 2009. Publisher: Nature Publishing Group.

[37] Julie Soutourina. Transcription regulation by the Mediator complex. *Nature Reviews Molecular Cell Biology*, 19(4):262–274, April 2018. Publisher: Nature Publishing Group.

[38] Yuichi Ichikawa, Nobuyuki Morohashi, Nobuyuki Tomita, Aaron P. Mitchell, Hitoshi Kurumizaka, and Mitsuhiro Shimizu. Sequence-directed nucleosome-depletion is sufficient to activate transcription from a yeast core promoter in vivo. *Biochemical and Biophysical Research Communications*, 476(2):57–62, July 2016.

[39] 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, 2017. Number: 9 Publisher: Nature Publishing Group.

[40] Andrew J. Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, March 2011.

[41] C. David Allis and Thomas Jenuwein. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics*, 17(8):487–500, August 2016. Publisher: Nature Publishing Group.

[42] Menno P. Creyghton, Albert W. Cheng, G. Grant Welstead, Tristan Kooistra, Bryce W. Carey, Eveline J. Steine, Jacob Hanna, Michael A. Lodato, Garrett M. Frampton, Phillip A. Sharp, Laurie A. Boyer, Richard A. Young, and Rudolf Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21931–21936, December 2010.

[43] Raphaël Margueron and Danny Reinberg. The Polycomb complex PRC2 and its mark in life. *Nature*, 469(7330):343–349, January 2011.

[44] J. Nakayama, J. C. Rice, B. D. Strahl, C. D. Allis, and S. I. Grewal. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science (New York, N.Y.)*, 292(5514):110–113, April 2001.

[45] Anja Ebert, Sandro Lein, Gunnar Schotta, and Gunter Reuter. Histone modification and the control of heterochromatic gene silencing in Drosophila. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, 14(4):377–392, 2006.

[46] Green P. Smit AFA, Hubley R. RepeatMasker open-3.0.

[47] Zhongling Jiang and Bin Zhang. On the role of transcription in positioning nucleosomes. *PLoS computational biology*, 17(1):e1008556, 2021.

[48] Stylianos Bakoulis, Robert Krautz, Nicolas Alcaraz, Marco Salvatore, and Robin Andersson. Endogenous retroviruses co-opted as divergently transcribed regulatory elements shape the regulatory landscape of embryonic stem cells. *Nucleic Acids Research*, 50(4):2111–2127, 2022.

[49] B. McCLINTOCK. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, 36(6):344–355, June 1950.

[50] Guillaume Bourque, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, Zsuzsanna Izsvák, Henry L. Levin, Todd S. Macfarlan, Dixie L. Mager, and Cédric Feschotte. Ten things you should know about transposable elements. *Genome Biology*, 19(1):199, 2018.

[51] Jiangping He, Xiuling Fu, Meng Zhang, Fangfang He, Wenjuan Li, Mazid Md Abdul, Jianguo Zhou, Li Sun, Chen Chang, Yuhao Li, He Liu, Kaixin Wu, Isaac A. Babarinde, Qiang Zhuang, Yuin-Han Loh, Jiekai Chen, Miguel A. Esteban, and Andrew P. Hutchins. Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. *Nature Communications*, 10(1):34, 2019. Number: 1 Publisher: Nature Publishing Group.

[52] Wayo Matsushima, Julien Duc, Shaoline Sheppard, Cyril Pulver, Delphine Grun, Sandra Offner, Charlène Raclot, Evarist Planet, and Didier Trono. Zinc-finger proteins with a co-opted capsid domain anchor nucleosomes over transposon sequences. Pages: 2025.03.03.638093 Section: New Results.

[53] Meng Zhou and Andrew D. Smith. Subtype classification and functional annotation of l1md retrotransposon promoters. *Mobile DNA*, 10:14, 2019.

[54] Jessica L. Elmer, Amir D. Hay, Noah J. Kessler, Tessa M. Bertozzi, Eve A. C. Ainscough, and Anne C. Ferguson-Smith. Genomic properties of variably methylated retrotransposons in mouse. *Mobile DNA*, 12(1):6, February 2021.

[55] Yanis Pelinski, Donia Hidaoui, Anne Stolz, François Hermetet, Rabie Chelbi, M'boyba Khadija Diop, Amir M. Chioukh, Françoise Porteu, and Emilie Elvira-Matelot. NF-*kappa*b signaling controls h3k9me3 levels at intronic LINE-1 and hematopoietic stem cell genes in cis. *Journal of Experimental Medicine*, 219(8):e20211356, 2022.

[56] Joanna W. Jachowicz, Xinyang Bing, Julien Pontabry, Ana Bošković, Oliver J. Rando, and Maria-Elena Torres-Padilla. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nature Genetics*, 49(10):1502–1510, October 2017.

[57] Tomoko Ichiyanagi, Hirokazu Katoh, Yoshinobu Mori, Keigo Hirafuku, Beverly Ann Boyboy, Masaki Kawase, and Kenji Ichiyanagi. B2 SINE copies serve as a transposable boundary of DNA methylation and histone modifications in the mouse. *Molecular Biology and Evolution*, 38(6):2380–2395, 2021.

[58] Francesco Gualdrini, Sara Polletti, Marta Simonatto, Elena Prosperini, Francesco Pileri, and Gioacchino Natoli. H3k9 trimethylation in active chromatin restricts the usage of functional CTCF sites in SINE b2 repeats. *Genes & Development*, 36(7):414–432, 2022.

[59] Maria Assunta Biscotti, Ettore Olmo, and J. S. (Pat) Heslop-Harrison. Repetitive DNA in eukaryotic genomes. *Chromosome Research*, 23(3):415–420, September 2015.

[60] Aleksey S Komissarov, Ekaterina V Gavrilova, Sergey Ju Demin, Alexander M Ishov, and Olga I Podgornaya. Tandemly repeated DNA families in the mouse genome. *BMC Genomics*, 12:531, October 2011.

[61] Andrew T.M. Bagshaw. Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes. *Genome Biology and Evolution*, 9(9):2428–2443, August 2017.

[62] J. L. Compton, R. Hancock, P. Oudet, and P. Chambon. Biochemical and electron-microscopic evidence that the subunit structure of Chinese-hamster-ovary interphase chromatin is conserved in mitotic chromosomes. *European Journal of Biochemistry*, 70(2):555–568, November 1976.

[63] Ada L. Olins and Donald E. Olins. Spheroid Chromatin Units (ν Bodies). *Science*, 183(4122):330–332, January 1974. Publisher: American Association for the Advancement of Science.

[64] R. D. Kornberg. Structure of chromatin. *Annual Review of Biochemistry*, 46:931–954, 1977.

[65] Maria Aurelia Ricci, Carlo Manzo, María Filomena García-Parajo, Melike Lakadamyali, and Maria Pia Cosma. Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell*, 160(6):1145–1158, 2015.

[66] Yoshinori Nishino, Mikhail Eltsov, Yasumasa Joti, Kazuki Ito, Hideaki Takata, Yukio Takahashi, Saera Hihara, Achilleas S. Frangakis, Naoko Imamoto, Tetsuya Ishikawa, and Kazuhiro Maeshima. Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *The EMBO journal*, 31(7):1644–1653, April 2012.

[67] Sandro Baldi, Philipp Korber, and Peter B. Becker. Beads on a string—nucleosome array arrangements and folding of the chromatin fiber. *Nature Structural & Molecular Biology*, 27(2):109–118, 2020. Publisher: Nature Publishing Group.

[68] Guo-Cheng Yuan, Yuen-Jong Liu, Michael F. Dion, Michael D. Slack, Lani F. Wu, Steven J. Altschuler, and Oliver J. Rando. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science (New York, N.Y.)*, 309(5734):626–630, July 2005.

[69] 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, March 2009. Publisher: Nature Publishing Group.

[70] Ho-Ryun Chung, Ilona Dunkel, Franziska Heise, Christian Linke, Sylvia Krobitsch, Ann E. Ehrenhofer-Murray, Silke R. Sperling, and Martin Vingron. The Effect of Micrococcal Nuclease Digestion on Nucleosome Positioning Data. *PLOS ONE*, 5(12):e15754, 2010. Publisher: Public Library of Science.

[71] Răzvan V. Chereji, Terri D. Bryson, and Steven Henikoff. Quantitative MNase-seq accurately maps nucleosome occupancy levels. *Genome Biology*, 20(1):198, September 2019.

[72] Kristin R. Brogaard, Liqun Xi, Ji-Ping Wang, and Jonathan Widom. Chapter fourteen - a chemical approach to mapping nucleosomes at base pair resolution in yeast. In Carl Wu and C. David Allis, editors, *Methods in Enzymology*, volume 513 of *Nucleosomes, Histones & Chromatin Part B*, pages 315–334. Academic Press, 2012.

[73] Lilien N. Voong, Liqun Xi, Ji-Ping Wang, and Xiaozhong Wang. Genome-wide mapping of the nucleosome landscape by micrococcal nuclease and chemical mapping. *Trends in Genetics*, 33(8):495–507, 2017.

[74] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nature methods*, 10(12):1213–1218, December 2013.

[75] Peter J Skene and Steven Henikoff. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*, 6:e21856, January 2017. Publisher: eLife Sciences Publications, Ltd.

[76] Hatice S. Kaya-Okur, Steven J. Wu, Christine A. Codomo, Erica S. Pledger, Terri D. Bryson, Jorja G. Henikoff, Kami Ahmad, and Steven Henikoff. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10(1):1930, April 2019. Publisher: Nature Publishing Group.

[77] Theresa K. Kelly, Yaping Liu, Fides D. Lay, Gangning Liang, Benjamin P. Berman, and Peter A. Jones. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Research*, 22(12):2497–2506, December 2012.

[78] Axel Delamarre, Benton Bailey, Jennifer Yavid, Richard Koche, Neeman Mohibullah, and Iestyn Whitehouse. Chromatin architecture mapping by multiplex proximity tagging. *bioRxiv*, page 2024.11.12.623258, January 2025.

[79] Matthew W. Snyder, Martin Kircher, Andrew J. Hill, Riza M. Daza, and Jay Shendure. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*, 164(1-2):57–68, January 2016.

[80] Lucas Penny, Sasha C. Main, Steven D. De Michino, and Scott V. Bratman. Chromatin- and nucleosome-associated features in liquid biopsy: implications for cancer biomarker discovery. *Biochemistry and Cell Biology*, 102(4):291–298, August 2024. Publisher: NRC Research Press.

[81] Sandra Tejerina-Miranda, Elisa Carral-Ibarra, Maria Gamella, Ana Montero-Calle, María Pedrero, José M. Pingarrón, Rodrigo Barderas, and Susana Campuzano. Determining and characterizing circulating nucleosomes in advanced cancer with electrochemical biosensors assisted by magnetic supports and proteomic technologies. *Biosensors and Bioelectronics*, 286:117582, October 2025.

[82] Fuhong Su, Anthony Moreau, Marzia Savi, Michele Salvagno, Filippo Annoni, Lina Zhao, Keliang Xie, Jean-Louis Vincent, and Fabio Silvio Taccone. Circulating Nucleosomes as a Novel Biomarker for Sepsis: A Scoping Review. *Biomedicines*, 12(7):1385, June 2024.

[83] José Luis García-Gimenez, Juan Carlos Ruiz-Rodríguez, Ricard Ferrer, Raquel Durá, Antonio Artigas, Iván Bajaña, David Bolado López de Andujar, Irene Cánovas-Cervera, Adrián Ceccato, Luis Chiscano-Camón, Elena Climent-Martinez, Georgia García Fernández, Gemma Goma, Verónica Monforte, Beatriz Quevedo-Sánchez, Adolf Ruiz-Sanmartín, Antonio Sierra-Rivera, and Nieves Carbonell Monleón. Circulating histones as clinical biomarkers in critically ill conditions. *FEBS Letters*, n/a(n/a), 2025. _eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1002/1873-3468.70093.

[84] Vladimir B. Teif. Nucleosome positioning: resources and tools online. *Briefings in Bioinformatics*, 17(5):745–757, 2016.

[85] 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):19–28, 2022.

[86] 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.

[87] Ji-Ping Z. Wang and Jonathan Widom. Improved alignment of nucleosome DNA sequences using a mixture model. *Nucleic Acids Research*, 33(21):6743–6755, 2005.

[88] 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, August 2006.

[89] Lewi Tonks. The Complete Equation of State of One, Two and Three-Dimensional Gases of Hard Elastic Spheres. *Physical Review*, 50(10):955–963, November 1936. Publisher: American Physical Society.

[90] Roger D. Kornberg and Lubert Stryer. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Research*, 16(14):6677–6690, July 1988.

[91] Wolfram Möbius and Ulrich Gerland. Quantitative Test of the Barrier Nucleosome Model for Statistical Positioning of Nucleosomes Up- and Downstream of Transcription Start Sites. *PLOS Computational Biology*, 6(8):e1000891, August 2010. Publisher: Public Library of Science.

[92] Răzvan V. Chereji, Denis Tolkunov, George Locke, and Alexandre V. Morozov. Statistical mechanics of nucleosome ordering by chromatin-structure-induced two-body interactions. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 83(5 Pt 1):050903, May 2011.

[93] Julien Riposo and Julien Mozziconacci. Nucleosome positioning and nucleosome stacking: two faces of the same coin. *Molecular BioSystems*, 8(4):1172–1178, April 2012. Publisher: The Royal Society of Chemistry.

[94] Vincent Miele, Cédric Vaillant, Yves d'Aubenton Carafa, Claude Thermes, and Thierry Grange. DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Research*, 36(11):3746–3756, June 2008.

[95] 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, August 2009.

[96] Federica Battistini, Christopher A. Hunter, Irene K. Moore, and Jonathan Widom. Structure-Based Identification of New High-Affinity Nucleosome Binding Sequences. *Journal of Molecular Biology*, 420(1):8–16, June 2012.

[97] Zhenhai Zhang, Christian J. Wippo, Megha Wal, Elissa Ward, Philipp Korber, and B. Franklin Pugh. A Packing Mechanism for Nucleosome Organization Reconstituted Across a Eukaryotic Genome. *Science (New York, N.Y.)*, 332(6032):977–980, May 2011.

[98] 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):346, June 2010.

[99] Robert Schöpflin, Vladimir B. Teif, Oliver Müller, Christin Weinberg, Karsten Rippe, and Gero Wedemann. Modeling nucleosome position distributions from experimental nucleosome positioning maps. *Bioinformatics*, 29(19):2380–2386, October 2013.

[100] Lasha Dalakishvili, Husain Managori, Anaïs Bardet, Věra Slaninová, Edouard Bertrand, and Nacho Molina. Biophysical Modeling Uncovers Transcription Factor and Nucleosome Binding on Single DNA Molecules. *bioRxiv*, May 2025. Pages: 2025.05.13.653852 Section: New Results.

[101] Pawan Kumar Mall, Pradeep Kumar Singh, Swapnita Srivastav, Vipul Narayan, Marcin Paprzycki, Tatiana Jaworska, and Maria Ganzha. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*, 4:100216, December 2023.

[102] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv*, (arXiv:1409.0473), May 2016.

[103] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to End Learning for Self-Driving Cars. *arXiv*, (arXiv:1604.07316), April 2016.

[104] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[105] Patrice Y. Simard, Dave Steinkrau, and Ian Buck. Using GPUs for Machine Learning Algorithms . In *Proceedings. Eighth International Conference on Document Analysis and Recognition*, pages 1115–1119, Los Alamitos, CA, USA, September 2005. IEEE Computer Society.

[106] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366, 2021. Number: 3 Publisher: Nature Publishing Group.

[107] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, March 1968.

[108] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980.

[109] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[110] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[111] Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, August 2015. Publisher: Nature Publishing Group.

[112] Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, October 2015. Publisher: Nature Publishing Group.

[113] David R. Kelley, Jasper Snoek, and John L. Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999, January 2016. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[114] Yunlong Wang, Siyuan Kong, Cong Zhou, Yanfang Wang, Yubo Zhang, Yaping Fang, and Guoliang Li. A review of deep learning models for the prediction of chromatin interactions with DNA and epigenomic profiles. *Briefings in Bioinformatics*, 26(1):bbae651, January 2025.

[115] Mattia Di Gangi, Giosuè Lo Bosco, and Riccardo Rizzo. Deep learning architectures for prediction of nucleosome positioning from sequences data. *BMC Bioinformatics*, 19(14):418, 2018.

[116] Domenico Amato, Giosue' Lo Bosco, and Riccardo Rizzo. CORENup: a combination of convolutional and recurrent deep neural networks for nucleosome positioning identification. *BMC Bioinformatics*, 21(8):326, 2020.

[117] Yiting Zhou, Tingfang Wu, Yelu Jiang, Yan Li, Kailong Li, Lijun Quan, and Qiang Lyu. DeepNup: Prediction of Nucleosome Positioning from DNA Sequences Using Deep Neural Network. *Genes*, 13(11):1983, October 2022.

[118] Guo-Sheng Han, Qi Li, and Ying Li. Nucleosome positioning based on DNA sequence embedding and deep learning. *BMC Genomics*, 23(1):301, 2022.

[119] Yosef Masoudi-Sobhanzadeh, Shuxiang Li, Yunhui Peng, and Anna R Panchenko. Interpretable deep residual network uncovers nucleosome positioning and associated features. *Nucleic Acids Research*, 52(15):8734–8745, August 2024.

[120] Stephen J. Mondo, Guifen He, Aditi Sharma, Doina Ciobanu, Robert Riley, William B. Andreopoulos, Anna Lipzen, Alan Kuo, Kurt LaButti, Jasmyn Pangilinan, Asaf Salamov, Hugh Salamon, Lili Shu, John Gladden, Jon Magnuson, M. Catherine Aime, Ronan O'Malley, and Igor V. Grigoriev. Consecutive low-frequency shifts in A/T content denote nucleosome positions across microeukaryotes. *iScience*, 28(5):112472, May 2025.

[121] Etienne Routhier, Alexandra Joubert, Alex Westbrook, Edgard Pierre, Astrid Lancrey, Marie Cariou, Jean-Baptiste Boulé, and Julien Mozziconacci. In silico design of DNA sequences for in vivo nucleosome positioning. *Nucleic Acids Research*, 52(12):6802–6810, June 2024.

[122] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv:1706.03762 [cs].

[123] Yilong Ji, Zhongchao Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[124] Simone Dalla-Torre, Julian Grewe, Zaid Hammoudeh, et al. The nucleotide transformer: Building and interpreting foundation models for biology. *bioRxiv*, 2023.

[125] Jin Zhou et al. Borzoi: A foundation model for genomic sequence-based prediction of chromatin features at single-nucleotide resolution. *bioRxiv*, 2024.

[126] P. T Lowary and J Widom. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning1. *Journal of Molecular Biology*, 276(1):19–42, February 1998.

[127] Astrid Lancrey, Alexandra Joubert, Evelyne Duvernois-Berthet, Etienne Routhier, Saurabh Raj, Agnès Thierry, Marta Sigarteu, Loic Ponger, Vincent Croquette, Julien Mozziconacci, and Jean-Baptiste Boulé. Nucleosome positioning on large tandem DNA repeats of the '601' sequence engineered in saccharomyces cerevisiae. *Journal of Molecular Biology*, 434(7):167497, 2022.

[128] 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.

[129] Cizhong Jiang and B. Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics*, 10(3):161–172, March 2009. Publisher: Nature Publishing Group.

[130] Hongyu Zhao, Yongqiang Xing, Guoqing Liu, Ping Chen, Xiujuan Zhao, Guohong Li, and Lu Cai. GAA triplet-repeats cause nucleosome depletion in the human genome. *Genomics*, 106(2):88–95, August 2015.

[131] Monica Tyagi, Nasir Imam, Kirtika Verma, and Ashok K. Patel. Chromatin remodelers: We are the drivers. *Nucleus*, 7(4):388–404, 2016. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19491034.2016.1211217.

[132] Karsten Rippe, Anna Schrader, Philipp Riede, Ralf Strohner, Elisabeth Lehmann, and Gernot Längst. DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 104(40):15635–15640, October 2007.

[133] Gwenael Badis, Esther T. Chan, Harm van Bakel, Lourdes Pena-Castillo, Desiree Tillo, Kyle Tsui, Clayton D. Carlson, Andrea J. Gossett, Michael J. Hasinoff, Christopher L. Warren, Marinella Gebbia, Shaheynoor Talukder, Ally Yang, Sanie Mnaimneh, Dimitri Terterov, David Coburn, Ai Li Yeo, Zhen Xuan Yeo, Neil D. Clarke, Jason D. Lieb, Aseem Z. Ansari, Corey Nislow, and Timothy R. Hughes. A Library of Yeast Transcription Factor Motifs Reveals a Widespread Function for Rsc3 in Targeting Nucleosome Exclusion at Promoters. *Molecular Cell*, 32(6):878–887, December 2008.

[134] Jeffrey N. McKnight, Toshio Tsukiyama, and Gregory D. Bowman. Sequence-targeted nucleosome sliding in vivo by a hybrid Chd1 chromatin remodeler. *Genome Research*, 26(5):693–704, May 2016.

[135] Elisa Oberbeckmann, Kimberly Quililan, Patrick Cramer, and A. Marieke Oudelaar. In vitro reconstitution of chromatin domains shows a role for nucleosome positioning in 3D genome organization. *Nature Genetics*, 56(3):483–492, March 2024. Publisher: Nature Publishing Group.

[136] Aurelio Balsalobre and Jacques Drouin. Pioneer factors as master regulators of the epigenome and cell fate. *Nature Reviews Molecular Cell Biology*, 23(7):449–464, July 2022. Publisher: Nature Publishing Group.

[137] Abdenour Soufi, Meilin Fernandez Garcia, Artur Jaroszewicz, Nebiyu Osman, Matteo Pellegrini, and Kenneth S. Zaret. Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. *Cell*, 161(3):555–568, April 2015.

[138] Ralph Stadhouders, Enrique Vidal, François Serra, Bruno Di Stefano, François Le Dily, Javier Quilez, Antonio Gomez, Samuel Collombet, Clara Berenguer, Yasmina Cuartero, Jochen Hecht, Guillaume J. Filion, Miguel Beato, Marc A. Marti-Renom, and Thomas Graf. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics*, 50(2):238–249, February 2018. Publisher: Nature Publishing Group.

[139] Abhijeet Pataskar, Johannes Jung, Pawel Smialowski, Florian Noack, Federico Calegari, Tobias Straub, and Vijay K Tiwari. NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program. *The EMBO Journal*, 35(1):24–45, January 2016. Num Pages: 45 Publisher: John Wiley & Sons, Ltd.

[140] 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:e47898, October 2019.

[141] Ralph Stefan Grand, Marco Pregnolato, Lisa Baumgartner, Leslie Hoerner, Lukas Burger, and Dirk Schübeler. Genome access is transcription factor-specific and defined by nucleosome position. *Molecular Cell*, 84(18):3455–3468.e6, September 2024.

[142] Catherine Do, Guimei Jiang, Giulia Cova, Christos C. Katsifis, Domenic N. Narducci, Theodore Sakellaropoulos, Raphael Vidal, Priscillia Lhoumaud, Aristotelis Tsirigos, Faye Fara D. Regis, Nata Kakabadze, Elphege P. Nora, Marcus Noyes, Anders S. Hansen, and Jane A. Skok. Binding domain mutations provide insight into CTCF's relationship with chromatin and its contribution to gene regulation. *Cell Genomics*, 5(4):100813, April 2025.

[143] Michael H. Nichols and Victor G. Corces. A CTCF Code for 3D Genome Architecture. *Cell*, 162(4):703–705, August 2015.

[144] Ya Guo, Quan Xu, Daniele Canzio, Jia Shou, Jinhuan Li, David U. Gorkin, Inkyung Jung, Haiyang Wu, Yanan Zhai, Yuanxiao Tang, Yichao Lu, Yonghu Wu, Zhilian Jia, Wei Li, Michael Q. Zhang, Bing Ren, Adrian R. Krainer, Tom Maniatis, and Qiang Wu. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, 162(4):900–910, August 2015. Publisher: Elsevier.

[145] Jie Yang, John R Horton, Bin Liu, Victor G Corces, Robert M Blumenthal, Xing Zhang, and Xiaodong Cheng. Structures of CTCF–DNA complexes including all 11 zinc fingers. *Nucleic Acids Research*, page gkad594, 2023.

[146] Catherine Do, Guimei Jiang, Paul Zappile, Adriana Heguy, and Jane A. Skok. A genome wide code to define cell-type specific CTCF binding and chromatin organization. *bioRxiv*, November 2024. Pages: 2024.11.02.620823 Section: New Results.

[147] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC Genomics*, 13:436, 2012.

[148] Jaspar matrix clusters (vertebrates). `https://jaspar.elixir.no/matrix-clusters/vertebrates/?detail=true`. Accessed: 2025-08-22.

[149] Ieva Rauluseviciute, Rafael Riudavets-Puig, Romain Blanc-Mathieu, Jaime A Castro-Mondragon, Katalin Ferenc, Vipin Kumar, Roza Berhanu Lemma, Jérémy Lucas, Jeanne Chèneby, Damir Baranasic, Aziz Khan, Oriol Fornes, Sveinung Gundersen, Morten Johansen, Eivind Hovig, Boris Lenhard, Albin Sandelin, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 52:D174–D182, 2024.

[150] Marion Dejosez, Joshua S. Krumenacker, Laura Jo Zitur, Marco Passeri, Li-Fang Chu, Zhou Songyang, James A. Thomson, and Thomas P. Zwaka. Ronin Is Essential for Embryogenesis and the Pluripotency of Mouse Embryonic Stem Cells. *Cell*, 133(7):1162–1174, June 2008. Publisher: Elsevier.

[151] Widia Soochit, Frank Sleutels, Gregoire Stik, Marek Bartkuhn, Sreya Basu, Silvia C. Hernandez, Sarra Merzouk, Enrique Vidal, Ruben Boers, Joachim Boers, Michael van der Reijden, Bart Geverts, Wiggert A. van Cappellen, Mirjam van den Hout, Zeliha Ozgur, Wilfred F. J. van IJcken, Joost Gribnau, Rainer Renkawitz, Thomas Graf, Adriaan Houtsmuller, Frank Grosveld, Ralph Stadhouders, and Niels Galjart. CTCF chromatin residence time controls three-dimensional genome organization, gene expression and DNA methylation in pluripotent cells. *Nature Cell Biology*, 23(8):881–893, 2021.

[152] W. Bao. Endogenous retrovirus from mouse. *Repbase Reports*, 20(6):1641, 2020. Repbase entry MT2_MM, LTR of an ERV3-type endogenous retrovirus (Mus musculus).

[153] Wei Ren, Leilei Gao, Yaling Mou, Wen Deng, Jinlian Hua, and Fan Yang. DUX: One Transcription Factor Controls 2-Cell-like Fate. *International Journal of Molecular Sciences*, 23(4):2067, February 2022.

[154] Marie Dewannieux, Anne Dupressoir, Francis Harper, Gérard Pierron, and Thierry Heidmann. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nature Genetics*, 36(5):534–539, May 2004. Publisher: Nature Publishing Group.

[155] Jafar Sharif, Yoichi Shinkai, and Haruhiko Koseki. Is there a role for endogenous retroviruses to mediate long-term adaptive phenotypic response upon environmental inputs? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1609):20110340, 2013.

[156] Y.I. Lee, K. Wang, A.F. Smit, J. Yu, K.G. Wong, P.S. Iadonato, L.C. Magness, P. Green, V.M. Olson, and L. Hood. Iapezi, an internal portion of a recent strain of iap provirus. Repbase Update, Release 11.08, 2006. Consensus sequence, Mus musculus, directly submitted to RepBase without submission.

[157] Léa Meneu, Christophe Chapard, Jacques Serizay, Alex Westbrook, Etienne Routhier, Myriam Ruault, Manon Perrot, Alexandros Minakakis, Fabien Girard, Amaury Bignaud, Antoine Even, Géraldine Gourgues, Domenico Libri, Carole Lartigue, Aurèle Piazza, Agnès Thierry, Angela Taddei, Frédéric Beckouët, Julien Mozziconacci, and Romain Koszul. Sequence-dependent activity and compartmentalization of foreign DNA in a eukaryotic nucleus. *Science*, 387(6734):eadm9466, February 2025. Publisher: American Association for the Advancement of Science.

[158] Laura C. Zárraga Vargas, Julio Ortíz-Ortíz, Yamelie A. Martínez, Gabriela E. Campos Viguri, Francisco I. Torres Rojas, and Pedro A. Ávila López. Identification of ZNF384 as a regulator of epigenome in leukemia. *Leukemia Research*, 153:107691, June 2025.

[159] Jochen Spiegel, Santosh Adhikari, and Shankar Balasubramanian. The Structure and Function of DNA G-Quadruplexes. *Trends in Chemistry*, 2(2):123–136, February 2020.

[160] Clayton K. Collings, Peter J. Waddell, and John N. Anderson. Effects of DNA methylation on nucleosome stability. *Nucleic Acids Research*, 41(5):2918–2931, March 2013.

[161] K. Hoogsteen. The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallographica*, 16(9):907–916, September 1963. Publisher: International Union of Crystallography.

[162] Hirotaka Nakahashi, Kyong-Rim Kieffer Kwon, Wolfgang Resch, Laura Vian, Marei Dose, Diana Stavreva, Ofir Hakim, Nathanael Pruett, Steevenson Nelson, Arito Yamane, Jason Qian, Wendy Dubois, Scott Welsh, Robert D. Phair, B. Franklin Pugh, Victor Lobanenkov, Gordon L. Hager, and Rafael Casellas. A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. *Cell reports*, 3(5):1678–1689, May 2013.

[163] Eliza C. Small, Liqun Xi, Ji-Ping Wang, Jonathan Widom, and Jonathan D. Licht. Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America*, 111(24):E2462–E2471, June 2014.

[164] Răzvan V. Chereji and David J. Clark. Major Determinants of Nucleosome Positioning. *Biophysical Journal*, 114(10):2279–2289, May 2018. Publisher: Elsevier.

[165] Chenfei Wang, Chuan Chen, Xiaoyu Liu, Chong Li, Qiu Wu, Xiaolan Chen, Lingyue Yang, Xiaochen Kou, Yanhong Zhao, Hong Wang, Yawei Gao, Yong Zhang, and Shaorong Gao. Dynamic nucleosome organization after fertilization reveals regulatory factors for mouse zygotic genome activation. *Cell Research*, 32(9):801–813, September 2022.

[166] Yi Xianfu, Zhisong He, Kuo-Chen Chou, and Xiang-Yin Kong. Nucleosome positioning based on the sequence word composition. *Protein and peptide letters*, 19:79–90, 09 2011.

[167] Raffaele Giancarlo, Simona E. Rombo, and Filippo Utro. Epigenomic k -mer dictionaries: shedding light on how sequence composition influences *in vivo* nucleosome positioning. *Bioinformatics*, 31(18):2939–2946, September 2015.

[168] Vladimir B. Teif, Daria A. Beshnova, Yevhen Vainshtein, Caroline Marth, Jan-Philipp Mallm, Thomas Höfer, and Karsten Rippe. Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Research*, 24(8):1285–1295, 2014. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[169] Rodolfo Ghirlando and Gary Felsenfeld. CTCF: making the right connections. *Genes & Development*, 30(8):881–891, 2016.

[170] Raphaël Romero, Christophe Menichelli, Christophe Vroland, Jean-Michel Marin, Sophie Lèbre, Charles-Henri Lecellier, and Laurent Bréhélin. TFscope: systematic analysis of the sequence features involved in the binding preferences of transcription factors. *Genome Biology*, 25(1):187, July 2024.

[171] Nikita S. Vassetzky, Olga R. Borodulina, Ilia G. Ustyantsev, Sergei A. Kosushkin, and Dmitri A. Kramerov. Analysis of SINE Families B2, Dip, and Ves with Special Reference to Polyadenylation Signals and Transcription Terminators. *International Journal of Molecular Sciences*, 22(18):9897, January 2021. Publisher: Multidisciplinary Digital Publishing Institute.

[172] Lucas J. T. Kaaij, Fabio Mohn, Robin H. van der Weide, Elzo de Wit, and Marc Bühler. The ChAHP complex counteracts chromatin looping at CTCF sites that emerged from SINE expansions in mouse. *Cell*, 178(6):1437–1451.e14, 2019.

[173] Wen Wang, Rui Gao, Dongxu Yang, Mingli Ma, Ruge Zang, Xiangxiu Wang, Chuan Chen, Xiaochen Kou, Yanhong Zhao, Jiayu Chen, Xuelian Liu, Jiaxu Lu, Ben Xu, Juntao Liu, Yanxin Huang, Chaoqun Chen, Hong Wang, Shaorong Gao, Yong Zhang, and Yawei Gao. ADNP modulates SINE B2-derived CTCF-binding sites during blastocyst formation in mice. *Genes & Development*, 38(3-4):168–188, March 2024.

[174] Mathys Grapotte, Manu Saraswat, Chloé Bessière, Christophe Menichelli, Jordan A. Ramilowski, Jessica Severin, Yoshihide Hayashizaki, Masayoshi Itoh, Michihira Tagami, Mitsuyoshi Murata, Miki Kojima-Ishiyama, Shohei Noma, Shuhei Noguchi, Takeya Kasukawa, Akira Hasegawa, Harukazu Suzuki, Hiromi Nishiyori-Sueki, Martin C. Frith, Clément Chatelain, Piero Carninci, Michiel J. L. de Hoon, Wyeth W. Wasserman, Laurent Bréhélin, and Charles-Henri Lecellier. Discovery of widespread transcription initiation at microsatellites predictable by sequence-based deep neural network. *Nature Communications*, 12(1):3297, June 2021. Publisher: Nature Publishing Group.

[175] Timothy L Bailey. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, 37(18):2834–2840, 2021.

[176] Yuan-Chen Sun, Wen-Jie Jiang, Kang-Wen Cai, Na-Na Wei, Fu-Ting Lai, Rui-Xiang Gao, Ze-Yu Kuang, Jia-Lu Zhou, An Liu, Han-Wen Zhu, Ming Xu, and Hua-Jun Wu. Hi-Compass resolves cell-type chromatin interactions by single-cell and spatial ATAC-seq data across biological scales, May 2025. Pages: 2025.05.14.654019 Section: New Results.

[177] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 34(suppl_1):D590–D598, January 2006.

[178] Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. MEME suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37:W202–W208, 2009.

[179] Gregory R. Lee, Ralf Gommers, Filip Waselewski, Kai Wohlfahrt, and Aaron O'Leary. PyWavelets: A python package for wavelet analysis. *Journal of Open Source Software*, 4(36):1237, 2019.

[180] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[181] Weidong Bao, Kenji K. Kojima, and Oleksiy Kohany. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1):11, June 2015.

[182] Matthew Parker. g4predict: Putative g quadruplex prediction using an extension of the quadparser method. `https://github.com/mparker2/g4predict`, 2017. Accessed: 2024-09-16, commit `<hash>`.

[183] Tsung-Han S. Hsieh, Claudia Cattoglio, Elena Slobodyanyuk, Anders S. Hansen, Oliver J. Rando, Robert Tjian, and Xavier Darzacq. Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Molecular cell*, 78(3):539–553.e8, May 2020.

[184] René Dreos, Giovanna Ambrosini, Romain Groux, Rouaïda Cavin Périer, and Philipp Bucher. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Research*, 45(D1):D51–D55, January 2017.

List of Figures

Appendix A

Complementary Chapter: Neural Network training

A.1 Architectures

Two CNN architectures were implemented to predict nucleosome occupancy profiles from DNA sequence, inspired by previous work [8]. The final choice of architectures, however, was guided by empirical testing.

CNN_simple5H A lightweight architecture designed for general nucleosome occupancy prediction. It consists of three consecutive convolutional blocks with ReLU activation (32 filters each, kernel sizes 3, 10, and 20bp), each followed by max-pooling (pool size 2) and batch normalization. The convolutional stack is flattened and passed through a dense layer (8 units, ReLU) with batch normalization, before the final output layer (5 sigmoid units corresponding to the 5 output positions).

Chemical_5H A deeper, higher-capacity model optimised for chemical cleavage data. This variant uses more filters in the first convolution (256 filters, kernel size 5bp) followed by two convolutional layers (64 filters, kernel sizes 11 and 21bp). Each convolution is followed by max-pooling (pool size 2) and batch normalization. As in CNN_simple5H, the flattened output is fed into a dense layer (8 units, ReLU) with batch normalization, then to the final 5-unit sigmoid output layer.

A.1.1 Loss function and metrics

The models were trained to minimize a custom loss function as described in Routhier *et al.* work, combining the mean absolute error (MAE) and the complement of the Pearson correlation coefficient between predicted ($\hat{y}$) and observed (y) nucleosome occupancy [8]:

$$\mathcal{L}(y, \hat{y}) = \mathrm{MAE}(y, \hat{y}) + (1 - r(y, \hat{y})) \tag{A.1}$$

where

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{A.2}$$

and

$$r(y, \hat{y}) = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}. \tag{A.3}$$

The metrics reported during training and evaluation were the Pearson correlation coefficient r and the MAE.

A.1.2 Target value reweighting

To ensure that all predicted occupancy values in the range $[0, 1]$ contributed equally to the loss, we applied a bin-wise reweighting strategy. The continuous interval $[0, 1]$ was divided into 100 equal-width bins. For each bin b, the weight w_b was set inversely proportional to its frequency in the training set:

$$w_b = \frac{1}{f_b} \tag{A.4}$$

where f_b is the fraction of training samples with target values in bin b. Each sample's contribution to the loss was then multiplied by the weight corresponding to its bin. This procedure compensates for imbalanced occupancy value distributions and prevents overrepresentation of highly frequent bins.

A.1.3 Number of heads

To stabilize the interpretation of the neural network, predictions were made at five positions along the input sequence: 501, 751, 1001, 1251, and 1501. Figure A.1 shows that the signal-to-noise ratio of the ISM score varies depending on the number of output heads. The best ratio is observed with an architecture using 21 heads. However, using 5 heads offers a good compromise between ISM score stability and the computational cost induced by the number of outputs.

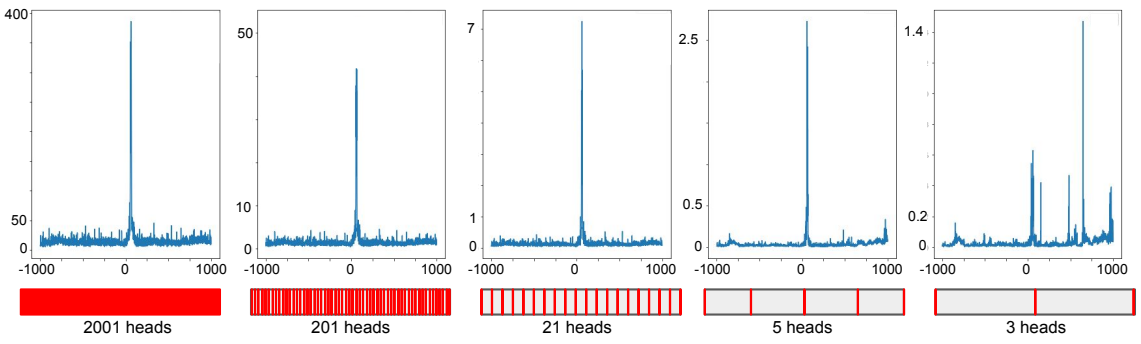


Figure A.1: **The number of head influence the noise-signal ratio of ISM computation.** Each panel is a 2kb ISM computation on a strongly reactive site

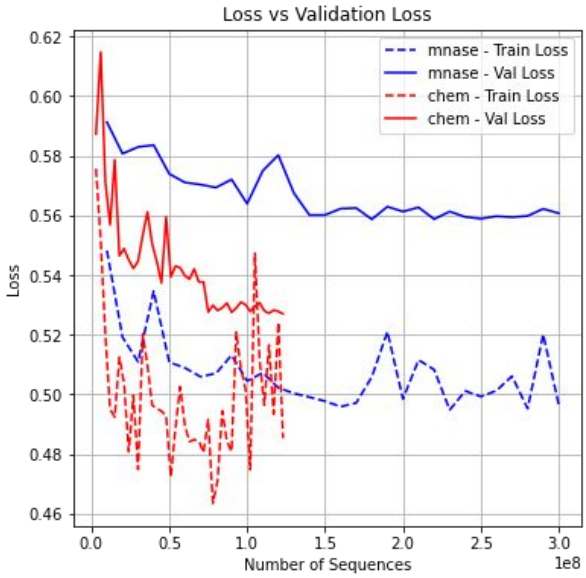


Figure A.2: **Loss and validation loss for chemical cleavage and mnase-seq model** Loss (dotted) and validation loss (solid) are shown for model trained on MNase-seq data (blue) and chemical cleavage (red)

A.2 Training strategy

A.2.1 Low-covered sequences disrupt CNN training

The main principle of any neural network is to fit the provided data. Keeping this in mind, it is crucial to carefully choose that data. In our work, we faced a challenge: the genome is largely composed of repeated elements, and mapping short reads results in ambiguous signals. To approach the biological truth of nucleosome positioning, we decided to use only reads with a unique mapping location. This process yields a nucleosome map with signal-depleted regions. To assess the performance of the network on this region we used a second signal issued from the same reads; the ambiguous mapping genes have been randomly mapped between the matching positions.

The training of the network is directly correlated with the selected windows for the training. It is neither possible to suppress totally the repeats, neither to include them. To assess the impact of low-coverage regions on model performance, we compared predictions obtained with the standard training procedure to those obtained after excluding low-coverage sequences. The threshold to exclude a window from the training set is 20 midpoints over the 2kb window. As shown in Figure A.3, filtering out such regions leads to a clearer agreement between experimental and predicted distributions. Notably, the network is able to produce high predicted values (above 0.8) which were absent in the standard training condition. This indicates that low-coverage regions introduce uncertainty into the predictions, limiting the model's ability to fully capture the signal distribution.

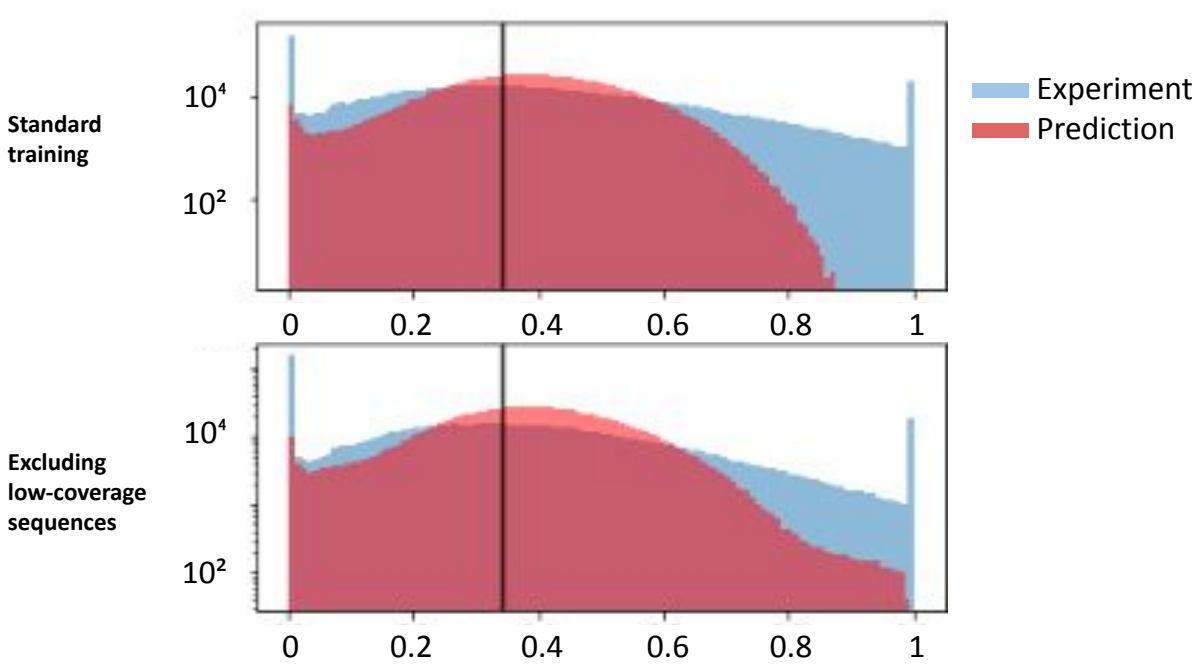


Figure A.3: **Comparison between the training strategies regarding the coverage**

A.2.2 Subsampling the signal

To estimate the amount of information contained in the MNase-seq data, we randomly subsampled the experimental signal at fractions ranging from 10 to 70% and assessed the impact on prediction quality (Figure A.4A). More specifically, we quantified the effect on the Kullback–Leibler (KL) divergence, which measures the dissimilarity between two distributions. As expected, the dissimilarity increased with the degree of subsampling.

We then fitted the empirical KL divergence values as a function of the subsampling fraction using a non-linear least-squares regression implemented in `scipy`, with `curve_fit`. The fitting function `logfit` was initialized with parameters $[2, 1, -0.2]$, and the optimization returned the best-fitting parameters along with their covariance estimates (Figure A.4B).

The extrapolation suggests that inflating the MNase-seq dataset by 400% would yield a null divergence. This conclusion, however, has important limitations. First, the experiment was performed without applying the coverage correction, which we showed to substantially improve model performance. Second, MNase-seq is subject to GC-bias, meaning that increasing sequencing depth could add redundant reads without providing additional information.

Nevertheless, these results indicate that our current MNase-seq data have not yet reached saturation, as there is no subsampling fraction at which information loss becomes negligible.

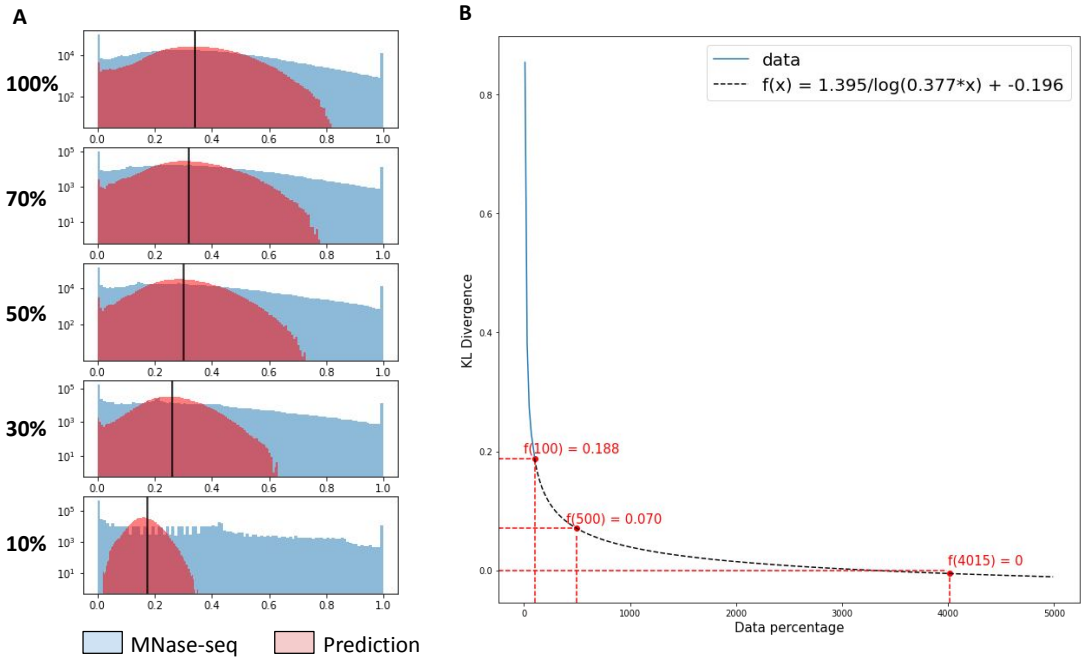


Figure A.4: **A. Distribution of prediction vs experimental MNase-seq data. B. Fit of the KL divergence between prediction and experimental data**

Appendix B

Datasets and Methods

B.1 MNase-sequencing

The nucleosomal landscape of mice embryonic stem cells has been obtained from MNase-sequencing [3] and is available under accession number GSE122589 (GEO; `https://www.ncbi.nlm.nih.gov/geo/`). The signal has been processed by Pablo Navarro's team in Institut Pasteur, notably by Luis Altamirano which provided the position of fragment midpoint (paired-end sequencing). The signal has then been constructed by performing a gaussian convolution of parameters $\mu = 1, \sigma = 15$ and truncating the signal with the 99th centile to get rid of the outliers values before normalizing it in [0-1].

B.1.1 Processing

B.1.1.1 Gaussian Kernel Construction

The Gaussian kernel is defined as a discrete approximation of the continuous Gaussian function, parameterized by the standard deviation σ and mean μ. The function is constructed over a symmetric range $[-3\sigma, 3\sigma]$ using the formula:

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{B.1}$$

B.1.1.2 Normalization

The preprocessing function applies the Gaussian convolution G to smooth the input signal. After convolution, a quantile-based thresholding is applied: the values are clipped at the $99 - th$ quantile and then normalized between 0 and 1. Formally, given an input signal S, the transformed signal is computed as:

$$\hat{S} = \frac{\min(S', \tau)}{\tau}, \quad \text{where } S' = S * G, \quad \tau = Q_{99th}(S') \tag{B.2}$$

where $*$ denotes convolution, and $Q_{99th}(S')$ represents the **99-th quantile** of the convolved signal.

B.2 Chemical cleavage

B.2.1 Processing

The Nucleosome Organization in Mouse Embryonic Stem Cells has been obtained from Chemical cleavage [4] and is available under accession number GSE82127 (GEO; https://www.ncbi.nlm.n These data were originally mapped to the mm9 assembly. We used the UCSC LiftOver tool [177] together with the provided mm9-to-mm10 chain file to remap the data to the mm10 assembly, which was used throughout the study. The nucleosome occupancy were computed as described: position k was calculated as the sum of the nucleosome core particle (NCP) scores over a local window centered on k, where S_{k+j} represents the NCP score at position $k + j$. Additionally, the center-weighted occupancy at position k was computed using a Gaussian weighting function:

$$w_j = \exp\left(-\frac{(j/20)^2}{2}\right)$$

This center-weighted approach results in a smoothed version of the NCP score, which helps refine nucleosome boundary identification.

B.3 In silico Mutagenesis

Following the mutasome approach [8], Given a set of nucleotides $B = A, C, G, T$ each single mutation is performed genome wide (3 possible mutations per position) and the result of the prediction is compared to the wild type sequence result using mean squared error. A mutascore is attributed to each position P such as

$$Mutascore = \sum_{B-Bp} \frac{MSE(Pred(b), Pred(Bp))}{3}. \tag{B.3}$$

with

$$MSE(\hat{Y}, Y) = \frac{1}{N}\sum_{n=1}(\hat{Y}_n - Y_n)^2 \tag{B.4}$$

The deliberate choice to use only the MSE to score the difference comes from the number of output head of the CNN: Performing the pearson correlation on 5 points can lead to noisy results and bad evaluation of the site.

B.3.1 Nucleosome Positioning Regions calling for sequence analysis

The motif discovery was performed on consensus peaks between MNase-seq and chemical cleavage. Specifically, we selected ISM peaks that were retrieved in both chemical-cleavage ISM and MNase-seq ISM. Both signals were z-scored and subsequently smoothed using a rolling mean with a window size of 12. The two signals were then multiplied element-wise

to highlight correlated regions, and peaks were called using scipy.signal.find_peaks with a prominence threshold of 1, yielding a set of Nucleosome Positioning Regions (NPRs).

Because the input windows were generated using a sliding procedure, fewer than 7% of sequences overlapped by at least 1 bp (Figure 5.6). The MEME/XSTREME framework tolerates such redundancy; the main effect is a potential inflation of significance estimates due to reduced sequence independence, without altering the identity of enriched motifs [178]. We therefore interpret the reported p-values/E-values with caution.

B.4 MEME suite

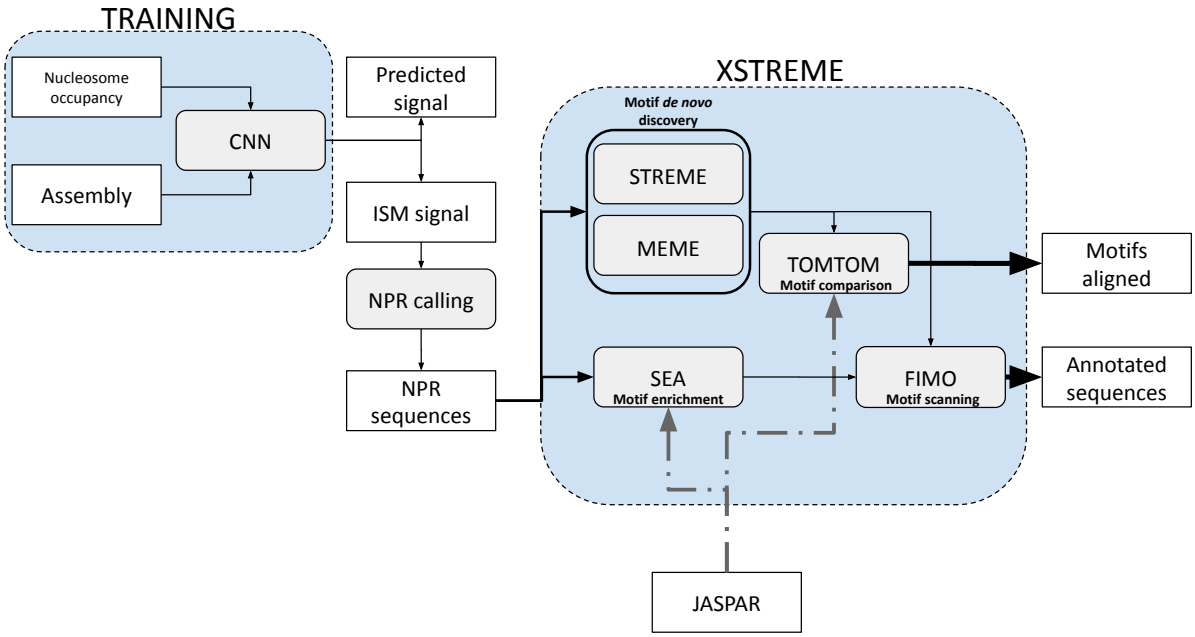


Figure B.1: **Graphical abstract of the pipeline used for sequence analysis.** Adapted from the Motif-based Sequence Analysis Tools (`https://meme-suite.org/meme/index.html`).

B.4.1 XSTREME analysis

De novo motif discovery was performed using **XSTREME** [175], part of the MEME suite. XSTREME identifies enriched motifs in input sequences by combining multiple algorithms (MEME, STREME, and motif clustering; Figure B.1) and by comparing the results to known motif databases. For our analysis, we used the set of nucleosome positioning regions (NPRs) as input and ran XSTREME with the JASPAR 2024 vertebrate non-redundant database (in MEME format) as reference motifs [149], and an E-value threshold of 1×10^{-5}. By default, the background model was estimated from the input sequences using `fasta-get-markov` with a Markov order of $k = 2$.

B.4.2 Motif filtering with SEA

We used the **Sequence Enrichment Analysis (SEA)** tool from the MEME suite to assess motif enrichment within our set of nucleosome positioning regions (NPRs). SEA evaluates whether known motifs occur more frequently in the input sequences than in a background model, providing statistical significance (E-value) and site-level scores. This approach is particularly suited to validate de novo motifs and to connect them with previously described transcription factor binding sites.

Motif instances reported by SEA were then subjected to additional filtering in order to retain only high-confidence sites:

- For each motif, we retrieved its consensus sequence and effective length from the `sea.tsv` output.
- We normalized the raw SEA site score by motif length, yielding a *normalized score*.
- Only motif instances with a normalized score ≥ 1 were kept.
- Motifs not overlapping with the subset of XSTREME-filtered motifs (E-value ≤ 0.01; MEME/STREME source only) were discarded to ensure consistency between de novo and enrichment analyses.

The resulting set, hereafter referred to as **full_sea_filtered**, represents the intersection between de novo motif discovery (XSTREME) and enrichment analysis (SEA), restricted to high-confidence motif instances with sufficient normalized score. This dataset was used as the input for subsequent overlap analyses with RepeatMasker annotations.

B.4.3 Enrichment of NPRs in repetitive-element families.

Let G denote the genomic coordinate space used as the background, here the whole reference genome without undefined bases (N). For each repeat family f, let $\mathcal{I}_f = \{I_{f,1}, \ldots, I_{f,m_f}\}$ be the set of its annotated intervals and $U_f = \bigcup_{k=1}^{m_f} I_{f,k}$ their union (non-overlapping). We write $|A|$ for the length (in bp) of a set of intervals A and define the family coverage fraction

$$\phi_f \;=\; \frac{|U_f|}{|G|}\,.$$

Let $\{J_1, \ldots, J_n\}$ be the set of NPRs and $L = \mathbb{E}[|J_i|]$ their mean length. We discretize G into $N = \big\lfloor |G|/L \big\rfloor$ non-overlapping bins of length L. Under the null hypothesis that NPRs are placed uniformly at random over G, the number of bins intersecting family f is $K_f \approx \phi_f N$, and the number of NPRs that intersect U_f by at least one base,

$$x_f \;=\; \Big|\{\, i \in \{1, \ldots, n\} : J_i \cap U_f \neq \varnothing \,\}\Big|,$$

is modelled by a right-tailed hypergeometric distribution,

$$X_f \sim \text{Hypergeom}(N,\, K_f,\, n), \qquad p_f \;=\; \Pr\Big[X_f \geq x_f \,\Big|\, N, K_f, n\Big].$$

We compute p_f numerically using the log survival function for stability (SciPy: `hypergeom.logsf`).

Let m be the number of families tested. We control the false discovery rate at level α using the Benjamini–Hochberg (BH) step-up procedure. Writing the p-values in ascending order $p_{(1)} \leq \cdots \leq p_{(m)}$, the BH q-values are

$$q_{(i)} \;=\; \min_{j \geq i} \frac{m}{j}\, p_{(j)} \quad (i = 1, \ldots, m),$$

with the usual monotonicity enforcement; in practice we implement this entirely in log-space to avoid underflow. For visualization we plot $-\log_{10}(q_f)$ on the x-axis (higher $\Rightarrow$ more significant); the vertical reference line corresponds to the chosen FDR threshold, e.g. $-\log_{10}(0.01) = 2$ for 1% FDR. On the y-axis we display the family size ("All Count"), and use a logarithmic scale for readability. The x-axis uses a symmetric log transform to display both small and very large values.

The assumptions and scope so that this formulation:

- uses the union coverage ϕ_f to account for heterogeneous repeat lengths and overlaps;
- treats NPRs as independent draws on a uniform background of bins of size L
- counts NPRs by occurrence ("touches ≥ 1 bp"), not by overlap length. Edge effects due to binning are negligible at the chosen resolution and do not affect our qualitative conclusions.

B.5 Wavelet analysis

The Wavelet analysis were performed using PyWavelet python package [179]. A complex Morlet wavelet is used and scaled to correspond to periods ranging from 150bp to 200bp in steps of 10bp. The wavelet is slid across the signal, producing higher values when the signal closely matches the wavelet.The resulting score is calculated as the sum of the magnitudes (modules) of these values (FigureB.2).

Nucleosome island are called using scipy.signal.find_peaks package with prominence of 4. [180].

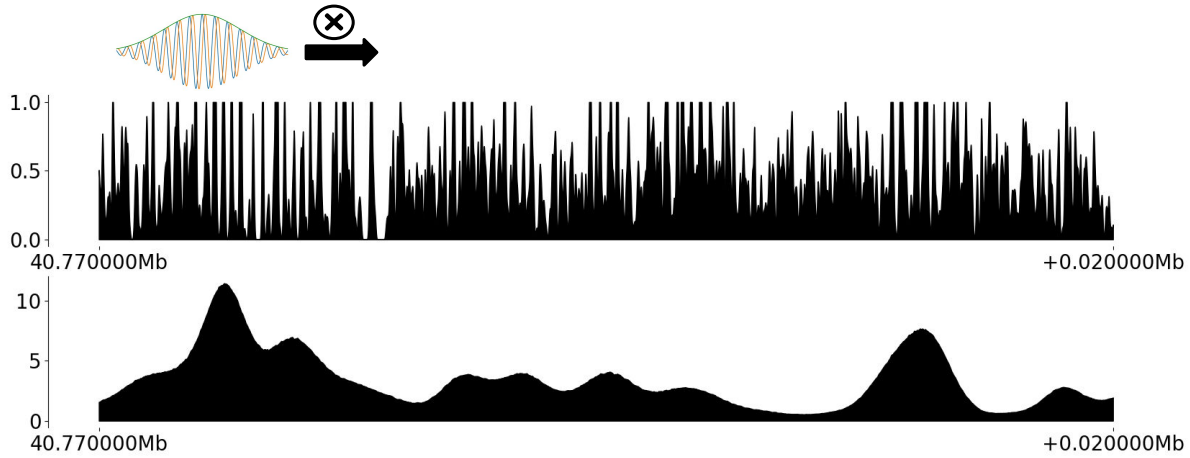


Figure B.2: **Principle of wavelet analysis.** The wavelet (blue: real part, orange: imaginary part, green:module) is slid across the signal (top), The resulting score (bottom) is calculated as the sum of the magnitudes (modules)

B.6 Repeated elements

Annotation of repeated elements was performed using RepeatMasker [46]. The annotation table is available on the UCSC Genome Browser (`https://genome.ucsc.edu/cgi-bin/hgTables?db=mm10&hgta_group=varRep&hgta_track=rmsk&hgta_table=rmsk`). The reference library of repetitive elements used for the annotation was Repbase [181].

B.7 ChIP-seq

ChIP-seq of CTCF has been provided by Navarro's team from the Institut Pasteur [3,140]

ChIP-seq of ZNF384 is available available on Encode under the accession ENCFF043AHA

B.8 G-quadruplex

Putative G-quadruplex prediction on mm10 was performed using g4predict (Matthew Parker, GitHub repository, `https://github.com/mparker2/g4predict` [182].

B.9 Micro-C

Micro-C data are available under GEO accession `GSE130275` [183]. Micro-c compartments were computed using SCN normalization, as described by Cournac *et al.* [147], using a 50kb bin size.

B.10 Transcription Start Sites

Coordinates of transcription start sites (TSS) were obtained from the **EPDnew** database [184], which provides a manually curated collection of experimentally validated promoters in eukaryotes. Each entry in EPDnew corresponds to a promoter region anchored at a high-confidence TSS, supported by high-throughput transcript mapping data (e.g. CAGE, RAMPAGE). The database is regularly updated to reflect current genome assemblies and integrates information from multiple sources to ensure accuracy.

For our analysis, we downloaded the `epdNewPromoter` table for the mouse genome (assembly mm10) using the UCSC Table Browser (`https://genome-euro.ucsc.edu/cgi-bin/hgTables`; group: regulation, track: epdNew, table: epdNewPromoter). This dataset provides strand-specific TSS positions with additional annotation fields, and was used as the reference set of promoters in our study.

Only TSS located on chromosome 1 were considered for the analysis shown in Figure 5.8. The dataset was stratified by strand orientation: for the positive strand, the annotated start coordinate was used; for the negative strand, the annotated end coordinate was taken. For each TSS, we extracted a ±2 kb window relative to the start coordinate. On the positive strand, this corresponds to positions $[TSS - 2000,\ TSS + 2000]$, while on the negative strand the same interval was extracted in reverse orientation to maintain a consistent 5'–3' transcriptional direction.

To characterize local nucleosome phasing, we computed the autocorrelation function of experimental nucleosome profiles within each TSS-centered window. For each profile, the lag corresponding to the maximum autocorrelation was identified in the range 100–500 bp, corresponding to expected nucleosome repeat lengths (NRL). Only windows with clear phasing, defined as having optimal lags between 170 and 230bp, were retained for downstream analysis. This deliberate filtering ensured that subsequent analyses focused on TSS showing well-positioned nucleosomal arrays, thereby facilitating robust comparisons between experimental and predicted signals.

B.11 Saliency

We computed base-resolution saliency maps to attribute the model's predictions to input sequence positions. We loaded the trained TensorFlow/Keras model and removed the final sigmoid activation by reconstructing the last dense layer with a linear activation, keeping all weights identical. Gradients were then computed with respect to one-hot encoded DNA input (A/C/G/T channels).

For an input batch $X \in \mathbb{R}^{B\times L\times 4}$ and a target output unit (class index) c, we used `tf.GradientTape` to obtain

$$\nabla_X \hat{y}_c = \frac{\partial \hat{y}_c}{\partial X} \in \mathbb{R}^{B\times L\times 4}.$$

Per-position, per-example saliency was defined as the channel-summed absolute gradient

after channel–mean centering,

$$S(b,\ell) \;=\; \sum_{k=1}^{4} \left| \nabla_X \hat{y}_c(b,\ell,k) \;-\; \frac{1}{4} \sum_{k'=1}^{4} \nabla_X \hat{y}_c(b,\ell,k') \right|,$$

yielding $S \in \mathbb{R}^{B \times L}$.

B.12 Data visualization

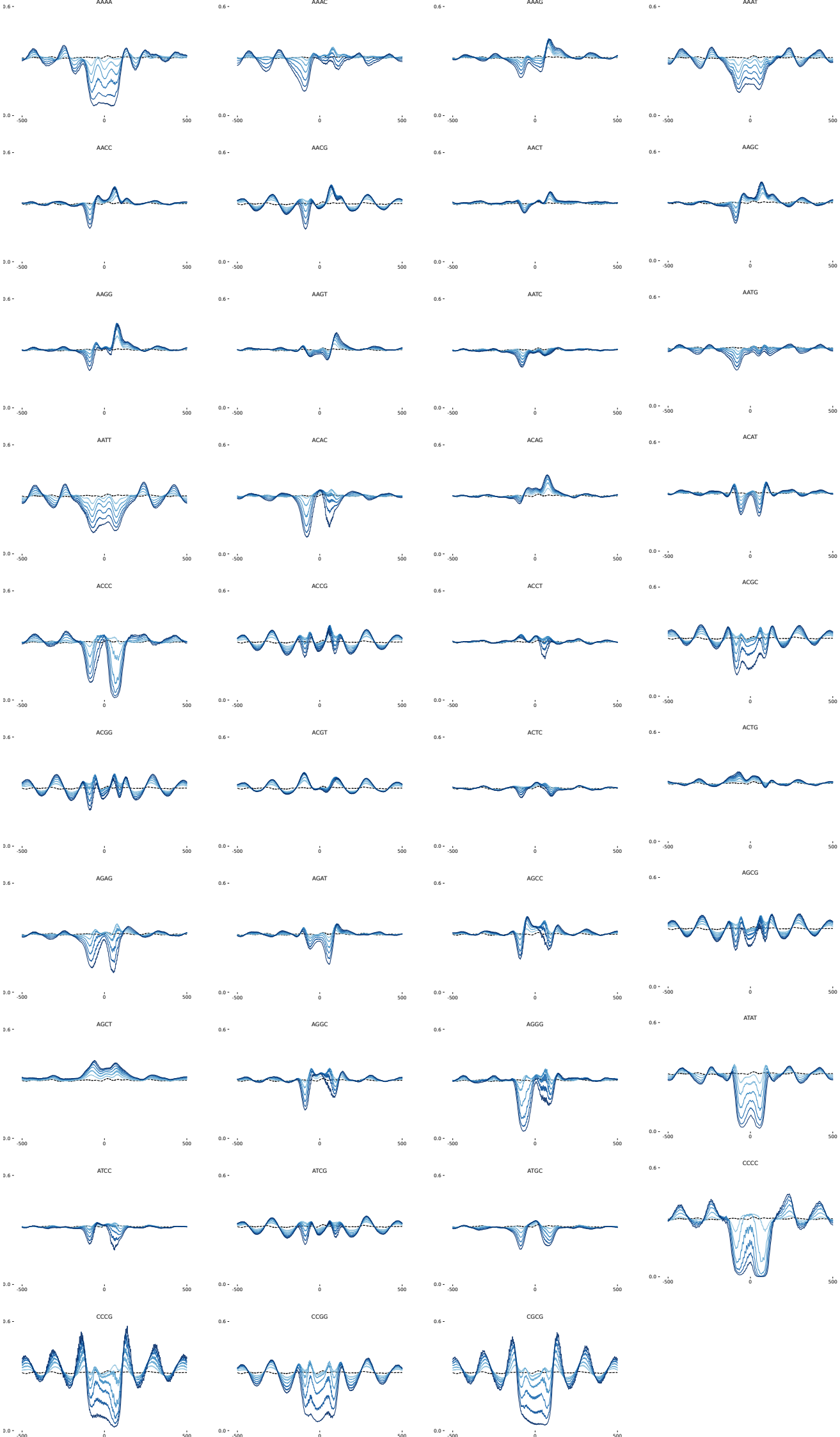
For data visualization, unless otherwise specified, sequences were oriented in the 5'–3' direction, using the reverse complement when necessary. In the same vein, repeated elements were aligned using RepeatMasker annotations: all instances were reoriented to match the consensus sequence, and their positions were expressed relative to the consensus, accounting for possible truncations.

B.13 Synthetic microsatellites

One thousand non-overlapping 3 kb-long sequences with low ISM scores were randomly selected as background. The test motifs used to simulate microsatellites consisted of all possible 4-mers (256 in total), each inserted as tandem repeats with a number of monomer units ranging from 2 to 14 (i.e., 2, 4, 6, 8, 10, 12, and 14). Predictions were performed for all combinations, yielding $256 \times 7 \times 1000$ synthetic observations.

Appendix C

Synthetic k-mers


AAAA
AAAC
AAAG
AAAT
AACC
AACG
AACT
AAGC
AAGG
AAGT
AATC
AATG
AATT
ACAC
ACAG
ACAT
ACCC
ACCG
ACCT
ACGC
ACGG
ACGT
ACTC
ACTG
AGAG
AGAT
AGCC
AGCG
AGCT
AGGC
AGGG
ATAT
ATCC
ATCG
ATGC
CCCC
CCCG
CCGG
CGCG
0.6
0.0
-500
0
500

Appendix D

Analysis of L1 monomer locus with tandem repeat finder

```
Tandem Repeats Finder Program written by:

                Gary Benson
      Program in Bioinformatics
          Boston University

Version 4.09

Sequence: chr10:27816950-27827001

Parameters: 2 7 7 80 10 50 500

Pmatch=0.80,Pindel=0.10
tuple sizes 0,4,5,7
tuple distances 0, 29, 159, 500

Length: 8051
ACGTcount: A:0.16, C:0.27, G:0.35, T:0.22


Found at i:792 original size:208 final size:208
```

Alignment explanation

```
    Indices: 437--7816  Score: 13997
    Period size: 208  Copynumber: 35.5  Consensus size: 208

        427 TGGCAGAAGT

                *         *             *  *    *   *   *      *             *
        437 TGTGTTCCACTCACTAGAGGTCTTAGGATCACGTGTGGAATCCTGTGTGGGCCCTTGCGGGTGTC
          1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT

            *          *                                        *       **
        502 AGGC--GACTCAGCTGGCAAGGTAG-CCGGGGCTCGAG--T---G--GAGTGGAAGGGTTTTGTG
         66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                     *                                                *
        557 CCCCAGATCAAGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCACGCGATTGGA
        131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA

                  *    *
        622 TTGGGGTAGGCGC
        196 TTGGGGCAGGCAC

                *                                                *
        635 TGTGTTCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGGCCCTTGCGGGTGTT
          1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT

                           *
        700 GGGCAAGACTCTGCTGTCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
         66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                        *
        765 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCGAGTTCCGCGCGATTGGA
        131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


        830 TTGGGGCAGGCAC
        196 TTGGGGCAGGCAC


        843 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
          1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


        908 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
         66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
```

```
                                                                 *
 973 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCACGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


1038 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


1051 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


1116 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                   *                *     *
1181 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCGTATGTACCGCAGTCTCGAGTTCTGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


1246 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


1259 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


1324 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                        *
1389 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCGAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


1454 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


1467 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


1532 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


1597 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


1662 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

                                  *
1675 TGTGATCCACTCACCAGAGGTCTTAGGGTCCGGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


1740 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                                 *
1805 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCACGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
```

```
1870 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


1883 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


1948 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

      *                             *               *     *
2013 CCACAGATCAGGCCCGGGTAGCCTGCTTCCGTATGTACCGCAGTCTCGAGTTCTGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


2078 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


2091 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


2156 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                      *
2221 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCGAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


2286 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

                                                 *
2299 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCTGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


2364 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


2429 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


2494 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

                                        *
2507 TGTGATCCACTCACCAGAGGTCTTAGGGTCCGGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


2572 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


2637 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


2702 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


2715 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
```

```
2780 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                        *
2845 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCGAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA

             *
2910 TTGGGGCAGGTAC
 196 TTGGGGCAGGCAC

         *
2923 TGTGGTCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


2988 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                        *
3053 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCGAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA

             *
3118 TTGGGGCAGGTAC
 196 TTGGGGCAGGCAC

         *
3131 TGTGGTCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


3196 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


3261 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


3326 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


3339 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


3404 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


3469 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


3534 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

                                                      *       *
3547 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGGCCCTTGCCGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


3612 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
```

```
3677 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


3742 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


3755 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


3820 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                                    *
3885 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGAGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


3950 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

        *        *                                        *
3963 TGTGGTCCACTCAGCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGGCCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


4028 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


4093 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


4158 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

                                     *
4171 TGTGATCCACTCACCAGAGGTCTTAGGGTCCGGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


4236 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                              *
4301 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCTGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


4366 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


4379 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


4444 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                         *
4509 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCGAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA

              *
4574 TTGGGGCAGGTAC
 196 TTGGGGCAGGCAC
```

```
        *
4587 TGTGGTCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


4652 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


4717 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


4782 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

                                    *
4795 TGTGATCCACTCACCAGAGGTCTTAGGGTCCTGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


4860 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


4925 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


4990 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

        *       *                                   *
5003 TGTGGTCCACTCAGCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGGCCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


5068 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


5133 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


5198 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

                                    *
5211 TGTGATCCACTCACCAGAGGTCTTAGGGTCCGGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


5276 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


5341 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


5406 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


5419 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
```

```
5484 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


5549 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


5614 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

        *        *                                     *
5627 TGTGGTCCACTCAGCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGGCCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


5692 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


5757 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


5822 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

                                    *
5835 TGTGATCCACTCACCAGAGGTCTTAGGGTCCGGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


5900 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


5965 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


6030 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


6043 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


6108 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                      *
6173 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCGAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA

             *
6238 TTGGGGCAGGTAC
 196 TTGGGGCAGGCAC

        *
6251 TGTGGTCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


6316 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


6381 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
```

```
6446 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


6459 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


6524 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


6589 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


6654 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

                                       *                              *
6667 TGTGATCCACTCACCAGAGGTCTTAGGGTCCGGTGGGGAGTCCCGTGTGGACCCTTGCCGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


6732 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


6797 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


6862 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC


6875 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


6940 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                              *
7005 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGAGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


7070 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

        *        *                                     *
7083 TGTGGTCCACTCAGCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGGCCCTTGCGGGTGTT
   1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


7148 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
  66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


7213 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
 131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA


7278 TTGGGGCAGGCAC
 196 TTGGGGCAGGCAC

                                       *
```

```
        7291 TGTGATCCACTCACCAGAGGTCTTAGGGTCCGGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
           1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


        7356 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
          66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG


        7421 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
         131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA

                      *
        7486 TTGGGGCAGGTAC
         196 TTGGGGCAGGCAC

                 *
        7499 TGTGGTCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
           1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


        7564 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
          66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG

                                                              *
        7629 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCGAGTTCCGCGCGATTGGA
         131 CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA

                      *
        7694 TTGGGGCAGGTAC
         196 TTGGGGCAGGCAC

                 *
        7707 TGTGGTCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
           1 TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT


        7772 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAG
          66 GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAG

        7817 AATATTACAG


Statistics
Matches: 7057,  Mismatches: 115, Indels: 10
        0.98            0.02        0.00

Matches are distributed among these distances:
 198   61  0.01
 200   17  0.00
 201   12  0.00
 203    1  0.00
 206    1  0.00
 208  6965  0.99

ACGTcount: A:0.16, C:0.27, G:0.36, T:0.21


Consensus pattern (208 bp):
TGTGATCCACTCACCAGAGGTCTTAGGGTCCCGTGGGGAGTCCCGTGTGGACCCTTGCGGGTGTT
GGGCAAGACTCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTG
CCCCAGATCAGGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGA
TTGGGGCAGGCAC

Done.
```

Appendix E

XSTREME analysis of Nucleosome Positioning Regions

CLUSTER	SOURCE	ALT_ID	CONSENSUS	WIDTH	SITES	EVALUE	SIM_MOTIF
1	MEME	MEME-1	CCASHAGRGGG	11	185421	0.00E+00	MA1930.2 (CTCF)
1	STREME	STREME-1	CCASHAGRGGGCRBY	15	136970	0.00E+00	MA0139.2 (CTCF)
1	JASPAR	CTCF	RCCASYAGRKGGCRS	15	123761	0.00E+00	MA0139.2
1	JASPAR	CTCFL	CAGGGGGC	8	107990	0.00E+00	MA1102.3
1	JASPAR	ZNF707	CCCCACTCCTGGTAC	15	35148	0.00E+00	MA1715.1
1	STREME	STREME-8	ACACACCAGAAGAGG	15	3408	5.50E-92	8-ACACACCAGAAGAGG
1	STREME	STREME-22	22-CCACTGGAGGGM	12	2748	2.85E-20	MA0139.2 (CTCF)
1	STREME	STREME-23	23-CACTNGGTGGCA	12	3774	7.28E-20	MA0139.2 (CTCF)
1	STREME	STREME-31	31-CCCTCTTCTGGCCTC	15	739	2.07E-15	MA1969.2 (THRA)
1	STREME	STREME-33	33-CCYCTAGRGG	10	7148	2.42E-15	MA0139.2 (CTCF)
1	STREME	STREME-50	50-CAGCAGAGG	9	1146	2.74E-07	MA1593.2 (ZNF317)
2	JASPAR	ZNF175	ACAGGAAGT	9	153970	0.00E+00	MA2332.1
2	JASPAR	ELF1	CAGGAAGTG	9	107803	0.00E+00	MA0473.4
2	JASPAR	Erg	ACAGGAAGTG	10	98333	0.00E+00	MA0474.4
2	JASPAR	Zbtb2	ACCGGAAGTG	10	75891	0.00E+00	MA2340.1
2	JASPAR	GABPA	CACTTCCTGT	10	67557	0.00E+00	MA0062.4
2	JASPAR	Spi1	AAAGAGGAAGTGG	13	63919	0.00E+00	MA0080.7
2	JASPAR	ZBTB11	CACTTCCGG	9	63073	0.00E+00	MA2329.1
2	JASPAR	ZBTB7A	CCGGAAGTG	9	62382	0.00E+00	MA0750.3
2	JASPAR	FOXJ2::ELF1	AAACMGGAAGT	11	37514	0.00E+00	MA1952.2
2	JASPAR	ETV2::FOXI1	AAACAGGAAGY	11	33725	0.00E+00	MA1942.2
2	JASPAR	ETV3	ACCGGAAGT	9	24692	0.00E+00	MA0763.2
2	JASPAR	ETV1	ACAGGAAGT	9	113034	5.30E-248	MA0761.3
2	JASPAR	SPIB	TCACTTCCTCTTT	13	47938	2.85E-215	MA0081.3
2	JASPAR	ETS1	ACCGGAART	9	6911	1.79E-96	MA0098.4
2	JASPAR	ELK1::HOXA1	ACCGGAAGTAATTA	14	4832	6.00E-91	MA1931.1
2	JASPAR	FOXO1::ELF1	RTMAACAGGAAGT	13	6896	1.87E-87	MA1953.2
2	JASPAR	FOXO1::ELK3	RWMAACAGGAAGT	13	1032	5.45E-84	MA1955.2
2	JASPAR	ETV5::DRGX	MGGAAGYAATTA	12	6771	2.79E-76	MA1944.2
2	JASPAR	ETS2	ACCGGAAGY	9	4939	3.68E-65	MA1484.2
2	JASPAR	ERF::NHLH1	VGCAGCTGCCGGAARY	16	1954	7.28E-60	MA1938.2
2	JASPAR	ETV2	ACCGGAAAT	9	6246	1.94E-57	MA0762.2
2	JASPAR	FOXO1::ELK1	RWMAACAGGAAGT	13	5676	2.82E-56	MA1954.2
2	JASPAR	SPDEF	AMCCGGATGT	10	8117	5.44E-55	MA0686.2
2	JASPAR	FEV	ACCGGAAGT	9	5501	5.21E-46	MA0156.4
2	JASPAR	ELK4	CRCTTCCGG	9	5824	1.31E-41	MA0076.3
2	JASPAR	Ikzf3	CAGGAAGTG	9	10180	5.23E-40	MA1992.2
2	JASPAR	ELK1::HOXB13	ACCGGAAGTHGTAAA	15	1927	2.62E-38	MA1932.2
2	JASPAR	ELK1::SREBF2	CCGGAAGTSRCGTGA	15	1439	4.53E-31	MA1933.2
2	JASPAR	ETV4	ACCGGAAGT	9	17362	3.23E-30	MA0764.4
2	JASPAR	ELF3	CACTTCCTG	9	7694	1.52E-27	MA0640.3
2	JASPAR	ETV6	SCGGAAGTR	9	17171	1.91E-25	MA0645.2
2	JASPAR	ELF2	AMCCGGAAGT	10	10936	3.98E-25	MA1483.3
2	JASPAR	FLI1	ACCGGAART	9	2919	1.75E-22	MA0475.3
2	JASPAR	FLI1::DRGX	ACCGGAAGTAATTA	14	1727	1.64E-15	MA1949.2
2	JASPAR	ETV5	ACCGGAAGT	9	5848	1.26E-12	MA0765.4
2	JASPAR	HOXB2::ELK1	RCCGGAAGTMRTTA	14	6287	1.23E-11	MA1957.1
2	JASPAR	ERF	ACCGGAAGT	9	4429	2.12E-10	MA0760.2
2	JASPAR	ELK3	ACCGGAAGT	9	3009	7.20E-06	MA0759.3
3	JASPAR	Ptf1A	ACAGATGTT	9	136286	0.00E+00	MA1618.2
3	JASPAR	Neurod2	ACAGATGG	8	123529	0.00E+00	MA0668.3
3	JASPAR	NEUROD1	ACAGATGG	8	123154	0.00E+00	MA1109.2
3	JASPAR	Atoh1	CAGATGG	7	120011	0.00E+00	MA1467.3
3	JASPAR	NEUROG2	CAGATGG	7	118271	0.00E+00	MA1642.2
3	JASPAR	HAND2	CAGATG	6	106601	0.00E+00	MA1638.2
3	JASPAR	TWIST1	CCAGATGT	8	78642	0.00E+00	MA1123.3
3	JASPAR	TAL1::TCF3	AMCATCTGKT	10	22751	0.00E+00	MA0091.2
3	JASPAR	Ptf1A	ACACCTGT	8	104015	2.84E-203	MA1620.2
3	JASPAR	Ptf1A	ACAGCTGT	8	60417	3.48E-45	MA1619.2
4	JASPAR	KLF9	GCCACACCCAC	11	23626	0.00E+00	MA1107.3
4	STREME	STREME-15	15-RRGCCMCACCTHCT	14	2336	1.01E-33	MA1107.3 (KLF9)
4	STREME	STREME-52	52-AGGTGGGGC	9	814	8.90E-07	MA0039.5 (KLF4)
5	JASPAR	ATOH7	AVCATATGBY	10	90067	0.00E+00	MA1468.1
5	JASPAR	Atoh1	RMCATATG	8	62827	0.00E+00	MA0461.3
5	JASPAR	NEUROG2	RACATATGTC	10	27136	0.00E+00	MA0669.1
5	JASPAR	OLIG1	AMCATATGKT	10	9209	0.00E+00	MA0826.1
5	JASPAR	BHLHE22	AMCATATGKY	10	6907	0.00E+00	MA0818.2

5	JASPAR	NEUROG1	RACATATGTY	10	3177	0.00E+00	MA0623.2
5	JASPAR	TCF21	ACCATATGKY	10	2943	0.00E+00	MA1568.2
5	JASPAR	MAX::MYC	ASCACGTGGT	10	2901	0.00E+00	MA0059.2
5	JASPAR	TFAP4	AHCATRTGDT	10	3210	0.00E-02	MA1570.1
5	JASPAR	BHLHA15	ACCATATGGT	10	8830	0.00E-02	MA0607.2
5	JASPAR	OLIG3	AMCATATGBY	10	9345	1.27E-302	MA0827.1
5	JASPAR	Msgn1	RACAAATGGT	10	43699	1.79E-302	MA1524.3
5	JASPAR	BHLHE23	AACATATGBT	10	6998	3.03E-282	MA0817.2
5	JASPAR	OLIG2	AMCATATGKT	10	3265	1.46E-238	MA0678.1
5	JASPAR	MXI1	CACATG	6	6076	3.17E-144	MA1108.3
5	JASPAR	FERD3L	GCRMCAGCTGTYAC	14	7005	6.76E-44	MA1485.1
5	JASPAR	Ascl2	ARCAGCTGCY	10	8409	9.73E-24	MA0816.1
5	JASPAR	MYF6	AACARCTGTT	10	553	9.15E-18	MA0667.1
6	JASPAR	GCM2	ATGCGGGT	8	123132	0.00E+00	MA0767.2
6	JASPAR	GCM1	ATGCGGGTAC	10	105688	0.00E+00	MA0646.2
7	JASPAR	ZNF317	ACAGCAGA	8	80266	0.00E+00	MA1593.2
8	JASPAR	KLF6	CCACGCCCH	9	146476	0.00E+00	MA1517.2
8	JASPAR	SP9	CCACGCCCMC	10	49132	0.00E+00	MA1564.2
8	JASPAR	KLF11	CCACGCCCMC	10	46758	0.00E+00	MA1512.2
8	JASPAR	SP8	CCACGCCCMCY	11	42975	0.00E+00	MA0747.2
8	JASPAR	KLF4	CCCCACCC	8	37517	0.00E+00	MA0039.5
8	JASPAR	KLF1	GGGYGGGG	8	23215	0.00E+00	MA0493.3
8	JASPAR	KLF7	GGGCGGGG	8	21044	0.00E+00	MA1959.2
8	JASPAR	KLF12	GGGGCGGGG	9	19892	0.00E+00	MA0742.2
8	JASPAR	SP3	GCCACGCCCMC	11	19755	0.00E+00	MA0746.3
8	JASPAR	KLF10	GGGGCGGGG	9	18432	0.00E+00	MA1511.2
8	JASPAR	KLF2	CCACRCCC	8	17536	0.00E+00	MA1515.2
8	JASPAR	KLF16	GMCACGCCCCC	11	17333	0.00E+00	MA0741.1
8	JASPAR	KLF5	GCCCCDCCCH	10	14622	0.00E+00	MA0599.1
8	JASPAR	KLF3	GRCCRCGCCC	10	13946	0.00E+00	MA1516.2
8	JASPAR	KLF14	KGGGCGGGG	9	12569	0.00E+00	MA0740.2
8	JASPAR	SP4	GGGGCGGGG	9	10940	0.00E+00	MA0685.2
8	JASPAR	SP2	GGGGCGGGG	9	8228	0.00E+00	MA0516.3
8	JASPAR	KLF15	CCCCGCCC	8	8207	0.00E+00	MA1513.2
8	JASPAR	SP1	GGGGCGGGG	9	8101	0.00E+00	MA0079.5
8	JASPAR	PATZ1	SGGGGMGGGGS	11	5183	3.15E-263	MA1961.2
8	STREME	STREME-2	GCCMCGCCC	9	19849	1.02E-244	MA1959.2 (KLF7)
8	JASPAR	KLF13	TGMCACGCCCCTTTTTG	17	3506	1.11E-244	MA0657.2
8	JASPAR	ZNF281	GGGGGAGGGG	10	49857	2.08E-232	MA1630.3
8	JASPAR	ZNF682	GGCYAAGCCCC	11	10002	4.94E-195	MA1599.2
8	JASPAR	MAZ	CCCCTCCC	8	35059	4.08E-86	MA1522.2
8	STREME	STREME-53	53-AGCGGGCGCGAC	12	468	1.36E-06	MA1727.2 (ZNF417)
8	STREME	STREME-57	57-AGCCCCCGCCCTGC	14	300	3.69E-06	MA1513.2 (KLF15)
9	JASPAR	NKX2-8	VCACTTSA	8	98585	0.00E+00	MA0673.2
9	JASPAR	NKX2-4	CCACTTSA	8	77805	0.00E+00	MA2003.2
9	JASPAR	NKX2-3	CCACTTRA	8	73098	0.00E+00	MA0672.2
9	JASPAR	Nkx3-1	CCACTTA	7	63934	0.00E+00	MA0124.3
9	JASPAR	Nkx2-1	CACTTGA	7	73116	5.51E-221	MA1994.2
9	JASPAR	Nkx3-2	AAMCACTTAA	10	3096	2.63E-41	MA0122.4
10	JASPAR	Hic1	RTGCCAAC	8	45961	0.00E+00	MA0739.2
10	JASPAR	HIC2	RTGCCC	6	20879	0.00E+00	MA0738.2
10	JASPAR	NFIX	TGCCAA	6	181106	3.54E-99	MA0671.2
11	JASPAR	ZIC4	GRCCCCCCGCKGYG	14	110930	0.00E+00	MA0751.2
11	JASPAR	ZIC1	GACCCCCYGCTGTG	14	26639	0.00E+00	MA0696.1
11	JASPAR	ZIC5	GACCCCCCGCTGHGM	15	16032	0.00E+00	MA1584.2
11	JASPAR	ZNF701	GAGCACYAARGGGGGRA	17	29936	2.76E-123	MA1987.2
11	JASPAR	GLIS2	GACCCCCCGCRAMG	14	3635	1.52E-26	MA0736.1
12	JASPAR	Yy1	CAAAATGG	8	114129	0.00E+00	MA0095.4
12	JASPAR	ZFP42	CAARATGGCTGCC	13	63400	0.00E+00	MA1651.2
13	JASPAR	Thap11	ACTACAABTCCCAG	14	7823	0.00E+00	MA1573.2
13	STREME	STREME-3	CTGGGATTTGAACTC	15	8579	6.81E-208	MA1573.2 (Thap11)
14	JASPAR	E2F6	GGCGGGAA	8	149790	0.00E+00	MA0471.3
14	JASPAR	TFDP1	SGCGGGAA	8	114503	6.42E-286	MA1122.2
14	JASPAR	E2F8	TTCCCGCCA	9	88648	2.69E-246	MA0865.3
15	JASPAR	OSR2	ACAGAAGC	8	52516	0.00E+00	MA1646.2
16	JASPAR	SCRT2	GCAACAGGTG	10	114849	0.00E+00	MA0744.3
16	JASPAR	SCRT1	KCAACAGGTG	10	80616	0.00E+00	MA0743.3
17	JASPAR	Six3	ATRGGGTATCA	11	36132	0.00E+00	MA0631.2

18	JASPAR	KLF17	MCCACGCACCCMTY	14	10870	0.00E+00	MA1514.2
18	STREME	STREME-4	ATACCCACCCACCCC	15	7210	3.72E-170	MA1514.2 (KLF17)
18	STREME	STREME-59	59-CTCCCCCACCCACCC	15	232	7.38E-06	MA0039.5 (KLF4)
19	JASPAR	ZNF85	GAGATTACWKCA	12	47175	0.00E+00	MA1720.2
20	JASPAR	CREB1	TGATGTCA	8	90707	0.00E+00	MA0018.5
20	JASPAR	FOSL1::JUN	ATGACGTCAT	10	80337	0.00E+00	MA1129.1
20	JASPAR	FOS::JUN	RTGACGTCAT	10	69536	0.00E+00	MA1126.2
20	JASPAR	FOSB::JUNB	RTGACGTCAT	10	67218	0.00E+00	MA1136.1
20	JASPAR	FOSB::JUN	GATGACGTCAT	11	64902	0.00E+00	MA1127.1
20	JASPAR	JUN::JUNB	KRTGACGTCAT	11	62420	0.00E+00	MA1133.2
20	JASPAR	FOSL2::JUND	RTGACGTCAY	10	60660	0.00E+00	MA1145.2
20	JASPAR	ATF2	ATGABGTCAT	10	32757	0.00E+00	MA1632.2
20	JASPAR	CREB3L4	RTGACGTCA	9	68728	9.59E-287	MA1475.2
20	JASPAR	JUNB	ATGACGTCAYC	11	9942	3.80E-198	MA1140.3
20	JASPAR	Nr1H4	AGGTCA	6	13107	6.81E-176	MA1110.3
20	JASPAR	Nr1H2	AGGTCA	6	10267	3.66E-165	MA1996.2
20	JASPAR	Creb5	ATGACGTCAY	10	2762	2.31E-138	MA0840.2
20	JASPAR	FOSL2::JUNB	ATGACGTCAT	10	2960	2.89E-129	MA1139.2
20	JASPAR	JUND	RATGABGTCAT	11	2245	1.25E-124	MA0492.2
20	JASPAR	FOSL2::JUN	RTGACGTMAT	10	3692	2.24E-100	MA1131.2
20	JASPAR	Atf1	RTGACGTA	8	8432	1.09E-92	MA0604.1
20	JASPAR	FOS	GATGACGTCATCR	13	1921	1.50E-85	MA1951.2
20	JASPAR	SREBF1	ATCACSTGAT	10	47830	7.03E-81	MA0829.3
20	JASPAR	FOSL1::JUND	RTGACGTMA	9	128109	2.83E-53	MA1143.2
20	JASPAR	CEBPG	ATGATGCAAT	10	3657	1.75E-24	MA1636.2
20	JASPAR	CREM	GTGACGTCAC	10	69	5.49E-12	MA0609.3
20	JASPAR	ATF4	ATGATGCAAT	10	12277	1.06E-09	MA0833.3
20	JASPAR	ATF7	ATGACGTCAT	10	73	3.81E-09	MA0834.2
20	JASPAR	SREBF1	VTCACCCCAY	10	1087	6.11E-08	MA0595.1
20	JASPAR	JDP2	ATGACGTCAY	10	123	1.27E-06	MA0656.2
21	JASPAR	DUX4	TAAYYYAATCA	11	15234	0.00E+00	MA0468.1
21	JASPAR	Dux	TGATTBAATCA	11	14117	1.49E-265	MA0611.3
21	JASPAR	PHOX2A	TAATYYAATTA	11	14752	1.05E-88	MA0713.1
21	JASPAR	PHOX2B	TAATCAAATTAW	12	1654	1.20E-32	MA0681.3
22	JASPAR	Hand1::Tcf3	RTCTGGMWT	9	29226	0.00E+00	MA0092.2
22	JASPAR	Hand1	TCCAGACCT	9	111203	9.63E-162	MA2123.1
23	JASPAR	ERF::FOXI1	AAACMGGAAR	10	48140	0.00E+00	MA1935.2
23	JASPAR	FLI1::FOXI1	TAAACAGGAAR	11	20064	0.00E+00	MA1950.2
23	JASPAR	ETV5::FOXO1	GTMAACAGGA	10	10650	0.00E+00	MA1947.2
23	JASPAR	ETV5::FOXI1	GTAAACAGGAWG	12	19053	3.67E-137	MA1946.2
23	JASPAR	IKZF1	AACAGGAA	8	6337	1.95E-114	MA1508.2
23	JASPAR	Stat2	GAAACAGAAA	10	977	3.68E-111	MA1623.2
23	JASPAR	ERF::FIGLA	CCGGAARCASCTG	13	1577	1.72E-58	MA1934.2
23	JASPAR	ERF::FOXO1	RTMAACAGGAAR	12	2150	2.78E-40	MA1936.2
23	JASPAR	ETV5::FIGLA	RGCGGAARCAGGTG	14	5366	1.74E-27	MA1945.2
24	JASPAR	PLAGL2	GGGCCCCC	8	4724	4.06E-107	MA1548.2
24	STREME	STREME-6	ACAGGGCCCCCAA	13	6178	2.42E-103	MA1548.2 (PLAGL2)
25	STREME	STREME-26	26-GGTTCGAACAC	11	978	4.53E-17	MA2096.1 (ZNF524)
26	JASPAR	RUNX3	AACCTCAA	8	134768	0.00E+00	MA0684.3
26	JASPAR	Bcl11B	AAACCACAA	9	86300	0.00E+00	MA1989.2
26	JASPAR	Runx1	YTGTGGTTT	9	57490	0.00E+00	MA0002.3
26	JASPAR	RUNX2	WAACCGCAA	9	2951	1.96E-147	MA0511.2
27	JASPAR	Nr1h3::Rxra	TGACCTNNAGTRACCY	16	47080	0.00E+00	MA0494.2
27	JASPAR	THRA	GTGTCCTCABRTGACCTY	18	9113	0.00E+00	MA1969.2
27	JASPAR	Rarg	AAGGTCAMSARAGGTCA	17	82	1.21E-06	MA0860.1
28	JASPAR	ZBTB26	CTCCAGAA	8	123530	0.00E+00	MA1579.2
29	JASPAR	LIN54	TTTRAAT	7	8449	0.00E+00	MA0619.2
30	STREME	STREME-5	AAAAAAAAAAAASAA	15	17715	2.77E-106	MA1125.2 (ZNF384)
30	STREME	STREME-27	27-AACAAAAAAAAAAAM	15	5891	7.51E-17	MA1125.2 (ZNF384)
30	JASPAR	ZNF384	AAAAAAAA	8	129120	3.62E-16	MA1125.2
30	STREME	STREME-37	37-AATWAAAAAAAAAAA	15	2944	3.72E-12	MA1125.2 (ZNF384)
30	STREME	STREME-60	60-AGAAAAAAACCAATA	15	292	1.48E-05	MA1125.2 (ZNF384)
31	JASPAR	FIGLA	CACCTG	6	105587	0.00E+00	MA0820.2
31	JASPAR	MLXIPL	RTCACGTG	8	53850	0.00E+00	MA0664.2
31	JASPAR	CREB3L4	GCCACGTCAY	10	51155	0.00E+00	MA1474.2
31	JASPAR	SNAI1	RCAGGTG	7	46068	0.00E+00	MA1558.2
31	JASPAR	USF2	RTCACGTGAY	10	4885	0.00E+00	MA0526.5
31	JASPAR	MYCN	CCACGTGG	8	4817	0.00E+00	MA0104.5

31	JASPAR	MLX	RTCACGTGAT	10	10988	2.34E-252	MA0663.1
31	JASPAR	TFEB	CACGTGAC	8	30951	3.82E-240	MA0692.2
31	JASPAR	TFE3	GTCACGTGAC	10	5054	3.22E-226	MA0831.3
31	JASPAR	MITF	RTCACGTGAY	10	6827	2.67E-203	MA0620.4
31	JASPAR	TFEC	CACGTGAC	8	12007	7.26E-162	MA0871.3
31	JASPAR	MNT	CACGTG	6	49297	7.04E-160	MA0825.2
31	JASPAR	MYC	CCACGTGC	8	77968	5.07E-157	MA0147.4
31	JASPAR	MAX	CACGTG	6	44257	3.05E-148	MA0058.4
31	JASPAR	ZEB1	CACCTG	6	66997	1.81E-145	MA0103.4
31	JASPAR	USF1	GTCATGTGAC	10	8198	1.34E-135	MA0093.4
31	JASPAR	CREB3L1	TGCCACGTCAYCA	13	12496	7.48E-129	MA0839.2
31	JASPAR	TCFL5	KCACGCGC	8	30027	1.87E-126	MA0632.3
31	JASPAR	Npas2	SCACGTGT	8	7339	1.43E-121	MA0626.2
31	JASPAR	HEY1	GRCACGTGCC	10	24177	3.31E-99	MA0823.1
31	JASPAR	Arntl	GTCACGTG	8	15281	2.21E-98	MA0603.2
31	JASPAR	XBP1	GMCACGTCATC	11	11903	1.75E-86	MA0844.2
31	JASPAR	CLOCK	ACACGTG	7	46454	2.65E-80	MA0819.3
31	JASPAR	HEY2	GRCACGTGY	9	30349	1.26E-56	MA0649.2
31	JASPAR	SREBF2	ATCACGTGAT	10	952	1.48E-55	MA0828.3
31	JASPAR	SOHLH2	GCACGTGC	8	14655	2.87E-26	MA1560.2
31	JASPAR	CREB3	TGCCACGTCAYC	12	5864	4.08E-26	MA0638.2
31	JASPAR	HES5	GRCACGTGYC	10	42092	3.64E-18	MA0821.2
31	JASPAR	ARNT2	GTCACGTG	8	15388	1.42E-15	MA1464.2
31	JASPAR	ARNT::HIF1A	ACGTG	5	28150	3.57E-15	MA0259.2
31	JASPAR	Creb3l2	GCCACGTGD	9	8679	8.86E-14	MA0608.1
31	JASPAR	EPAS1	CGCACGTAS	9	1573	1.88E-07	MA2325.1
32	JASPAR	TCF7	CTTTGAW	7	11611	0.00E+00	MA0769.3
32	JASPAR	Hnf1A	CCTTTGATST	10	11184	0.00E+00	MA1991.2
32	JASPAR	TCF7L1	AAAGATCAAAGG	12	5735	0.00E+00	MA1421.1
32	JASPAR	Lef1	CCTTTGAT	8	132153	9.45E-12	MA0768.3
33	JASPAR	NR4A2	AAAGGTCA	8	89476	0.00E+00	MA0160.3
33	JASPAR	NR4A1	AAAGGTCA	8	73686	0.00E+00	MA1112.3
33	JASPAR	Esrrg	TCAAGGTCA	9	63491	0.00E+00	MA0643.2
33	JASPAR	Nr1h3	AGGKCA	6	53779	0.00E+00	MA2337.1
33	JASPAR	NR1I3	TGAACTTT	8	6518	0.00E+00	MA1534.2
33	JASPAR	PPARA::RXRA	AWNTRGGTYAAAGGTCA	17	3146	0.00E+00	MA1148.2
33	JASPAR	NR2F2	AAGGTCA	7	68091	6.20E-279	MA1111.2
33	JASPAR	Nr2f6	RGGTCAAAGGTCA	13	9240	1.18E-204	MA0677.2
33	JASPAR	Nr5A2	TGACCTTGA	9	442	7.73E-73	MA0505.3
33	JASPAR	ESRRA	TCAAGGTCA	9	2209	2.45E-44	MA0592.4
33	JASPAR	NR5A1	AGTTCAAGGTCA	12	13754	2.50E-41	MA1540.3
33	JASPAR	Nr2e1	AAAAGTCAA	9	18963	5.47E-32	MA0676.1
33	JASPAR	NR2C2	GRGGTCARAGGTCA	14	4096	1.82E-23	MA0504.2
33	JASPAR	RARA	AGGTCAHSYAAAGGTCA	17	1253	1.87E-17	MA0730.1
33	JASPAR	THRB	RGGTCAAAGGTCA	13	14685	9.69E-16	MA1574.2
33	JASPAR	Ppara	AAGGTCA	7	24198	1.48E-11	MA2338.1
34	JASPAR	NR2C2	RGGTCA	6	233962	0.00E+00	MA1536.2
34	JASPAR	NR2C1	RGGTCA	6	158335	0.00E+00	MA1535.2
34	JASPAR	RARG	AGGTCATGACCTY	13	12709	0.00E+00	MA1553.2
34	JASPAR	RARA::RXRG	RGGTCAHNNNRRGGTCA	17	14044	9.45E-206	MA1149.2
34	JASPAR	RXRB	RRGGTCATGACCYY	14	14713	1.14E-182	MA1555.1
34	JASPAR	RARB	RAGGTCATGACCT	13	14846	8.21E-143	MA1552.2
34	JASPAR	RORC	AWNTRGGTCA	10	9396	2.48E-108	MA1151.2
34	JASPAR	RORB	AWWTRGGTCA	10	99047	5.85E-80	MA1150.2
34	JASPAR	RORA	AWMWAGGTCA	10	6272	1.00E-72	MA0071.1
34	JASPAR	NR1H4::RXRA	AGKTCATTGACCY	13	2936	1.03E-42	MA1146.2
34	JASPAR	NR4A2::RXRA	RGGTCRTTGACCY	13	15838	4.49E-33	MA1147.2
34	JASPAR	BCL11A	CTGACCA	7	66745	1.27E-19	MA2324.1
34	JASPAR	ESR1	RGGTCACSRTGACCT	15	62	4.28E-07	MA0112.4
34	JASPAR	ESR2	AGGTCASVNTGMCCY	15	62	4.28E-07	MA0258.2
35	JASPAR	ZNF184	KAGAAAGDNGMAT	13	44182	0.00E+00	MA2120.1
36	JASPAR	EGR4	MCGCCCACGCA	11	112791	0.00E+00	MA0733.2
36	JASPAR	EGR1	MCGCCCACGC	10	37863	0.00E+00	MA0162.5
36	JASPAR	EGR2	MCGCCCACGCA	11	19636	3.15E-12	MA0472.2
37	JASPAR	Vdr	GAGTTCA	7	7122	0.00E+00	MA0693.4
38	STREME	STREME-10	10-CTGACACCMK	10	2712	8.08E-54	MA0498.3 (MEIS1)
39	JASPAR	Gfi1B	AAATCWCWGC	10	7358	0.00E+00	MA0483.2
40	JASPAR	Rfx6	CCTAGCAAC	9	3287	0.00E+00	MA1724.2

40	JASPAR	RFX4	GTTGCYWRGCAAC	13	1352	1.60E-38	MA0799.3
41	STREME	STREME-7	CAMWTCYTCRMCDDM	15	4532	7.38E-99	MA1589.2 (ZNF140)
41	STREME	STREME-30	30-CCACATCCWCGCCAG	15	567	1.70E-15	MA1107.3 (KLF9)
42	JASPAR	GSC2	TAATCC	6	23571	0.00E+00	MA0891.2
42	JASPAR	OTX1	TAATCC	6	19291	0.00E+00	MA0711.2
42	JASPAR	Dmbx1	RWNMGGATTA	10	17611	0.00E+00	MA0883.2
42	JASPAR	PITX3	TAATCC	6	17379	0.00E+00	MA0714.2
42	JASPAR	GSC	TAATCC	6	9926	1.75E-114	MA0648.2
42	JASPAR	OTX2	RGGATTA	7	27512	1.30E-93	MA0712.3
42	JASPAR	Crx	GGATTA	6	32912	6.29E-83	MA0467.3
42	JASPAR	PITX1	TAATCC	6	6124	1.86E-26	MA0682.3
43	JASPAR	Zfp809	WTCCCAGCC	9	5137	0.00E+00	MA2125.1
44	JASPAR	ASCL1	GCAGCTGC	8	168101	0.00E+00	MA1100.3
44	JASPAR	TCF3	CACCTGC	7	13318	4.42E-217	MA0522.4
44	JASPAR	TFAP4	AWCAGCTGWT	10	18207	6.43E-200	MA0691.1
44	JASPAR	TCF12	CACCTGC	7	11988	5.79E-172	MA1648.2
44	JASPAR	MSC	AACAGCTGTT	10	12698	4.30E-153	MA0665.1
44	JASPAR	Tcf12	CAGCTG	6	51494	7.16E-116	MA0521.3
44	JASPAR	SNAI3	RCAGGTGYA	9	24871	1.40E-112	MA1559.2
44	JASPAR	NHLH2	GGNCGCAGCTGCGYCC	16	635	8.79E-54	MA1529.2
44	JASPAR	TGIF2LY	TGACAGCTGTCA	12	1183	2.92E-31	MA1572.1
44	JASPAR	TFAP4::ETV1	SCGGAAGCAGSTG	13	1806	4.33E-26	MA1966.2
44	JASPAR	ASCL1	GCACCTGCY	9	22975	1.17E-24	MA1631.2
44	JASPAR	Tcf21	AACAGCTGTT	10	2994	3.77E-21	MA0832.2
44	JASPAR	TGIF2LX	TGACAGCTGTCA	12	1257	6.27E-20	MA1571.1
44	JASPAR	MYB	CAACTG	6	38987	1.57E-18	MA0100.4
44	JASPAR	TCF4	GCACCTGC	8	34287	1.54E-13	MA0830.3
45	JASPAR	ZNF524	CTCGRACCC	9	11815	0.00E+00	MA2096.1
46	JASPAR	ZNF667	TTAAGAGCTCA	11	15954	0.00E+00	MA1984.2
47	JASPAR	Zic3	CAGCAGG	7	195963	0.00E+00	MA0697.3
47	JASPAR	Zic1::Zic2	CAGCAGG	7	17045	0.00E+00	MA1628.2
47	JASPAR	Zic2	CACAGCAGG	9	48570	4.18E-240	MA1629.2
48	JASPAR	ZBTB12	CTRGAAC	7	134443	0.00E+00	MA1649.2
49	JASPAR	MZF1	AATCCCCA	8	77770	0.00E+00	MA0056.3
50	STREME	STREME-12	12-AGGATCMAGY	10	3521	3.23E-49	MA1581.2 (ZBTB6)
51	STREME	STREME-11	11-AGTAGACGGCAGG	13	2120	8.01E-52	11-AGTAGACGGCAGG
52	STREME	STREME-9	CTCCCTCCCTCCCTC	15	2021	8.49E-56	MA1965.2 (SP5)
52	JASPAR	SP5	CCTCCC	6	44789	6.30E-45	MA1965.2
52	STREME	STREME-20	20-CCTCCCCCCCCCCCC	15	2299	8.67E-21	MA1630.3 (ZNF281)
52	STREME	STREME-54	54-CCCCCCCCCACTCC	14	1273	1.41E-06	MA0753.3 (ZNF740)
52	STREME	STREME-56	56-CCCCCCCCCCCRC	13	11068	1.94E-06	MA1630.3 (ZNF281)
53	STREME	STREME-13	13-GGYACTGCA	9	5634	1.02E-47	MA0019.2 (Ddit3::Cebpa)
54	STREME	STREME-58	58-AGCTCCAAAC	10	369	3.81E-06	58-AGCTCCAAAC
55	STREME	STREME-14	14-AGTTCTTTATA	11	2334	1.31E-42	14-AGTTCTTTATA
56	JASPAR	ZNF549	TGCTGCCC	8	24263	0.00E+00	MA1728.2
57	JASPAR	ZNF135	CCTCGACCTCCYRR	14	8084	0.00E+00	MA1587.1
57	JASPAR	ZNF460	GCCTCMGCCTCCCRAG	16	8380	1.14E-31	MA1596.1
58	JASPAR	Ddit3::Cebpa	RTGCAATMCC	10	60460	0.00E+00	MA0019.2
59	JASPAR	Zfp961	GGGCGCCA	8	111769	0.00E+00	MA2126.1
59	JASPAR	ZNF417	GGCGCCA	7	42965	6.40E-174	MA1727.2
60	JASPAR	HLF	TTATGCAAC	9	4121	0.00E+00	MA0043.4
60	JASPAR	NFIL3	TTATGYAAT	9	4358	3.50E-229	MA0025.3
60	JASPAR	TEF	RTTACRTAAY	10	3310	1.74E-138	MA0843.2
60	JASPAR	DBP	RTTACGTAAY	10	3671	8.71E-86	MA0639.2
60	JASPAR	CEBPA	ATTGCACAAT	10	2622	6.06E-72	MA0102.5
60	JASPAR	CEBPD	TTGCACAA	8	7897	2.31E-39	MA0836.3
60	JASPAR	CEBPG	ATTRCGCAAY	10	7859	1.19E-10	MA0838.1
61	JASPAR	POU3F2	WTATGCWAATKA	12	21898	0.00E+00	MA0787.1
61	JASPAR	POU2F3	TATGCAAAT	9	10508	7.31E-267	MA0627.3
61	JASPAR	POU3F1	WTATGCWAAT	10	24686	1.84E-230	MA0786.2
61	JASPAR	POU5F1B	TATGCWAAT	9	24885	1.96E-150	MA0792.1
61	JASPAR	POU2F1	TATGCWAAT	9	22394	4.68E-115	MA0785.2
61	JASPAR	Pou5f1::Sox2	CWTTGTYATGCAAAT	15	4928	1.28E-112	MA0142.1
61	JASPAR	POU2F2	WTATGCAAATKAG	13	1811	5.82E-97	MA0507.3
61	JASPAR	POU2F1::SOX2	MATTTRCATMACAATRG	17	5872	1.20E-90	MA1962.1
61	JASPAR	POU3F4	TATGCWAAT	9	14786	3.25E-86	MA0789.1
61	JASPAR	POU3F3	WWTATGCWAATTW	13	26018	6.12E-23	MA0788.1
62	JASPAR	HNF4A	CAAAGTCCA	9	11992	0.00E-02	MA0114.5

62	JASPAR	HNF4G	CAAAGTCCA	9	11062	2.24E-178	MA0484.3
63	JASPAR	PAX6	TTCACGCWTSANTK	14	3067	0.00E-02	MA0069.1
63	JASPAR	PAX2	BCAVTSRAGCGTGACG	16	280	1.77E-08	MA0067.3
64	STREME	STREME-16	16-TCAACGAATTAGAA	14	1439	8.66E-32	16-TCAACGAATTAGAA
65	STREME	STREME-48	48-TCCCTTGTCGAAAA	14	302	9.32E-08	48-TCCCTTGTCGAAAA
66	JASPAR	Tbx6	RGGTGTGAA	9	62411	1.12E-283	MA1567.3
66	JASPAR	TBX2	AGGTGTGAA	9	4432	1.74E-272	MA0688.2
66	JASPAR	TBX15	AGGTGTGA	8	17518	2.07E-253	MA0803.1
66	JASPAR	TBX18	AGGTGTGAA	9	63779	1.06E-242	MA1565.2
66	JASPAR	TBX21	TTCACACCTT	10	19389	1.34E-242	MA0690.3
66	JASPAR	TBX5	AGGTGTKA	8	10777	1.08E-223	MA0807.1
66	JASPAR	TBX3	AGGTGTGAA	9	5616	7.69E-181	MA1566.3
66	JASPAR	TBR1	AGGTGTGAA	9	1139	4.97E-139	MA0802.2
66	JASPAR	TBX1	AGGTGTGA	8	34435	1.32E-129	MA0805.1
66	JASPAR	EOMES	AGGTGTGAA	9	1172	4.19E-128	MA0800.2
66	JASPAR	TBX4	AGGTGTGA	8	36240	2.40E-93	MA0806.1
66	JASPAR	MGA	AGGTGTGA	8	79918	1.66E-83	MA0801.1
66	JASPAR	TBXT	TCACACMTAKGTGTGA	16	4247	1.29E-72	MA0009.2
66	JASPAR	MGA::EVX1	RGGTGWTAATK	11	2280	8.43E-50	MA1960.2
66	JASPAR	TBX19	TMRCACMTAGGTGTGAA	17	3641	8.42E-48	MA0804.2
66	JASPAR	TBX20	WAGGTGTGAAR	11	8903	2.23E-27	MA0689.1
67	STREME	STREME-18	18-CACCCGMAAR	10	1165	2.48E-26	MA0641.1 (ELF4)
68	JASPAR	TFAP2A	TGCCCYSRGGGCA	13	16469	1.20E-277	MA0872.1
68	JASPAR	TFAP2B	TGCCCYBRGGGCA	13	14004	3.49E-226	MA0813.1
68	JASPAR	TFAP2C	TGCCCYSRGGGCA	13	13476	8.91E-203	MA0815.1
69	STREME	STREME-19	19-TAGAGGGCACAGGGA	15	920	4.67E-25	MA2337.1 (Nr1h3)
70	JASPAR	POU6F2	ASCTMATTA	9	4297	4.56E-262	MA0793.2
70	JASPAR	PAX4	CTAATTAG	8	2280	4.61E-193	MA0068.2
70	JASPAR	GSX2	SYMATTA	7	2358	1.40E-184	MA0893.3
70	JASPAR	LMX1B	TTAATTAA	8	2340	3.48E-176	MA0703.3
70	JASPAR	VSX1	YTAATTA	7	2326	1.84E-170	MA0725.2
70	JASPAR	MEOX1	STAATTA	7	2404	1.86E-169	MA0661.2
70	JASPAR	VSX2	CTAATTA	7	2333	2.46E-169	MA0726.2
70	JASPAR	Isl1	CCATTAG	7	82501	4.79E-148	MA1608.2
70	JASPAR	MEOX2	STAATTA	7	2444	2.35E-144	MA0706.2
70	JASPAR	Lhx1	GCTAATTAGC	10	1087	1.61E-135	MA1518.3
70	JASPAR	BARX2	WAAYMATTA	9	2333	2.19E-133	MA1471.2
70	JASPAR	ARGFX	CTAATTAR	8	1011	5.40E-94	MA1463.2
70	JASPAR	mix-a	TTAATTA	7	2988	5.38E-90	MA0621.2
70	JASPAR	POU6F1	TAATGAG	7	8418	1.06E-88	MA1549.2
70	JASPAR	LHX6	CTAATTAR	8	2319	2.36E-75	MA0658.2
70	JASPAR	PRRX2	CTAATTA	7	2342	6.61E-67	MA0075.4
70	JASPAR	NOTO	YTAATTA	7	2532	7.02E-61	MA0710.2
70	JASPAR	VAX1	YTAATTA	7	2633	2.35E-56	MA0722.2
70	JASPAR	EVX1	TAATTA	6	6743	2.76E-56	MA0887.2
70	JASPAR	EVX2	TAATTA	6	6753	6.91E-54	MA0888.2
70	JASPAR	GBX1	CYAATTA	7	4996	2.86E-50	MA0889.2
70	JASPAR	DRGX	YAATTA	6	2836	4.50E-48	MA1481.2
70	JASPAR	NKX6-2	YMATTA	6	4881	8.74E-47	MA0675.2
70	JASPAR	LMX1A	TTAATTA	7	2571	6.95E-46	MA0702.3
70	JASPAR	GSX1	YMATTA	6	4748	5.36E-45	MA0892.2
70	JASPAR	HOXD4	TMATTA	6	15102	2.65E-39	MA1507.2
70	JASPAR	TLX2	YAATTA	6	2867	1.04E-38	MA1577.2
70	JASPAR	GBX2	YAATTA	6	4822	4.01E-37	MA0890.2
70	JASPAR	POU6F1	TAATTA	6	9502	1.40E-35	MA0628.2
70	JASPAR	RAX	YAATTA	6	3128	1.25E-34	MA0718.2
70	JASPAR	UNCX	YAATTA	6	2879	2.98E-34	MA0721.2
70	JASPAR	BSX	YVATTA	6	6419	4.13E-33	MA0876.2
70	JASPAR	VAX2	YAATTA	6	18639	7.45E-32	MA0723.3
70	JASPAR	PRRX1	YAATTA	6	2892	2.42E-31	MA0716.2
70	JASPAR	HOXB3	YMATTA	6	5537	2.84E-31	MA0903.2
70	JASPAR	ALX3	YAATTA	6	2892	2.93E-31	MA0634.2
70	JASPAR	Shox2	YAATTA	6	2933	7.66E-30	MA0720.2
70	JASPAR	MIXL1	YAATTA	6	2896	1.69E-29	MA0662.2
70	JASPAR	HOXC4	TMATTA	6	15848	1.58E-28	MA1504.2
70	JASPAR	BARX1	CMATTA	6	15609	2.60E-28	MA0875.2
70	JASPAR	HOXD3	CTAATTAC	8	10323	1.76E-25	MA0912.2
70	JASPAR	HOXB4	TMATTA	6	16526	5.45E-25	MA1499.2

70	JASPAR	LHX9	CYAATTA	7	17911	2.24E-24	MA0701.3
70	JASPAR	EN2	SYAATTA	7	12181	2.64E-24	MA0642.3
70	JASPAR	HOXA7	YMATTA	6	3699	6.62E-22	MA1498.3
70	JASPAR	HOXB1	STAATTA	7	5566	3.50E-20	MA2093.1
70	JASPAR	DLX6	YAATTA	6	5553	1.99E-14	MA0882.2
70	JASPAR	SHOX	TAATTR	6	5284	1.56E-11	MA0630.2
70	JASPAR	HOXA2	TAATTA	6	11508	2.86E-11	MA0900.3
70	JASPAR	HOXA4	RTMATTA	7	6500	6.39E-10	MA1496.2
70	JASPAR	Arx	CRYTAATTAR	10	3834	2.86E-09	MA0874.2
70	JASPAR	HOXB6	TAATKRC	7	17594	1.01E-08	MA1500.2
70	JASPAR	HOXA5	GYMATTAS	8	7342	3.16E-08	MA0158.2
70	JASPAR	HOXB5	TAATTA	6	12602	3.84E-06	MA0904.3
71	JASPAR	RARA::RXRA	RGKTCANVGRSAGGTCA	17	7848	1.61E-260	MA0159.1
71	JASPAR	Pparg::Rxra	RGGGCARAGGKCA	13	3418	5.92E-217	MA0065.3
71	JASPAR	THRB	TGACCTBRNYVAGGTCA	17	832	5.53E-14	MA1575.2
72	JASPAR	NR6A1	CAAGKTCAAGKTCA	14	1369	2.34E-248	MA1541.2
73	JASPAR	ZNF530	GMARGGMRAGGGGC	14	30063	6.62E-239	MA1981.2
74	JASPAR	ZNF354C	MTCCAC	6	16373	1.37E-236	MA0130.1
75	JASPAR	RHOXF1	TRATCC	6	13730	3.88E-228	MA0719.2
76	JASPAR	ZNF680	CCAAGAAGAAT	11	2340	7.80E-228	MA1729.2
77	STREME	STREME-17	17-GCAATCKGCARAT	13	1193	1.13E-26	17-GCAATCKGCARAT
78	JASPAR	Plagl1	TGGGGCCA	8	109789	4.02E-217	MA1615.2
79	JASPAR	Gmeb1	KACGTM	6	105504	1.54E-216	MA0615.2
80	JASPAR	PGR	ACANNNTGT	9	46238	1.51E-207	MA2327.1
81	JASPAR	ZBED2	CGAAACC	7	22375	7.99E-207	MA1971.2
82	STREME	STREME-32	32-AGGTGTGTKGG	11	1166	2.33E-15	MA0803.1 (TBX15)
83	STREME	STREME-21	21-CACGACAAAAC	11	903	1.12E-20	21-CACGACAAAAC
84	JASPAR	THAP1	GCAGGGCA	8	36819	2.72E-202	MA0597.3
85	JASPAR	MYBL2	RRCCGTTAAACBGYY	15	2267	5.76E-195	MA0777.1
85	JASPAR	MYBL1	ACCGTTAACSGY	12	2959	1.04E-190	MA0776.1
86	JASPAR	ZNF677	RATAAGAACAGC	12	56391	1.06E-192	MA2101.1
87	JASPAR	MEIS1	TGACA	5	122495	2.50E-186	MA0498.3
87	JASPAR	MEIS3	DTGACAG	7	13381	3.35E-22	MA0775.2
88	JASPAR	ZNF274	TRTGAGTTCTCG	12	2058	9.88E-184	MA1592.2
89	JASPAR	Prdm5	GTTCTCCATCT	11	2647	1.85E-181	MA1999.2
90	STREME	STREME-28	28-CCCATAATCAGCWTC	15	826	1.28E-16	MA0752.2 (ZNF410)
91	JASPAR	ZNF684	ACAGTCCACCCCTT	14	969	9.66E-57	MA1600.2
91	STREME	STREME-34	34-AAGGGCTGGAT	11	890	1.40E-14	MA1600.2 (ZNF684)
92	JASPAR	GATA4	CCTTATCT	8	63783	3.22E-178	MA0482.3
92	JASPAR	GATA1	CTAATCT	7	6120	5.79E-41	MA0035.5
92	JASPAR	Gata3	TCTTATCT	8	326	2.15E-25	MA0037.5
92	JASPAR	GATA6	TCTTATCT	8	329	1.11E-24	MA1104.3
92	JASPAR	TRPS1	TCTTATCT	8	337	2.04E-21	MA1970.2
92	JASPAR	GATA5	WGATAASR	8	2198	1.07E-17	MA0766.3
92	JASPAR	GATA2	CTTATCT	7	753	3.37E-12	MA0036.4
93	JASPAR	ZBTB6	SCTTGAGCC	9	1551	1.16E-176	MA1581.2
94	JASPAR	HOXD12	RGTCGTAAAA	10	2268	1.76E-176	MA0873.2
94	JASPAR	HOXC13	CTCGTAAAA	9	3190	4.40E-82	MA0907.2
94	JASPAR	HOXC12	RGTCGTAAAA	10	6407	2.33E-31	MA0906.2
94	JASPAR	Hoxa11	GTCGTAAAA	9	6849	2.96E-23	MA0911.2
94	JASPAR	HOXC10	GTCRTAAAA	9	6945	3.71E-11	MA0905.2
95	JASPAR	TBP	TATAAAW	7	25029	5.91E-175	MA0108.3
96	JASPAR	ZBED4	CCCGCYCCGC	10	6355	7.39E-172	MA2328.1
97	JASPAR	Sox17	AGAACAATGG	10	2183	5.63E-169	MA0078.3
97	JASPAR	Sox7	AGAACAATGG	10	2284	1.65E-123	MA2095.1
97	JASPAR	SOX14	CGAACAATG	9	2603	8.16E-56	MA1562.2
97	JASPAR	SOX13	ACAATGG	7	50133	3.02E-43	MA1120.2
97	JASPAR	Sox3	ACAATGG	7	52078	2.74E-27	MA0514.3
97	JASPAR	SOX2	ACAATGG	7	54576	2.40E-23	MA0143.5
97	JASPAR	SOX18	AACAATDV	8	2786	4.84E-17	MA1563.2
98	STREME	STREME-24	24-CGTGTGGRCC	10	783	1.74E-18	24-CGTGTGGRCC
99	MEME	MEME-2	HMAAAAAAAAAAAAM	15	389127	8.40E-93	MA1978.2 (ZNF354A)
99	JASPAR	FOXD3	WAWGHAAATAAACA	14	2754	1.64E-19	MA0041.3
99	JASPAR	ZNF354A	WAADWATAAATGRAYWAWTT	20	22104	1.24E-18	MA1978.2
100	JASPAR	RELB	ATTCCCC	7	121494	1.09E-163	MA1117.2
101	STREME	STREME-40	40-GTTAAGGA	8	348	5.60E-11	40-GTTAAGGA
102	STREME	STREME-61	61-TGCATACA	8	950	7.62E-05	61-TGCATACA
103	JASPAR	Dmrt1	TACAAAGTA	9	2547	1.82E-157	MA1603.2

104	JASPAR	ZFP57	TGCCGCA	7	61359	2.00E-157	MA1583.2
105	JASPAR	Tfcp2l1	CCAGYYYVADCCRG	14	72182	1.51E-152	MA0145.2
106	JASPAR	DMRTA2	GHTACA	6	118469	1.64E-151	MA1478.2
107	STREME	STREME-25	25-ACTTCGCCCAGTCCA	15	755	4.40E-17	25-ACTTCGCCCAGTCCA
108	STREME	STREME-47	47-GATTCAA	7	1734	2.32E-08	MA0611.3 (Dux)
109	JASPAR	ZBTB7C	CGACCACC	8	148955	1.75E-139	MA0695.2
109	JASPAR	Gli2	GACCACCCA	9	20291	2.64E-84	MA0734.4
109	JASPAR	ZBTB7B	CGACCACCGA	10	34361	1.03E-37	MA0694.2
110	STREME	STREME-35	35-ATATCCAG	8	3204	1.83E-14	35-ATATCCAG
111	JASPAR	ZKSCAN1	AYAGTAGGT	9	2071	5.04E-138	MA1585.2
112	JASPAR	PAX3	SGTCACGSYWATTA	14	4170	1.30E-136	MA1546.2
113	JASPAR	Wt1	CCTCCCCCAC	10	140193	1.81E-203	MA1627.2
113	STREME	STREME-29	29-CGACTCCCCCACC	13	645	9.42E-16	MA1627.2 (Wt1)
114	JASPAR	ZNF449	AAGCCCAACC	10	11706	8.33E-131	MA1656.2
115	JASPAR	RFX7	CGTTGCYA	8	2703	1.71E-130	MA1554.2
116	STREME	STREME-55	55-GGATCCACCCC	11	263	1.84E-06	55-GGATCCACCCC
117	JASPAR	ZNF784	ACYTACCG	8	26886	5.33E-123	MA1717.2
118	JASPAR	ZNF770	CGGCCTCA	8	121151	4.14E-122	MA2099.1
119	JASPAR	ZNF410	MCATCCCATAATAHTC	16	545	6.49E-119	MA0752.2
120	JASPAR	ZNF75A	GCTTTTCCCACA	12	2156	7.60E-119	MA2097.1
121	JASPAR	ZNF343	CCGCTTCMCCDCGGCM	16	1396	6.38E-117	MA1711.2
122	STREME	STREME-39	39-GAGGGAGAAACCAAA	15	566	2.50E-11	39-GAGGGAGAAACCAAA
123	JASPAR	ZNF582	TCTGTTACTTGCAGCCAAA	19	2051	2.07E-114	MA1983.2
124	STREME	STREME-44	44-ATAGGGCGGACCT	13	565	7.46E-09	MA0131.3 (HINFP)
125	JASPAR	Zfx	GCCBVGGCCT	10	24744	6.79E-113	MA0146.3
126	JASPAR	ZBTB17	AATCGATT	8	65180	1.64E-112	MA2102.1
126	JASPAR	Pax7	TAATCAATTA	10	1663	1.51E-80	MA0680.3
126	JASPAR	PAX3	TAATYRATTA	10	1698	2.04E-42	MA0780.1
126	JASPAR	CUX1	TAATCGATA	9	152	8.50E-11	MA0754.3
127	STREME	STREME-49	49-GRGGAACCGMC	11	672	2.56E-07	MA1725.2 (ZNF189)
128	JASPAR	Hmga1	ATTTTTAW	8	15354	4.04E-111	MA2124.1
129	JASPAR	ZNF652	AAGRGTTAA	9	3426	1.74E-109	MA1657.2
130	JASPAR	Ebf4	TCCCCAGGGGA	11	10681	9.04E-109	MA2122.1
130	JASPAR	Ebf2	CCCAAGGGA	9	4838	2.38E-72	MA1604.2
130	JASPAR	EBF1	TCCCCAGGGGA	11	12365	1.56E-71	MA0154.5
130	JASPAR	EBF3	CCCAAGGGA	9	6213	1.10E-54	MA1637.2
131	JASPAR	ISL2	CAMTTA	6	187083	3.31E-100	MA0914.2
132	JASPAR	SIX2	TGAAACCTGAT	11	64879	2.31E-98	MA1119.2
132	JASPAR	SIX1	GWAACCTGA	9	3078	3.73E-93	MA1118.2
132	JASPAR	Six4	AACCTGA	7	1395	4.11E-79	MA2001.2
133	STREME	STREME-36	36-ATGCCCAGGA	10	2647	9.73E-14	MA1979.2 (ZNF416)
134	JASPAR	TEAD1	ACATTCCAG	9	33761	6.85E-95	MA0090.4
134	JASPAR	TEAD3	RCATTCCW	8	13870	7.97E-48	MA0808.1
134	JASPAR	TEAD4	ACATTCCA	8	16731	1.92E-27	MA0809.3
135	JASPAR	FOS::JUN	ATGAGTCAY	9	4086	1.89E-93	MA0099.4
135	JASPAR	FOSL2::JUN	RTGAGTCAY	9	4045	6.41E-91	MA1130.2
135	JASPAR	FOSL2::JUND	ATGACTCAT	9	4676	1.66E-85	MA1144.2
135	JASPAR	FOSL2	RRTGASTCAB	10	5217	8.51E-83	MA0478.2
135	JASPAR	FOSL2::JUNB	RTGASTCAT	9	4809	7.72E-79	MA1138.2
135	JASPAR	FOS::JUND	ATGAGTCAT	9	4183	2.44E-76	MA1141.2
135	JASPAR	FOSB::JUNB	RTGASTCAT	9	4774	6.22E-72	MA1135.2
135	JASPAR	BNC2	TGAGTCA	7	24407	2.10E-67	MA1928.2
135	JASPAR	BACH1	ATGACTCAT	9	4531	3.09E-67	MA1633.2
135	JASPAR	NFE2	ATGACTCATS	10	2193	3.04E-66	MA0841.2
135	JASPAR	JUN::JUNB	ATGACKCA	8	4741	9.91E-58	MA1132.2
135	JASPAR	FOSL1::JUN	ATGACTCAT	9	4504	1.99E-52	MA1128.2
135	JASPAR	Atf3	TGACTCA	7	53331	7.20E-52	MA1988.2
135	JASPAR	FOS::JUNB	ATGASTCAT	9	7907	1.91E-42	MA1134.2
135	JASPAR	FOSL1::JUNB	RTGACTCAT	9	5657	5.55E-41	MA1137.2
135	JASPAR	MAFK	MTGACTCAGC	10	5923	1.75E-34	MA0496.4
135	JASPAR	BATF3	TGACTCA	7	20921	1.47E-33	MA0835.3
135	JASPAR	Jun	TGACTCAT	8	27303	1.91E-33	MA0489.3
135	JASPAR	FOSL1::JUND	RTGACTCA	8	4785	7.45E-30	MA1142.2
135	JASPAR	BATF::JUN	TGACTCA	7	18381	1.96E-28	MA0462.3
135	JASPAR	BATF	TGACTCA	7	1520	1.69E-22	MA1634.2
135	JASPAR	Nfe2l2	ATGACTCAGCA	11	6697	3.28E-15	MA0150.3
135	JASPAR	FOSL1	ATGACTCAT	9	14364	7.45E-09	MA0477.3
135	JASPAR	JUND	ATGACTCAT	9	8288	4.58E-06	MA0491.3

136	JASPAR	ZNF148	CCCCTCCCCC	10	128502	0.00E+00	MA1653.2
136	JASPAR	VEZF1	CCCCCC	6	129998	9.61E-198	MA1578.2
136	MEME	MEME-3	SGGGGGGGGGGGGGG	15	16917	3.20E-154	MA0753.3 (ZNF740)
136	JASPAR	ZNF740	CCCCCCCCAC	10	93599	3.88E-129	MA0753.3
136	JASPAR	RREB1	CCCCMAAMCAMCCMCMMMC	19	1833	1.27E-30	MA0073.2
136	JASPAR	PRDM9	RGDGGGVAGGGRGGVRRMA	20	19583	3.46E-22	MA1723.2
137	JASPAR	REST	TCAGCACCATGGACAGCDCC	20	1624	4.27E-87	MA0138.3
138	STREME	STREME-45	45-AAATGGCTATCT	12	428	1.42E-08	45-AAATGGCTATCT
139	STREME	STREME-41	41-CCCATATYA	9	321	8.98E-10	MA0752.2 (ZNF410)
139	STREME	STREME-43	43-CCCAKATTAA	10	683	6.91E-09	MA0752.2 (ZNF410)
140	JASPAR	POU4F3	TGMATWATTAAT	12	4622	2.25E-83	MA0791.2
140	JASPAR	HNF1A	RTTAATNATTAAC	13	2597	1.27E-76	MA0046.3
140	JASPAR	POU4F1	TGMATAATTAAT	12	2273	4.04E-48	MA0790.2
140	JASPAR	HNF1B	GTTAATNATTAAY	13	2684	4.25E-32	MA0153.2
140	JASPAR	Arid3b	TATTAAT	7	139	8.70E-09	MA0601.2
141	STREME	STREME-42	42-CGGGGCTCGAGT	12	392	4.49E-09	MA1581.2 (ZBTB6)
142	JASPAR	Stat5b	TTCCCAGAA	9	16641	6.62E-76	MA1625.2
142	JASPAR	RBPJ	TGGGAA	6	31924	1.82E-52	MA1116.2
142	JASPAR	Stat5a	TTCCAAGAA	9	3002	6.64E-47	MA1624.2
142	JASPAR	Stat4	TTCYRGGAAR	10	3250	1.99E-35	MA0518.2
142	JASPAR	BCL6	GCTTTCKAGGAAY	13	2865	6.94E-22	MA0463.3
143	JASPAR	TFCP2	AAACCGGTT	9	42143	5.51E-73	MA1968.2
143	JASPAR	GRHL2	AACAGGTT	8	1771	5.88E-34	MA1105.3
143	JASPAR	GRHL1	AAACCGGTTT	10	2882	6.79E-31	MA0647.2
144	STREME	STREME-38	38-AATAGAAGCAGT	12	425	5.46E-12	38-AATAGAAGCAGT
145	JASPAR	YY2	ATGGCGG	7	8417	2.49E-71	MA0748.3
146	JASPAR	Arid3a	ATYAAA	6	1782	3.24E-70	MA0151.1
147	JASPAR	Foxq1	ATTGTTTATW	10	4084	5.26E-68	MA0040.2
147	JASPAR	Foxj2	RTAAACAA	8	1259	6.98E-07	MA0614.1
148	JASPAR	Ahr::Arnt	GCGTG	5	60010	1.34E-67	MA0006.2
149	JASPAR	ZNF416	TATCTGGGCA	10	619	1.87E-67	MA1979.2
150	JASPAR	Nanog	GCAATCA	7	87453	1.47E-66	MA2339.1
151	JASPAR	Znf423	GSMMCCYARGGKKBM	15	4922	2.63E-59	MA0116.1
152	STREME	STREME-46	46-AAAGAAAAGGGGGAK	15	283	1.44E-08	46-AAAGAAAAGGGGGAK
153	JASPAR	ZNF136	RWATTCTTGGTTGRC	15	2774	1.38E-57	MA1588.1
154	JASPAR	ZBTB24	CCCAGGACCC	10	2407	1.34E-55	MA2330.1
155	STREME	STREME-62	62-TAGAAAAAACA	11	490	1.11E-03	MA1125.2 (ZNF384)
156	JASPAR	HOXA6	GYMATTA	7	3883	1.04E-54	MA1497.2
156	JASPAR	LHX2	CAATTA	6	3052	1.19E-34	MA0700.3
156	JASPAR	HOXA9	RTCGTWA	7	128445	2.04E-34	MA0594.3
156	JASPAR	Dlx2	GCAATTAG	8	2401	4.73E-32	MA0885.3
156	JASPAR	LBX2	YAATTA	6	2893	3.02E-31	MA0699.2
156	JASPAR	RAX2	YAATTA	6	2897	1.74E-29	MA0717.2
156	JASPAR	Dlx5	GCAATTAG	8	2423	5.45E-27	MA1476.3
156	JASPAR	Hmx3	AGCAATTAA	9	3362	3.88E-22	MA0898.2
156	JASPAR	Dlx3	YAATTA	6	5665	3.04E-08	MA0880.2
157	STREME	STREME-51	51-GGACTCCCCA	10	341	6.01E-07	MA0105.4 (NFKB1)
158	JASPAR	ZNF263	GGGAGGA	7	19451	1.64E-53	MA0528.3
159	JASPAR	MEIS2	TTGACAGS	8	9916	4.86E-53	MA0774.1
160	JASPAR	Prdm15	GAAAACCTGGA	11	31039	9.06E-47	MA1616.2
161	JASPAR	BARHL1	CGTTTA	6	51580	2.30E-46	MA0877.4
162	JASPAR	Prdm14	GGTCTCTA	8	5742	3.17E-46	MA1998.2
163	JASPAR	Nfatc2	AATGGAAA	8	37600	3.52E-44	MA0152.3
163	JASPAR	Nfat5	ATGGAAAA	8	384	2.53E-15	MA0606.3
163	JASPAR	Nfatc1	TGGAAA	6	72006	2.47E-08	MA0624.3
164	JASPAR	ZSCAN21	AAGCACT	7	154702	1.32E-41	MA2336.1
165	JASPAR	Rhox11	CGCTGTWAW	9	3499	1.04E-38	MA0629.2
166	JASPAR	ZNF35	AATTCTA	7	19472	8.45E-37	MA2333.1
167	JASPAR	MAFA	TGCTGASTCAGCA	13	2201	1.38E-36	MA1521.2
167	JASPAR	MAF	TGCTGASTCAGCA	13	2230	6.89E-19	MA1520.2
168	JASPAR	NRL	AWWNTGCTGACG	12	3324	1.95E-33	MA0842.3
168	JASPAR	Mafb	AWWNTGCTGAC	11	2596	4.03E-25	MA0117.3
168	JASPAR	MAFF	GTCAGCATTTT	11	4319	5.58E-08	MA0495.4
169	JASPAR	HINFP	RCGTCCGC	8	1695	1.93E-32	MA0131.3
170	JASPAR	TFAP2A	GCCTCAGGC	9	88646	1.01E-30	MA0003.5
171	JASPAR	Npas4	TCGTGAC	7	107474	2.98E-30	MA1995.2
172	JASPAR	HSF2	TTCTAGAAYRTTC	13	3837	4.52E-30	MA0770.1
173	JASPAR	IRF5	CCGAAACCGAAACY	14	12731	7.46E-30	MA1420.1

173	JASPAR	IRF7	CGAAARYGAAAVT	13	2031	8.06E-18	MA0772.2
173	JASPAR	IRF3	RRAAMGGAAACCGAAAC	17	258	2.19E-12	MA1418.2
174	JASPAR	Zfp335	TCAGGCA	7	3041	1.71E-29	MA2002.2
175	JASPAR	ZNF341	GAACAGCC	8	155554	1.33E-26	MA1655.2
176	JASPAR	ZNF257	GAGGCRAGRG	10	5644	3.84E-26	MA1710.2
177	JASPAR	ZNF214	TCATCAABGTCCT	13	2924	7.78E-25	MA1975.2
178	JASPAR	SMAD2	CCAGAC	6	674	1.88E-24	MA1964.2
179	JASPAR	Stat6	TTCCWSAGAA	10	343	2.00E-22	MA0520.2
180	JASPAR	ZNF692	GGGCCCAS	8	269	1.01E-20	MA1986.2
181	JASPAR	NKX2-2	CCACTCAA	8	2627	1.18E-20	MA1645.2
181	JASPAR	NKX2-5	CACTCAA	7	3229	3.31E-19	MA0063.3
182	JASPAR	NFYA	CCAATCAG	8	28750	1.46E-20	MA0060.4
182	JASPAR	NFYC	CCAATCA	7	25361	5.90E-20	MA1644.2
183	JASPAR	MEF2B	RCTAWAAATAGC	12	88	4.78E-20	MA0660.1
183	JASPAR	MEF2D	DCTAWAAATAGM	12	88	4.78E-20	MA0773.1
183	JASPAR	MEF2A	CTAAAAATAG	10	114	7.72E-11	MA0052.5
184	JASPAR	ZKSCAN5	GGARGTGAG	9	1338	1.75E-18	MA1652.2
185	JASPAR	Foxl2	WATGTAAACA	10	35204	2.66E-17	MA1607.2
186	JASPAR	ZSCAN29	YGTCTACRCNG	11	327	2.95E-17	MA1602.2
187	JASPAR	SOX12	ACCGAACAAT	10	1272	6.30E-17	MA1561.2
188	JASPAR	ZNF24	CATTCATTCATTC	13	4798	1.01E-15	MA1124.1
189	JASPAR	REL	BGGRNWTTCC	10	29971	1.49E-15	MA0101.1
190	JASPAR	E2F2	WTTTGGCGCCAWW	13	834	5.84E-15	MA0864.3
190	JASPAR	E2F3	TTTTGGCGCCAAAA	14	1078	1.59E-06	MA0469.4
191	JASPAR	Hoxa13	CCAATAAA	8	7934	5.54E-14	MA0650.4
192	JASPAR	ZNF211	YATATACCAY	10	16908	3.02E-12	MA1974.2
193	JASPAR	Prdm4	CCTTGAAACYG	11	5919	3.54E-12	MA1647.3
194	JASPAR	ZNF282	CTTTCCCMYAACACG	15	225	4.57E-12	MA1154.2
195	JASPAR	OSR1	GCTACYGT	8	10895	5.25E-11	MA1542.2
196	JASPAR	ZKSCAN3	CCCAGGCTAGCCCA	14	1292	1.03E-10	MA1973.2
197	JASPAR	Nrf1	CTGCGCMTGCGC	12	108	1.66E-08	MA0506.3
198	JASPAR	Arid5a	YAATATTG	8	7001	1.28E-07	MA0602.2
199	JASPAR	IRF6	ACCGAAACT	9	4144	7.77E-07	MA1509.1
200	JASPAR	CEBPB	ATTGCGCAAT	10	549	9.20E-07	MA0466.4
201	JASPAR	ZNF140	AGGAGYGGAATTGCTGGGT	19	724	3.09E-06	MA1589.2