UNIVERSITAT DE BARCELONA

Exploring the chromatin landscape and gene expression mechanisms

Alba Sala Huerta

Exploring the chromatin landscape and gene regulatory mechanisms

Alba Sala Huerta

Universitat de Barcelona

Faculty of Biology

Doctoral program in Biomedicine

Exploring the chromatin landscape and gene expression mechanisms

The work presented in this thesis has been carried out in the Molecular Modelling and Bioinformatics laboratory of the Institute de Recerca Biomèdica (IRB Barcelona)

Thesis submitted by

Alba Sala Huerta

to qualify for the

Doctorate degree by the Universitat de Barcelona

Director and Tutor		Co-Director
Modesto Orozco López	Alba Sala Huerta	Federica Battistini

Acknowledgements

This thesis would have not been possible without the help from many people, and it is difficult to express my gratitude with just a few words. I would like to start by thanking my supervisor Modesto, who gave me the opportunity to undertake this thesis at his lab. Gràcies Modesto per donar-me la llibertat d'explorar noves idees, i per ajudar-me sempre a reconduir la història. Gràcies per tota la supervisió, l'ajuda i la constant motivació per la ciència que tant m'han inspirat.

Next, I want to thank my other supervisor without who this whole experience wouldn't have had any sense. Fede, gracias por estar siempre allí, guiándome y ayudándome cuando algo no tiraba hacia delante. Gracias por dejarme formar parte del dúo dinámico y hacer que esta tesis haya sido un placer. For all the guidance and support from day one, this thesis is as much yours as it is mine!

Thirdly, to my two non-official supervisors. Isabelle, thank you for always being there to explain the biological relevance of the results and to motivate me to make sense of the data. Adam, gràcies no només per formar part del meu procés de selecció (sense la nostra entrevista no hi seria aquí) sinó per sempre ajudar en qualsevol cosa que necessités encara que estiguessis més ocupat que ningú.

I also want to thank everybody in the MMB lab, from the first generation of people to the last one. To my current lab mates, the Modest Family FC, Luca, Milosz, Kazi, Vero, Santi, Isra, David, Paula, Jose, Subhamoy, Agus, Javier, Dani, Elena, Montse…This is the happiest I have ever been at a workplace, and it is thanks to all of you. Meeting you guys has been the best part of this journey. We might not be the best football team nor the best after lunch quiz team, but I am very lucky to have shared this experience with all of you. També al Juan Pablo, al Genís i a la Diana, por su ayuda con sus códigos para todo, la informàtica i els petits trucs que sempre comparteixen. Y a ti Marga, por siempre ayudar con todas las gestiones y estar pendiente de cómo estaba en todo momento. També a l'equip de l'EBL, especialment a la Mireia i en Rafa, qui junt amb el Jose, han proporcionat una gran part de les dades experimentals d'aquesta tesi.

To the former members, Marc, Diego, Alessio, Sandro, Kim, Juan, Francesco, I am grateful for having had the chance to share part of this journey with you.

Also, to the people with whom we crossed paths at the IRB. Thanks to all my fellow PhD riders, basketball players, IRB rock band members, administration and outreach team. Thank you for sharing this ride with me and making this workplace a better place.
To my sisters Laura, Maria, my brothers in law, Joan, Xavi and all my nephews. Perquè tinc moltíssima sort de rebre tanta felicitat i suport per part vostra i això sempre m'ha ajudat a tirar cap endavant.
I a tu Sandra. Perquè has sigut el meu pilar en aquests anys, ajudant-me de la millor manera possible amb tot. No només amb la portada d'aquesta tesi sinó amb tots els moments on has tirat de mi i m'has ajudat a gaudir del procés. Soc molt afortunada amb tu.
I per últim a vosaltres. Inma, Manu, gracias por creer en mí, animarme a perseguir las cosas y apoyarme en todo. Sois y seréis mis referentes en la vida, y llegar hasta aquí ha sido gracias a todas las oportunidades y los esfuerzos que habéis hecho.

Per tots vosaltres i totes les persones que m'han acompanyat sempre.

Abstract

The interplay between chromatin, transcription factors and genes defines a complex gene regulatory system whose study is essential to understand cell differentiation and how relevant cell functions are maintained or disrupted during biological processes. This is the main area of interest of what is known as 4D genomics, a field that has been evolving on the wave of high-throughput methods producing extensive amounts of data on gene location, chromatin structure and gene expression. 4D genomics data is each day more accessible, the challenge being to interpret it. In this sense Machine Learning (ML) and Artificial Intelligence (AI) methods are becoming crucial to transform noisy experimental data into biological information. The aim of this thesis is to untangle some of the mechanisms that constitute regulatory networks while integrating various state-of-the-art techniques to further understand some of the currently unanswered challenges.

During my PhD thesis I have been exploring the chromatin landscape and gene expression mechanisms at different levels of detail and this volume summarizes the main results obtained. In the Introduction, a brief overview of gene expression and transcriptional regulation mechanisms is provided. I also discuss the underpinnings of chromatin organization and stress conditions.

Chapters 1 through 5 are a compendium of articles where I describe the different research projects that I undertook during my PhD thesis. These projects have either been published in peer-review journals, are currently under review or in preparation.

More specifically in Chapter 1 we started by studying the first regulatory layer, the double-stranded helical structure of DNA and its binding to effector proteins. We developed a ML model that predicted with high accuracy the *in vitro* affinities and binding sites of various transcription factors based on physical properties of the DNA. Our method also successfully reproduced *in vivo* data when combined with a second layer of information, the chromatin organization of the nucleosomes.

In Chapter 2 we explored nucleosome positioning preferences in yeast genomic DNA by first developing a predictor of nucleosome free regions around the transcription start and terminating sites which are known to comprise critical binding sites. Our method allowed us

to predict the nucleosome architecture within gene bodies by using signal theory from two strongly positioned nucleosomes referred to as +1 and -last (the nucleosomes immediately downstream of the TSS or upstream of the TTS respectively). We additionally studied the link between nucleosome arrangements and gene expression mechanisms.

In Chapter 3, the effects of oxidative stress damage on nucleosome organization and overall chromatin structure are described. In order to clarify the effect of these lesions, we performed statistical analysis on a series of gene expression mechanisms through different experimental techniques such as MNase-Seq, Hi-C and Micro-C experiments.

Chapters 4 and 5 introduce the study of RNA as a distinct structure and discuss its properties and capability of playing a key role in some regulatory mechanisms such as triplex forming oligonucleotides. A general discussion that encompasses the significance and future perspectives of these 5 projects is presented in Chapter 6, together with the main conclusions of this work in Chapter 7.

Index

PhD Supervisor Report

Ms. Alba Sala Huerta did her doctoral thesis at IRB under our direction. The work of her doctoral thesis is reflected in a series of scientific publications, out of which two are published, a third one is accepted for publication, a fourth one is in revision, and one is in preparation. The publications are listed in the same order as the chapters in this thesis:

Publication 1

Barissi S.*, **Sala A**.*, Wieczór M., Battistini F., Orozco M., DNAffinity: a machine-learning approach to predict DNA binding affinities of transcription factors, *Nucleic Acids Research*, Volume 50, Issue 16, 9 September 2022, Pages 9105–9114, https://doi.org/10.1093/nar/gkac708

JIF of the JCR SCIE/SSCI6: 19.160: Quartile JCR SCIE/SSCI: Q1; Category JCR SCIE/SSCI: BIOCHEMISTRY & MOLECULAR BIOLOGY

(*) co-first authors
This is one of Ms. Alba Sala Huerta main projects. She was involved in the development of the AI algorithm and in the processing of the experimental data for transcription factor binding affinity. Ms. Alba Sala Huerta was the author in charge to compare and benchmark the developed algorithm with previously published ones. She is the first co-author of this work, contributing equally as the other first author, and the only one presenting this study for a PhD thesis.

Publication 2

Sala A.*, Labrador M.*, Buitrago D., De Jorge P., Battistini F., Brun Heath I. and Orozco M., An integrated Machine-Learning model to predict nucleosome architecture.
Accepted for publication - Nucleic Acids Research

(*) co-first authors
This is one of Ms. Alba Sala Huerta main work. She was the main developer of the neural network algorithm to identify and predict nucleosomes free regions along the genome. She also processed the *in silico* nucleosome positioning and performed statistical analyses. Ms. Alba Sala Huerta is the first co-author of this work, contributing equally as the other first author.

Publication 3

Arcon J.P. *, Lema R. *, Caballe A., Buitrago D., **Sala A**., Álvarez-Meythaler J.G., Villegas N., Blanc J., Reina O., Gut M., Dans P.D., Stephan-Otto Attolini C., Brun Heath I., Orozco M, Effect of oxidative stress on 3D genome structure. *(In preparation)*

(*) co-first authors
This is one of the main works in which Ms. Alba Sala Huerta worked on during her PhD. She analyzed experimental data from chromatin structure (MNase-seq, Hi-C and Micro-C) with the corresponding statistical analyses. She is the fifth contributor in order of signature of this work and the first one to present this work in a thesis.

Publication 4

Genna V.*****, Portella G.*****, **Sala A**.*****, Terrazas M.*****, Villegas N., Mateo L., Castellazzi C., Labrador M., Aviño A., Hospital A., Gandioso A., Aloy P., Brun-Heath I., Gonzalez C., Eritja R. and Orozco M. Systematic study of hybrid triplex topology and stability suggests a general triplex-mediated regulatory mechanism, bioRxiv 2024.05.28.596189, https://doi.org/10.1101/2024.05.28.596189.
In revision - Nucleic Acids Research

(*) co-first authors
This is another of Ms. Alba Sala Huerta main projects. In this work she contributed with the development and application of an algorithm to predict triplex formation along the genome. Ms. Alba Sala Huerta is a co-first author of this work, and the only one presenting this study in a PhD thesis.

Publication 5

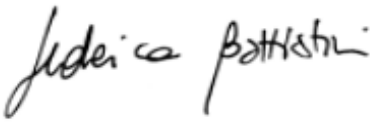
Battistini, F.; **Sala, A**.; Hospital, A.; Orozco, M., Sequence-Dependent properties of the RNA duplex, Journal of Chemical Information and Modeling 2023 63 (16), 5259-5271, DOI: 10.1021/acs.jcim.3c00741

JIF of the JCR SCIE/SSCI3: 5.6; Quartile JCR SCIE/SSCI: Q1; Category JCR SCIE/SSCI: Chemistry

Ms. Alba Sala Huerta is the second author of this work. She performed the statistical analyses of the sequence-dependent parameters describing RNA structure in the helical space. She is the only author using this article for a PhD thesis.

Prof. Modesto Orozco
modesto.orozco@irbbarcelona.org

Dr. Federica Battistini
federica.battistini@irbbarcelona.org



Introduction

The complexity of living organisms is the result of a combination of processes starting with the information encoded in the DNA and the consequent expression of genes. The regulation of gene expression is essential for normal development and cellular differentiation. A better understanding of the regulation and miss-regulation of genes can facilitate the comprehension of phenotypic effects, modes of evolution and even cancer and other diseases. In this Introduction, I start by exploring the structural characteristics of DNA and RNA that play a crucial role in many biological processes. I will then describe the key elements in gene expression mechanisms and the role that chromatin structure plays at a higher level. Finally, different elements that can affect gene regulation, such as DNA damage drivers, will be introduced followed by a description of the foundational elements of machine learning.

1. The Structure of DNA

The discovery of the structure of DNA (1) was a scientific breakthrough that had profound biological implications. The central dogma - DNA makes RNA and RNA makes proteins- was formulated shortly after (2). This was then followed by the discovery of DNA replication, the responses to DNA damage, DNA compaction into chromatin, or the regulation of gene expression before translating the encoded sequences, among other relevant biological processes.

At the core of biology lies DNA, a polymeric molecule composed of repeating units called nucleotides, which at physiological conditions arrange forming a complementary right-handed duplex containing the genetic information necessary to build life. Each nucleotide base consists of one nitrogenous base (adenine (A), cytosine (C), guanine (G) or thymine (T)), a deoxyribose sugar, and a phosphate group defining the backbone. These bases are planar heterocyclic molecules that can be classified into two groups: purines (A, G) and pyrimidines (C, T). In the canonical Watson-Crick duplex, purines, A or G, will interact with pyrimidines, T or C respectively, through specific hydrogen bonds. Two hydrogen bonds constitute the A-T pairing while three comprise the G-C pairing, making the latter more stable given the additional bond (see Figure 1). Phosphodiester bonds

hold together nucleotides within the same strand, resulting in an alternating sugar-phosphate backbone, while aromatic moieties stack one another generating stabilizing stacking interactions.

The canonical double-helical structure of the DNA at physiological conditions is named the B-form. This duplex forms a right-handed helix, with about ten to eleven base pairs forming a helical turn. Two types of grooves arise from the DNA helix: the major groove and the minor groove. These indentations differ from each other in terms of widths and depths given the asymmetric attachment of the base pairs to the sugar-phosphate backbone, with the major groove being wider and deeper, able to recognize protein elements (see Figure 2).

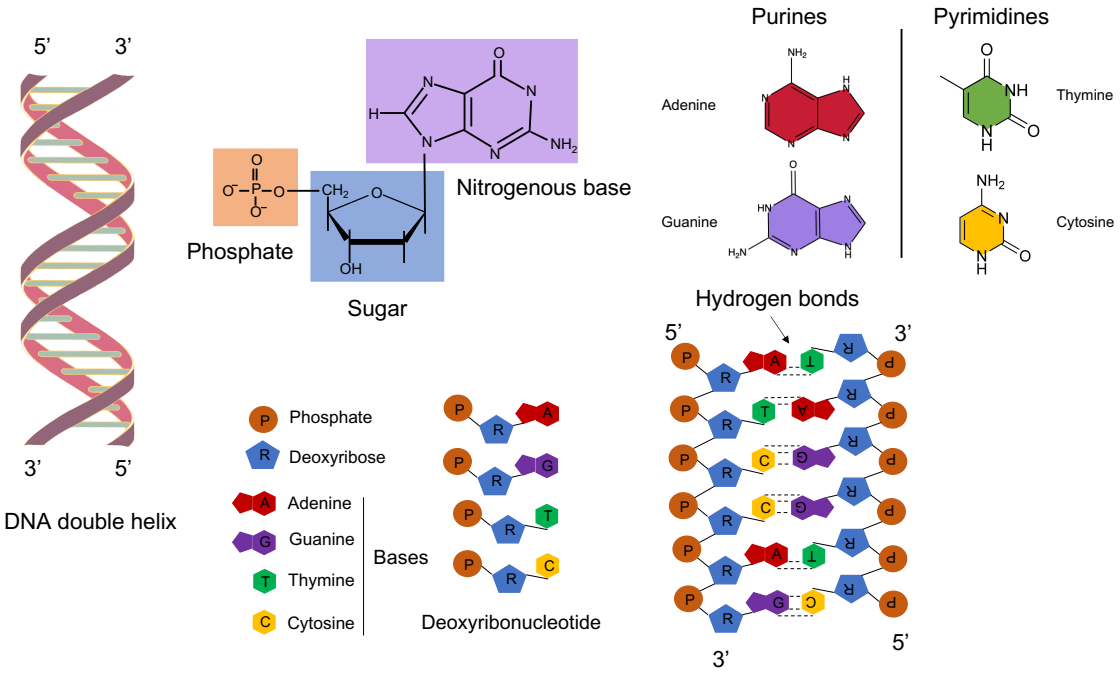


Figure 1. Scheme of the structure of DNA. Figure of the double helix of the DNA, the nucleotide bases, and their structure together with their base pairing modes.

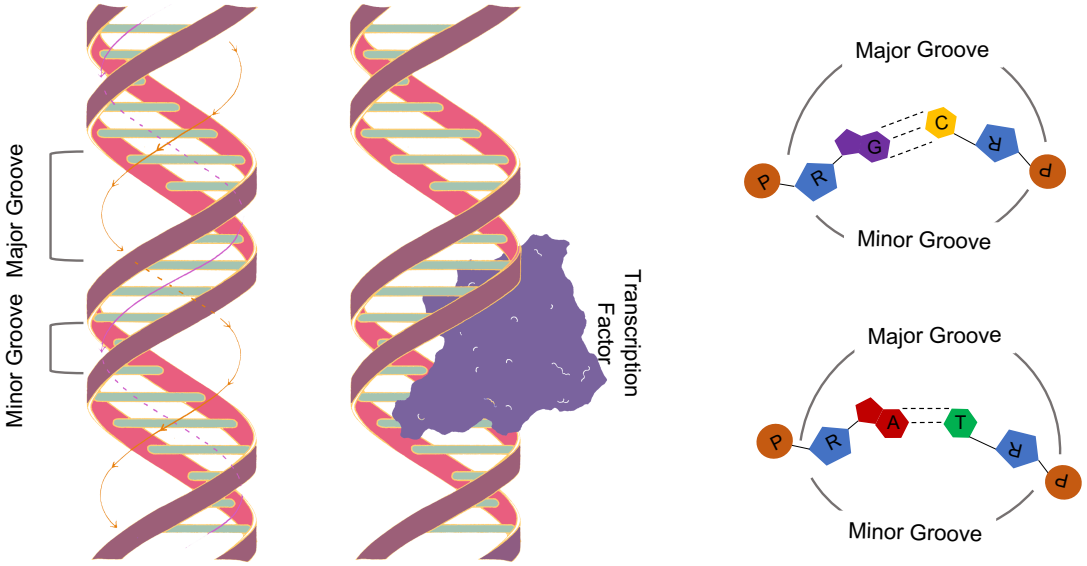


Figure 2. Scheme and definition of the major and minor grooves conforming the DNA. A schematic major groove binding by a transcription factor is also shown.

1.1. Helical Parameters: a rigid base model

Contrary to first thought, the duplex structure is not perfectly regular in biological conditions, but rather changes depending on the sequence (3, 4). A convenient way to measure such variability is at the base and base-pair level, measuring the movements using helical internal coordinates. These coordinated were defined in 1988 at an EMBO meeting in Cambridge, also known as the "Cambrige Accord", and standardized at the Tsukuba Workshop on Nucleic Acid Structure and Interactions (5). The definition of the parameters put together a single reference frame that would be used to calculate base morphology parameters. This definition assured the generation of consistent values across studies.

The parameters are defined either locally with respect to a local coordinate system attached to each individual base pair, or globally, with respect to a global curvilinear helical axis (see Figure 3) (6–8).

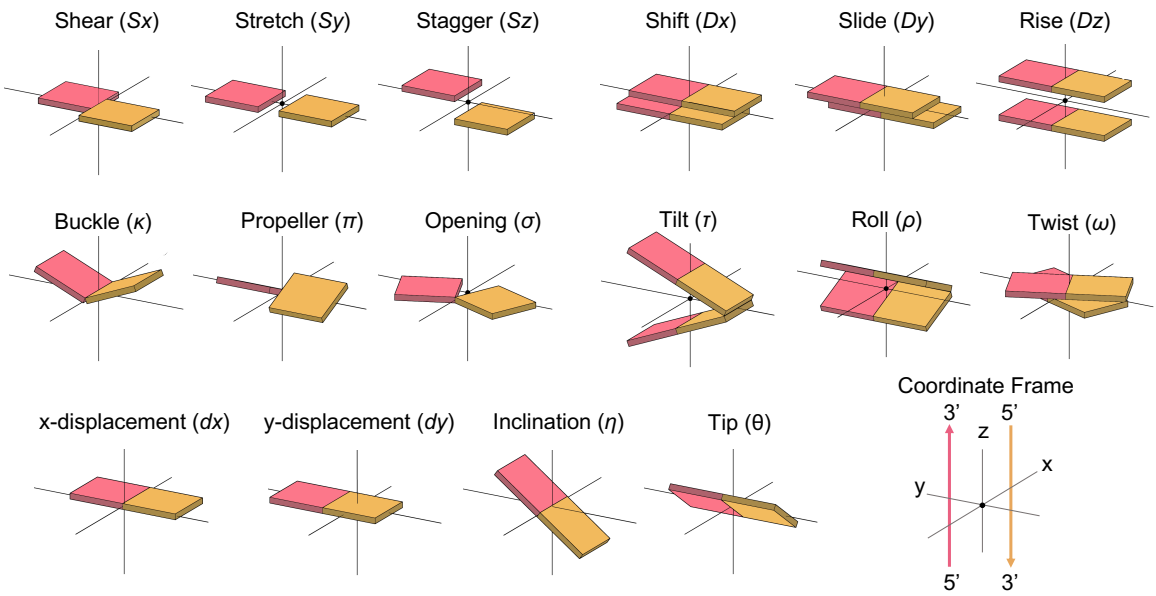


Figure 3. The helical parameters that define the spatial location of bases and the DNA conformations.

Through 6 intra-base pair parameters, 3 translational and 3 rotational, we define the relative geometry of two rigid bases conforming a base pair:

- Shear, stretch, and stagger (translational parameters), called the relative displacements of the bases along their x-, y- and helical axis (z-axis) respectively.
- Buckle, propeller twist, and opening (rotational parameters), called the relative torsions of the base planes around their x-, y- and helical axis (z-axis) respectively.

The orientation in space of a base-pair modeled as a rigid body is characterized by 10 coordinates, 6 of which (3 rotational and 3 translational) are defined relative to the previous base-pair in a dimer reference frame (inter-base pair parameters or base pair step parameters):

- Shift, slide, and rise (translational parameters) define the relative displacement of one base pair against another one in the direction of the x-, y- and z- helical axis respectively.
- Tilt, roll, and twist are the 3 rotational parameters. Tilt is the corresponding dihedral angle along the x-axis of the base pair. Similarly, roll is the dihedral angle for rotation about the y-axis. Its positive value opens a base pair step towards the minor groove and its negative towards the major groove.

Twist is the angle between successive base pairs about the helical axis (z-axis).

The remaining 4 coordinate parameters describe the geometry of a rigid base pair with respect to a local helical axis:

- Inclination is the angle between the y-axis of the rigid base pair and a plane perpendicular to the helical axis. Tip is the angle between the x-axis of the base pair and a plane perpendicular to the helical axis. Both take positive values for a right-handed rotation.
- X-displacement and Y-displacement define translations, along the x- or y-axis respectively, of the midpoint of the base pair mean plane with respect to the helical axis. A base pair with positive X-displacement is translated towards the major groove and a positive Y-displacement is towards the first nucleic acid strand of the duplex.

The helical parameters allow us to describe the basic structural features of DNA. However, a growing understanding of molecular cell biology led by advancements in experimental (4, 9) and theoretical methods (10–12), required the need to additionally take into account the dynamics, and more specifically, the flexibility of DNA (vide infra).

1.2. DNA dynamics and flexibility

Experimental evidence demonstrates that on top of the one-dimensional sequence information and the sequence-dependent DNA structure, the deformability (or flexibility) of DNA plays an important role and has a significant impact on gene regulatory mechanisms (13, 14). One way to evaluate the flexibility of DNA is to generate an ensemble of structures and use the inverse of the covariance matrix (either in the Cartesian or the helical spaces), to calculate force constants associated with the elastic deformation modes (15). This assumes that helical properties are normally distributed and that known structures provide a dense enough sampling of the accessible conformational space. The latter is achieved by using dense sampling that can be obtained from molecular dynamics simulations (16), which can be used to explore a large variety of sequences.

The Ascona B-DNA (ABC) consortium have undertaken initiatives to provide information on the conformational properties of the 136 unique tetranucleotide sequences (17, 18). Based on systematic analyses of databases and molecular dynamic simulations it is evident

that the sequence plays a key role in the conformation of DNA by a combination of energetic and structural factors. The information obtained from these analyses for the different unique tetrameric sequences transforms the genetic information from its linear sequence into a global structural arrangement of the double helix. This allows a deeper understanding of the functional implications associated with sequence-dependent conformational variability. We will show how the information derived from these analyses can help to understand gene regulation mechanisms.

2. Structure of RNA

Similarly to DNA, RNA is a polymer composed of sugar and nitrogenous bases with phosphate backbones. Unlike DNA, RNA has been found as a single strand that can adopt different topological conformations such as internal-loops, duplexes, triplexes, hairpins, or bulges. RNA not only carries genetic information but also performs various regulatory and enzymatic functions, highlighting the diversity of its conformational possibilities.

In RNA a hydroxyl group is found attached to the ribose sugar at the 2′ position, which is not present in the DNA deoxyribose sugars. The additional hydroxyl group allows the RNA to interact in different ways with itself and a variety of ligands. This allows it to be readily hydrolyzed and cleaved, making it a more versatile but more unstable molecule (19, 20); it also introduces a bias towards the North-type puckering, which impacts the global helical structure. In addition, thymine bases are replaced by uracils which follow the same base-pairing characteristics but do not have the methyl group at the 5′ position (see Figure 4).

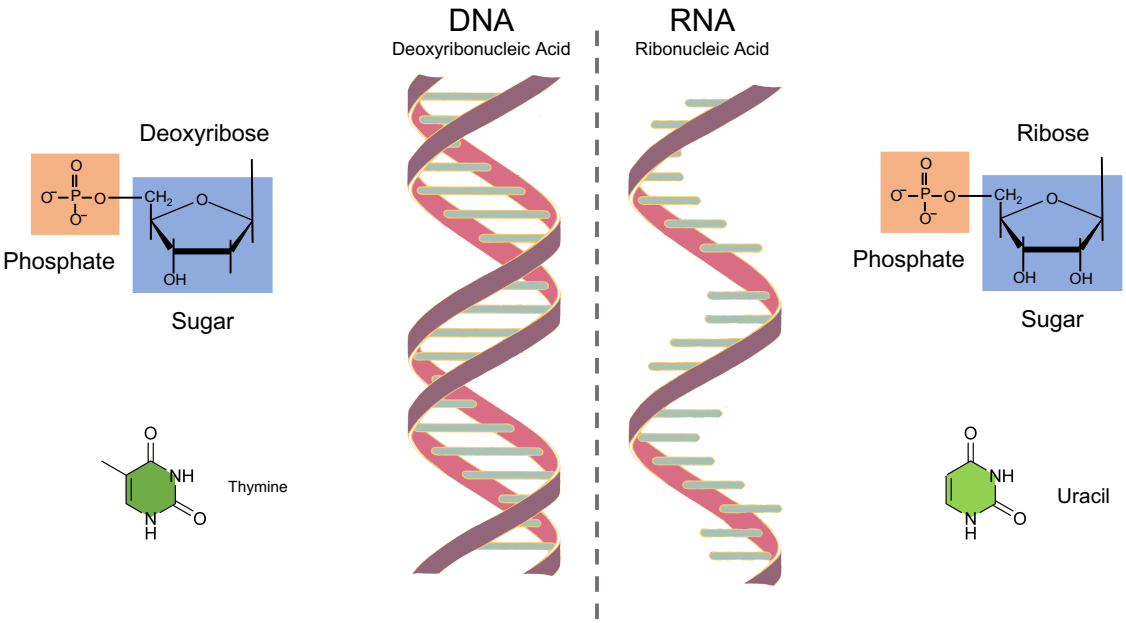


Figure 4. Main differences between double-stranded DNA and single-stranded RNA base composition

The physiological RNA structure is composed by a single-stranded polymer, where the ribonucleotides tend to interact with each other through base stacking and if topologically possible, by hydrogen bonding forming a hairpin-like structure. In RNA even though we observe the canonical DNA Watson-Crick base pairings, we commonly observe an additional pairing mode that has a similar stability than A•U, the G•U wobble pair (21, 22).

While the sequence-dependent properties of DNA have been widely studied and accurately defined, the molecular simulations and experimental data describing the RNA structure are not as broadly available. Similarly to the previously described motivation of understanding the DNA conformations, it is possible to study the structure and flexibility of RNA in order to fully characterize the bases of the polymer and to better understand the mechanisms of biological processes. The different parameters can also be defined locally with respect to a local coordinate system or globally, with respect to a global curvilinear helical axis. Further details on a massive simulation effort undertaken during this thesis is described in Chapter 5.

RNA does not only interact with itself but rather with other molecules such as proteins or DNA. One example of this is the formation of triple helices, the most interesting one being the hybrid triplex, when a single-stranded RNA interacts with a duplex DNA.

2.1. The triplex structure

The ability of nucleic acids to form a triple helix has been known since the 1950s-60s (23–25); however, in recent years, a substantial body of research has gained interest in this structure giving its implications in biological processes and its potential application to anti-gene therapies and biotechnological applications (26). Triplexes are formed when a poly-purine segment of a duplex molecule is recognized by a third oligonucleotide strand, which binds it by means of specific hydrogen bond interactions (27, 28). The binding of the third strand, also known as triplex forming oligonucleotide (TFO), happens at the major groove of the double helix, also known as the TTS (triplex target site), by means of Hoogsteen (or reverse Hoogsteen) bonds (see Figure 5).

The TFO can be a molecule of DNA, RNA or a xenonucleotide (29). Based on the orientation of the TFO with respect to the polypurine TTS we define two types of triplexes: parallel or antiparallel triplexes. Parallel triplexes occur when a TFO (typically pyrimidine-rich) binds a central Watson-Crick purine by means of Hoogsteen hydrogen bonds while anti-parallel triplexes are characterized by reverse-Hoogsteen hydrogen bonds and a purine-rich TFO (see Figure 6).

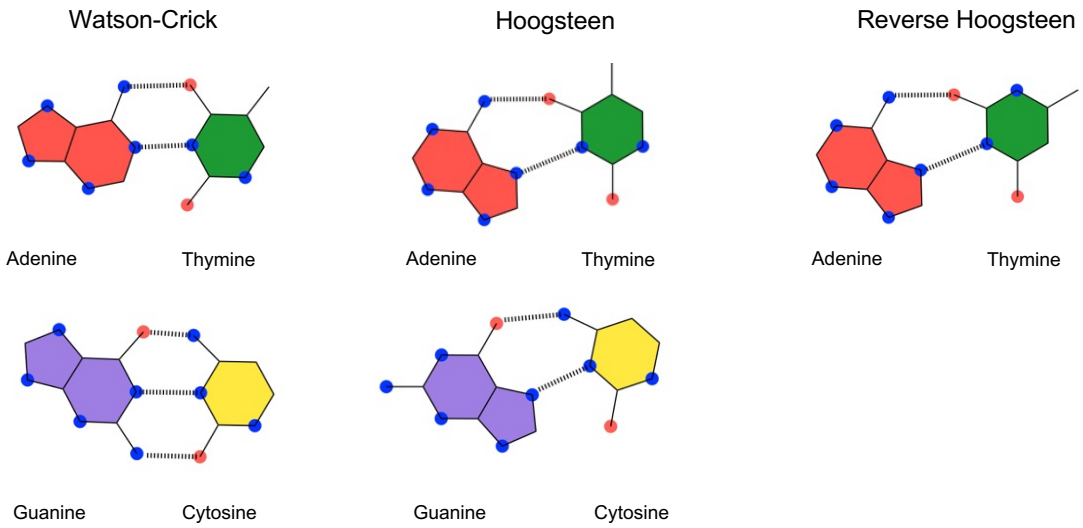


Figure 5. Schematic representation of the canonical Watson-Crick vs. the Hoogsteen and the reverse Hoogsteen A-T and G-C base pairings.

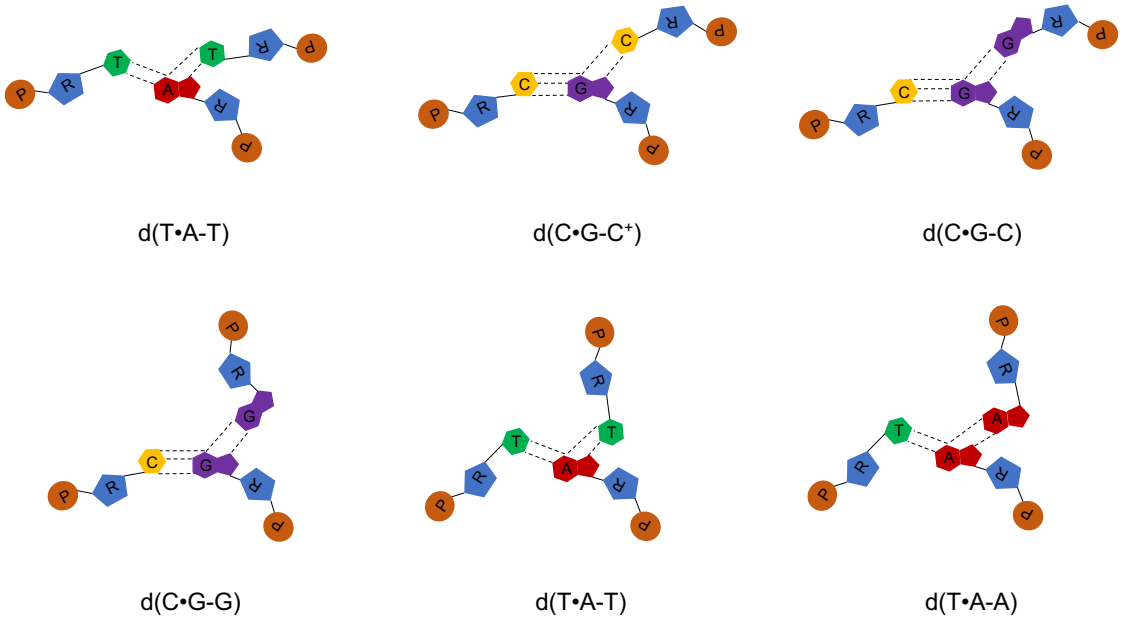


Figure 6. Schematic representation of the six nucleotide triads which form the basis of canonical triplex formation.

Very interestingly, TFOs are over-represented in the human genome (30, 31), especially in regulatory regions such as promoters, which suggest a potential regulatory role of these structures.

3. Gene Regulation Mechanisms

Gene expression defines the process by which the encoded information in DNA is turned into a function. The first step in this process starts with a copy of the DNA and the transcription of RNA into messenger RNA (mRNA) by an enzyme called RNA polymerase. Both coding and non-coding regions of DNA are transcribed. The introns are removed through an initial processing step and the exons get spliced together. In the second step of this process an intercellular structure made of both RNA and proteins, known as the ribosome, translates the mRNA into coded amino acids which end up forming a polypeptide chain that will carry out a specific function (see Figure 7). Gene regulation dictates when, where and how many RNA molecules and proteins are made, which constantly changes under different conditions and cell types. Each step in this process from DNA to RNA to protein provides a potential control point. The regulation of these processes is fundamental to the correct operation of a cell. Defining and understanding the structure and dynamics of gene regulatory networks is of great interest but represents one of the most complex biological problems.

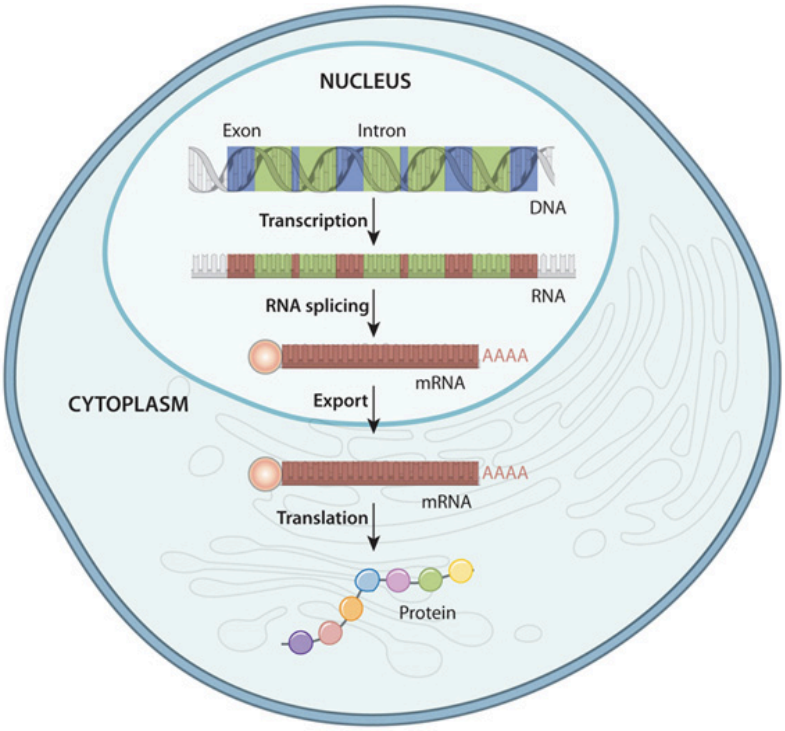


Figure 7. An overview of the flow of information from DNA to protein in a eukaryote (picture adapted from https://www.nature.com/scitable/content/an-overview-of-the-flow-of-information-14711098/).

Advancements in current technologies have allowed us to measure the genome-wide expression of mRNA for many cells, and investigate which genes are activated or repressed in different situations. The most popular technique used to measure the expression of genes is RNA sequencing (RNA-seq). This technique uses next-generation sequencing to quantify the amounts of transcript data, also known as the transcriptome. In a first step, RNA is reversely transcribed into complementary DNA (cDNA) and adaptors are attached to both ends of the fragments. Secondly, the library is high-throughput sequenced following manufacturer protocols. Lastly, RNA-seq reads are aligned to a reference genome for downstream usage.

What exactly determines the expression of a gene is not a straightforward route, but rather can be modulated by many factors. Nonetheless, something that is clear today is the relevant role that protein-DNA interactions, especially those involving transcription factors, play in expression mechanisms and transcriptional responses.

3.1. Transcription Factors

Transcription Factors (TFs) are effector proteins that bind to DNA in order to activate or repress the transcription of genes. They act as buttons to turn on/off the translation of RNA into proteins while activating/inhibiting the function of genes. Therefore, proteins can be seen as the main regulators of gene expression as they can directly or indirectly activate or repress gene activity.

Predicting the binding patterns of TFs during regular cellular processes presents a challenge given the intricate interplay of many key factors, such as the presence of nucleosomes, which generally hinder TF binding (32, 33), the formation of DNA-protein clusters promoting cooperative or anti-cooperative interactions, chromatin compaction, or even phase separation (34–36). Nonetheless, an essential prerequisite for effective *in vivo* binding involves high affinities with the naked DNA by transcription factors (37). We can distinguish the binding of a transcription factor to its target DNA based on direct and indirect interactions. Depending on the type of contacts used for DNA recognition we differentiate TFs where the residues contact the DNA bases and get stabilized by the contacts at the sugar-phosphate backbone, those establishing hydrogen bonds with either the minor or the major groove, and those that induce a fit and cause a geometry disruption in the structure to create contacts.

While TFs bind DNA in regulatory regions, TF co-factors (co-repressors and co-activators) do not necessarily have a specific binding motif but are required to catalyze enzymatic reactions. In higher eukaryotes, transcription by the pol II molecular machine is initiated through the assembly of the pre-initiation complex (PIC) near the transcription start site (TSS), where transcription factors get recruited (see Figure 8). Gene expression mechanisms are controlled in time and space through these complex processes. The interaction of regulatory elements and DNA determines the transcriptional levels and the activities of individual enhancers and promoters.

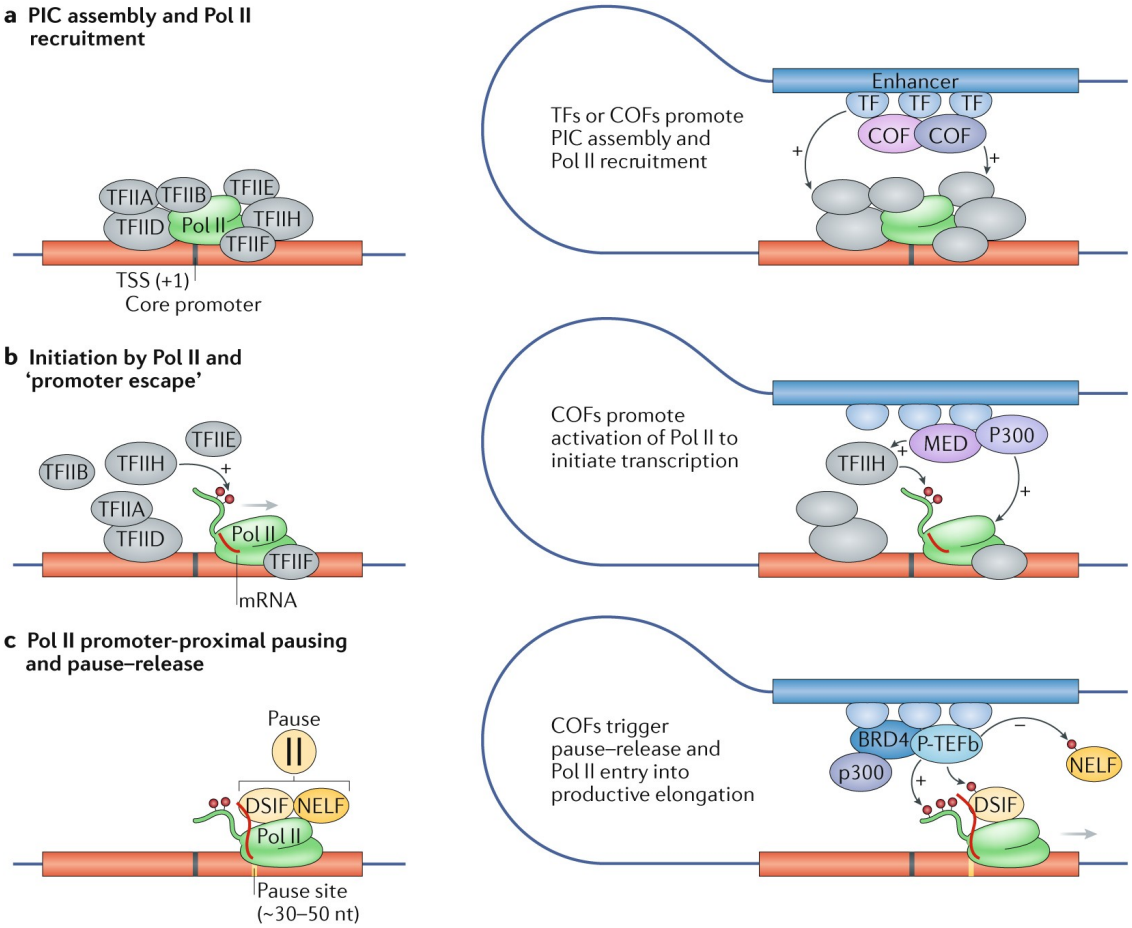


Figure 8. Steps of transcription from core promoters. A) The first step of transcription initiation is the assembly of the pre-initiation complex (PIC), the RNA polymerase II (Pol II) recruitment and six general transcription factors (GTFs). B) After PIC assembly, the DNA duplex at core promoters melts (not shown) and allows Pol II to initiate transcription at the transcription start site (TSS). To continue transcribing, Pol II must dissociate (escape) from the TSS-binding GTFs. C) After escaping from the TSS, Pol II synthetizes a short stretch of nascent RNA (~30–50 nucleotides) and then pauses downstream of the TSS. Image adapted from (38).

3.2. DNA-Protein Binding Mechanisms

The recognition of a DNA by a protein can either be highly specific, uniquely recognizing a defined sequence (consensus sequence) with high affinity, or non-specific, without a significant preference for binding to a specific nucleotide sequence. The recognition of a DNA sequence by a protein is a very complex process

that often involves many elements. Up to this day, it has not been possible to define a simple code for DNA recognition mechanisms (37). However, some patterns have been observed and some characteristics have been defined based on structure-determination experiments and molecular simulations.

Non-specific protein-DNA binding has been extensively observed across genomes of different organisms (39). Two fundamental mechanisms characterize this type of binding (40): (i) the overall electrostatic attraction between protein and DNA, and (ii) the geometry of DNA. In the first mechanism cationic residues of the protein bind to either the phosphates or the negative electrostatic potential present within the DNA grooves. In the second mechanism, the helical arrangement of DNA serves as a template for the non-specific binding of recognition proteins which exhibit a complementary helical configuration.

Specific binding occurs through two strategies broadly known as "direct" and "indirect" readouts (37). In a direct readout the DNA sequence is identified through specific interactions between amino acid sidechains and base groups exposed at the protein-DNA interface. It typically occurs through the formation of hydrogen-bonds between the bases and proteins through either the major or the minor groove (41). The unique arrangement of hydrogen bond donor and acceptor sites for each dinucleotide within the grooves provides the specificity utilized by proteins to discriminate specific DNA sequences. In the major groove more specifically, we observe the presence of two hydrogen bond acceptors and one donor group in all four dinucleotide pairs (see Figure 9). In addition, a methyl group located at the C5 position of thymine can engage in van der Waals interactions. Discrimination amongst DNA sequences is achieved on the basis of differences in hydrogen bonding patterns between base pairs (i.e. the C•G Watson-Crick base pair has a distinct pattern than the G•C base pair). The A•T and T•A base pairs have symmetrical donor/acceptor patterns in their major groove, but the presence of the C5 methyl group of thymine introduces variations in the groove that can facilitate effective distinction between these two base-pair sequences. Conversely, the hydrogen bonding patterns in the minor grooves of each pair of sequences are symmetric (i.e. the N2 atom of G provides a hydrogen bond donor at the center of the minor groove for both C•G and G•C dinucleotide pairs), making the sequence discrimination based on the minor groove direct reading a more complex process. Although the major groove is usually preferred for binding in a direct readout, some regulatory and structural proteins exhibit a target

binding site in the minor groove, including proteins that are able to deform DNA and expand the groove (42).

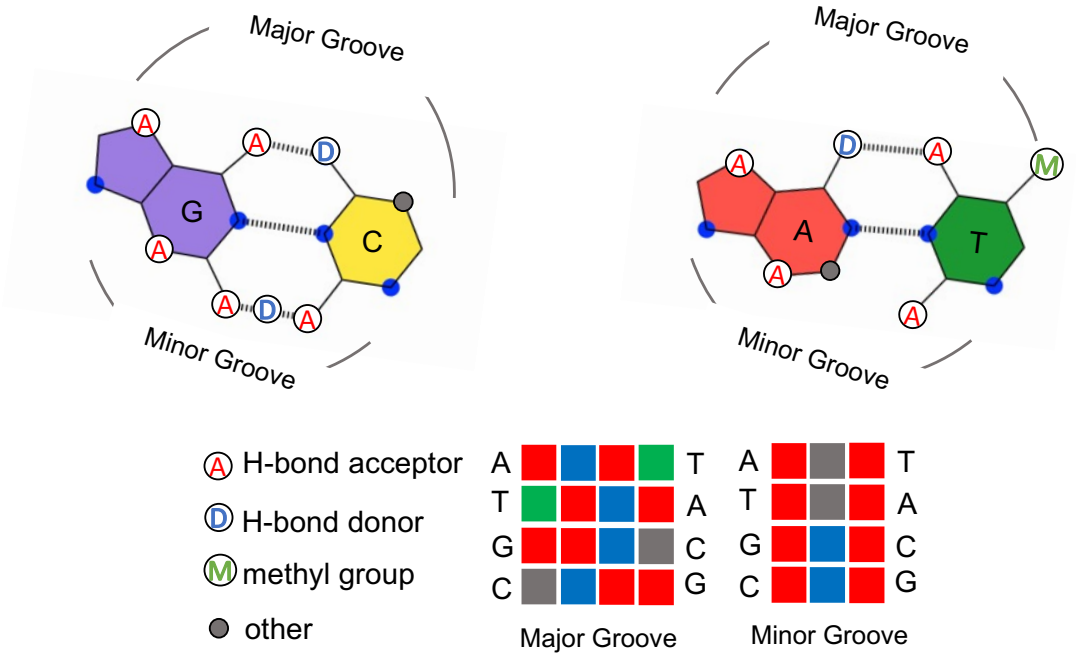


Figure 9. Base readout in the major and minor grooves

Contrary to the direct readout, in the indirect readout, proteins read out DNA sequences by detecting sequence-dependent variations in structure and flexibility. This readout is defined by the interaction of structural elements of DNA that do not involve base pair hydrogen bonds. The indirect readout it is characterized by the ability of the DNA to adopt a "bioactive conformation" which is required to interact with the target protein. One of the key elements is the size of DNA grooves, specially the major one. Groove-width fluctuations are themselves sequence-dependent and describe the DNA accessibility. Many experimental techniques have shown narrower A/T-rich minor grooves through a large number of oligonucleotide crystal structures. Other studies provided evidence that A/T regions have a higher flexibility than G/C ones, and that particular features such as preferential ion binding often result in narrower minor grooves for A-tracks (43–46).

Considering all the variables involved and the different kinds of proteins, the overall rules governing DNA-protein recognition are still unclear. Despite the existence of cases where binding can be explained by the direct or indirect readout mechanism, in the majority of cases we find a combination of both readouts during specific protein binding events (47).

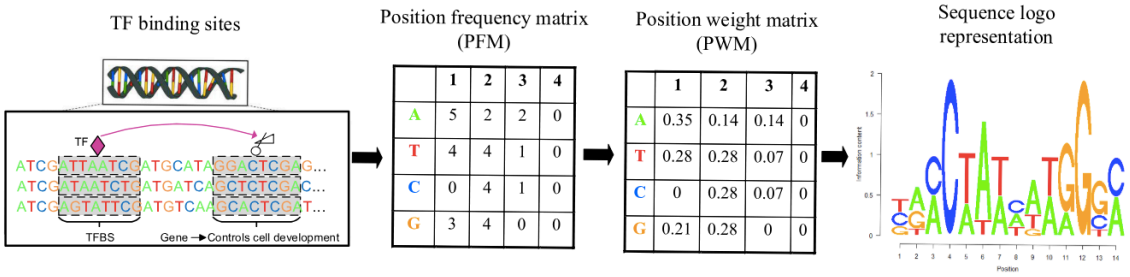
3.3. Experimental study of DNA-protein binding

Large advancements in high-throughput sequencing techniques have exponentially increased the available experimental data from known effector protein binding sites. Some of these techniques include high-throughput SELEX experiments (HT-SELEX) (48–50) and protein binding microarrays (PBM) using either synthetic sequences (uPBM) (51) or sequences coming from the genomic context (gcPBM) (52). These *in vitro* methods will output a set of sequences of varying lengths and provide the necessary information in order to classify binding affinities for a given protein. ChIP-seq approaches (52–54) allow the attainment of the *in vivo* preferences through the immunoprecipitation of chromatin by specific protein-antibodies followed by the sequencing of the retained fragments. This technique provides direct information on the active binding sites under physiological conditions, but it also comes with a downside as it presents a poorer resolution, having ChIP-seq scaling down to 200bp ± 300 and ChIP-exo down to 50bp (53–55).

3.4. Theoretical methods to study DNA-protein binding

TFs typically recognize and bind short sequences (6-20 bp), which can then be summarized as binding motifs (54, 56). Motifs are usually visualized using sequence logos where at each position nucleotide letters are stacked with their heights being proportional to their frequencies (see Figure 10). Conventional theoretical methods to predict TF binding sites often rely on the positional weight matrices (PWM) based on these motifs. Initial methods assumed that the nucleotide preferences at each position were independent from each other, but the latest PWM models have increased their capability to capture interdependence amongst the positions (57, 58). However, these models present some limitations (59, 60) which led to the development of alternative approaches aiming to reproduce *in vivo* and *in vitro* experimental binding sites. Models designed for *in vitro* binding site predictions utilize nucleotide sequences and diverse DNA shape descriptors as input parameters (60–62). Conversely, methods for *in vivo* binding site predictions integrate naked DNA descriptors with those associated with chromatin structure and dynamics (RNAseq, DNase, etc.). The increasing amount of experimental data has also motivated the use of a variety of ML and AI methods (63–73) to predict protein binding sites. In this thesis we will study the sequence-preferences and the rules governing TF-protein recognition

and we will investigate the prevalence of base versus shape readout mechanisms. This will allow us not only to predict the binding motifs but also to better understand sequence specificity analyzing the conformational selection versus an induced fit binding that could explain conformational changes in DNA. In order to do so we will be using a combination of experimental data coming from various techniques and a machine learning algorithm for the prediction.



	1	2	3	4
A	5	2	2	0
T	4	4	1	0
C	0	4	1	0
G	3	4	0	0

	1	2	3	4
A	0.35	0.14	0.14	0
T	0.28	0.28	0.07	0
C	0	0.28	0.07	0
G	0.21	0.28	0	0

Figure 10. TF binding sites are transformed into a position frequency matrix (PFM) which gives the PWM used to generate the motif plot from a binding site. Image adapted from (74)

4. Chromatin Structure and its role in transcription

Transcription factor binding is often highly correlated with DNA accessibility (34–36). Much like proteins, DNA exhibits a quaternary structure in its compaction which results in a higher-level nucleic acid organization that defines chromatin. This motivates the need to further study the compaction of DNA at various levels of details.

At the simplest level, chromatin is defined by a double-stranded helical structure of DNA. The compaction of DNA takes place through the recruitment of ions and proteins which guide the interactions between DNA and histones, leading to the assembly of nucleosomes. Even though the human DNA fiber is about 2m long, these structures fold-up in a tight manner to produce chromatin fibers which then form loops in an organized manner to fit inside a small space defined by the cell nucleus with a diameter of approximately 10μm (75) (see Figure 11).

Experimental evidence has shown that chromatin is dynamic and undergoes changes in its organization through different cellular processes such as differentiation (76) cell cycle progression (77) or DNA damage response (78). Chromatin is known not to be randomly distributed in the nucleus but rather differently concentrated in different nuclei regions (76, 79, 80). Consequently, DNA organization

does not only provide space-restricted compaction but also protects the genome from damage and generates entry points for biologically relevant processes such as transcription, DNA replication or DNA repair (81). Therefore, the relevance of understanding the landscape of DNA compaction is clear and in the following sections the different levels of compaction will be explained in detail.

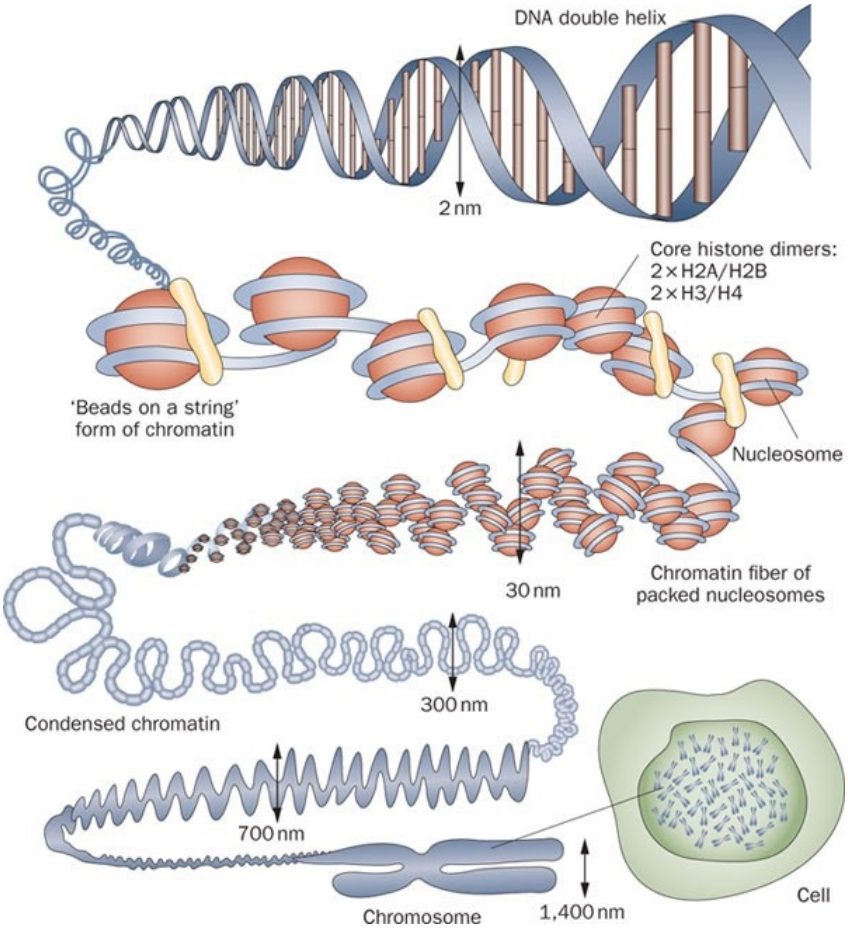


Figure 11. DNA organization and chromatin compaction at different levels of details, from the simplest DNA double helix up to the chromosome level. Image adapted from (82).

4.1. The nucleosome

DNA organization is aided by proteins that guide its folding in eukaryotic cells. Its helix coils up around histone proteins to form the primary units of genome organization, the nucleosomes. A canonical nucleosome consists of an octamer of histones, two copies of each histone H2A, H2B, H3 and H4, around which the DNA wraps 1.65 times (see Figure 12). The central bases of the DNA stretch align with a pseudo 2-fold symmetry axis, known as the dyad axis. Nucleosome

formation requires a significant bending energy due to the high curvature of the DNA within the nucleosome. This structure is stabilized by the interaction between the positively charged histones and the negatively charged DNA backbone that form interactions every 10 base pairs (83). In addition, early X-ray crystal structures of the nucleosome (84, 85) revealed that histone proteins are defined by a globular domain that constitutes the nucleosome core and N-terminal histone tails. These histone tails, which are known to be unstructured and flexible, can undergo post-translational modifications (methylation, acetylation, ubiquitination) which can alter chromatin accessibility.

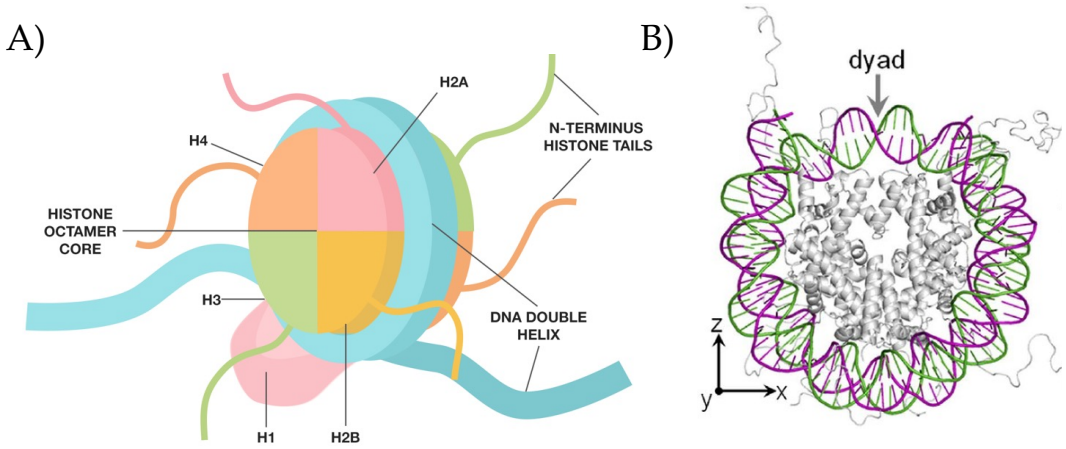


Figure 12. A) Schematic representation of the nucleosome structure with the dimer of tetramers, the respective tails, the linker histone, and the DNA stretched (picture adapted from https://eloisenewman.myportfolio.com/how-dna-is-organised-nucleosome-structure) . B) X-ray structure of the nucleosome core particle adapted from (86). The structure shows the two strands of the double-helix in purple and green, with the protein core in grey and the curvature of DNA around the histone core, with the dyad at the top

The DNA wraps around the histones in parallel, except at the entry and exit points of higher eukaryotes where linker histones (H1 or H5) bind. Connected by stretches of linker DNA, several nucleosomes are aligned into arrays of 10 nm in diameter. During this thesis I have mainly focused on the chromatin organization of yeast, where no additional histones are found. Nonetheless, in higher order organisms we find that nucleosomes are complemented by an additional H1 histone to form the chromatosome.

4.1.1. Other complex organisms

High resolution genomic maps of the budding yeast have emerged in recent years allowing comprehensive studies which helped clarify some of the mechanisms underlying the link between chromatin structure and gene regulatory mechanisms. However, such context has not yet been established in higher order eukaryotes, with many questions left to be answered and other factors coming into play. For example, the linker H1 histone binding to the DNA in order to form the chromatosome in more complex organisms. Previous research has proved that the H1 histone plays an important role in the organization of chromatin and consequently in the regulation of genes (87). The linker histone binding is positively correlated with nucleosome stability and the H1 position deviation relative to the dyads seems to be connected with a shift in nucleosome positions. Furthermore, H1 occupancy presents an inverse correlation against the nucleosome distance to the transcription start sites, with a lower correlation in genes with lower transcription levels. The precise role of the linker histone remains elusive, but something that is well conserved across species is the relevant role that nucleosome positioning and chromatin accessibility play in gene expression, regulation and other fundamental processes.

4.2. Chromatin at the 2D level

This thesis will shed some light on the conservation of underlying rules encompassing chromatin structure across eukaryotes. In order to do so it is essential to study the mechanisms governing the 2D and 3D architectures of chromatin to understand the role that chromatin plays in biological systems.

4.2.1. Nucleosome positioning across the genome

Nucleosome positioning is determined by an interplay of multiple factors including the DNA sequence, transcription, chromatin remodelers, histone modifications or non-histone barriers amongst others. Physical-chemical properties are known to regulate the nucleosome preferences in genomic DNA (33). These define a hidden physical code which indirectly controls gene expression. However, the physical fingerprint of a DNA sequence can undergo significant modifications by its sequence or by epigenetic changes. Consequently, the physical code not only directly determines high affinity positions

for binding transcription factors and transcription machinery, but also does so in an indirect way by influencing nucleosome positioning.

The positioning of nucleosomes within genomes *in vivo* has been determined through a variety of experimental methodologies, including FAIR, ATAC-seq, and MNase-seq (88–90), with the latter method giving the most comprehensive insights into nucleosome architecture (see Figure 13). This technique is based on an initial cross-linking of histones and nucleosome DNA using formaldehyde, followed by enzymatic treatment with Micrococcal nuclease (MNase) to cleave linker fragments. After the removal of the enzyme and the reversal of the cross-linking, the remaining undigested segments are then sequenced in order to determine nucleosome positions (90).

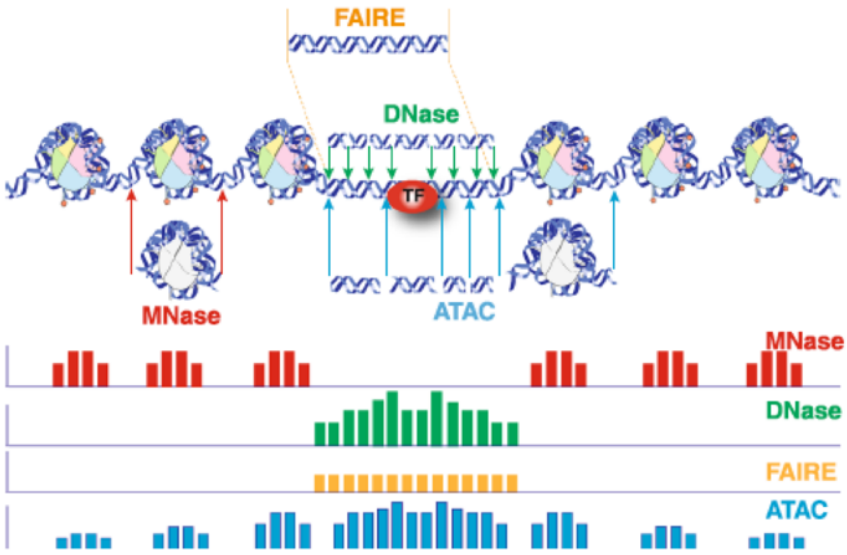


Figure 13. Various experimental techniques which allow us to quantify the positioning of nucleosomes across a genome. Image adapted from (91).

These experimental techniques often involve a population of cells which define rather noisy profiles (32), described by two attributes: occupancy and positioning (illustrated in Figure 14). The former corresponds to the proportion of cells in an experiment that contain a specific nucleosome, while the latter defines the variability of genomic positions it occupies among all cells. A nucleosome is classified as well-positioned (W) when it is prevalent across a substantial proportion of cells, and its fragments across different cells exhibit minimal variance regarding their genomic positioning. Conversely, a nucleosome with low coverage and/or substantial variability in its positioning is denoted as fuzzy (F).

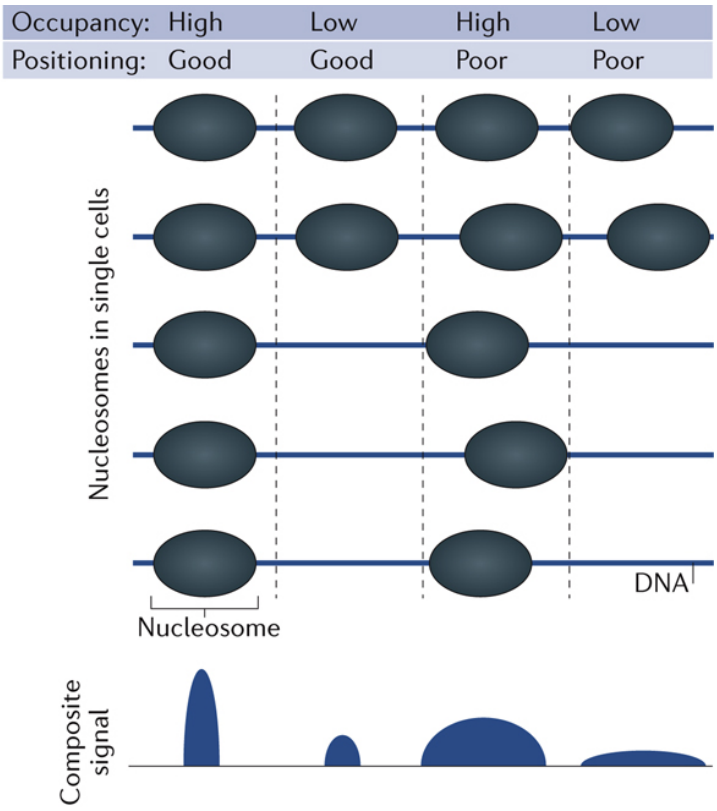


Figure 14. Schematic representation of nucleosome occupancy and positioning. A nucleosome in a pool of cells can be characterized by the relative number of cells that contain it (occupancy) and the variability between cells in the sequence position (positioning). Image adapted from (92).

4.2.2. Determining nucleosome positions

After performing an experimental sequencing technique, the reads are often stored in a FASTA or FASTQ file format. These reads are then mapped to a reference genome with an aligner such as Bowtie (93) to obtain a SAM or BAM file with the corresponding genomic positions of the aligned fragments. Mapped fragments that pass the quality control filters are then processed with a nucleosome positioning software. Our research group previously developed nucleR (94, 95) for this purpose, and through this thesis all nucleosome position calls have been performed using this R package. The following steps summarize the pipeline that nucleR follows when processing the mapped fragments:

1. In order to only keep mono-nucleosomes, fragments wider than 170 are discarded (see Figure 15A-B).

2. MNase digestion can present variability among cells. That is why as a second step fragments are trimmed to 50bp maintaining the original center to remove noise from regions in the nucleosome coverage profile (see Figure 15C).
3. The nucleosome coverage per base pair is computed genome-wide and transformed to reads per million mapped.
4. A Fast Fourier Transformation is used to filter noise but to still keep 1% or 2% of the principal components in human and yeast experiments, respectively (see Figure 15D).
5. Lastly, nucleosome peak calling is performed often using the default parameters: peak width 147 bp, peak detection threshold 35%, maximum overlap 80 bp and dyad length 50 bp. Nucleosome calls are considered well-positioned (W) when nucleR peak width score and height score are higher than 0.6 and 0.4, respectively, and as fuzzy (F) otherwise (see Figure 15E).

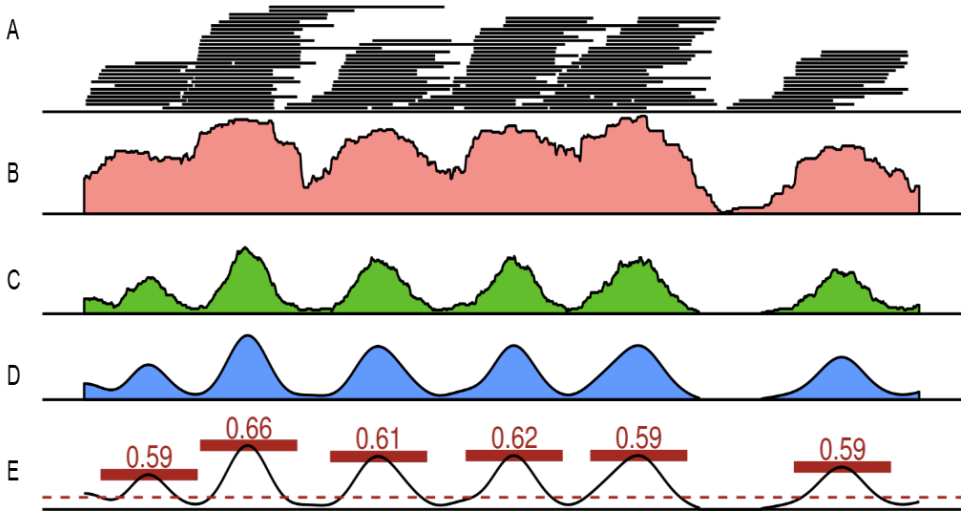


Figure 15. Nucleosome Positioning quantification steps from MNase-seq data using nucleR (94, 95).

4.3. Chromatin structure at higher level

Nucleosomes are connected by linkers creating a bead-on-a-string array that folds and assembles into the 3D space. Early electron microscopy experiments and *in vitro* reconstituted nucleosomes claimed a regular folding into 30-nm fibers, but different folding motifs were identified (96–100). However, *in vivo* studies questioned this claim and found evidence that the folding occurred in a more irregular

manner (101, 102). Recent super-resolution microscopy has revealed that indeed the chromatin fiber is not a regular structure but rather presents irregular regions distributed in an uneven way, formed by groups of varying sizes that can be linked to different cell types (103, 104). Furthermore, other techniques have been developed to analyze the 3D structure of the nucleosome fibers within the nucleus. Chromatin conformation capture (3C) techniques have shed some light on the arrangement of the nucleosomes by cross-linking interacting chromatin regions, digesting them with restriction enzymes, and then ligating the cross-linked fragments to identify physical interactions between distant genomic regions. Hi-C and Micro-C (75, 105, 106) are some of the 3C technologies that are able to quantify the frequency of genome-wide interaction between pairs of loci. Hi-C allows the study of the chromatin structure at kilo base scale whereas Micro-C can study the contacts at the nucleosome level.

The study of chromatin at a higher level of organization has shown a hierarchical organization in the nuclear space. Hi-C experiments exemplified the separation of chromosomes into territories as we observe larger contact frequencies between regions in the same chromosome, also known as cis-contacts, in comparison to those found amongst regions from different chromosomes, trans-contacts (107). As previously observed by FISH experiments, contact frequencies revealed the segregation of A/B compartments which correlate to an actively transcribed chromatin (euchromatin) and an inactive counterpart (heterochromatin) (108) (see Figure 16).

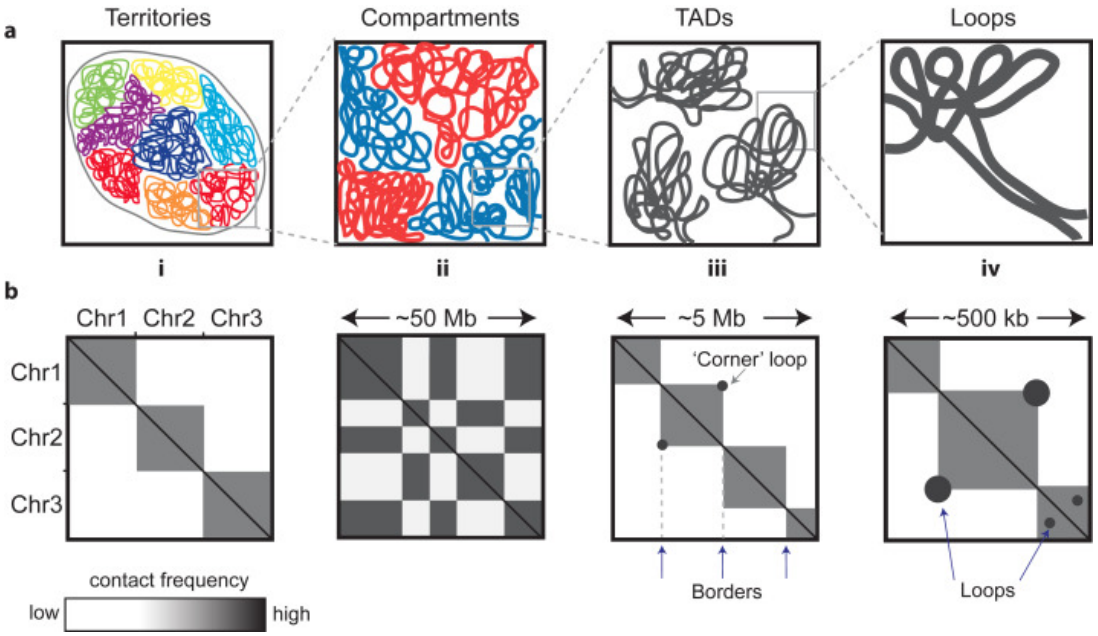


Figure 16. A higher-level view of chromatin organization. A) Schema of the features of chromatin folding at different scales, and B) the appearance of the underlying Hi-C contact map that allows deduction of the topological feature Image adapted from (109).

4.3.1. TAD formation

These studies have observed that chromatin is organized into topologically associated domains (TADs), regions which highly interact with themselves but are insulated from neighboring regions (107). TADs have been linked to loops formed in order to bring closer in proximity regulatory elements that are separated by a large genomic distance, such as enhancers and their target sites (110–113). In smaller genomes such as S. cerevisiae, Hi-C initially failed to detect the presence of TADs (114) but Micro-C allowed the detection of chromosomal interaction domains (CIDs) (105, 106). TAD borders in mammalian cells are known to strongly colocalize with CTCF (a zinc finger protein, CCCTC-binding factor) target sites (115) but not in organisms such as D. melanogaster, where the role of CTCF in TAD insulation is lower, or S. cerevisiae, where we do not find this protein nor a homolog. In these organisms, we usually find TAD or CID borders associated with promoters of actively transcribed genes, typically bound by the RSC remodeling complex (105, 116). Enhancer-promoter interactions and the co-localization of functionally related genes seem to be favored by TADs and thus get associated with transcription activation (80, 117). Furthermore, Micro-C contact

frequencies have shown an anticorrelation between chromatin compaction and transcriptional activity (105). Altogether, the study of TAD and CID formations is of relevant importance as it contains a straight link with regulatory mechanisms.

4.3.2. Chromosome territories

Chromatin is organized into TADs which form functional compartments (Figure 16). As previously mentioned at the whole nucleosome level, 3C techniques have shown that these compartments further aggregate into distinct chromosomal territories (see Figure 16). More specifically, chromatin clusters into two types of conformations: (i) an open 10 nm conformation, termed euchromatin, which is predominantly localized in the inside of the nucleus, where it allows gene transcription, and (ii) compacted regions, termed heterochromatin, which cluster in the nuclear periphery (118, 119). This division separates the nucleus into transcriptionally active and silenced regions, generating a fundamental level of gene regulation (81) (see Figure 17).

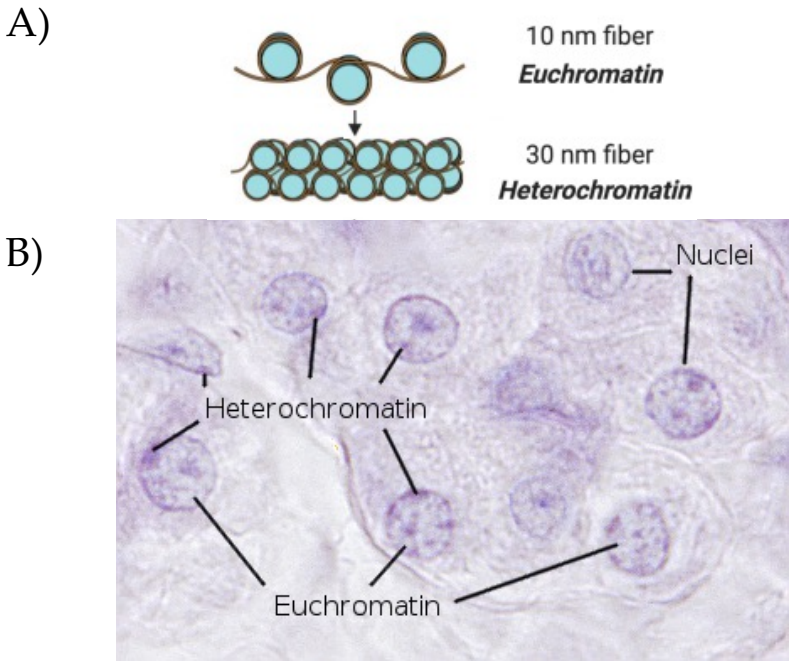


Figure 17. (A) Chromatin division into active euchromatin which occurs as an open 10 nm fiber and an inactive 30 nm condensed heterochromatin seen as a graphical illustration and (B) light microscopy image of mammal cells (image adapted from https://mmegias.webs.uvigo.es/02-english/5-celulas/4-cromatina.php).

The positioning of nucleosomes and the higher order chromatin organization are key elements in determining the accessibility of chromatin and consequently what elements affect gene expression, regulation and other fundamental processes. It is thus essential to study the mechanisms governing the 2D and 3D architectures of chromatin to understand the role that chromatin plays in biological systems.

4.4. Chromatin structure models

The capacity to gather comprehensive chromatin contact information through Hi-C and Micro-C experiments has motivated the development of various physical models aiming to explain the 3D genome conformation. These models originate from the observation that contact frequencies show similar decay patterns than those observed in polymer physics (120). We classify these models into two groups, 'bottom-up' versus 'top-down' models. The first set tries to reproduce the observed contact frequencies from 3C models by formulating a hypothetical mechanism of chromatin folding (121–125). These methods propose to model chromatin as a co-polymer given the clear correlation between TADs and chromatin modifications (see Figure 18A). Larger attractions or repulsions can be dictated by many factors such as epigenetic marks or concentration of protein binding sites acting as looping factors. The 'top-down' approach, starts by using the information from the 3C experiments, using the contact frequencies as restraints, and explores potential configurations of the chromatin in 3D. These models aim to deduce the mechanisms of chromatin folding and obtain structures that resemble the original contact data (114, 120, 126). One example of such models tries to search for a consensus structure that will minimize the deviations between the distances in the model and those derived from the experimental data (127–130) (see Figure 18B). This strategy assumes that higher frequencies of contacts mean shorter distances between loci. Since 3C-based data is obtained from a population of cells, the consensus structure represents an average that does not necessarily reflect the observed structures at a single cell resolution. One alternative to these strategies comprises resampling models, which similarly make a conversion from contact frequencies but obtain an ensemble of structures by optimizing multiple minima (131–133). Another alternative, deconvolution methods, transforms the contacts into single structures utilizing a subset of the original frequencies to provide an ensemble of possible configurations (134, 135).

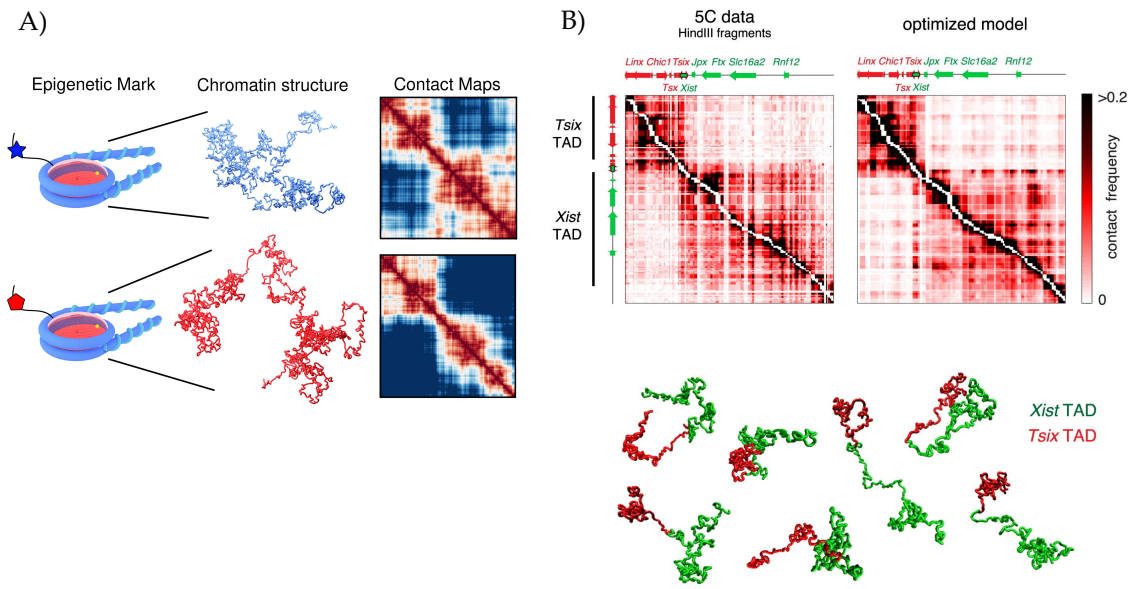


Figure 18. A) 'Bottom-up' Coarse-grained modeling of a chromatin fiber. Adapted from (130). B) 'Top-down' coarse-grained modeling of a chromatin fiber. Image adapted from (134).

Chromatin modeling provides us a deeper understanding on the role that chromatin plays in biological processes. These models can be used to better comprehend the elements that can affect chromatin organization not just under normal physiological conditions but also under DNA damage states. This broader understanding will give us a clearer idea onto the main factors that affect not just regulatory mechanisms but also cell integrity.

5. Elements that can affect chromatin organization

5.1. DNA Lesions

Gene expression mechanisms and chromatin organization can be severely altered in response to DNA damage agents which can disrupt cell homeostasis and promote diseases. The sources of DNA damage can be categorized into endogenous, which occur naturally within the cell, and exogenous, which are caused by external factors. Some examples of endogenous sources include oxidation through reactive oxygen species (ROS), hydrolysis, alkylation or methylation (136, 137). External environmental factors comprise various lifestyle components, as well as viral exposure, ionizing (IR) radiation, ultraviolet (UV) radiation or chemical agents (137). Globally, DNA damage cells undergo changes in their characteristics during DNA

damage conditions while DNA repair mechanisms are promoted through modifications in chromatin organization. For example, some studies have shown a high concordance between epigenetic marks and UV radiation response (138, 139) and a clear link between irradiation, a disruption of the chromatin architecture and reorganization of nucleosomes (105, 140). However, the exact alterations and chromatin dynamics upon DNA damage response are still poorly understood.

In this thesis we will study one of the major DNA damage mechanisms that are currently known, oxidative stress, and its effects upon chromatin structure in the yeast genome.

5.1.1. Oxidative Stress

An excess in Reactive Oxygen Species (ROS) produces oxidative stress to the cell. Concretely, this occurs when antioxidant molecules or enzymatic systems are not capable of removing the excess of chemically reactive species which contain oxygen (141, 142) (i.e., hydrogen peroxide (H_2O_2), hydroxyl radical (•OH), or superoxide anions (O_2^-)). This causes DNA damage which has been observed to be involved in more than 100 types of DNA base modifications that cause single or double strand breaks (143) and ultimately affect cell integrity (144). When antioxidant defenses are unable to counteract ROS, oxidative stress can damage nucleic acids, oxidize amino acids, and affect co-factors of proteins (145, 146). In addition, oxidative stress has been associated with aging and other complex conditions such as cancer and neurodegenerative diseases. Antioxidant mechanisms have been characterized in a series of studies in order to better understand the endogenous and exogenous responses to the stress (147–151). The molecular mechanisms which underlie the reactions to oxidative stress are still unclear. Furthermore, even though some studies have analyzed the expression of genes related to stress response, the exact signaling mechanisms that activate the response remain unclear (152). Nonetheless, it is clear that the correlation links between oxidative stress, chromatin reorganization and transcription are of great interest.

5.2. Cell Cycle

Throughout the cell cycle (G1 (growth) -> S (DNA synthesis) -> G2 (growth) -> M (mitosis), see Figure 19A), the chromatin undergoes several dynamic changes and modifications in its structure.

Some studies (153) have explored the underlying functional significance of these conformational changes, observing an increase in

fuzziness during the S and M phases in comparison to the gap G1 and G2 phases. Furthermore, a clear correlation between the dynamic changes and changes in gene expression has been found, suggesting an association between chromatin organization and cell cycle-dependent gene regulatory mechanisms. At a higher structural level, Hi-C experiments revealed a compaction of the chromatin between the G1 and M phases (77, 154). These studies showed the dependence of the increased compaction achieved in mitosis on cohesin but not condensin (see Figure 19B) (154).

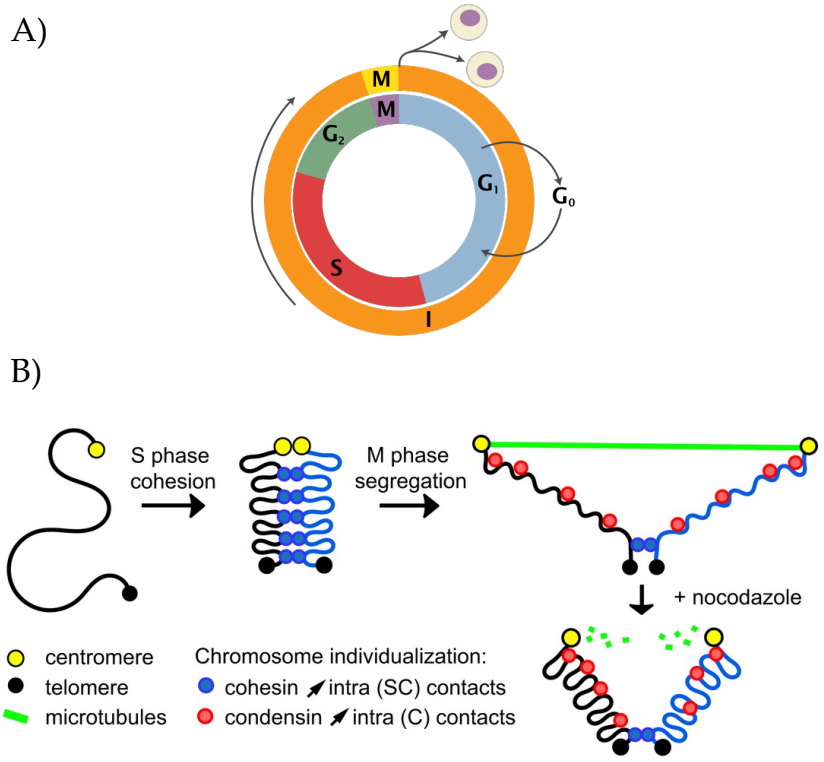


Figure 19. A) Schematic figure of the canonical cell cycle. Image adapted from https://en.wikipedia.org/wiki/Cell_cycle. B) Chromatin organization throughout the cell cycle in yeast. Chromosome individualization occurs through an increase of intra-contacts at S phase. Spindle elongation and condensin loading lead to more elongated chromosomes during the M phase. Adapted from (77).

Many factors can affect chromatin organization and this thesis will try to define some of the main characteristics that define the mechanisms that chromatin undergoes under oxidative stress DNA damage responses. Given the exponential increase in resources and data, it is important to take into consideration the use of state-of-the-art methods that can approach efficiently and accurately the large amounts of existing experimental data. One approach that has received

a growing attention giving its performance, has been to use machine learning and artificial intelligence to answer complex biological questions.

6. Machine Learning & Artificial Intelligence

The amount of available biological data has exponentially increased in recent years due to advancements such as the appearance of high-throughput sequencing. In addition, the large complexity and combinatorial analysis of nucleotide DNA and RNA sequences present an extremely complex sequence conformational space that, to this day, cannot be solved by experimental techniques alone. The analysis of these large amounts of data being generated can be quite complex but well suited for a series of computational models known as Artificial Intelligence (AI).

AI encompasses a wide range of methodologies that simulate intelligent behavior. Machine learning (ML) is a subset of AI where a model learns how to perform a task without explicitly being programmed to solve it. It is broadly said that the machine learns once it improves with training data while extracting patterns and features from it. The definitions of the task and the evaluation of performance are subject to the specific challenge that is being treated. Many types of problems can be solved by using different ML algorithms, which can be divided into three groups: supervised learning, unsupervised learning, and reinforcement learning (see Figure 20).

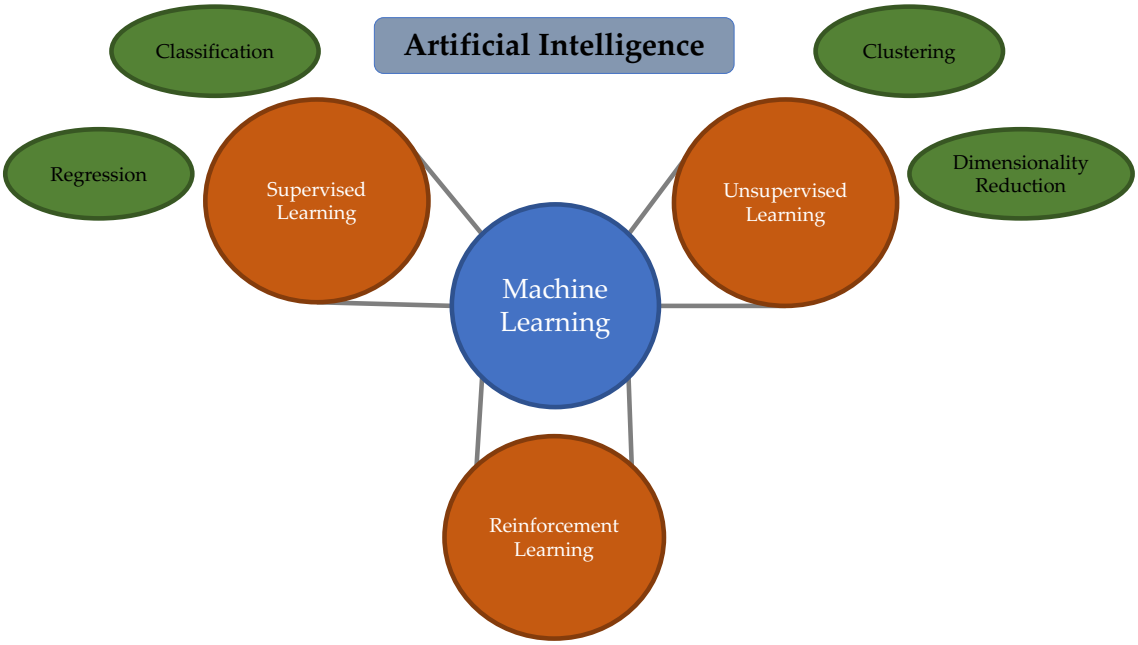


Figure 20. ML is an area of artificial intelligence that fits a model to previously observed data. ML can in turn be divided into supervised, unsupervised or reinforcement learning. Throughout this thesis we have been exploring this landscape and selecting the method that better suited our data.

In supervised learning the algorithm models, through training data, the link between the features and known output labels with the goal to be able to predict previously unseen data. In unsupervised learning the output labels are unknown, and the goal is to find patterns without human supervision. This can correspond to algorithms that will group training examples into clusters, while balancing intra-group homogeneity and model complexity, or reducing the dimensionality of the analyzed data for different purposes. Lastly, reinforcement learning entails methods where an agent rewards a desired action while undesired ones get punished. This last set of methods is however beyond the scope of this thesis.

6.1. The supervised learning overview

During our supervised learning methods, regardless of the task that we are performing, we will always find an input **x** that is mapped to a prediction **y**. The mapping occurs through a function $\mathbf{f[x, \phi]}$, having $\mathbf{y = f[x, \phi]}$ where ϕ denotes the parameters of the model. When we compute the predictions **y** from the inputs **x**, we call this *inference*. In the process of training a model we attempt to find the parameters ϕ

that will map each training input to its associated output as closely as possible, and we do so by quantifying the degree of mismatch with a *loss L*. More specifically, the loss returns a value that calibrates how well a model with parameters ϕ predicts the training outputs. We can thus summarize the loss as a function $L[\phi]$ of a model's parameters, and seek the $\hat{\phi}$ that will minimize the function:

$$\hat{\phi} = \underset{\phi}{\operatorname{argmin}}[L[\phi]] \qquad (1)$$

This overall quantity that we are trying to minimize is known as the *cost* or *objective function*. Once we have found the parameters ϕ that minimize our objective function, we must assess the performance of our model. To do so, we predict new values for a separate test set data in order to see how well the model *generalizes* to examples that were not previously observed.

6.2. Classification vs. Regression tasks

Two types of learning tasks can be done by ML algorithms: classification and regression. In both cases, a real-world input is encoded as a vector of numbers and the model tries to map it to an output vector, the label. Classification tasks are those where the model assigns a discrete label to a set of features either through a supervised or unsupervised manner. When the model tries to assign the input to one of two categories, we define it as a binary classification given that the model only assigns one label or another. If the model assigns the input to one of N > 2 categories, we depict this as a multiclass classification problem. In this case the model will return a vector of size N that contains the probabilities for each of the N categories.

In regression problems the output is no longer discrete but rather continuous and approximated by a real-valued function as accurately as possible. This is done by minimizing what is known as the cost function which is based on a set of parameters that are iteratively optimized with methods such as gradient descent. Similarly, if the model predicts more than one continuous output, we define this as a multivariate regression problem. Figure 21 shows some examples of the different tasks.

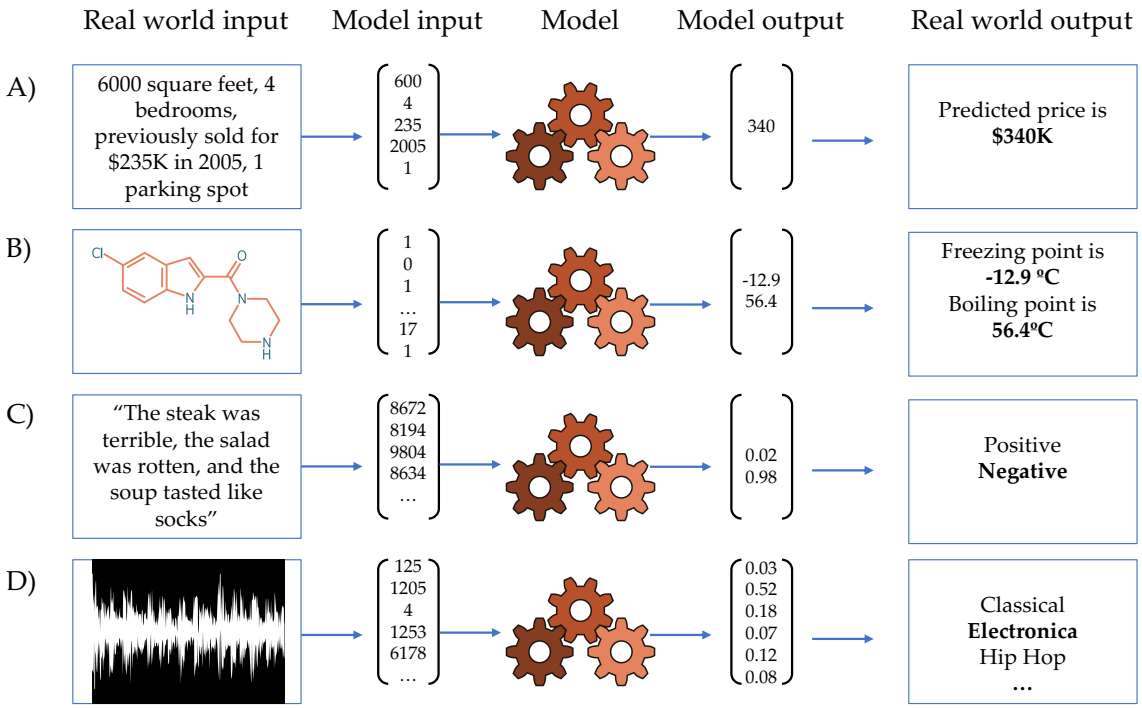


Figure 21. Regression and classification problems. A) This regression model takes a vector of numbers that characterize a property and predicts its price. B) This multivariate regression model takes the structure of a chemical molecule and predicts its melting and boiling points. C) This binary classification model takes a restaurant review and classifies it as either positive or negative. D) This multiclass classification problem assigns a snippet of audio to one of N genres. Image adapted from (155).

6.3. Cost functions

Different cost functions are used for the many different ML architectures that exist, and the objective can be to either maximize or minimize the selected function. In other words, the cost function is defined depending on what we are interested in predicting. These, however, can be classified into three groups:

1. Regression cost functions: This group includes functions used in regression problems where we minimize the error between the predicted and the actual outputs. Some examples include the Least squares loss, Mean Error (ME), the Mean Absolute Error (MAE), the Mean Squared Error (MSE) or the Root Mean Squared Error (RMSE).

2. Binary classification cost functions: These are commonly used cost functions for binary classification problems. They measure the differences between the predicted and actual probabilities of a binary output variable. Some examples include the binary cross-entropy (also known as the log loss), or the Hinge loss.
3. Multi-class classification cost functions: These include the cost functions used for multi-class classification problems. Similarly to the binary log loss, they measure the differences between the actual and the predicted probabilities of the multiple output classes. The most commonly used is called the categorical cross-entropy.

6.4. Fitting models

The learning process of a machine learning algorithm consists on training and fitting the model to find the parameters that will minimize or maximize the objective function. This step is known as the optimization process, in which we iteratively update the parameters to improve the performance of a model. Several algorithms have been developed to adjust the parameters of a model's architecture including Newton's and Quasi-Newton's methods, Gradient Descent techniques (GD), Stochastic Optimization techniques, Evolutionary algorithms, Genetic algorithms, Grid Search or the Adaptive Moment Estimation (ADAM) algorithm amongst many others.

There are several techniques that can be used to improve a model's robustness and better assess the choice of parameters. A common approach consists on using k-fold cross validation, where the training set is divided into k smaller sets and the optimization of parameters occurs each time on a different k-sized validation subset. The final evaluation of the model will then be done on a never before used testing set (see Figure 22).

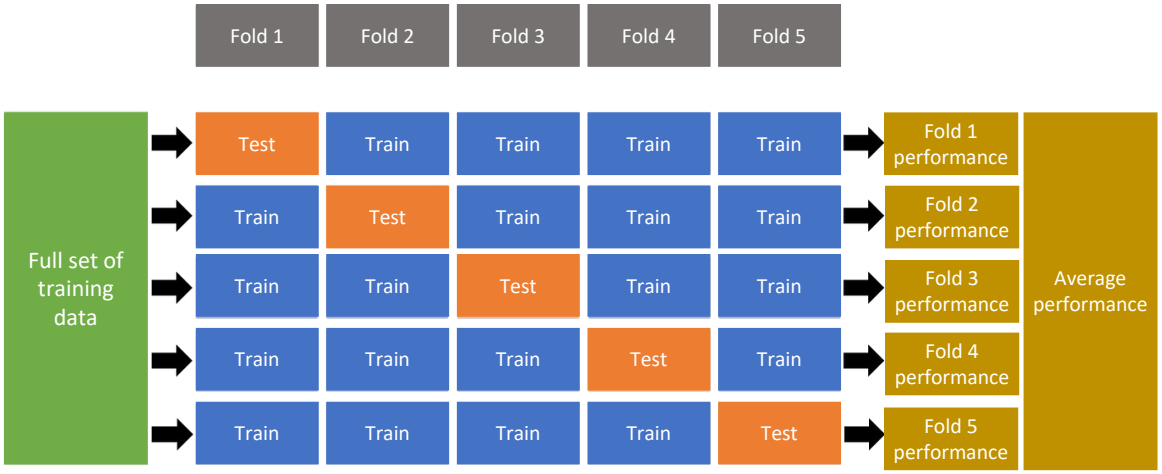


Figure 22. Cross validation scheme illustrated here for the case of K = 5. The technique involves taking the available data and partitioning it into K groups, using K – 1 groups to train the model and evaluating it on the remaining group. Lastly, the performances are averaged out (161).

6.5. Underfitting vs. overfitting, the bias-variance trade off

A machine learning model will learn to reduce a *cost function* or accurately predict classes for a training set. This however does not guarantee that the model will perform accurately when predicting unseen testing data. The degree to which the model *generalizes* to the test data, partially depends on how representative and complete the training data is. When the model fails to predict new data, it often means that the model is overfitting or underfitting. When it learns to characterize the training examples but these are not a representation of the underlying data, the model describes the statistical peculiarities of the training data but simultaneously overfits it. Overfitting occurs when predictions cannot be generalized, and the cost of the testing set is much higher than the cost of the training set. Contrarily, a very simple model that is unable to capture the true relationship between the inputs and outputs of the data is said to underfit.

One of the motivations when optimizing a model is to balance the cost of overfitting versus underfitting in order to obtain the best representation (see some examples in Figure 23). This is known as the bias-variance trade-off. As we add capacity to a model, the bias decreases, but the variance increases for a fixed-size training dataset.

The goal when constructing a ML algorithm is to find the optimal balance between its bias and variance.

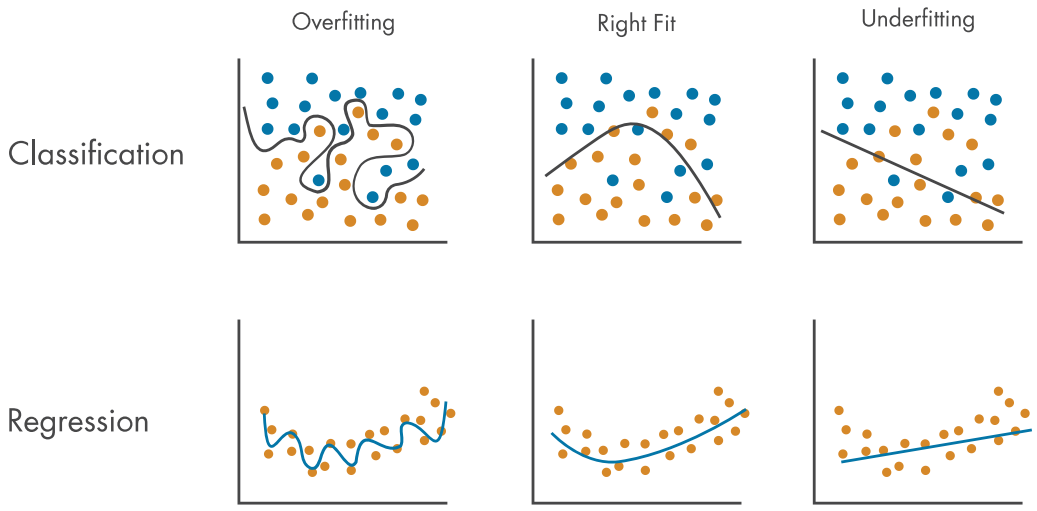


Figure 23. Overfitted (left panels) and underfitted (right panels) classification (top panels) and regression (bottom panels) models in comparison to correctly fitted models (middle panels). Image adapted from https://www.mathworks.com/discovery/overfitting.html.

6.6. Regularization

In order to reduce the generalization gap between the training and test sets we can also add explicit or implicit terms to the chosen loss function that will help favor or avoid certain parameters. This process is known as *regularization*.

The most broadly known techniques are L1 and L2 regularization. The first method, also named *LASSO* (least absolute shrinkage and selection operator) regularization, penalizes the absolute values of the weights (see Eq. 2) while the second one, also termed *Ridge* regularization, imposes a penalty on the sum of the squares of the parameter values (see Eq. 3).

$$\hat{\phi} = \underset{\phi}{\operatorname{argmin}}\left[\sum_{i=1}^{I} l_i[x_i, y_i] + \lambda \sum_j |\phi_j|\right] \quad (2)$$

$$\hat{\phi} = \underset{\phi}{\operatorname{argmin}}\left[\sum_{i=1}^{I} l_i[x_i, y_i] + \lambda \sum_j \phi_j^2\right] \quad (3)$$

Several techniques have been described to help improve generalization, including early stopping, dropout, adding noise,

transfer learning or data augmentation amongst others. Nonetheless, all of them can be grouped into four principles; those that force the function to be smoother (e.g., L2 regularization), those that increase the amount of data (e.g., data augmentation), the combination of models (e.g., ensembling), or the search for a wider minima (e.g., applying noise) (see Figure 24).

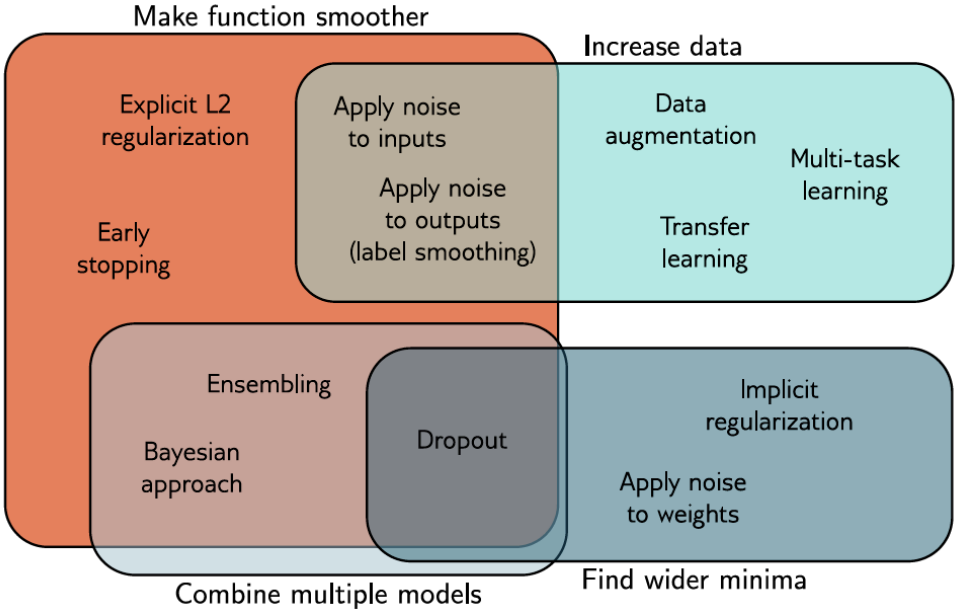


Figure 24. Regularization methods divided into four groups. The first group of methods aim to make the modeled function smoother. The second ones increase the effective amount of data. The third group of methods combine multiple models and hence mitigate against uncertainty in the fitting process. Finally, the fourth group of methods encourages the training process to converge to a wide minimum where small errors in the estimated parameters are less important. Image adapted from (155).

Throughout this thesis several ML and AI methods will be used to analyze large amounts of data and disentangle the underlying elements in gene regulatory mechanisms. More specifically, Chapters 1 will give a brief overview of the use of a Random Forest Regressor to model binding affinities while Chapter 2 the use of a shallow neural network allows the prediction of nucleosome depleted regions in essential gene regulatory regions.

6.7. Random Forest Regressor

A Random Forest Regressor (156, 157) is an ensemble learning method used for regression tasks that combines the predictive power of multiple decision trees to improve accuracy and mitigate overfitting. The forest is built during training, defining each tree from a random subset of the training data and features. The final prediction is obtained by averaging the outputs of these individual trees in order to enhance the generalization capabilities of the model. The inherent randomness introduced in the data sampling and feature selection processes contributes to the diversity among trees, thereby strengthening the model's adaptability to noise and variance. In constructing each decision tree within the random forest, the model utilizes metrics such as *entropy* and *information gain* to determine the optimal splits at each node. The *entropy* (equation 4) is defined as a measure of uncertainty or disorder, and calculated using the formula below:

$$E(S) = \sum_{i=1}^{c} -p_i \, (log_2 \, p_i) \qquad (4)$$

where S is the set of samples, c is the number of classes, and p_i is the proportion of samples belonging to class i. Information gain quantifies the reduction in entropy achieved by partitioning the data according to a given attribute. By maximizing the information gain at each node, the decision trees within the random forest ensure that the splits are highly informative, leading to more accurate and reliable predictions.

6.8. Shallow Neural Network Classifier

A shallow neural network, first motivated by (158) and often referred to as a single-layer neural network (157), is a fundamental model in the field of ML. In summary, it consists of an input layer, a hidden layer and an output layer that are defined by neurons (see Figure 25).

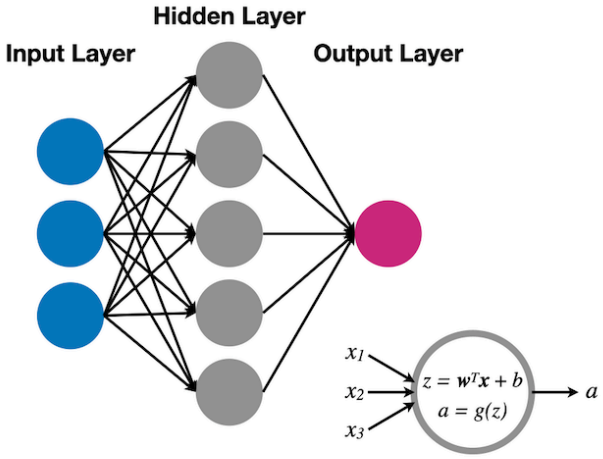


Figure 25. Schematic architecture of a shallow neural network with 3 input neurons, a hidden layer of 5 neurons and a final output neuron. Image adapted from (https://scipython.com/blog/a-shallow-neural-network-for-simple-nonlinear-classification/).

Despite its simplicity, a shallow neural network can approximate a wide range of functions and effectively solve various predictive tasks. Each neuron in the hidden layer receives an input signal x_i from the i^{th} input neuron, which is linearly combined using weights w_i and biases b. This linear combination is then passed through an activation function $g(z)$ to introduce non-linearity into the model, enabling the network to capture complex patterns in the data (see Figure 25). Lastly, the outputs of the activation functions are fed into the output layer, which produces a final prediction.

Training a shallow neural network involves adjusting the weights and biases to minimize a loss function. This optimization is commonly performed using gradient descent algorithms, where the gradients of the loss with respect to the weights are computed and used to update the new weights iteratively.

Throughout this thesis we will investigate how ML methods have the potential to model the whole yeast genome at a very high resolution to help untangle essential elements of chromatin structure.

References

1. Watson,J.D. and Crick,F.H.C. (1953) Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature, 171.
2. Crick,F.H.C. (1958) On Protein Syntesis. Symp. Soc. Exp. Biol. XII.
3. Muirhead,H. and Perutz,M.F. (1963) Structure of hæemoglobin: A three-dimensional fourier synthesis of reduced human haemoglobin at 5.5 Å resolution. Nature, 199.
4. Arnott,S. and Hukins,D.W.L. (1973) Refinement of the structure of B-DNA and implications for the analysis of X-ray diffraction data from fibers of biopolymers. J Mol Biol, 81.
5. Olson,W.K., Bansal,M., Burley,S.K., Dickerson,R.E., Gerstein,M., Harvey,S.C., Heinemann,U., Lu,X.J., Neidle,S., Shakked,Z., et al. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. J Mol Biol, 313.
6. Blanchet,C., Pasi,M., Zakrzewska,K. and Lavery,R. (2011) CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. Nucleic Acids Res, 39.
7. Lu,X.J. and Olson,W.K. (2003) 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res, 31.
8. Pasi,M., Maddocks,J.H. and Lavery,R. (2015) Analyzing ion distributions around DNA: Sequence-dependence of potassium ion distributions from microsecond molecular dynamics. Nucleic Acids Res, 43.
9. Ts'o,P.O.P. and Kan,L.-S. (1979) Nuclear Magnetic Resonance Studies of Nucleic Acids and Proteins. In Chromatin Structure and Function.
10. Pérez,A., Luque,F.J. and Orozco,M. (2012) Frontiers in molecular dynamics simulations of DNA. Acc Chem Res, 45.
11. Cheatham,T.E. and Case,D.A. (2013) Twenty-five years of nucleic acid simulations. Biopolymers, 99.
12. Dans,P.D., Walther,J., Gómez,H. and Orozco,M. (2016) Multiscale simulation of DNA. Curr Opin Struct Biol, 37.
13. Hamelberg,D., McFail-Isom,L., Williams,L.D. and David Wilson,W. (2000) Flexible structure of DNA: Ion dependence of minor-groove structure and dynamics. J Am Chem Soc, 122.
14. Kanhere,A. and Bansal,M. (2005) Structural properties of promoters: Similarities and differences between prokaryotes and eukaryotes. Nucleic Acids Res, 33.
15. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A, 95.

16. Lankaš,F., Šponer,J., Hobza,P. and Langowski,J. (2000) Sequence-dependent elastic properties of DNA. J Mol Biol, 299.
17. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C., et al. (2014) μABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. Nucleic Acids Res, 42.
18. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dixit,S., Jayaram,B., Lankas,F., Laughton,C., et al. (2009) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. Nucleic Acids Res, 38.
19. Neidle,S. (2007) Principles of Nucleic Acid Structure.
20. Klostermeier,D. and Hammann,C. (2013) RNA structure and folding: Biophysical techniques and prediction methods.
21. Varani,G. and McClain,W.H. (2000) The G·U wobble base pair: A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. EMBO Rep, 1.
22. Xu,D., Landon,T., Greenbaum,N.L. and Fenley,M.O. (2007) The electrostatic characteristics of G · U wobble base pairs. Nucleic Acids Res, 35.
23. Pauling,L. and Corey,R.B. (1953) A Proposed Structure For The Nucleic Acids. Proceedings of the National Academy of Sciences, 39.
24. Felsenfeld,G., Davies,D.R. and Rich,A. (1957) Formation of a Three-Stranded Polynucleotide Molecule. J Am Chem Soc, 79.
25. Felsenfeld,G. and Rich,A. (1957) Studies on the formation of two- and three-stranded polyribonucleotides. BBA - Biochimica et Biophysica Acta, 26.
26. Mergny,J.L., Sun,J.S., Rougée,M., Montenay-Garestier,T., Chomilier,J., Hélène,C. and Barcelo,F. (1991) Sequence Specificity in Triple-Helix Formation: Experimental and Theoretical Studies of the Effect of Mismatches on Triplex Stability. Biochemistry, 30.
27. Potaman,V.N. and Sinden,R.R. (1995) Stabilization of Triple-Helical Nucleic Acids by Basic Oligopeptides. Biochemistry, 34.
28. Scaria,P. V and Shafer,R.H. (1996) Calorimetric analysis of triple helices targeted to the d(G3A4G3)·d(C3T4C3) duplex. Biochemistry, 35.
29. Robles,J., Grandas,A., Pedroso,E., Luque,F., Eritja,R. and Orozco,M. (2005) Nucleic Acid Triple Helices: Stability Effects of Nucleobase Modifications. Curr Org Chem, 6.
30. Goñi,J.R., de la Cruz,X. and Orozco,M. (2004) Triplex-forming oligonucleotide target sequences in the human genome. Nucleic Acids Res, 32.
31. Goñi,J.R., Vaquerizas,J.M., Dopazo,J. and Orozco,M. (2006) Exploring the reasons for the large density of triplex-forming

oligonucleotide target sequences in the human regulatory regions. BMC Genomics, 7.
32. Flores,O., Deniz,Ö., Soler-López,M. and Orozco,M. (2014) Fuzziness and noise in nucleosomal architecture. Nucleic Acids Res, 42, 4934–4946.
33. Deniz,Ö., Flores,O., Battistini,F., Pérez,A., Soler-López,M. and Orozco,M. (2011) Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. BMC Genomics, 12, 489.
34. D'Oliveira Albanus,R., Kyono,Y., Hensley,J., Varshney,A., Orchard,P., Kitzman,J.O. and Parker,S.C.J. (2021) Chromatin information content landscapes inform transcription factor and DNA interactions. Nat Commun, 12.
35. Voss,T.C. and Hager,G.L. (2014) Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat Rev Genet, 15.
36. Li,B., Carey,M. and Workman,J.L. (2007) The Role of Chromatin during Transcription. Cell, 128.
37. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. Annu Rev Biochem, 79.
38. Haberle,V. and Stark,A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol, 19.
39. Afek,A. and Lukatsky,D.B. (2012) Nonspecific protein-DNA binding is widespread in the yeast genome. Biophys J, 102.
40. Von Hippel,P.H. and Berg,O.G. (1986) On the specificity of DNA-protein interactions. Proc Natl Acad Sci U S A, 83.
41. Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) Sequence specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci U S A, 73.
42. Bewley,C.A., Gronenborn,A.M. and Clore,G.M. (1998) Minor groove-binding architectural proteins: Structure, function, and DNA recognition. Annu Rev Biophys Biomol Struct, 27.
43. Harteis,S. and Schneider,S. (2014) Making the bend: DNA tertiary structure and protein-DNA interactions. Int J Mol Sci, 15.
44. Wecker,K., Bonnet,M.C., Meurs,E.F. and Delepierre,M. (2002) The role of the phosphorus BI-BII transition in protein-DNA recognition: The NF-κB complex. Nucleic Acids Res, 30.
45. Hunter,C.A. (1993) Sequence-dependent DNA structure the role of base stacking interactions. J Mol Biol, 230.
46. Lavery,R. and Pullman,B. (1982) The electrostatic field of DNA: The role of the nucleic acid conformation. Nucleic Acids Res, 10.

47. Siggers,T. and Gordân,R. (2014) Protein-DNA binding: Complexities and multi-protein codes. Nucleic Acids Res, 42.
48. Smaczniak,C., Angenent,G.C. and Kaufmann,K. (2017) SELEX-seq: A method to determine DNA binding specificities of plant transcription factors. In Methods in Molecular Biology.Vol. 1629.
49. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G., et al. (2013) DNA-binding specificities of human transcription factors. Cell, 152.
50. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M.J., et al. (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res, 20.
51. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol, 24.
52. Mundade,R., Ozer,H.G., Wei,H., Prabhu,L. and Lu,T. (2014) Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle, 13.
53. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell, 147.
54. Mahony,S. and Pugh,B.F. (2015) Protein-DNA binding in high-resolution. Crit Rev Biochem Mol Biol, 50.
55. Rhee,H.S. and Pugh,B.F. (2008) ChIP-exo: A Method to Identify Genomic Location of DNA-binding proteins at Near Single Nucleotide Accuracy. Curr Protoc Mol Biol, 141.
56. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. In Bioinformatics.Vol. 15.
57. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: A new way to display consensus sequences. Nucleic Acids Res, 18.
58. Benos,P. V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: How good an approximation is it? Nucleic Acids Res, 30.
59. Zhao,Y., Ruan,S., Pandey,M. and Stormo,G.D. (2012) Improved models for transcription factor binding site identification using nonindependent interactions. Genetics, 191.
60. Zhou,T., Shen,N., Yang,L., Abe,N., Horton,J., Mann,R.S., Bussemaker,H.J., Gordân,R. and Rohs,R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. Proc Natl Acad Sci U S A, 112.

61. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordân,R. and Rohs,R. (2014) Absence of a simple code: How transcription factors read the genome. Trends Biochem Sci, 39.
62. Wang,S., Zhang,Q., Shen,Z., He,Y., Chen,Z.H., Li,J. and Huang,D.S. (2021) Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture. Mol Ther Nucleic Acids, 24.
63. Peng,P.C. and Sinha,S. (2016) Quantitative modeling of gene expression using DNA shape features of binding sites. Nucleic Acids Res, 44.
64. Koo,P.K. and Ploenzke,M. (2020) Deep learning for inferring transcription factor binding sites. Curr Opin Syst Biol, 19.
65. Cevost,J., Vaillant,C. and Meyer,S. (2018) ThreaDNA: Predicting DNA mechanics' contribution to sequence selectivity of proteins along whole genomes. Bioinformatics, 34.
66. Chen,C., Hou,J., Shi,X., Yang,H., Birchler,J.A. and Cheng,J. (2021) DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks. BMC Bioinformatics, 22.
67. Dai,H., Umarov,R., Kuwahara,H., Li,Y., Song,L. and Gao,X. (2017) Sequence2Vec: A novel embedding approach for modeling transcription factor binding affinity landscape. Bioinformatics, 33.
68. Li,J., Sagendorf,J.M., Chiu,T.P., Pasi,M., Perez,A. and Rohs,R. (2017) Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. Nucleic Acids Res, 45.
69. Fu,L., Zhang,L., Dollinger,E., Peng,Q., Nie,Q. and Xie,X. (2020) Predicting transcription factor binding in single cells through deep learning. Sci Adv, 6.
70. Park,S., Koh,Y., Jeon,H., Kim,H., Yeo,Y. and Kang,J. (2020) Enhancing the interpretability of transcription factor binding site prediction using attention mechanism. Sci Rep, 10.
71. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol, 33.
72. Zhang,Q., Shen,Z. and Huang,D.S. (2021) Predicting in-vitro Transcription Factor Binding Sites Using DNA Sequence + Shape. IEEE/ACM Trans Comput Biol Bioinform, 18.
73. Asif,M. and Orenstein,Y. (2020) DeepSELEX: Inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs. Bioinformatics, 36.
74. Zeng,Y., Gong,M., Lin,M., Gao,D. and Zhang,Y. (2020) A Review about Transcription Factor Binding Sites Prediction Based on Deep Learning. IEEE Access, 8.

75. Lajoie,B.R., Dekker,J. and Kaplan,N. (2015) The Hitchhiker's guide to Hi-C analysis: Practical guidelines. Methods, 72.
76. Cavalli,G. and Misteli,T. (2013) Functional implications of genome topology. Nat Struct Mol Biol, 20.
77. Lazar-Stefanita,L., Scolari,V.F., Mercy,G., Muller,H., Guérin,T.M., Thierry,A., Mozziconacci,J. and Koszul,R. (2017) Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. EMBO J, 36.
78. Vizcaya-Molina,E., Klein,C.C., Serras,F. and Corominas,M. (2020) Chromatin dynamics in regeneration epithelia: Lessons from Drosophila imaginal discs. Semin Cell Dev Biol, 97.
79. van Steensel,B. and Furlong,E.E.M. (2019) The role of transcription in shaping the spatial organization of the genome. Nat Rev Mol Cell Biol, 20.
80. Rowley,M.J. and Corces,V.G. (2018) Organizational principles of 3D genome architecture. Nat Rev Genet, 19.
81. Misteli,T. (2007) Beyond the Sequence: Cellular Organization of Genome Function. Cell, 128.
82. Tonna,S., El-Osta,A., Cooper,M.E. and Tikellis,C. (2010) Metabolic memory and diabetic nephropathy: Potential role for epigenetic mechanisms. Nat Rev Nephrol, 6.
83. Battistini,F., Hunter,C.A., Gardiner,E.J. and Packer,M.J. (2010) Structural mechanics of DNA wrapping in the nucleosome. J Mol Biol, 396.
84. Richmond,T.J., Finch,J.T., Rushton,B., Rhodes,D. and Klug,A. (1984) Structure of the nucleosome core particle at 7 resolution. Nature, 311.
85. Luger,K., Mäder,A.W., Richmond,R.K., Sargent,D.F. and Richmond,T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature, 389.
86. Reynolds,S.M., Bilmes,J.A. and Noble,W.S. (2010) Learning a weighted sequence model of the nucleosome core and linker yields more accurate predictions in Saccharomyces cerevisiae and Homo sapiens. PLoS Comput Biol, 6.
87. Hu,J., Gu,L., Ye,Y., Zheng,M., Xu,Z., Lin,J., Du,Y., Tian,M., Luo,L., Wang,B., et al. (2018) Dynamic placement of the linker histone H1 associated with nucleosome arrangement and gene transcription in early Drosophila embryonic development. Cell Death Dis, 9.
88. Giresi,P.G., Kim,J., McDaniell,R.M., Iyer,V.R. and Lieb,J.D. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res, 17.
89. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and

sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods, 10.
90. Schones,D.E., Cui,K., Cuddapah,S., Roh,T.Y., Barski,A., Wang,Z., Wei,G. and Zhao,K. (2008) Dynamic Regulation of Nucleosome Positioning in the Human Genome. Cell, 132.
91. Tsompana,M. and Buck,M.J. (2014) Chromatin accessibility: A window into the genome. Epigenetics Chromatin, 7.
92. Lai,W.K.M. and Pugh,B.F. (2017) Understanding nucleosome dynamics and their links to gene expression and DNA replication. Nat Rev Mol Cell Biol, 18, 548–562.
93. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods, 9.
94. Flores,O. and Orozco,M. (2011) nucleR: a package for non-parametric nucleosome positioning. Bioinformatics, 27, 2149–2150.
95. Buitrago,D., Codó,L., Illa,R., de Jorge,P., Battistini,F., Flores,O., Bayarri,G., Royo,R., Del Pino,M., Heath,S., et al. (2019) Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning. Nucleic Acids Res, 47, 9511–9523.
96. Ohno,M., Priest,D.G. and Taniguchi,Y. (2018) Nucleosome-level 3D organization of the genome. Biochem Soc Trans, 46.
97. Finch,J.T. and Klug,A. (1976) Solenoidal model for superstructure in chromatin. Proc Natl Acad Sci U S A, 73.
98. Woodcock,C.L.F., Frado,L.L.Y. and Rattner,J.B. (1984) The higher-order structure of chromatin: Evidence for a helical ribbon arrangement. Journal of Cell Biology, 99.
99. Williams,S.P., Athey,B.D., Muglia,L.J., Schappe,R.S., Gough,A.H. and Langmore,J.P. (1986) Chromatin fibers are left-handed double helices with diameter and mass per unit length that depend on linker length. Biophys J, 49.
100. Dorigo,B., Schalch,T., Kulangara,A., Duda,S., Schroeder,R.R. and Richmond,T.J. (2004) Nucleosome arrays reveal the two-start organization of the chromatin fiber. Science (1979), 306.
101. Fussner,E., Ching,R.W. and Bazett-Jones,D.P. (2011) Living without 30nm chromatin fibers. Trends Biochem Sci, 36.
102. Maeshima,K., Hihara,S. and Eltsov,M. (2010) Chromatin structure: Does the 30-nm fibre exist in vivo? Curr Opin Cell Biol, 22.
103. Boettiger,A.N., Bintu,B., Moffitt,J.R., Wang,S., Beliveau,B.J., Fudenberg,G., Imakaev,M., Mirny,L.A., Wu,C.T. and Zhuang,X. (2016) Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature, 529.
104. Ricci,M.A., Manzo,C., García-Parajo,M.F., Lakadamyali,M. and Cosma,M.P. (2015) Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. Cell, 160.

105. Hsieh,T.H.S., Weiner,A., Lajoie,B., Dekker,J., Friedman,N. and Rando,O.J. (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell, 162.
106. Hsieh,T.H.S., Fudenberg,G., Goloborodko,A. and Rando,O.J. (2016) Micro-C XL: Assaying chromosome conformation from the nucleosome to the entire genome. Nat Methods, 13.
107. Cremer,T. and Cremer,C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet, 2.
108. Lieberman-Aiden,E., Van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O., et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (1979), 326.
109. Sikorska,N. and Sexton,T. (2020) Defining Functionally Relevant Spatial Chromatin Domains: It is a TAD Complicated. J Mol Biol, 432.
110. Pueschel,R., Coraggio,F. and Meister,P. (2016) From single genes to entire genomes: The search for a function of nuclear organization. Development (Cambridge), 143.
111. Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., Van Berkum,N.L., Meisig,J., Sedat,J., et al. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature, 485.
112. Tan-Wong,S.M., Zaugg,J.B., Camblong,J., Xu,Z., Zhang,D.W., Mischo,H.E., Ansari,A.Z., Luscombe,N.M., Steinmetz,L.M. and Proudfoot,N.J. (2012) Gene loops enhance transcriptional directionality. Science (1979), 338.
113. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Leblanc,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. Cell, 148.
114. Duan,Z., Andronescu,M., Schutz,K., McIlwain,S., Kim,Y.J., Lee,C., Shendure,J., Fields,S., Blau,C.A. and Noble,W.S. (2010) A three-dimensional model of the yeast genome. Nature, 465.
115. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S., et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159.
116. Ramírez,F., Bhardwaj,V., Arrigoni,L., Lam,K.C., Grüning,B.A., Villaveces,J., Habermann,B., Akhtar,A. and Manke,T. (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun, 9.
117. Sexton,T. and Cavalli,G. (2015) The role of chromosome domains in shaping the functional genome. Cell, 160.

118. Brown,S.W. (1966) Heterochromatin. Science (1979), 151.
119. Weintraub,H. and Groudine,M. (1976) Chromosomal subunits in active genes have an altered conformation. Science (1979), 193.
120. Varoquaux,N., Ay,F., Noble,W.S. and Vert,J.P. (2014) A statistical approach for inferring the 3D structure of the genome. In Bioinformatics.Vol. 30.
121. Jost,D., Carrivain,P., Cavalli,G. and Vaillant,C. (2014) Modeling epigenome folding: Formation and dynamics of topologically associated chromatin domains. Nucleic Acids Res, 42.
122. Bianco,S., Lupiáñez,D.G., Chiariello,A.M., Annunziatella,C., Kraft,K., Schöpflin,R., Wittler,L., Andrey,G., Vingron,M., Pombo,A., et al. (2018) Polymer physics predicts the effects of structural variants on chromatin architecture. Nat Genet, 50.
123. Racko,D., Benedetti,F., Dorier,J. and Stasiak,A. (2018) Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes. Nucleic Acids Res, 46.
124. Fudenberg,G., Imakaev,M., Lu,C., Goloborodko,A., Abdennur,N. and Mirny,L.A. (2016) Formation of Chromosomal Domains by Loop Extrusion. Cell Rep, 15.
125. Sanborn,A.L., Rao,S.S.P., Huang,S.C., Durand,N.C., Huntley,M.H., Jewett,A.I., Bochkov,I.D., Chinnappan,D., Cutkosky,A., Li,J., et al. (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A, 112.
126. Polles,G., Hua,N., Yildirim,A. and Alber,F. (2019) Genome Structure Calculation through Comprehensive Data Integration. In Modeling the 3D Conformation of Genomes.
127. Buitrago,D., Labrador,M., Arcon,J.P., Lema,R., Flores,O., Esteve-Codina,A., Blanc,J., Villegas,N., Bellido,D., Gut,M., et al. (2021) Impact of DNA methylation on 3D genome structure. Nat Commun, 12, 3243.
128. Hu,M., Deng,K., Qin,Z., Dixon,J., Selvaraj,S., Fang,J., Ren,B. and Liu,J.S. (2013) Bayesian Inference of Spatial Organizations of Chromosomes. PLoS Comput Biol, 9.
129. Zhang,Z., Li,G., Toh,K.C. and Sung,W.K. (2013) 3D chromosome modeling with semi-definite programming and Hi-C data. Journal of Computational Biology, 20.
130. Moller,J. and de Pablo,J.J. (2020) Bottom-Up Meets Top-Down: The Crossroads of Multiscale Chromatin Modeling. Biophys J, 118.
131. Serra,F., Baù,D., Goodstadt,M., Castillo,D., Filion,G. and Marti-Renom,M.A. (2017) Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput Biol, 13.

132. Di Stefano,M., Paulsen,J., Lien,T.G., Hovig,E. and Micheletti,C. (2016) Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. Sci Rep, 6.
133. Paulsen,J., Sekelja,M., Oldenburg,A.R., Barateau,A., Briand,N., Delbarre,E., Shah,A., Sørensen,A.L., Vigouroux,C., Buendia,B., et al. (2017) Chrom3D: Three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. Genome Biol, 18.
134. Giorgetti,L., Galupa,R., Nora,E.P., Piolot,T., Lam,F., Dekker,J., Tiana,G. and Heard,E. (2014) Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell, 157.
135. Tjong,H., Li,W., Kalhor,R., Dai,C., Hao,S., Gong,K., Zhou,Y., Li,H., Zhou,X.J., Le Gros,M.A., et al. (2016) Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proc Natl Acad Sci U S A, 113.
136. Sall,S.O., Johann To Berens,P. and Molinier,J. (2022) DNA damage and DNA methylation. In Epigenetics and DNA Damage.
137. Hakem,R. (2008) DNA-damage repair; the good, the bad, and the ugly. EMBO Journal, 27.
138. Shen,Y., Stanislauskas,M., Li,G., Zheng,D. and Liu,L. (2017) Epigenetic and genetic dissections of UV-induced global gene dysregulation in skin cells through multi-omics analyses. Sci Rep, 7.
139. Zentner,G.E., Tesar,P.J. and Scacheri,P.C. (2011) Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res, 21.
140. Van Eijk,P., Nandi,S.P., Yu,S., Bennett,M., Leadbitter,M., Teng,Y. and Reed,S.H. (2019) Nucleosome remodeling at origins of global genome–nucleotide excision repair occurs at the boundaries of higher-order chromatin structure. Genome Res, 29.
141. Finkel,T. and Holbrook,N.J. (2000) Oxidants, oxidative stress and the biology of ageing. Nature, 408.
142. Herrero,E., Ros,J., Bellí,G. and Cabiscol,E. (2008) Redox control and oxidative stress in yeast cells. Biochim Biophys Acta Gen Subj, 1780.
143. Girard,P.M. and Boiteux,S. (1997) Repair of oxidized DNA bases in the yeast saccharomyces cerevisiae. In Biochimie.Vol. 79.
144. Costa,V. and Moradas-Ferreira,P. (2001) Oxidative stress and signal transduction in Saccharomyces cerevisiae: Insights into ageing, apoptosis and diseases. Mol Aspects Med, 22.
145. Imlay,J.A. (2015) Transcription Factors That Defend Bacteria Against Reactive Oxygen Species. Annu Rev Microbiol, 69.

146. Imlay,J.A. (2015) Diagnosing oxidative stress in bacteria: not as easy as you might think. Curr Opin Microbiol, 24.
147. Storz,G., Tartaglia,L.A., Farr,S.B. and Ames,B.N. (1990) Bacterial defenses against oxidative stress. Trends in Genetics, 6.
148. Storz,G., Tartaglia,L.A. and Ames,B.N. (1990) The OxyR regulon. Antonie Van Leeuwenhoek, 58.
149. Charizanis,C., Juhnke,H., Krems,B. and Entian,K.D. (1999) The oxidative stress response mediated via Pos9/Skn7 is negatively regulated by the Ras/PKA pathway in Saccharomyces cerevisiae. Molecular and General Genetics, 261.
150. Charizanis,C., Juhnke,H., Krems,B. and Entian,K.D. (1999) The mitochondrial cytochrome c peroxidase Ccp1 of Saccharomyces cerevisiae is involved in conveying an oxidative stress signal to the transcription factor Pos9 (Skn7). Molecular and General Genetics, 262.
151. Jamieson,D.J. (1998) Oxidative stress responses of the yeast Saccharomyces cerevisiae. Yeast, 14.
152. Morano,K.A., Grant,C.M. and Moye-Rowley,W.S. (2012) The response to heat shock and oxidative stress in saccharomyces cerevisiae. Genetics, 190.
153. Deniz,Ö., Flores,O., Aldea,M., Soler-López,M. and Orozco,M. (2016) Nucleosome architecture throughout the cell cycle. Sci Rep, 6, 19729.
154. Schalbetter,S.A., Goloborodko,A., Fudenberg,G., Belton,J.M., Miles,C., Yu,M., Dekker,J., Mirny,L. and Baxter,J. (2017) SMC complexes differentially compact mitotic chromosomes according to genomic context. Nat Cell Biol, 19.
155. Prince,S.J.D. (2023) Understanding Deep Learning MIT Press.
156. Breiman,L. (2001) Random forests. Mach Learn, 45.
157. Bishop,C.M. (2004) Pattern Recognition and Machine Learning Chris Bishop.
158. McCulloch,W.S. and Pitts,W. (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys, 5.

Objectives

During this thesis I have used computational approaches to study the mechanisms of gene regulation through the study of key DNA-protein bindings and chromatin structure definitions. I intend to better understand the interplay between DNA sequence, physical and mechanical properties, transcription, and chromatin structure, in order to extend the current knowledge of genome organization and expression processes. The main objective of this thesis is to study the gene expression mechanisms and the role that chromatin plays at different resolutions both at the two-dimensional and three-dimensional levels. For this purpose, the following precise objectives are proposed and grouped into five sections:

- To characterize Transcription Factor (TF)-DNA binding affinities and assess mechanisms for protein-DNA recognition.

- To build a machine learning model to predict *in vitro* TF binding preferences integrating sequence-dependent physical properties and extrapolating the predictions to detect binding sites in *in vivo* cases.

- To combine different computational tools for the analysis of nucleosome positioning while deciphering organizational patterns and incorporating the results into a genomic context.

- To develop an algorithm that predicts nucleosome architectures, such as nucleosome depleted regions, using a neural network and statistical positioning techniques.

- To study the conformational changes that chromatin undergoes at the 3D level under one of the most common DNA lesions, oxidative stress, leveraging a combination of experimental techniques, including MNase-seq, ChIP-seq, Hi-C, micro-C and micro-C XL.

- To study and calibrate the properties of a set of hybrid triplexes and build an empirical-based stability predictor to scan candidate triplex forming oligonucleotides (TFOs) and their role in gene expression mechanisms along the human genome.

- To enhance the understanding of the sequence-dependent properties of double-stranded RNA (dsRNA) conformation and flexibility through the use of molecular dynamics simulations and compare them with DNA properties to develop a mesoscopic coarse-grained model capable of reproducing long dsRNA conformational ensembles.

Chapter 1. DNA-protein binding

Transcription Factors (TFs) are proteins that bind specific DNA motifs and either repress or activate genes by modulating the recruitment of RNA polymerases: the main machinery for gene transcription. As explained in the Introduction, TFs can interact with and recognize a DNA sequence through different mechanisms that are divided into direct (hydrogen-bond interactions) and indirect (shape-related effects) readouts. The structural modifications required for a fruitful DNA-TF interaction can be explained by means of two models based on the extend of the conformational changes required to form the complex (1): the conformational selection and the induced fit. Previous work from our group (1) showed that DNA follows a double mechanism to undertake conformational changes in its structure: some initial moves along the easiest deformation modes to approach the bioactive conformation (following the conformational selection model) are then followed by final adjustments which require localized rearrangements at the base pair step and backbone level (induced fit paradigm). Altogether, previous works suggest that TF-binding could be explained by considering the gain in specific TF-DNA contacts, most likely modulated by hydrogen bonds across the grooves, and the physical and deformability properties of DNA that control the accessibility to the "bioactive conformation" and that can be defined by sequence-dependent helical parameters (equilibrium values and stiffness constants) derived from atomistic MD simulations (2–4).

In this first chapter we discuss the performance of a machine learning (ML) based method to reproduce *in vitro* and *in vivo* Transcription Factor Binding Sites (TFBS). Sequence-based and structural descriptors are used as inputs for a Random Forest (RF) Regressor trained to predict binding affinities obtained from a variety of experimental techniques such as high-throughput SELEX (HT-SELEX) or binding microarrays (PBM) (5–13). To test the predictor out of *in vitro* conditions, the trained model was used to reproduce *in vivo* data collected by Chromatin immunoprecipitation sequencing (ChIP-seq) together with in house collected data from nucleosome positioning for the same cellular model (yeast). Our results shed light into the myriad of factors that can help disentangle the rules governing TFs binding affinity and the hidden DNA code regulating genome

organization. The final methods reproduce quantitative HTSELEX and PBM experimental results with an astonishing accuracy. Ideas to enrich the model to approach to *in vivo* conditions are discussed.

1.1. Experimental Techniques

1.1.1. *In vitro* preferences

1.1.1.1. PBM

Protein binding microarrays encompass high-throughput sequencing techniques that measure the *in vitro* affinities between a protein and various DNA sequences. The arrays measure the preferences of a protein towards thousands of sequences at once, producing a clear overview of binding specificities. Universal PBMs (uPBMs), constitute an example of these techniques which utilize custom-designed microarrays containing all 10-mer sequence variants (11). This methodology starts by double-stranding a synthetic DNA oligonucleotide array, binding a protein directly to the microarray and labeling the protein-bound arrays with fluorophore-conjugated antibodies. Some variations involve using custom-designed genomic context protein binding microarrays (gcPBMs), where the sequences of interest are placed within genomic flanking sequences (12) avoiding detecting sequences in unrealistic contexts.

1.1.1.2. HT-SELEX

The systematic evolution of ligands by exponential enrichment (SELEX) provides a powerful technique to determine the *in vitro* binding preferences of proteins (8). Specifically, this method is based on the enrichment of small populations of bound DNAs from a random pool of sequences by PCR amplification. The first protocol of this technique was then enhanced by high-throughput sequencing (HT-SELEX) to generate a robust version of the previous technique (9, 10). Firstly, a random set of DNA sequences of typically 16-24bps, is mixed with the protein of interest and flanked by primers to allow the posterior PCR amplification. Several rounds of selection are done with the oligonucleotides being sequenced after each round to determine the sequences that are preferentially bound.

1.1.2. *In vivo* preferences

1.1.2.1. ChIP-seq

The *in vivo* binding preferences of proteins *in vivo* are generally measured with Chromatin immunoprecipitation followed by sequencing approaches (or variants such as ChIP-exo) (5, 6, 14). In order to find the relevant binding sites, an antibody selects the region targeted by the target protein before being immunoprecipitated and massively parallel sequenced. As a first step, the protocol starts with the cross-linking of the protein and DNA with formaldehyde. Secondly, the DNA is broken to obtain short fragments that are then immunoprecipitated with the selected antibody. Lastly, the cross-linking of the protein and DNA is reversed, and the DNA gets purified and sequenced (fragment lengths typically range between 150 and 300 bps). This experimental technique has limited accuracy in terms of the location of the specific binding site (typically less than 10 bp long) and tends to present some biases such as the specificity of the antibody or the uneven fragmentation in open versus closed chromatin (15). It is thus always important to include a control sample to account for these biases. This can include an input DNA before the immunoprecipitation, samples immunoprecipitated without antibodies or a sample immunoprecipitated using an antibody against a protein that does not bind to DNA and is not involved in chromatin modifications.

1.2. Theoretical Techniques

In addition to experimental techniques, the protein binding preferences to given DNA sequences can be theoretically studied through computational methods which tend to be more cost and time effective. Some of the features that aid computational models account for DNA sequence encodings, sequence properties modeled as rigid base pairs, deformation energy or the electrostatics defined by the major and minor grooves.

1.2.1. Positional Weight Matrices (PWM)

Binding affinities have been described for many TFs in different organisms and summarized in several databases such as TRANSFAC (16) and JASPAR (17). Classical theoretical approaches to predict binding affinities use these databases to generate positional weight

matrices. Each matrix contains the frequency of every nucleotide in every position where the TF is bound and can be expressed as a motif logo. The original methods regarded the preferences for each position as independent events, whereas later models incorporated sequence interdependence to compute binding motifs (18, 19). The information provided by the PWMs can then be used to estimate the binding affinity of a particular TF to any given DNA sequence by adding the corresponding scores for the different nucleotide values.

1.2.2. Machine Learning and Artificial Intelligence

PWMs present well-known limitations (20, 21), fueling the need for developing better predictors. Given the large amounts of binding data and myriad of descriptors, ML models were posed as a potential solution to the problem, but methods available to date are far from the desired accuracy. In this thesis we have investigated the capability of a Random Forest predictor to model various experimental techniques.

1.2.2.1. Random Forests

Random Forests (RFs) are a ML algorithm based essentially on a randomized ensemble of decision trees (22). In order to understand how RFs work, it is essential to learn some of the characteristics they share with tree-based methods which make them a good fit for many tasks:

- Decision trees can model complex relations between inputs and outputs without any a priori assumption.
- Decision trees intrinsically implement feature selection and are robust to outliers.
- Decision trees are easily interpretable.

Learning a decision tree involves optimizing the node splits with the descriptors that will maximize the information gained from a specific split. In the learning process where we try to reproduce the observed data, the goal is to optimize a cost function that is determined by the true labels and the predicted ones (more details can be found in the Introduction). The randomized ensemble of trees allows the algorithm to be more robust and less prone to overfitting the data while still allowing a straight-forward interpretation of the model.

In the first publication of this thesis, we demonstrate how TF-DNA binding can be predicted based on DNA sequence-dependence

and physical properties derived from MD simulations. A Random Forest Regressor has been built to model binding affinities for a large pool of TFs. Using a set of pre-processed data and theoretically defined descriptors, we are able to investigate the binding mechanism preferences, considering both shape and sequence readouts. The chosen regressor is capable of accurately reproducing *in vitro* binding preferences of many TFs, showing that the combination of the chosen sequence-based and structural descriptors improves performance over previously published methods. After training on 80% of the data, we calculated the determination coefficient (R^2) between our predictions and the experimental values on the remaining 20% of the data. We found that our model could reproduce gcPBM data with astonishing quality, obtaining an average R^2 of 0.93 ± 0.02. When reproducing uPBM data we predicted affinities with an average determination coefficient R^2 = 0.69 ± 0.17. We then applied our ML predictor to two datasets from HT-SELEX to obtain an average R^2 of 0.70 ± 0.14.

When comparing our predictor with previously published methods (21, 23, 24) DNAffinity outperformed all of them for the different experimental techniques. The developed architecture trained to reproduce *in vitro* binding properties, was extended to reproduce *in cellulo* TFs preferences through the introduction of extrinsic factors from nucleosome positioning data. Our method predicted many potential *in vivo* binding sites where no experimental evidence of binding exists, obtaining an overall True Positive (TP) rate of 94% and a False Positive (FP) rate of 7%. More importantly, most FP cases were explained by nucleosome occupancy.

Overall, this first publication sheds some light on the extrinsic and intrinsic factors that define the first layer of gene regulatory networks and illustrate the power of ML approaches to decipher crucial elements of chromatin structure.

Publication:

Sandro Barissi*, Alba Sala*, Miłosz Wieczór, Federica Battistini, Modesto Orozco, DNAffinity: a machine-learning approach to predict DNA binding affinities of transcription factors, Nucleic Acids Research, Volume 50, Issue 16, 9 September 2022, Pages 9105–9114, https://doi.org/10.1093/nar/gkac708

*Equally contributing authors
Supplementary material for this article can be found in the Annex.

References

1. Battistini,F., Hospital,A., Buitrago,D., Gallego,D., Dans,P.D., Gelpí,J.L. and Orozco,M. (2019) How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition. J Mol Biol, 431.
2. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A., et al. (2015) Parmbsc1: A refined force field for DNA simulations. Nat Methods, 13.
3. Lankaš,F., Šponer,J., Langowski,J. and Cheatham,T.E. (2003) DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. Biophys J, 85.
4. Pérez,A., Lankas,F., Luque,F.J. and Orozco,M. (2008) Towards a molecular dynamics consensus view of B-DNA flexibility. Nucleic Acids Res, 36.
5. Mundade,R., Ozer,H.G., Wei,H., Prabhu,L. and Lu,T. (2014) Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle, 13.
6. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell, 147.
7. Mahony,S. and Pugh,B.F. (2015) Protein-DNA binding in high-resolution. Crit Rev Biochem Mol Biol, 50.
8. Smaczniak,C., Angenent,G.C. and Kaufmann,K. (2017) SELEX-seq: A method to determine DNA binding specificities of plant transcription factors. In Methods in Molecular Biology.Vol. 1629.
9. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M.J., et al. (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res, 20.
10. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G., et al. (2013) DNA-binding specificities of human transcription factors. Cell, 152.
11. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol, 24.
12. Gordân,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. Cell Rep, 3.

13. Nutiu,R., Friedman,R.C., Luo,S., Khrebtukova,I., Silva,D., Li,R., Zhang,L., Schroth,G.P. and Burge,C.B. (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat Biotechnol, 29.
14. Rhee,H.S. and Pugh,B.F. (2008) ChIP-exo: A Method to Identify Genomic Location of DNA-binding proteins at Near Single Nucleotide Accuracy. Curr Protoc Mol Biol, 141.
15. Park,P.J. (2009) ChIP-seq: Advantages and challenges of a maturing technology. Nat Rev Genet, 10.
16. Matys,V., Kel-Margoulis,O. V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K., et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res, 34.
17. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., Van Der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G., et al. (2018) JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res, 46.
18. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: A new way to display consensus sequences. Nucleic Acids Res, 18.
19. Benos,P. V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: How good an approximation is it? Nucleic Acids Res, 30.
20. Zhao,Y., Ruan,S., Pandey,M. and Stormo,G.D. (2012) Improved models for transcription factor binding site identification using nonindependent interactions. Genetics, 191.
21. Zhou,T., Shen,N., Yang,L., Abe,N., Horton,J., Mann,R.S., Bussemaker,H.J., Gordân,R. and Rohs,R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. Proc Natl Acad Sci U S A, 112.
22. Breiman,L. (2001) Random forests. Mach Learn, 45.
23. Li,J., Sagendorf,J.M., Chiu,T.P., Pasi,M., Perez,A. and Rohs,R. (2017) Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. Nucleic Acids Res, 45.
24. Asif,M. and Orenstein,Y. (2020) DeepSELEX: Inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs. Bioinformatics, 36.

Published online 26 August 2022

Nucleic Acids Research, 2022, Vol. 50, No. 16 ***9105–9114***
https://doi.org/10.1093/nar/gkac708

DNAffinity: a machine-learning approach to predict DNA binding affinities of transcription factors

Sandro Barissi[1,†], Alba Sala[1,†], Miłosz Wieczór[1,2], Federica Battistini[1,3,*] and Modesto Orozco[1,3,*]

[1]Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Baldiri Reixac 10–12, 08028 Barcelona, Spain, [2]Department of Physical Chemistry. Gdansk University of Technology, 80-233 Gdańsk, Poland and [3]Department of Biochemistry and Molecular Biology. University of Barcelona, 08028 Barcelona, Spain

Received January 20, 2022; Revised July 21, 2022; Editorial Decision July 27, 2022; Accepted August 08, 2022

ABSTRACT

We present a physics-based machine learning approach to predict *in vitro* transcription factor binding affinities from structural and mechanical DNA properties directly derived from atomistic molecular dynamics simulations. The method is able to predict affinities obtained with techniques as different as uPBM, gcPBM and HT-SELEX with an excellent performance, much better than existing algorithms. Due to its nature, the method can be extended to epigenetic variants, mismatches, mutations, or any non-coding nucleobases. When complemented with chromatin structure information, our *in vitro* trained method provides also good estimates of *in vivo* binding sites in yeast.

INTRODUCTION

Proteins are the main regulators of gene expression as they can directly or indirectly inactivate, activate, or enhance the transcription of DNA. Central elements in this regulatory system are transcription factors (TFs): modular proteins that recognize sequences of DNA (typically 6–20 bp long), helping to recruit RNA polymerases that trigger the subsequent transcription of a nearby gene (1,2). The binding of TFs during normal cell life is difficult to predict (3,4) as it is modulated by a myriad of effects, such as the presence of nucleosomes (which in general hinders TF binding) (5,6), or the formation of clusters that foster cooperativity, general chromatin compaction, or even phase separation (7–9). However, a key requirement for *in vivo* binding is a good binding to the targeted naked DNA.

The recognition of naked DNA by transcription factors is complex and does not follow a common code or a single mechanism (1). Based on the degree of structural distortion that protein induces in DNA, we can distinguish three binding paradigms (10): (i) Fischer's lock and key theory (no distortion on DNA from its canonical B-form); (ii) conformational selection (small to medium deformation that aligns with intrinsic deformation patterns of DNA) and (iii) induced fit (large deformations of DNA that are unlikely to happen in the absence of the protein). Based on the type of contacts used for DNA recognition, we can distinguish between TFs interacting mainly with the DNA backbone, those establishing hydrogen bond interactions with the nucleobases in either major or minor groove, and finally those disrupting the duplex geometry to generate stacking contacts. In a similar vein, Rohs and coworkers (1,11,12) defined two main mechanisms for 'TF–DNA reading': the direct readout, related to the formation of specific interactions of the TF with the nucleobases (typically by means of hydrogen bonds), and the indirect readout, related to the sequence-dependent shape of DNA. Recently, our group extended these ideas by also considering the sequence-dependent energy cost for changing the DNA conformation from the unbound to the bound state (10), a concept that introduces sequence-dependent flexibility as a determinant of sequence-dependent binding.

Most data on the sequence preferences of TFs, i.e. the TFBSs (transcription factor binding sites), rely on high-throughput experimental techniques such as high-throughput SELEX experiments (HT-SELEX; (13–15)), protein binding microarrays (PBM) using either synthetic sequences (uPBM; (16)) or sequences from the genomic context (gcPBM; (17)), or fluidic engines such as HITS-FLIP (18). The final output of these *in vitro* techniques is a list of DNA sequences of different lengths (depending on the experimental method, from 10 in HT-SELEX to 36 in PBM) with an associated estimate of their binding affinity for a given TF. The *in vivo* preferences are typically derived from ChIP-seq approaches (or variants such as ChIP-exo) (19–21), where the chromatin is immunoprecipitated

*To whom correspondence should be addressed. Tel: +34 934 037 156; Email: federica.battistini@irbbarcelona.org
Correspondence may also be addressed to Modesto Orozco. Tel: +34 934 037 156; Email: modesto.orozco@irbbarcelona.org
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

9106 Nucleic Acids Research, 2022, Vol. 50, No. 16

by TF-specific antibody and the retained chromatin is then sequenced. The technique is noisy, as it uses cell populations, and the resolution is poor (200 bp ± 300 in ChIP-seq, and down to 50 bp in the case of ChIP-exo) (20,21), but it provides direct evidence of the active TFBS under physiological conditions.

Traditional theoretical approaches to predicting TFBS are based on positional weight matrices (PWM) and their associated logos (22,23). While original versions assume the independence of nucleotide preferences at each position, the newest PWM models can capture some of the sequence interdependence (23,24), increasing their reliability. However, the limitations of PWM models are well known (25,26), which has fuelled the development of alternative methods relying on an extended set of descriptive parameters and last-generation learning models. We can broadly classify these new predictive models based on whether they aim to predict *in vivo* or *in vitro* TFBS. Those for *in vitro* TFBS prediction use the nucleotide sequences and a variety of DNA shape descriptors as input parameters (12,26,27). Methods for *in vivo* TFBS prediction are typically focused on human genomes, and complement the descriptors of naked DNA (typically sequences) with ones related to chromatin structure and dynamics (RNAseq, DNase, conservation profiles etc.). Experimental data have been widely used to train a variety of machine learning and deep learning methods (28–38) to predict TFBS.

We present here a physics-based ML approach to TF binding affinity prediction *in vitro* that uses the physical properties of DNA directly derived from molecular dynamics (MD) simulations (39–41). These properties consider both the equilibrium geometry and the flexibility, as defined by Olson (42), to describe DNA conformation with sequence-dependence at a base-pair step (bps) resolution. The importance of these conformational properties for the study and prediction of DNA behavior and preferences has been proved in numerous studies (42–48).

Our method uses these DNA physical properties to train a random forest regressor to reproduce uPBM, gcPBM and HT-SELEX data for a large variety of protein families, and yields results that outperform currently available methods. We then use this *in vitro* trained method to explore *in vivo* binding sites of the transcription factor CBF1 in yeast, one of the very few TFs for which we have experimentally available data for both *in vitro* (PB-exo) and *in vivo* (CHIP-exo) binding. When *in vitro* predictions were combined with chromatin structure as determined by nucleosome positioning maps, our method has shown state-of-the-art predictive power in identifying *in vivo* TPBSs.

MATERIALS AND METHODS

Datasets for training and testing

There is a variety of data on *in vitro* TF binding preferences, covering a variety of proteins and measuring techniques. HT-SELEX data were taken from 2 different studies (European Nucleotide Archive ENA PRJEB14744 and PRJEB29730) and processed using the associated package (https://bioconductor.org/packages/release/bioc/html/SELEX.html). The combined databases contain information on 600 TFs from >30 different protein families. TFs with no k-mer (10 bases) with at least 100 counts in 0th order SELEX cycle were removed from the training set as they have no clear binding motif (as discussed in (13)). As reported in the literature (13), training and testing were done using data from the penultimate SELEX cycle provided in the databases. uPBM-data were taken from the DREAM5 challenge containing information on binding preferences of 35-mer oligos for 66 mouse TFs (49). The 50 oligos with the highest affinity to a given TF were used to define a PWM (50), which was then used to align the sequences and derive the most probable binding site (a shorter k-mer, typically 12-mer). Finally, for gcPBM data, sequences already aligned around the putative binding site placed at the center of 36mer genomic were used; more specifically, sequences for the TF dimers Mad1/Max ('Mad'), Max/Max ('Max'), c-Myc/Max ('Myc') and CFB1 (Gene Expression Omnibus accession numbers GSE59845 and GSE44604 respectively) (17,50–52). Possible multiple binding sites were removed using a previously published protocol (26).

In vivo binding data for CBF1 were taken from ChIP-exo maps of *Saccharomyces cerevisiae* (GSE44604 and GSE147927) (17,52,53) and were used to perform a proof of concept of the ability of the *in vitro* CFB1 binding predictor to detect *in vivo* binding sites. Data on chromatin used to discuss the differences between *in vitro* and *in vivo* binding profiles were taken from previously published nucleosome maps in the same cellular model (5,54).

All ID and references to the datasets used for training and testing are summarized in Supplementary Table S1.

Feature classes

For the training of the machine learning (ML) algorithm, different classes of features were used:

- ***Sequence composition: presence***. For each *k*-mer, a vector of counts for the 256 possible tetramers that show up in a given k-mer was calculated, using a sliding window of length 4 and simply adding occurrences. For instance, the 6-mer 'AAAAAT' would have two counts for the tetramer 'AAAA', one for 'AAAT' and none for the remaining 254 tetramers.
- ***Indirect readout: base pair parameters***. The base pair parameters (equilibrium values and the diagonal components of the stiffness constant matrix, called AVG and DIAG respectively, (37–39)) for each individual base pair step movement (translational: shift, slide and rise, and rotational: tilt, roll and twist) were considered. The values were retrieved from a dataset that covers all the unique base pair steps in all the possible tetranucleotide environments from microsecond-long molecular dynamics simulations (10,41). All data used to characterize tetramers are available in our BigNASim database (55).
- ***Direct readout***. The electrostatic patterns of each base (hbond acceptor/donor or hydrophobic) were considered using the scheme below (see Supplementary Figure S1) (12). Our method assigns integers (–1, 0, +1) to acceptors, hydrophobic sites and donors respectively, and for each

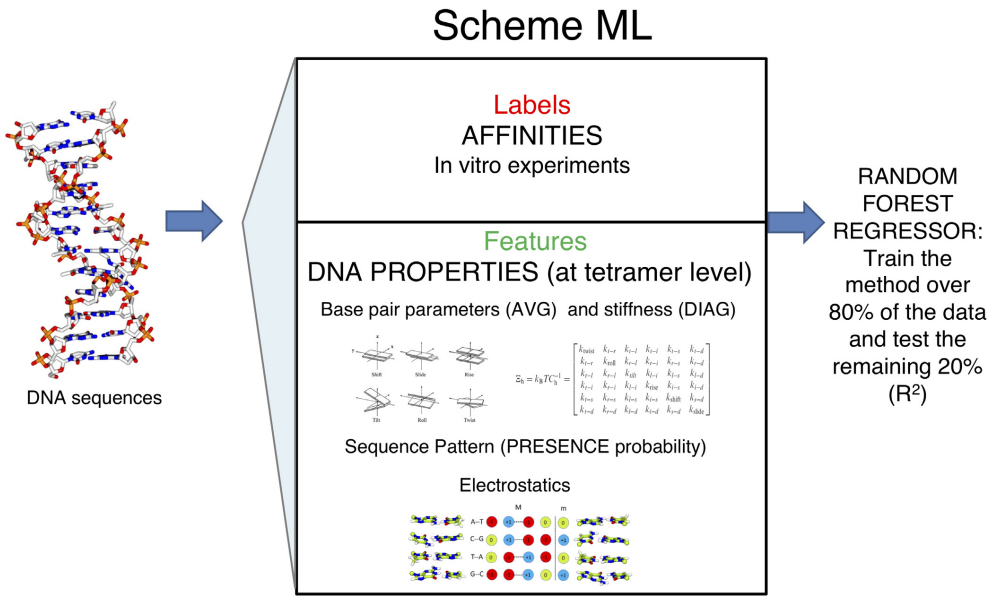


Figure 1. General scheme of the ML training strategy.

overlapping tetramer along the DNA sequence sums the relative values along columns. For the minor groove, since the flanking sites are always –1, they are omitted and we keep only the middle value. In total, for each tetramer the electrostatics are explained by five values, for example, for 'AACT' [–1, +1, –1, 0, 0] + [–1, +1, –1, 0, 0] +[0, +1, –1, –1, +1] + [0, –1, +1, –1, 0] = [–2, 2, –2, –2, 1].

All the references to the features used are summarized in Supplementary Table S1.

Machine learning training

Features described above were used as descriptors, and they were attributed to each overlapping tetramer in the k-mer sequence under study (see Supplementary Figure S2). Experimental affinities in databases were used as labels for a Random Forest regressor (56). A train/test ratio of 80/20 on the experimental data (HT-SELEX, uPBM and gcPBM) (see Scheme) was used to reduce overtraining artifacts. The R^2 regression scoring function, Pearson correlation (r) and MSE (mean squared error; see Supplementary Methods for details) between the predicted and experimental affinity (56) was used to validate the accuracy of the model (Figure 1). For the training process, we randomly selected 80% of the data, and we performed bootstrap to avoid biases. This method allowed us to perform the training multiple times even if the chosen training and testing sets could have contained repeated entries. For validation of the method, we also tested the choice of the training set using cross-validation, which performs the simulation K times dividing the data into K partitions and using each time one different partition as a test set. Changing the algorithm to K-fold cross-validation ($K = 10$; 90/10 randomly chosen), we obtained very similar results as previously (see Supplementary Table S2).

To improve the accuracy and efficiency of the training process and to avoid training artifacts, we applied several data pre-processing steps:

Undersampling

The uPBM experimental datasets are usually very noisy with an overpopulation of k-mers with a low affinity value. To obtain a more balanced set, we applied different undersampling approaches for each dataset type (details and an example in Supplementary Methods and and an example in Supplementary Figure S3).

Weighting

For the uPBM dataset, because of the lack of high-affinity sequences, we assigned uniform higher weights (10) to sequences matching the PWM data for each TF being considered. We also calculated the importance of each class feature in the regressor (56).

The purpose of undersampling and weighting (or oversampling of high affinity values) is to remove noisy data by stratifying the affinity profile and picking only a few samples in each stratum.

HT-SELEX data quality assessment

For HT-SELEX data, we considered the target affinity values as those reported in the cycles. In principle, binding affinities of the different k-mers should grow exponentially at each cycling step until saturation, and any deviation from this behaviour signals inconsistency of the data.

9108 Nucleic Acids Research, 2022, Vol. 50, No. 16

Thus, to guarantee the quality of the experimental data, we performed a quality check with a support vector machine (SVM) discriminant using the correlation between selected counts across different cycles as descriptors (see Supplementary Methods).

Genomic testing

For the *in vivo* validation on the yeast genome, we generated discrete maps using the relative score obtained from the ML training for each possible TFBS and comparing the predictions with *in vitro* (PB-exo) (17) and *in vivo* maps (ChIP-exo) (17,52,53). We define:

- True positive (TP) when the experimental peaks (both PB-exo and ChIP-exo) overlap with our prediction.
- False positive (FP) when TFBSs predicted by our model do not correspond to any experimental peak. We also applied this category if our prediction corresponds to just one of the experimental dataset (either PB-exo or ChIP-exo)
- Nucleosome occupied locations (Nuc) when comparing our prediction with nucleosome maps one TFBS predicted by our model overlaps with a nucleosome.
- False negative (FN) when experimental TFBS peaks have not been predicted by our model.
- True negative (FN) when both experiments and predictive model agree that there is not a TFBS.

The program and the full database of feature parameters are available in the GitHub repository: https://github.com/Jalbiti/DNAffinity.

RESULTS AND DISCUSSION

For each experimental dataset (gcPBM, uPBM and HT-SELEX), we trained our machine learning regressor to predict experimental binding affinities using three classes of features informative of the three DNA–protein recognition modes: sequence, direct and indirect readout. After training on 80% of the data, we calculated the determination coefficient (R^2) between our predictions and the experimental values (see Materials and Methods) on the remaining 20% of the data. With that, we found that our model is able to reproduce gcPBM data with astonishing quality, as shown in an average R^2 of 0.93 ± 0.02 (see Figure 2).

Using the uPBM data as reference (see Materials and Methods), we could predict affinities with an average determination coefficient $R^2 = 0.69 \pm 0.17$ (Figure 3).

We then applied our ML predictor to two datasets from HT-SELEX (see Materials and Methods and Supplementary Table S1): the first using the results based on 5-cycles of HT-SELEX experiments, and the second on 7-cycles. In the first case, we achieved an R^2 of 0.63 ± 0.19, and in the second 0.71 ± 0.21, which yielded a total average of 0.66 ± 0.19. Supplementary Figures S4–S7 detail the results for each HT-SELEX experiment and each transcription factor. In a further step, we used SVM to remove those cases displaying inconsistency between the enrichments at different HT-SELEX cycles (see Materials and Methods and Supplementary Methods) as they are suspicious cases whose inclusion can bias the training and testing. The improvement obtained after SVM-filtering of data is clear, as seen from the average R^2 of 0.70 ± 0.14 and the dramatic reduction of cases with low R^2 (see Figure 4).

In summary, DNAffinity is able to accurately predict relative binding affinities of transcription factors with a common set of descriptors, irrespectively of the source of experimental approach used to determine the binding (see Figure 4). The logo plots of the most favorable TFBSs, calculated using the top 100 predicted sequences, for each TF studied are presented in Supplementary Table S3.

DNAffinity's performance compared to existing ML predictors

We compared our predictor with a larger variety of previously published methods that combine different learning approaches and include DNA sequence and DNA shape features (27,35,36,49) (Figure 5 and Supplementary Figures S8 and S9). In all cases comparisons are done using the same number of cases (TFs) and the same dataset as used by the original developers of the methods.

For gcPBM our performance (Figure 2) is nearly identical to that obtained with the shape-augmented ML predictor (32), but this outstanding performance should be taken with caution as there are only three transcription factors for which gcPBM information is available. Here, the probes are very well aligned based on the central TFBS and the sequences used do not show high variability, making the problem relatively easy to solve.

A more reliable and challenging benchmark can be obtained using uPBM and HT-SELEX data as: i) there are more datasets available and ii) there is a larger variety of trained models to benchmark against.

In order to evaluate the performance of DNAffinity, we compared its predictive power on 66 uPBM datasets and compared our R^2 with the ones obtained using the same dataset by: CRPTS/CRPT (a hybrid convolutional recurrent neural network (CNN/RNN) architecture that combines DNA sequence and DNA shape features) (27); Deepbind (a CNN model primarly based on DNA sequences) (36); two kernel-based methods (spectrum + shape kernel, di-mismatch + shape kernel) (57); a deep learning-based DLBSS (37); and a shape-based ML regressor DNAShapeR (33). The data used for the comparison were previously reported (27). In Figure 5A and Supplementary Figure S8, we show the results of the comparison. The improvement with respect to all other predictive algorithms is very clear: our algorithm has a stronger predictive power when compared to methods based on shape and neural network, probably thanks to the properties considered and the pre-processing of the 36mers.

For HT-SELEX data, we compared our results with a shape-based ML regressor, DNAShapeR, and DeepSELEX (38) trained and tested on the same TFs (Figure 5B). Running DNAShapeR, we used their refined set of sequences (M-word, https://rohslab.usc.edu/MSB2017/) and their latest parameters. The comparison shows how our method could better predict the TFBS affinities, and that the results we get for those selected protein is in the range of our prediction using all the cases (DNAffinity all in Figure 5B) available for HT-SELEX (see Supplementary Table S1). We tested the rigorousness of our method calculat-

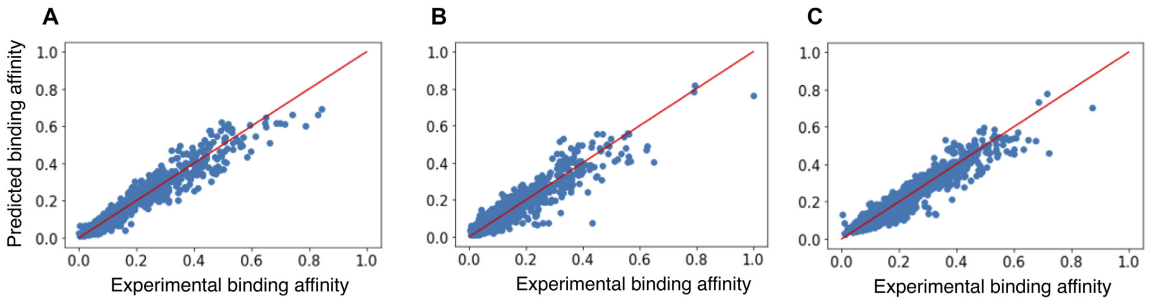


Figure 2. Correlation between the predicted and experimental TF binding affinity for the three cases under study: (**A**) Mad1/Max ('Mad'), (**B**) Max/Max ('Max') and (**C**) c-Myc/Max ('Myc'). Correlation with experimental affinities for the transcription factors: MAD1 (left panel): $R^2 = 0.951$, MYC (central panel): $R^2 = 0.905$, MAX (right panel): $R^2 = 0.922$.

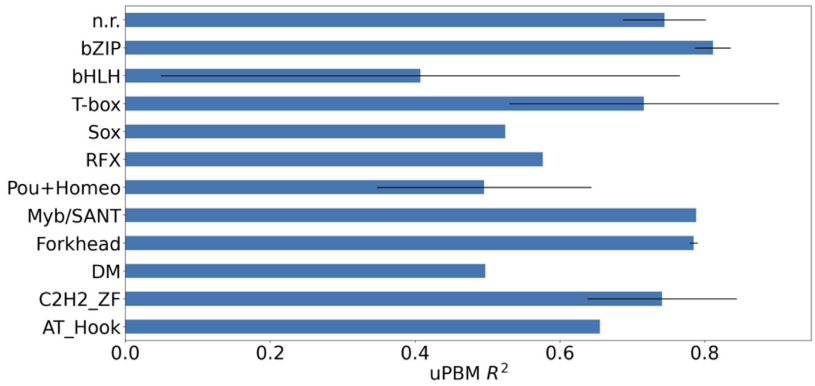


Figure 3. Determination coefficient (R^2) between predicted and experimental data for different protein families using uPBM data.

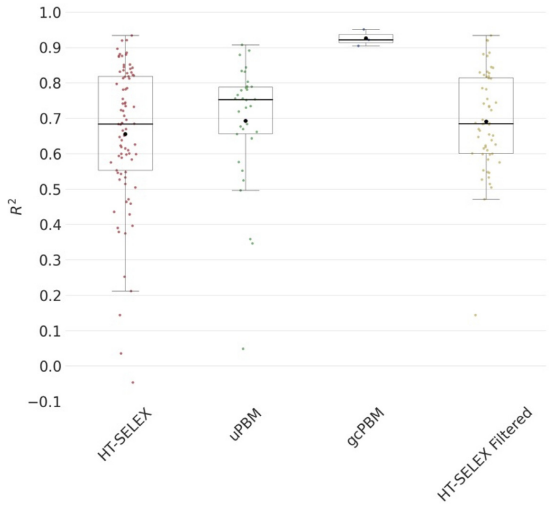


Figure 4. Determination coefficient (R^2) between predicted and experimental affinities for different transcription factors (black circle marks the mean, black bar marks the median value). Data used were: HT-SELEX on the entire dataset, uPBM, gcPBM and HT-SELEX on the filtered dataset (after applying our reliability filter that uses an SVM classifier on the raw data; see Materials and Methods and Supplementary Methods).

ing the MSE (see Supplementary Methods) and comparing it to the other algorithms that had the best (newest) and the worst performance compared to ours: CRPT (27) and DNAShapeR (33) for uPBM data and both DNAShapeR (33) and DeepSELEX (38) for HT-SELEX data (see Supplementary Table S4); confirming that also using this metric the results obtained by our algorithm are consistent.

Finally, many previously developed methods verified the transferability of their predictive algorithm by first training with one dataset (HT-SELEX) and subsequently testing on another dataset (uPBM). These two experimental techniques differ in the variety and number of sequences that can be studied, and on the length of the different probes: HT-SELEX considers a large amount of different short sequences with one possible binding site, while uPBM has fewer and longer probes with multiple candidate binding sites, and mainly low affinity. Consequently we also compared the ability of our method to inter-cross between datasets. The results obtained (Figure 5C) show that when comparing our results to previously published methods, including neural network and deep learning algorithms (data taken from (38)), DNAffinity outperforms all of them. We obtained an average Pearson correlation of 0.47 versus 0.41 (DeepSELEX) (38), 0.35 (DeepBind) (36), 0.36 (BEESEM)(58) and 0.20 (BindSpace) (59).

9110 Nucleic Acids Research, 2022, Vol. 50, No. 16

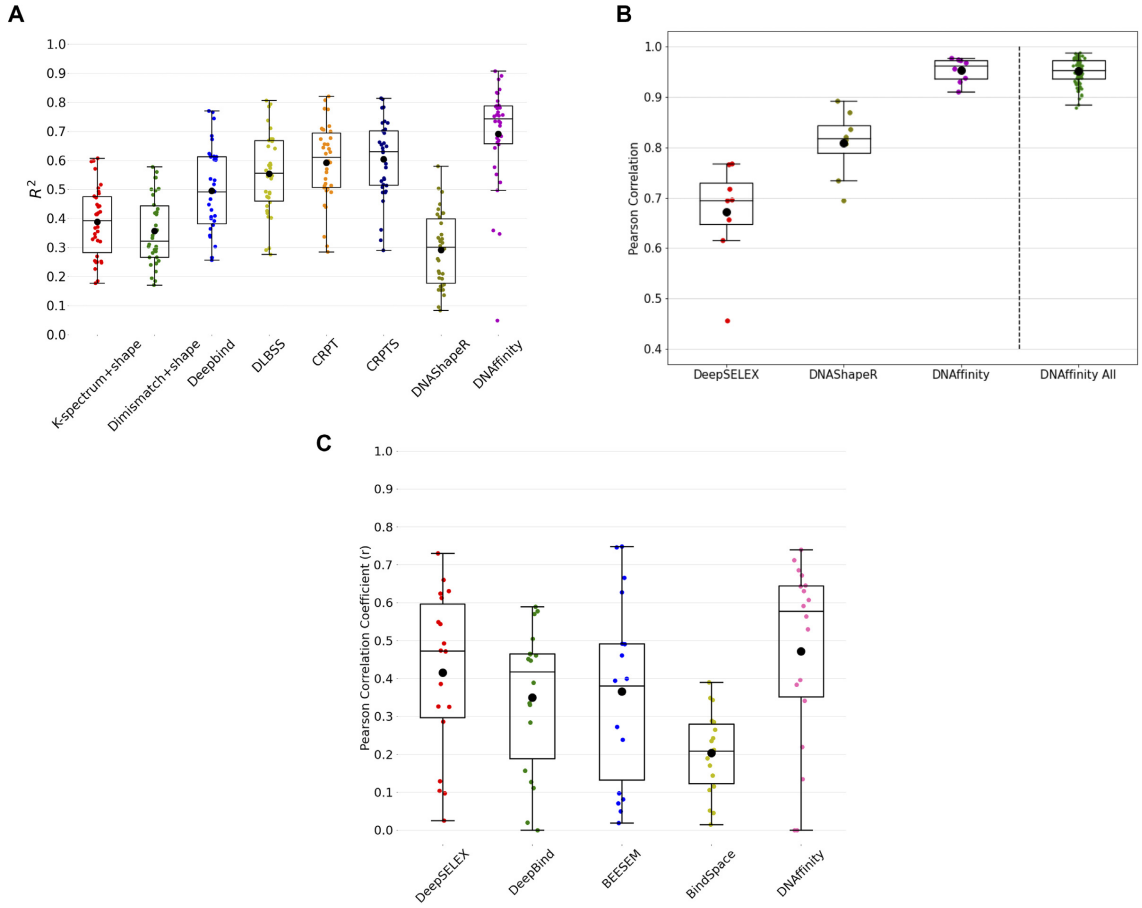


Figure 5. Comparison between our predictions and previously reported ones. (**A**) Determination coefficient (R^2) between predicted and experimental data for different protein families using uPBM data, data retrieved from (27) and DNAShapeR (33). (**B**) Pearson correlation using in common HT-SELEX data obtained in the literature by previously developed methods, DeepSELEX and DNAShapeR (33,38) and ours (DNAffinity) respectively, values obtained using all the HT-SELEX data available using our method (DNAffinity all). (**C**) Pearson correlation between predicted and experimental data for different protein families using HT-SELEX for training and testing on uPBM data, using our method (DNAffinity) and previously developed methods: DeepSELEX (38), DeepBind (36), BEESEM (58) and BindSpace (59). The data reported were taken from (38). Black circle: mean, black bar: median.

Importance of the features

Contrary to our original expectations, we found that the impact of the different features on the predictive power of the method depends dramatically on the type of experimental data used for training (Figure 6).

For gcPBM data, characterized by a very low sequence variability and a very well defined TFBS, the physical properties of the tetramers have a higher importance, probably because they can differentiate between otherwise very similar sequences.

For uPBM, we trimmed the sequence based on MEME suite result, so the variability of the sequence diminishes. For this reason, in part like in case of gcPBM, the motifs present a common pattern and shape seems to be the best class of feature capable of accentuating their differences (see the contribution of different feature classes in Supplementary Figures S10 and S11). However, because the sequence variability is broader than in gcPBM, also sequence and electrostatics features seem to gain importance.

On the contrary, for HT-SELEX, where a wide range of sequences is reported, the predictive power is equally divided across the three feature classes (Figure 6). This can be explained considering that the method explores a larger range of sequences, including these that are not physiologically accessible. Also, in this case we studied the importance of every class of feature for the prediction and detected that all of them contribute to the prediction (Supplementary Figures S10 and S11). Interestingly, the electrostatic descriptor gained importance when using HT-SELEX data compared to uPBM. The addition of this new 'direct-reading' feature to the prediction scheme introduces a new dimensionality in our method to discriminate

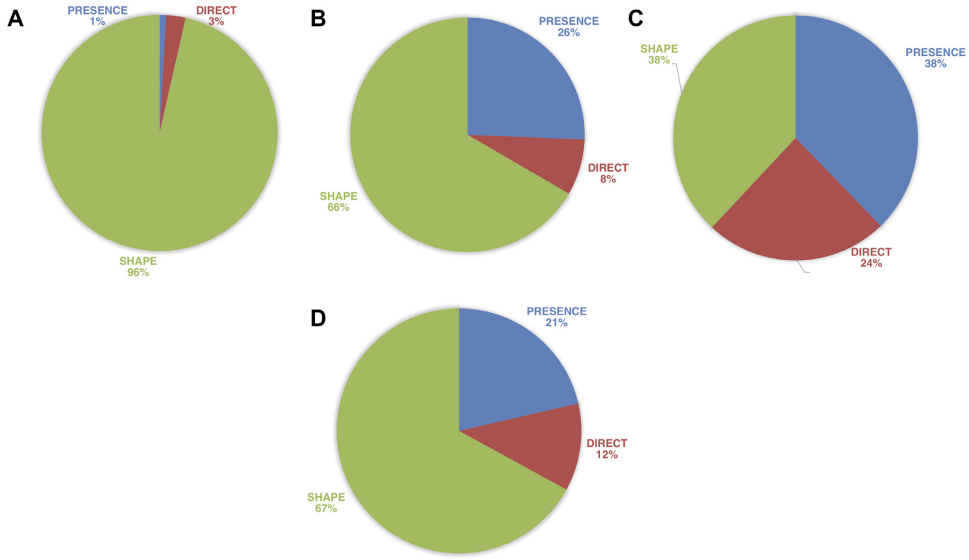


Figure 6. Relative importance (%) of each feature class in the prediction. (**A–C**) Results regarding prediction of gcPBM, uPBM and HT-SELEX data respectively. (**D**) Average of the relative importance (%) considering the prediction on all the datasets (A–C).

among largely variable sequences. The HT-SELEX dataset seems pushing the predictive power of the models to their limits.

Above all, by analyzing the effect of each feature on the different protein families it could be possible detect the ones that are mostly affected by indirect-readout (shape and force constants) or direct-readout (sequence and electrostatics) descriptors (see Supplementary Figure S11). Our results also raise concerns about attempts to train ML methods with a narrow set of sequences, and give us confidence that DNAffinity provides results of very similar quality when reference data come from uPBM or HT-SELEX with a common set of features.

In vivo testing

We finally applied our method to simultaneously predict *in vitro* and *in vivo* datasets describing the TFBS for the protein CBF1 (one case for which both *in vitro* and *in vivo* data are available, see Materials and Methods). After training our regressor using gcPBM data ($R^2 = 0.80$), we applied our model to predict PB-exo and ChIP-exo peaks along the yeast genome. We considered the consensus exo peaks, because being in common they are independent on the experimental technique/conditions. Each method (ChIP and PB) has some intrinsic noise due to non-specific or spurious interactions and that using consensus peaks ensures that each signal is genuine. To account for the impact of chromatin structure, we include accurate nucleosome maps collected for yeast in the G1 phase (5,54). Quite encouragingly, we were able to predict almost all consensus exo peaks, defined as locations where PB-exo and ChIP-exo signals coincide (Figure 7). Our true positive (TP) rate (TP/total number of exo experiments peaks) was over 94% (TP case example in Supplementary Figure S12A), meaning that only < 6% of the consensus exo peaks are not detected by our method (see Figure 7). Although these TPs entail only 14% of all predictions of our model, a vast majority of the theoretically false positive are at locations occupied by nucleosomes (see Figure 7 and example in Supplementary Figure S12B). As those chromatin sites would not have been accessible for the binding of a transcription factor (occupied by nucleosomes, Nuc in Figure 7B), they correspond to cases where intrinsic (*in vitro*) binding can be favourable, but chromatin structure precludes effective *in vivo* binding. Thus, when nucleosome maps are included as descriptors, the resulting false positive (FP) rate is just 11%. In fact, of the 146 'bona fide' FPs, 37 (FP2 in Figure 7) correspond to sequences that matches one exo signal (PB-exo or CHIP-exo) and 15 have evidence of activity based on polymerase maps (fourth column in the classification scheme Figure 7, FP case examples in Supplementary Figure S12C and D). It means that the real FP rate can be as low as 7% (FP1 and third column scheme in Figure 7). Due to the lack of simultaneous *in vivo* and *in vitro* binding data, it is difficult to generalize our conclusions; we consider here nucleosome occupancy as a proxy for chromatin structure, but there are many other means by which cells can hide regions that would otherwise be bound by transcription factors. Even though we will never be able to make a prediction taking into consideration all the possible variables to *in vivo* TF binding, we transferred the *in silico* prediction to *in vivo* conditions. We think it is important to determine how the intrinsic sequence-dependent binding properties *in vitro* are affected by chromatin accessibility. Our results may suggest that *in vivo* binding may be understood as in vitro binding corrected by high-resolution (nucleosome-scale) chromatin structure.

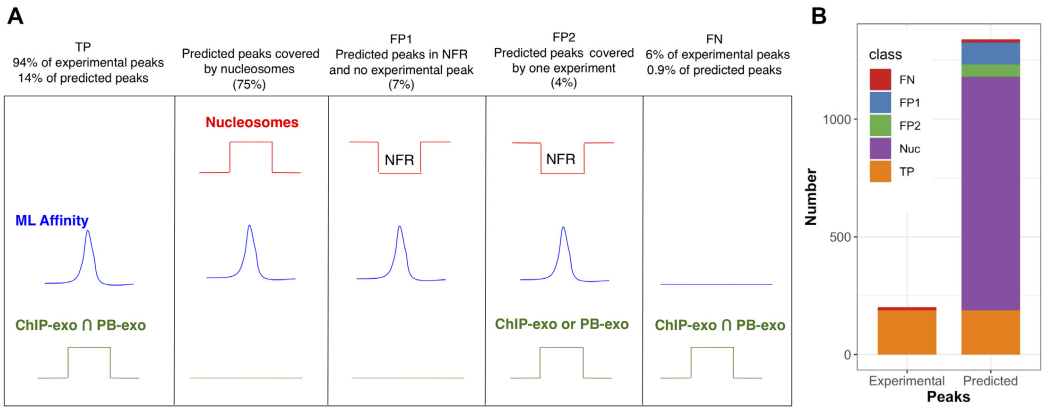


Figure 7. (**A**) Scheme representing the prediction scores along the yeast genome. (**B**) Statistics of our prediction along the yeast genome. On the right, distribution over all the predicted peaks: 188/1326 true positive (TP, orange); 13/1326 false negative (FN, red); 992/1326 in locations occupied by nucleosomes (in purple); the false positive (FP) cases are divided into FP1 that correspond to predicted peaks at nucleosome free region (NFR) that do not correspond to any experimental peak ((94/1326, in blue) and FP2 (52/1326, green) that correspond to sequences that either match one exo signal (PB-exo or CHIP-exo) or have evidence of activity based on polymerase maps. On the right, distribution over all the consensus experimental peaks (PB-exo and CHIP-exo): 188/201 true positive (TP, orange); 13/201 false negative (FN, red).

CONCLUSIONS

Prediction of transcription factor binding sites is the next grand challenges in genomic research. Development of efficient predictive algorithm requires solving a series of intrinsic problems: on the one hand, the concept of transcription factor binding site is not uniquely defined, as it deeply depends on the intrinsically noisy and low-resolution experimental technique used to detect it, making it impossible to create a universal predictor. On the other, transcription factors use a repertoire of mechanisms for selecting target DNA sequences, and the most informative parameters describing these mechanisms largely depend on the sequence variability explored by the experiment. The complexity of the problem increases even more if *in vitro* predictions are tried to be extrapolated to *in vivo* settings, where other factors besides intrinsic transcription factor affinity play a role.

Our predictive model (DNAffinity) is based on a simple machine learning algorithm trained on *ab initio* parameters derived from first-principle molecular dynamics simulations. One of the advantages of using theoretically derived descriptors is that they can be in principle obtained for any non-coding DNA, including epigenetic variants or lesions. Despite the '*ab initio*' nature of the descriptors and the simplicity of the training, the method provides excellent results, outperforming all available competitors when predicting *in vitro* transcription factor binding sites irrespective of the experiment used for validation. Very encouragingly, DNAffinity trained on *in vitro* data showed an excellent ability to detect the binding sites of the same transcription factor *in vivo*. Thus, even though DNAffinity predicts many potential binding sites where no experimental evidence of *in vivo* binding exists, a grand majority of these seemingly false positives are trivially explained by chromatin structure and nucleosome occupancy. When combining DNAffinity and nucleosome maps, our method was able to locate *in vivo* TFBS with a high accuracy.

DATA AVAILABILITY

The program and the full database of feature parameters are available in the GitHub repository: https://github.com/Jalbiti/DNAffinity.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Center of Excellence for HPC H2020 European Commision; BioExcel-2 Centre of Excellence for Computational Biomolecular Research [823830]; Spanish Ministry of Science [RTI2018-096704-B-100]; Instituto de Salud Carlos III–Instituto Nacional de Bioinformatica [ISCIII PT 17/0009/0007 co-funded by the Fondo Europeo de Desarrollo Regional]; European Regional Development Fund under the framework of the ERFD Operative Programme for Catalunya; Catalan Government AGAUR [SGR2017-134]; the IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from the MINECO; Modesto Orozco is an ICREA Academy scholar; S.B. is a M4L student. Funding for open access charge: Universitat de Barcelona.
Conflict of interest statement. None declared.

REFERENCES

1. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
2. Thorne,J., Campbell,M. and Turner,B. (2009) Transcription factors, chromatin and cancer. *Int. J. Biochem. Cell Biol.*, **41**, 164–175.
3. Shlyueva,D., Stampfel,G. and Stark,A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.***15**, 272–286.
4. Levo,M. and Segal,E. (2014) In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.*, **15**, 453–468.

5. Flores,O., Deniz,Ö., Soler-López,M. and Orozco,M. (2014) Fuzziness and noise in nucleosomal architecture. *Nucleic Acids Res.*, **42**, 4934–4946.
6. Deniz,Ö., Flores,O., Battistini,F., Pérez,A., Soler-López,M. and Orozco,M. (2011) Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC Genomics*, **12**, 489.
7. D'Oliveira Albanus,R., Kyono,Y., Hensley,J., Varshney,A., Orchard,P., Kitzman,J.O. and Parker,S.C.J. (2021) Chromatin information content landscapes inform transcription factor and DNA interactions. *Nat. Commun.*, **12**, 1307.
8. Voss,T.C. and Hager,G.L. (2013) Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet.*, **15**, 69–81.
9. Li,B., Carey,M. and Workman,J.L. (2007) The role of chromatin during transcription. *Cell*, **128**, 707–719.
10. Battistini,F., Hospital,A., Buitrago,D., Gallego,D., Dans,P.D., Gelpí,J.L. and Orozco,M. (2019) How B-DNA dynamics decipher sequence-selective protein recognition. *J. Mol. Biol.*, **431**, 3845–3859.
11. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–1253.
12. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordân,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
13. Smaczniak,C., Angenent,G. and Kaufmann,k (2017) SELEX-Seq: a method to determine DNA binding specificities of plant transcription factors. *Methods Mol. Biol.*, **1629**, 67–82.
14. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-Binding specificities of human transcription factors. *Cell*, **152**, 327–339.
15. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
16. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. 3rd and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429.
17. Gordân,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.*, **3**, 1093.
18. Nutiu,R., Friedman,R., Luo,S., Khrebtukova,I., Silva,D., Li,R., Zhang,L., Schroth,G. and Burge,C. (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.*, **29**, 659–664.
19. Mundade,R., Ozer,H.G., Wei,H., Prabhu,L. and Lu,T. (2014) Role of chip-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle.*, **13**, 2847–2852.
20. Rhee,H.S., Pugh,B.F., McCord,R.P., Shi,Z., Berger,M.F., Newburger,D.E., Saulrieta,K., Smith,Z., Shah,M.V., Radhakrishnan,M. *et al.* (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
21. Mahony,S. and Pugh,B.F. (2015) Protein–DNA binding in high-resolution. *Crit. Rev. Biochem. Mol. Biol.*, **50**, 269–283.
22. Hertz,G.Z., Stormo,G.D., Gordon,D.B., Gifford,D.K., Stormo,G.D., Fraenkel,E., Hannett,N., Harbison,C., Thompson,C., Simon,I. *et al.* (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
23. Schneider,T. and Stephens,R. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
24. Benos,P., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it?*Nucleic Acids Res.*, **30**, 4442–4451.
25. Zhao,Y., Ruan,S., Pandey,M. and Stormo,G.D. (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, **191**, 781.
26. Zhou,T., Shen,N., Yang,L., Abe,N., Horton,J., Mann,R.S., Bussemaker,H.J., Gordân,R. and Rohs,R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 4654–4659.
27. Wang,S., Zhang,Q., Shen,Z., He,Y., Chen,Z.-H., Li,J. and Huang,D.-S. (2021) Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture. *Mol. Ther. - Nucleic Acids*, **24**, 154–163.
28. Peng,P.-C. and Sinha,S. (2016) Quantitative modeling of gene expression using DNA shape features of binding sites. *Nucleic Acids Res.*, **44**, e120.
29. Koo,P.K. and Ploenzke,M. (2020) Deep learning for inferring transcription factor binding sites. *Curr. Opin. Syst. Biol.*, **19**, 16–23.
30. Cevost,J., Vaillant,C. and Meyer,S. (2018) ThreaDNA: predicting DNA mechanics' contribution to sequence selectivity of proteins along whole genomes. *Bioinformatics*, **34**, 609–616.
31. Chen,C., Hou,J., Shi,X., Yang,H., Birchler,J.A. and Cheng,J. (2021) DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks. *BMC Bioinforma.***22**, 38
32. Dai,H., Umarov,R., Kuwahara,H., Li,Y., Song,L. and Gao,X. (2017) Sequence2Vec: a novel embedding approach for modeling transcription factor binding affinity landscape. *Bioinformatics*, **33**, 3575–3583.
33. Li,J., Sagendorf,J.M., Chiu,T.-P., Pasi,M., Perez,A. and Rohs,R. (2017) Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Res.*, **45**, 12877–12887.
34. Fu,L., Zhang,L., Dollinger,E., Peng,Q., Nie,Q. and Xie,X. (2020) Predicting transcription factor binding in single cells through deep learning. *Sci. Adv.*, **6**, eaba9031.
35. Park,S., Koh,Y., Jeon,H., Kim,H., Yeo,Y. and Kang,J. (2020) Enhancing the interpretability of transcription factor binding site prediction using attention mechanism. *Sci. Rep.*, **10**, 13413.
36. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
37. Zhang,Q., Shen,Z. and Huang,D.S. (2021) Predicting in-vitro transcription factor binding sites using DNA sequence + shape. *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, **18**, 667–676.
38. Asif,M. and Orenstein,Y. (2020) DeepSELEX: inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs. *Bioinformatics*, **36**, i634–i642.
39. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A. *et al.* (2015) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
40. Orozco,M., Noy,A. and Pérez,A. (2008) Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.*, **18**, 185–193.
41. Dans,P.D., Balaceanu,A., Pasi,M., Patelli,A.S., Petkevičiūtė,D., Walther,J., Hospital,A., Bayarri,G., Lavery,R., Maddocks,J.H. *et al.* (2019) The static and dynamic structural heterogeneities of B-DNA: extending calladine-dickerson rules. *Nucleic Acids Res.*, **47**, 11090–11102.
42. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11163–11168.
43. Cui,F. and Zhurkin,V.B. (2010) Structure-based analysis of DNA sequence patterns guiding nucleosome positioning in vitro. *J. Biomol. Struct. Dyn.*, **27**, 821–841.
44. Miele,V., Vaillant,C., Aubenton-Carafa,Y., Thermes,C. and Grange,T. (2008) DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.*, **36**, 3746–3756.
45. R,R., X,J., SM,W., R,J., B,H. and RS,M. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
46. Fujii,S., Kono,H., Takenaka,S., Go,N. and Sarai,A. (2007) Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.*, **35**, 6063–6074.
47. Schiessel,H., Gelbart,W.M. and Bruinsma,R. (2001) DNA folding: structural and mechanical properties of the two-angle model for chromatin. *Biophys. J.*, **80**, 1940–1956.
48. Mergell,B., Ejtehadi,M.R. and Everaers,R. (2003) Modeling DNA structure, elasticity, and deformations at the base-pair level. *Phys. Rev. E*, **68**, 021911.

49. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
50. Bailey,T., Boden,M., Buske,F., Frith,M., Grant,C., Clementi,L., Ren,J., Li,W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
51. Afek,A., Schipper,J.L., Horton,J., Gordân,R. and Lukatsky,D.B. (2014) Protein−DNA binding in the absence of specific base-pair recognition. *Proc. Natl. Acad. Sci.*, **111**, 17140–17145.
52. Badjatia,N., Rossi,M.J., Bataille,A., Mittal,C., Lai,W.K.M. and Pugh,B.F. (2021) Acute stress drives global repression through two independent RNA polymerase II stalling events in saccharomyces. *Cell Rep.*, **34**, 180640.
53. Rossi,M.J., Lai,W.K.M. and Pugh,B.F. (2018) Genome-wide determinants of sequence-specific DNA binding of general regulatory factors. *Genome Res.*, **28**, 497–508.
54. Deniz,Ö., Flores,O., Aldea,M., Soler-López,M. and Orozco,M. (2016) Nucleosome architecture throughout the cell cycle. *Sci. Rep.*, **6**, 19729.
55. Hospital,A., Andrio,P., Cugnasco,C., Codo,L., Becerra,Y., Dans,P.D., Battistini,F., Torres,J., Goñi,R., Orozco,M. *et al.* (2016) BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.*, **44**, D272–D278.
56. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Müller,A., Nothman,J., Louppe,G. *et al.* (2012) Scikit-learn: Machine Learning in Python.
57. Ma,W., Yang,L., Rohs,R. and Noble,W.S. (2017) DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding. *Bioinformatics*, **33**, 3003–3010.
58. Ruan,S., Swamidass,S.J. and Stormo,G.D. (2017) BEESEM: estimation of binding energy models using HT-SELEX data. *Bioinformatics*, **33**, 2288–2295.
59. Yuan,H., Kshirsagar,M., Zamparo,L., Lu,Y. and Leslie,C.S. (2019) BindSpace decodes transcription factor binding signals by large-scale sequence embedding. *Nat. Methods*, **16**, 858–861.

Chapter 2. A nucleosome positioning predictor

In the previous chapter we developed a machine learning model based on a Random Forest Regressor that was able to predict the binding affinities of a large pool of TFs for different DNA sequences. The model was trained with *in vitro* experimental techniques, and we additionally analyzed its applicability to an *in vivo* context by introducing a new factor, nucleosome occupancy. This first work led us to introduce the role of nucleosomes on the modulation of protein recognition by DNA and in the regulatory network.

Nucleosomes constitute the first unit of chromatin organization, each unit wrapping around 147 bps of DNA in a quite packed cylinder (around 20 nm diameter). This compaction occurs through 1 and ¾ turns wrapping of DNA around a histone octamer. Nucleosomes cover most of the chromatin complex, where no obvious direct-reading mechanism for DNA positioning exists.

However, despite that, there is a maintenance in nucleosome positioning, part of which can be recovered from *in vitro* reconstitution, suggesting that there are some unknown mechanisms controlling nucleosome formation and positioning. Furthermore, nucleosome maps along different genomes showed the presence of very well-defined nucleosome depleted regions in the vicinities of transcription start and terminating sites (the nucleosome free regions; NFRs).
In this second chapter we will study the intrinsic role of DNA physical properties to NFRs, nucleosome array periodicity along gene bodies, the role of transcription, as well as other extrinsic determinants of nucleosome positioning in yeast.

2.1. Nucleosome Free Regions at the beginning and end of genes

As previously observed (1, 2) and as we have previously noted from our *in vivo* binding predictor paper described in the previous chapter, there is a high correlation between nucleosome architecture and TF binding sites at promoters. Nucleosome maps have shown that nucleosome depleted regions are often observed at the Transcription Start Sites (TSSs) of actively transcribed genes together with a strongly positioned +1 nucleosome followed by fuzzier nucleosomes as we

move downstream of the TSS (3–6). Additionally, some studies support the presence of NFRs around the Transcription Terminating Sites (TTSs), while others claim that these are only an artefact of neighboring TSSs (7). In this work we performed MNase-seq experiments processing the reads with a previously developed nucleosome peak caller, nucleR (see Introduction), has been used to prove the presence of NFRs at TTSs, in both tandem (when there is a nearby TSS) and convergent (when there is a nearby TTS) genes, and to study the distribution of nucleosome arrays.

2.2. Determinants of NFRs

We investigate the determinants of the observed nucleosome depleted regions at the 5′ and 3′ ends of genes. We considered the combination of intrinsic (sequence-dependent physical properties of the DNA) and extrinsic (mainly TF-DNA binding sites) factors in influencing NFRs. We selected two factors to describe and identify a nucleosome-depleted region: the deformation energy and the density of transcription factor binding sites (TFBSs). The deformation energy is defined by the elastic energy associated with the harmonic deformation of the DNA from a naked state to a nucleosome-bound state and was derived by a mesoscopic model with equilibrium and stiffness parameters derived from Molecular Dynamic simulations (8). On the other hand, the TFBSs densities were taken from annotated sources (9) and aimed to model the extrinsic effects i.e. the competition of nucleosomes with other effector proteins (see Figure 2.1).

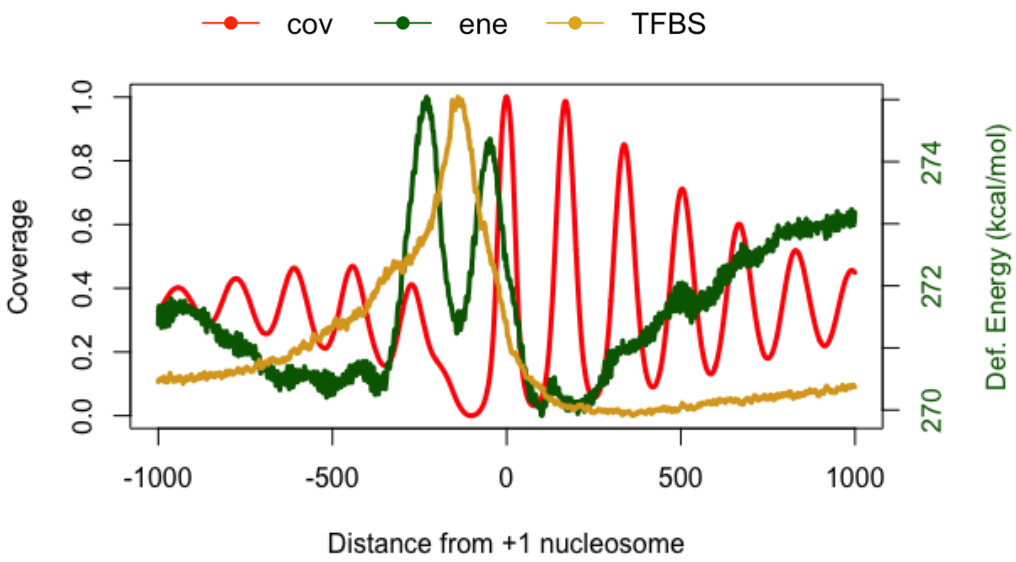


Figure 2.1. The deformation energy (green line) and transcription binding site densities (yellow line) against the experimental nucleosome coverage (red line), centered at around the +1 nucleosome for all TSSs.

2.3. Neural Network Predictor

Using the intrinsic and extrinsic features described above we built a machine learning classifier that would predict the probability of a region being nucleosome free. Neural networks were first motivated by the functionality of the human brain, developed in an attempt to model a biological neuron (10). A network of neurons is built by linking multiple neurons together in a way that the output of one neuron forms an input to another. The original idea by McCulloch and Pitts was that a neuron receives input signals that are combined with corresponding weights. If this combination exceeds a threshold, then a neuron fires, otherwise it remains inactive (see Introduction for more details). In a neural network architecture, we will always find a first input layer with as many neurons as descriptors, one or more hidden layers, and a final output layer, which corresponds to the given solutions. We chose a neural network architecture for our model given the following two characteristics:

- Neural networks can find patterns within unstructured data.
- Neural networks efficiently reproduce high dimensional functions when other classical approaches cannot.

Our predictor was trained using experimental maps for all TSS-NFRs and TTS-NFRs defined by well-positioned nucleosomes using the previously developed nucleR software (11). The classifier obtained an Area Under the Curve (AUC) of 0.96 in the test set and was used through the whole genome.

2.4. Intragenic Nucleosome Positioning

Having seen the existence of NFRs at the beginning and end of genes, we state the possibility of having a barrier-like positioning of the first nucleosome from a gene (+1) and the last one (-last) which determine the intragenic positions along the gene body. The precise positions are established following statistical positioning as determined by signal transmission theory with two emitters located at the vicinities of the TSS and TTS of a gene. We tested a simple signal decay model which is based on a periodicity of 165 bps (optimized for yeast) and showed that this simple model can predict with accuracy the nucleosome coverage within gene bodies (see Figure 2.2), provided that the +1 and -last nucleosomes are well placed contiguous to their corresponding neighboring NFRs.

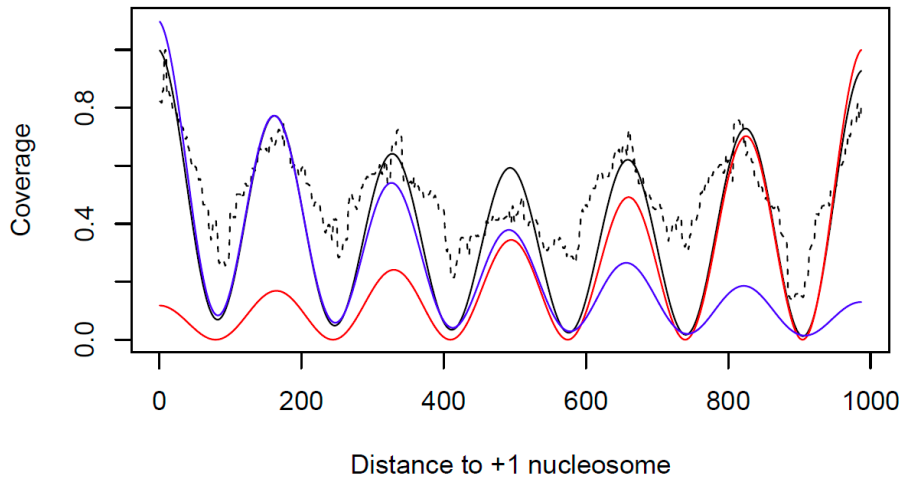


Figure 2.2. Signal Transmission Theory model to reproduce intragenic nucleosome architectures. Two signals are emitted, the first one from the +1 nucleosome (blue signal) and the second one from the -last nucleosome (red signal) resulting in a periodic signal (black line) that could reproduce the experimental mapping (dashed black line).

2.5. Phase and Autocorrelation of genes

Following our Signal Transmission Theory (STT), genes can be categorized, based on the distance between the +1 and the -last nucleosomes, into two classes according to their coverage profile: a set of genes where the distances are within a multiple of the period and therefore their signals overlap significantly, and a second set of genes where their distance is not in phase and therefore the signals do not overlap, presenting a fuzzier nucleosome architecture along the gene body. In this project we experimentally explored the effect of phase on nucleosomes by adding an 81-nucleotide (81-nt) sequence to eight selected genes (four of which were phased and four control not phased).

2.6. The role of transcription in nucleosome positioning

Finally, we investigated the correlation between phase and transcription to decipher the causality link between them. This was motivated by the observation that genes presenting a well-positioned nucleosome architecture showed higher levels of expression in comparison to genes with fuzzier profiles (12). However, this link was questioned by some studies claiming the opposite, suggesting that more active genes present less ordered nucleosome architectures (13). For this reason, we analyzed the impact on transcription from the addition of an 81-nt sequence in our eight selected genes. To further investigate this, we also analyzed data on the effect of transcription upon nucleosome positioning using both in house and previously published experiments, to avoid making conclusions that could be biased by the experimental technique. In our lab, we carried out MNase-seq experiments in cells treated by 1,10-phenantroline, a metal chelator that stalls the polymerase at the promoters and inhibits transcription (14–16). To avoid any bias from this experimental technique and make our conclusions more robust, we additionally analyzed previously published results (3) where transcription inhibition occurred through a different experimental set up, in principle less prone to artefacts than phenanthroline treatment.

We have shown that nucleosome positioning can be regulated by the combination of intrinsic and extrinsic factors, which help us predict nucleosome depleted regions, which in combination with statistical positioning, through STT, can reproduce the nucleosome architecture in yeast. Herein, we demonstrated that the positions of the

+1 and -last nucleosomes can be accurately predicted (AUC of 0.96) by a neural network classifier that utilizes an energetic contribution related to DNA deformability and protein binding densities. Simple statistical positioning allows the reconstitution of intragenic nucleosomes with higher accuracy than experimental noise. Furthermore, with the insertion of the 81-nt sequence we observed a change in the overall nucleosome organization with a fuzzier and less periodic profile in the originally phased set of genes. However, these changes were not observed in the control genes that were previously not classified as phased.

Finally, we did not find any significant effects of the 81-nt insert on transcription levels, including in those originally phased genes, which led us to formulate the causal relationship: nucleosome periodicity is affected by transcription, but alteration of periodicity or phasing does not change in a systematic manner gene expression. Finally, our data show that the addition of 1,10-phenantroline for transcription inhibition led to larger NFRs (mostly from the displacement of the -1 nucleosome), an overall increase in nucleosome fuzziness and a decrease in the proportion of phased genes. This strongly suggests that the presence of RNA polymerase affects nucleosome architecture and not the reverse, thus confirming our causal relationship.

This second publication clarifies the main determinants of nucleosome positioning and exemplifies the effectiveness a neural network to reconstitute chromatin organization.

Publication:

Alba Sala*, Mireia Labrador*, Diana Buitrago, Pau De Jorge, Federica Battistini, Isabelle Brun Heath and Modesto Orozco. An integrated machine-learning model to predict nucleosome architecture (*accepted for publication* - Nucleic Acids Research).

*Equally contributing authors
Supplementary material for this article can be found in the Annex.

References

1. Jiang,C. and Pugh,B.F. (2009) Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet, 10, 161–172.
2. Lai,W.K.M. and Pugh,B.F. (2017) Understanding nucleosome dynamics and their links to gene expression and DNA replication. Nat Rev Mol Cell Biol, 18, 548–562.
3. Weiner,A., Hughes,A., Yassour,M., Rando,O.J. and Friedman,N. (2010) High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Res, 20, 90–100.
4. Chereji,R. V. and Morozov,A. V. (2015) Functional roles of nucleosome stability and dynamics. Brief Funct Genomics, 14.
5. Deniz,Ö., Flores,O., Battistini,F., Pérez,A., Soler-López,M. and Orozco,M. (2011) Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. BMC Genomics, 12, 489.
6. Deniz,Ö., Flores,O., Aldea,M., Soler-López,M. and Orozco,M. (2016) Nucleosome architecture throughout the cell cycle. Sci Rep, 6, 19729.
7. Chereji,R. V., Ramachandran,S., Bryson,T.D. and Henikoff,S. (2018) Precise genome-wide mapping of single nucleosomes and linkers in vivo. Genome Biol, 19.
8. Battistini,F., Hospital,A., Buitrago,D., Gallego,D., Dans,P.D., Gelpí,J.L. and Orozco,M. (2019) How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition. J Mol Biol, 431.
9. Pachkov,M., Balwierz,P.J., Arnold,P., Ozonov,E. and van Nimwegen,E. (2012) SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Res, 41, D214–D220.
10. McCulloch,W.S. and Pitts,W. (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys, 5.
11. Flores,O. and Orozco,M. (2011) nucleR: a package for non-parametric nucleosome positioning. Bioinformatics, 27, 2149–2150.
12. Vaillant,C., Palmeira,L., Chevereau,G., Audit,B., D'Aubenton-Carafa,Y., Thermes,C. and Arneodo,A. (2010) A novel strategy of transcription regulation by intragenic nucleosome ordering. Genome Res, 20, 59–67.
13. Jiang,Z. and Zhang,B. (2021) On the role of transcription in positioning nucleosomes. PLoS Comput Biol, 17.
14. Grigull,J., Mnaimneh,S., Pootoolal,J., Robinson,M.D. and Hughes,T.R. (2004) Genome-Wide Analysis of mRNA Stability Using Transcription Inhibitors and Microarrays Reveals Posttranscriptional Control of Ribosome Biogenesis Factors. Mol Cell Biol, 24, 5534–5547.

15. Kim,T.S., Liu,C.L., Yassour,M., Holik,J., Friedman,N., Buratowski,S. and Rando,O.J. (2010) RNA polymerase mapping during stress responses reveals widespread nonproductive transcription in yeast. Genome Biol, 11.
16. McClure,W.R., Cech,C.L. and Johnston,D.E. (1978) A steady state assay for the RNA polymerase initiation reaction. Journal of Biological Chemistry, 253.

AN INTEGRATED MACHINE-LEARNING MODEL TO PREDICT NUCLEOSOME ARCHITECTURE

Alba Sala[1,&], Mireia Labrador[1,&], Diana Buitrago[1], Pau De Jorge[1], Federica Battistini[1,2], Isabelle Brun Heath[1] and Modesto Orozco[1,2*]

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain; [2] Departament de Bioquímica i Biomedicina, Universitat de Barcelona, Barcelona, Spain

[&] These authors contributed equally to this work

[*] Correspondence to M.Orozco: modesto.orozco@irbbarcelona.org

Abstract

We demonstrate that nucleosomes placed in the gene body can be accurately located from signal decay theory assuming two emitters located at the beginning and at the end of genes. These generated wave signals can be in phase (leading to well defined nucleosome arrays) or in antiphase (leading to fuzzy nucleosome architectures). We found that the first (+1) and the last (-last) nucleosomes are contiguous to regions signaled by transcription factor binding sites and unusual DNA physical properties that hinder nucleosome wrapping. Based on these analyses, we developed a method that combines Machine Learning and signal transmission theory able to predict the basal locations of the nucleosomes with an accuracy similar to that of experimental MNaseq-based methods.

Introduction

Nucleosomes (the basic units of eukaryotic chromatin) are formed by 147bp of duplex DNA wrapped around an octamer of histones(1), followed by a linker DNA where, in complex eukaryotic organisms, an additional histone (H1) can be bound(2). Nucleosomes are not randomly placed but maintain a defined architecture along the genome, with certain positions occupied by well-positioned nucleosomes while others are nucleosome-free(3–9). Most significant nucleosome free regions (NFRs) are associated with the promoter regions of genes (upstream of the Transcription Start Sites, TSSs), the replication origins (ORIs) and the Transcription Termination Sites (TTSs)(10, 11). The general consensus is that NFRs at TSSs are preferentially recognized by effector proteins involved in the regulation of gene activity, and the widths of these regions correlate with gene expression(12). Furthermore, perturbation in nucleosome architectures associated to stress, changes in cell cycle phases, source of nutrients, or the cell metabolic cycle(6, 11, 13, 14) proved the connection between nucleosome architecture and gene activity. The causality in this relationship is however unclear.

Over the last two decades, many efforts have been made to discover the main determinants of nucleosome positioning(15–20). Several studies have suggested that DNA physical properties are crucial for defining nucleosome positioning, with NFRs characterized by sequences where the mechanical cost of wrapping DNA around nucleosomes is very high(10, 13, 21). On the contrary, others have suggested that nucleosome positioning is dictated exclusively by cellular machinery involving a complex interplay between chromatin remodelers, transcription factors and RNA polymerase activity(22–25). Chromatin reconstitution experiments(13, 22, 26, 27) demonstrated that NFRs are well reproduced *in vitro*, but their boundaries are not precise in the absence of cellular effectors. These findings suggest that while physical principles can signal NFRs, cellular machinery is required for the correct definition of their boundaries(18, 25, 28). Nonetheless, it is unclear how these intrinsic and extrinsic signals are combined to define nucleosome architecture.

We explore here whether the basal nucleosome architecture can be determined by easily available DNA descriptors such as sequence-dependent physical properties(29, 30) and sequence annotations of transcription factor binding sites (TFBSs) (31). Additionally, we investigated whether changes in nucleosome architecture are a reason for, or a consequence of gene expression. With this aim, we developed a method that combines Machine Learning (ML) and signal transmission theory (STT) able to predict the most probable nucleosome architectures in yeast with accuracy comparable to experimental techniques. Furthermore, synthetic biology experiments demonstrate that the structural fingerprint of active genes (characterized by wide NFRs and phased nucleosome arrays) is a consequence, rather than a reason for their gene expression level(14, 32).

Material & Methods

Yeast strains and growth conditions. The *Saccharomyces cerevisiae* PPY1 strain (*MATa his3Δ0 leu2Δ0 met15Δ0 ura3Δ0 bar1::leu2)* was transformed with the appropriate DNA fragments to generate all the mutant strains used in this work. The PPY1 strain was obtained from Oscar Aparicio's lab at the University of Southern California, USA. For the selection of the mutant strains, we used YPD with or without 5-FOA (5-Fluoroorotic acid) and SD (Synthetic Defined) with the required amino acids.

Mutant strains generation. We generated 4 mutant strains, with the 81-nt DNA sequence (5'GCGTGTTGTGTTTTCTCCGAGGAGAAACATTCAAATCTTGTGCTATGGCTTTGCCTACCGTCTGCG CC ATCCATCTTTCGC-3') inserted in the coding sequence of 2 selected genes per strain (Table 1). We selected four non-essential genes which showed phased nucleosomes (UBX5, CKB2, PPT1, TRP4; see phase definition below) together with four non-essential control genes which were unphased (BSP1, DGK1, SLM3, PAN5). The 81-nt sequence was designed not to match any existing yeast sequence and not to favor nor disfavor nucleosome formation or affect the reading frame (see Results). The strains were produced using the *Delitto Perfetto* strategy described in (33).

Gene	Strain	Strand	Start	End	Chromosome	Insert Position	Phasing
UBX5	1	+	1127872	1129374	chrIV	1128586	Phased
CKB2	2	+	405768	406544	chrXV	406248	Phased
PPT1	3	-	736662	738203	chrVII	737615	Phased
TRP4	4	+	1184747	1185889	chrIV	1185196	Phased
BSP1	1	+	883828	885558	chrXVI	884578	Control (not-phased)
DGK1	2	-	899056	899928	chrXV	899667	Control (not-phased)
SLM3	3	-	392659	393912	chrIV	393462	Control (not-phased)
PAN5	4	-	224030	225169	chrVIII	224581	Control (not-phased)

Table 1. Genes modified with the 81-nt sequence in each strain.

RNA extraction and RT-qPCR. Three independent colonies from each yeast strain were grown until exponential phase and then arrested at late G1 by alpha-factor. RNA was obtained from 10ml yeast cultures (OD_{600} 0.8) using the hot-phenol method. cDNA synthesis was done with the First Strand cDNA Synthesis Kit (Roche) using oligo dT and following the provider instructions. Gene expression levels were determined by quantitative PCR using the LightCycler 480 sybr green I master (Roche). The in Suppl. Table S1.sed for the qPCR are listed

Transcription inhibition. In order to determine the correct incubation time to inhibit transcription without killing the cells, we selected 2 genes with low RNA stability (RPA135 and NMD3) and 2 genes with high RNA stability (ACT1 and DGK1) to serve as controls(34). We then measured their mRNA level by qPCR after incubation with 10-phenanthroline at 100 µg/ml at 30ºC during 0, 5, 15, 30 and 45 minutes. Using this approach, we observed that the amount of RPA135 and NMD3 mRNA started to decrease after 30 min. This incubation time was selected to perform the MNase-seq experiments on cells with inhibited transcription.

MNase digestion. The Micrococcal nuclease (MNase) digestion was performed on semi-intact yeast cells prepared as described elsewhere(35). We optimised the MNase digestion conditions for each sample to obtain about 80% of mononucleosomes. The integrity and size

distribution of digested fragments were determined using the microfluidics-based platform Bioanalyzer (Agilent) prior to sample preparations and sequencing. The sample preparation was done using the Illumina TruSeq DNA sample preparation kit for whole genome sequencing, following the Illumina standard protocol. The libraries (paired-end) were sequenced paired-end on a HiSeq2000, v4, 2x75bp, with approximately 10 M PE reads/sample.

Nucleosome calling. MNase-seq paired-end reads were mapped to customized versions of the yeast genome (SacCer3, UCSC), containing the inserted sequences in the modified genes, using the Bowtie(36) aligner, allowing up to two mismatches. Output files were imported in to R where reads were trimmed to 50bp maintaining the original center and transformed to reads per million bp. Peak calling was performed, after noise filtering, with the nucleR package implemented in the Nucleosome Dynamics platform(Buitrago et al., 2019; Flores & Orozco, 2011) using the standard parameters for yeast: peak width of 147 bp, peak detection threshold of 35% and maximum overlap of 80 bp. Nucleosome calls were considered well-positioned when nucleR's peak width score and height score were higher than 0.6 and 0.4(39) respectively, and fuzzy otherwise.

Nucleosome periodicity and phasing. Periodicity in nucleosome positioning was determined for each gene by computing the autocorrelation coefficient, as seen in (40):

$$R(T) = \int_{X_1}^{X_2} I(x) \cdot I(x - T) dx \qquad (1)$$

where X_1 and X_2 stand for the limits of a sampling window (e.g. the position of TSS and TTS), I is the function representing nucleosome coverage for all genes and T is the period. This thus reflects the continuity of a nucleosome repeat length and will have a maximum when all the nucleosome peaks are T units apart as $R(T)$ will be the highest. In other words, autocorrelation is defined by the correlation between a profile and shifted versions of itself. This method can uncover hidden patterns in the signal that wouldn't be clear by just examining the strength of the signal itself. Autocorrelation coefficients for different periods were normalized as shown in:

$$\grave{R}(T) = \frac{R(T)}{R(0)} \quad (2)$$

Nucleosome period is defined as the value of T that optimizes $\grave{R}(T)$, and periodic genes are those showing large autocorrelation coefficient values (eq. 1). Phased genes are defined as those where the +1 to –last distance (L) is a multiple of the period (T). Anti-phased genes are those where the distance from integer (DFI) score, defined as the modulus of the ratio length/periodicity, is close to $T/2$. Unphased genes refer to intermediate values.

$$DFI = L - T \cdot round\left(\frac{L}{T}\right) \quad (3)$$

Signal Transmission Theory for nucleosome positions. Having observed experimentally two clear NFRs at the beginning and end of genes, we propose a simple signal decay model, where the coverage at a given position is given by the addition of two independent positioning signals emitted from the two ends of a gene, one starting from the +1 (i.e., the nucleosome right after the NFR at promoter region), and another one from the –last nucleosome (i.e., right before the TTS). The strength of the signal emitted from the two ends has been chosen to better reproduce the nucleosome coverage pattern. For this reason, the emitted value is strongest at the position of the +1 ($Cov^{+1}(X)$) compared to the emitter from the -last nucleosome ($Cov^{-last}(X')$), as can be seen in:

$$Cov^{+1}(X) = \left(1 + \alpha + sin\left(\frac{\pi}{2} + 2 \cdot \frac{\pi}{T} X\right)\right)\sigma^{\left(\frac{|X|}{T}\right)} \quad (4)$$

$$Cov^{-last}(X') = \left(1 + sin\left(\frac{\pi}{2} + 2 \cdot \frac{\pi}{T} X'\right)\right)\sigma^{\left(\frac{|X'|}{T}\right)} \quad (5)$$

where X is the distance from the +1 nucleosome and X' is the distance from the -last nucleosome, *X'=L-X*. The shifting factor corrects for the higher density of reads at the +1 nucleosome and the decay factor accounts for the reduced coverage as we move away from the NFR. We evaluated different values for and and selected those that maximized the correlation between the observed experimental coverage and the predicted (α was set to

0.2 and σ was set to 0.7). The total coverage is then normalized to guarantee an effective decay of the signal:

$$Cov(X) = \frac{Cov^{+1}(X)+Cov^{-last}(X')}{Cov^{+1}(0)+Cov^{-last}(0)} \quad (6)$$

where, $Cov^{+1}(0)$,$Cov^{-last}(0)$ are the values of the two emitting signals at the +1 nucleosome dyad, which are used as denominator to normalize the *Cov(X)* to 1 at this position. The difference in strengths will be determined by the factor ⍰⍰which affects the normalization of the total coverage.

Deformation energy. The elastic energy associated to the DNA deformation from the naked to the nucleosome DNA was calculated in the harmonic regime using:

$$Energy = \frac{\sum_{j=1}^{146} E_j}{146} \quad with \quad E_j = \frac{1}{2}\sum_{s=1}^{6}\sum_{t=1}^{6} k_{st}^{j}\Delta X_s^j \Delta X_t^j \quad (7)$$

where j stands for each of the 146 base pair steps of the DNA stretches. E_j is the elastic energy required at each base pair step determined using the stiffness matrix (K), and ΔX_s^j and ΔX_t^j are the differences between the nucleosome and equilibrium values for the 6 base pair step helical parameters (roll, twist, tilt, slide, rise or shift). The equilibrium values and stiffness constants for each individual base pair step were taken from MD simulations that cover all the unique base pair steps in all the possible tetranucleotide environments from microsecond-long parmbsc1 simulations(29, 30).

Machine Learning. A Neural Network (NN) was developed to predict NFRs (and non-NFRs, i.e., regions occupied by nucleosomes) based exclusively on sequence information. The selected NN was defined by three layers: an *input layer* which depended on the selected centered window length, a 30-neuron *hidden layer* and a 2-neuron *output layer* to obtain the respective probabilities for the two studied classes. We used a *ReLU* activation function for the *hidden layer* and a *sigmoid* activation function for the output layer as defined in Keras(41). The *binary crossentropy* function was minimized using a stochastic gradient descent optimizer with parameters 0.001 for the learning rate and 0.003 for the momentum. The NN was built

using the Python library scikit-learn (v 0.20.3)(42), the software library TensorFlow(43) and its API Keras.

Data on nucleosome positions were obtained from MNase-seq experiments of yeast cells synchronized at the G1 phase and processed with nucleR as implemented in our Nucleosome Dynamics package(37, 38). NFRs defined by nucleR proximal to TSS or TTS were used for training. As non-NFRs are more present in the data than NFRs, we randomly removed points from non-NFRs to obtain a balanced data set corresponding to the NFRs/one and non-NFRs/zero classes. A train/test ratio of 80/20 was used to reduce overtraining artifacts and to choose the best ML algorithm. To explore the generality of the model some additional tests were done by training the model only with data from one chromosome (chrIV) and validating the model on the whole genome. Additionally, to avoid any potential bias training and testing on the same set of data, we re-trained our model with data from DANPOS (44) and tested on nucleR defined maps.

Results

NFRs are characterized by unique DNA physical properties and by high density of protein-recognition sequences.

Our analysis of the entire yeast genome reveals the placement of NFRs at the TSS and TTS of genes; the latter being present in different classes of gene, tandem and convergent, showing that the NFRs at the TTS is not a duplication of a neighboring TSS (see Suppl. Figure S1). Interestingly, both NFRs correspond to regions where the harmonic deformation energy (see Methods) required to wrap the DNA around the histone core is unusually high and where there is a large density of potential TFBS (Figure 1). Note that this behavior was found for both open (Figure 1A-B) and closed (Figure 1C-D) NFRs, suggesting the existence of a sequence-coded fingerprint characterizing all NFRs (at least at functionally relevant gene positions).

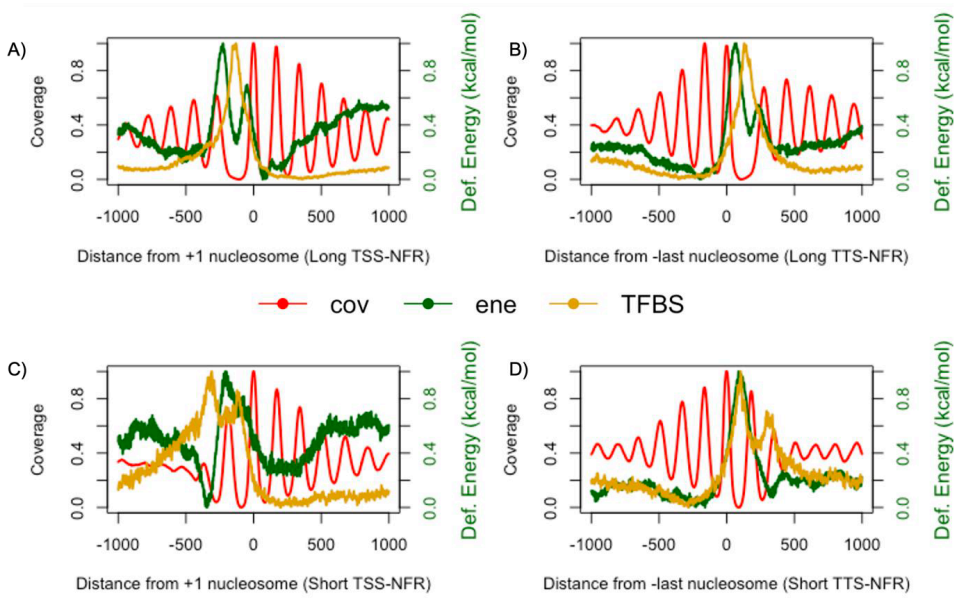


Figure 1. Average nucleosome coverage (red), TFBS density (yellow) and deformation energy (green) around well positioned +1 and –last nucleosomes for open NFRs (wider than 215 bps): Long TSSs-NFR (panel A , 2749 genes), Long TTSs-NFR (panel B, 1134 genes); and closed NFRs (shorter than 215 bps): Short TSSs-NFR (panel C, 644 genes) and Short TTSs-NFR (panel D, 942 genes).

The differential characteristics of NFRs allowed us to train a Neural Network (NN) classifier to predict NFRs using as predictive features the deformation energy and experimental TFBS density profiles through the entire genome (see Methods). These features are taken as stacked vector windows of size *N* around a center point, which define the first input layer of our neural network consisting of *N x 2* neurons (see Suppl. Figure S2). The resulting method has a good NFR prediction power as shown by the Area under the ROC Curve (AUC) of 96% and accuracy of 92% (Figure 2A-B; see also some examples of NFR predictions along the genome shown in Suppl. Figure S3). As described in Methods, in order to demonstrate the consistency of our model we trained our predictor using only one large chromosome (chrIV) and tested it on all remaining chromosomes with good results (accuracy of 90% on the remaining chromosomes).

The NN-prediction peaks were then fitted to a Gaussian curve to define the 95% range (2 standard deviations of the mean; see Figure 2C-F) at which we place the +1 (TSS) and –last (TTS) nucleosomes. In order to test the precision of this procedure, we then considered the overall averaged experimental maps and located (45) the first peak before and after the NFR to which we assigned the dyad of the -last and +1 nucleosomes, respectively. After the positioning we calculated the absolute difference between the experimental positions and our predicted ones considering the Gaussian fitting. This simple method allowed us to position the +1 and -last nucleosomes with striking accuracy: only 4bp (+1) and 17bp (-last) away from the detected peaks (45)(Figure 2C-D for TSSs and 2E-F for TTSs). The results from increasing and decreasing our predetermined standard deviation from our Gaussian fitting allowed us to observe a decrease in accuracy as we move away from two standard deviations. This was also the case when considering separately and optimizing for different classes of genes (tandem or convergent). Additionally, training our model and performing the same analysis for larger and shorter windows (600bps and 250bps), we observed similar or worse results (data not shown). We repeated the same procedure but training only using a single chromosome (chrIV) and obtained similar results (see Suppl. Figure S4A-D). Quite interestingly, our method works for both open (the distance between the two dyads is wider than 215bp) and closed NFRs (the distance between the two dyads is shorter than 215bp)(39), especially when predicting the crucial +1 nucleosome (see Suppl. Figure S4E-H). In order to further validate our model and discard possible biases given by training and testing the algorithm on the same dataset, we trained our algorithm with nucleosome positioning maps obtained from a widely used calling tool DANPOS (44) and tested it with our experimental data. The results obtained showed that our model trained with DANPOS nucleosome positions was still able to position the +1 and -last nucleosomes, respectively 9bps and 11bps away from the experimental average peak (see Suppl. Figure S5). This is similar to the distances obtained when the algorithm was trained with maps from nucleR.

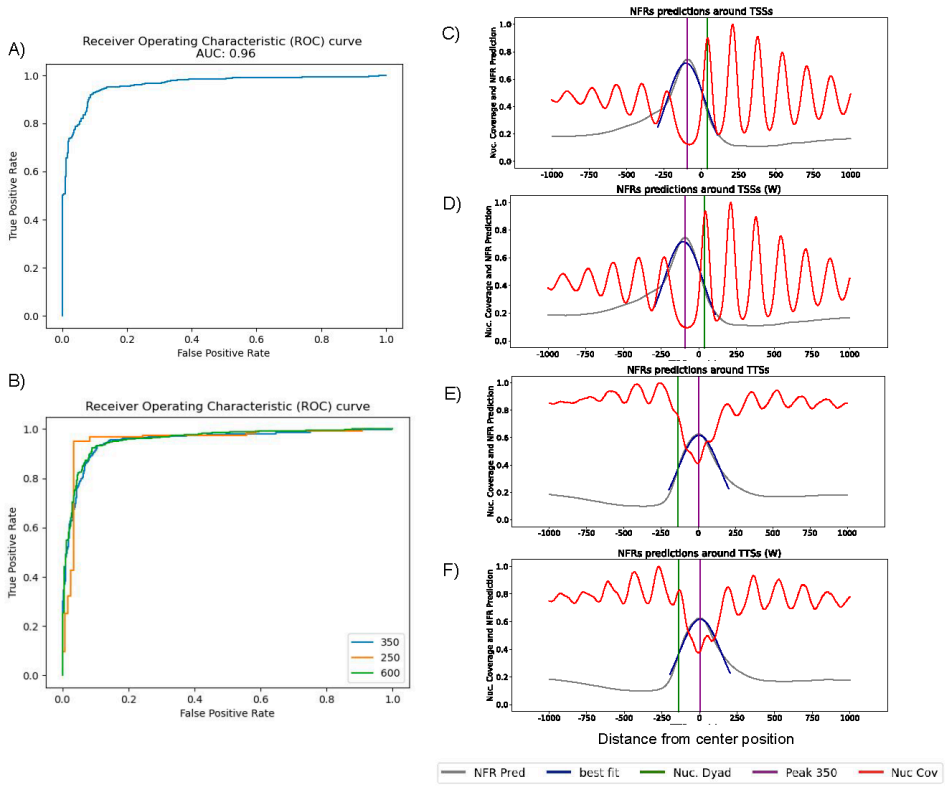


Figure 2. A-B panels showing the receiver operating characteristic (ROC) curve results from our Neural Network for A) a 350 bp window and B) different window sizes ROC curves (350bp in blue, 250bp in orange and 600 bp in green). NFR prediction (grey) against nucleosome experimental coverage (red) for C) all TSSs (5676 genes), D) well defined TSSs (3393 genes), E) all TTSs (5676 genes) and F) well defined TTSs (2076 genes). Green lines denote the average prediction of the +1 (in C and D) and –last (in E and F) nucleosomes, 2 stds from a fitted Gaussian distribution (dark blue). Purple lines mark the peak of the NFR probability prediction. Around 19% of the genes were excluded from the analysis given that they were missing the+1 and/or -last nucleosome experimental calls (see Suppl. Table S2).

Nucleosome positioning along gene body is determined by distance-decayed periodic signals.

We showed above how nucleosome positioning at the beginning and end of genes can be defined by a combination of intrinsic and extrinsic properties coded in the DNA sequence. Our next step was to predict the placement of nucleosomes in the gene body and to understand why some of these nucleosomes appear well-positioned giving clear signal in the experimental MNase-seq maps, while others appear quite delocalized leading to fuzzy signals. Firstly, we determined nucleosome periodicity by computing the autocorrelation coefficient (see Methods for details) from the nucleosome coverage from one of our MNase-seq experiments for different given periods (T; see eq. 1-2), finding a clear peak at 165 bp (the average nucleosome repeat length in yeast(46); see Suppl. Figure S6). This establishes reasonably well the distance between the +1 and the -last nucleosomes (DFI, see Methods) as a multiple of T, but as expected not the distance between the TSS and TTS which we find to be more uniform (Suppl. Figure S7).

Taking the experimental positions of the +1 and –last nucleosomes as emitting sites of a distance decaying signal (see Methods, eq. 4-6), we can predict the positions of the intragenic nucleosomes (this is defined throughout the paper as our *combined prediction*) with good accuracy (Figure 3). As expected, the model performs best when predicting those nucleosomes which are well positioned close to the TSS, while the largest deviations are found in central regions, where nucleosomes show more cell-dependent variability (see the green error bars in Suppl. Figure S8A), leading to fuzzier signals. Similarly, at the TTS our prediction improves when the experimental nucleosomes are well positioned, and we perform worse when we observe fuzzier architectures.

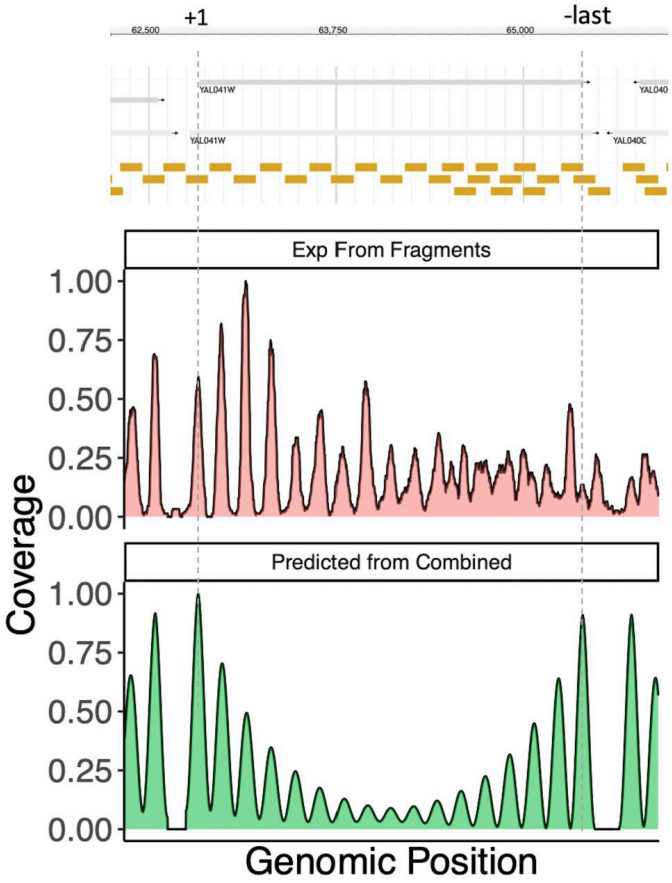


Figure 3. Scheme of resulting calls using nucleR (top golden boxes) with the coverage coming from the experimental mapping (red profile) against the predicted coverage from our signal theory combined prediction; see Methods (green profile). Note the cell variability detected as multiple peak callings detected by nucleR in the top plot.

Globally, we could reproduce well the nucleosome architecture within the gene (green box in Suppl. Figure S8B) with 85% of the experimental nucleosomes correctly predicted, and an average distance to the experimental peak (as determined by nucleR) of 19bp. Note that, matching the prediction from our model, experimental MNase-seq maps (Figure 4A) show the presence of phased genes, where the +1 and –last signals add up to define clear and periodic nucleosome patterns, and unphased genes, where signals can partially cancel out in the middle of the gene, leading to diffuse nucleosome patterns (see examples in Figure 4B, and

profiles in Figure 4C), illustrating cell variability in these regions. Additionally, it is also worth noting the stronger intensity detected experimentally of the +1 nucleosome emitter compared to the –last one, something that is considered and well reproduced by our signal-decay model (see Methods).

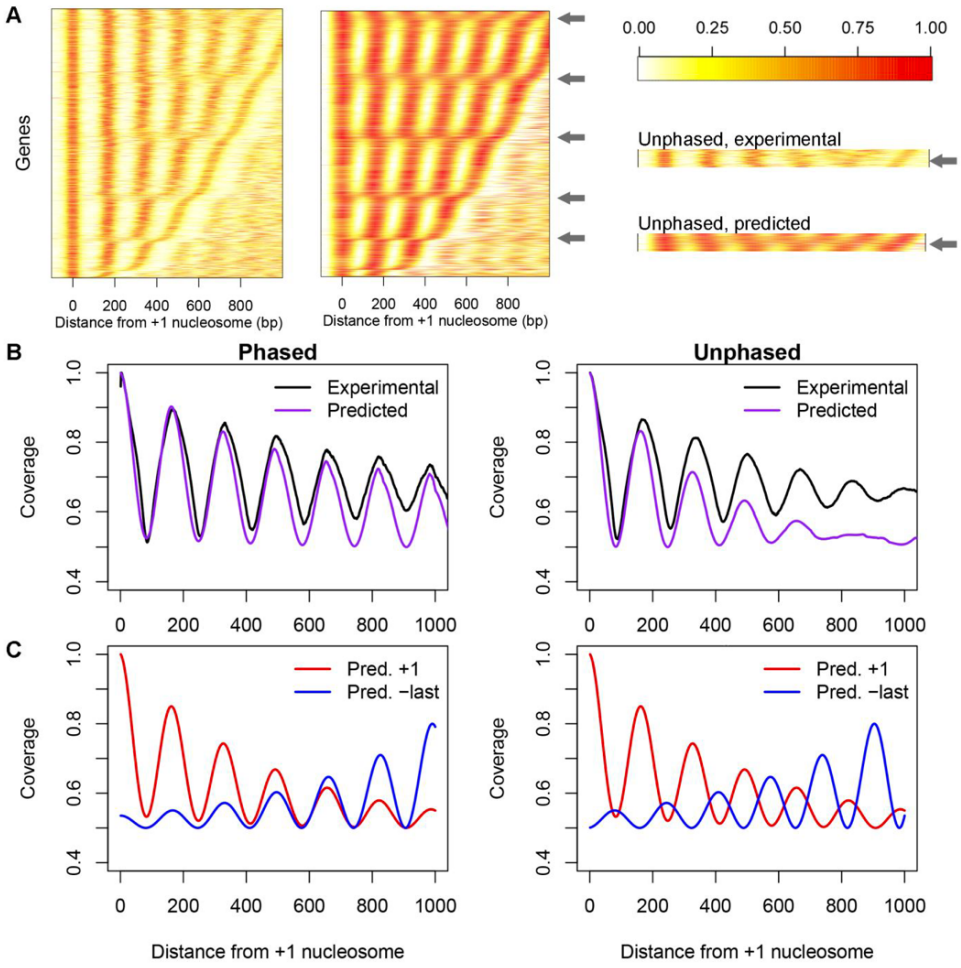


Figure 4. Signal decay model of nucleosome positioning. A) Experimental (left panel) and predicted (right panel) nucleosome coverage for each gene, with respect to the +1 nucleosome. Genes are sorted by the distance between the +1 and the –last nucleosomes. Colour scale corresponds to normalized nucleosome coverage, from 1 (red) to 0 (white). B) Nucleosome coverage, experimental (black) and predicted (purple, see Methods eq. 6) from the +1 nucleosome, averaged across all genes. Genes are split into phased or unphased based

on DFI<10 and DFI>40, respectively. C) Signals from the +1 (red) and the –last nucleosomes (blue) to predict the experimental coverage (see Methods eq. 4-5).

Predicting Nucleosome positioning along the entire genome

We can now join the NFR predictor and the periodic signals from STT to reconstruct the nucleosome architecture at the genome level (*full prediction* throughout the paper). This method allowed us to reproduce 78% of the nucleosome profile with an average distance of 32±22bp from the experimental peaks determined by nucleR (blue box in Suppl. Figure S8B), without any experimental information on the position of the +1 and -last nucleosomes. The average distance to nucleR peaks compares well to the average experimental distance of 37bp found between the centers of each individual read and the corresponding peak obtained from the coverage of all the reads (37) (red box Suppl. Figure S8B), indicating that our nucleosome position estimate is within the intrinsic experimental noise derived from cellular heterogeneity. In our *full prediction* we detected a milder increase in noise as we displaced to the center of the gene than that found with our *combined prediction* (i.e. STT prediction based on experimental +1 and –last nucleosomes (blue bars in Suppl. Figure S8A)). Finally, our *full prediction* model is able to distinguish well between phase and unphased genes as detected experimentally (Suppl. Figure S9).

Next, we explored the robustness of the predicted nucleosome arrays to changes in gene expression(10). To this end, we checked the ability of our G1-trained model to reproduce nucleosome arrays within cells collected at different cell cycle phases or determined from completely different experiments(11). As seen in Suppl. Figure S10 the model performs very well in reproducing not only G1, but also M and S data. This suggests that there is a basal nucleosome architecture, which can be reasonably well reproduced by our simple predictive model, despite of specific changes related to the action of the cellular machinery.

Nucleosome architecture and gene expression are coupled in a complex way

Results above strongly suggest that nucleosomes along the genes are located based on periodicity rules from signals derived from the presence of NFRs, which position the +1 and –last nucleosomes, organizing the rest of the nucleosome string, which can be well ordered in

the case of phased genes, or fuzzier in the case of the non-phased ones. Analysis of MNase-seq and RNA-seq datasets from(47) show that, as expected(14, 32), transcriptionally active genes are associated with wider NFRs. Interestingly, while nucleosome positions are not altered dramatically, the coverage is more periodic along the gene bodies in transcribed regions in comparison to the inactive counterparts (Figure 5A and 5C) as was previously observed (40). This finding, which is also clear when looking at predicted nucleosome coverage (Figure 5B and 5D), reveals a correlation between expression and periodicity, a result that agrees with the "crystal-like" behavior of nucleosome in active genes suggested by Vaillant et al(7).

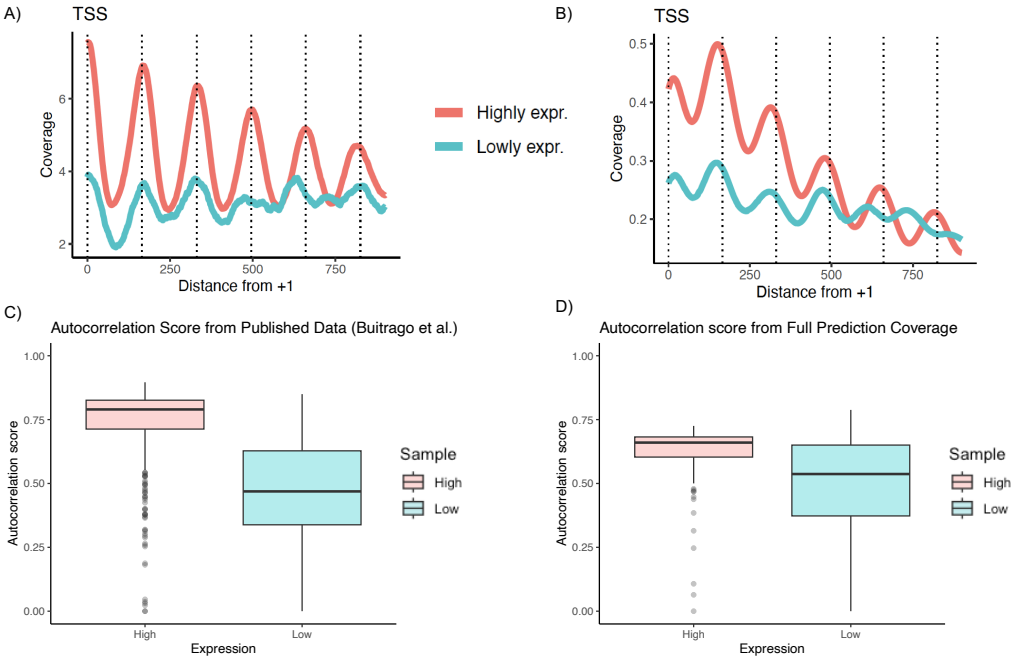
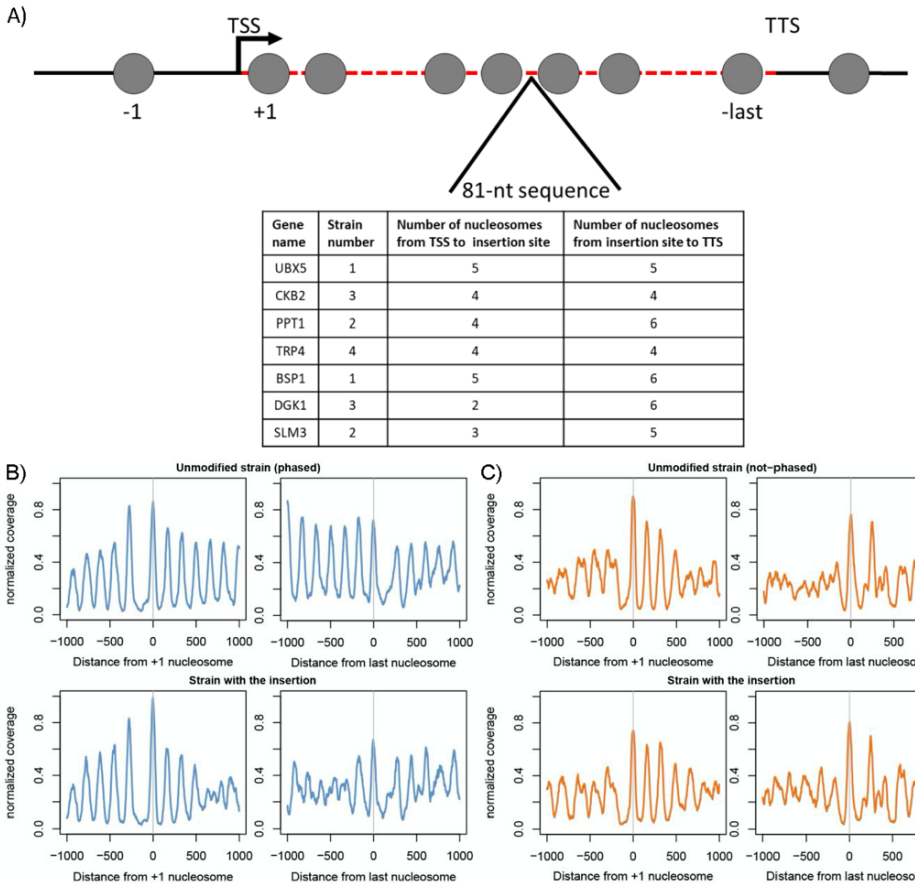


Figure 5. Coverage for highly (713 genes) and lowly (669 genes) expressed genes against a periodic nucleosome repeat length for A) our experimental method and B) combined predicted coverage. Autocorrelation scores for highly (red) and lowly (blue) expressed genes derived from C) MNase-seq experimental data and D) our coverage from our full prediction (see Methods).

While the correlation between the architecture of the nucleosome array and gene activity is clear, the causality link is not so obvious. To clarify this point, we inserted an innocuous 81-nt sequence in a linker region approximately at the middle of the coding sequence of 8 non-essential genes (4 genes with phased nucleosomes and 4 with unphased nucleosomes). The insert was placed in a linker region to avoid direct interference with specific nucleosomes (see Figure 6A and Suppl. Figure S11). Additionally, we tested the effect on nucleosome positioning and gene expression. For technical reasons, we only modified 2 genes per strain so we built 4 strains in total, with one phased and one unphased gene modified per strain (see Fig 6, Table 1 and Methods). In principle, we could expect three different scenarios: i) a displacement of the +1 and –last nucleosomes to recover the original phasing; ii) a coordinated small displacement of all nucleosomes to recover phasing; and iii) an increase in the fuzziness of the nucleosome string. Results in Table 2, Figure 6B and 6C, Suppl. Figures S12, S13 and S14 demonstrate that the introduction of the DNA segment in the phase genes leads to the generation of fuzzier nucleosome arrays, but neither to a significant change in the placement of the +1 and –last nucleosomes, nor to a coordinate sliding of nucleosomes. This suggests that irrespectively of other nucleosomes, the +1 and –last are placed in well-defined regions marked by sequence-dependent intrinsic and extrinsic factors, while the rest of the nucleosomes are placed based on periodicity considerations as described above.

	Unmodified strain		Strain with the 81-nt insert	
Gene	**DFI**	**R**	**DFI**	**R**
UBX5	7	0.79424	79	0.76687
CKB2	1	0.81862	79	0.68771
PPT1	13	0.88553	71	0.82500
TRP4	6	0.80702	80	0.73771
BSP1	77	0.73841	7	0.77164
DGK1	46	0.69007	31	0.81921
SLM3	37	0.61947	42	0.60019
PAN5	44	0.73341	46	0.81708

Table 2. Phase score (DFI) and autocorrelation (R) in the unmodified strain and in the strain with the 81-nt insertion in the selected genes. Genes UBX5, CKB2, PPT1 and TRP4 are phased, and genes BSP1, DGK1, SLM3 and PAN5 are unphased in the unmodified strain.



Gene name	Strain number	Number of nucleosomes from TSS to insertion site	Number of nucleosomes from insertion site to TTS
UBX5	1	5	5
CKB2	3	4	4
PPT1	2	4	6
TRP4	4	4	4
BSP1	1	5	6
DGK1	3	2	6
SLM3	2	3	5

Figure 6. A) Schematic representation of the experimental design. The exact location of the 81-nt insertion for each of the 8 genes is indicated in Table 1 and represented in Supplementary Figures S13 and S14. B) and C) Nucleosome coverage for the selected genes in the unmodified strain (top panels) and the strain with the 81-nt insertion (bottom panels). Average of B) the four genes phased in the unmodified strain (UBX5, CKB2, PPT1 and TRP4)

(blue line) and C) the four genes unphased in the unmodified strain (BSP1, DGK1, SLM3 and PAN5) (orange line).

Interestingly, analysis of mRNA levels for the 8 modified genes with or without the 81-nt segment only showed significant changes for 2 of the 8 genes (a 2.2 $\log_2$ fold change for PPT1 transcript, and a mild decrease for CKB2 transcript) (Suppl. Figure S15) suggesting that changes in nucleosome architecture do not necessarily lead to changes in expression. In order to investigate this further, the treatment with 1,10-pt was performed in conditions that lead to a decrease in RNAPII signals in ChIP-seq experiments(48). Inhibition of transcription by treatment with 1,10 phenanthroline (1,10-pt) led to an increased fuzziness of the nucleosome array and a slight displacement of the +1 and (specially) the –last nucleosomes (Figure 7A-B, Table 3 and Suppl. Figure S16), with a significant decrease of the autocorrelation score observed in the 4 strains tested (Figure 7C and Suppl. Table S3). However, to discard an artefact caused by the use of 1,10-pt, we repeated the analysis using data from a previously published work on transcription and nucleosome positioning (49). We observed that inactivation of RNAPII using the temperature-sensitive (ts) allele rpb1-1 causes an increase in nucleosome fuzziness and leads to a wider NFR at the TSS due mostly to a shift of the -1 nucleosome, in agreement with our results obtained with 1,10-pt (see Table 3 and Suppl. Figure S17).

In conclusion, we can outline a preferential (most likely not unique) causality arrow: expression activity ? changes in nucleosome architecture, with a direct role of polymerase or the elongation complex in reinforcing positioning signals within the gene body.

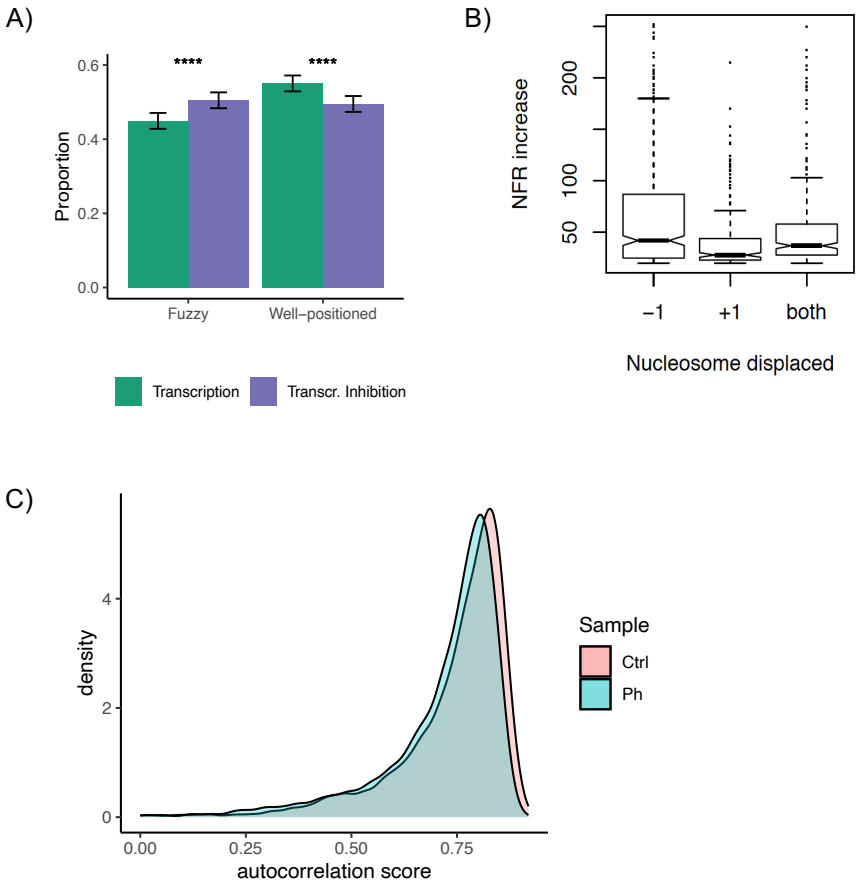


Figure 7. Effect of transcription on nucleosome positioning. A) Change in the proportion of Fuzzy and Well-positioned nucleosomes upon transcription inhibition, with bars indicating relative standard error. B) Change in NFRs' width at the TSS (-1 to +1 nucleosome distance) upon transcription inhibition in the presence of 1,10-Phenantroline (only cases with significant displacements (> 20 bps) are considered in the box plots). C) Mean autocorrelation scores of control (Ctrl) and phenanthroline (Ph) samples for the 4 strains previously described (see Figure 6A and Table 1) and for all the genes having a well-defined +1 and a -last nucleosome.

NFR width mean with std deviation (bp)	1,10-Phenantroline inhibition	ts-allele inhibition
Transcription (1)	242.5 ± 3.9	196.5± 65
Transcription Inhibition (2)	251.1 ± 1.2	214.5 ± 69
Difference (2-1)	8.6 ± 4.6	18.0 ± 73

Table 3. NFRs' width increase at the TSS for genes displacing -1, +1 or both nucleosomes upon transcription inhibition from our 1,10-Phenantroline inhibition and published work on the inactivation of RNAPII by a ts-allele (49).

Finally, we benchmarked our predictor to the state-of-the-art NuPoP (50) method, taking as reference our experimental MNase nucleosome maps (grey profiles in Suppl. Figure 18). The profile calculated using our full predictor (blue profile in Suppl. Figure 18) could position with high accuracy the +1 and –last nucleosomes and the periodicity in the middle of the genes, the largest differences being in the definition of the intensity of the peaks. NuPoP (green profile in Suppl. Figure 18) while being quite accurate, has more problems in the localization of the +1 nucleosome and generates a very fuzzy and irregular profile, overpopulating some regions of nucleosomes, while fully depleting others, and leading to flat peaks that do not correspond to the read distributions found experimentally.
While our model reproduces with high accuracy the nucleosome architecture in yeast, other elements come in to play when understanding more complex organisms such as mammalian genome architectures. We performed a similar analysis to predict the probability of having a NFR in human data using annotated TFBS from the UCSC Genome Browser (51) and we calculated the deformation energy. The model was trained and tested on data from a single chromosome (chrI) obtaining an AUC of almost 70% (see Suppl. Figure S19) outperforming what would be expected from random.

Discussion

Nucleosome positioning in the gene body can be predicted with good accuracy by signal transmission theory (STT), assuming the existence of two well-positioned nucleosomes at the +1 and –last positions, which emit periodic signals whose intensities decay with distance. Phased genes (i.e., those whose distances between the +1 and -last is a multiple of 165) have periodic nucleosome signals, while unphased genes (and at a lower extend non-phased genes) tend to have fuzzy nucleosomes in the middle of the gene body. Change in distance between the +1 and -last leads to changes in nucleosome periodicity fully predictable by the theory and in the fuzziness of nucleosomes in the middle of the gene. Very interestingly, the placement of the +1 and –last nucleosomes can be defined by the vicinity of NFRs, i.e., segments of DNA depleted of nucleosomes, which in turn, can be predicted by a simple neural network considering physical descriptors of DNA and TFBS densities. This is the case for both tandem and convergent genes, where the overall effects of neighboring genes are still captured by the model even though we consider each gene independently. Future models can investigate if a more specific gene level model that takes into account neighboring effects/cooperativity would benefit from a substantial increase in accuracy or contrary, not compensate for the increase in complexity. The combination of our machine learning algorithm to localize the NFR, and consequently the +1 and -last nucleosomes, with the emission of two periodic signals on opposite direction allows us to predict the ground state positioning of the nucleosomes through the gene body with an accuracy similar to the experimental noise associated to MNAse-seq. This suggests that nucleosome positions are quite well defined in the absence of complex mechanisms involving chromatin remodelers. Obviously, this "ground state" nucleosome architecture can be modified to satisfy cellular needs by a myriad of factors, including among others, epigenetic signals, effector proteins or remodelers. However, the predictive power of the G1-trained model is maintained for nucleosome architectures detected in phase S and M, suggesting the existence of a basal nucleosome configuration, which can be modified to adapt to the cellular needs.

When we applied our model to the human genome, we obtained more accurate results than what would be expected from a random model. Considering the additional layers to deconvolute when studying more complex organisms and that the current model and architecture was optimized for yeast, our current approach with a fine tuning methodology and additional experimental data, shows the potential to be used to study any nucleosome positioning array.

Results presented here show the existence of a clear connection between expression level and the organization of nucleosome arrays, but while changes in nucleosome phasing do not lead to alteration in gene activity, transcription inhibition results in loss of order in the nucleosome string. Thus, our results suggest a causal order expression level nucleosome architecture, with a role of RNA polymerase, and/or the transcription elongation complex, in refining nucleosome strings that goes beyond unfolding nucleosomes (51), supporting the mechanism of pausing and histone transfer suggested by recent cryoEM studies of RNA polymerase II and the histone chaperone FACT (FAcilitates Chromatin Transcription)(52–54).

Data Availability

All relevant data supporting the key findings of this study are available within the article and the Supplementary Information. The datasets generated and/or analyzed during the current study are available in the ArrayExpress repository under the following accession number EMTAB-13613 and in GEO under GSE255857. For a detailed description of the generated data see Suppl. Table S4. The code for the subsequent analysis is available in the GitHub repository: https://github.com/Jalbiti/NucleosomePeriodicity.

Acknowledgements

This work was supported by the Center of Excellence for HPC H2020 European Commission; "Bi oExcel2– Centre of Excellence for Computational Biomolecular Research" [823830]; BioExcel-3: Centre of Excellence for Computational Biomolecular Research [European Union: 101093290; Ministerio de Ciencia e Innovación: PCI2022-134976-2]; Spanish Ministry of Science [PID2021-122478NB-I00]; Instituto de Salud Carlos III–Institu to Nacional de Bioinformatica, Fondo Europeo de Desarrollo Regional [ISCIII PT 17/0009/0007]; European

Regional Development Fund, ERFD Operative Programme for Catalunya, the Catalan Government AGAUR [SGR2021 00863]. The IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from the MINECO. M.O. is an ICREA Academy scholar.

Competing interests

The authors declare no competing interests.

References

1. Richmond,T.J. and Davey,C.A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145–150.
2. Izzo,A., Kamieniarz,K. and Schneider,R. (2008) The histone H1 family: specific members, specific functions? *bchm*, **389**, 333–343.
3. Yuan,G.-C., Liu,Y.-J., Dion,M.F., Slack,M.D., Wu,L.F., Altschuler,S.J. and Rando,O.J. (2005) Genome-Scale Identification of Nucleosome Positions in S. cerevisiae. *Science (1979)*, **309**, 626–630.
4. Mavrich,T.N., Ioshikhes,I.P., Venters,B.J., Jiang,C., Tomsho,L.P., Qi,J., Schuster,S.C., Albert,I. and Pugh,B.F. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res*, **18**, 1073–1083.
5. Mavrich,T.N., Jiang,C., Ioshikhes,I.P., Li,X., Venters,B.J., Zanton,S.J., Tomsho,L.P., Qi,J., Glaser,R.L., Schuster,S.C., *et al.* (2008) Nucleosome organization in the Drosophila genome. *Nature*, **453**, 358–362.
6. Shivaswamy,S., Bhinge,A., Zhao,Y., Jones,S., Hirst,M. and Iyer,V.R. (2008) Dynamic Remodeling of Individual Nucleosomes Across a Eukaryotic Genome in Response to Transcriptional Perturbation. *PLoS Biol*, **6**, e65.
7. Vaillant,C., Palmeira,L., Chevereau,G., Audit,B., D'Aubenton-Carafa,Y., Thermes,C. and Arneodo,A. (2010) A novel strategy of transcription regulation by intragenic nucleosome ordering. *Genome Res*, **20**, 59–67.
8. Valouev,A., Johnson,S.M., Boyd,S.D., Smith,C.L., Fire,A.Z. and Sidow,A. (2011) Determinants of nucleosome organization in primary human cells. *Nature*, **474**, 516–520.
9. Baldi,S., Krebs,S., Blum,H. and Becker,P.B. (2018) Genome-wide measurement of local nucleosome array regularity and spacing by nanopore sequencing. *Nat Struct Mol Biol*, **25**, 894–901.
10. Deniz,Ö., Flores,O., Battistini,F., Pérez,A., Soler-López,M. and Orozco,M. (2011) Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC Genomics*, **12**, 489.
11. Deniz,Ö., Flores,O., Aldea,M., Soler-López,M. and Orozco,M. (2016) Nucleosome architecture throughout the cell cycle. *Sci Rep*, **6**, 19729.
12. Weiner,A., Hughes,A., Yassour,M., Rando,O.J. and Friedman,N. (2010) High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res*, **20**, 90–100.

13. Kaplan,N., Moore,I.K., Fondufe-Mittendorf,Y., Gossett,A.J., Tillo,D., Field,Y., LeProust,E.M., Hughes,T.R., Lieb,J.D., Widom,J., *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.
14. Nocetti,N. and Whitehouse,I. (2016) Nucleosome repositioning underlies dynamic gene expression. *Genes Dev*, **30**, 660–672.
15. Jiang,C. and Pugh,B.F. (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet*, **10**, 161–172.
16. Segal,E. and Widom,J. (2009) What controls nucleosome positions? *Trends in Genetics*, **25**, 335–343.
17. Clark,D.J. (2010) Nucleosome Positioning, Nucleosome Spacing and the Nucleosome Code. *J Biomol Struct Dyn*, **27**, 781–793.
18. Struhl,K. and Segal,E. (2013) Determinants of nucleosome positioning. *Nat Struct Mol Biol*, **20**, 267–273.
19. Lieleg,C., Krietenstein,N., Walker,M. and Korber,P. (2015) Nucleosome positioning in yeasts: methods, maps, and mechanisms. *Chromosoma*, **124**, 131–151.
20. Chereji,R. V and Clark,D.J. (2018) Major Determinants of Nucleosome Positioning. *Biophys J*, **114**, 2279–2289.
21. Suter,B. (2000) Poly(dAmiddle dotdT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. *Nucleic Acids Res*, **28**, 4083–4089.
22. Zhang,Y., Moqtaderi,Z., Rattner,B.P., Euskirchen,G., Snyder,M., Kadonaga,J.T., Liu,X.S. and Struhl,K. (2009) Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat Struct Mol Biol*, **16**, 847–852.
23. Hughes,A.L., Jin,Y., Rando,O.J. and Struhl,K. (2012) A Functional Evolutionary Approach to Identify Determinants of Nucleosome Positioning: A Unifying Model for Establishing the Genome-wide Pattern. *Mol Cell*, **48**, 5–15.
24. Lorch,Y., Maier-Davis,B. and Kornberg,R.D. (2014) Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions. *Genes Dev*, **28**, 2492–2497.
25. Kubik,S., Bruzzone,M.J., Challal,D., Dreos,R., Mattarocci,S., Bucher,P., Libri,D. and Shore,D. (2019) Opposing chromatin remodelers control transcription initiation frequency and start site selection. *Nat Struct Mol Biol*, **26**, 744–754.
26. Krietenstein,N., Wal,M., Watanabe,S., Park,B., Peterson,C.L., Pugh,B.F. and Korber,P. (2016) Genomic Nucleosome Organization Reconstituted with Pure Proteins. *Cell*, **167**, 709–721.e12.
27. Zhang,Z., Wippo,C.J., Wal,M., Ward,E., Korber,P. and Pugh,B.F. (2011) A Packing Mechanism for Nucleosome Organization Reconstituted Across a Eukaryotic Genome. *Science (1979)*, **332**, 977–980.
28. Zhang,Z. and Pugh,B.F. (2011) High-Resolution Genome-wide Mapping of the Primary Structure of Chromatin. *Cell*, **144**, 175–186.
29. Dans,P.D., Balaceanu,A., Pasi,M., Patelli,A.S., Petkevičiūtė,D., Walther,J., Hospital,A., Bayarri,G., Lavery,R., Maddocks,J.H., *et al.* (2019) The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. *Nucleic Acids Res*, **47**, 11090–11102.
30. Walther,J., Dans,P.D., Balaceanu,A., Hospital,A., Bayarri,G. and Orozco,M. (2020) A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Res*, **48**, e29–e29.

31. Pachkov,M., Balwierz,P.J., Arnold,P., Ozonov,E. and van Nimwegen,E. (2012) SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res*, **41**, D214–D220.
32. Bai,L. and Morozov,A. V. (2010) Gene regulation by nucleosome positioning. *Trends in Genetics*, **26**, 476–483.
33. Storici,F. and Resnick,M.A. (2006) The Delitto Perfetto Approach to In Vivo Site-Directed Mutagenesis and Chromosome Rearrangements with Synthetic Oligonucleotides in Yeast. In.pp. 329–345.
34. Grigull,J., Mnaimneh,S., Pootoolal,J., Robinson,M.D. and Hughes,T.R. (2004) Genome-Wide Analysis of mRNA Stability Using Transcription Inhibitors and Microarrays Reveals Posttranscriptional Control of Ribosome Biogenesis Factors. *Mol Cell Biol*, **24**, 5534–5547.
35. Schlenstedt,G., Hurt,E., Doye,V. and Silver,P.A. (1993) Reconstitution of nuclear protein transport with semi-intact yeast cells. *J Cell Biol*, **123**, 785–798.
36. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.
37. Flores,O. and Orozco,M. (2011) nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics*, **27**, 2149–2150.
38. Buitrago,D., Codó,L., Illa,R., de Jorge,P., Battistini,F., Flores,O., Bayarri,G., Royo,R., Del Pino,M., Heath,S., *et al.* (2019) Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning. *Nucleic Acids Res*, **47**, 9511–9523.
39. Flores,O., Deniz,Ö., Soler-López,M. and Orozco,M. (2014) Fuzziness and noise in nucleosomal architecture. *Nucleic Acids Res*, **42**, 4934–4946.
40. Wan,J., Lin,J., Zack,D.J. and Qian,J. (2009) Relating periodicity of nucleosome organization and gene regulation. *Bioinformatics*, **25**, 1782–1788.
41. Chollet,F. (2015) keras. *Github*.
42. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Müller,A., Nothman,J., Louppe,G., *et al.* (2012) Scikit-learn: Machine Learning in Python. https://doi.org/10.48550/arXiv.1201.0490.
43. Abadi,M., Barham,P., Chen,J., Chen,Z., Davis,A., Dean,J., Devin,M., Ghemawat,S., Irving,G., Isard,M., *et al.* (2016) TensorFlow: A system for large-scale machine learning. https://doi.org/10.48550/arXiv.1605.08695.
44. Chen,K., Xi,Y., Pan,X., Li,Z., Kaestner,K., Tyler,J., Dent,S., He,X. and Li,W. (2013) DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res*, **23**.
45. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J., *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*, **17**, 261–272.
46. Ocampo,J., Chereji,R. V, Eriksson,P.R. and Clark,D.J. (2016) The ISW1 and CHD1 ATP-dependent chromatin remodelers compete to set nucleosome spacing in vivo. *Nucleic Acids Res*, **44**, 4625–4635.
47. Buitrago,D., Labrador,M., Arcon,J.P., Lema,R., Flores,O., Esteve-Codina,A., Blanc,J., Villegas,N., Bellido,D., Gut,M., *et al.* (2021) Impact of DNA methylation on 3D genome structure. *Nat Commun*, **12**, 3243.

48. Martin,B.J.E., Brind'Amour,J., Kuzmin,A., Jensen,K.N., Liu,Z.C., Lorincz,M. and Howe,L.J. (2021) Transcription shapes genome-wide histone acetylation patterns. *Nat Commun*, **12**, 210.
49. Weiner,A., Hughes,A., Yassour,M., Rando,O.J. and Friedman,N. (2010) High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res*, **20**, 90–100.
50. Wang,J.-P., Fondufe-Mittendorf,Y., Xi,L., Tsai,G.-F., Segal,E. and Widom,J. (2008) Preferentially Quantized Linker DNA Lengths in Saccharomyces cerevisiae. *PLoS Comput Biol*, **4**, e1000175.
51. Nassar,L.R., Barber,G.P., Benet-Pagès,A., Casper,J., Clawson,H., Diekhans,M., Fischer,C., Gonzalez,J.N., Hinrichs,A.S., Lee,B.T., *et al.* (2023) The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res*, **51**.
52. Schwabish,M.A. and Struhl,K. (2004) Evidence for Eviction and Rapid Deposition of Histones upon Transcriptional Elongation by RNA Polymerase II. *Mol Cell Biol*, **24**, 10111–10117.
53. Žumer,K., Maier,K.C., Farnung,L., Jaeger,M.G., Rus,P., Winter,G. and Cramer,P. (2021) Two distinct mechanisms of RNA polymerase II elongation stimulation in vivo. *Mol Cell*, **81**, 3096–3109.e8.
54. Kujirai,T., Ehara,H., Fujino,Y., Shirouzu,M., Sekine,S. and Kurumizaka,H. (2018) Structural basis of the nucleosome transition during RNA polymerase II passage. *Science (1979)*, **362**, 595–598.
55. Kujirai,T., Ehara,H., Sekine,S. and Kurumizaka,H. (2023) Structural Transition of the Nucleosome during Transcription Elongation. *Cells*, **12**, 1388.

Chapter 3. The effect of Oxidative Stress Damage on Chromatin

Chapters 1 and 2 explain the first two projects of this thesis, where we try to deconvolute gene regulatory networks at the protein-DNA interaction level and introduce nucleosomes through exploring their role in regulating gene expression. In this third chapter we introduce another layer of complexity, that affects chromatin structure and dynamics and hinders chromatin-templated processes such as DNA replication or gene regulation: DNA damage.

It is well characterized that agents that alter the structure of DNA, can severely affect the physiological processes of cells, but the changes induced by these agents in chromatin structure are unknown. In order to clarify the changes that chromatin undergoes upon DNA lesions, we have studied the impact of oxidative stress, one of the most common origins of DNA damage. Efficient DNA repair requires coordinated chromatin dynamics and while many steps in repair pathways are well characterized, how they affect chromatin structure is unclear. This is the objective of Chapter 3, where we explore the 2D nucleosome organization (see Introduction on nucleosome positioning for the methodologies) as well as the 3D structure of the chromatin (from the nucleosome fiber to the entire chromatin), when it is subjected to oxidative stress conditions.

3.1. An overview of oxidative stress

Reactive oxygen species (ROS) are generated when single electrons are transferred to oxygen and produce superoxide radical ($O_2^{\bullet -}$), hydrogen peroxide (H_2O_2) or the highly reactive and damaging hydroxyl radical ($^{\bullet}OH$). The accumulation of ROS leads to a disruption of signaling pathways that play a key role in cell stability (1–5). Mechanistically, ROS can affect lipids, proteins, lipoproteins and DNA, damaging membranes, and organelles (6, 7). The effect upon DNA and RNA can introduce severe damages that under normal conditions are prevented by antioxidant systems that scavenge additional levels of oxidants. ROS-induced DNA damage includes base modifications (primarily 8-oxoG), apurinic site formation, and in severe cases, DNA

strand breaks (single or double) leading to genome instability and problems that can extent to the systemic level.

3.2. DNA repair mechanisms

The specificity that characterizes the many distinct DNA lesions that occur, requires the existence of dedicated DNA repair pathways. Some lesions such as oxidative stress or alkylating agents directly modify DNA bases. Others, such as UV light, can result in the crosslinking of neighboring bases while ionizing radiation can generate the most difficult lesions to repair, double-stranded breaks. This means that repairing mechanisms should be diverse in order to cover the complete portfolio of lesions. Overall, the main DNA damage repair mechanisms can be divided into three categories: base excision repair, nucleotide excision repair and double-stranded break repair mechanisms (see Figure 3.1).

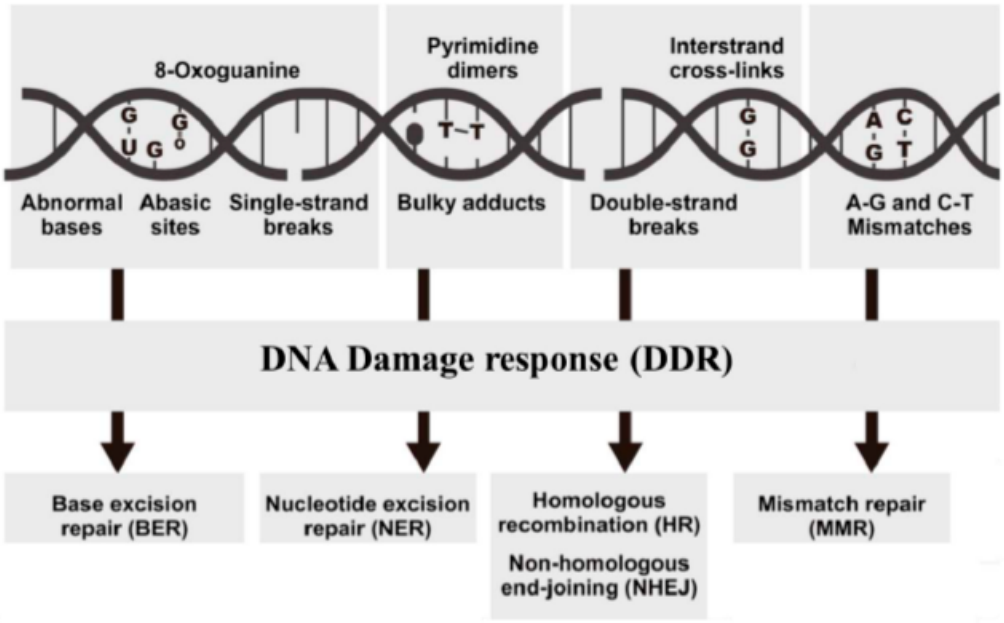


Figure 3.1. Schematic view of the DNA damage responses upon DNA lesions (top) and their associated repair pathways (bottom). Adapted from (8).

3.2.1. Base excision repair

The base excision repair (BER) mechanism specifically corrects DNA damage from oxidation, deamination and alkylation which cause small base lesions that do not significantly alter the DNA helix structure (9), but that can trigger mutations with cellular and systemic

impact. The process starts when a DNA glycosylase recognizes and excises a damaged base, leaving a base free site that is further processed by patch repair mechanisms. The short-patch mechanism (also called single-nucleotide BER) fills and ligates a single nucleotide gap whereas the long-patch mechanism involves the repair of at least two nucleotides. When inducing oxidative stress upon the cells, DNA base lesions that cause minor structural changes are mainly repaired by the BER mechanism.

3.2.2. Nucleotide excision repair

Bulky DNA lesions, which can distort the DNA helix, such as those generated by UV radiation or environmental mutagens, are repaired through a process called nucleotide excision repair (NER). Eukaryotic cells have developed a specific pathway where a nuclease enzyme recognizes and removes a segment of DNA containing the lesion (10). NER can occur through two different pathways: one which can occur anywhere in the genome (known as the global genome NER or GG-NER) and a second one responsible for the accelerated repair of lesions in transcribed regions (the transcription-coupled NER or TC-NER). The initiation of the repair pathway is thus dependent on the position of the lesion in the genome. Nonetheless, both pathways require the core NER factors to undertake the excision mechanisms.

3.2.3. Double-stranded break repair

Double-stranded breaks can be caused by the exposure to exogenous factors such as ionizing radiation or endogenous events like oxidative stress. These damage events comprehend different signaling pathways that try to repair the strand split. There exist two distinct pathways involved in this repair: the Homologous Recombination Repair (HR) and the Non-Homologous End Joining (NHEJ) mechanism. The HR mechanism copies the information from the DNA and restores the complementary sequence whereas the NHEJ repairs the free DNA ends by gluing them back together.

3.3. The 3D Study of Chromatin

In order to investigate the chromatin dynamics upon DNA damage we introduce a new dimension of study: the 3D study of the chromatin structure and how it changes upon oxidative stress. To this end we use two variants of Chromosome Conformation Capture (3C)

techniques named HiC and microC, by which we can quantify the frequency of interactions between pairs of loci by crosslinking, DNA fragmentation and the ligation of proximal ends (11–14).

3.3.1. The experimental data

3.3.1.1. Hi-C

Hi-C consists in the quantification of genome-wide contacts (15), derived from the Chromosome Conformation Capture (3C) technique. In this technique a sample gets cross-linked with formaldehyde and posteriorly fragmented with a restriction enzyme. Next, the fragmented ends are filled with nucleotides marked with biotin in order to facilitate the posterior quantification of interactions. The proximal fragments are ligated forming chimeric molecules that contain the two regions that were previously cross-linked. Finally, the DNA is purified and sheared by sonification before getting sequenced (see Figure 3.2). Finally, the reads are mapped into the genome, with chimeric segments signaling the presence of close contacts between two regions of the genome.

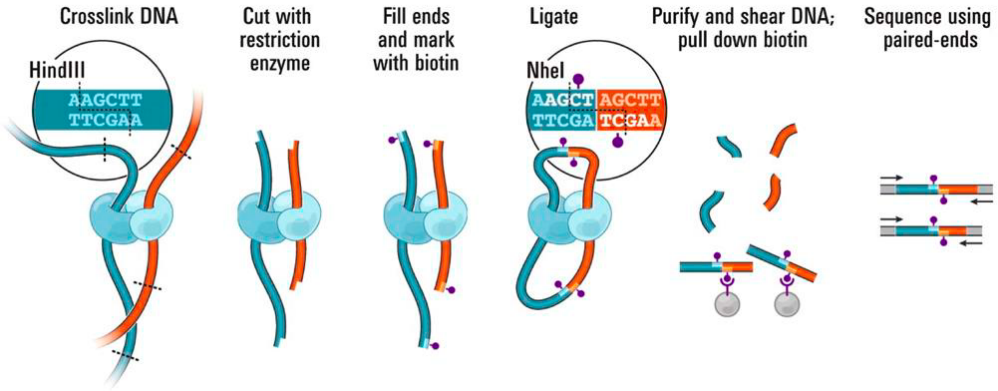


Figure 3.2. Scheme of a Hi-C experiment adapted from (15).

Capture Hi-C is a variation of the Hi-C protocol but instead of computing all genome-wide contacts it restricts the analysis to those between regions targeted by designed probes (16, 17). In this technique we can design a pool of primers to selectively purify different regions and enrich Hi-C ligation product libraries. This reduces the cost and allows an increase in coverage.

3.3.1.2. Micro-C

Micro-C is another 3C-based technique used to quantify genome-wide contacts. The difference resides in the level of resolution, given that Micro-C allows us to obtain the contacts at a nucleosome level resolution. In this method we cleave DNA by using MNase instead of a restriction enzyme. Linker DNA is preferentially fragmented, and the chimeric molecules get sequenced to define two nucleosomes that are close in space. This technique can utilize crosslinkers of different sizes to capture proximal and distal nucleosome contacts.

3.3.2. Hi-C and Micro-C data processing

In order to study the 3D structure of chromatin we processed our obtained Hi-C and Micro-C data using TADbit (18) and HiC-Pro (19). We performed the corresponding quality control, mapping and filtering of our sequencing data following the previously published protocols. For both types of data, non-informative contacts (self-circle, dangling-end, error, duplicated and random breaks) were identified and filtered out. Lastly, following the mentioned processing steps, we generated contact matrices from the reads at different resolutions which allowed us to infer the changes in the chromatin conformation and visualize those using available tools (20–22).

When comparing two conditions we are interested in quantifying the regions where the most prominent changes in contacts appear. To do so, we performed a differential analysis of both Hi-C and Micro-C data for different conditions using the R Bioconductor package diffHic (23). Additionally, we investigated the differences in chromosomal interaction domains (CIDs) and computed the insulation scores (24, 25) to assess the differences in DNA packaging and its organization into domains (i.e., regions which are more likely to self-interact). Lastly, the 3D study of the chromatin was complemented with a chromatin fiber coarse-grained model (see Introduction) to further study the conformation changes under oxidative stress.

In this chapter, we studied the response to oxidative stress from a low- to high-order chromatin structure point of view. Altogether, our 2D and 3D analyses of chromatin structure enhance our understanding of the changes that chromatin undergoes during one of the most common DNA damage lesions. We showed that chromatin experiences complex structural changes under oxidative stress with a gain of interactions at very short distances (< 600 bp), a loss of interactions at

distances between 600 bp and 15 kb and then a gain of interactions at large distances (> 15 kb). Simultaneously, this was shown by the chromatin fiber coarse-grained model where nucleosome clutches were enhanced and reduced in size correlating with the increased number of CIDs that presented smaller sizes under stress. We also observed an increase in fuzziness in the overall nucleosome architecture with a stronger effect in upregulated genes. In line with this finding, our study revealed a loss of periodicity in the nucleosome arrangement and a slight redistribution of the NFR length at the TSS with a higher proportion of shorter NFRs in oxidative stress samples. These results agree with our previously presented results where higher fuzziness was linked to nucleosome architectures with less periodic profiles, which might be linked to the need to have more dynamic nucleosomes in order to allow repair mechanisms to access damaged regions. Moreover, the different results at various levels of chromatin dynamics could reflect the different type of DNA lesions caused by oxidative stress that activate different repair pathways. Base lesions are repaired essentially by the Base Excision Repair (BER) pathway (26) and Nucleotide Excision Repair (NER) while Non Homologous End Joining (NHEJ) and Homologous Recombination (HR) pathways are triggered when single and double strand breaks occur (27).

Examining chromatin behavior in both 2D and 3D provides a broader insight into the mechanisms regulating chromatin organization within the genome. Furthermore, this helps clarify the significant role that chromatin plays in biological processes and gene expression mechanisms.

Publication:

Juan Pablo Arcon*, Rafael Lema*, Adrià Caballe, Diana Buitrago, Alba Sala, José Gabriel Álvarez-Meythaler, Nuria Villegas, Julie Blanc, Oscar Reina, Marta Gut, Pablo D. Dans, Camille Stephan Otto Attolini, Isabelle Brun Heath, Modesto Orozco. Effect of oxidative stress on 3D genome structure (*in preparation*).

*Equally contributing authors
Supplementary material for this article can be found in the Annex.

References

1. Jones,D.P. and Sies,H. (2015) The Redox Code. Antioxid Redox Signal, 23.
2. Sies,H. (1986) Biochemistry of Oxidative Stress. Angewandte Chemie International Edition in English, 25.
3. Sies,H. and Jones,D.P. (2020) Reactive oxygen species (ROS) as pleiotropic physiological signalling agents. Nat Rev Mol Cell Biol, 21.
4. Sies,H. (1997) Oxidative stress: Oxidants and antioxidants. Exp Physiol, 82.
5. Sies,H. (2016) Oxidative stress: impact in redox biology and medicine. Archives of Medical and Biomedical Research, 2.
6. Schieber,M. and Chandel,N.S. (2014) ROS function in redox signaling and oxidative stress. Current Biology, 24.
7. Reichmann,D., Voth,W. and Jakob,U. (2018) Maintaining a Healthy Proteome during Oxidative Stress. Mol Cell, 69.
8. Tasaki,E., Mitaka,Y., Nozaki,T., Kobayashi,K., Matsuura,K. and Iuchi,Y. (2018) High expression of the breast cancer susceptibility gene BRCA1 in long-lived termite kings. Aging, 10.
9. Krokan,H.E. and Bjørås,M. (2013) Base excision repair. Cold Spring Harb Perspect Biol, 5.
10. Schärer,O.D. (2013) Nucleotide excision repair in Eukaryotes. Cold Spring Harb Perspect Biol, 5.
11. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. Science (1979), 295.
12. Simonis,M., Klous,P., Splinter,E., Moshkin,Y., Willemsen,R., De Wit,E., Van Steensel,B. and De Laat,W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet, 38.
13. Zhao,Z., Tavoosidana,G., Sjölinder,M., Göndör,A., Mariano,P., Wang,S., Kanduri,C., Lezcano,M., Sandhu,K.S., Singh,U., et al. (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet, 38.
14. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C., et al. (2006) Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res, 16.
15. Lieberman-Aiden,E., Van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O.,

et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (1979), 326.
16. Hughes,J.R., Roberts,N., Mcgowan,S., Hay,D., Giannoulatou,E., Lynch,M., De Gobbi,M., Taylor,S., Gibbons,R. and Higgs,D.R. (2014) Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet, 46.
17. Schoenfelder,S., Furlan-Magaril,M., Mifsud,B., Tavares-Cadete,F., Sugar,R., Javierre,B.M., Nagano,T., Katsman,Y., Sakthidevi,M., Wingett,S.W., et al. (2015) The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res, 25.
18. Serra,F., Baù,D., Goodstadt,M., Castillo,D., Filion,G. and Marti-Renom,M.A. (2017) Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput Biol, 13.
19. Servant,N., Varoquaux,N., Lajoie,B.R., Viara,E., Chen,C.J., Vert,J.P., Heard,E., Dekker,J. and Barillot,E. (2015) HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. Genome Biol, 16.
20. Durand,N.C., Robinson,J.T., Shamim,M.S., Machol,I., Mesirov,J.P., Lander,E.S. and Aiden,E.L. (2016) Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst, 3.
21. Kerpedjiev,P., Abdennur,N., Lekschas,F., McCallum,C., Dinkla,K., Strobelt,H., Luber,J.M., Ouellette,S.B., Azhir,A., Kumar,N., et al. (2018) HiGlass: Web-based visual exploration and analysis of genome interaction maps. Genome Biol, 19.
22. Reiff,S.B., Schroeder,A.J., Kırlı,K., Cosolo,A., Bakker,C., Lee,S., Veit,A.D., Balashov,A.K., Vitzthum,C., Ronchetti,W., et al. (2022) The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. Nat Commun, 13.
23. Lun,A.T.L. and Smyth,G.K. (2015) diffHic: A Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics, 16.
24. Mizuguchi,T., Fudenberg,G., Mehta,S., Belton,J.M., Taneja,N., Folco,H.D., FitzGerald,P., Dekker,J., Mirny,L., Barrowman,J., et al. (2014) Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. Nature, 516.
25. Open2C, Abdennur,N., Abraham,S., Fudenberg,G., Flyamer,I.M., Galitsyna,A.A., Goloborodko,A., Imakaev,M., Oksuz,B.A. and Venev,S. V. (2022) Cooltools: enabling high-resolution Hi-C analysis in Python. bioRxiv.
26. Maynard,S., Schurman,S.H., Harboe,C., de Souza-Pinto,N.C. and Bohr,V.A. (2009) Base excision repair of oxidative DNA damage and association with cancer and aging. Carcinogenesis, 30.

27. Ciccia,A. and Elledge,S.J. (2010) The DNA Damage Response: Making It Safe to Play with Knives. Mol Cell, 40.

Effect of oxidative stress on 3D genome structure

Juan Pablo Arcon[1,#], Rafael Lema[1,#], Adrià Caballe[1], Diana Buitrago[1], Alba Sala[1], José Gabriel Álvarez-Meythaler[1], Nuria Villegas[1], Julie Blanc[3], Oscar Reina[1], Marta Gut[3], Pablo D. Dans[1,4], Camille Stephan Otto Attolini[1], Isabelle Brun Heath[1], Modesto Orozco[1,2].

[1]Institute for Research in Biomedicine (IRB Barcelona)-The Barcelona Institute of Science and Technology. Baldiri i Reixach 10. Barcelona 08028, Spain; [2]Departament de Bioquímica i Biomedicina, Universitat de Barcelona, Barcelona, Spain; [3]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain; [4]Department of Biological Sciences, CENUR North Coast. University of the Republic, Salto, Uruguay.

[#] Equal contribution

Abstract

The genome is constantly exposed to damaging agents which can be either extrinsic (e.g., UV light, pollution) or intrinsic (e.g., reactive oxygen species). This can lead to mutations, and eventually to disorders such as cancer if the DNA lesions are not properly detected and repaired. Recently, it has been shown that mutation signatures in cancer genomes are not randomly located suggesting an important role of the chromatin structure on either the formation of lesions or on the repair process (due to the accessibility to the damage by the repair machinery). In order to study the role of chromatin on the localization of DNA damage, we studied the impact of oxidative stress on 3D genome organization and on the chromatin fiber structure. Using Hi-C and Micro-C techniques combined with Molecular Dynamic simulations, we observed a local decondensation of the chromatin fiber with a clear rearrangement of the nucleosomes in damaged regions and in up-regulated genes. This local decondensation is associated with an increase in long range interactions that maintain the overall organization of the genome. These results suggest that nucleosomes are sufficiently dynamic to allow the repair machinery to access the damaged regions and that the overrepresentation of mutations in nucleosomal DNA is more likely the result of lesions occurring at the surface of the nucleosomes, as suggested by Wu et al. 2018, rather than because the nucleosomes are physically blocking the repair process.

Introduction

Oxidative stress is caused by an imbalance between production and accumulation of reactive oxygen species (ROS) in cells, which affects lipids, proteins, lipoproteins, and DNA, damaging cellular structures such as membranes and organelles (Schieber and Chandel 2014; Reichmann et al. 2018). Oxidative DNA damage is linked to base modifications, 8-oxoG being the most abundant one, and formation of apurinic sites (reviewed in (Dizdaroglu et al. 2002; Wallace 2002; Cooke et al. 2003)). The failure to fix these DNA lesions leads to mutagenesis and ultimately to cancer, accelerated aging and neurodegenerative disorders (Kreuz and Fischle 2016). Over the last few years, several techniques profile (OxiDIP-seq (Amente et al. 2019) is

based on immuno-precipitation using 8-oxodG antibodies, or entrap-seq (Fang and Zou 2020) OG-seq (Ding et al. 2017) , click-code-seq (Wu et al. 2018)) have been developed to map 8-oxoG (reviewed in (Mingard et al. 2020; Poetsch 2020). While most of the studies suggest an over-representation of 8-oxoG in 5'UTR, promoters, and more generally, in euchromatin, click-code-seq (Wu et al. 2018) reveals an over-representation of 8-oxoG in telomeres, in nucleosomal DNA and in region with low polII signal. Hydroxyl radicals can also induce intra-strand crosslinking, DNA single-strand breaks (SSBs) and double-strand breaks (DSBs) that lead to genome instability (Ye et al. 2016; Zhang et al. 2019). Interestingly, 8-OxoG (mapped using OxiDIP-seq) seem to strongly correlate with DSBs revealed by H2AX phosphorylation in both human and mouse epithelial cells (Amente et al. 2019). Several studies have shown that oxidative stress affects histone post translational modifications (PTM) and DNA methylation (Niu et al. 2015; Casali et al. 2022) and that several chromatin regulators are involved in the diamide stress response (Weiner et al. 2012). A systematic study of 26 histone PTM after diamide stress showed that most of the changes in histone marks correlated with RNA Pol II transcription (e.g., H3K56ac) and DNA damage (e.g. H2AS129ph), which suggest some impact in chromatin structure (Weiner et al. 2015).

Chromosome conformation capture (3C)(Dekker et al. 2002; Bohn and Heermann 2010; van Steensel and Dekker 2010; Bau and Marti-Renom 2011; Kalhor et al. 2011; Fudenberg and Mirny 2012; Hakim and Misteli 2012; Mifsud et al. 2015; Marti-Renom et al. 2018) and ultra-resolution microscopy alone (Otterstrom et al. 2019; Xu et al. 2020; Miriklis et al. 2021) or the two combined (Neguembor et al. 2022) have provided the community the possibility to look at sub-nucleosome resolution at chromatin structure and how it changes as consequence of cellular needs (Handoko et al. 2011; de Wit et al. 2015; Lupianez et al. 2015; Dehingia et al. 2022), cell cycle (Deniz et al. 2016; Lazar-Stefanita et al. 2017), senescence (Sati et al. 2020), quiescence (Swygert et al. 2019) or differentiation (Pekowska et al. 2018; Neguembor et al. 2022). However, while several risk loci have been mapped structurally (Dryden et al. 2014; Jager et al. 2015; Martin et al. 2015; Baxter et al. 2018), and some connections between nucleosomal architecture and cancer-mutations have been made (Pich et al. 2018) there is not a general picture of the way in which chromatin reacts to DNA lesions (Hauer et al. 2017).

In this paper, we combine gene expression analysis, MNase-Seq, Hi-C, Micro-C experiments with coarse-grained and data-driven models of the 3D genome structure (Buitrago et al. 2021; Neguembor et al. 2022) to explore how the changes produced by oxidative stress affect chromatin structure. We demonstrate that DNA lesions lead to different levels of condensation and decondensation of chromatin which correlated with transcriptional changes and DNA damage.

Results

Cellular response to oxidative stress

Oxidative stress (OS) (see Methods) slightly reduces cell viability (Figure 1A) and causes cells to arrest in G1 during 3 hours before they can re-enter in S phase and resume cell division (Figure 1B). To remove any cell cycle bias in our study, all experiments presented hereafter were performed comparing cells collected 30 minutes after induction of oxidative stress, with control cells synchronized in G1 (see protocol scheme in Supplementary Figure S1A). Indeed, their DNA content and cell morphology (Supplementary Figure S1B) as well as their microtubule

organization (Supplementary Figure S1C) showed that cells were in G1 phase in both samples, allowing a fair comparison of data.

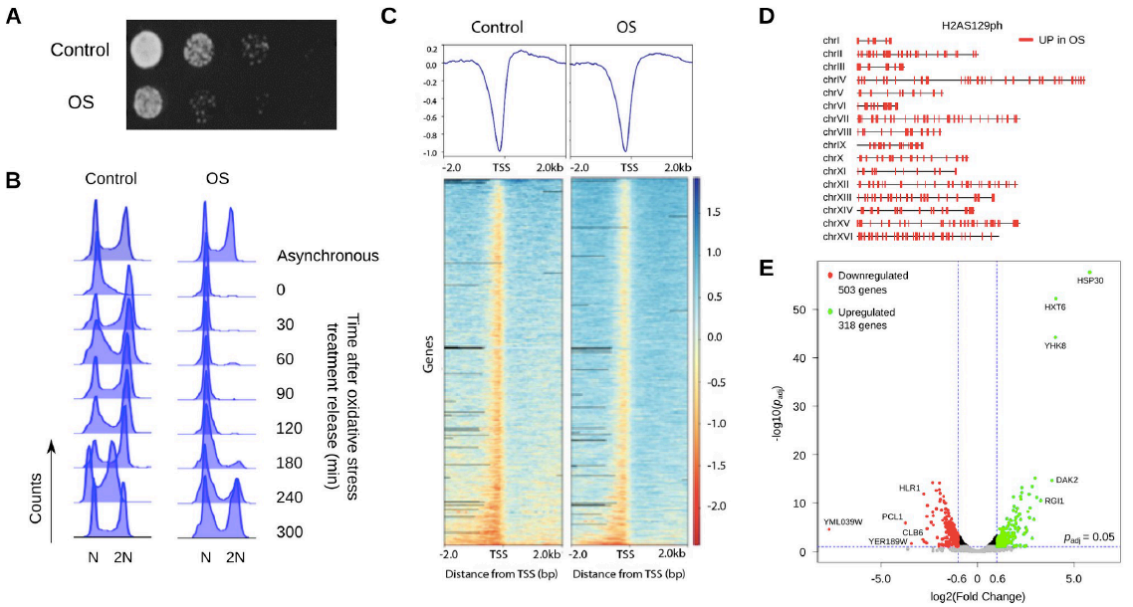


Figure 1. Cellular response to oxidative stress (OS). (A) Spot viability test under control (top row) and OS (bottom row) conditions. Cells were diluted and seeded at several concentrations (from left to right, OD600 = 10-1, 10-2, 10-3, 10-4). (B) Fluorescence intensity after flow cytometry of control (left panel) and OS (right panel) samples, collected at different time points after α-factor removal. (C) Heat map of the H2AS129ph ChIP-seq signals, centered at the transcription start site (TSS) and covering a region of +/- 2 kb. The color label indicates the level of enrichment. (D) Localization of differential (DiffBind package, $p < 0.05$) H2AS129ph peaks along the genome. No weaker signals were detected in OS with respect to control. (E) Gene expression difference between OS and control samples. RNA level difference is plotted on the x axis and the adjusted p-value (padj) on the y axis. The selected differential threshold is +/- 0.6 log2(fold change) for the RNA level and 0.05 for padj. Downregulated (N = 503) and upregulated (N = 318) genes are show in red and green, respectively. In each case, the 5 genes with the highest change are identified.

Our H_2O_2 treatment produced DNA damage, including double strand breaks (DSBs) as evidenced by the increase of H2AS129ph immunostaining signal (Supplementary Figure S1D) and of H2AS129ph ChIP-seq signal (Figure 1C) with 391 genomic regions showing a stronger signal upon OS (Figure 1D). A stronger signal of RAP1, a transcription factor involved in telomere maintenance and chromatin silencing, is also observed suggesting some lesions at telomeres, which are guanine-rich and therefore more susceptible to OS (Supplementary Figure S1E). This effect is corroborated by a significant increase of H2AS129 phosphorylation at telomeres with 29 ChIP-seq peaks located on those regions (p-value = 0.0022).

Effect of oxidative stress on gene expression

Transcriptomic analysis indicates that 12% of the genes changed expression in response to oxidative stress with 318 upregulated genes that include well characterized stress response genes (e.g. HSP30, DDR2) and 503 downregulated genes, including *CLN1* and *CLN2* as expected by the arrest of the cells at the checkpoint G1/S (Figure 1E). Gene Set Enrichment Analysis (GSEA)

revealed that translation, cellular response to oxidative stress and electron transport chain pathways were upregulated; while cell cycle, DNA replication and telomere maintenance pathways were downregulated (Supplementary Figure S1F). Among all the environmental stress response genes (ESR) known to be induced by various stress agents (Gasch et al. 2000), the genes playing a role in redox regulation and described in (Jamieson 1998; Gasch et al. 2000) presented a 1.5 to 4 fold change increase of their expression. However, the expression of the transcription factor YAP1, one of the key players in the regulation of the oxidative stress response, did not increase. Previous studies showed that nuclear import of Yap1p protein is enhanced in response to oxidative stress (Kuge et al. 1997), and our results suggest that the oxidative stress response is not due to an increase in YAP1 transcription but rather an increase in the import of cytoplasmic Yap1p protein into the nucleus.

Oxidative stress affects the structure of the chromatin fiber

Apart from the -1 and +1 nucleosomes at the Transcription Start Site (TSS), and the last nucleosome at the Transcription Termination Site (TTS), nucleosome occupancy seems globally reduced under OS (Figure 2A-C), with a stronger effect in upregulated genes than in downregulated genes. Oxidative stress also leads to a global increase in nucleosome fuzziness (i.e., decrease in well-localized nucleosomes) revealed by higher fuzziness scores (Figure 2D). This increase of fuzziness is larger in upregulated genes (Figure 2E) than in downregulated genes (Figure 2F), and it is also observed in regions with stronger H2AS129ph signal (Figure 2G). In line with this finding, we also observed a loss of periodicity in the nucleosome arrangement (Supplementary Figure S2A) and a slight redistribution of the Nucleosome Free Regions (NFRs) length at the TSS with a higher proportion of shorter NFRs in OS samples (Supplementary Figure S2B).

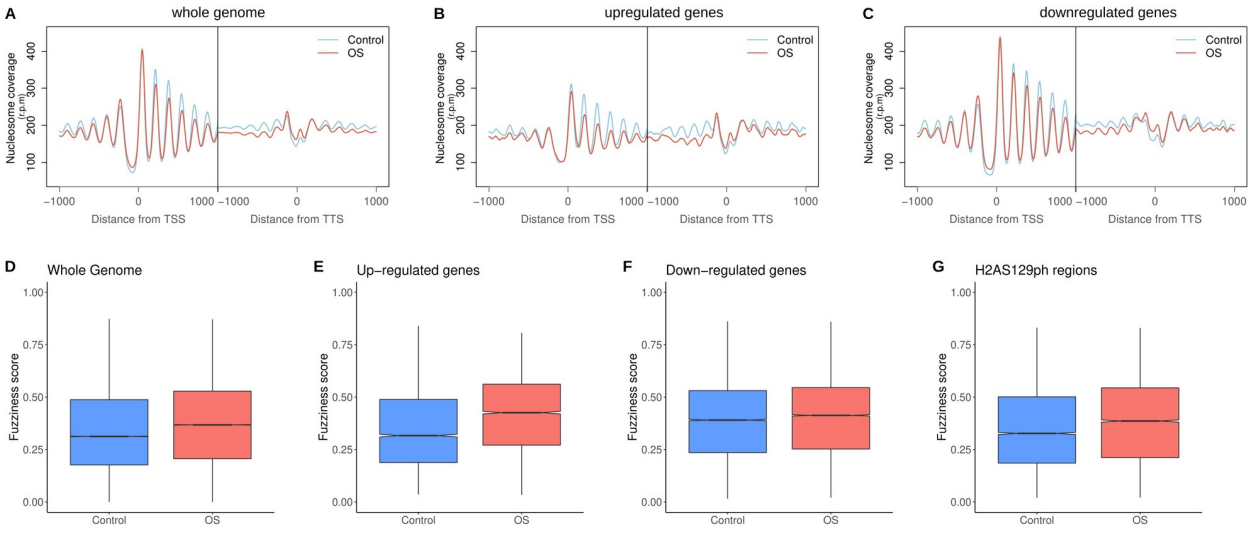


Figure 2. Nucleosome organization under oxidative stress (OS). Nucleosome positioning along the gene in the whole genome (A), and in upregulated (B) and downregulated (C) genes. Distribution of fuzziness score for nucleosomes in the whole genome (D), upregulated genes (E), downregulated genes (F) and regions with stronger H2AS129ph signal (G). Median, quartiles and extreme values are represented as box plots.

Nucleosome positioning along the DNA contributes to the structure of the chromatin fiber and the increase of fuzziness caused by oxidative stress strongly suggests an alteration of the chromatin conformation. Analysis of the differential interactions between conditions at a fine Micro-C scale (nucleosome resolution) shows that the interactions between nucleosomes occurring at very short distances (up to 600 bp) are mainly gained in OS, while interactions further apart than 600 bp are mainly lost (Figure 3A). This tendency is more evident in upregulated genes, but also observed in downregulated genes and, to a lesser extent, in damaged genes where H2AS129ph signal is enriched (Figure 3B).

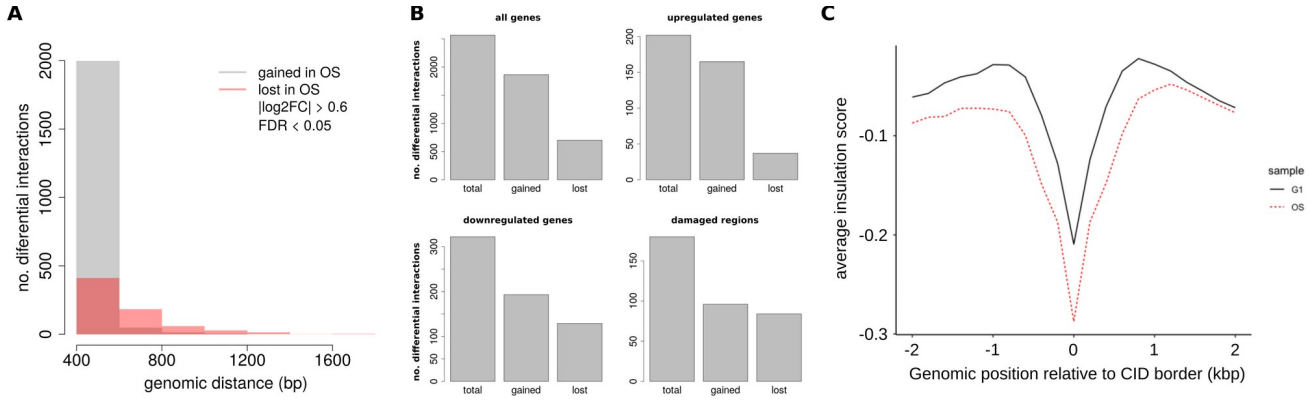


Figure 3. Micro-C differential interactions between oxidative stress and control samples. (A) Distribution of differential chromatin interactions as a function of the genomic distance. Gray bars represent interactions gained and red bars represent interactions lost in oxidative stress samples. Only interactions with counts above 10, absolute log2 fold change greater than 0.6 and false discovery rate < 0.05 were considered as differential. (B) Number of differential chromatin interactions for all, up regulated and down regulated genes, and damaged regions (with enriched H2AS129ph). (C) Insulation score in a window of 4 kb around CID borders in control (black) and oxidative stress (red) samples.

This curious OS-induced alteration of the pattern of nucleosome contacts (global increase in proximal contacts and reduction in remote ones) leads to an interesting change in chromatin interactions domains (CIDs) whose number and definition increase (Suppl. Table S1, and Figure 3C), while their size decreases (Supplementary Figure S2C). To illustrate these results with specific genes, contact maps for three representative chromatin regions containing the upregulated gene *APJ1,* the downregulated gene *ERG5* and the damaged region *HOL1-BIO3* are presented (Supplementary Figure S3). We observe a reduction of nucleosome interactions between the *APJ1* and the neighboring genes, associated with an increase of the insulation score on both sides of the gene, the effect being stronger at its promoter region (Supplementary Figure S3A). Nucleosomes also appear fuzzier along the gene. As a consequence, upon transcription activation, the gene *APJ1* is no longer part of a larger CID but forms a smaller CID on its own. This reorganization could facilitate transcription re-initiation to reach a higher level of expression by co-localizing the TTS and TSS closer in space, as seen in modelled 3D structures obtained with our coarse-grained model with base pair resolution (see Supplementary Figure S4). In contrast, the down-regulated gene *ERG5* has nucleosomes with similar fuzziness in both conditions, even showing in the OS sample better positioned +1/+2 nucleosomes and a smaller previous NFR. In addition, a clear re-organization of the CID including *ERG5* can be observed with reduced insulation score between neighboring regions and *ERG5* interacting with more distant genes and being included in a larger CID in the OS condition (Supplementary Figure S3B), leading

to a smaller decrease in chromatin interactions than those noted for *APJ1*. We also analyzed a representative region containing 4 genes (*HOL1* and *BIO3*, *4* and *5*) which are damaged, but whose expression is not changing (Supplementary Figure S3C). Clearly, DNA damage alone can trigger a re-organization of the chromatin, in this case with an increase in the locality of the CIDs similar to that of upregulated genes. In summary, despite global analysis suggest a general effect of oxidative stress tends increase the number of CIDs, but decreasing their size, gene variability is large and detected changes depends on the expression level and the level of damage.

Finally, MNase-Seq and Micro-C data were combined with our coarse-grained model (Neguembor et al. 2022) to simulate the chromatin fiber corresponding to different genes containing at least six nucleosomes. Figure 4 shows representative structures from the generated ensembles of a damage region including the *YER107W-A* gene, along with the contact maps resulting from the whole ensemble of structures (right) highly correlated to Micro-C experimental maps (left) for both control and OS samples. Integration of the data derived from deconvolution of MNase-seq signals (including cell variability) gives a general decrease in the number of nucleosomes in response to oxidative stress. The fibers appear more extended in the stressed samples than in the control, especially for upregulated and damaged genes (Figure 4A and Supplementary Figure S7). Interestingly, the structures from Figure 4A and the distance distributions between nucleosomes (N/N+x) from Figure 4B show that in OS samples smaller and better-defined nucleosome clutches of 3-4 nucleosomes are formed, quantified by increased overlaps between N/N+1 distance distribution and N/N+x distance distributions (for x = 2, 3, 4) detailed in Figure 4B. These results are general as evidenced by the internucleosome distance distributions and overlaps for all simulated upregulated and damaged genes (Figure 4D and Supplementary Figure S5). Representative structures for upregulated genes size displaying defined nucleosome clutches are shown on Supplementary Figures S4 and S6. This is in line with the aforementioned gain in chromatin interactions at short genomic distances (< 600 bp) and the loss of periodicity in the nucleosome arrangement across the genome (Supplementary Figure S2A). On the other hand, the extension of the chromatin fiber, arising from the loss of interactions beyond 600 bp (and until 15 kbp, see the following section) is represented by an increase in the radius of gyration observed for all upregulated genes that were simulated, as well as in genes with increased H2AS129ph signal (Supplementary Figure S7). Again, the effect is less evident or even reversed in downregulated genes (Supplementary Figure S7).

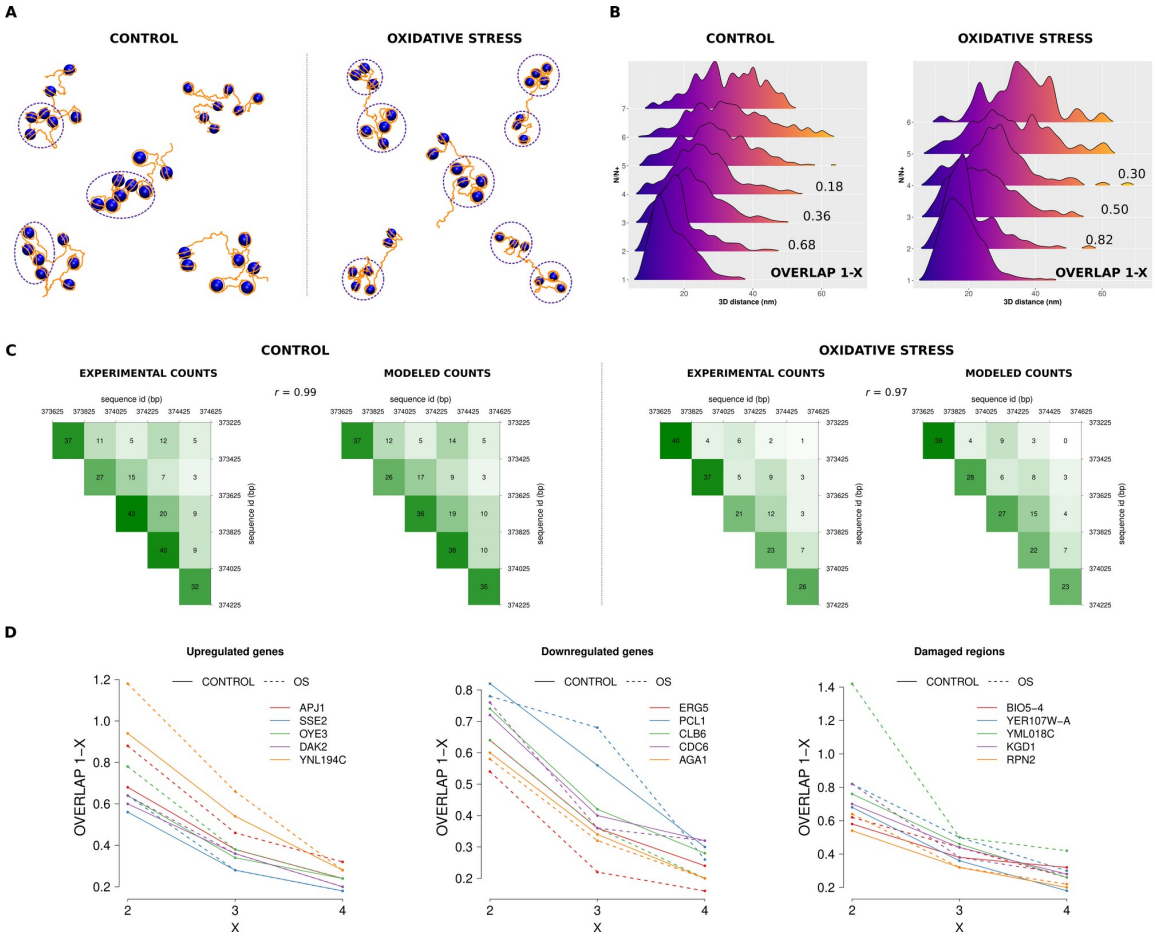


Figure 4. A-C) Chromatin fiber models for YER107W-A gene obtained with the coarse-grained approach at the nucleosome level. A) Representative structures extracted from the ensembles for control and oxidative stress samples. Nucleosome clutches are highlighted with dashed circles. B) 3D distance distributions between nucleosome N and nucleosome N+x (x = 1, 2, … representing subsequent positions in the sequence) obtained for the ensemble structures in control and oxidative stress samples. Overlap between N/N+1 and N/N+x distributions are shown (x = 2, 3, 4). Overlap = normalized proportion of the unitary area under the N/N+x curve below the median of the N/N+1 distribution (overlap = 1 means equal median for both distributions). C) Micro-C contact matrices from experiment and model for control and oxidative stress samples. Pearson correlation coefficients are shown. D) Overlap values between N/N+1 and N/N+x distance distributions (x = 2, 3, 4) for modelled genes (upregulated, downregulated, and damaged) for both conditions (control and oxidative stress).

Effect of oxidative stress on 3D whole genome structure

The loss of interactions at the longer genomic distances observed by Micro-C (beyond 600 bp) is confirmed by Hi-C data, which show that chromatin interactions at intermediate genomic distances (between 1-15 kb) largely decrease upon oxidative stress. However, the contrary trend is found for large genomic distances (> 15 kb), where oxidative stress leads to a global gain of intra- and inter- chromosomal contacts (Figure 5A-C), mainly related to centromere contacts. Globally, 75% of the differential interactions occurring in gene regions are lost but this effect is even more pronounced for the upregulated genes where 95% of the differential interactions are lost in oxidative stress . This percentage decreases to 55% for downregulated genes and to close to 60% for the damaged genes. Out of the 763 genes that significantly lost or gained chromatin

interactions, 18% are differentially expressed and 4% have a damage causing H2AS129 phosphorylation. The other genes could have suffered milder damage (not mapped by H2AS129ph) or could be already under repair.

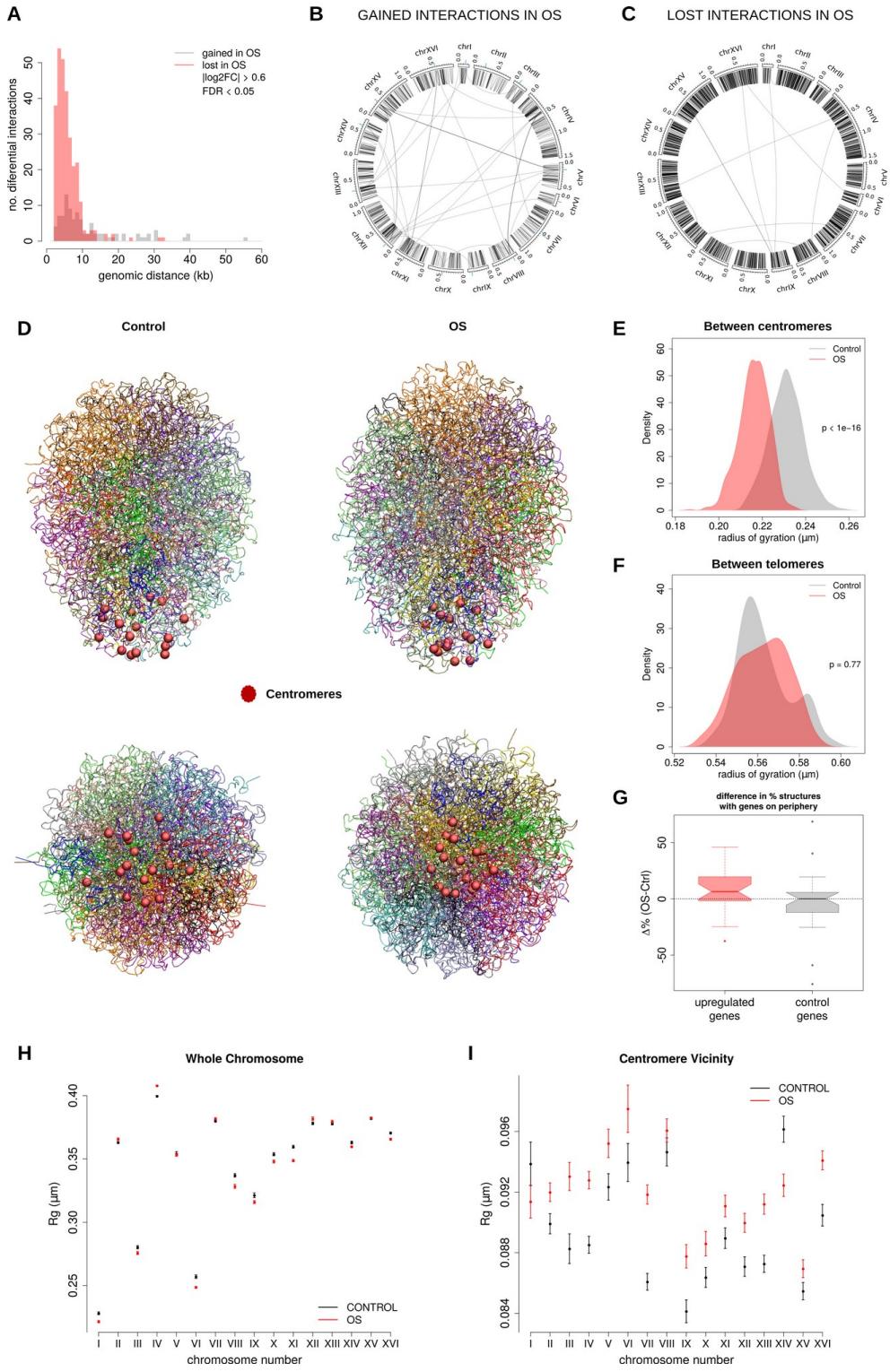


Figure 5. Hi-C differential interactions between oxidative stress and control samples and 3D modelling of chromatin structures. (A) Distribution of differential chromatin interactions as a function of the genomic distance. Gray bars represent interactions gained and red bars

represent interactions lost in oxidative stress samples. Only interactions with counts above 10, absolute log2 fold change greater than 0.6 and false discovery rate < 0.05 were considered as differential. (B, C) Circos plots displaying the positions of the gained (B) and lost (C) interactions in oxidative stress across the genome. (D) Representative 3D chromatin structures for the whole genome for control (left) and OS (right) samples. Centromeres are indicated as red spheres. (E,F) Distribution of radius of gyration for (E) centromeres and (F) telomeres in control (gray) and OS (red) samples. (G) Difference between OS and control samples in percentage of ensemble structures locating upregulated (red) or control (gray) genes on the periphery of the genome structure (accessible surface). 25 top upregulated genes, and 25 genes with the highest expression among samples and without difference in expression between samples (OS and control) were considered, respectively.(H,I) Radius of gyration (mean +/- standard deviation) computed (H) around centromeres (+/- 10 kb) and (I) on the whole chromosome for the control (black) and oxidative stress (red) ensemble models.

Differential binding analysis of the SCC4 ChIP-Seq data (cohesion loader complex bound to chromatin) revealed that 291 sites were depleted of SCC4 signal and 28 showed a stronger signal in the OS sample (Supplementary Figure S8). As can be seen, many regions exhibiting loss of cohesin are correlated with the loss of chromatin interactions in adjacent regions (compare red to orange sites on Supplementary Figure S8).

To model the effect of oxidative stress on the spatial organization of the whole genome, we first generated ensembles of 3D structures using our data-driven model of DNA (Neguembor et al. 2022) at 1 kb resolution. As shown on representative structures (Figure 5D) and quantified for the whole ensembles (Figure 5E), centromeres appear to be more clustered in the treated samples, while telomeres remain more or less spread to a similar extent (Figure 5F). Actually, in control cells, telomeres are mainly clustered compared to OS, but exhibit some structures with “open arms”. Also, upregulated genes under OS show a tendency to migrate to the periphery of the whole genome structure (Figure 5G). When analyzing each chromosome structure separately, several chromosomes tend to be globally more compact as indicated by the decrease of their radius of gyration (Figure 5H). However, some of them, especially the larger ones (e.g. chromosomes IV and XII) are decondensed and practically all the pericentromeric regions present an inverted tendency (Figure 5I), consequent with the loss of interactions found in the shorter -below 15 kb- genomic distances (Figure 5A). Globally, the yeast genome organization leads to a bigger nucleus with 11% increase of its volume upon stress (1,64 μm^3 vs 1,47 μm^3 for the control cells, Supplementary Figure S9). This increase is relatively small considering the local decondensation of the chromatin fiber and could be a consequence of a steric constraint imposed by the nuclear membrane to control the volume of the nucleus. This global constrain also explain the increase of long-range cis and trans interactions (Figure 5B).

Our integrative approach has allowed us to study in detail the 3D genome re-organization from low- to high-order chromatin structure in response to oxidative stress. Altogether, our results suggest that the chromatin is experiencing complex structural changes under oxidative stress with a gain of interactions at very short distances (< 600 bp), a loss of interactions at distances between 600 bp and 15 kb and then a gain of interactions at large distances (> 15 kb).

Discussion

Combining state-of-the-art modelling and experimental techniques we were able to relate a series of changes on chromatin structure generated by oxidative stress to differences in nucleus condensation, protein expression and DNA damage. First, oxidative stress induces an increase

in the fuzziness of the nucleosomes that seems to be stronger in up-regulated genes and regions showing DNA damage, generally accompanied by a decrease in nucleosome occupancy in the gene body and a decrease in the length of nucleosome free regions. These results suggest that oxidative stress has an impact on chromatin structure and could lead to a relaxation of the fiber, in agreement with chromatin decondensation reported by Hauer et al (Hauer et al. 2017) who observed histone degradation upon DNA damage in yeast cells. The effect observed in nucleosome positioning and nucleosome occupancy is correlated with changes in the chromatin fiber structure where we observed an increase of nucleosome interactions at distances below 600 bp and a decrease of interactions at longer distances (> 600 bp). This was transduced to the chromatin fiber coarse-grained model where nucleosome clutches, as described by (Ricci et al. 2015), were enhanced and reduced in size (no more than 4 nucleosomes), especially in upregulated genes and regions with DNA damage, also leading to more extended fibers. The correlation between gene expression, DNA damage and differential chromatin interactions was studied and revealed that close to 20% of all differential chromatin interactions were related with differentially expressed genes, while around 7% corresponded to regions experiencing DSB, which represent only one type of DNA damage induced by ROS (Wallace 2002). These results indicate that transcription and DNA damage were involved in the changes observed in the chromatin structure.

The organization of the chromosomes into CIDs was also affected with an increase in the number of CIDs correlated with an increase in the insulation score and decrease in the size.

The effects observed at the level of the chromatin fiber also have consequences at the level of the whole genome organization, where an increase of *trans* contacts was detected, while *cis* contacts were reduced, especially at distances < 15 kb. The increase of *trans* interactions and long range *cis* interactions might be due to physical constrains introduced by the nuclear membrane, despite the general opening of the chromatin fiber. We also observed that centromeres were more condensed, while telomeres were more spread out in response to oxidative stress, maybe due to the extensive DNA damage they suffer revealed by H2AS129ph. The phosphorylation of H2AS129 could promote some repulsion forces between the nucleosomes and the DNA backbone to create space for the DNA repair machinery, as suggested by Herbert et al. (Herbert et al. 2017).

Altogether, our results show that chromatin is modulated by oxidative stress at two different levels, the first one at short-range distances where we observed a chromatin decondensation while long-range interactions and *trans* interactions increase. This suggests that the response to stress is a combination of events occurring in favor of the open-state of the chromatin, but with a complex and differential behavior depending on the resolution level (short internucleosome contact, long internucleosome contact, medium chromatin and long chromatin contacts). Also, the multiple level of response could reflect the different type of DNA lesions caused by oxidative stress (OxoG, single strand break and double strand break, etc) that activate different repair pathways. Oxidative base lesions are repaired essentially by the Base Excision Repair (BER) pathway (Maynard et al. 2009) and Nucleotide Excision Repair (NER) while Non Homologous End Joining (NHEJ) and Homologous Recombination (HR) pathways are triggered when SSBs and DSBs are detected (Friedberg 2003; Ciccia and Elledge 2010). It appears that the mechanism of 3D genome organization upon oxidative damage is more complex than one single model as previously proposed for several studies in chromatin response after DNA damage (Hauer et al. 2017; Herbert et al. 2017).

Material and methods.

Yeast Strain

In this study, we used the yeast Strain BCY123 (Mat a, pep4::HIS3 prb1::LEU2 bar1::HISG lys2::GAL1/10-GAL4 can1 ade2 trp1 ura3 his3 leu2-3, 112 Dlys2cir+GAL+RAF+SUC).

Oxidative Stress treatment

S.cerevisiae cells were cultivated in 350 ml of YPD medium to OD600=0.3. Then, cells were synchronized in G1 phase with 100 nM α-factor mating pheromone (GenScript). α-factor was removed by washing the cells 3 times before they were re-cultured in new YPD medium during 10 min. Control cells were collected and while the rest of the cells were treated with 10 min with 10 mM H2O2, as previously described in (Azevedo et al. 2011). Treated cells were centrifuged 5 min at 1500 g, washed and cultivated 30 min more in fresh YPD medium before being harvested.

Spot assay

To test the effect of our treatments on cell viability, the spot assay was performed as described in (Zechmann et al. 2011). Briefly, cells were diluted to OD_{600} 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}. 5 µl of each dilution were spotted onto a YPD agar plate and cultivated at 30ºC for two days.

Flow Cytometry analysis

Monitoring of cell cycle synchrony was followed by flow cytometry and fluorescence microscopy. Flow Cytometry was performed using 1 ml yeast cells at OD_{600} =0.6 fixed with 1 ml of 100% cold Ethanol that was added slowly during vigorous mixing. After 1h at 4ºC, cells are fixed. The cells were centrifuged 5 min at 13200 g at 4ºC and washed with 500 µl of 1X PBS. Half of the cells were collected for observation by fluorescence microscopy.

The other half of the cells were washed with 500 µl of 1X SCC buffer, centrifuged, and the pellets were resuspended in 250 µl of 1 X SCC buffer, supplemented with 12.5 µl of 10 mg/ml RNase A and incubated for 90 min at 50ºC. 6.25 µl of 20 mg/ml Proteinase K were added and cells were incubated 90 min at 50ºC. Cells were then sonicated for 10 min (at intervals of 10 sec on -20 sec off) at medium power (Bioruptor® Pico (Diagenode)). To avoid aggregates, 1 ml of PBS-Triton 0.1% and 2.5 µl of EDTA 0.5M were added and cells were stained with 1 µM SytoxGreen (Invitrogen). Finally, the fluorescence emitted was measured by Beckman Coulter Gallios® flow cytometer.

Fluorescence Microscopy

Cells were stained with 1 µg/ml DAPI for 5 min at RT and washed three times with 1X PBS. Pellets were resuspended in 20 µl of 1 X PBS. 5 µl of sample and 5 µl of Prolong Gold anti-fade mounting medium were mixed on a glass slide and stored in the dark at room temperature at least 24 hours before observation with a SP5 microscope (Leica).

Determination of nucleus surface by IMARIS

Confocal pictures were used to calculate the volume and the length of the nuclei using Bitplain IMARIS® (Oxford instruments).

Immunofluorescence

Immunofluorescence was performed with the protocol described in (Silver 2009). Briefly, 25 ml yeast cells were incubated with Zymolase (Genotech, Inc) to digest the cell wall and obtain spherosplasts. 50 µl of spheroplasts were mixed with 200 µl Methanol for 6 min at RT to permeabilize the membranes. After centrifugation, pellets were resuspended with 500 µl of 1X

PBS-BSA 1% incubated for 1 hour at RT and washed again three times with 1X PBS. Spheroplasts were then incubated with 1 µg/ml of primary antibody in 1X PBS-BSA 1%, during 1 hour at RT. Then, samples were washed 3X and incubated 1 hour at RT with secondary antibody diluted (1:1000) in 1X PBS-BSA 1%, washed and DNA was labelled by incubating the samples with 1 µg/ml DAPI for 5 min at RT. The excess of DAPI was removed by three washes and samples were mounted as previously described with Prolong Gold. Finally, samples were observed using a SP5 microscope (Leica, Inc) or a LSM 880 Airyscan microscope (Zeiss, Inc).

Quantification of Histone marks signal using Immunofluorescence

For the automated detection of H2AS129ph foci inside the yeast nuclei cells, we developed a macro in the ImageJ software. Images taken with the confocal microscope were z maximum projected. A threshold was applied to the DAPI signal and the images of the nuclei were converted to binary images that were used as a mask. A mathematical operation was performed between the mask and the H2AS129ph foci channel. ROI manager tool was used to measure the mean intensity values of each foci for each nucleus independently.

Yeast RNA Isolation

25 ml of yeast cells were collected by centrifugation at 1500 g for 5 min at 4ºC. Pellets were frozen in dry ice and stored at -80ºC. The next day, samples are left to defrost on ice and cells were resuspended in 400 µl AE buffer (NaAc 50 mM, EDTA 10 mM, pH5) supplemented with 40 µl of Sodium Dodecyl Sulfate (SDS) and 300 µl of glass beads was added. 400 µl of Phenol Acid was added and the samples were vigorously vortexed, alternating 1 min vortexing and 1 min on ice, repeated 3 times. Samples were then incubated 10 min at 65ºC with shaking and 5 min on ice, centrifuged 5 min at 13200 g at 4ºC and supernatants were collected. After 2 chloroform extractions, RNA was Ethanol precipitated and suspended in 100 µl of nuclease free water.

Stranded mRNA library preparation and sequencing

To determine the total RNA quality and quantity was used Qubit® RNA HS Assay (Life Technologies, Inc) and RNA 6000 Nano Assay on a Bioanalyzer 2100 (Agilent, Inc).

The RNASeq libraries were prepared following KAPA Stranded mRNA-Seq Illumina® Platforms Kit (Roche,Inc) following the manufacturer´s recommendations. Total RNA (500 ng) was enriched for the polyA mRNA fraction and fragmented by divalent metal cations at high temperature. In order to achieve the directionality, the second strand cDNA synthesis was performed in the presence of dUTP. The blunt-ended double stranded cDNA was 3´adenylated and Illumina platform compatible adaptors with unique dual indexes and unique molecular identifiers (Integrated DNA Technologies,Inc) were ligated. The ligation product was amplified with 15 PCR cycles and the final library was validated on an Agilent 2100 Bioanalyzer with the DNA 7500 assay (Agilent, Inc). The libraries were sequenced to 20 million reads on HiSeq2500 (Illumina, Inc) using TruSeq SBS Kit v4-HS (Illumina,Inc), in paired-end mode with a read length of 2x76bp following the manufacturer's protocol. Images analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (1.18.66.3).

Mapping and quantification

Around 95% of the reads were mapped against the reference genome (*S.cerevisiae* release 74 + artificial plasmids) with the GEM software (v1.7.0) (Marco-Sola et al. 2012) allowing for split maps. As expected most of the reads mapped to exonic regions (92%). Genes were quantified using Flux-Capacitor (v1.6.1) (Montgomery et al. 2010) and normalized by the TMM method of the edgeR software (Robinson and Oshlack 2010).

Gene Set Enrichment Analysis (GSEA)

Using the list of genes ranked by expression level, we run GSEA software (Subramanian, Tamayo, et al. 2005) to identify gene sets (http://ge-lab.org/gskb/) enriched in the up and down regulated genes. GSEA was executed with the following parameters: score='weighted', summarize='Median_of_probes', minSize=5, maxSize=2000, numplots=25, permutations=1000.

Micrococcal Nuclease Digestion.

Micrococcal Nuclease (MNase) digestion was performed on the Semi-Intact yeast cells as described in (Deniz et al. 2011). The optimal conditions of MNase digestion necessary to obtain more than 80 % of mono-nucleosomes were first established using a small aliquot of cells before preparing the large-scale reaction required for the MNase-Seq experiment. See supplementary material for detailed protocol. Once the optimal conditions were set up (30 min incubation with 0.02 U of MNase often gives the best results), the reaction was scaled up to digest 0.6 x 10^9 cells to obtain at least 3 µg of DNA for the sequencing.

Chromatin immunoprecipitation (ChIP)

ChIP experiment for H2AS129ph was performed, as previously described in (van Attikum et al. 2004) while ChIP experiment for the SCC2/SCC4 complex was performed as reported in (Lopez-Serra et al. 2014). The detailed protocols are described in supplementary material.

Micro-C methodology

Micro-C analysis were performed as described in (Hsieh et al. 2016) with few modifications . Cells were crosslinked with 3% formaldehyde for 15 min and the reaction was quenched with 125 mM Glycine 5 min. Spheroplasts were prepared as described in section (Semi-Intact yeast cell preparation) and MNase digestion was performed using the conditions defined previously. Reactions were stopped adding 2 mM EGTA and incubated 10 min at 65ºC, centrifuged 5min at 1000 g at 4ºC.

Micro-C protocol was performed using pellet as a sample as reported in (Hsieh et al. 2016), but removing the BSA in the end-labelling reaction that was performed with 100 µM of each nucleotide. The ligation reaction was incubated 1hour at RT and then at 16ºC overnight, shaking at 300 rpm. After proteinase K incubation, samples were purified with phenol extraction and Ethanol precipitated as previously described, resuspending the sample in 50 µl of sterile H_2O. The quality of the Micro-C libraries was tested by agarose electrophoresis to check the proportion of fragments around 300 bp (the size of two nucleosomes ligated) in the Micro-C sample compared to the initial MNase digestion pattern.

The paired-end Micro-C-sequencing libraries were prepared with KAPA Library Preparation kit (Roche, Inc) with some modifications. The biotin marked and de-crosslinked DNA was sheared to a size of 300-500 bp on Covaris™ LE220 (Covaris, Inc) focused-ultrasonicator. The fragmented DNA was end-repaired, adenylated and the biotin-tagged DNA was pulled down using the Dynabeads™ MyOne™ Streptavidin C1 beads (Thermo Fisher Scientific, Inc). The biotinylated fragments were ligated to Illumina platform compatible adaptors with unique dual indexes and unique molecular identifiers (Integrated DNA Technologies, Inc) and enriched by 12 PCR cycles by KAPA HiFi PCR Kit (Roche, Inc).

The libraries were sequenced to 40 million reads pairs on HiSeq2500 (Illumina, Inc) using TruSeq SBS Kit v4-HS (Illumina, Inc), in paired-end mode with a read length of 2x76bp following the manufacturer's protocol. Images analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (1.18.66.3).

Hi-C library preparation in budding yeast cells

Hi-C was performed as previously described in (Belton and Dekker 2015) with some modifications described in supplementary material. After formaldehyde crosslinking, samples were poured into a pre-chilled mortar with liquid nitrogen. Once the sample were frozen, cells were crushed with the pestle during 30 min, adding liquid nitrogen every 5 min to keep the cells frozen. The broken cells were then transferred to a 50 ml falcon tube and 45 ml of ice-cold 1X NEBuffer 2 (NEB Biolabs, Inc) was added. Samples were centrifuged for 5 min at 1800g at 4ºC, supernatants were removed and pellets were resuspended in 1X NEBuffer 2 (NEB Biolabs, Inc) to have a final OD_{600} 10. At this point, the sample can be stored at -80ºC for several years. Once the sample is unfrozen, pellet was resuspended in 2.7 ml 1X NEBuffer 2 and distributed 456 µl of cell lysate into each six 1.5 ml LoBind tube (Eppendorf, Inc). The digestion of the chromatin was performed as described in the original protocol. The libraries were sequenced to 60 million reads pairs on HiSeq2500 (Illumina, Inc) using TruSeq SBS Kit v4-HS (Illumina, Inc), in paired-end mode with a read length of 2x76bp following the manufacturer's protocol. Images analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (1.18.66.3).

Nucleosome calling

MNase-Seq paired-end reads were mapped to yeast genome (sacCer3, Apr. 2011) using Bowtie (Langmead et al. 2009) aligner, allowing a maximum of 2 mismatches and maximum insert size of 500 bp. Output BAM files were imported in R (Team 2011)and quality control was performed with htSeqTools package to remove PCR artifacts (Planet et al. 2012). Filtered reads were processed with nucleR package (Flores and Orozco 2011) as follows: mapped fragments were trimmed to 50 bp maintaining the original center and transformed to reads per million. Then, noise was filtered through Fast Fourier Transform, keeping 2% of the principal components, and peak calling was performed using the parameters: peak width 147 bp, peak detection threshold 35%, maximum overlap of 80 bp, dyad length 50 bp. Nucleosome calls were considered well-positioned when nucleR peak width score and height score were higher than 0.6 and 0.4, respectively, and fuzzy otherwise. A global score of nucleosome positioning for each nucleosome is computed as 1 – the sum of averaged width and height scores.

Nucleosome Dynamics

Nucleosome Dynamics R package (Buitrago et al. 2019) was used to find changes in nucleosome organization between control and treated samples. Nucleosome changes were obtained running Nucleosome Dynamics with the following parameters: maximum difference of 70, maximum length of 140, minimum number of reads to report a shift of 3, shifts threshold of 0.1, indels minimum number of reads to report evictions and inclusions (indels) of 3, indels threshold of 0.05.

Whole genome detection of binding ChIP-Seq signal

The FastQC 0.11 software was used to perform an initial quality control of ChIP-Seq raw FASTQ files. Afterwards, reads were trimmed to remove Illumina adapter sequences identified with FastQC using AdapterRemoval v2.1.7 (Schubert et al. 2016). Trimmed reads were then aligned against the *Saccharomyces cerevisiae* sacCer3 genome assembly, using Bowtie 0.12.9 (Langmead et al. 2009) allowing 1 mismatch (-n 1) and discarding reads aligning to more than one genomic location. Duplicated reads (potential PCR over-amplification artifacts) were identified and removed with sambamba v.0.5.1 (Feng et al. 2012) using default options. Genomic tracks for visual inspection were generated using IGVTools version 2.2.23 (Thorvaldsdottir et al. 2013). Coverage ratio tracks were generated in R version 3.4.4 with the

coverage function from the GenomicRanges package version 1.30.3 (Lawrence et al. 2013) and exported using rtracklayer version 1.38.3 (Lawrence et al. 2009). Additional quality control (PCA-like plots, Gini/Lorenz coverage distribution plots) were done using the htSeq Tools package version 1.26.0 (Planet et al. 2012). MACS 1.4.2 was used to identify putative binding sites with options -g 12e06 (*S.cerevisiae* genome size) and –keep-dup to maintain those duplicated reads not identified as artifacts by sambamba. The reported binding regions were annotated using the ChIPSeeker package version 1.14.2 (Yu et al. 2015), with the TxDb *S.cerevisiae* UCSC sacCer3 SGD Gene annotation package version 3.2.2, using options TSS Region=c (-100,100), overlap=all. Coverage profiles around annotated genes were generated using functions regions Coverage and plot Mean Coverage from the htSeq Tools package.

For the SCC4 ChIP-Seq samples, the analysis was performed using the ChIP-Seq analysis pipeline described in (Afgan et al. 2018).

Whole genome differential binding analysis of ChIP-Seq data between OS and Control conditions.

The DiffBind package version 2.6.6 (Ross-Innes et al. 2012) was used to identify differential binding signal between OS and control conditions from the set of previously identified binding sites, using functions dba.count with default parameters, and function dba.analyze with options method=EDGER, bFull Library Size=FALSE, bSubControl=TRUE, bTagwise =FALSE. The reported differentially bound regions were annotated using the ChIPSeeker package as described above.

The differential analysis between OS and control condition in SCC4 ChIP-Seq samples was performed using the web platform Galaxy as described in (Afgan et al. 2018).

Whole genome and differential binding analysis of ChIP-Seq data between OS and Control conditions over Telomeric and other repeated regions.

In order to study whole genome distribution in both conditions, and differential binding signal of H2A129ph in OS against control with relation to telomeric regions and other repeated elements, we performed an additional round of alignment as described in step 1, but this time allowing all possible alignment sites per read so that proportion of reads aligning with these regions could be compared between immunoprecipitated and control samples (Bayona-Feliu et al. 2017). Afterwards, whole genome and differential binding sites were identified and annotated using the same procedure described above. Statistical significance of co-localization of differential binding sites with telomeric regions was assessed using the RegioneR package version 1.10.0 (Gel et al. 2016) using 5000 permutations.

Hi-C and Micro-C data processing and normalization

We processed Hi-C and Micro-C data using TADbit (Serra et al. 2017) (https://github.com/3DGenomes/tadbit) for quality control, mapping and filtering. First, quality control was performed with the FastQC protocol implementation in TADbit. Then, reads were mapped to the reference yeast genome (sacCer3, Apr. 2011) with a fragment-based strategy. For the Hi-C data, non-informative contacts (self-circle, dangling-end, error, duplicated and random breaks) identified by TADbit were filtered-out. Off-target contacts (neither end of the read mapped to one of the capture regions) were also discarded. Finally, contact matrices were created from valid reads at different resolutions with the corresponding TADbit module, and low frequency bins were removed. Raw contact maps were normalized using the iterative correction approach described in (Imakaev et al. 2012). Contact map visualization and analyses of Hi-C data considered such balanced matrices if not stated otherwise. Intra-arms from regions with no

reads mapped at pairs that occur in the same fragment (diagonal element equal to zero) were set to zero.

Differential Hi-C and Micro-C interactions

Differential Hi-C/Micro-C analysis were performed using the R/Bioconductor package diffHic (Lun and Smyth 2015). The mapped data were filtered and the differential interaction analysis between the control and oxidative stress samples (using the two replicates for each treatment) was performed using the recommended procedure.

Chromosomal interaction domains

Chromosomal interaction domains (CIDs) were estimated using the rGMAP R package (Yu et al. 2017) with default parameters. Normalized contact maps were considered with 1kb resolution binning. Regions with no reads mapped at pairs that occur in the same fragment (diagonal element equal to zero) were ignored.

Insulator score

A measure of the contact frequency along the chromosome at specific distances is calculated as in (Mizuguchi et al. 2014). The insulator score is found as the average over distances between 2 kb to 15 kb. Normalized contact maps (Yu et al. 2017) are considered at 1 kb resolution binning.

Contact probability plot

Contact probability as function of intra-chromosomal distance P(s) was calculated following the strategy proposed in (Naumova et al. 2013) using the Hi-C normalized contact maps (Yu et al. 2015) at 1 kb binning.

Genomic separations were divided into logarithmically spaced bind from 2 kb and increasing by a factor of 1.2 as proposed in (Naumova et al. 2013) establishing 10 kb as a short-range threshold as published in (Lazar-Stefanita et al. 2017). For each bin, it was computed the number of observed Hi-C reads within the newly bin distance range and the number of fragment pairs separated by that distance range. The total number of reads in each bin was divided by the given number of possible fragment pairs. Normalization was done so that the P(s) curve under range distances integrates to 1. In this way, curves resulting from different observations can be compared.

Subtraction plots

Difference OS – Control contact maps were computed as follows. For any chromosome, log2 of one plus the normalized counts was calculated in the 4 samples. For each bin, the average difference of OS samples against Control samples was found. For visualization reasons, values higher than the 95% percentile were set at such 95% percentile, and values lower than the 5% percentile were set at such 5% percentile.

Chromatin models using experimental data

Hi-C based chromatin 3D structure

High resolution Hi-C data at 1 kb was used to obtain the 3D structure, conformation and dynamics of entire yeast chromosomes and their context inside the nucleus. The Hi-C technique provides interaction contacts between DNA fragments. The interaction counts or frequencies

between two *loci i* and *j* (f_{ij}) can be converted to spatial 3D distances between those *loci* (d_{ij}) by an inverse relationship (equation 1),

$$d_{ij} = \gamma / f_{ij}^{\alpha} \qquad 1$$

where γ represents the scale of the structure and is usually taken to match experimental distances between selected genomic regions, and the precise value of α depends on the organism under study, the genomic distance, and the resolution of the HiC map and needs to be fitted (Zhang et al. 2013; Varoquaux et al. 2014; Adhikari et al. 2016). In the present work, γ was taken to reproduce the ellipsoidal axis lengths of the nucleus measured by microscopy and α was fitted to maximize the correlation between experimental and modeled contact maps.

Since Hi-C interaction counts are known to present several biases, such as mappability of fragments, GC content, and fragment length, they were normalized using iterative correction and eigenvector decomposition (Imakaev et al. 2012). Finally, the output of the conversion procedure was a matrix containing equilibrium distances (r_0) for the different interacting *loci*. To remove the background noise, a cutoff of two times the median of all trans contacts (i.e., between *different* chromosomes) was applied to the Hi-C contact map to define *interacting regions*.

The chromosome model was built as a chain of beads, each bead representing a genomic region that corresponds to a bin from the Hi-C map. Spatial equilibrium distances were obtained from equation 1 as explained above. The distances between interacting beads (r) were restrained near their equilibrium length (r_0) during the simulations by penalizing with a flat-welled parabola potential (equation 2) when approaching at shorter distances or moving away at longer distances than the equilibrium. A tolerance of one bead radius (rb) was applied.

$$E = \begin{cases} k(r - r_1)^2 & r < r_1 \\ 0 & r_1 \leq r \leq r_2 \\ k(r - r_2)^2 & r > r_2 \end{cases} \qquad 2$$

where k is the spring constant, $r_1 = r_0 - r_b$ and $r_2 = r_0 + r_b$.

To ensure proper connectivity of the fiber, consecutive beads were bound by a harmonic potential with a force constant five orders of magnitude stronger than that applied to interacting non-consecutive beads. An excluded volume was defined for each bead to avoid interpenetration. Additional repulsive restraints were added for *non-interacting* beads, forced to remain at a distance longer than the maximum equilibrium distance obtained from equation 1. The initial structure of the chromosome fiber was varied between an extended conformation and a random localization of initially unbound beads in different replicas. The system was

allowed to sample the conformational space using pmemd simulation engine for GPU from the Amber 18 package. Different conformations of the fibers were determined by attraction and repulsion forces arising from the distance restraints between beads.

In the end, an ensemble of structures was obtained by minimizing the number of experimental restraint violations (equilibrium distances input). A method yielding a population of structures with different conformations was chosen since Hi-C maps are derived from a population of cells with variable chromatin structure.

Chromatin coarse-grained model at the nucleosome level

The starting point for the 3D chromatin model at the nucleosome level is the coverage of the MNase-seq signal obtained using nucleR software (Deniz et al. 2011; Flores and Orozco 2011). Different families bearing nucleosomes in locations compatible with the MNase-seq signal and DNA/histone stoichiometry are derived by deconvolution of the coverage signal by using a composite Gaussian approximation. For each of the resulting families with unique nucleosome arrangements, an ideal 3D chromatin structure is prepared and further simulated by a coarse-grained Monte Carlo sampling approach with flexible linkers and rigid nucleosomes. Linker DNA is represented at the base pair level by a pseudo-harmonic potential expressed in helical parameters (rise, slide, shift, twist, roll, tilt (Walther et al. 2020). Debye-Huckel electrostatics and excluded volume potentials were added to avoid overlaps (exact details of the simulation procedure will be described elsewhere). The results of the different simulations are clustered to select the minimum number of nucleosome structural families that makes physical sense and that together reproduce MNase-seq experiments. Finally, the different fiber structures from the ensemble of each family are selected and properly weighed to reproduce Micro-C contact maps.

Bibliography

Adhikari B, Trieu T, Cheng J. 2016. Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. *BMC Genomics* **17**: 886.

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Gruning BA et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* **46**: W537-W544.

Amente S, Di Palo G, Scala G, Castrignano T, Gorini F, Cocozza S, Moresano A, Pucci P, Ma B, Stepanov I et al. 2019. Genome-wide mapping of 8-oxo-7,8-dihydro-2'-deoxyguanosine reveals accumulation of oxidatively-generated damage at DNA replication origins within transcribed long genes of mammalian cells. *Nucleic Acids Res* **47**: 221-236.

Azevedo F, Marques F, Fokt H, Oliveira R, Johansson B. 2011. Measuring oxidative DNA damage and DNA repair using the yeast comet assay. *Yeast* **28**: 55-61.

Bau D, Marti-Renom MA. 2011. Structure determination of genomic domains by satisfaction of spatial restraints. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **19**: 25-35.

Baxter JS, Leavy OC, Dryden NH, Maguire S, Johnson N, Fedele V, Simigdala N, Martin LA, Andrews S, Wingett SW et al. 2018. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nature communications* **9**: 1028.

Bayona-Feliu A, Casas-Lamesa A, Reina O, Bernues J, Azorin F. 2017. Linker histone H1 prevents R-loop accumulation and genome instability in heterochromatin. *Nature communications* **8**: 283.

Belton JM, Dekker J. 2015. Hi-C in Budding Yeast. *Cold Spring Harbor protocols* **2015**: 649-661.

Bohn M, Heermann DW. 2010. Topological interactions between ring polymers: Implications for chromatin loops. *J Chem Phys* **132**: 044904.

Buitrago D, Codo L, Illa R, de Jorge P, Battistini F, Flores O, Bayarri G, Royo R, Del Pino M, Heath S et al. 2019. Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning. *Nucleic Acids Res* **47**: 9511-9523.

Buitrago D, Labrador M, Arcon JP, Lema R, Flores O, Esteve-Codina A, Blanc J, Villegas N, Bellido D, Gut M et al. 2021. Impact of DNA methylation on 3D genome structure. *Nature communications* **12**: 3243.

Casali C, Siciliani S, Galgano L, Biggiogera M. 2022. Oxidative Stress and Nuclear Reprogramming: A Pilot Study of the Effects of Reactive Oxygen Species on Architectural and Epigenetic Landscapes. *Int J Mol Sci* **24**.

Ciccia A, Elledge SJ. 2010. The DNA damage response: making it safe to play with knives. *Molecular cell* **40**: 179-204.

Cooke MS, Evans MD, Dizdaroglu M, Lunec J. 2003. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J* **17**: 1195-1214.

de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H, Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity Determines Chromatin Looping. *Molecular cell* **60**: 676-684.

Dehingia B, Milewska M, Janowski M, Pekowska A. 2022. CTCF shapes chromatin structure and gene expression in health and disease. *EMBO reports* **23**: e55146.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* **295**: 1306-1311.

Deniz O, Flores O, Aldea M, Soler-Lopez M, Orozco M. 2016. Nucleosome architecture throughout the cell cycle. *Scientific reports* **6**: 19729.

Deniz O, Flores O, Battistini F, Perez A, Soler-Lopez M, Orozco M. 2011. Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC Genomics* **12**: 489.

Ding Y, Fleming AM, Burrows CJ. 2017. Sequencing the Mouse Genome for the Oxidatively Modified Base 8-Oxo-7,8-dihydroguanine by OG-Seq. *J Am Chem Soc* **139**: 2569-2572.

Dizdaroglu M, Jaruga P, Birincioglu M, Rodriguez H. 2002. Free radical-induced damage to DNA: mechanisms and measurement. *Free Radic Biol Med* **32**: 1102-1115.

Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, Nagano T, Andrews S, Wingett S, Kozarewa I et al. 2014. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res* **24**: 1854-1868.

Fang Y, Zou P. 2020. Genome-Wide Mapping of Oxidative DNA Damage via Engineering of 8-Oxoguanine DNA Glycosylase. *Biochemistry* **59**: 85-89.

Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. *Nature protocols* **7**: 1728-1740.

Flores O, Orozco M. 2011. nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics* **27**: 2149-2150.

Friedberg EC. 2003. DNA damage and repair. *Nature* **421**: 436-440.

Fudenberg G, Mirny LA. 2012. Higher-order chromatin structure: bridging physics and biology. *Curr Opin Genet Dev* **22**: 115-124.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell* **11**: 4241-4257.

Gel B, Diez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. 2016. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**: 289-291.

Hakim O, Misteli T. 2012. SnapShot: Chromosome confirmation capture. *Cell* **148**: 1068 e1061-1062.

Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F et al. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**: 630-638.

Hauer MH, Seeber A, Singh V, Thierry R, Sack R, Amitai A, Kryzhanovska M, Eglinger J, Holcman D, Owen-Hughes T et al. 2017. Histone degradation in response to DNA damage enhances chromatin dynamics and recombination rates. *Nat Struct Mol Biol* **24**: 99-107.

Herbert S, Brion A, Arbona JM, Lelek M, Veillet A, Lelandais B, Parmar J, Fernandez FG, Almayrac E, Khalil Y et al. 2017. Chromatin stiffening underlies enhanced locus mobility after DNA damage in budding yeast. *EMBO J* **36**: 2595-2608.

Hsieh TH, Fudenberg G, Goloborodko A, Rando OJ. 2016. Micro-C XL : Assaying chromosome conformation from the nucleosome to the entire genome. *Nature Methods* **13**: 1009-1011.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**: 999-1003.

Jager R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, Heindl A, Whiffin N, Carnicer MJ, Broome L, Dryden N et al. 2015. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature communications* **6**: 6178.

Jamieson DJ. 1998. Oxidative stress responses of the yeast Saccharomyces cerevisiae. *Yeast* **14**: 1511-1527.

Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. 2011. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* **30**: 90-98.

Kreuz S, Fischle W. 2016. Oxidative stress signaling to chromatin in health and disease. *Epigenomics* **8**: 843-862.

Kuge S, Jones N, Nomoto A. 1997. Regulation of yAP-1 nuclear localization in response to oxidative stress. *EMBO J* **16**: 1710-1720.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Lawrence M, Gentleman R, Carey V. 2009. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**: 1841-1842.

Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**: e1003118.

Lazar-Stefanita L, Scolari VF, Mercy G, Muller H, Guerin TM, Thierry A, Mozziconacci J, Koszul R. 2017. Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. *EMBO J* **36**: 2684-2697.

Lopez-Serra L, Kelly G, Patel H, Stewart A, Uhlmann F. 2014. The Scc2-Scc4 complex acts in sister chromatid cohesion and transcriptional regulation by maintaining nucleosome-free regions. *Nat Genet* **46**: 1147-1151.

Lun AT, Smyth GK. 2015. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16**: 258.

Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R et al. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**: 1012-1025.

Marco-Sola S, Sammeth M, Guigo R, Ribeca P. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* **9**: 1185-1188.

Marti-Renom MA, Almouzni G, Bickmore WA, Bystricky K, Cavalli G, Fraser P, Gasser SM, Giorgetti L, Heard E, Nicodemi M et al. 2018. Challenges and guidelines toward 4D nucleome data and model standards. *Nat Genet* **50**: 1352-1358.

Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, Cooper NJ, Barton A, Wallace C, Fraser P et al. 2015. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nature communications* **6**: 10069.

Maynard S, Schurman SH, Harboe C, de Souza-Pinto NC, Bohr VA. 2009. Base excision repair of oxidative DNA damage and association with cancer and aging. *Carcinogenesis* **30**: 2-10.

Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA et al. 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47**: 598-606.

Mingard C, Wu J, McKeague M, Sturla SJ. 2020. Next-generation DNA damage sequencing. *Chem Soc Rev* **49**: 7354-7377.

Miriklis EL, Rozario AM, Rothenberg E, Bell TDM, Whelan DR. 2021. Understanding DNA organization, damage, and repair with super-resolution fluorescence microscopy. *Methods Appl Fluoresc* **9**.

Mizuguchi T, Fudenberg G, Mehta S, Belton JM, Taneja N, Folco HD, FitzGerald P, Dekker J, Mirny L, Barrowman J et al. 2014. Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. *Nature* **516**: 432-435.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**: 773-777.

Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny LA, Dekker J. 2013. Organization of the mitotic chromosome. *Science* **342**: 948-953.

Neguembor MV, Arcon JP, Buitrago D, Lema R, Walther J, Garate X, Martin L, Romero P, AlHaj Abed J, Gut M et al. 2022. MiOS, an integrated imaging and computational strategy to model gene folding with nucleosome resolution. *Nat Struct Mol Biol* **29**: 1011-1023.

Niu Y, DesMarais TL, Tong Z, Yao Y, Costa M. 2015. Oxidative stress alters global histone modification and DNA methylation. *Free Radic Biol Med* **82**: 22-28.

Otterstrom J, Castells-Garcia A, Vicario C, Gomez-Garcia PA, Cosma MP, Lakadamyali M. 2019. Super-resolution microscopy reveals how histone tail acetylation affects DNA compaction within nucleosomes in vivo. *Nucleic Acids Res* **47**: 8470-8484.

Pekowska A, Klaus B, Xiang W, Severino J, Daigle N, Klein FA, Oles M, Casellas R, Ellenberg J, Steinmetz LM et al. 2018. Gain of CTCF-Anchored Chromatin Loops Marks the Exit from Naive Pluripotency. *Cell Syst* **7**: 482-495 e410.

Pich O, Muinos F, Sabarinathan R, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N. 2018. Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell* **175**: 1074-1087 e1018.

Planet E, Attolini CS, Reina O, Flores O, Rossell D. 2012. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics* **28**: 589-590.

Poetsch AR. 2020. The genomics of oxidative DNA damage, repair, and resulting mutagenesis. *Comput Struct Biotechnol J* **18**: 207-219.

Reichmann D, Voth W, Jakob U. 2018. Maintaining a Healthy Proteome during Oxidative Stress. *Molecular cell* **69**: 203-213.

Ricci MA, Manzo C, Garcia-Parajo MF, Lakadamyali M, Cosma MP. 2015. Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell* **160**: 1145-1158.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**: R25.

Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, Brown GD, Gojis O, Ellis IO, Green AR et al. 2012. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**: 389-393.

Sati S, Bonev B, Szabo Q, Jost D, Bensadoun P, Serra F, Loubiere V, Papadopoulos GL, Rivera-Mulia JC, Fritsch L et al. 2020. 4D Genome Rewiring during Oncogene-Induced and Replicative Senescence. *Molecular cell* doi:10.1016/j.molcel.2020.03.007.

Schieber M, Chandel NS. 2014. ROS function in redox signaling and oxidative stress. *Current biology : CB* **24**: R453-462.

Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC research notes* **9**: 88.

Serra F, Bau D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. 2017. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol* **13**: e1005665.

Silver P. 2009. Indirect immunofluorescence labeling in the yeast Saccharomyces cerevisiae. *Cold Spring Harbor protocols* **2009**: pdb prot5317.

Swygert SG, Kim S, Wu X, Fu T, Hsieh TH, Rando OJ, Eisenman RN, Shendure J, McKnight JN, Tsukiyama T. 2019. Condensin-Dependent Chromatin Compaction Represses Transcription Globally during Quiescence. *Molecular cell* **73**: 533-546 e534.

Team RDC. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**: 178-192.

van Attikum H, Fritsch O, Hohn B, Gasser SM. 2004. Recruitment of the INO80 complex by H2A phosphorylation links ATP-dependent chromatin remodeling with DNA double-strand break repair. *Cell* **119**: 777-788.

van Steensel B, Dekker J. 2010. Genomics tools for unraveling chromosome architecture. *Nat Biotechnol* **28**: 1089-1095.

Varoquaux N, Ay F, Noble WS, Vert JP. 2014. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* **30**: i26-33.

Wallace SS. 2002. Biological consequences of free radical-damaged DNA bases. *Free Radic Biol Med* **33**: 1-14.

Walther J, Dans PD, Balaceanu A, Hospital A, Bayarri G, Orozco M. 2020. A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Res* **48**: e29.

Weiner A, Chen HV, Liu CL, Rahat A, Klien A, Soares L, Gudipati M, Pfeffner J, Regev A, Buratowski S et al. 2012. Systematic dissection of roles for chromatin regulators in a yeast stress response. *PLoS Biol* **10**: e1001369.

Weiner A, Hsieh TH, Appleboim A, Chen HV, Rahat A, Amit I, Rando OJ, Friedman N. 2015. High-resolution chromatin dynamics during a yeast stress response. *Molecular cell* **58**: 371-386.

Wu J, McKeague M, Sturla SJ. 2018. Nucleotide-Resolution Genome-Wide Mapping of Oxidative DNA Damage by Click-Code-Seq. *J Am Chem Soc* **140**: 9783-9787.

Xu J, Ma H, Ma H, Jiang W, Mela CA, Duan M, Zhao S, Gao C, Hahm ER, Lardo SM et al. 2020. Super-resolution imaging reveals the evolution of higher-order chromatin folding in early carcinogenesis. *Nature communications* **11**: 1899.

Ye B, Hou N, Xiao L, Xu Y, Xu H, Li F. 2016. Dynamic monitoring of oxidative DNA double-strand break and repair in cardiomyocytes. *Cardiovasc Pathol* **25**: 93-100.

Yu G, Wang LG, He QY. 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**: 2382-2383.

Yu W, He B, Tan K. 2017. Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. *Nature communications* **8**: 535.

Zechmann B, Liou LC, Koffler BE, Horvat L, Tomasic A, Fulgosi H, Zhang Z. 2011. Subcellular distribution of glutathione and its dynamic changes under oxidative stress in the yeast Saccharomyces cerevisiae. *FEMS Yeast Res* **11**: 631-642.

Zhang K, Zheng DQ, Sui Y, Qi L, Petes TD. 2019. Genome-wide analysis of genomic alterations induced by oxidative DNA damage in yeast. *Nucleic Acids Res* **47**: 3521-3535.

Zhang Z, Li G, Toh KC, Sung WK. 2013. 3D chromosome modeling with semi-definite programming and Hi-C data. *J Comput Biol* **20**: 831-846.

Chapter 4. Triplex, a new regulatory player

In this Chapter we introduce a new factor that can affect the regulatory network, the formation of hybrid RNA•DNA-DNA triplex (triple-stranded helix). These triplexes can be formed when an RNA sequence binds the DNA duplex blocking its ability to interact with effector proteins interfering the regulatory mechanisms. Previous studies have demonstrated that, indeed, the formation of triplexes in the promoter regions could impact gene regulation (1–3), generating the term "anti-gene" therapies. Accordingly, a potential therapeutic effect can be derived by the addition of a Triplex Forming Oligonucleotide (TFO) designed to bind the promoter region of a pathological gene (4, 5). Another interesting possibility exists: triplex formation might form an ancient regulatory mechanism (6, 7) when the TFO is in fact a cellular RNA (coding, non-coding, miRNA, or even a messenger RNA) which recognizes a region of its own genome (the Triplex Target Sequence; TTS).

DNA triplexes were first theoretically suggested by Pauling and Corey in 1953 (8) and investigated experimentally 4 years later (9). A triplex is formed when a single-stranded polynucleotide binds a polypurine-rich duplex through specific major groove interactions. These interactions are defined by means of Hoogsteen hydrogen bonds in what is known as parallel triplexes, and reverse-Hoogsteen bonds for the antiparallel ones (see Introduction and Figure 4.1). Parallel triplexes are known to be (despite their pH-dependence) more stable under physiological conditions than the antiparallel counterpart (10, 11). Early studies of the conformation of triplexes through fiber diffraction models suggested an A-type conformation (12) but later studies using NMR and Molecular Dynamics showed that the DNA triplex presents in fact an intermediate A- and B-like conformation (13–15).

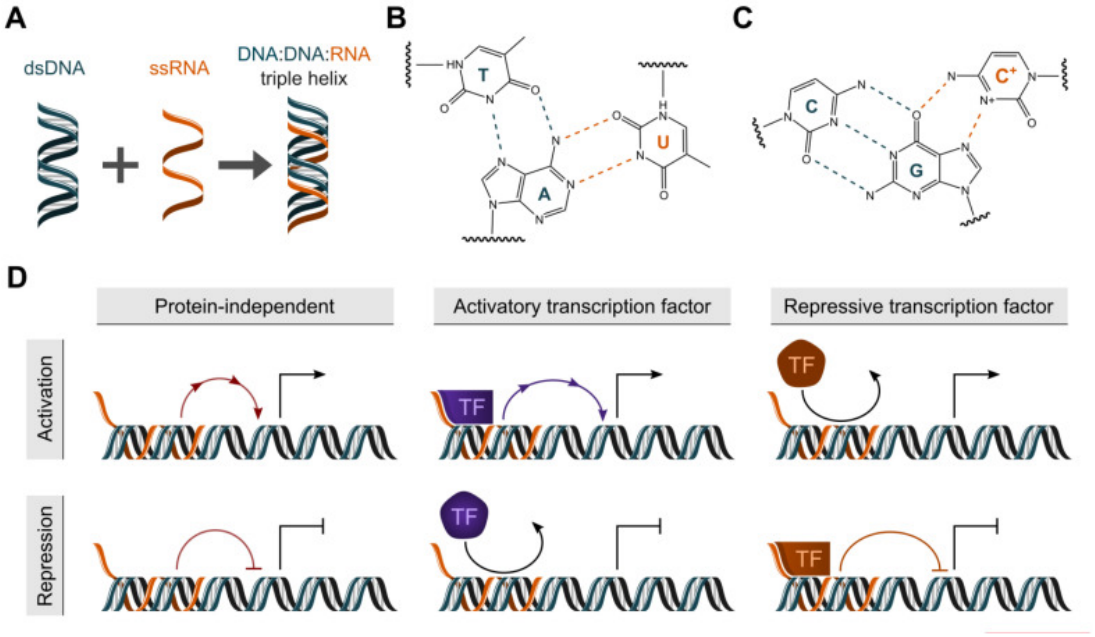


Figure 4.1. Overview of the hybrid RNA•DNA-DNA triplex formation. (A) Schematic image of the triple helix formation between double-stranded DNA (dsDNA) and single-stranded RNA (ssRNA). (B, C) Canonical Watson–Crick and Hoogsteen (red) base pairings which permit the formation of DNA:DNA:RNA triple helices. (D) Putative mechanisms by which the hybrid triplex permits the control of gene expression via interactions with gene loci and transcription factors. Image adapted from (16).

4.1. Theoretical study of triplex stability

A variety of parallel triplexes can be formed by mixing complementary DNA and RNA strands. The combination of extensive *in silico* MD simulations and biophysical experiments allowed the study of the dynamics and stability of hybrid-parallel triplexes. Through the MD exploration of 6 different hybrid triplexes, we found a general order of stability that was further confirmed by melting experiments. The agreement between the theoretical and experimental results provided robustness to the reliability of the obtained stability scale for the different topologies. From these stability results the most stable triplex appeared to be the RNA (pyrimidine) – DNA (purine) · RNA (pyrimidine). Nonetheless, this topology is not expected to have a large prevalence in the cell given R-loop formations. For this reason, we centered our attention on the second most stable triplex: RNA (pyrimidine) – DNA (purine) · DNA (pyrimidine) that can be easily formed by pairing an RNA segment with genomic DNA. This topology, assumed to be more prevalent in cellular conditions, was

chosen to develop and validate a stability predictor which allowed the scanning of stable triplexes under a range of conditions.

4.1.1. Melting temperature as an indicator of stability

The stability of a triplex can be studied through melting experiments (17, 18), where the melting temperature (Tm) can be determined from the inflexion point curve of UV or CD vs. temperature curve. The triplex thermodynamics can be determined from many measures at different oligonucleotide concentrations from the Van't Hoff's equation (eq. 1). Roberts and Crothers (19), realized that different thermodynamic parameters could be well approximated using a nearest neighbor model, with parameters being fitted to a series of duplex-single strand melting experiments systems:

$$\mathrm{Tm} = \frac{310 \cdot \Delta H}{\Delta H - \Delta G - 310 \cdot R \cdot \ln\left({}^{4}/_{C_t}\right)} \qquad (1)$$

where C_t denotes the concentration of the (hairpin) duplex and RNA strands, and *R* is the ideal gas constant. Thus, following Roberts and Crothers (19), the enthalpy of the triplex formation was calculated with a simple nearest-neighbor model (ΔH, equation 2):

$$\Delta H = -\alpha_{cc}(CC) - \alpha_{uc}(UC + CU) - \alpha_{uu}(UU) \qquad (2)$$

where (XX) refers to the number or dinucleotide steps of the type XpX in the TFO and α_{xx} are fitted parameters. While the free energy of triplex formation was calculated by a combination model consisting of five parameters (ΔG, equation 3):

$$\Delta G(pH) = -\alpha_c(C) - \alpha_u(U) - \alpha'_{cc}(CC) + \delta + (C)(pH - 5.6)(\beta - \alpha''_{cc} \cdot (CC)) \qquad (3)$$

The ΔG is defined as a function of the nucleotide content and is dependent on the pH. All symbols in Greek letters are fitted parameters. Note that this equation accounts for the pH dependence of d(C^+-G·C) triads. The model was reparametrized by non-linear fitting using ΔH and ΔG values obtained from 105 melting experiments. These experiments performed at different pH and TFO concentrations were used for training and the new model was latter validated with 52 additional experiments.

4.1.2. Bioinformatics scanning of candidate TFOs and TTSs

The reparameterization of the stability predictor opened the door for an accurate and comprehensive scanning of the human genome and transcriptome for potential TFO sequences. Previous studies in the literature demonstrated the potential role of non-coding RNAs in triplex formation (20–27). Motivated by these studies we used our predictor to screen potential TFOs amongst annotated human long noncoding RNAs (lncRNAs) and microRNAs (miRNAs) from the GENCODE and miRbase databases (28, 29). More specifically, we selected potential TFOs with a minimum length of 10 and a maximum of 30 bps. We defined the default variables for our Tm calculation as follows: pH value at 7.0 and the C_T at a value of 12 μM. Furthermore, a Tm of 30°C was set as a threshold to classify stable fragments (to account for uncertainties in the model). Our triplex target sites (TTSs) were extracted from the list of stable triplexes found in each sample and mapped to the reference human genome.

After obtaining the annotated positions we then investigated relevant features that could explain the regulatory role of triplexes and their mechanisms of action. To further understand the potential regulatory role of triplexes we performed a Gene Ontology (GO) analysis of the genes associated to the promoters that were selected as potential target sites for our candidate TFOs. Lastly, we investigated the co-localization of candidate target sites in comparison to chromatin accessibility, using nucleosomes as a proxy, to further understand the potential effects that TFOs binding can have on chromatin dynamics.

In the fourth publication of this thesis a combination of melting temperature experimental techniques and MD simulations were assessed to characterize the stability of different hybrid triplexes. From consensus stability results we focused on one of the most stable triplexes, both computationally and experimentally, and probably most prevalent in the cell, the RNA (pyrimidine) – DNA (purine) · DNA (pyrimidine).

Melting experiments on our chosen topology allowed the calibration and validation of a thermodynamics-based model which allowed us to design the first triplex stability predictor able to reproduce experimental data for different conditions (pH, concentration and sequence composition). The developed predictor allowed us to demonstrate the enrichment of potential triplexes in human lncRNAs and miRNAs (TFO: expressed RNAs, TTS: genomic DNA). More specifically, we observed a large prevalence of TTSs in

regulatory regions (promoters and 5'UTRs) for our miRNAs in comparison to the expected population from a random distribution. The GO analysis showed that, as previously published in literature (30–32), these genes are frequently related to complex biological processes, such as development. Moreover, these TTSs co-localize in nucleosome free regions, supporting the hypothesis of an ancient RNA-based triplex-mediated regulatory mechanism together with the fact that the main effect of triplex formation in promoter regions is the inactivation of DNA transcription.

In conclusion, candidate TFOs have been found to concentrate in regulatory regions and UTRs, suggesting a potential involvement of triplexes in gene regulatory mechanisms. Furthermore, linking triplex formation to data from chromatin accessibility we found evidence of a broader role for miRNAs in transcription regulation and a potential role for both sets in maintaining chromatin structure by fixing nucleosome arrays. In summary, this fourth project highlights the potentiality regulatory role that triplexes have and further analyzes its link to chromatin structure.

Publication:

Vito Genna*, Guillem Portella*, Alba Sala*, Montserrat Terrazas*, Núria Villegas, Lidia Mateo, Chiara Castellazzi, Mireia Labrador, Anna Aviño, Adam Hospital, Albert Gandioso, Patrick Aloy, Isabelle Brun-Heath, Carlos Gonzalez, Ramon Eritja, and Modesto Orozco. Systematic study of hybrid triplex topology and stability suggests a general triplex-mediated regulatory mechanism, bioRxiv 2024.05.28.596189, https://doi.org/10.1101/2024.05.28.596189.

*Equally contributing authors
Supplementary material for this article can be found in the Annex.

References

1. Durland,R.H., Kessler,D.J., Gunnell,S., Hogan,M.E., Duvic,M. and Pettitt,B.M. (1991) Binding of Triple Helix Forming Oligonucleotides to Sites in Gene Promoters. Biochemistry, 30.
2. Guntaka,R. V, Varma,B.R. and Weber,K.T. (2003) Triplex-forming oligonucleotides as modulators of gene expression. International Journal of Biochemistry and Cell Biology, 35.
3. Joseph,J., Kandala,J.C., Veerapanane,D., Weber,K.T. and Guntaka,R. V (1997) Antiparallel polypurine phosphorothioate oligonucleotides form stable triplexes with the rat α1(I) collagen gene promoter and inhibit transcription in cultured rat fibroblasts. Nucleic Acids Res, 25.
4. Hélène,C. (1991) The anti-gene strategy: control of gene expression by triplex-forming-oligonucleotides. Anticancer Drug Des, 6.
5. Li,C., Zhou,Z., Ren,C., Deng,Y., Peng,F., Wang,Q., Zhang,H. and Jiang,Y. (2022) Triplex-forming oligonucleotides as an anti-gene technique for cancer therapy. Front Pharmacol, 13.
6. Goñi,J.R., Vaquerizas,J.M., Dopazo,J. and Orozco,M. (2006) Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. BMC Genomics, 7.
7. Goñi,J.R., de la Cruz,X. and Orozco,M. (2004) Triplex-forming oligonucleotide target sequences in the human genome. Nucleic Acids Res, 32.
8. Pauling,L. and Corey,R.B. (1953) A Proposed Structure For The Nucleic Acids. Proceedings of the National Academy of Sciences, 39.
9. Felsenfeld,G., Davies,D.R. and Rich,A. (1957) Formation of a Three-Stranded Polynucleotide Molecule. J Am Chem Soc, 79.
10. Scaria,P. V and Shafer,R.H. (1996) Calorimetric analysis of triple helices targeted to the d(G3A4G3)·d(C3T4C3) duplex. Biochemistry, 35.
11. Chandler,S.P. and Fox,K.R. (1996) Specificity of antiparallel DNA triple helix formation. Biochemistry, 35.
12. Arnott,S., Bond,P.J., Selsing,E. and Smith,P.J.C. (1976) Models of Triple-Stranded Polynucleotides with Optimised Stereochemistry. Nucleic Acids Res, 3.
13. Macaya,R.F., Schultze,P. and Feigon,J. (1992) Sugar Conformations in Intramolecular DNA Triplexes Determined by Coupling Constants Obtained by Automated Simulation of P. COSY Cross Peaks. J Am Chem Soc, 114.

14. Raghunathan,G., Miles,H.T. and Sasisekharan,V. (1993) Symmetry and Molecular Structure of a DNA Triple Helix: d(T)n·d(A)n·d(T)n. Biochemistry, 32.
15. Soliva,R., Laughton,C.A., Luque,F.J. and Orozco,M. (1998) Molecular dynamics simulations in aqueous solution of triple helices containing d(G·C·C) trios. J Am Chem Soc, 120.
16. Warwick,T., Brandes,R.P. and Leisegang,M.S. (2023) Computational Methods to Study DNA:DNA:RNA Triplex Formation by lncRNAs. Noncoding RNA, 9.
17. Darby,R.A.J., Sollogoub,M., McKeen,C., Brown,L., Risitano,A., Brown,N., Barton,C., Brown,T. and Fox,K.R. (2002) High throughput measurement of duplex, triplex and quadruplex melting curves using molecular beacons and a LightCycler. Nucleic Acids Res, 30.
18. Jaumot,J., Aviñó,A., Eritja,R., Tauler,R. and Gargallo,R. (2003) Resolution of parallel and antiparallel oligonucleotide triple helices formation and melting processes by multivariate curve resolution. J Biomol Struct Dyn, 21.
19. Roberts,R.W. and Crothers,D.M. (1996) Prediction of the stability of DNA triplexes. Proc Natl Acad Sci U S A, 93.
20. Postepska-Igielska,A., Giwojna,A., Gasri-Plotnitsky,L., Schmitt,N., Dold,A., Ginsberg,D. and Grummt,I. (2015) LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure. Mol Cell, 60.
21. Leisegang,M.S., Bains,J.K., Seredinski,S., Oo,J.A., Krause,N.M., Kuo,C.C., Günther,S., Cetin,N.S., Warwick,T., Cao,C., et al. (2022) HIF1α-AS1 is a DNA:DNA:RNA triplex-forming lncRNA interacting with the HUSH complex. Nat Commun, 13.
22. Lin,T.C., Liu,Y.L., Liu,Y.T., Liu,W.H., Liu,Z.Y., Chang,K.L., Chang,C.Y., Ni,H.C., Huang,J.H. and Tsai,H.K. (2023) TRIPBASE: a database for identifying the human genomic DNA and lncRNA triplexes. NAR Genom Bioinform, 5.
23. Soibam,B. and Zhamangaraeva,A. (2021) LncRNA:DNA triplex-forming sites are positioned at specific areas of genome organization and are predictors for Topologically Associated Domains. BMC Genomics, 22.
24. Cicconetti,C., Lauria,A., Proserpio,V., Masera,M., Tamburrini,A., Maldotti,M., Oliviero,S. and Molineris,I. (2023) 3plex enables deep computational investigation of triplex forming lncRNAs. Comput Struct Biotechnol J, 21.
25. Li,Y., Syed,J. and Sugiyama,H. (2016) RNA-DNA Triplex Formation by Long Noncoding RNAs. Cell Chem Biol, 23.
26. Paugh,S.W., Coss,D.R., Bao,J., Laudermilk,L.T., Grace,C.R., Ferreira,A.M., Waddell,M.B., Ridout,G., Naeve,D., Leuze,M., et al.

(2016) MicroRNAs Form Triplexes with Double Stranded DNA at Sequence-Specific Binding Sites; a Eukaryotic Mechanism via which microRNAs Could Directly Alter Gene Expression. PLoS Comput Biol, 12.
27. Conde,J., Oliva,N., Atilano,M., Song,H.S. and Artzi,N. (2016) Self-assembled RNA-triple-helix hydrogel scaffold for microRNA modulation in the tumour microenvironment. Nat Mater, 15.
28. Frankish,A., Diekhans,M., Ferreira,A.M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J., Armstrong,J., et al. (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res, 47.
29. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) MiRBase: From microRNA sequences to function. Nucleic Acids Res, 47.
30. Pasquier,C., Agnel,S. and Robichon,A. (2017) The mapping of predicted triplex DNA: RNA in the Drosophila genome reveals a prominent location in development- and morphogenesis-related genes. G3: Genes, Genomes, Genetics, 7.
31. Grote,P., Wittler,L., Hendrix,D., Koch,F., Währisch,S., Beisaw,A., Macura,K., Bläss,G., Kellis,M., Werber,M., et al. (2013) The Tissue-Specific lncRNA Fendrr Is an Essential Regulator of Heart and Body Wall Development in the Mouse. Dev Cell, 24.
32. Ali,T., Rogala,S., Krause,N.M., Bains,J.K., Melissari,M.T., Währisch,S., Schwalbe,H., Herrmann,B.G. and Grote,P. (2023) Fendrr synergizes with Wnt signalling to regulate fibrosis related genes during lung development via its RNA:dsDNA triplex element. Nucleic Acids Res, 51.

Systematic study of hybrid triplex topology and stability suggests a general triplex-mediated regulatory mechanism

Vito Genna[1,2#], Guillem Portella[1,3#], Alba Sala[1#], Montserrat Terrazas [1#], Núria Villegas,[1] Lidia Mateo[1], Chiara Castellazzi[1], Mireia Labrador,[1] Anna Aviño[4], Adam Hospital[1], Albert Gandioso,[1] Patrick Aloy[1], Isabelle Brun-Heath[1], Carlos Gonzalez[5], Ramon Eritja[4], and Modesto Orozco[1,6]*

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, E-08028 Barcelona, Spain.
[2]Nostrum Biodiscovery, SL. Barcelona, Spain
[3]Department of Chemistry, University of Cambridge, Cambridge, UK.
[4]Institute for Advanced Chemistry of Catalonia (IQAC), CSIC, Networking Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), E-08034 Barcelona, Spain
[5] Instituto de Química Física Blas Cabrera. CSIC. E-28006. Madrid
[6] Department of Biochemistry and Biomedicine, University of Barcelona, E-08028 Barcelona, Spain.

# Equally contributing authors

ABSTRACT

By combining *in-silico*, biophysical and *in-cellulo* experiments, we decipher the topology, physical and potential biological properties of hybrid-parallel nucleic acids triplexes; an elusive structure at the basis of life. We found that hybrid triplex topology follows a stability order: r(Py)-d(Pu)·r(Py)> r(Py)-d(Pu)·d(Py)> d(Py)-d(Pu)·d(Py)> d(Py)-d(Pu)·r(Py). The r(Py)-d(Pu)·d(Py) triplex is expected to be the preferred in the cell as it avoids the need to open the duplex reducing the torsional stress required for triplex formation in the r(Py)-d(Pu)·r(Py) topology. Upon a massive collection of melting data, we have created the first predictor for hybrid triplex stability. Leveraging this predictor, we conducted a comprehensive scan to assess the likelihood of the human genome and transcriptome to engage in triplex formation. Our findings unveil a remarkable inclination - of both the human genome and transcriptome - to generate hybrid triplex formation, particularly within untranslated (UTRs) and regulatory regions, thereby corroborating the existence of a triplex-mediated regulatory mechanism. Furthermore, we found a correlation between nucleosome linkers and TFS which agree with a putative role of triplexes in arranging chromatin structure and local/global level.

INTRODUCTION

Triplexes are formed when a poly-purine segment of a duplex is recognized by a third oligonucleotide strand (the TFO; triplex forming oligonucleotide) by means of specific hydrogen bond interactions along the major groove [1–5]. The TFO can be arranged parallel or antiparallel to the purine (Pu) strand. The triads (T-A·T, C^+-G·C and G-G·C) present in parallel triplexes are stabilized by means of Hoogsteen hydrogen bonds, while reverse Hoogsteen hydrogen bond pattern stabilizes triads (A-A·T, G-G·C and T-A·T) in antiparallel triplexes (where "-" refers to Hoogsteen/reverse Hoogsteen and "·" refers to Watson-Crick pairings). Isosteric consideration favors triplexes where the third strand is either homopyrimidine (parallel triplexes; pyrimidine (Py) motif) or homopurine (antiparallel triplexes; purine (Pu) motif) [4–6]. Despite the pH dependence of the C^+-G·C triad, the parallel triplexes are more stable than the anti-parallel ones under physiological conditions [6–9].

Early fiber diffraction models suggest an A-type conformation for the DNA triplex[10], but several NMR experiments and exhaustive molecular dynamics (MD) simulations

demonstrated that the DNA triplex shows a "B-like" conformation, with sugars in the South conformation and triads perpendicular to the helix axis [11–18]. The duplex major groove (MG) is divided by the TFO in two grooves[14,17,18]: one very narrow with purine C8 in the bottom (mMG in Figure 1), and the other very wide (MMG in Figure 1), covering all the region between the TFO and the third strand of the duplex (Figure 1). The presence of the TFO blocks the major-groove recognition pattern between transcription factors and the DNA duplex, and while the MMG can be recognized by some proteins [18–20] the general and main effect of triplex formation is the inactivation of DNA transcription; the effect being maximized if the triplex is formed in the regulatory regions [4,21–24]. Very interestingly [25,26], promoters in most organisms, including humans, are highly enriched in poly-Pu sequences (triplex target sequences; TTS), suggesting that a large number of genes could be inactivated by triplex formation [25] if a suitable TFO is available.

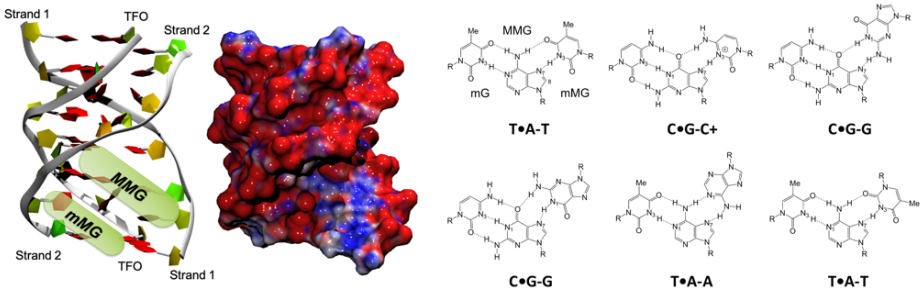


Figure 1. Left panel: tridimensional representation of a triplex with the two sections of the Major Groove highlighted. Right panel. Schematic representation of T·A-T, C·G-C$^+$, C·G-G, C·G-G, T·A-A and T·A-T triads in parallel (top row) and antiparallel (bottom row) triplexes.

The possibility to block DNA expression by adding a DNA-based TFO has been exploited in the so-called anti-gene therapy [21–24,27–38]. However, in normal cellular conditions, single stranded DNA are rare, and putative TFOs are RNA sequences, which act in some cases as regulators of gene expression [39–45]. These experimental findings, combined with bioinformatic analysis [25,26] suggest the existence of an ancient feedback regulatory mechanism based on the formation of triplexes with TTS in a regulatory gene and TFO defined by the RNA of a regulated gene or an accessory regulatory RNA. Very recently, bioinformatics data have been published showing a significant correlation between human HiC contact maps and regions that can form triplexes with a third strand

made of long non-coding RNAs [46,47] suggesting a correlation between triplex formation and genome structural organization [46,47].

Understanding the role of hybrid triplexes in gene regulation and chromatin structure first requires a good knowledge of their structural characteristics. Unfortunately, contrary to pure DNA triplexes [1–18,48–51], little and contradictory information exists on the structure of hybrid triplexes [48–52]. We combine here MD simulations and biophysical experiments to explore the stoichiometry, topology, stability, structure and dynamics of hybrid triplexes. We found, both theoretically and experimentally, a general order of stability: r(Py)-d(Pu)·r(Py)> r(Py)-d(Pu)·d(Py)> d(Py)-d(Pu)·d(Py)> d(Py)-d(Pu)·r(Py), the rest showing little stability, except when the d(Pu) is made of a poly-d(A), in which case the ordering is d(Py)-r(Pu)·r(Py)> d(Py)-d(Pu)·d(Py)> r(Py)-r(Pu)·r(Py)> r(Py)-r(Pu)·d(Py), with little structural differences between the stable hybrid triplexes. We centered our attention in the triplex that is likely to be more prevalent in cellular conditions: r(Py)-d(Pu)·d(Py), developing and validating a stability predictor which allows us to scan for the stability of these triplexes under a range of conditions. Applying this predictor to genomic and transcriptomic data, the likelihood of hybrid triplex formation in human cells is analyzed. A large prevalence of these triplexes is found, being very abundant in regulatory regions (promoters and 5'UTR) and involving mainly miRNAs as TFOs. These findings provide strong support to the hypothesis of an ancient RNA-based triplex-mediated regulatory mechanism. Furthermore, triplexes are located at positions where they can help to fix chromatin structure, both locally and globally.

RESULTS AND DISCUSSION

Stability of Homo-polymers and Hybrid Triplexes. Melting experiments were first performed using different homopyrimidine triplexes as TFO and homopurine-homopyrimidine hairpins as TTSs. The use of hairpins (polyethylene glycol was used as loop) has the advantage to minimize the formation of other competing structures in the TTS, such as reverse Watson-Crick [53,54], Hoogsteen duplexes [55–59], quadruplexes and others [60–62] which will introduce noise in the T_m estimates). We consider 3 compositions of the hairpin: (100% A·T/U (**I-IV**); 70% A·T/U (**V-VIII**) and 50% A·T/U (**IX-XII**)), we do not consider higher percentages of guanines as this will increase the risk of quadruplex formation. With these compositions we create all the combinations of DNA

and RNA in the hairpin: d(Pu)·d(Py) (**I**, **V** and **IX**); r(Pu)·d(Py) (**II**, **VI** and **X**); d(Pu)·r(Py) (**III**, **VII** and **XI**) and r(Pu)·r(Py) (**IV**, **VIII** and **XII**)) and incubate them with the corresponding homopyrimidine TFO (DNA with 100%, 70% or 50% T (**1**, **3** and **5**); RNA with 100%, 70% or 50% U (**2**, **4** and **6**). Combination of all TFOs with all TTS leads to 24 potential triplexes whose stability was measured by the corresponding melting curves (recorded in all cases at pH 6.0; see Methods). Results (Figure 2) show melting temperatures in the range T<15ºC (not detectable) to 52 ºC. Triplexes with 100% A·T/U show in general a poor stability with a decreasing order of stability d(Py)-r(Pu)·r(Py)> d(Py)-d(Pu)·d(Py)> r(Py)-r(Pu)·r(Py)> r(Py)-r(Pu)·d(Py) the rest being not detectable (Figure 2). When the ratio of G·C increases the triplexes become more stable, showing a quite well-defined order of stability (Figure 2): r(Py)-d(Pu)·r(Py)> r(Py)-d(Pu)·d(Py)> d(Py)-d(Pu)·d(Py)> d(Py)-d(Pu)·r(Py)> r(Py)-r(Pu)·d(Py) ≈ r(Py)-r(Pu)·r(Py). At the studied pH, the increase in the ratio G·C/A·T implies an increase in the stability of the triplex. For a given G·C/A·T ratio the two most stable triplexes are those with RNA in the TFO with d(Pu)·r(Py) preferred over d(Pu)·d(Py) in the TTS.

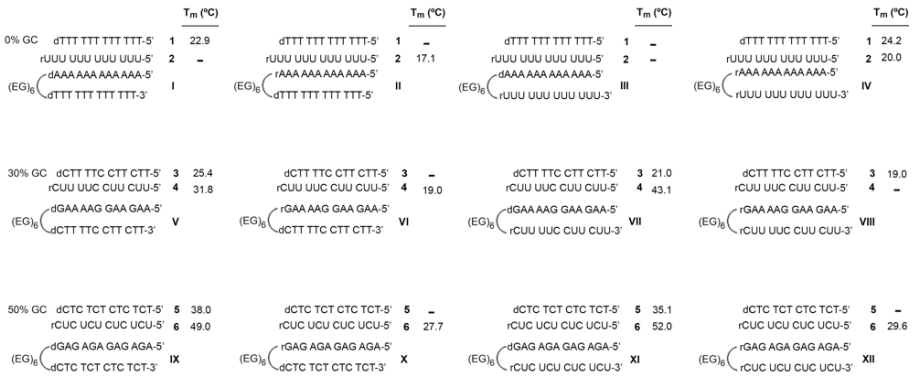


Figure 2. Melting temperatures of d(Py)-d(Pu)·d(Py), r(Py)-d(Pu)·d(Py), d(Py)-r(Pu)·d(Py), r(Py)-r(Pu)·d(Py), d(Py)-d(Pu)·r(Py), r(Py)-d(Pu)·r(Py), d(Py)-r(Pu)·r(Py) and r(Py)-r(Pu)·r(Py) 12mer triplexes of 100%, 70% and 50% A·T/U content in 10 mM sodium cacodylate buffer (pH 6.0) containing 100 mM NaCl and 10 mM $MgCl_2$ (see Materials and Methods for details). Only the T_m of the triplex transition is indicated in all cases. For results at higher pH see below.

In order to confirm that melting experiments were really analyzing triplex→duplex

transitions instead of other processes related to strand invasion with R-loop formation, we repeated melting experiments for triplexes d(Py)-d(Pu)·d(Py), r(Py)-d(Pu)·d(Py), d(Py)-d(Pu)·r(Py) and r(Py)-d(Pu)·r(Py) for the case of 70% A·T/U in the TTS) at higher pH values (from 6.0 to 8.0; see Figure 3a-h) finding a reduction of T_m, consistent with triplex formation. The triplex nature of the structures was further confirmed by means of ^{1}H NMR spectroscopy. Thus, ^{1}H-NMR spectra of hairpin **VII** and its equimolar mix with **4** different temperatures show the expected decay of imino signal intensities as temperature increases (panels i and j of Figure 3). In the spectra of **VII**, four guanine imino signals involved in WC base pairs (12-13 ppm) are observed at T = 5ºC, being two of them still observed at 45ºC. Most probably, they correspond with the two central GC base-pairs. In addition, uracil imino signals forming Watson-Crick AU base-pairs are observed around 14 ppm. These NMR data are fully consistent with formation of a hairpin structure as shown in the insert of the Figure 3i. Upon addition of triplex forming RNA **4**, the number of exchangeable proton signals increases drastically. Formation of GC Hoogsteen base pairs is demonstrated by the observation of protonated cytosine imino signals around 15 ppm and their corresponding amino signals around 10 ppm. Additional sharp and well-dispersed imino signals can be observed in the 13 to 14 ppm region, consistent with the formation of additional AU base pairs. Most of these signals are observed at 45ºC indicating the formation of a very stable structure (Figure 3j). Overall, the NMR spectra of the mix are consistent with formation of the expected parallel triplexes (see Figure 1). Overall, our experiments explain apparently contradictory previous data, agreeing with Crother's estimates[51] obtained for 66% GC triplexes and with Dervan's data [51,52]collected for 81% AT triplexes.

Very interestingly, hybrid triplexes d(Py)-d(Pu)·r(Py) are quite unstable compared to the r(Py)-d(Pu)·d(Py) ones, which suggests that triplexes with a 2:1 (DNA:RNA) stoichiometry show a topology r(Py)-d(Pu)·d(Py). Furthermore, triplexes r(Py)-d(Pu)·r(Py) which are intrinsically very stable, are expected to be disfavored in the cell, as its formation requires strand invasion, which would generate a strong topological stress in the DNA duplex and an unstable unpaired d(Py) strand. So, present results suggest that r(Py)-d(Pu)·d(Py), with TTS being the genomic DNA and the TFO being expressed RNAs will be the most stable triplex topology in the cell.

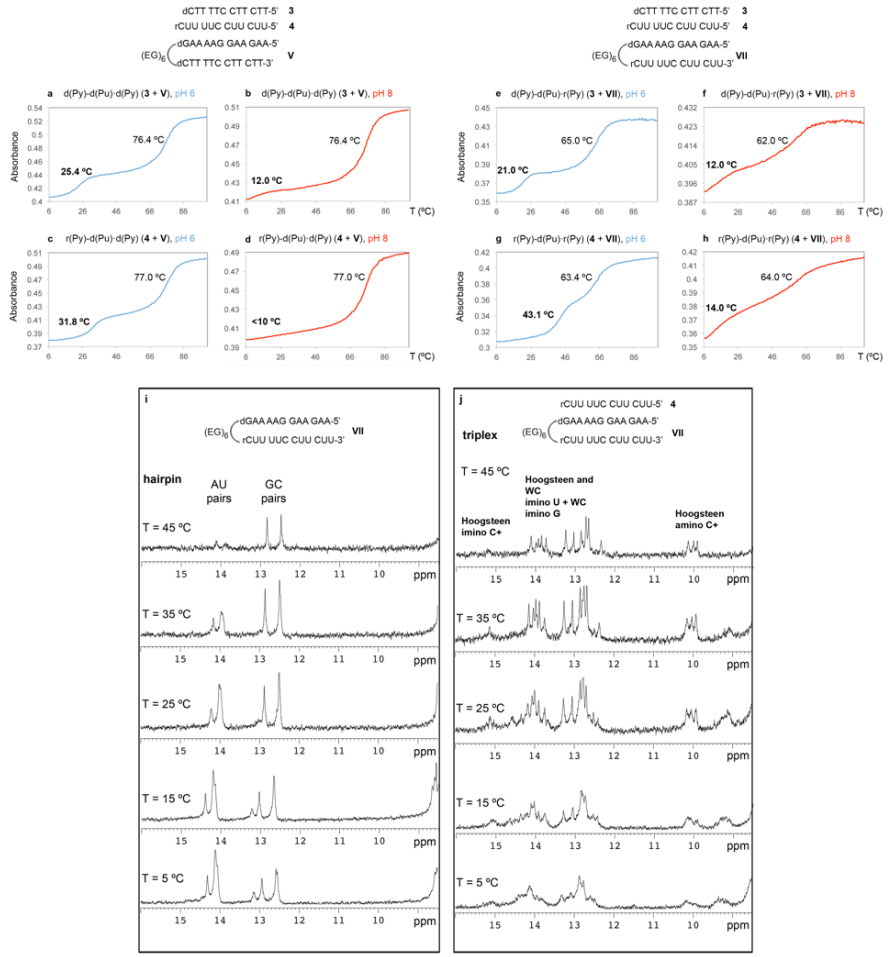


Figure 3. a-h. Thermal stability of d(Py)-d(Pu)·d(Py) (**3** + **V**), r(Py)-d(Pu)·d(Py) (**4** + **V**), d(Py)-d(Pu)·r(Py) (**3** + **VII**) and r(Py)-d(Pu)·r(Py) (**4** + **VII**) 12mer triplexes in 10 mM sodium cacodylate buffer (pH 6.0) containing 100 mM NaCl and 10 mM $MgCl_2$ (**a, c, e, g**) and in 10 mM sodium cacodylate buffer (pH 8.0) containing 100 mM NaCl and 10 mM $MgCl_2$ (**b, d, f, h**). Melting temperatures (T_m) for the duplex and triplex are indicated in each case (in bold: T_m corresponding to the triplex). **(i, j)** ^{1}H NMR spectra of the imino region of D·R hairpin **VII** (**i**) and r(Py)-d(Pu)·r(Py) triplex **4** + **VII (j)** acquired at 5 ºC, 15 ºC, 25 ºC, 35 ºC and 45 ºC in 30 mM phosphate buffer (pH 6.0) containing 100 mM NaCl and 10 mM $MgCl_2$.

The Structure and Dynamics of the Hybrid Triplexes. To gain structural and mechanistic insights on the structure and stability of hybrid triplexes, we performed a set of extensive molecular dynamics (MD) simulations (see Methods) of 6 triplexes in aqueous solution: r(Py)-d(Pu)·r(Py), r(Py)-d(Pu)·d(Py), d(Py)-d(Pu)·d(Py), d(Py)-d(Pu)·r(Py), r(Py)-r(Pu)·r(Py) and r(Py)-r(Pu)·d(Py). Hydrogen bonds (H-bonds) are not equally conserved in all triplexes, the differences being especially remarkable for the Hoogsteen ones, which agree with the experimental difference in stability between them. Thus, for the stable r(Py)-d(Pu)·r(Py) and r(Py)-d(Pu)·d(Py) triplexes (Figure 2), around 96% of Watson-Crick and 95% of Hoogsteen hydrogen bonds are preserved, while for "low-stability" r(Py)-r(Pu)·r(Py) and r(Py)-r(Pu)· d(Py) triplexes massive disruption of H-bonds is found, leading to structural disruption (Figure 4A). In fact, looking at the conservation of H-bonding we can define a "theoretical stability ordering": r(Py)-d(Pu)·r(Py)≥ r(Py)-d(Pu)·d(Py)> d(Py)-d(Pu)·d(Py) ≥ d(Py)-d(Pu)·r(Py)> d(Py)-d(Pu)·r(Py) ≈ r(Py)-r(Pu)·r(Py), which matches the experimental one (Figure 2). RMSd distributions (Figure 4B) confirm the stability ranking derived from H-bond analysis. Overall, agreement between theoretical and experimental estimates provide very strong confidence on the reliability of the atomistic MD simulations [63–66]. Analyzing the different sampled structures, we found that all hybrid triplexes show similar structures, which are not far from the homopolymeric ones, in terms of groove geometries and helical coordinates (see Figure 4; see Suppl. Table S1 and references [10–17,63]). There are however interesting differences depending on the stoichiometry and topology of the hybrid triplexes. For example, 2(RNA):1(DNA) triplexes are in general more "A-like" as reflected in lower twist-higher roll than 1(RNA):2(DNA) triplexes (see Suppl. Table S1). As expected from previous studies on DNA·RNA duplexes [63] RNA-strands maintain N-puckering, while DNA strands sample South and East regions (Suppl. Table S2). The situation does not change much depending on the placement of the nucleic acid strand in the Watson-Crick or Hoogsteen strands.

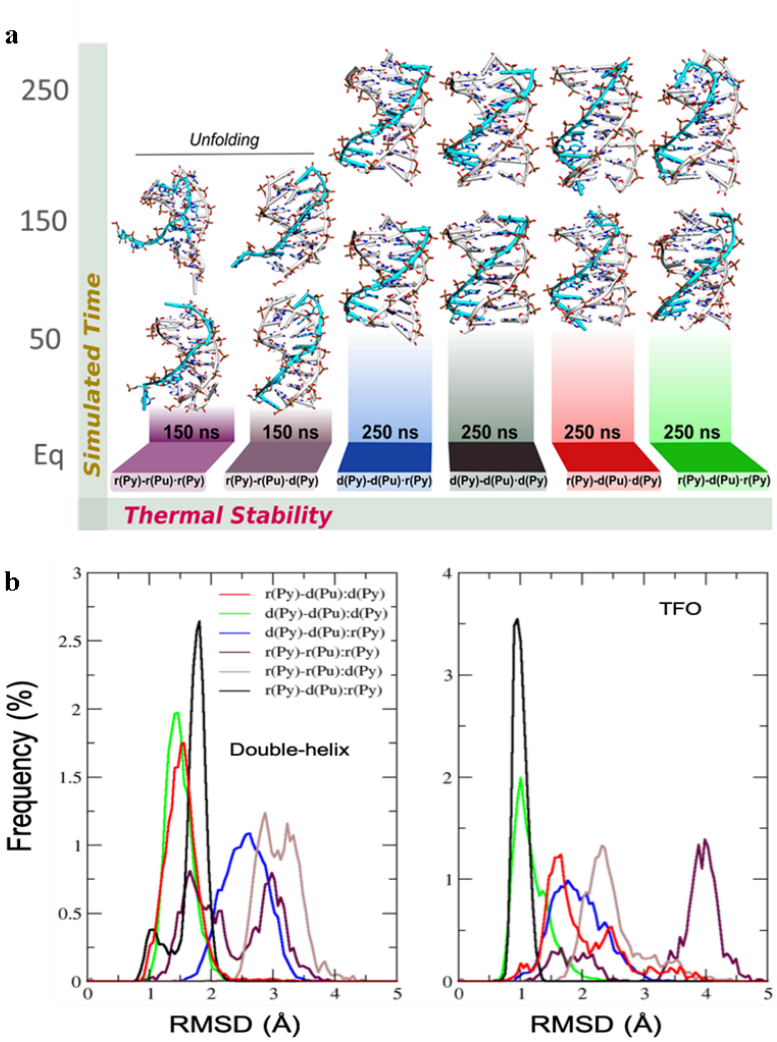


Figure 4. Selected results of the MD simulation of different hybrid triplexes. **a.** Evolution of the structure of the different triplexes along trajectories; **b.** Histograms of the RMSd in the TTS (left) and TFO (right) along the trajectories.

Development and Validation of a Predictor of the Stability of r(Py)-d(Pu)·d(Py) Triplexes. As discussed above, the most stable triplex (r(Py)-d(Pu)·r(Py)) is not expected to have a large prevalence in the cell out of R-loop constructs. However, the second most stable triplex: r(Py)-d(Pu)·d(Py) can be easily formed by pairing an RNA segment with genomic DNA. Following Robert and Crothers approach [50] we trained a simple nearest

neighbor model for r(Py)-d(Pu)·d(Py) to reproduce experimental data in a variety of triplexes (see Methods) in a variety of conditions. The refined method predicts experimental melting observables with root mean square errors around: 4.8 degrees (T_m), and 0.7 kcal/mol (melting free energy), improving dramatically the accuracy obtained by transferring Roberts-Crothers DNA triplex method (see Figure 5A). Our predictions also outperform the widely used Triplexator software [67] which is unable to detect all stable triplexes at a given temperature; see Figure 5B.

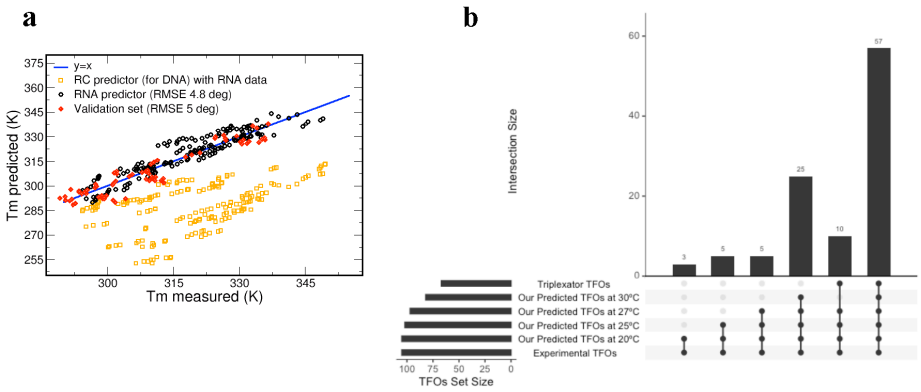


Figure 5. **a.** Predicted vs measured melting temperatures of triplexes. Values in yellow correspond to estimates obtained using Roberts-Crothers method for triplexes. Values in black to the data for the training set and that in red to data for a completely independent validation set (see Methods for details). **b.** Predicted TFOs from our model in comparison to Triplexator in the evaluation of stable triplexes at various temperatures. Results are shown as an intersecting upset plot.

To further validate the predictive power of our model, we designed a 50 nt polypyrimidine TFO (TFO **7**; Figure 6a) which, according to our method, should form stable triplexes (T_m values = 57 ºC at pH 6.5 and 48 ºC at pH 7.0) in the promoter region of the BRD7 gene. As shown in Figure 6a, synthetic TFO 7 interacts with a radiolabeled synthetic double-stranded DNA hairpin (**XIII**) comprising the target BRD7 polypurine sequence, forming a low-mobility complex. To validate the triplex nature of this complex, chromatin extracted from HeLa cells (see Materials and Methods for details) was treated with DNase I and proteinase K and sonicated to yield fragments of 200-300 nt (see Suppl. Figure S3). Two aliquots of this purified DNA were incubated with TFO 8, a biotinylated

version of TFO 7, at pH 5.5 and pH 7.0 in the presence of RNase H to digest putative R-loops. The streptavidin-retained DNA was eluted and identified by qPCR amplification, finding significant DNA recovery when using BRD7 promoter-specific primers amplifying a 92 nt region just around the target triplex region (92Nt primers; Figure 6b; red panel, Suppl. Figure S4 and Suppl. Table S3). Interestingly, a decrease in the pH led to an increase in DNA recovery, which is consistent with pH-dependent stability in C-C·G triplex formation as captured by our predictor. Similar results were obtained when using promoter-specific primers amplifying a larger region (217 nt) around the target triplex region (217Nt primers; Figure 6b; yellow panel). On the contrary, no recovery was observed when using intronic-specific primers (Figure 6b; green panel), or when genomic DNA fragments were incubated with a TFO lacking the sequence matching the target region (TFO **9**), confirming the specificity of the above-described results and the triplex nature of the complex predicted by our model.

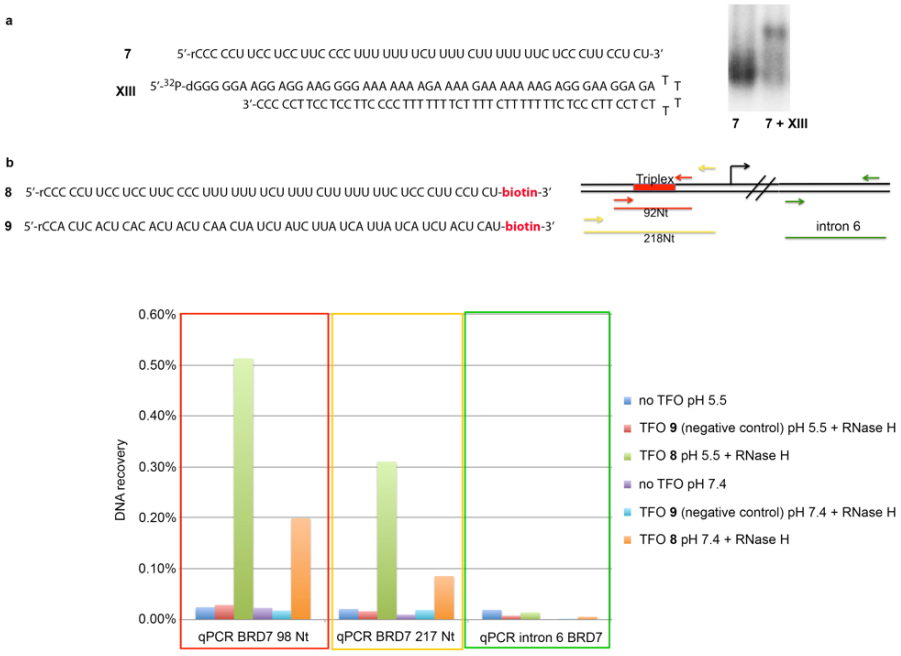


Figure 6. a. Electrophoretic mobility shift assays to analyze triplex formation between TFO **7** and hairpin **XIII**. **b.** Biotinylated TFOs used in this study (**8**: TFO targeting the polypurine A/G) site of BRD7 promoter region; **9**: negative control). Schematic depicting the position of the target triplex forming region and the primers used for DNA

amplification. Upon binding to streptavidin beads, associated DNA was analyzed by qPCR using promoter-specific [92 NT (in red) or 217 NT (in yellow)] or intronic-specific (in green) primers.

Formation of r(Py)-d(Pu)·d(Py) Triplexes in Human Cells. We used our predictor to screen for potential TFOs amongst annotated human lncRNAs and miRNAs from the gencode and miRbase [68,69] databases respectively (see Methods). We found a strong enrichment of TFO candidates (triplex Tm>30º) in both dataset in comparison to the population of expected TTSs from a random distribution (see random model in Methods) (Figure 7), suggesting potential r(Py)-d(Pu)·d(Py) triplex formation *in vivo*. When compared with the DNA-associated RNA isolated by Grummt and coworkers [45], we found that 44% or our predicted TFOs from lncRNAs and 51% from miRNAs were indeed found associated with DNA in a triplex structure. Both sets of TTS (from miRNA and lncRNA TFOs) were mapped to the human genome, and we observed an over-representation in promoters when analyzing stable triplexes (Tm > 30º), and in 5'UTR (in the case of miRNA's TFO) when analyzing very stable Triplexes (Tm>45ºC), suggesting that parallel r(Py)-d(Pu)·d(Py) triplexes form preferentially in regions important for the control of gene expression.

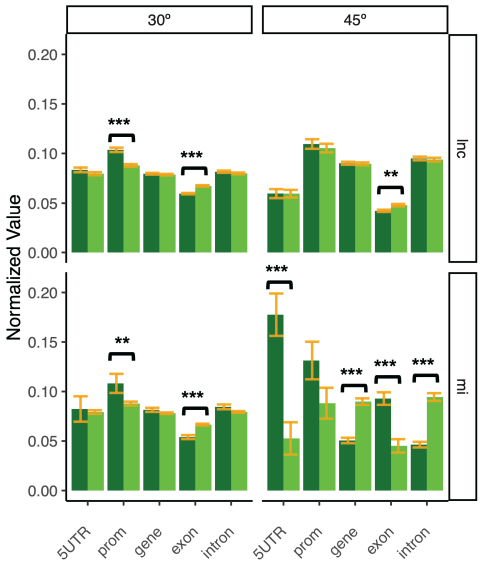


Figure 7. Overall representation across different genomic classes in lncRNAs (upper plots – dark green) and miRNAs (lower plots – dark green) against a random background model (light green), with the corresponding standard error bars (orange). RNA-DNA·DNA triplexes are shown as stable at 30° (left-most plots) or 45 °C (right-most plots) in the different regions. When a p-value is less than 0.05, it is flagged with one star (*). If a p-value is less than or equal to 0.01, it is flagged with two stars (**). If a p-value is less than or equal to 0.001 it is flagged with three stars (***).

GO analysis of the genes potentially controlled by RNA-DNA·DNA triplex formation with miRNAs and lncRNAs showed that these genes are frequently related to complex processes, such as development (see Suppl. Figures S5), with very significant hits in the development of the nervous system. It is tempting to speculate that stable triplexes generated by the binding of transcribed RNAs with genomic DNA can be involved in a fine-tuning regulatory mechanism, which was inherited from an ancient triplex-mediated DNA-RNA regulatory network. Note that this finding agrees well with the work from Pasquier et al. in Drosophila that showed that the genes targeted by TFOs were involved in development and morphogenesis [70].

Final comment on role in chromatin structure. To investigate triplex formation in the context of chromatin, we predicted the putative TFOs from lncRNA and miRNAs expressed in lymphoblastic cells [71] and compared the location of their target sites along the human genome with a genome-wide map of nucleosome occupancy in human lymphoblastoid cell line [72]. We observed a local minimum which coincides with the nucleosome dyads (Figure 8), suggesting a correlation between chromatin accessibility and triplex formation. These results which agree with previous findings by Maldonado et al [73], showed that triplexes could form away from the dyad, at the entry-exit site of the nucleosome, helping to fix the nucleosome array and in the case of very long lncRNA helping to approach in the space distant regions as suggested by Marti and co-workers [46].

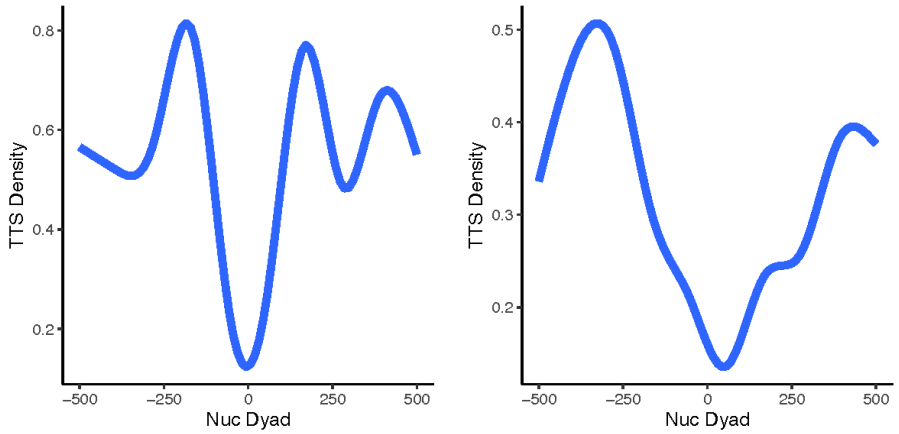


Figure 8. TTS densities from candidate TFOs centered at nucleosome dyads for lncRNAs (left) and miRNAs (right). The nucleosome maps are obtained from lymphoblastoid cells [72] and TFOs are originated from lncRNAs and miRNAs expressed in lymphoblastic cells.

DISCUSSION

A variety of parallel triplexes can be formed mixing complementary DNA and RNA strands, and a significant number of them can be stable under physiological conditions as predicted by state-of-the-art atomistic MD simulations and confirmed by melting and NMR experiments. In general, r(Py)-d(Pu)·r(Py) (*i.e.*, a poly-pyrimidine RNA as TFO and a hybrid DNA(Pur)·RNA(Pyr) as TTS) leads to the most stable structures, followed very closely by the r(Py)-d(Pu)·d(Py) triplex. The formation of the first triplex requires strand invasion of the DNA duplex, and the exposure of an unpaired pyrimidine-rich DNA strand, which would be disfavored in the cell. On the contrary, the r(Py)-d(Pu)·d(Py) triplex can be easily formed without the need for disruption of the DNA duplex, taking as TFO an expressed RNA sequence complementary with the purine strand of the duplex. A massive experimental effort allowed us to develop the first predictor for hybrid triplexes, which despite its simplicity, show quantitative accuracy. This predictor was used to determine all the potential triplexes in human lnc and miRNAs (TFO: expressed RNAs and TTS: genomic DNA). Calculations show a very large number of possible stable triplexes, much more than those predicted by random models. Potential

triplexes are concentrated in regulatory regions and UTRs, quite interestingly in genes that are related to the development, morphology and functioning of central nervous system, suggesting a potential role of triplexes in a RNA←→DNA mediated regulatory network. This work also suggests that, despite the fact that miRNAs are commonly known as post transcriptional regulators, their nuclear function as transcription regulators via triplex formation is more widespread than at first thought [74]. Furthermore, mapping potential triplex formation with chromatin structure, we found evidence suggesting a role of triplex formation in fixing nucleosome array probably protecting nucleosome from eviction and in the case of lncRNA helping, as suggested by others, to compact chromatin.

MATERIALS AND METHODS

Oligonucleotide Synthesis and Melting Experiments. Hairpins **I-XII** were synthesized as previously described [63]. Oligonucleotides **XIII** and **7-9** (Figure 7) were synthesized via solid phase synthesis using standard phosphoramidite methods (see Suppl. Methods for details).

Samples containing the required strands were heated to 90 °C and slowly cooled down to allow triplex formation in suitable buffers (see Suppl. Method for details). Melting experiments were performed by heating from low temperature to 100 °C at 0.5 °C/min, monitoring absorbance at 260 nm every 0.5 °C. Experiments were repeated for 5 μM, 8 μM, 12 μM, 18 μM and 22 μM oligonucleotide concentration to derive melting thermodynamic parameters from Van't Hoff analysis (see Suppl. Methods for details).

NMR spectroscopy. NMR spectra of hairpin **VII** were first recorded at a range of temperatures (5-45 °C). Later, the TFO was added and the mixture was heated (95 °C) and cooled down slowly, collecting spectra in the same temperature range. Spectra were acquired in a Bruker spectrometer operating at 600 MHz, equipped with cryoprobe and processed with the TOPSPIN software. Water suppression was achieved by including an excitation sculpting module in the pulse sequence [75], see Suppl. Methods for additional details.

Electrophoretic Mobility Shift Assays to Analyze Triplex Formation. TFO **7** was heated at 65 °C for 10 min to prevent self-aggregation and then quickly cooled on ice. Triplex formation was obtained by incubating the TFO with ^{32}P-labeled hairpin DNA **XVII** in a suitable buffer for 6 hours at 35 °C (see Suppl. Methods). Electrophoresis was done in a 15% native polyacrylamide gel at 8V/cm for 16 h at pH 5.5 (see Suppl. Methods for details). The gels were analyzed by phosphor-imaging.

Chromatin Preparation and gDNA purification. HeLa cells were grown to 90% confluency in T75 flask with Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% Fetal Bovine Serum (FBS) and 1% penicillin/streptomycin. Cells were trypsinized and nuclei were isolated using standard procedures (see Suppl. Methods) and lysed to obtain chromatin, which was then subjected to treatment with Proteinase K and DNAse I to yield fragments with an average size >10 Kb (Suppl. Figure S3a) followed by phenol/chloroform extraction and ethanol precipitation (see Suppl. Methods for details).

***In vitro* triplex Pull-down Assay.** Purified genomic DNA was sheared into 200-300 bp fragments by sonicating using a Bioruptor Pico (see Suppl. Methods for details). The resulting DNA mixtures were incubated with biotinylated TFO (**8** or **9**; Figure 7) at two different pH values (pH 7.4 and pH 5.5). TFO-associated DNA was captured by incubation with MyOne Streptavidin DynaBeads followed by treatment with RNase H and elution with suitable buffer (see Suppl. Methods for details).

Parameterization of the Nearest Neighbor Model for RNA-DNA$_2$ Triplex Stability. Following Roberts and Crothers [50] we determined the enthalpy of triplex formation by (eq. 1):

$$\Delta H = -\alpha_{cc}(CC) - \alpha_{uc}(UC + CU) - \alpha_{uu}(UU)$$

where (XX) refers to the number or dinucleotide steps of the type XpX in the TFO (CC, UC, CU or UU), and αs are fitted parameters.

The ΔG is determined as a function of the nucleotide content and the pH (eq. 2):

$$\Delta G(pH) = -\alpha_c(C) - \alpha_u(U) - \alpha'_{cc}(CC) + \delta + (C)(pH - 5.6)(\beta - \alpha''_{cc} \cdot (CC))$$

where all symbols in Greek letters are fitted parameters.

From ΔH and ΔG, we can extract the T_m using (eq. 3) [76].

$$Tm = \frac{310 \cdot \Delta H}{\Delta H - \Delta G - 310 \cdot R \cdot \ln\left(4/C_t\right)}$$

where R is the ideal gas constant and C_t is the concentration of the (hairpin) duplex and RNA strands.

The model was parametrized by non-linear fitting using ΔH and ΔG values obtained from our training set (Suppl. Figure S1) at different pH and concentrations (see Suppl. Methods for additional details).

Bioinformatics Scanning of Potential RNA-DNA·DNA Formation in Humans. We analyzed the triplex potential of annotated lncRNA and miRNA sequences from GENCODE [68] and miRbase [69] respectively. All annotated sequences were scanned with our stability predictor defining potential TFOs with a minimum length of 10 and a maximum of 30bps. The pH value was set at a default value of 7.0 and the c_{TFO} was set at a value of 12 μM. A T_m of 30°C was set as a threshold to classify stable fragments, while T_m of 45°C was considered to detect highly stable triplexes. In order to detect the formation of potential parallel triplex cores (see Suppl. Methods for additional details) we defined our TTSs as polypurine segments with perfect parallel alignment to the previously found TFOs. The extension of the core was evaluated by the melting predictor (see above) with a penalty equal to 10°C decrease in T_m per mismatch. The population of potential RNA-DNA·DNA triplexes in the human genome was compared with a random background model. In this model we randomly generated 1 million sequences which followed the base distributions found in the human genome. This allowed us to get a large enough sample for scanning candidate TFOs. We then obtained the target sites from our randomly generated TFOs and used them in the downstream analysis. The potential formation of triplexes in the human genome with our predicted TFOs was validated using

the DNA-associated RNA dataset published by Grummt et al. [45] and available in GEO repository under the accession number GSE120849.

Triplex Forming Oligonucleotide Fragment Analysis. Aiming to find the enriched regions where the RNA candidate TFOs would bind, we re-aligned the complementary sequences of our TFOs against the human genome. We used STAR (version 2.5.3a, [77]) mapping the candidate TTSs to the hg38 assembly of the human genome. The aligned reads were mapped to the corresponding annotation files and classified accordingly. The annotations for genes, exons, transcripts, and UTR regions were obtained from GENCODE (Release 35 (GRCh38.p13)). Promoters were defined as the regions from the transcription start sites up until 1kb upstream. Counts for different features at the gene, promoter regions, exon, transcript and UTR-level resulting from mapping based on the reads were determined separately using the Bioconductor package Rsubread's function featureCounts (version 2.0.1) [78]. To further investigate the role of our TTSs, we mapped the obtained promoter sites to their associated genes and performed a Gene Ontology (GO) Analysis using g:Profiler [79]. A Benjamin-Hochberg FDR index < 0.05 was set to assess significance corrected from multiple test biases. The GO biological processes of the annotated genes were investigated for terms with size > 15 and < 2700 in order to avoid mappings to large pathways that are of limited value and increase statistical value when removing small pathways [80].

Triplex formation in the context of chromatin. The nucleosome map in human lymphoblastoid cell lines was obtained analyzing MNase-seq data (Accession number GSE36979) from Gaffney et al.[81]. The reads were processed with the nucleR package [82] as follows: mapped fragments were trimmed to 50 bp maintaining the original center and transformed to reads per million. Noise was filtered through Fast Fourier Transform, keeping 2% of the principal components, and peak calling was performed using the following parameters: peak width 147 bp, peak detection threshold 35%, maximum overlap of 45 bp, dyad length 60 bp. The lncRNA and miRNAs expressed in lymphoblastic cells were obtained from RNA-seq and small RNA-seq data (accession numbers E-MATB-8300 and E-MTAB-8301)[83].

Structural Models and Molecular Dynamics (MD) Simulations. Starting conformations for the six triplexes considered here: r(Py)-d(Pu)·r(Py), r(Py)-d(Pu)·d(Py),

d(Py)-d(Pu)·d(Py), d(Py)-d(Pu)·r(Py), r(Py)-r(Pu)·r(Py) and r(Py)-r(Pu)·d(Py) were built from DNA triplex structures [17,18,84,85]. Systems were solvated with waters, neutralized with Na^+ adding 100 mM additional NaCl. The size of the final triclinic box was approximately 60 Å × 60 Å × 60 Å. Simulation systems were optimized and slowly heated and equilibrated for 50 ns prior to production that extended from 250 ns in the isothermal isobaric ensemble (NPT; T=310 K and P= 1atm). Long-range electrostatic interactions were calculated with the particle mesh Ewald method (PME) with a real space cut-off of 12 Å and periodic boundary conditions in the three directions of Cartesian space were used [86]. Constant temperature was imposed using Langevin dynamics [87] with a damping coefficient of 1 ps, while pressure was maintained with Langevin-Piston dynamics [88] with a 200 fs decay period and a 50 fs time constant. LINCS [89] was used to maintain covalent bonds at equilibrium distance, allowing the use of 2 fs integration step. Parmbsc1 was used to describe DNA interactions [64–66], while RNA was described using recent RNA force-field published by Tan, D. et *al* [90]. Water molecules were represented by the TIP3P [91] model, while ions were modeled by Dang's parameters [92].

All MD simulations were performed using *GRO*ningen *MA*chine for *C*hemical *S*imulations (GROMACS) 2016 code [93]. Analysis of the trajectories were perfomed using GROMACS. Coordinates of the systems were collected every 5 ps of the production trajectory. Analyses were carried out using GROMACS analysis tools, VMD 1.9 Software [94], Curves+ [95] and NAFlex [96] and BIGNAsim analysis tools [97,98]. Trajectories were stored in our BIGNAsim database [97,98] following FAIR data standards as described elsewhere [98].

Acknowledgments

V.G. thanks the European Molecular Biology Organization (EMBO) for financial support (ALTF 103-2018) and "Juan De La Cierva Fellowship" (IJC2019-040468-I / 25A04100). M.O. thanks Spanish Ministry of Science [RTI2018-096704-B-100]; European Research Council (ERC SimDNA), MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona); the Biomolecular and Bioinformatics Resources Platform (ISCIII PT 13/000/0030 co-funded by the Fondo Europeo de Desarrollo Regional [FEDER]) and the H2020 BioExcel Center of Excellence.

REFERENCES

1. Pauling, L. & Corey, R. B. A Proposed Structure For The Nucleic Acids. *Proceedings of the National Academy of Sciences* **39**, (1953).
2. Felsenfeld, G. & Rich, A. Studies on the formation of two- and three-stranded polyribonucleotides. *BBA - Biochimica et Biophysica Acta* **26**, (1957).
3. Frank-Kamenetskii, M. D. & Mirkin, S. M. Triplex DNA structures. *Annual Review of Biochemistry* vol. 64 Preprint at https://doi.org/10.1146/annurev.bi.64.070195.000433 (1995).
4. Potaman, V. N. & Sinden, R. R. Stabilization of Triple-Helical Nucleic Acids by Basic Oligopeptides. *Biochemistry* **34**, (1995).
5. Robles, J. *et al.* Nucleic Acid Triple Helices: Stability Effects of Nucleobase Modifications. *Curr Org Chem* **6**, (2005).
6. Waring, M. J. *DNA-Targeting Molecules as Therapeutic Agents. DNA-targeting Molecules as Therapeutic Agents* (2018). doi:10.1039/9781788012928.
7. Scaria, P. V. & Shafer, R. H. Calorimetric analysis of triple helices targeted to the d(G3A4G3)·d(C3T4C3) duplex. *Biochemistry* **35**, (1996).
8. Chandler, S. P. & Fox, K. R. Specificity of antiparallel DNA triple helix formation. *Biochemistry* **35**, (1996).
9. Jaumot, J., Aviñó, A., Eritja, R., Tauler, R. & Gargallo, R. Resolution of parallel and antiparallel oligonucleotide triple helices formation and melting processes by multivariate curve resolution. *J Biomol Struct Dyn* **21**, (2003).
10. Arnott, S., Bond, P. J., Selsing, E. & Smith, P. J. C. Models of Triple-Stranded Polynucleotides with Optimised Stereochemistry. *Nucleic Acids Res* **3**, (1976).
11. Macaya, R. F., Schultze, P. & Feigon, J. Sugar Conformations in Intramolecular DNA Triplexes Determined by Coupling Constants Obtained by Automated Simulation of P. COSY Cross Peaks. *J Am Chem Soc* **114**, (1992).
12. Raghunathan, G., Miles, H. T. & Sasisekharan, V. Symmetry and Molecular Structure of a DNA Triple Helix: d(T)n·d(A)n·d(T)n. *Biochemistry* **32**, (1993).
13. Howard, F. B. *et al.* Structure of d(T)n·d(A)n·d(T)n: The DNA Triple Helix Has B-Form Geometry with C2′-Endo Sugar Pucker. *Biochemistry* **31**, (1992).
14. Radhakrishnan, I. & Patel, D. J. Solution structure and hydration patterns of a Pyrimidine·Purine·Pyrimidine DNA triplex containing a novel T·CG base-triple. *J Mol Biol* **241**, (1994).
15. Radhakrishnan, I. & Patel, D. J. Hydration sites in purine·purine·pyrimidine and pyrimidine·purine·pyrimidine DNA triplexes in aqueous solution. *Structure* **2**, (1994).
16. Bomet, O. & Lancelot, G. Solution structure of a selectively 13c-labeled intramolecular dna triplex. *J Biomol Struct Dyn* **12**, (1995).
17. Shields, G. C., Laughton, C. A. & Orozco, M. Molecular dynamic simulations of the d(T·A·T) triple helix. *J Am Chem Soc* **119**, (1997).
18. Soliva, R., Laughton, C. A., Luque, F. J. & Orozco, M. Molecular dynamics simulations in aqueous solution of triple helices containing d(G·C·C) trios. *J Am Chem Soc* **120**, (1998).
19. Guieysse, A. L., Praseuth, D. & Helene, C. Identification of a triplex DNA-binding protein from human cells. *J Mol Biol* **267**, (1997).
20. Jiménez-García, E. *et al.* The GAGA factor of Drosophila binds triple-stranded DNA. *Journal of Biological Chemistry* **273**, (1998).

21. Knauert, M. P. & Glazer, P. M. Triplex forming oligonucleotides: Sequence-specific tools for gene targeting. *Human Molecular Genetics* vol. 10 Preprint at https://doi.org/10.1093/hmg/10.20.2243 (2001).
22. Giovannangeli, C. & Hélène, C. Triplex technology takes off. *Nature Biotechnology* vol. 18 Preprint at https://doi.org/10.1038/82348 (2000).
23. Bacolla, A., Wang, G. & Vasquez, K. M. New Perspectives on DNA and RNA Triplexes As Effectors of Biological Activity. *PLoS Genetics* vol. 11 Preprint at https://doi.org/10.1371/journal.pgen.1005696 (2015).
24. Buske, F. A., Mattick, J. S. & Bailey, T. L. Potential in vivo roles of nucleic acid triple-helices. *RNA Biology* vol. 8 Preprint at https://doi.org/10.4161/rna.8.3.14999 (2011).
25. Goñi, J. R., de la Cruz, X. & Orozco, M. Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res* **32**, (2004).
26. Goñi, J. R., Vaquerizas, J. M., Dopazo, J. & Orozco, M. Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. *BMC Genomics* **7**, (2006).
27. Van Dongen, M. J. P. *et al.* Structure and mechanism of formation of the H-y5 isomer of an intramolecular DNA triple helix. *Nat Struct Biol* **6**, (1999).
28. Rogers, F. A., Lloyd, J. A. & Glazer, P. M. Triple-forming oligonucleotides as potential tools for modulation of gene expression. *Current Medicinal Chemistry - Anti-Cancer Agents* vol. 5 Preprint at https://doi.org/10.2174/1568011054222300 (2005).
29. Alvarez-Salas, L. Nucleic Acids as Therapeutic Agents. *Curr Top Med Chem* **8**, (2008).
30. Guntaka, R. V., Varma, B. R. & Weber, K. T. Triplex-forming oligonucleotides as modulators of gene expression. *International Journal of Biochemistry and Cell Biology* vol. 35 Preprint at https://doi.org/10.1016/S1357-2725(02)00165-6 (2003).
31. Duval-Valentin, G., Thuong, N. T. & Hélène, C. Specific inhibition of transcription by triple helix-forming oligonucleotides. *Proc Natl Acad Sci U S A* **89**, (1992).
32. Cooney, M., Czernuszewicz, G., Postel, E. H., Flint, S. J. & Hogan, M. E. Site-specific oligonucleotide binding represses transcription of the human c-myc gene in vitro. *Science (1979)* **241**, (1988).
33. Grigoriev, M. *et al.* A triple helix-forming oligonucleotide-intercalator conjugate acts as a transcriptional repressor via inhibition of NF κB binding to interleukin-2 receptor αRegulatory sequence. *Journal of Biological Chemistry* **267**, (1992).
34. Conde, J., Oliva, N., Atilano, M., Song, H. S. & Artzi, N. Self-assembled RNA-triple-helix hydrogel scaffold for microRNA modulation in the tumour microenvironment. *Nat Mater* **15**, (2016).
35. Besch, R., Giovannangeli, C., Schuh, T., Kammerbauer, C. & Degitz, K. Characterization and quantification of triple helix formation in chromosomal DNA. *J Mol Biol* **341**, (2004).
36. Devi, G., Zhou, Y., Zhong, Z., Toh, D. F. K. & Chen, G. RNA triplexes: From structural principles to biological and biotech applications. *Wiley Interdisciplinary Reviews: RNA* vol. 6 Preprint at https://doi.org/10.1002/wrna.1261 (2015).
37. Joseph, J., Kandala, J. C., Veerapanane, D., Weber, K. T. & Guntaka, R. V. Antiparallel polypurine phosphorothioate oligonucleotides form stable triplexes

with the rat α1(I) collagen gene promoter and inhibit transcription in cultured rat fibroblasts. *Nucleic Acids Res* **25**, (1997).

38. Postel, E. H., Flint, S. J., Kessler, D. J. & Hogan, M. E. Evidence that a triplex-forming oligodeoxyribonucleotide binds to the c-myc promoter in HeLa cells, thereby reducing c-myc mRNA levels. *Proc Natl Acad Sci U S A* **88**, (1991).
39. Li, Y., Syed, J. & Sugiyama, H. RNA-DNA Triplex Formation by Long Noncoding RNAs. *Cell Chemical Biology* vol. 23 Preprint at https://doi.org/10.1016/j.chembiol.2016.09.011 (2016).
40. Postepska-Igielska, A. *et al.* LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure. *Mol Cell* **60**, (2015).
41. Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, (2007).
42. Li, T., Mo, X., Fu, L., Xiao, B. & Guo, J. Molecular mechanisms of long noncoding RNAs on gastric cancer. *Oncotarget* **7**, (2016).
43. Grote, P. *et al.* The Tissue-Specific lncRNA Fendrr Is an Essential Regulator of Heart and Body Wall Development in the Mouse. *Dev Cell* **24**, (2013).
44. Sridhar, B. *et al.* Systematic Mapping of RNA-Chromatin Interactions In Vivo. *Current Biology* **27**, (2017).
45. Cetin, N. S. *et al.* Isolation and genome-wide characterization of cellular DNA:RNA triplex structures. *Nucleic Acids Res* **47**, (2019).
46. Farabella, I., Di Stefano, M., Soler-Vila, P., Marti-Marimon, M. & Marti-Renom, M. A. Three-dimensional genome organization via triplex-forming RNAs. *Nat Struct Mol Biol* **28**, (2021).
47. Soibam, B. & Zhamangaraeva, A. LncRNA:DNA triplex-forming sites are positioned at specific areas of genome organization and are predictors for Topologically Associated Domains. *BMC Genomics* **22**, (2021).
48. Mergny, J. L. *et al.* Sequence Specificity in Triple-Helix Formation: Experimental and Theoretical Studies of the Effect of Mismatches on Triplex Stability. *Biochemistry* **30**, (1991).
49. Singleton, S. F. & Dervan, P. B. Equilibrium Association Constants for Oligonucleotide-Directed Triple Helix Formation at Single DNA Sites: Linkage to Cation Valence and Concentration. *Biochemistry* **32**, (1993).
50. Roberts, R. W. & Crothers, D. M. Prediction of the stability of DNA triplexes. *Proc Natl Acad Sci U S A* **93**, (1996).
51. Roberts, R. W. & Crothers, D. M. Stability and properties of double and triple helices: Dramatic effects of RNA or DNA backbone composition. *Science (1979)* **258**, (1992).
52. Han, H. & Dervan, P. B. Sequence-specific recognition of double helical RNA and RNA·DNA by triple helix formation. *Proc Natl Acad Sci U S A* **90**, (1993).
53. Otto, C., Thomas, G. A., Rippe, K., Jovin, T. M. & Peticolas, W. L. The Hydrogen-Bonding Structure in Parallel-Stranded Duplex DNA Is Reverse Watson-Crick. *Biochemistry* **30**, (1991).
54. Cubero, E., Luque, F. J. & Orozco, M. Theoretical studies of d(A:T)-based parallel-stranded DNA duplexes. *J Am Chem Soc* **123**, (2001).
55. Cubero, E., Abrescia, N. G. A., Subirana, J. A., Luque, F. J. & Orozco, M. Theoretical Study of a New DNA Structure: The Antiparallel Hoogsteen Duplex. *J Am Chem Soc* **125**, (2003).

56. Cubero, E. *et al.* Hoogsteen-based parallel-stranded duplexes of DNA. Effect of 8-amino-purine derivatives. *J Am Chem Soc* **124**, (2002).
57. Radwan, M. M. & Wilson, H. R. Fibre and molecular structure of thymidylyl-3′,5′-deoxyadenosine. *Int J Biol Macromol* **4**, (1982).
58. Abrescia, N. G. A., Thompson, A., Huynh-Dinh, T. & Subirana, J. A. Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. *Proc Natl Acad Sci U S A* **99**, (2002).
59. Aishima, J. *et al.* A Hoogsteen base pair embedded in undistorted B-DNA. *Nucleic Acids Research* vol. 30 Preprint at https://doi.org/10.1093/nar/gkf661 (2002).
60. Spiegel, K., Rothlisberger, U. & Carloni, P. Duocarmycins binding to DNA investigated by molecular simulation. *Journal of Physical Chemistry B* **110**, (2006).
61. Cubero, E., Luque, F. J. & Orozco, M. Theoretical study of the Hoogsteen-Watson-Crick junctions in DNA. *Biophys J* **90**, (2006).
62. Brahms, S., Brahms, J. & Van Holde, K. E. Nature of conformational changes in poly[d(A-T).d(A-T)] in the premelting region. *Proc Natl Acad Sci U S A* **73**, (1976).
63. Terrazas, M. *et al.* The Origins and the Biological Consequences of the Pur/Pyr DNA·RNA Asymmetry. *Chem* **5**, (2019).
64. Ivani, I. *et al.* Parmbsc1: A refined force field for DNA simulations. *Nat Methods* **13**, (2015).
65. Dans, P. D., Walther, J., Gómez, H. & Orozco, M. Multiscale simulation of DNA. *Current Opinion in Structural Biology* vol. 37 Preprint at https://doi.org/10.1016/j.sbi.2015.11.011 (2016).
66. Dans, P. D. *et al.* How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res* **45**, (2017).
67. Buske, F. A., Bauer, D. C., Mattick, J. S. & Bailey, T. L. Triplexator: Detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res* **22**, (2012).
68. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, (2019).
69. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. MiRBase: From microRNA sequences to function. *Nucleic Acids Res* **47**, (2019).
70. Pasquier, C., Agnel, S. & Robichon, A. The mapping of predicted triplex DNA: RNA in the Drosophila genome reveals a prominent location in development- and morphogenesis-related genes. *G3: Genes, Genomes, Genetics* 7, (2017).
71. Florian, R. T. *et al.* Unstable TTTTA/TTTCA expansions in MARCH6 are associated with Familial Adult Myoclonic Epilepsy type 3. *Nat Commun* **10**, (2019).
72. Gaffney, D. J. *et al.* Controls of Nucleosome Positioning in the Human Genome. *PLoS Genet* **8**, (2012).
73. Maldonado, R., Schwartz, U., Silberhorn, E. & Längst, G. Nucleosomes Stabilize ssRNA-dsDNA Triple Helices in Human Cells. *Mol Cell* **73**, (2019).
74. Paugh, S. W. *et al.* MicroRNAs Form Triplexes with Double Stranded DNA at Sequence-Specific Binding Sites; a Eukaryotic Mechanism via which microRNAs Could Directly Alter Gene Expression. *PLoS Comput Biol* **12**, (2016).
75. Stott, K., Keeler, J., Hwang, T. L., Shaka, A. J. & Stonehouse, J. Excitation Sculpting in High-Resolution Nuclear Magnetic Resonance Spectroscopy:

Application to Selective NOE Experiments. *Journal of the American Chemical Society* vol. 117 Preprint at https://doi.org/10.1021/ja00119a048 (1995).

76. Marky, L. A. & Breslauer, K. J. Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves. *Biopolymers* **26**, (1987).
77. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, (2013).
78. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* **47**, (2019).
79. Raudvere, U. *et al.* G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**, (2019).
80. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc* **14**, (2019).
81. Gaffney, D. J. *et al.* Controls of Nucleosome Positioning in the Human Genome. *PLoS Genet* **8**, (2012).
82. Flores, O. & Orozco, M. nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics* **27**, 2149–2150 (2011).
83. Florian, R. T. *et al.* Unstable TTTTA/TTTCA expansions in MARCH6 are associated with Familial Adult Myoclonic Epilepsy type 3. *Nat Commun* **10**, (2019).
84. Gotfredsen, C. H., Schultze, P. & Feigon, J. Solution structure of an intramolecular: Pyrimidine-purine-pyrimidine triplex containing an RNA third strand. *J Am Chem Soc* **120**, (1998).
85. Ruszkowska, A., Ruszkowski, M., Hulewicz, J. P., Dauter, Z. & Brown, J. A. Molecular structure of a U•A-U-rich RNA triple helix with 11 consecutive base triples. *Nucleic Acids Res* **48**, (2020).
86. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* **98**, (1993).
87. Grest, G. S. & Kremer, K. Molecular dynamics simulation for polymers in the presence of a heat bath. *Phys Rev A (Coll Park)* **33**, (1986).
88. Feller, S. E., Zhang, Y., Pastor, R. W. & Brooks, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. *J Chem Phys* **103**, (1995).
89. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for molecular simulations. *J Comput Chem* **18**, (1997).
90. Tan, D., Piana, S., Dirks, R. M. & Shaw, D. E. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc Natl Acad Sci U S A* **115**, (2018).
91. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, (1983).
92. Dang, L. X. Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study. *J Am Chem Soc* **117**, (1995).
93. Páll, S., Abraham, M. J., Kutzner, C., Hess, B. & Lindahl, E. Tackling exascale software challenges in molecular dynamics simulations with GROMACS. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 8759 (2015).
94. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J Mol Graph* **14**, (1996).

95. Blanchet, C., Pasi, M., Zakrzewska, K. & Lavery, R. CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res* **39**, (2011).
96. Hospital, A. *et al.* NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res* **41**, (2013).
97. Hospital, A. *et al.* BIGNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res* **44**, (2016).
98. Hospital, A., Battistini, F., Soliva, R., Gelpí, J. L. & Orozco, M. Surviving the deluge of biosimulation data. *Wiley Interdisciplinary Reviews: Computational Molecular Science* vol. 10 Preprint at https://doi.org/10.1002/wcms.1449 (2020).

Chapter 5. The physical properties of RNA duplexes

DNA and its sequence-dependent and physical properties have been widely described through experimental techniques and extensive molecular dynamic (MD) simulations (1). These studies allowed the development of a systematic database that described the sequence-dependent properties of duplex DNA. These mesoscopic descriptors which help untangle the structure and flexibility of DNA together with its role in many biological processes (2, 3), have been used throughout this thesis for predicting TF binding sites and nucleosome positioning along the genome. Meanwhile, the A-form of the right-handed double helical RNA (RNA_2), the most studied and common conformation for RNA, lacks the same level of detail in its understanding, which hinders the theoretical prediction of the properties of regulatory RNAs, such as those by RNA interference. It is thus of great interest to perform a systematic effort to describe with the same level of detail and accuracy the properties of double-stranded RNA (dsRNA; (4–7)).

In this chapter we present the results obtained by MD simulations to decipher the RNA_2 sequence-dependent properties, derived from statistically accurate analyses of the sampled cartesian and helical spaces. We also compared these results with previously obtained data from the DNA_2 properties and developed a very simple mesoscopic-correlated harmonic model to reproduce long-RNA_2 conformations, mimicking that previously developed for DNA_2. Overall, this project addressed some of the differences found in the structure and flexibility of the RNA duplex in comparison to the properties found in the DNA counterpart.

5.1. Differences between RNA and DNA

For many years RNA was only thought to carry the genetic information needed to translate into proteins. However, recent studies have shown that RNA acts as a global regulator of cellular processes (8), has a structural role in chromatin organization (9–11) and even presents catalytic properties (12, 13). All these effects are coded in its sequence, and more interesting in its structure and physical properties of its most abundant form: the A- duplex.

It has been known for decades that significant differences exist between DNA and RNA duplexes (4, 5). The first difference resides in

the fact that the RNA duplex is intramolecular, which means that no perfect self-pairing is possible, thus the duplex can present many mismatches and even loops that modulate its structure and physical properties. The second difference obeys to the difference in the sugar, while the DNA contains deoxyribose that favors a S-puckering of the furanose ring, the RNA contains ribose which prefers a N-puckering. As described elsewhere (14) this difference justifies the adoption of two different canonical conformations: the B- form for DNA_2 and the A-form for RNA_2, see Figure 5.1B). The A-form is more compact that the B- one and has completely different groove architectures (5, 14) which generate dramatic differences in the pattern of interaction with other molecules, particularly with proteins. Previous experimental assays supported the idea that the B-form DNA structure was more flexible than the A-form RNA counterpart (15, 16). However, some more recent studies have shown that this is not necessarily the case as it depends on the way in which flexibility is defined (4). Thus, as protein-nucleic acids are very dependent on nucleic acids rearrangements (see Chapter 1) accurate description of RNA equilibrium properties and flexibility characteristics is of paramount importance.

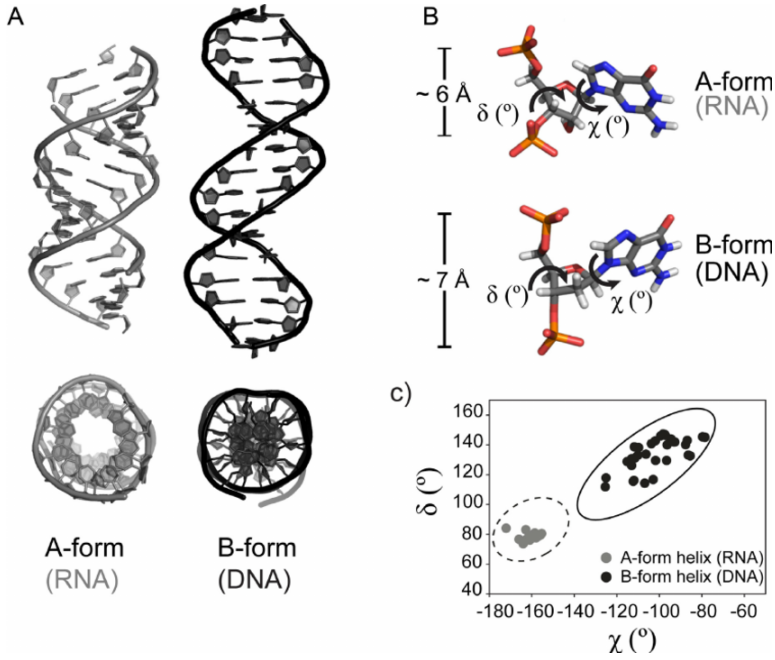


Figure 5.1. A) Helical geometries for the canonical RNA A-form (left) and DNA B-form (right). A view from the top is presented underneath each structure. B) Differences in sugar puckering and P-P distances between A- and B-form helices. A close-up view of a single nucleotide in a canonical A-helix (top panel, RNA) and a B-helix (bottom panel, DNA) color coded by atom type. Image adapted from (17).

5.2. Molecular Dynamic Simulations

In a former study carried out by the ABC (Ascona B-DNA Consortium), physical properties from all the 136 unique DNA tetranucleotide base were derived from MD simulations, using a set of 13 sequences called miniABC (1). These studies allowed the community to determine that DNA properties depend on the underlying sequence, elucidating polymorphism, detecting deviations from the canonical B-DNA form, and portraying nearest neighbor effects of the central base pair step (4, 18). Improvements in the force fields that described RNA (19, 20) opened the door to perform similar efforts and acquire the same level of detail and accuracy to describe double-stranded RNA. To this end, we performed 1 μs long MD simulations for the same set of 13 sequences but for RNA duplexes, starting from the canonical A-form, using the state-of-the-art FF and simulation protocols. These MD simulations were analyzed to detect the stability of the RNA duplexes and the variations at the global and local level of structure and dynamics. The analysis of the ensembles of trajectories obtained for RNA sequences has helped to characterize the sequence-dependent properties of RNA duplexes by mapping the cartesian ensembles into helical coordinates. After a detailed statistical analysis on the data, equilibrium parameters and stiffness matrices were compared to the DNA counterpart. The comparison between DNA_2 and RNA_2 physical properties allowed us to detect different behaviors in the distributions of the helical parameters and to develop for the first time, an RNA specific mesoscopic model.

In this chapter we present a massive computational effort by using MD simulations to better characterize and expand the knowledge on the physical and mechanical properties of the RNA_2 and its sequence-dependent properties at the tetramer level. We characterized the base movements discovering that, unlike DNA duplexes, RNA did not show any polymorphisms in helical geometries in the tetranucleotide sequence contexts, following a more harmonic behavior which is less dependent on the sequence context. We also compared the flexibility in the cartesian and helical parameter space, discovering that surprisingly RNA is not stiffer than DNA, as usually claimed, but rather depends on the sequence context and the considered movement.

Finally, the more harmonic behavior of RNA allowed us to adapt and simplify the already existing mathematical formalism of the

DNA mesoscopic model (21, 22) and develop a new mesoscopic-correlated harmonic model to predict the structure of long RNA duplexes. This model could reproduce with high accuracy the local and global conformation and deformability of long-RNA2, with a precision comparable to atomistic conformational ensembles.

Publication:

Federica Battistini, Alba Sala, Adam Hospital, and Modesto Orozco. Sequence-Dependent Properties of the RNA Duplex. J. Chem. Inf. Model. 2023, 63, 16, 5259–5271, https://doi.org/10.1021/acs.jcim.3c00741.

Supplementary material for this article can be found in the Annex.

References

1. Dans,P.D., Balaceanu,A., Pasi,M., Patelli,A.S., Petkevičiūtė,D., Walther,J., Hospital,A., Bayarri,G., Lavery,R., Maddocks,J.H., *et al.* (2019) The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. *Nucleic Acids Res*, **47**, 11090–11102.
2. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C., *et al.* (2014) μABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res*, **42**.
3. Walther,J., Dans,P.D., Balaceanu,A., Hospital,A., Bayarri,G. and Orozco,M. (2020) A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Res*, **48**, e29–e29.
4. Noy,A., Pérez,A., Lankas,F., Javier Luque,F. and Orozco,M. (2004) Relative flexibility of DNA and RNA: A molecular dynamics study. *J Mol Biol*, **343**.
5. Pérez,A., Noy,A., Lankas,F., Luque,F.J. and Orozco,M. (2004) The relative flexibility of B-DNA and A-RNA duplexes: Database analysis. *Nucleic Acids Res*, **32**.
6. Faustino,I., Pérez,A. and Orozco,M. (2010) Toward a consensus view of duplex RNA flexibility. *Biophys J*, **99**.
7. Dans,P.D., Gallego,D., Balaceanu,A., Darré,L., Gómez,H. and Orozco,M. (2019) Modeling, Simulations, and Bioinformatics at the Service of RNA Structure. *Chem*, **5**.
8. Spitale,R.C. and Incarnato,D. (2023) Probing the dynamic RNA structurome and its functions. *Nat Rev Genet*, **24**.
9. Li,X. and Fu,X.D. (2019) Chromatin-associated RNAs as facilitators of functional genomic interactions. *Nat Rev Genet*, **20**.
10. Sridhar,B., Rivas-Astroza,M., Nguyen,T.C., Chen,W., Yan,Z., Cao,X., Hebert,L. and Zhong,S. (2017) Systematic Mapping of RNA-Chromatin Interactions In Vivo. *Current Biology*, **27**.
11. Farabella,I., Stefano,M. Di, Soler-Vila,P., Marti-Marimon,M. and Marti-Renom,M.A. (2021) Three-dimensional genome organization via triplex-forming RNAs. *Nat Struct Mol Biol*, **28**.
12. Alonso,D. and Mondragón,A. (2021) Mechanisms of catalytic RNA molecules. *Biochem Soc Trans*, **49**.
13. Park,S. V., Yang,J.S., Jo,H., Kang,B., Oh,S.S. and Jung,G.Y. (2019) Catalytic RNA, ribozyme, and its applications in synthetic biology. *Biotechnol Adv*, **37**.

14. Soliva,R., Luque,F.J., Alhambra,C. and Orozco,M. (1999) Role of sugar re-puckering in the transition of a and b forms of dna in solution. a molecular dynamics study. *J Biomol Struct Dyn*, **17**.
15. Hagerman,P.J. (1997) FLEXIBILITY OF RNA. *Annu Rev Biophys Biomol Struct*, **26**, 139–156.
16. Hagerman,P.J. (1988) Flexibility of DNA. *Annu Rev Biophys Biophys Chem*, **17**, 265–286.
17. Anosova,I., Kowal,E.A., Dunn,M.R., Chaput,J.C., Horn,W.D.V. and Egli,M. (2016) SURVEY AND SUMMARY the structural diversity of artificial genetic polymers. *Nucleic Acids Res*, **44**.
18. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A., *et al.* (2015) Parmbsc1: A refined force field for DNA simulations. *Nat Methods*, **13**.
19. Banáš,P., Hollas,D., Zgarbová,M., Jurečka,P., Orozco,M., Cheatham,T.E., Šponer,J. and Otyepka,M. (2010) Performance of molecular mechanics force fields for RNA simulations: Stability of UUCG and GNRA hairpins. *J Chem Theory Comput*, **6**.
20. Zgarbová,M., Otyepka,M., Šponer,J., Mládek,A., Banáš,P., Cheatham,T.E. and Jurečka,P. (2011) Refinement of the Cornell et al. Nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J Chem Theory Comput*, **7**.
21. Gonzalez,O., Petkevičiute,D. and Maddocks,J.H. (2013) A sequence-dependent rigid-base model of DNA. *Journal of Chemical Physics*, **138**.
22. López-Güell,K., Battistini,F. and Orozco,M. (2023) Correlated motions in DNA: Beyond base-pair step models of DNA flexibility. *Nucleic Acids Res*, **51**.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

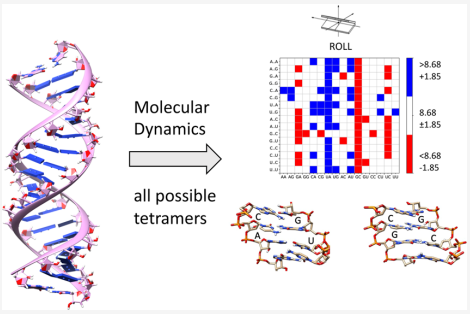
Sequence-Dependent Properties of the RNA Duplex

Federica Battistini, Alba Sala, Adam Hospital, and Modesto Orozco*

Cite This: J. Chem. Inf. Model. 2023, 63, 5259−5271 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Sequence-dependent properties of the DNA duplex have been accurately described using extensive molecular dynamics simulations. The RNA duplex meanwhile—which is typically represented as a sequence-averaged rigid rod—does not benefit from having equivalent molecular dynamics simulations. In this paper, we present a massive simulation effort using a set of ABC-optimized duplexes from which we derived tetramer-resolution properties of the RNA duplex and a simple mesoscopic model that can represent elastic properties of long RNA duplexes. Despite the extreme chemical similarity between DNA and RNA, the local and global elastic properties of the duplexes are very different. DNA duplexes show a complex and nonelastic pattern of flexibility, for instance, while RNA duplexes behave as an elastic system whose deformations can be represented by simple harmonic potentials. In RNA duplexes (RNA_2), not only are intra- and interbase pair parameters (equilibrium and mechanical) different from those in the equivalent DNA duplex sequences (DNA_2) but the correlations between movements also differ. Simple statements on the relative flexibility or stability of both polymers are meaningless and should be substituted by a more detailed description depending on the sequence and the type of deformation considered.



INTRODUCTION

Cellular RNA appears as a single strand, which may adopt a variety of secondary structures. The most abundant of these secondary structures is a right-handed double helix known as the A-form,[1,2] which constitutes almost 40% of naked ribonucleotides in the Protein Data Bank (PDB, data taken from NDB[3,4]). The similarity in chemical composition of DNA and RNA would suggest similar structural and physical properties for DNA duplex (DNA_2) and RNA duplex (RNA_2) helices, but in reality, they form duplexes that are notably different. The origin of the main structural difference in these duplexes is due to the presence of the 2′OH group in RNA, which biases the sugar puckering preferences toward the North conformation, leading the transition from the B- to A-form.[5] Inspection of PDB[6] and a myriad of biophysical data such as electron micrography, hydrodynamic experiments, gel electrophoresis, and NMR spectra supports the idea that the RNA_2 helix is more rigid than the DNA one.[7−12] Early molecular dynamics (MD) simulations throw the generality of this claim into question[13,14] and raise the need to perform detailed analysis on the response of RNA_2 to different deformations.

A series of contributions made over time by the ABC (Ascona B-DNA Consortium) uncovered the sequence-dependent variability of the physical properties of DNA_2.[15−22] The latest studies were performed using state-of-the-art force fields (FFs[23−25]) and simulation conditions.[20,25] These studies provide very accurate physical descriptions of the 136 unique tetramers,[17,19] and the ongoing work aims to extend such descriptions to the hexamer level.

In doing the following, ABC studies have changed our view on the structural and physical properties of DNA and have been instrumental in FF development. Thus, ABC simulations have highlighted the extreme sequence dependence of the duplex, which expand out of the traditional nearest neighbor level[17−20,26,27] and have revealed the presence of subtle correlations between base pair step (bps) helical degrees of freedom and backbone conformations[18,20,26] and the existence of bimodalities, which could not be explained by harmonic deformations.[15,17,19,20,28] Finally, ABC studies have provided publicly available data, which have been used to derive a variety of mesoscopic sequence-dependent descriptors of the DNA structure and flexibility,[19,21,28−32] which are useful to reproduce flexibility of long linear and circular DNAs,[19,21,29−33] nucleosome positioning,[34,35] chromatin structure,[36,37] the susceptibility of DNA to oncogenic mutations,[38] the impact of methylation on the DNA structure and function,[36,39] or the binding of DNA to proteins.[40−42]

Received: May 16, 2023
Published: August 14, 2023

Journal of Chemical Information and Modeling pubs.acs.org/jcim Article

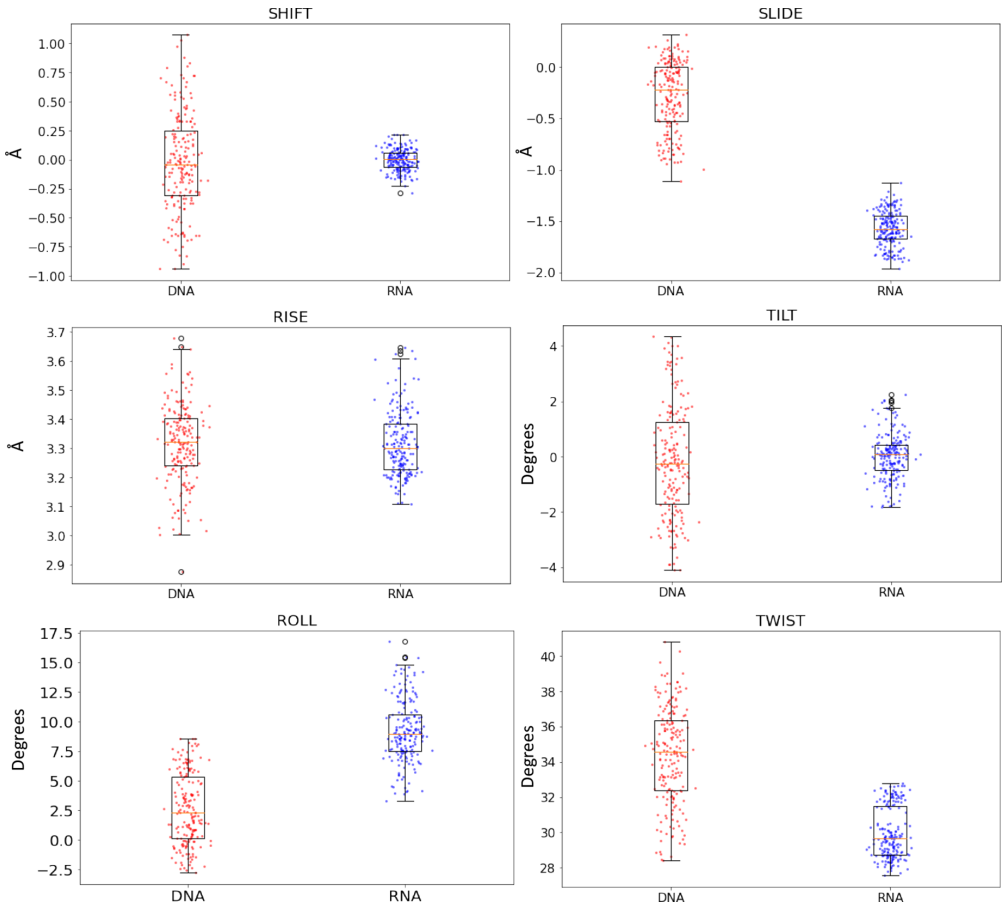


Figure 1. Box plots of the average helical parameters of the central bps of the136 unique tetramers. Values in red correspond to DNA and those in blue to RNA, in each case the average value is marked with a red line and outliers with black circles. Translational parameters (shift, slide, and rise) are in angstrom and rotational ones (tilt, roll, and twist) in degrees.[43]

Unfortunately, these impressive efforts for DNA have not been translated to the RNA world, where the duplex is considered simply as a sequence-independent rigid rod.

By using the state-of-the-art FF and simulation protocols, we aim here to expand our knowledge on the different modes and the sequence-dependent structural and mechanical properties of the RNA_2 deformability through a detailed analysis. To this end, we have simulated 13 duplexes containing the 136 unique tetramers,[20] simultaneously analyzing the global and local structure and dynamics. Analysis of the trajectories provides detailed information of the local and global sequence-dependent properties of the RNA_2 and allows us to develop a simple mesoscopic model, which reproduces the deformability of the RNA_2 surprisingly well.

RESULTS AND DISCUSSION

In the past decades, the ABC has studied the physical/conformational properties of B-DNA.[22] The latest studies[17,20] have made it clear that individual dinucleotide steps can have different conformational statistics depending on their tetranucleotide sequence contexts (5′ and 3′ flanking base pairs). This finding means that the conformational characteristics of the 10 distinct dinucleotide steps surrounded by all possible flanking base pairs—representing 136 distinct tetranucleotides—had to be studied. To do so, the Consortium designed an approach for studying these 136 distinct tetranucleotides with the least possible computational cost.[20] In this work, the team described the structural and dynamical properties of duplex B-DNA under physiological conditions and determined the sequence-dependent structural properties of DNA as expressed in the equilibrium distribution of its stochastic dynamics. The team detected polymorphisms in certain helical geometries for certain tetranucleotide sequence contexts.

Taking advantage of the detailed analyses done on B-DNA, we investigated the sequence-dependent properties of RNA duplexes using the same strategy as done by the ABC in the latest work about DNA duplexes. We decided to study the same set of sequences, as previously done for DNA (see Supporting Information Table S1). This set of sequences includes an optimized number of oligomers that enabled us to sample the conformational space of every possible tetramer and to obtain multimicrosecond trajectories in a practical way, given the not excessive length of the oligomers (18mers).

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

Journal of Chemical Information and Modeling pubs.acs.org/jcim Article

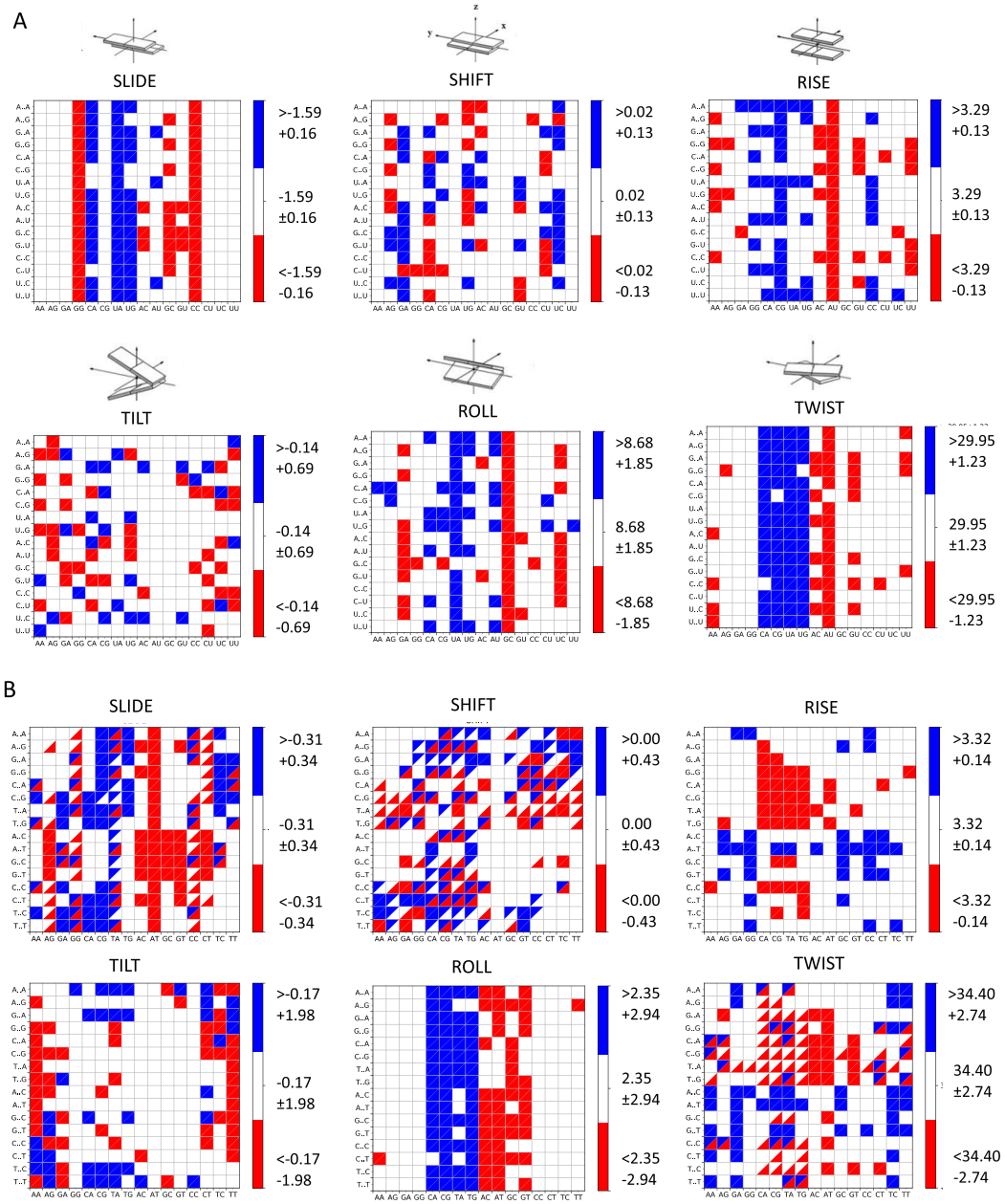


Figure 2. Average values of bps helical coordinates for the central bps (*x*-axis) in all possible tetranucleotide contexts (*y*-axis) for (A) RNA_2 and (B) DNA_2 (data from previous study[20]). Translational parameters (shift, slide, and rise) are in angstrom and rotational ones (tilt, roll, and twist) in degrees. The blue squares mean that a specific step has an average value above the global average plus one standard deviation, while the red squares mean an average value below the global average minus one standard deviation (see the legend at the right of each plot). When bimodality exists, the cell is divided into portions and the value refers to each one of the maxima as found by the BIC analysis.

In depth, this library is composed of 13 double helices each containing 18 bps (including GC terminals on each end), covering the complete tetranucleotide space (see Supporting Information Table S1 for a list of the designed sequences). The GC termini were chosen to reduce "fraying" and were not considered in the calculation of the base pair parameters, to avoid terminal effects.

Using the same set of sequences allowed us to make a direct comparison between RNA and DNA duplexes on their static and flexibility conformational properties. In this work for each sequence, we run a 1 μs simulation (see Methods), and for

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

Journal of Chemical Information and Modeling pubs.acs.org/jcim Article

Table 1. RNA Average Helical Parameter Values, with Standard Deviations, for dsRNA PDB Structures and for MD Simulation Ones, Averaged among All the Tetramers Studied

	shift (Å)	slide (Å)	rise (Å)	tilt (deg)	roll (deg)	twist (deg)
PDB	0.05 ± 0.42	−1.35 ± 0.47	3.21 ± 0.20	0.41 ± 2.37	9.75 ± 4.06	31.92 ± 3.24
MD	−0.01 ± 0.09	−1.58 ± 0.16	3.31 ± 0.11	−0.09 ± 0.71	8.95 ± 2.69	29.88 ± 1.40

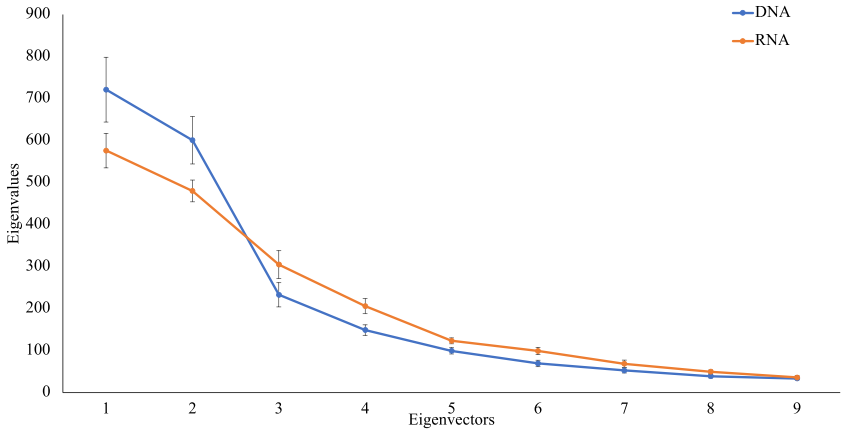


Figure 3. Variance associated with the different eigenvectors (values determined considering only the common backbone for all the duplexes, Å^2) for DNA (blue[20]) and RNA (orange) duplexes. The error bars correspond to the standard deviations associated with averaging values for 13 duplexes.

each tetramer, we analyzed the properties of the central base pair and the relative effect of the neighboring bases.

Convergence and the Overall Structure. MD simulations provide stable and converged trajectories in the 1 μs regime, with small RMSd with respect to the averaged structure (around 1.7 ± 0.4 Å) and the canonical A-form (RMSd around 2 ± 0.5 Å) (see Supporting Information Table S2 and Supporting Information Figures S1 and S2). No drifts in RMSd are observed and nearly identical RMSds are reported at the initial and final part of the trajectories, suggesting that the trajectories are well converged and stable. This is also confirmed by looking at the root mean square inner product (RMSIP, see Supporting Information Table S2) between the essential deformation modes for each sequence obtained from the first (100−500 ns) and last part (600−1000 ns) of the trajectories with values between 0.97 and 1.0. This means that essential deformation modes are identical in two separated parts of the trajectory. Furthermore, the profiles of variance vs the essential deformation modes are identical when computed from first and second halves of the trajectories (see some examples in Supporting Information Figure S3).

Helical Parameter Distributions. Average helical properties collected for all tetramers confirm the known differences between the local structure of DNA and RNA duplexes: lower twist, higher roll, and more negative slide of RNA_2 compared to DNA_2 (see Figure 1, for DNA data taken from ref 20). It is noteworthy that—as previously reported from database analysis—the tetramer-dependent variability of shift, slide, tilt, and twist found in DNA_2 is reduced for RNA_2. Only rise and roll show sizeable sequence-dependent variability in RNA duplexes, which in any case is lower than that found for DNA.

Interestingly, the helical degrees of freedom generating more sequence-dependent polymorphic behaviors in DNA_2 (twist and shift, which have a marked bimodality) show very little dispersion in RNA_2 (Figure 2). Detailed analysis of the individual tetramers shows that sequence variability in the RNA_2 (which is small in any case) is related to the central bps (see Figure 2, *x*-axis), with little impact of the neighboring bps in the tetramer (see Figure 2, neighboring bases in each tetramer *y*-axis). In the detailed analyses of the effect of the sequence context on the average parameters, we saw that tetramers containing the UpA step at the central position always show a higher roll (as well as slide and twist) than the average, while tetramers containing the GpC step at the central position display the opposite trend, irrespective of the neighboring base pairs.

We detected that tetramers with central base pairs CpA, CpG, UpA, and UpG, independently of the tetrameric context, are characterized by a high twist, while tetramers with central CpC/GpG show a low slide. Slide and twist appeared as the least context-dependent movements, but we could detect common patterns given only by the central base pair also for rise and roll, while in DNA, these patterns are present only for the roll parameter (see Supporting Information Figure S4).

The Bayesian Information Criterion (BIC)-Helguero analysis (see Methods) failed to detect cases of multimodality in the 136 tetramers of the RNA_2 studied here (see Figure 2). The differences with respect to DNA_2 in terms of central bps induced variability, nearest neighbor effects, and complexity of the distributions that were dramatic (see Figure 2 and specific examples can be seen in Supporting Information Figure S5). As expected from the small tetramer-dependent polymorphism, hexamer effects are negligible (see some examples in Supporting Information Figure S6), indicating a strong locality in the definition of equilibrium helical properties in the RNA_2.

Finally, we compared simulation averages to experimental values obtained by mining the Protein Data Bank (PDB) for naked RNA_2. The agreement is remarkable (see Table 1),

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

suggesting that despite potential caveats, the currently used RNA force field accurately represents RNA conformational preferences.

Global Flexibility. The global RNA_2 deformability is mainly defined by global twisting and bending deformations, mimicking the movement found for DNA_2 [see the first component videos in the principal component analysis (PCA) of the trajectories stored in BIGNASim]. The differences are significant, however, when looking at the distribution of variance among the different essential deformation modes. While in the case of DNA_2, a large part of the variance is distributed along the first two modes, a much smoother decay in the importance of the variance along modes is found in the case of RNA_2 (see Figure 3). This confirms previous suggestions[6,13,14] that the dynamics of the elongated DNA and RNA duplexes are different. In this instance, DNA shows a simpler dynamic than the more globular RNA_2, which involves many independent movements that contribute to the definition of the global dynamics.

Local Flexibility. We explored two types of local deformations of the RNA_2: those disrupting the hydrogen bond pattern and base pair distortions within each base pair step. As shown in Table 2, the H-bond scheme is well

Table 2. Loss of Hydrogen Bonds for A-T(U) and G-C Pairs in DNA (in Italic[20]) and RNA (Bold) Duplexes[a]

pair	loss of one HB (average % time)	loss of two HB (average % time)	loss of three HB (average % time)	solvent exchange (average % time)
G·C bp terminal	*3.73*	*2.55*	*1.73*	*2.14*
	1.48	**0.59**	**0.29**	**0.29**
G·C bp terminal (−1)	*0.33*	*0.01*	*<0.01*	*<0.01*
	0.92	**0.41**	**0.03**	**0.05**
G·C bp central	*0.45*	*0.03*	*0.01*	*0.01*
	0.83	**0.18**	**0.01**	**0.01**
A·T (bp central)	*1.67*	*0.06*		*0.03*
A·U (bp central)	**4.12**	**0.42**		**0.07**

[a]See Supporting Information Methods for a detailed description on h-bond determination.

preserved for both duplexes. For DNA_2, the terminal bases (G·C in our sequence set, Supporting Information Table S1) are those more frequently unpaired and those where the open state has a residence time large enough to allow proton interchange with solvent. Such terminal fraying is dramatically reduced for RNA_2, where states that can justify proton exchange of the terminal bases with the solvent are scarce. As described previously,[18,23,25] openings in the middle of the DNA double helix are rare, short lived, and partial (see Table 2; see examples in Supporting Information Figure S7), while these distortions are not so infrequent in the context of RNA_2, always related to fast A·U breathing events. This might appear unexpected considering the general, larger stability of RNA duplexes but agrees perfectly with the greater stability of d(A·T) compared to r(A·U) pairs.[44,45] In fact, relative DNA_2 vs RNA_2 melting temperatures for the oligos studied here show a general reduction in the stability gap between the two duplexes when the number of A·T(U) pairs increase (see Supporting Information Figure S8), in perfect alignment with our calculations.

As described above, independently from the sequence context, the RNA_2 shows Gaussian distributions for all helical parameters. This suggests that local deformations in the helix can be accurately captured by a simple harmonic model using a 6×6 stiffness matrix, where the diagonal terms represent the stiffness associated with pure deformations (slide, shift, rise, tilt, roll, and twist) and the out-off diagonal represent the coupling terms. Interestingly, sequence variability—which has little impact in defining equilibrium helical coordinates (Figure 1)—has a non-negligible effect in defining helical stiffness (Figure 4). In fact, dispersion of stiffness values along sequences in RNA_2 is similar or even greater than that found for DNA_2. The only exception being twist—which due to the absence of sequence-dependent bimodality—shows a much more reduced dependence with the sequence in RNA_2 (Figure 4).

Note that the lack of variability in helical equilibrium values, compared to the variability in the corresponding stiffness matrices, suggests that the analysis of the dispersion of helical coordinates in the experimental structures of different sequences, which was a reasonable approach to obtain stiffness matrices in DNA duplexes,[46] is not acceptable for RNA_2.

Finally, looking at the average values in the box plots in Figure 4, we can rule out the common assumption that RNA_2 is stiffer than DNA_2 also at the microscopic level. In fact, only two of the helical deformations (slide and twist) are systematically stiffer in RNA compared to DNA duplexes, and some helical deformations, like those affecting the tilt, are easier for RNA_2. Therefore, as already suggested from the analysis of global deformability above, no general claims on the relative stiffness of DNA_2 and RNA_2 should be made, as it depends on the sequence and the type of deformation considered.

Most of the sequence-dependent variability of helical stiffness of the RNA_2 is related to the central bps (see Figure 5, *x*-axis), showing a reduced dependence from the first and fourth bps of the tetramer (see Figure 5, neighboring bases in each tetramer *y*-axis). There are some common trends between DNA_2 and RNA_2: for example, analysis on the sequence context effect suggests than in general, Pyr-Pur steps are stiffer and Pur-Pyr steps are softer than the average (Figure 5). However, the rest of the sequence dependence found for DNA_2 cannot be translated to RNA_2. For example, the "flexible" r(UA) step is stiffer than the average in terms of shift and slide deformations (Figure 5).

Interestingly, roll and rise flexibilities in RNA_2 seem to go together, as partially seen also for DNA_2. Tetramers with central base pairs CA, CG, UA, and UG are not only characterized by high twist but also seem to be the most flexible base pairs independently on the sequence context. Tetramers with central CC/GG showed a low slide and are the stiffest for slide−shift−tilt movements. Differently from DNA_2, in RNA_2, the tetramers with central GC and CC steps are stiff and tetramer-independent. Clearly, the code linking sequence with deformability is quite different in DNA and RNA duplexes.

Correlations. Stiffness matrices are nondiagonal for both DNA_2 and RNA_2, indicating coupling between pure helical deformations. These couplings are the main reason for the well-characterized connection between entire deformation modes of the helices (ex. twist, stretch[48]). Understanding the individual correlation between the different helical deformations would require discussion of 136 stiffness matrices (for

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

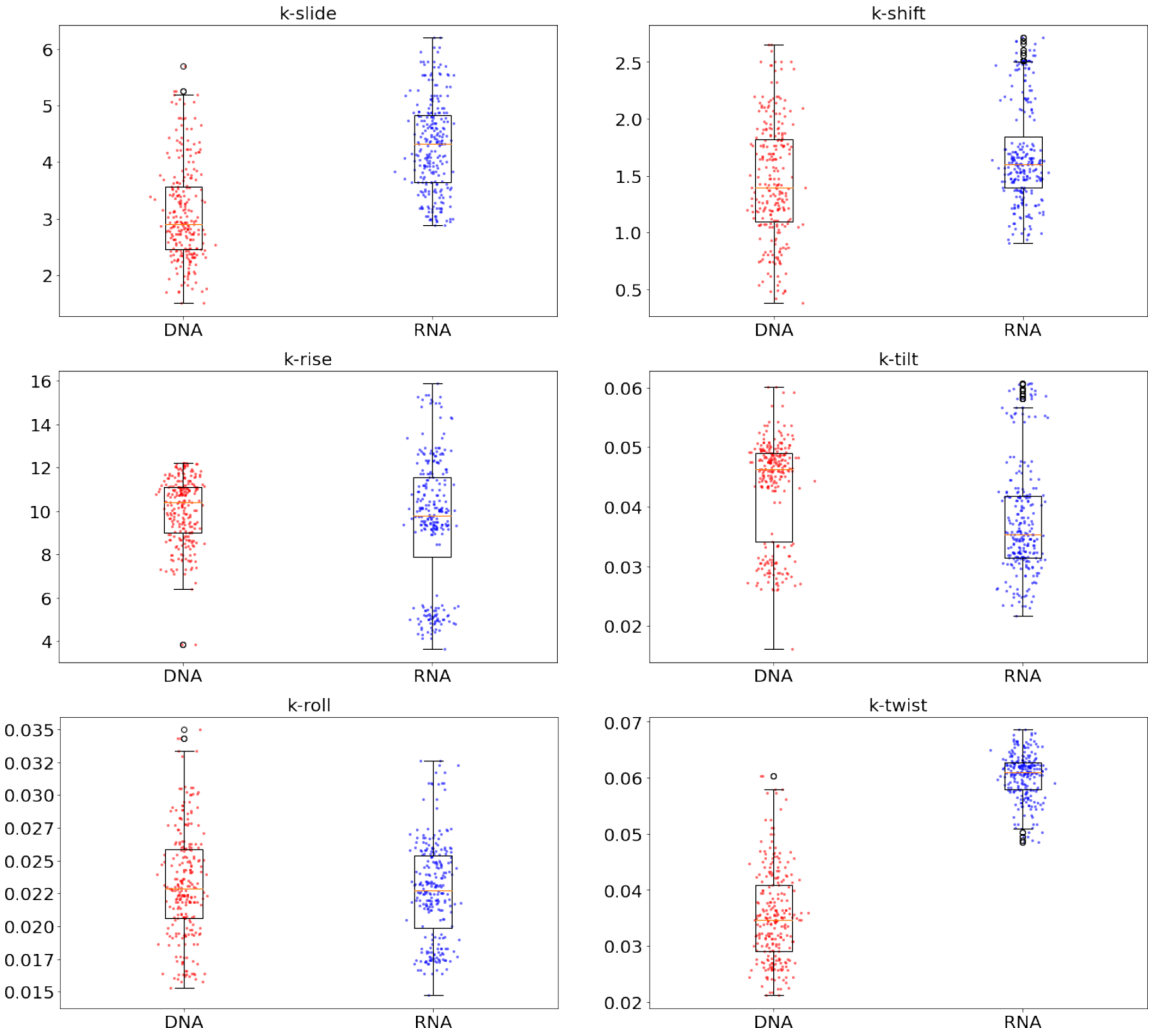


Figure 4. Box plots of diagonal stiffness matrices for the central bps of the 136 unique tetramers in DNA (blue, data taken from ref 20) and RNA (red). Translational constants (*k*-slide, *k*-shift, and *k*-rise) are in kcal/molÅ^2 and rotational constants (k-tilt, k-roll, and k-twist) are in kcal/mol·deg^2.

DNA_2 and RNA_2), but a general picture can be obtained by averaging them (after normalization using Lankaš' transformation[47]). Results in Table 3 show the existence of strong positive couplings between the slide and rise for both DNA and RNA duplexes. Furthermore, rise is negatively coupled with all the torsional deformations for RNA_2, but the strong rise−roll negative coupling found in the RNA_2 is lost in DNA_2. The slide−roll negative coupling is more intense in RNA_2 than in DNA_2, and the opposite happens for the twist−roll positive coupling that is more intense for DNA_2. In summary, the pattern of correlation between helical deformations in the DNA and RNA helix is quite different, but both polymers show a significant degree of interconnection between the different helical movements at the bps level.

While the intra-bps correlations have been described since the original studies of Zhurkin and co-workers,[46] the correlation between the movement of the neighboring bps has been ignored for decades due to an implicit assumption that global deformation is a consequence of independent bps movements. Different studies with the DNA_2 have demonstrated that contrary to this assumption, the deformations at one given bps are correlated with those in the neighboring ones (see above and detailed discussion in refs 29−31, 49). These connections are often negative (anticorrelation), typically when the same helical parameter is considered in two neighboring bps and might be slightly positive in a few cases when different helical deformations are explored in the neighboring bps (see Figure 6). For DNA_2, the strongest anticorrelations are found for shift/shift, twist/twist, and tilt/tilt, while positive correlations are found for shift/twist, slide/shift, and slide/tilt (see Figure 6), with significant variability depending on the tetramer. For RNA_2, the strength and complexity of the couplings between bps deformations are largely reduced compared to those of the DNA_2 (Figure 6),

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

Journal of Chemical Information and Modeling pubs.acs.org/jcim Article

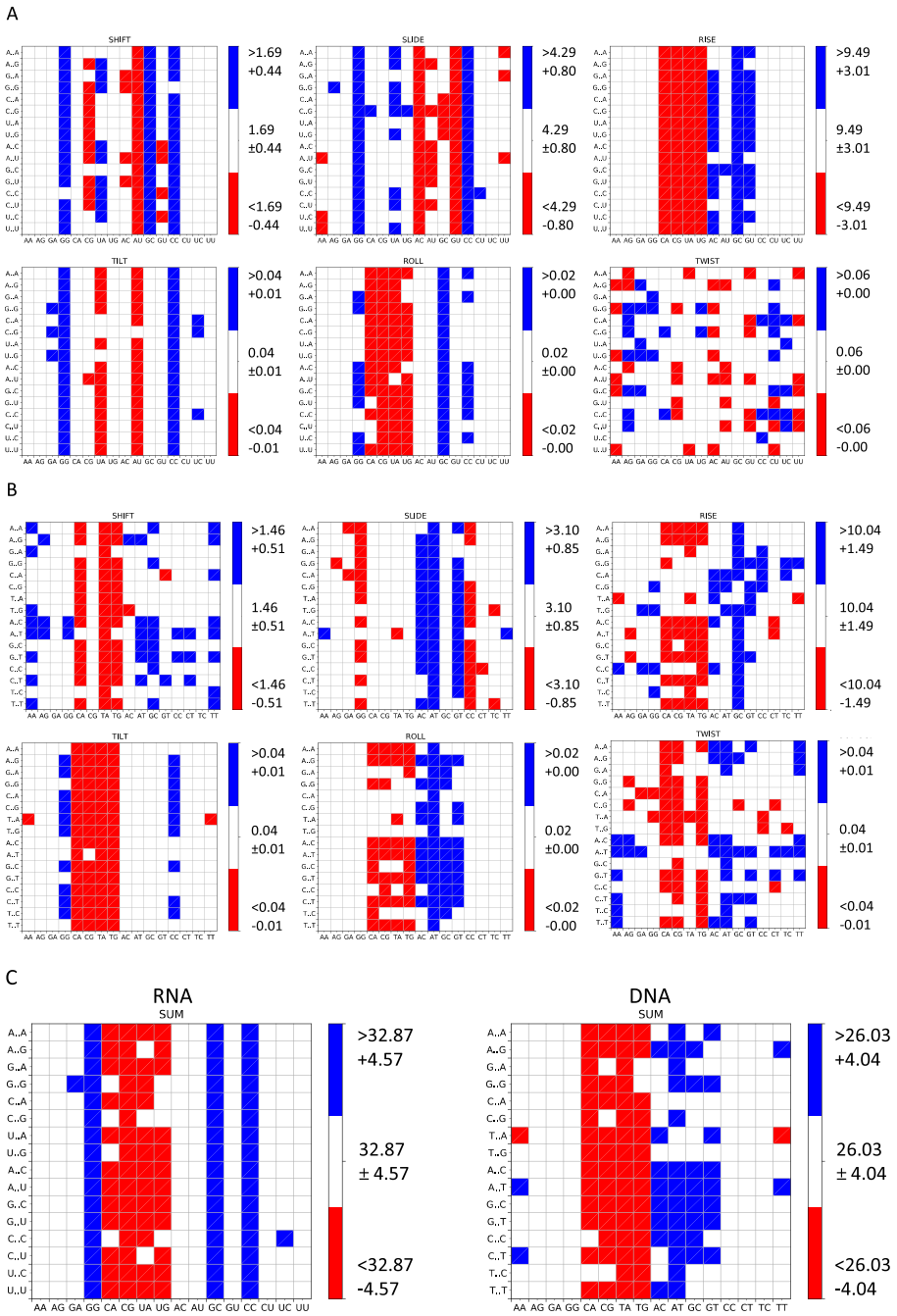


Figure 5. Average values of bps helical stiffness for the central bps (*x*-axis) in all possible tetranucleotide contexts (*y*-axis). The blue squares mean that a specific step has an average value above the global average plus one standard deviation, while the red squares mean an average value below the global average minus one standard deviation (see the legend at the right of each plot). For comparison purposes, we considered here all distributions represented by a single Gaussian. (A) RNA_2 and (B) DNA_2.[20] Translational constants (*k*-slide, *k*-shift, and *k*-rise) are in kcal/molÅ^2 and rotational constants (*k*-tilt, *k*-roll, and *k*-twist) are in kcal/mol·deg^2. (C) Overall sums of the stiffness constants for RNA (left) and DNA(right[20]) computed by applying Lankaš' transformation to the original stiffness matrix.[47]

but negative correlations are found for shift/shift, tilt/tilt, and twist/twist, as already described for DNA_2, even at a reduced magnitude compared with DNA_2. On the other hand, positive correlations are mild in all cases except rise/tilt, which can be

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

Journal of Chemical Information and Modeling pubs.acs.org/jcim Article

Table 3. RNA (in Bold) and DNA (in Italic) Stiffness Constants Normalized,[47] Averaged among All the Tetramers Studied

DNA/RNA	shift	slide	rise	roll	twist	tilt
shift	*1.452*/**1.691**	**−0.090**	**0.085**	**0.022**	**0.093**	**−0.288**
slide	*0.064*	*3.120*/**4.281**	**1.497**	**−0.432**	**−1.919**	**0.112**
rise	*0.162*	*1.725*	*10.028*/**9.441**	**−1.781**	**−1.953**	**−0.808**
roll	*0.099*	*−0.141*	*−0.197*	*2.607*/**2.555**	**0.447**	**−0.006**
twist	*0.099*	*−1.292*	*−2.167*	*0.985*	*3.970*/**6.789**	**−0.019**
tilt	*−0.802*	*−0.007*	*−0.767*	*−0.021*	*0.102*	*4.824*/**4.211**

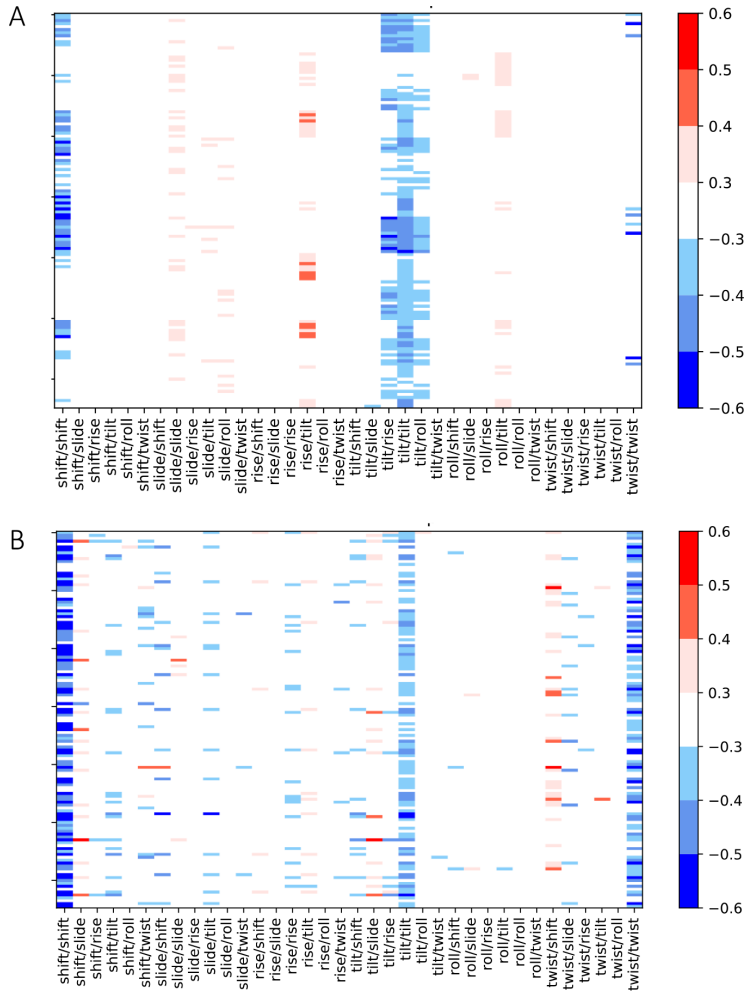


Figure 6. Heat maps showing correlations between neighboring base pairs, for each possible bps parameter[43] combination for the 136 unique tetramers (*y*-axis) for (A) RNA and (B) DNA, respectively.

sizeable for some tetramers. Finally, and quite unique of RNA_2, is the presence of anticorrelation between tilt/rise and tilt/roll movements in neighboring steps. Altogether, these suggest that the short-range flexibility of DNA and RNA duplexes and accordingly their ability to react against perturbation and transfer their information is quite different.

Correlated Harmonic Mesoscopic Model for RNA_2. We compared the distributions for helical parameters obtained using a mesoscopic-correlated harmonic model, inspired by that developed by Maddocks and co-workers for DNA_2[49,50] (CHM; see Methods for details) and atomistic MD simulations for a 36 bps long random sequence.

Thanks to this method, we could reproduce both the intra- and interbase pair ensembles (see their densities for some tetramers along the sequence in Figure 7 and in Supporting Information Figure S9).

As a last check, we compared the global helical properties of the long duplex RNA_2 simulated by our CHM with respect to the atomistic MD simulations. We calculated the average twist and roll, parameters that vary the most in long polymers, for

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

Journal of Chemical Information and Modeling pubs.acs.org/jcim Article

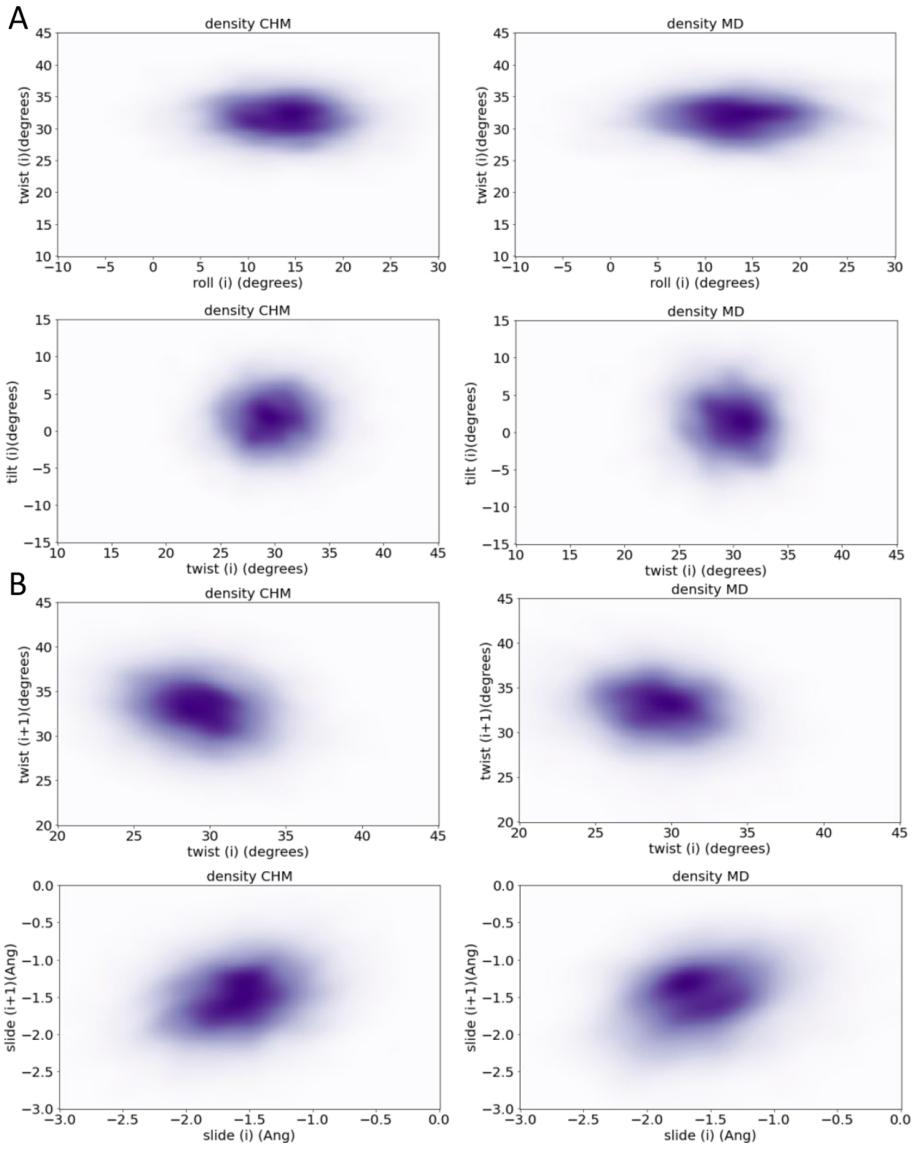


Figure 7. (A) Selected examples of the density of intrabase pairs ensembles (in helical space) using the correlated harmonic model (left panels, CHM) and atomistic MD simulations (the reference, right panels). (B) Selected examples of the inter-bps parameters for consecutive steps, i against the neighboring step $i + 1$, using the correlated harmonic model (left panels) and atomistic MD simulations (the reference, right panels).

the central 34mer for both CHM and MD structures, and RMSd, finding an excellent agreement (Figure 8).

In summary, we present here, for the first time, a comprehensive study of the sequence-dependent structural and mechanical properties of the RNA_2. The pattern of flexibility of RNA_2 is simpler than that of the DNA_2, showing a more harmonic behavior and a reduced sequence-dependent variability in terms of equilibrium properties. Claims that RNA_2 is "stiffer" than DNA_2 are incorrect, as it depends on the type of mechanical deformation that is introduced. The patterns of global deformability, signal transduction, and sequence-dependent deformation rules are different for DNA_2 and RNA_2. Finally, we were able to develop a very simple mesoscopic-correlated harmonic model that seems to have a good ability to reproduce RNA conformational ensembles, opening the possibility to perform mesoscopic simulations of RNA duplexes similar to those done for decades for the DNA ones.

METHODS

The duplexes of the miniABC sequence library[20] (translated to RNA, T → U) were built in the A-conformation using Arnott's fiber conformation.[1] Each duplex was placed in a solvated-truncated octahedral box of SPC/E water molecules,[51] which extends for more than 10 Å from the nearest atom of the RNA.

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

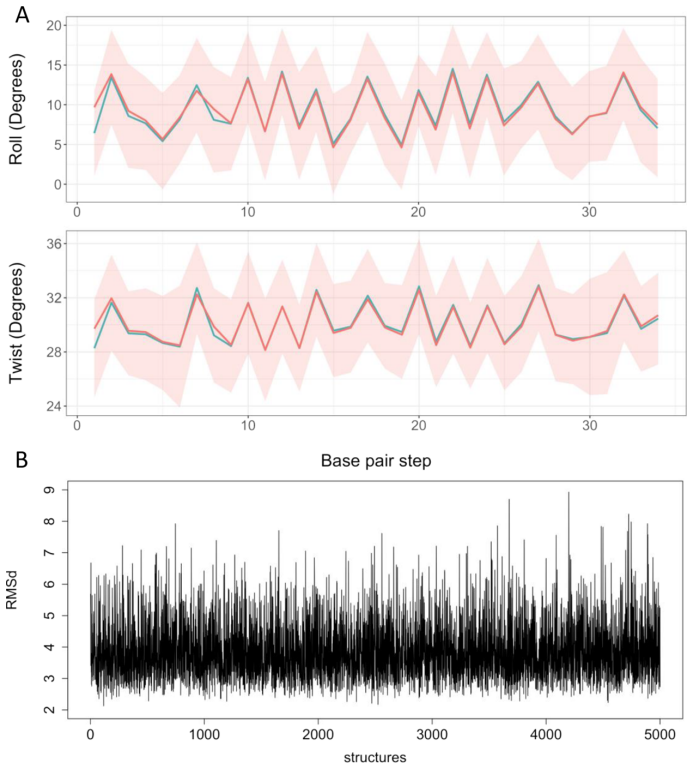


Figure 8. (A) Average roll (top) and twist (bottom) parameters along the MD simulation (blue) compared to the average among the structures of the CHM with the relative standard deviation (red line and shadow). (B) RMSd values (in Å) of the CHM structures using as a reference the MD average structure. The average RMSd along the simulation taking as a reference the MD average structure is 2.9 ± 0.9 Å.

Systems were neutralized by K^+, adding then 150 mM KCl using Dang's parameters.[52] χOL3 FF was used to represent RNA.[53,54] All systems were optimized, thermalized, and equilibrated using the standard ABC-simulation protocol.[20] Production simulations extend for 1 μs using the state-of-the-art simulation conditions as described elsewhere.[20] Data were collected every picosecond and, if not otherwise stated, analyses shown correspond to the last 0.5 μs. Trajectories were processed using the CPPTRAJ[55] module of the AmberTools 18 package (https://ambermd.org/AmberTools.php).

For each duplex, physical and geometrical descriptors derived from MD simulations were calculated to study the RNA (and DNA) average conformation and deformability at the base pair step level. Instead of using six Cartesian coordinates, the geometry of two consecutive base pairs, a DNA base pair step, is described with a set of six helical movement parameters (base pair step helical parameters): three translations (rise, slide, and shift) and three rotations (twist, roll, and tilt). The deformability along these movements is described by the stiffness constants (k_i) associated with the displacements with respect to the equilibrium values of the helical parameters[56] and determined by inversion of the covariance matrix computed in the helical space.[47] The helical parameters were calculated using the program Curves+ and Canal.[43]

The BIC was used to quantify the normal or binormal (i.e., a mixture of two normal functions) nature of the distributions of the helical parameters. Helguero's theorem was used to determine the normality and modality of the different distributions.[57,58] PCA was carried out using a python suite and RMSIP was calculated using the Bio3D package in R.[55,59] Melting temperatures were calculated using Biopython.[60]

All new RNA trajectories and DNA trajectories used for comparison are accessible and can be retrieved from the BIGNASim database (https://mmb.irbbarcelona.org/BIGNASim/);[61,62] data for DNA were taken from ref 20. The same protocol was followed for the 36 base pair long sequence (ACUAGAUCGAUGUACGCUAGCGUACAUCGAUCUAGU) used in the comparison between MD and the mesoscopic-correlated harmonic model.

We recently developed a general model to capture flexibility including nonharmonic and correlated Hamiltonians,[49] which simplifies to a correlated harmonic model when deviation from normality of the different helical distributions is not present. This simplified model is very close to that developed by Maddocks and co-workers for DNA_2,[49,50] which include nonlocality in the elastic response of DNA. In detail, the deformation energy considers the couplings between the nearest neighboring bps using a banded rather than a block stiffness matrix (K, see Figure S10):

$$E(X) = \frac{1}{2}Y^T KY = \frac{1}{2}K\Delta Y^2$$

where the sum extends to all the length of the DNA composed of N bps, and K and ΔY_j are the banded stiffness matrix (see

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

Supporting Information Figure S10) and the deformation vector, respectively. Maddocks and co-workers presented a simple procedure for inversion of the banded covariance from which banded stiffness matrices can be derived and built a coarse-grained model where they treated the bases separately, not as rigid units within the bps, leading to a complex stiffness matrix accounting for inter−inter, intra−inter, and intra−intra base pair contributions. Contrary to Maddocks's model, which considers base deformations, our simplified CHM method considers the base pair step as the minimum conformational unit.

Base pair parameters and stiffness and the mesoscopic descriptors calculated are available at https://github.com/Jalbiti/NaStruc_db.

■ ASSOCIATED CONTENT

Data Availability Statement

The data set is available in the repository https://github.com/Jalbiti/NaStruc_db.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00741.

> Supplementary methods for hydrogen bond calculation details; the RNA sequences studied with the relative links to visualize the structures and the values of RMSD and RMSIP along the MD, respectively; images of the structures studied through MD simulations; additional data regarding, respectively, PCA analyses, base pair parameter distributions, hydrogen bond detection, melting temperature data, and intrabase and interbase pair ensembles; and the scheme to calculate the stiffness matrix (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Modesto Orozco − *Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona 08028, Spain; Departament de Bioquímica i Biomedicina. Facultat de Biologia, Universitat de Barcelona, Barcelona 08028, Spain;* orcid.org/0000-0002-8608-3278; Email: modesto.orozco@irbbarcelona.org

Authors

Federica Battistini − *Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona 08028, Spain; Departament de Bioquímica i Biomedicina. Facultat de Biologia, Universitat de Barcelona, Barcelona 08028, Spain;* orcid.org/0000-0002-7544-0938

Alba Sala − *Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona 08028, Spain;* orcid.org/0000-0003-1046-7432

Adam Hospital − *Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona 08028, Spain;* orcid.org/0000-0002-8291-8071

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00741

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are indebted to Lucia Fabio for help with analysis and programming for the multimodality study and to Dr. Juan Pablo Arcón for help in hydrogen bond detection. This work was supported by the Center of Excellence for HPC H2020 European Commission; "BioExcel-2− Centre of Excellence for Computational Biomolecular Research" [823830]; BioExcel-3: Centre of Excellence for Computational Biomolecular Research [European Union: 101093290; Ministerio de Ciencia e Innovación: PCI2022-134976-2]; Spanish Ministry of Science [RTI2018-096704-B-100 and PID2021-122478NB-I00]; Instituto de Salud Carlos III−Instituto Nacional de Bioinformatica, Fondo Europeo de Desarrollo Regional [ISCIII PT 17/0009/0007]; European Regional Development Fund, ERFD Operative Programme for Catalunya, the Catalan Government AGAUR [SGR2021 00863]. European Union "MDDB: MOLECULAR DYNAMICS DATA BANK. THE EUROPEAN REPOSITORY FOR BIOSIMULATION DATA." [101094651]. The IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from the MINECO. M.O. is an ICREA Academy scholar.

■ REFERENCES

(1) Arnott, S.; Hukins, D. W. L.; Dover, S. D. Optimised Parameters for RNA Double-Helices. *Biochem. Biophys. Res. Commun.* **1972**, *48*, 1392−1399.

(2) Grille, L.; Gallego, D.; Darré, L.; Da Rosa, G.; Battistini, F.; Orozco, M.; Dans, P. D.; Pablo, P.; Dans, D. The Pseudo-Torsional Space of RNA. *bioRxiv* **2022**, No. 2022.06.24.497007.

(3) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(4) Berman, H. M.; Lawson, C. L.; Schneider, B. Developing Community Resources for Nucleic Acid Structures. *Life* **2022**, *12*, 540.

(5) Soliva, R.; Luque, F. J.; Alhambra, C.; Orozco, M. Role of Sugar Re-Puckering in the Transition of A and B Forms of DNA in Solution. A Molecular Dynamics Study. *J. Biomol. Struct. Dyn.* **1999**, *17*, 89−99.

(6) Pérez, A.; Noy, A.; Lankas, F.; Luque, F. J.; Orozco, M. The Relative Flexibility of B-DNA and A-RNA Duplexes: Database Analysis. *Nucleic Acids Res.* **2004**, *32*, 6144−6151.

(7) Hagerman, P. J. FLEXIBILITY OF RNA. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *26*, 139−156.

(8) Hagerman, P. J. Flexibility of DNA. *Annu. Rev. Biophys. Biophys. Chem.* **2003**, *17*, 265−286.

(9) Thomas, G. J.; Benevides, J. M.; Overman, S. A.; Ueda, T.; Ushizawa, K.; Saitoh, M.; Tsuboi, M. Polarized Raman Spectra of Oriented Fibers of A DNA and B DNA: Anisotropic and Isotropic Local Raman Tensors of Base and Backbone Vibrations. *Biophys. J.* **1995**, *68*, 1073−1088.

(10) Matsumoto, U.; Fujiwara, T.; Akutsu, H.; Kyogoku, Y.; Shindo, H. Phosphorus-31 Nuclear Magnetic Resonance of Highly Oriented DNA Fibers. 1. Static Geometry of DNA Double Helices. *Biochemistry* **1985**, *24*, 887−895.

(11) Fujiwara, T.; Shindo, H. Phosphorus-31 Nuclear Magnetic Resonance of Highly Oriented DNA Fibers. 2. Molecular Motions in Hydrated DNA. *Biochemistry* **1985**, *24*, 896−902.

(12) Cheatham, T. E.; Kollman, P. A. Molecular Dynamics Simulations Highlight the Structural Differences among DNA:DNA, RNA:RNA, and DNA:RNA Hybrid Duplexes. *J. Am. Chem. Soc.* **1997**, *119*, 4805−4825.

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

(13) Noy, A.; Pérez, A.; Lankas, F.; Javier Luque, F.; Orozco, M. Relative Flexibility of DNA and RNA: A Molecular Dynamics Study. *J. Mol. Biol.* **2004**, *343*, 627−638.
(14) Faustino, I.; Pérez, A.; Orozco, M. Toward a Consensus View of Duplex RNA Flexibility. *Biophys. J.* **2010**, *99*, 1876−1885.
(15) Beveridge, D. L.; Barreiro, G.; Suzie Byun, K.; Case, D. A.; Cheatham, T. E.; Dixit, S. B.; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Seibert, E.; Sklenar, H.; Stoll, G.; Thayer, K. M.; Varnai, P.; Young, M. A. Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. I. Research Design and Results on d(CpG) Steps. *Biophys. J.* **2004**, *87*, 3799−3813.
(16) Dixit, S. B.; Beveridge, D. L.; Case, D. A.; Cheatham, T. E.; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Sklenar, H.; Thayer, K. M.; Varnai, P. Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. II: Sequence Context Effects on the Dynamical Structures of the 10 Unique Dinucleotide Steps. *Biophys. J.* **2005**, *89*, 3721.
(17) Pasi, M.; Maddocks, J. H.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T.; Dans, P. D.; Jayaram, B.; Lankas, F.; Laughton, C.; Mitchell, J.; Osman, R.; Orozco, M.; Pérez, A.; Petkevičiute, D.; Spackova, N.; Sponer, J.; Zakrzewska, K.; Lavery, R. MABC: A Systematic Microsecond Molecular Dynamics Study of Tetranucleotide Sequence Effects in B-DNA. *Nucleic Acids Res.* **2014**, *42*, 12272−12283.
(18) Dans, P. D.; Pérez, A.; Faustino, I.; Lavery, R.; Orozco, M. Exploring Polymorphisms in B-DNA Helical Conformations. *Nucleic Acids Res.* **2012**, *40*, 10668−10678.
(19) Walther, J.; Dans, P. D.; Balaceanu, A.; Hospital, A.; Bayarri, G.; Orozco, M. A Multi-Modal Coarse Grained Model of DNA Flexibility Mappable to the Atomistic Level. *Nucleic Acids Res.* **2020**, *48*, No. e29.
(20) Dans, P. D.; Balaceanu, A.; Pasi, M.; Patelli, A. S.; Petkevičiūtė, D.; Walther, J.; Hospital, A.; Bayarri, G.; Lavery, R.; Maddocks, J. H.; Orozco, M. The Static and Dynamic Structural Heterogeneities of B-DNA: Extending Calladine−Dickerson Rules. *Nucleic Acids Res.* **2019**, *47*, 11090.
(21) Lavery, R.; Zakrzewska, K.; Beveridge, D. L.; Bishop, T. C.; Case, D. A.; Cheatham, T.; Dixit, S.; Jayaram, B.; Lankas, F.; Laughton, C.; Maddocks, J. H.; Michon, A.; Osman, R.; Orozco, M.; Perez, A.; Singh, T.; Spackova, N.; Sponer, J. A Systematic Molecular Dynamics Study of Nearest-Neighbor Effects on Base Pair and Base Pair Step Conformations and Fluctuations in B-DNA. *Nucleic Acids Res.* **2010**, *38*, 299−313.
(22) da Rosa, G.; Grille, L.; Calzada, V.; Ahmad, K.; Arcon, J. P.; Battistini, F.; Bayarri, G.; Bishop, T.; Carloni, P.; Cheatham, T.; Collepardo-Guevara, R.; Czub, J.; Espinosa, J. R.; Galindo-Murillo, R.; Harris, S. A.; Hospital, A.; Laughton, C.; Maddocks, J. H.; Noy, A.; Orozco, M.; Pasi, M.; Pérez, A.; Petkevičiūtė-Gerlach, D.; Sharma, R.; Sun, R.; Dans, P. D. Sequence-Dependent Structural Properties of B-DNA: What Have We Learned in 40 Years? *Biophys. Rev.* **2021**, *13*, 995−1005.
(23) Dans, P. D.; Danilāne, L.; Ivani, I.; Dršata, T.; Lankaš, F.; Hospital, A.; Walther, J.; Pujagut, R. I.; Battistini, F.; Gelpí, J. L.; Lavery, R.; Orozco, M. Long-Timescale Dynamics of the Drew−Dickerson Dodecamer. *Nucleic Acids Res.* **2016**, *44*, 4052−4066.
(24) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; Portella, G.; Battistini, F.; Gelpí, J. L.; González, C.; Vendruscolo, M.; Laughton, C. A.; Harris, S. A.; Case, D. A.; Orozco, M. Parmbsc1: A Refined Force Field for DNA Simulations. *Nat. Methods* **2016**, *13*, 55−58.
(25) Dans, P. D.; Ivani, I.; Hospital, A.; Portella, G.; González, C.; Orozco, M. How Accurate Are Accurate Force-Fields for B-DNA? *Nucleic Acids Res.* **2017**, *45*, No. gkw1355.
(26) Balaceanu, A.; Pasi, M.; Dans, P. D.; Hospital, A.; Lavery, R.; Orozco, M. The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J. Phys. Chem. Lett.* **2017**, *8*, 21−28.
(27) Balaceanu, A.; Buitrago, D.; Walther, J.; Hospital, A.; Dans, P. D.; Orozco, M. Modulation of the Helical Properties of DNA: Next-to-Nearest Neighbour Effects and Beyond. *Nucleic Acids Res.* **2019**, *47*, 4418−4430.
(28) Dans, P. D.; Walther, J.; Gómez, H.; Orozco, M. Multiscale Simulation of DNA. *Curr. Opin. Struct. Biol.* **2016**, *37*, 29−45.
(29) Liebl, K.; Zacharias, M. Accurate Modeling of DNA Conformational Flexibility by a Multivariate Ising Model. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, No. e2021263118.
(30) Lanka, F.; Gonzalez, O.; Heffler, L. M.; Stoll, G.; Moakher, M.; Maddocks, J. H. On the Parameterization of Rigid Base and Basepair Models of DNA from Molecular Dynamics Simulations. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10565−10588.
(31) Sharma, R.; Patelli, A. S.; Bruin, L.; De Maddocks, J. H. CgNA +web: A Visual Interface to the CgNA+ Sequence-Dependent Statistical Mechanics Model of Double-Stranded Nucleic Acids. *J. Mol. Biol.* **2023**, *435*, No. 167978.
(32) De Bruin, L.; Maddocks, J. H. CgDNAweb: A Web Interface to the CgDNA Sequence-Dependent Coarse-Grain Model of Double-Stranded DNA. *Nucleic Acids Res.* **2018**, *46*, W5−W10.
(33) Basu, A.; Bobrovnikov, D. G.; Cieza, B.; Arcon, J. P.; Qureshi, Z.; Orozco, M.; Ha, T. Deciphering the Mechanical Code of the Genome and Epigenome. *Nat. Struct. Mol. Biol.* **2022**, *29*, 1178−1187.
(34) Flores, O.; Deniz, Ö.; Soler-López, M.; Orozco, M. Fuzziness and Noise in Nucleosomal Architecture. *Nucleic Acids Res.* **2014**, *42*, 4934−4946.
(35) Deniz, Ö.; Flores, O.; Battistini, F.; Pérez, A.; Soler-López, M.; Orozco, M. Physical Properties of Naked DNA Influence Nucleosome Positioning and Correlate with Transcription Start and Termination Sites in Yeast. *BMC Genomics* **2011**, *12*, 489.
(36) Buitrago, D.; Labrador, M.; Arcon, J. P.; Lema, R.; Flores, O.; Esteve-Codina, A.; Blanc, J.; Villegas, N.; Bellido, D.; Gut, M.; Dans, P.; Heath, S.; Gut, I.; Heath, I. B.; Orozco, M. Impact of DNA Methylation on 3D Genome Structure. 2020, DOI: 10.21203/rs.3.rs-36311/v1.
(37) Neguembor, M. V.; Arcon, J. P.; Buitrago, D.; Lema, R.; Walther, J.; Garate, X.; Martin, L.; Romero, P.; AlHaj Abed, J.; Gut, M.; Blanc, J.; Lakadamyali, M.; Wu, C. T.; Brun Heath, I.; Orozco, M.; Dans, P. D.; Cosma, M. P MiOS, an Integrated Imaging and Computational Strategy to Model Gene Folding with Nucleosome Resolution. *Nat. Struct. Mol. Biol.* **2022**, *29*, 1011−1023.
(38) Dziubańska-Kusibab, P. J.; Berger, H.; Battistini, F.; Bouwman, B. A. M.; Iftekhar, A.; Katainen, R.; Cajuso, T.; Crosetto, N.; Orozco, M.; Aaltonen, L. A.; Meyer, T. F. Colibactin DNA-Damage Signature Indicates Mutational Impact in Colorectal Cancer. *Nat. Med.* **2020**, *26*, 1063−1069.
(39) Pérez, A.; Castellazzi, C. L.; Battistini, F.; Collinet, K.; Flores, O.; Deniz, O.; Ruiz, M. L.; Torrents, D.; Eritja, R.; Soler-López, M.; Orozco, M. Impact of Methylation on the Physical Properties of DNA. *Biophys. J.* **2012**, *102*, 2140−2148.
(40) Barissi, S.; Sala, A.; Wieczór, M.; Battistini, F.; Orozco, M. DNAffinity: A Machine-Learning Approach to Predict DNA Binding Affinities of Transcription Factors. *Nucleic Acids Res.* **2022**, *50*, 9105−9114.
(41) Battistini, F.; Hospital, A.; Buitrago, D.; Gallego, D.; Dans, P. D.; Gelpí, J. L.; Orozco, M. How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition. *J. Mol. Biol.* **2019**, *431*, 3845−3859.
(42) Cuppari, A.; Fernández-Millán, P.; Battistini, F.; Tarrés-Solé, A.; Lyonnais, S.; Iruela, G.; Ruiz-López, E.; Enciso, Y.; Rubio-Cosials, A.; Prohens, R.; Pons, M.; Alfonso, C.; Tóth, K.; Rivas, G.; Orozco, M.; Solà, M. DNA Specificities Modulate the Binding of Human Transcription Factor A to Mitochondrial DNA Control Region. *Nucleic Acids Res.* **2019**, *47*, 6519−6537.
(43) Blanchet, C.; Pasi, M.; Zakrzewska, K.; Lavery, R. CURVES+ Web Server for Analyzing and Visualizing the Helical, Backbone and Groove Parameters of Nucleic Acid Structures. *Nucleic Acids Res.* **2011**, *39*, W68.

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

(44) Pérez, A.; Sponer, J.; Jurecka, P.; Hobza, P.; Luque, F. J.; Orozco, M. Are the Hydrogen Bonds of RNA (A·U) Stronger Than Those of DNA (A·T)? A Quantum Mechanics Study. *Chem. − Eur. J.* **2005**, *11*, 5062−5066.
(45) Terrazas, M.; Genna, V.; Portella, G.; Villegas, N.; Sánchez, D.; Arnan, C.; Pulido-Quetglas, C.; Johnson, R.; Guigó, R.; Brun-Heath, I.; Aviñó, A.; Eritja, R.; Orozco, M. The Origins and the Biological Consequences of the Pur/Pyr DNA·RNA Asymmetry. *Chem* **2019**, *5*, 1619−1631.
(46) Olson, W. K.; Gorin, A. A.; Lu, X. J.; Hock, L. M.; Zhurkin, V. B. DNA Sequence-Dependent Deformability Deduced from Protein-DNA Crystal Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 11163−11168.
(47) Dohnalová, H.; Lankaš, F. Deciphering the Mechanical Properties of B-DNA Duplex. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022**, *12*, No. e1575.
(48) Liebl, K.; Drsata, T.; Lankas, F.; Lipfert, J.; Zacharias, M. Explaining the Striking Difference in Twist-Stretch Coupling between DNA and RNA: A Comparative Molecular Dynamics Analysis. *Nucleic Acids Res.* **2015**, *43*, 10143.
(49) Lopez-Güell, K.; Battistini, F.; Orozco, M. Correlated Motions in DNA: Beyond Base-Pair Step Models of DNA Flexibility. *Nucleic Acids Res.* **2023**, *51*, 2633−2640.
(50) Gonzalez, O.; Petkevičiūtė, D.; Maddocks, J. H. A Sequence-Dependent Rigid-Base Model of DNA. *J. Chem. Phys.* **2013**, *138*, No. 055102.
(51) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269−6271.
(52) Dang, L. X. The Nonadditive Intermolecular Potential for Water Revised. *J. Chem. Phys.* **1992**, *97*, 2659.
(53) Banáš, P.; Hollas, D.; Zgarbová, M.; Jurečka, P.; Orozco, M.; Cheatham, T. E.; Šponer, J.; Otyepka, M. Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J. Chem. Theory Comput.* **2010**, *6*, 3836−3849.
(54) Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham, T.; Jurečka, P. Refinement of the Cornell et Al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886−2902.
(55) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084−3095.
(56) Lankaš, F.; Šponer, J.; Hobza, P.; Langowski, J. Sequence-Dependent Elastic Properties of DNA. *J. Mol. Biol.* **2000**, *299*, 695−709.
(57) de Helguero, F. Sui Massimi Delle Curve Dimorfiche. *Biometrika* **1904**, *3*, 84.
(58) Bhat, H. S.; Kumar, N. *On the Derivation of the Bayesian Information Criterion*, 2010.
(59) Skjærven, L.; Yao, X.-Q.; Scarabelli, G.; Grant, B. J. Integrating Protein Structural Dynamics and Evolutionary Analysis with Bio3D. *BMC Bioinf.* **2014**, *15*, 399.
(60) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; De Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422−1423.
(61) Hospital, A.; Andrio, P.; Cugnasco, C.; Codo, L.; Becerra, Y.; Dans, P. D.; Battistini, F.; Torres, J.; Goñi, R.; Orozco, M.; Gelpí, J. L. BIGNASim: A NoSQL Database Structure and Analysis Portal for Nucleic Acids Simulation Data. *Nucleic Acids Res.* **2016**, *44*, D272−D278.
(62) Hospital, A.; Battistini, F.; Soliva, R.; Gelpí, J. L.; Orozco, M. Surviving the Deluge of Biosimulation Data. *WIREs Comput. Mol. Sci.* **2020**, *10*, No. e1449.

https://doi.org/10.1021/acs.jcim.3c00741
J. Chem. Inf. Model. 2023, 63, 5259−5271

Chapter 6. Discussion

The primary focus of the research presented in this thesis has been to understand the connection between nucleic acid structures and gene regulation. To this end, we have had to explore nucleic acids, and in particular DNA, from 1D to 3D, putting emphasis on the connection between pure sequence information and chromatin conformation. This has forced us to navigate in the many resolution levels and to use a variety of theoretical techniques, from classical physics to artificial intelligence methods.

This thesis started from a first layer of regulation through the study of transcription factor (TF) binding mechanisms. As a second step, an additional layer of regulation was added by considering the nucleosome architecture, i.e. how nucleosomes are distributed: nucleosome depleted regions at the transcription start and terminating sites, as well as the positioning along the gene body. Thirdly, we explored the conformational changes that chromatin can undergo as a response to pathological conditions by analyzing the impact of oxidative stress on its three-dimensional structure, and how this modulates gene expression. We also explored the possibility of a protein-independent regulatory mechanism related to the formation of RNA-DNA·DNA triplexes, where an RNA transcript recognizes a target sequence in the genome and interferes its normal regulation. Finally, we turn our attention to RNAi regulatory mechanisms, conducting a foundational study on the physical properties of RNA duplexes, hoping they will be as useful for understanding biology as the properties of DNA duplexes.

6.1. Transcription Factor - DNA Binding

Transcription Factors (TFs) are key regulators of gene expression given that they can directly or indirectly activate, inactivate, or enhance transcription of DNA by interacting with specific sequences of DNA. The specific binding preferences of TFs result from a combination of factors such as intrinsic sequence preference, chromatin accessibility, nucleosome presence, cooperative effects with other effectors, or even phase separation (1). Nonetheless in order to predict TF binding sites *in vivo*, a first step is to understand and predict the sequence-dependent binding of TFs to naked DNA *in vitro*. This was the main objective in the first chapter of this thesis.

The development of accurate force fields for Molecular Dynamics (MD) simulations (2, 3) has enabled the structural study and characterization of the sequence-dependent DNA physical properties with an accuracy comparable with that of experiments. These properties, together with sequence-dependent groove interaction properties, allowed us to develop a predictor of DNA-TF binding sites. The predictor, called DNAffinity and based on a Random Forest Regressor, accurately reproduces experimental data from SELEX (4, 5) (R^2 of 0.70 ± 0.14), gcPBM (6) (R^2 of 0.93 ± 0.02) and uPBM (7) (R^2 of 0.69 ± 0.17), outperforming all state-of-the-art methods (8–12) for all sources of experimental data. One of the unique advantages of our predictor is that by using theoretically derived descriptors it could be extended to consider epigenetic variants or lesions. Very encouragingly, DNAffinity trained on *in vitro* data showed an excellent ability to detect the binding sites of the same transcription factor *in vivo* when excluding nucleosome occupied regions. The fact that the presence of a nucleosome explains the exclusion *in vivo* of otherwise good binding sites shows that nucleosomes are a good proxy for chromatin accessibility. However, due to the lack of concurrent *in vitro* and *in vivo* binding data the generalization of our model remains challenging. In any case, for the few instances where data exist, we found that combining detected potential binding sites together with nucleosome occupancy profiles trivially explains *in vivo* TF binding sites.

6.2. Nucleosome Positioning and its determinants

In the second publication of this thesis, we analyzed nucleosome positioning firstly by examining experimental nucleosome maps, and secondly by trying to understand the intrinsic (DNA physical properties) and extrinsic (effector proteins) that control the placement of the nucleosomes along the chromatin fiber. The analysis of reads obtained by MNase-seq experiments (13, 14) revealed nucleosome free regions (NFRs) at the transcription start site (TSS) and transcription termination site (TTS), as well as two well positioned nucleosomes at the start (+1) and at the end (-last) of the genes. We developed a model based on signal transmission theory (STT), which assuming two signal emitters at the first and last nucleosomes and a distance decay of such signals, allowed us to determine the intragenic nucleosome arrangement by statistical positioning. Genes with phased nucleosome arrangements (distances between +1 and last nucleosomes being multiples of 165 bps) exhibit periodic signals, while unphased genes show fuzzier architectures. Changes in the distance between +1 and -

last nucleosomes can alter nucleosome periodicity and fuzziness, as predicted by STT.

We then developed a neural network to predict NFRs, key determinants in the placement of nucleosomes. The model was based on two descriptors: the DNA deformation energy required to bend DNA from its naked to nucleosome-bound conformation and experimental TF binding site densities (15). The model successfully characterized (AUC of 0.96) the NFRs at both the TSS and the TTS of genes, outperforming previous predictive models (16). The combination of NFR predictions with STT enables highly accurate nucleosome positioning predictions with an average distance of 19bps from true peaks. These results are not only more accurate than those obtained by any existing predictive model but are in fact comparable to experimental noise levels in MNase-seq data, indicating that nucleosome positioning is well-defined even without complex chromatin remodeling mechanisms. This does not preclude the role of many other cellular mechanisms modulating nucleosome positioning which are not considered in our model like effectors cooperative effects, global chromatin structure, histone variants, epigenetic signals and indeed chromatin remodelers.

The ability to predict NFRs and nucleosome organization in yeast can help understanding the mechanisms of regulation in higher order organisms. As a proof of concept, we tested the applicability of our method on the human genome, where we obtained an AUC close to 0.70. Even with a reduced performance we were still able to use our descriptive features to predict NFRs with a higher accuracy than what would be expected at random. Given the additional complexity in studying higher organisms and that our current model was optimized for yeast, our approach—enhanced through fine-tuning and supplemented with additional experimental data—demonstrates the potential for studying nucleosome positioning arrays in any organism.

Finally, the use of synthetic biology experiments and computational analyses revealed that altering periodicity does not significantly affect gene expression. However, while changes in nucleosome phasing do not alter gene activity, inhibiting transcription leads to a loss of nucleosome periodicity. Thus, our findings suggest a causal relationship where expression levels influence nucleosome architecture, but not necessarily the way around.

.

6.3. The chromatin conformational changes upon Oxidative Stress

The question on how chromatin dynamics are modulated during gene regulatory mechanisms became then the central

motivation of this thesis. To fully comprehend its characteristics, the 2D study of the chromatin landscape was complemented with a 3D study, not just under physiological conditions but also under one of the most common DNA damage sources, oxidative stress (OS). To study its effect on chromatin a combination of gene expression analyses, MNase-Seq, Hi-C and Micro-C experiments (13, 17, 18) together with coarse-grained and data-driven models of the 3D genome structure (19, 20), were used. The presented integrative approach has allowed the study in detail of the 3D genome re-organization from low- to high-order chromatin structure in response to DNA damage.

DNA lesions showed different levels of condensation and decondensation of chromatin which correlated with transcriptional changes and DNA damage. The observed effects on nucleosome positioning and occupancy correlate with changes in chromatin fiber structure, with increased nucleosome interactions at distances below 600 bp and decreased interactions at longer distances. This observation was reflected in two analyses. Firstly, chromatin interaction domains were altered with an increase in number and a decrease in sizes upon oxidative stress. Secondly, nucleosome clutches (21), were both enhanced in number and reduced in size, particularly in upregulated genes and regions with DNA damage, leading to overall more extended fibers. Additionally, oxidative stress damage increased nucleosome positioning fuzziness, particularly in upregulated genes and damaged regions, reducing the length of nucleosome free regions and decreasing nucleosome occupancy in the gene body.

Overall, our results show that oxidative stress affects chromatin at two levels: a short-range chromatin decondensation and increased long-range and trans interactions. This suggests that the stress response favors an open chromatin state but involves complex, differential behaviors depending on the resolution level. The mechanisms of 3D genome organization in response to oxidative stress damage appear more intricate than previously proposed models for chromatin response after DNA damage, but in any case, the picture obtained suggests a massive rearrangement of chromatin, probably related to the activation of lesion repairing mechanisms and the activation of OS-response genes.

6.4. Triplexes as means of a regulatory mechanism

The fourth project explored the possibility of a protein-less DNA expression regulation based on the formation of RNA-DNA·DNA triplexes (22–24). Experimental and computational studies analyzed the stability of a variety of parallel triplexes that can be formed by

mixing complementary DNA and RNA strands. Many remained stable under physiological conditions, as predicted by advanced atomistic MD simulations (2, 25), and confirmed by melting and NMR experiments. The two most stable triplexes are the RNA·DNA·RNA and the RNA-DNA·DNA, the latter being the most compatible with the biology of the cell. Massive melting experiments were used to train a simple predictor that can determine melting free energy with an unprecedented accuracy (0.7 kcal/mol). This predictor was then used to identify all potential triplexes forming oligonucleotides (TFOs) in human long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) with triplex target sequences (TTSs) along the genome. Calculations revealed a large number of possible stable triplexes, significantly higher than those predicted by random models. The potential TTSs are enriched in regulatory regions and untranslated regions (UTRs), particularly in genes related to the development, morphology, and functioning of the central nervous system, suggesting a potential role of triplexes in an RNA-DNA mediated regulatory network. Furthermore, this work suggests that miRNAs, commonly known as post-transcriptional regulators, may have a more widespread nuclear function as transcription regulators via triplex formation than previously thought (26). Furthermore, mapping potential triplex formation with chromatin structure revealed evidence suggesting that triplex formation might play a role in stabilizing nucleosome arrays, potentially protecting nucleosomes from eviction and helping, as suggested by others, to compact chromatin. Future work is necessary to fully confirm these suggestions.

6.5 The extensive study of the properties of RNA

RNAs have been observed to be key regulators of numerous cellular processes (27–30), including gene expression. This has motivated us to focus efforts on expanding the understanding of the structure of RNA and its sequence-dependent physical properties. The final objective is to derive a foundational model that could be used to predict RNAi properties.

In the last work of this thesis, we presented a comprehensive analysis of the sequence-dependent structural and mechanical properties of the RNA duplex (RNA_2) using MD simulations (31). Overall, the studies of all the 136 unique tetramers showed that patterns of global deformability, signal transduction, and sequence-dependent deformation rules are different for RNA_2 than for DNA_2 (3). Sequence-dependent RNA_2 flexibility was observed to be simpler than that of the DNA_2, showing a more harmonic behavior and a reduced

sequence-dependent variability in terms of equilibrium properties. Previous claims that RNA_2 is "stiffer" than DNA_2 have been proven incorrect, needing a more detailed description given that its stiffness depends on the type of mechanical deformation introduced and the sequence environment. The collected data allowed the development of a simple harmonic-correlated mesoscopic model (32, 33), able to simulate long-RNAs and predict conformational ensembles with atomistic precision. This model will be of great utility to understand the behavior of long-RNA segments and RNA recognition in future predictive models.

6.6 Summary of integrative study of the chromatin landscape and gene expression mechanisms

The research conducted in this PhD thesis addresses the clear need to introduce complementary knowledge on the regulatory mechanisms that modulate gene expression and to understand the contributions of the chromatin landscape beyond DNA sequence. Initially, we discovered how sequence-dependent descriptors can accurately characterize TF-DNA binding interactions, revealing intrinsic and extrinsic modulators of TF binding across the genome, and highlighting the first regulation layer. We then turned our focus to nucleosomes, another crucial factor in chromatin accessibility. Our analysis of nucleosome positioning determinants uncovered the presence of two emitting nucleosome barriers, essential for the 2D organization of the genome within gene regulatory networks. Building on this, we examined the 3D chromatin conformation, demonstrating the complex mechanisms that chromatin undergoes in response to oxidative stress-induced DNA damage and how these changes correlate with gene expression. Additionally, we investigated triplex structures to understand their regulatory roles and conducted a detailed characterization of the properties of RNA duplexes using MD simulations as these improvements will be crucial for advancing RNA mediated networks.

In conclusion, this thesis suggests that gene regulation is mediated by a competitive interplay among factors such as sequence-dependent properties, nucleosome depleted regions, periodicity, protein effector binding, global and local chromatin structure, and the presence of interacting RNAs. This interplay assessed during this thesis characterizes the complex regulatory network.

6.7 Limitations and Future Perspectives

During the study of gene regulatory mechanisms in this PhD thesis we faced several limitations. Firstly, there is a need for additional experimental *in vivo* data on TF binding due to current data scarcity. Furthermore, our current predictive method, despite its accuracy, could benefit from additional descriptors such as epigenetic marks on DNA, lesions or mismatches to study their effect on binding mechanisms. Cooperative effects and multiple binding would require the integration of neighboring effects into sequence-dependent and deformation energy calculations.

Similarly, the presented reconstitution of nucleosome architectures currently does not investigate the cooperativity effects induced by neighboring genes, which might be a source of errors with some kind of gene architectures when the intergenic distance is small. Furthermore, while nucleosome positioning in yeast yielded satisfactory results, its accuracy diminished when applied to more complex organisms, necessitating reassessment of training, parameters, and variables, including longer genes and ultimately the effect of epigenetic marks on deformation energy calculations.

In the 3D study of the structure of chromatin, further assessment of other sources of DNA damage and their effects on chromatin dynamics are needed to fully understand the chromatin DNA damage landscape. Higher-resolution mapping of DNA damage could shed light on its relationship with transcription disruption and chromatin dynamic changes. Finally, a time-dependent series of experiment would help to monitor the changes on chromatin structure when damage occurs and along the repairing process.

The regulatory role of triplex forming oligonucleotides is compelling, and to confirm the possibility of an ancient regulatory mechanism, additional experimental data would be needed for the training and assessment of its regulatory impact upon *in vivo* binding. Synthetic biology type experiments seem the direction to follow to confirm the existence of a triplex-mediated regulatory mechanism. Additional experiments to determine the possibility of knocking down genes by adding a TFO-like oligo will be helpful to explore the concept of anti-gene therapies.

References

1. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. Annu Rev Biochem, 79.
2. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A., et al. (2015) Parmbsc1: A refined force field for DNA simulations. Nat Methods, 13.
3. Dans,P.D., Balaceanu,A., Pasi,M., Patelli,A.S., Petkevičiūtė,D., Walther,J., Hospital,A., Bayarri,G., Lavery,R., Maddocks,J.H., et al. (2019) The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. Nucleic Acids Res, 47, 11090–11102.
4. Smaczniak,C., Angenent,G.C. and Kaufmann,K. (2017) SELEX-seq: A method to determine DNA binding specificities of plant transcription factors. In Methods in Molecular Biology.Vol. 1629.
5. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M.J., et al. (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res, 20.
6. Gordân,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. Cell Rep, 3.
7. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol, 24.
8. Li,J., Sagendorf,J.M., Chiu,T.P., Pasi,M., Perez,A. and Rohs,R. (2017) Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. Nucleic Acids Res, 45.
9. Asif,M. and Orenstein,Y. (2020) DeepSELEX: Inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs. Bioinformatics, 36.
10. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol, 33.
11. Ruan,S., Swamidass,S.J. and Stormo,G.D. (2017) BEESEM: Estimation of binding energy models using HT-SELEX data. Bioinformatics, 33.

12. Yuan,H., Kshirsagar,M., Zamparo,L., Lu,Y. and Leslie,C.S. (2019) BindSpace decodes transcription factor binding signals by large-scale sequence embedding. Nat Methods, 16.
13. McKnight,L.E., Crandall,J.G., Bailey,T.B., Banks,O.G.B., Orlandi,K.N., Truong,V.N., Donovan,D.A., Waddell,G.L., Wiles,E.T., Hansen,S.D., et al. (2021) Rapid and inexpensive preparation of genome-wide nucleosome footprints from model and non-model organisms. STAR Protoc, 2.
14. Flores,O. and Orozco,M. (2011) nucleR: a package for non-parametric nucleosome positioning. Bioinformatics, 27, 2149–2150.
15. Pachkov,M., Balwierz,P.J., Arnold,P., Ozonov,E. and van Nimwegen,E. (2012) SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Res, 41, D214–D220.
16. Wang,J.-P., Fondufe-Mittendorf,Y., Xi,L., Tsai,G.-F., Segal,E. and Widom,J. (2008) Preferentially Quantized Linker DNA Lengths in Saccharomyces cerevisiae. PLoS Comput Biol, 4, e1000175.
17. Belton,J.M., McCord,R.P., Gibcus,J.H., Naumova,N., Zhan,Y. and Dekker,J. (2012) Hi-C: A comprehensive technique to capture the conformation of genomes. Methods, 58.
18. Hsieh,T.H.S., Weiner,A., Lajoie,B., Dekker,J., Friedman,N. and Rando,O.J. (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell, 162.
19. Neguembor,M.V., Arcon,J.P., Buitrago,D., Lema,R., Walther,J., Garate,X., Martin,L., Romero,P., AlHaj Abed,J., Gut,M., et al. (2022) MiOS, an integrated imaging and computational strategy to model gene folding with nucleosome resolution. Nat Struct Mol Biol, 29.
20. Buitrago,D., Labrador,M., Arcon,J.P., Lema,R., Flores,O., Esteve-Codina,A., Blanc,J., Villegas,N., Bellido,D., Gut,M., et al. (2021) Impact of DNA methylation on 3D genome structure. Nat Commun, 12, 3243.
21. Ricci,M.A., Manzo,C., García-Parajo,M.F., Lakadamyali,M. and Cosma,M.P. (2015) Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. Cell, 160.
22. Potaman,V.N. and Sinden,R.R. (1995) Stabilization of Triple-Helical Nucleic Acids by Basic Oligopeptides. Biochemistry, 34.
23. Knauert,M.P. and Glazer,P.M. (2001) Triplex forming oligonucleotides: Sequence-specific tools for gene targeting. Hum Mol Genet, 10.
24. Bacolla,A., Wang,G. and Vasquez,K.M. (2015) New Perspectives on DNA and RNA Triplexes As Effectors of Biological Activity. PLoS Genet, 11.
25. Dans,P.D., Walther,J., Gómez,H. and Orozco,M. (2016) Multiscale simulation of DNA. Curr Opin Struct Biol, 37.

26. Paugh,S.W., Coss,D.R., Bao,J., Laudermilk,L.T., Grace,C.R., Ferreira,A.M., Waddell,M.B., Ridout,G., Naeve,D., Leuze,M., et al. (2016) MicroRNAs Form Triplexes with Double Stranded DNA at Sequence-Specific Binding Sites; a Eukaryotic Mechanism via which microRNAs Could Directly Alter Gene Expression. PLoS Comput Biol, 12.
27. Spitale,R.C. and Incarnato,D. (2023) Probing the dynamic RNA structurome and its functions. Nat Rev Genet, 24.
28. Li,X. and Fu,X.D. (2019) Chromatin-associated RNAs as facilitators of functional genomic interactions. Nat Rev Genet, 20.
29. Sridhar,B., Rivas-Astroza,M., Nguyen,T.C., Chen,W., Yan,Z., Cao,X., Hebert,L. and Zhong,S. (2017) Systematic Mapping of RNA-Chromatin Interactions In Vivo. Current Biology, 27.
30. Farabella,I., Stefano,M. Di, Soler-Vila,P., Marti-Marimon,M. and Marti-Renom,M.A. (2021) Three-dimensional genome organization via triplex-forming RNAs. Nat Struct Mol Biol, 28.
31. Zgarbová,M., Otyepka,M., Šponer,J., Mládek,A., Banáš,P., Cheatham,T.E. and Jurečka,P. (2011) Refinement of the Cornell et al. Nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. J Chem Theory Comput, 7.
32. López-Güell,K., Battistini,F. and Orozco,M. (2023) Correlated motions in DNA: Beyond base-pair step models of DNA flexibility. Nucleic Acids Res, 51.
33. Gonzalez,O., Petkevičiute,D. and Maddocks,J.H. (2013) A sequence-dependent rigid-base model of DNA. Journal of Chemical Physics, 138.

Chapter 7. Conclusions

- The first layer of gene regulatory networks, TF-DNA binding, can be accurately predicted by the use of sequence-dependent structural and physical properties of the naked DNA that define *in vitro* readout mechanisms. DNAffinity, the developed algorithm, provides accurate predictions that out-performed current state-of-the art methods. Putative *in vitro* binding sites can be extrapolated to *in vivo* binding sites by adding a chromatin accessibility proxy: the nucleosome coverage.

- The second layer of regulatory elements, nucleosome arrangements, can be accurately disentangled by the combination of a neural network for the detection of nucleosome free regions, and statistical positioning, as defined by signal transmission theory, for intragenic nucleosome arrays. The precision of the predicted nucleosome arrays surpasses that of previous methods and is within the range of experimental noise. This proves that the basal nucleosome array can be predicted from the basis of intrinsic (physical properties) and extrinsic (protein binding) descriptors.

- Lesions can produce changes in chromatin structure, either by itself or by the activation of repairing mechanisms. For example, oxidative stress-induced DNA damage affects the 3D structure of the chromatin, especially in regions where we observe a change in the expression of genes or where the damage, as detected by lesion-dependent epigenetic signals, is more prominent. The overall changes observed in these regions, accounting for the simultaneous damage and repair, show a gain of interactions at very short distances, a loss of mid-range interactions and a gain of interactions at large distances, with significant changes in nucleosome arrangements and the associated nucleosome free regions (NFRs).

- Triplex forming oligonucleotides (TFOs) are enriched in human long noncoding RNAs and microRNAs in comparison to those predicted by random models. These TFOs recognize with high affinity triplex target sequences (TTS) located in regulatory regions and UTRs, portraying their regulatory role.

- The study of all the unique RNA_2 tetramers through extensive MD simulations reveals unique conformational and dynamic properties which differ from the DNA_2. In essence, RNA_2 follows a more harmonic behavior and a reduced sequence-dependent variability in terms of equilibrium properties. Accepted claims on the generally higher stiffness of the RNA duplex compared with the DNA one should be re-evaluated. The developed mesoscopic model might help to explore the properties of RNAi and other regulatory ribonucleotide acids.

Annex

1. DNAffinity: a machine-learning approach to predict DNA binding affinities of transcription factors

Supplementary Info

DNAffinity: A MACHINE-LEARNING APPROACH TO PREDICT DNA BINDING AFFINITIES OF TRANSCRIPTION FACTORS

Sandro Barissi[1&], Alba Sala[1&], Milosz Wieczor[1,2], Federica Battistini[1,3*] and Modesto Orozco[1,3*]

[1] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Baldiri Reixac 10–12, 08028 Barcelona, Spain.
[2] Department of Physical Chemistry. Gdansk University of Technology, 80-233 Gdańsk, Poland.
[3] Department of Biochemistry and Molecular Biology. University of Barcelona, 08028 Barcelona, Spain.
[&] Equally contributing authors
* Correspondence to: federica.battistini@irbbarcelona.org or modesto.orozco@irbbarcelona.org

---

Supplementary Methods

Supplementary Tables

Supplementary Table 1 Datasets of the features and labels used in this work and corresponding post processing methods applied.

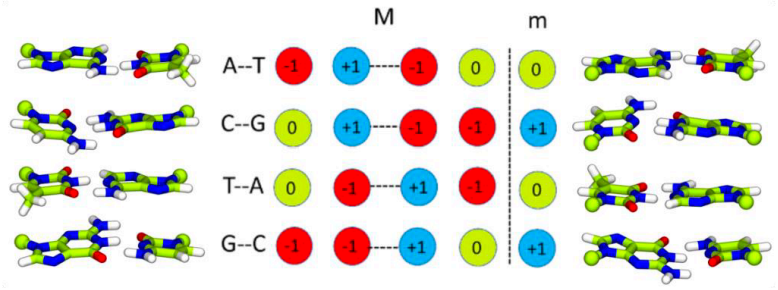
Labels dataset	**Post processing methods**
uPBM (DREAM5 (41))	Input 35-mer oligos for 66 mouse TFs. Alignment and trimming using motif detected by MEME suite (k-mer, typically 12-mer).
gcPBM (GSE59845 and GSE44604)	gcPBM data already aligned. When detected, removal of multiple binding sites.
HT-SELEX (PRJEB14744) (first dataset)	Bioconductor Package SELEX Removed TFs with no k-mer counts > 100 in the 0th cycle Trained and tested data from the penultimate SELEX cycle
HT-SELEX (PRJEB29730)	Bioconductor Package SELEX

(second dataset)	Removed TFs with no k-mer counts > 100 in the 0th cycle Trained and tested data from the penultimate SELEX cycle

Feature dataset	Link
Base pair parameters (average and flexibilities) and Electrostatics	MiniABC for each tetramer (1) Simulations stored in BigNASim (2) https://github.com/Jalbiti/DNAffinity

Supplementary Table S2. Average correlation (R^2) and MSE for dataset (uPBM and HT-SELEX) among all the studied proteins, using cross-validation (CV) and bootstrap respectively.

Method	Metric	CV	Bootstrap
HT-SELEX	R^2	0.62 ± 0.21	0.66 ± 0.19
HT-SELEX	MSE	0.002 ± 0.001	0.001 ± 0.001
uPBM	R^2	0.63 ± 0.12	0.69 ± 0.17
uPBM	MSE	0.007 ± 0.009	0.011 ± 0.004

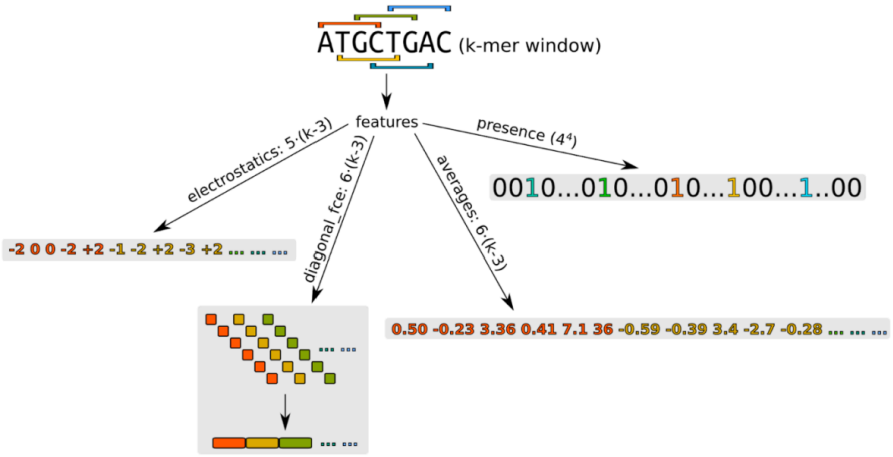


Supplementary Figure S1. Scheme for describing the different electrostatics of each possible base pair at the major (M) and minor (m) groove. At each position a value of +1, -1 or 0, has

been assigned depending on the presence of a hydrogen bond acceptor (red), and donor (blue), or a non-polar hydrogen or methyl group (green) respectively.

Sequence vector

Each sequence considered as a potential TFBS was read as a k-mer vector in our algorithm. For processing, each k-mer was broken down into overlapping tetramers, and each tetramer was assigned the corresponding 4 classes of features (see Supplementary Figure S2). We considered 17 per-tetramer features and 256 presence features whose number is not dependent on the k-mer length; for example, for a 10-mer we considered a total of 375 = 17*(10-3) + 256 parameters.



Supplementary Figure S2. Scheme of the sequence vector and the division into overlapping tetramers, with the attribution of each class of parameters.

Undersampling approaches

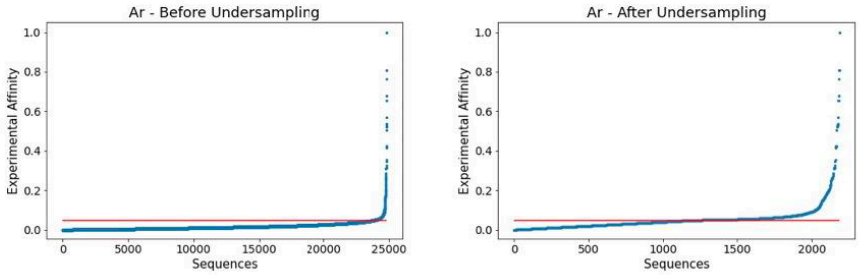
Undersampling was applied to the uPBM data to have a clear distinction between high and low affinity sequence without the extreme redundancy of low affinity points. More specifically, we applied undersampling only to uPBM data because in these datasets almost all TFs had only about 50-100 high-affinity k-mers (with affinities above 0.2) out of 25,000 data points, making the training extremely skewed so that the important data was lost among the big amount of low-affinity kmers (as an example, see the undersampling of the Ar dataset in Supplementary Figure S3).

To perform the undersampling, we divided the interval [0,1] of affinities into N smaller subintervals, namely:

$$[0,1] = k = 0N - 1\left[\frac{k}{N}, \frac{k+1}{N}\right]$$

If *N* is large (usually N=25000), the intervals closer to 0 will have a large density of points as compared to the intervals closer to 1. We can use this fact to pick samples from the low-affinity points and hence reduce the excess of these data points. In other words, if k ≤ N/2 we only picked $n = 1$ point from the interval $\left[\frac{k}{N}, \frac{k+1}{N}\right]$ and if k > N/2 we picked all the available data in the interval, if any.

By using a revisited version of the k nearest neighbors method (https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761) we made sure that no range of affinities was missing and at the same time we got rid of possible noise.



Supplementary Figure S3. Illustrative example of the affinity profile before (left) and after (right) undersampling of the uPBM dataset for the Ar TF. We divided the range of affinities into 25,000 intervals, and set a threshold not to undersample the top affinity 1,000 sequences, which translates into an affinity threshold around 0.05. This yielded 1196 intervals (affinity < 0.05) for which we sampled the data. In each one of these intervals we randomly take into consideration k=1 point, and from the interval above the threshold we take the remaining 1000 points. In short, we have 1,000 points to cover almost the whole range of affinity (range 0.05-1) and 1,196 of very low affinity (range 0-0.05).

For the HT-SELEX data, we removed noisy data using the 'Probability' column given by the Markov Model to under sample the dataset: we only used the top 10% points regarding probabilities to train the model and we removed the lower tail. The efficiency and reliability of this method has been previously proven (3).

Removal of multiple binding sites

For the gcPBM datasets (GSE59845 and GSE44604), where each probe in the array is aligned at the central TF binding site, sequences that included possible multiple binding sites were removed. For this purpose, we first generated a 6 bps PWM using the top affinity 36-mers and used it to scan the flanking sequence surrounding the central TFBS The sequence was discarded when comparing the flanking sequences with the PWM generated we found more than 4 continuous bps in common. This procedure is an adaptation of the protocol described in (4), which provides a final number of curated sequences similar to ours.

Feature importance

To calculate the importance of each feature, we used a function called "feature importances" integrated in the sklearn package that we used for the Random Regressor algorithm; the higher the value, the more important the feature. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance (https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3).

HT-SELEX data quality assessment

Analysis is based on HT-SELEX raw observed counts. We were able to classify whether the raw data would allow us to predict the binding affinities with high accuracy, using the correlation between the counts across the different cycles and the statistical distribution across the penultimate cycle, which was used for prediction. The SVM was then trained to classify all proteins, analysing whether our DNA protein-affinity predictor would perform better than a random predictor ($R^2 > 0.5$) and all proteins were reassessed using this classification as a filter. A radial-basis function kernel (5), a regularization parameter of 10 and a gamma value of 0.1 were used to define our SVM parameters.

Prediction quality

For the R^2, we used the r2_score function in scikit-learn that computes the coefficient of determination; the formula used is the following:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Where $\hat{y}_i$ is the predicted value of the i-th sample, y_i is the corresponding true value for total n samples and $\overline{y}$ is the average value.

For MSE we used the mean_squared_error function in scikit-learn that computes the coefficient of determination; the formula used is the following:

$$MSE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2$$

Where $\hat{y}_i$ is the predicted value of the i-th sample, y_i is the corresponding true value for total n samples.

For Pearson correlation we used the stats.pearsonr function in scikit-learn (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html), the formula used is the following:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \quad (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Where y_i is the predicted value of the i-th sample, x_i is the corresponding true value for total n samples, $\overline{x}$ and $\overline{y}$are the respective mean values of the true and predicted samples.

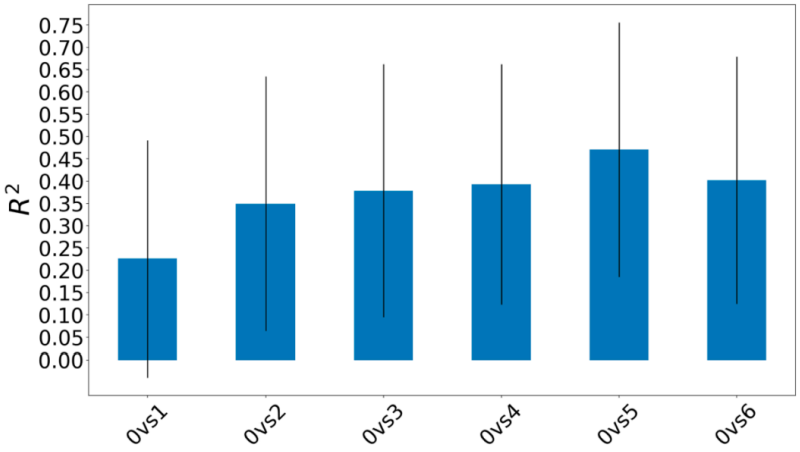
We also tested the choice of the training set using cross-validation, which performs the simulation K times dividing the data into K partitions and using each time one different partition as a test set. Changing the algorithm to k-fold cross-validation **(K=10; 90/10 random split**), we found very similar results are obtained (see Supplementary Table S2).

Supplementary References

1. Dans,P.D., Balaceanu,A., Pasi,M., Patelli,A.S., Petkevičiūtė,D., Walther,J., Hospital,A., Bayarri,G., Lavery,R., Maddocks,J.H., *et al.* (2019) The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. *Nucleic Acids Res.*, **47**, 11090.
2. Hospital,A., Andrio,P., Cugnasco,C., Codo,L., Becerra,Y., Dans,P.D., Battistini,F., Torres,J., Goñi,R., Orozco,M., *et al.* (2016) BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.*, **44**, D272–D278.
3. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G., *et al.* (2013) DNA-Binding Specificities of Human Transcription Factors. *Cell*, **152**, 327–339.
4. Zhou,T., Shen,N., Yang,L., Abe,N., Horton,J., Mann,R.S., Bussemaker,H.J., Gordân,R. and Rohs,R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 4654–4659.
5. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Müller,A., Nothman,J., Louppe,G., *et al.* (2012) Scikit-learn: Machine Learning in Python.

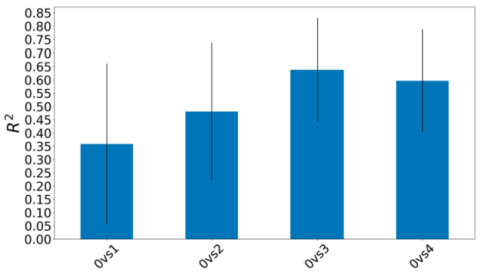
6. Wang,S., Zhang,Q., Shen,Z., He,Y., Chen,Z.-H., Li,J. and Huang,D.-S. (2021) Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture. *Mol. Ther. - Nucleic Acids*, **24**, 154–163.
7. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
8. Ma,W., Yang,L., Rohs,R. and Noble,W.S. (2017) DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding. *Bioinformatics*, **33**, 3003–3010.
9. Zhang,Q., Shen,Z. and Huang,D.S. (2021) Predicting in-vitro Transcription Factor Binding Sites Using DNA Sequence + Shape. *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, **18**, 667–676.
10. Li,J., Sagendorf,J.M., Chiu,T.-P., Pasi,M., Perez,A. and Rohs,R. (2017) Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Res.*, **45**, 12877–12887.
11. Asif,M. and Orenstein,Y. (2020) DeepSELEX: inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs. *Bioinformatics*, **36**, i634–i642.

Supplementary Results

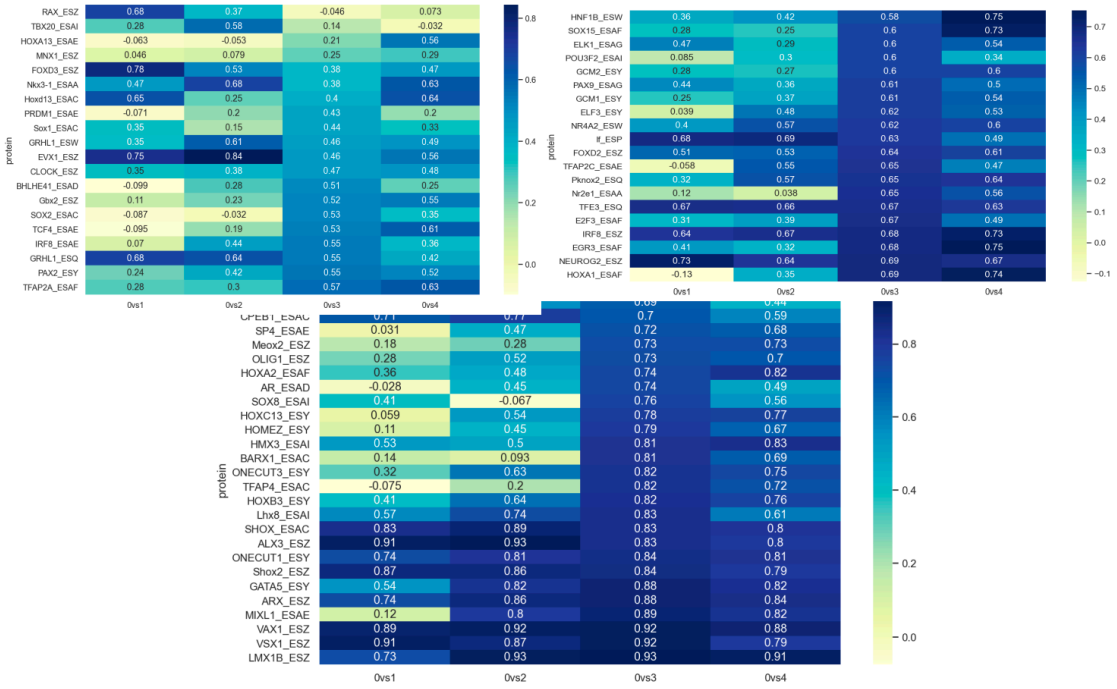


Supplementary Figure S4. Average correlation between experimental and predicted affinities when training on different cycle pairs among all the studied proteins, with standard deviations for the second HT-SELEX dataset (see Methods and Supplementary Table S1).

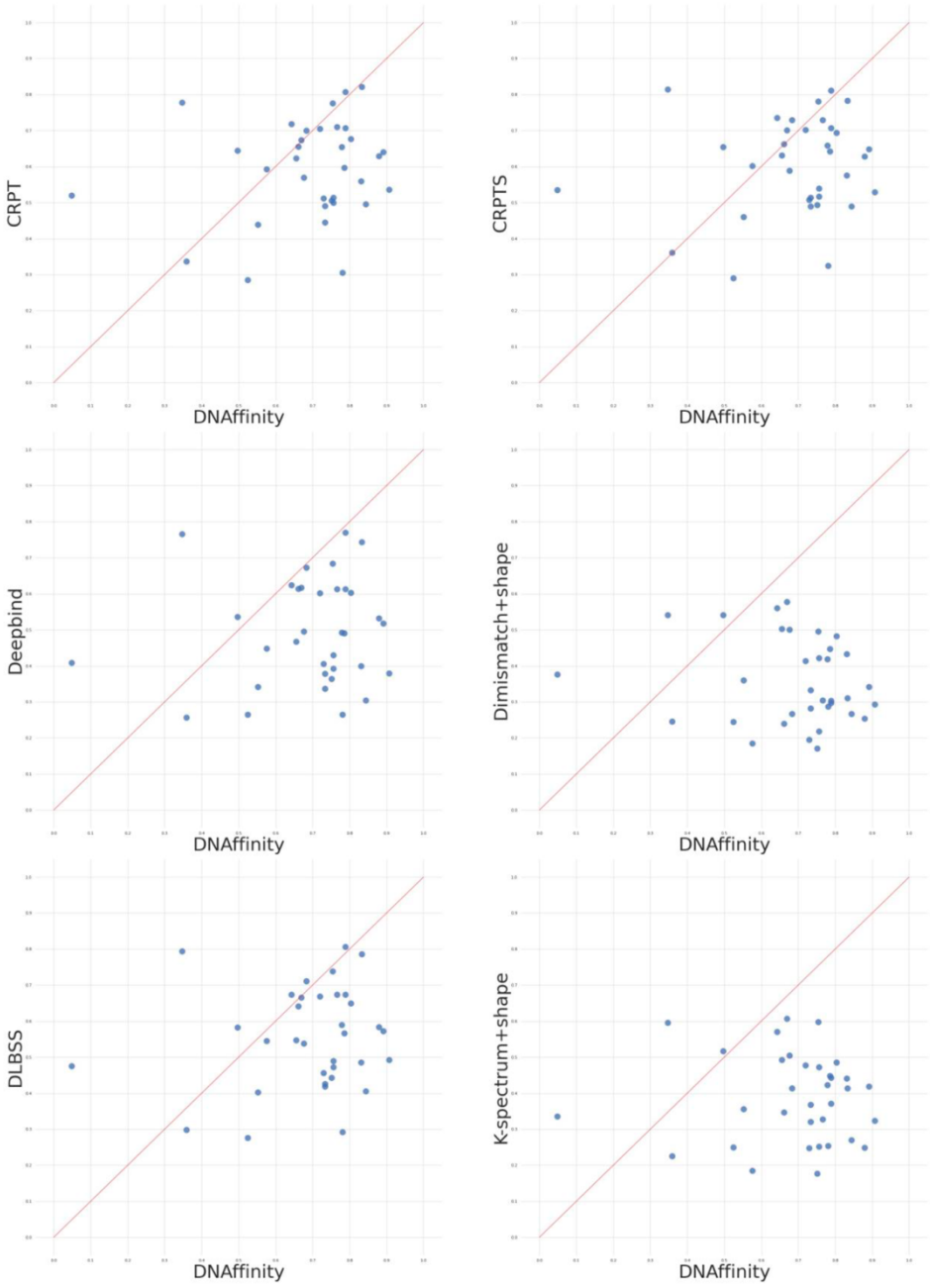
Supplementary Figure S5. Correlations between experimental affinities and ML predictions when model was trained on different pairs of cycles (0vs) from HT-SELEX for the second HT-SELEX dataset (see Methods and Supplementary Table S1).



Supplementary Figure S6. Figure. Average correlation between experimental and predicted affinities when training on different cycle pairs among all the studied proteins, with standard deviations for the first HT-SELEX dataset (see Methods and Supplementary Table S1).

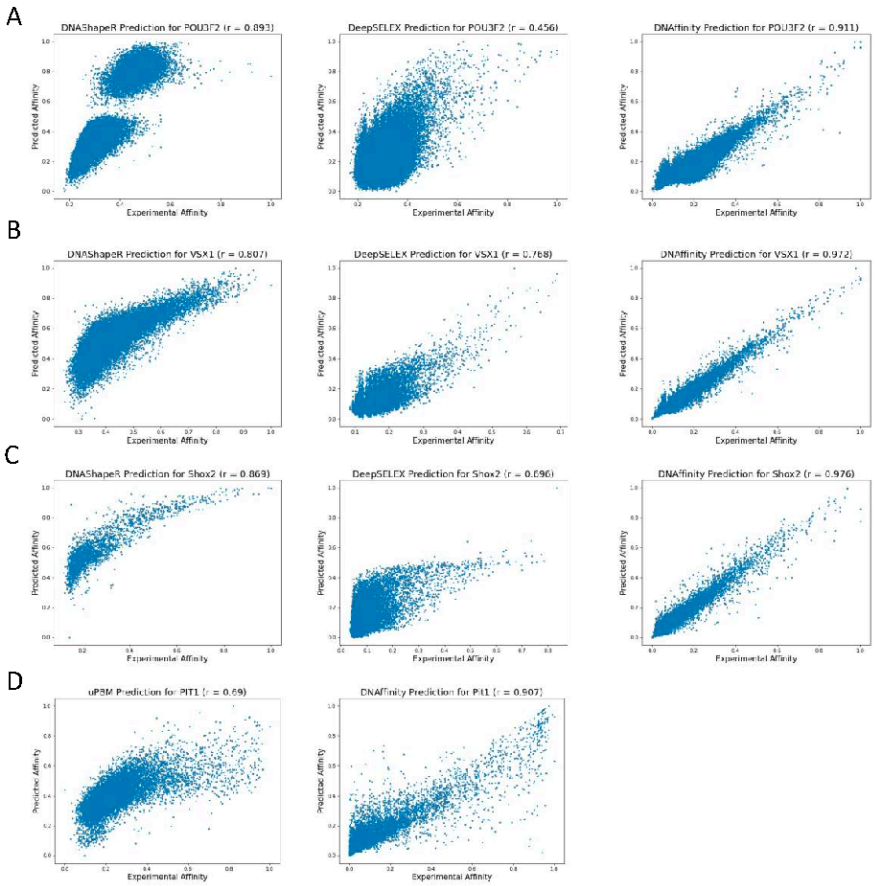


Supplementary Figure S7. Correlations between experimental affinities and ML predictions when model was trained on different pairs of cycles (0vs) from HT-SELEX and predicted affinity for the first dataset (see Methods and Supplementary Table S1).

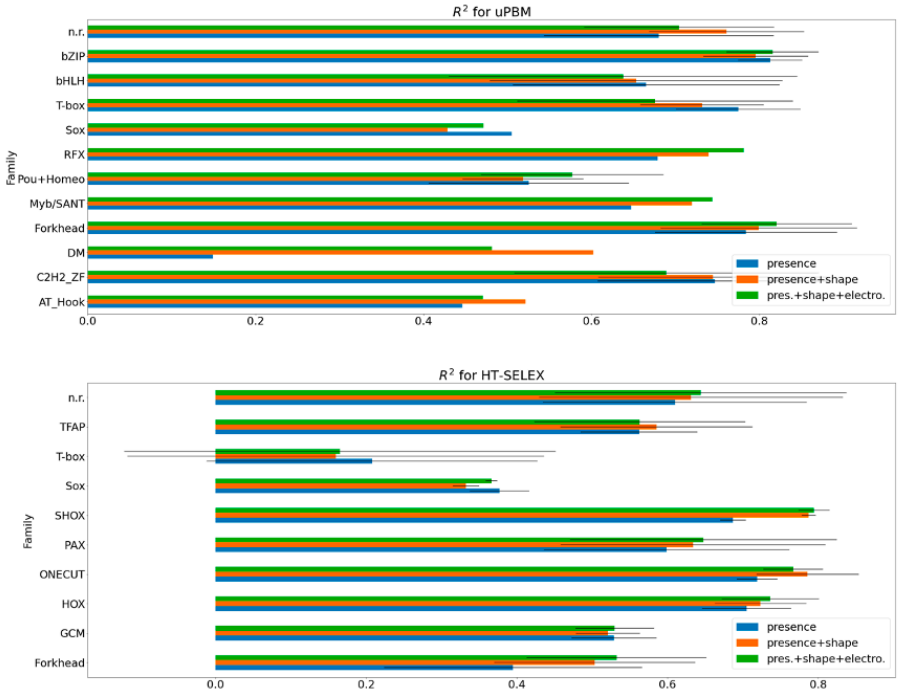


Supplementary Figure S8. Correlation between determination coefficients obtained by previously developed predictors and our method (DNA affinity). Methods used for

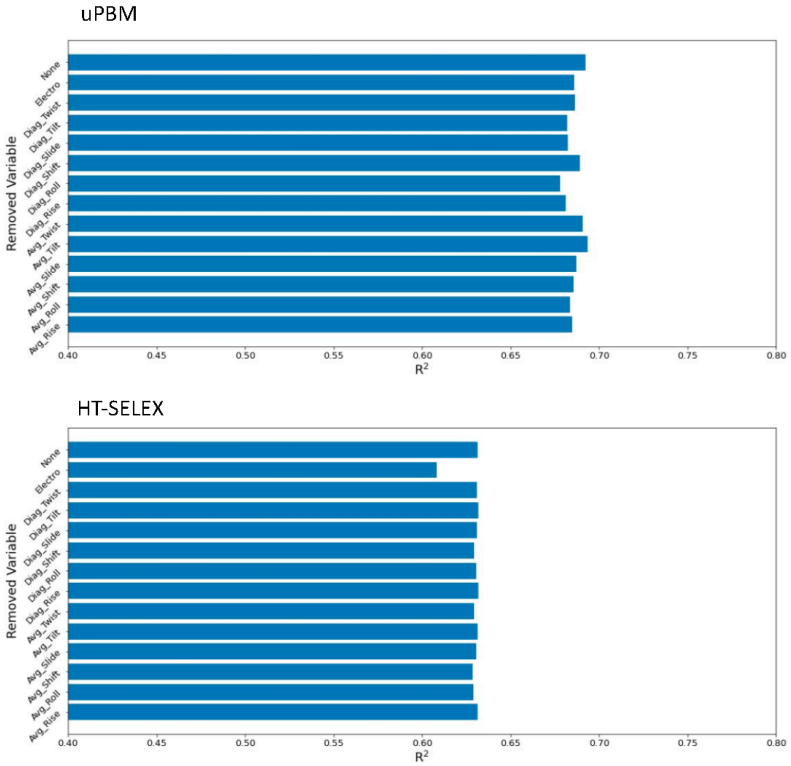
benchmarking are: CRPTS/CRPT (a hybrid convolutional recurrent neural network (CNN/RNN) combining DNA sequence and DNA shape features)(6); Deepbind (CNN model based on primary DNA sequences) (7); two kernel-based methods (spectrum + shape kernel, di-mismatch + shape kernel) (8); and a deep learning method DLBSS (9). The data were taken from (6).



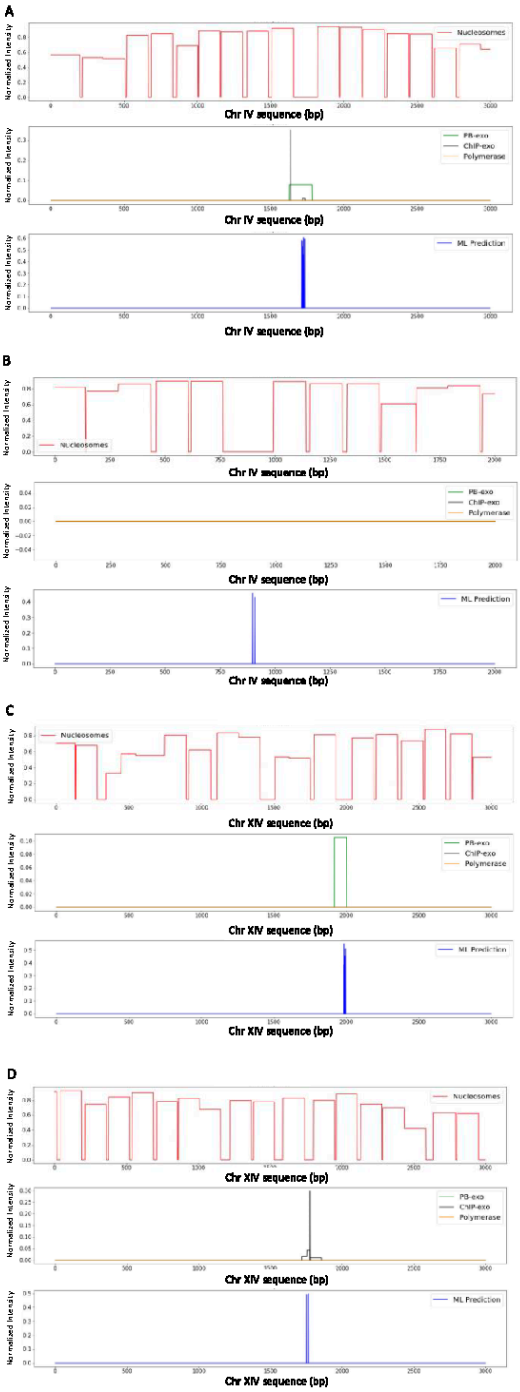
Supplementary Figure S9. Comparison of the affinity values predicted for 3 TFs using: A-C) DNAShapeR (left panels, (10)), DEEPSELEX (central panels, (11)) and our predictor (DNAffinity, right panel). The 3 cases selected are among all the ones studied using HT-SELEX data, for which the predictions of DNAShapeR (A), DEEPSELEX (B), DNAffinity (C) perform at its best respectively. D) CRPTS/CRPT (left panel) and DNAffinity (right panel) for a TF among all the ones studied using uPBM data with high performance.



Supplementary Figure S10. Average correlations by protein families, with standard deviation, using three different combinations of features for uPBM: presence (in blue), base pair parameters (presence + shape, in orange) and all combined (presence+shape+electrostatics, in green).



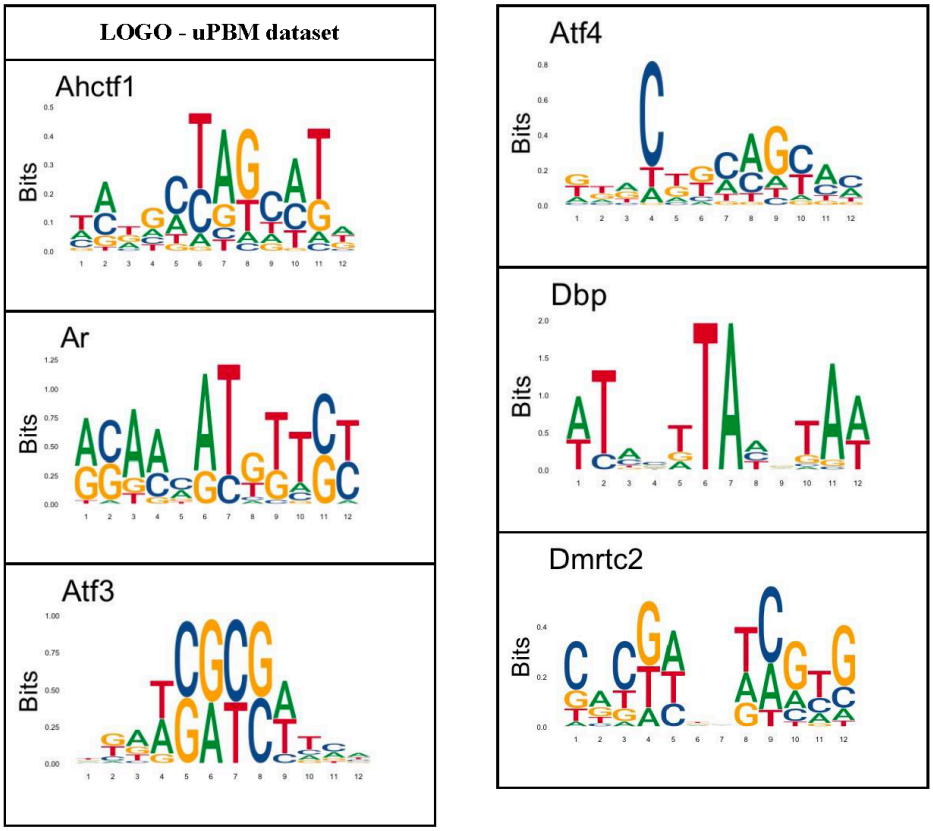
Supplementary Figure S11. Determination coefficient (R^2) between predicted and experimental affinities using uPBM (top panel) and first HT-SELEX (bottom panel) dataset, considering all the descriptors used in the machine learning algorithm (none) and removing each variable one by one (average, stiffness parameters and electrostatics).

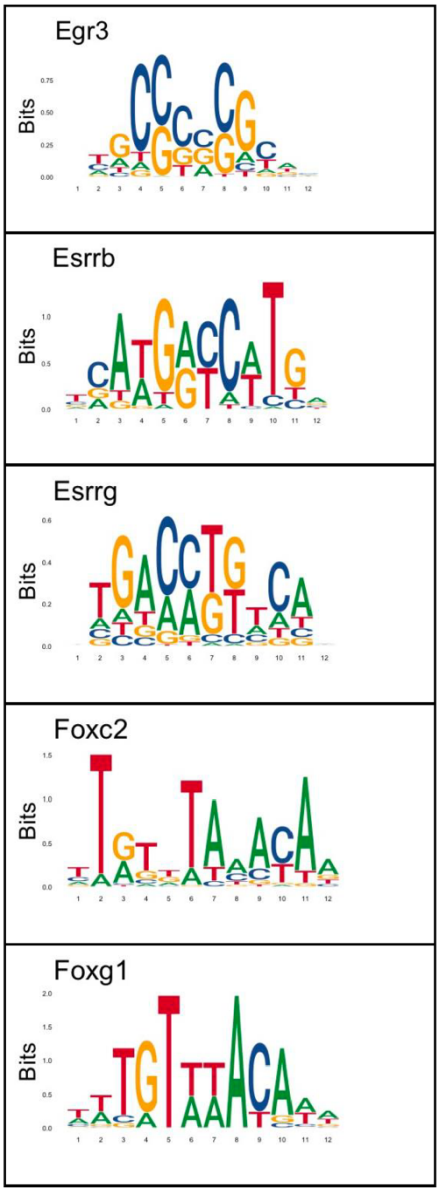

A
Normalized Intensity
Nucleosomes
Chr IV sequence (bp)
PE-exo
ChIP-exo
Polymerase
ML Prediction
B
C
Chr XIV sequence (bp)
D

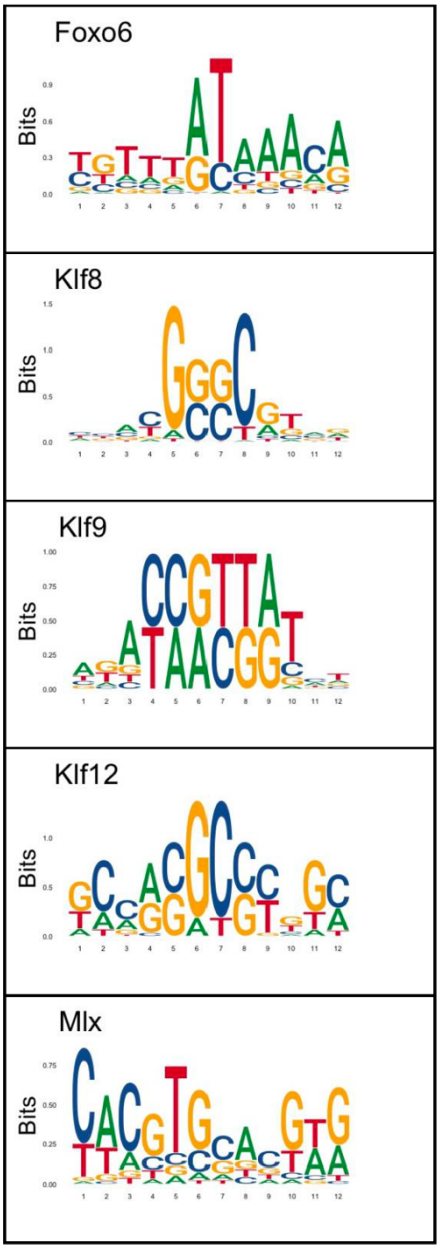
Supplementary Figure S12. Examples of the TF positioning prediction compared with experimental data (CHiP-exo, PB-exo and Polymerase; green, black and yellow lines, respectively) and nucleosome positioning (red line). A) True positive case, the ML prediction coincides with experimentally detected TFBS in a nucleosome free region. B) False positive case, the ML prediction coincides with a nucleosome free region, but no TFBS has been detected experimentally. C-D) "Contentious" false positive case, the ML prediction coincides with a nucleosome free region and a TFBS only detected by one experimental technique (PB-exo and Chip-exo respectively).

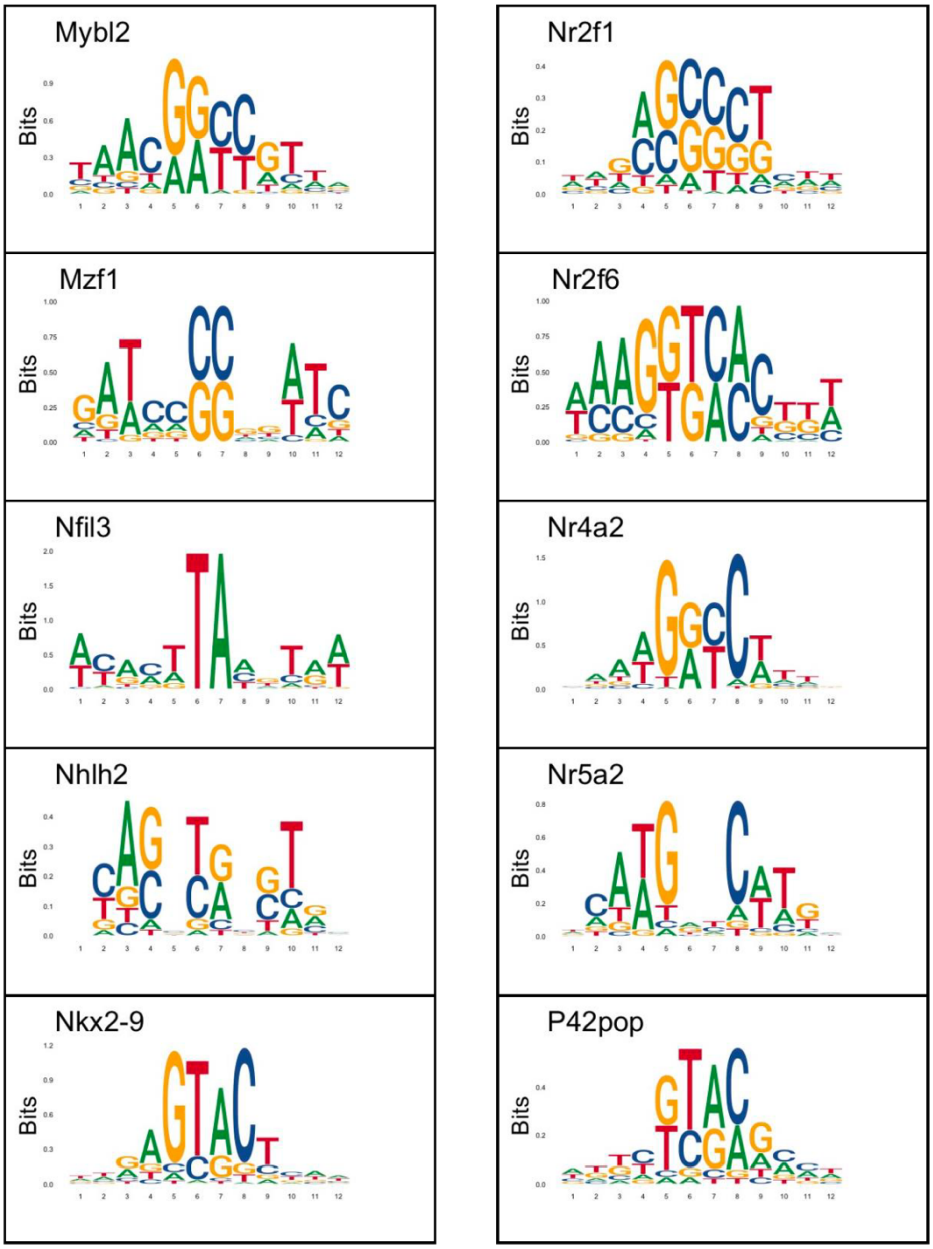
Supplementary Tables

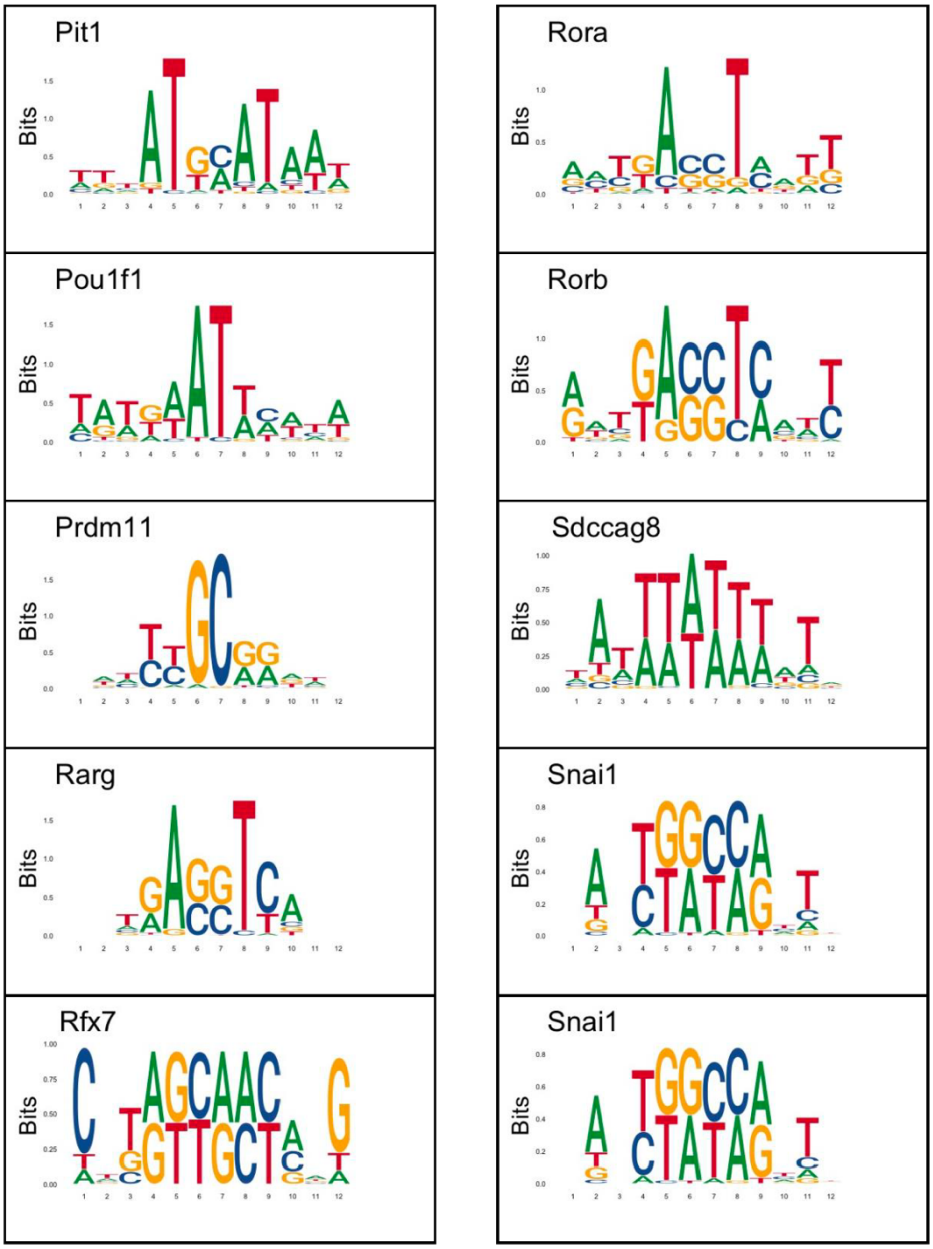
Supplementary Table S3. LOGOs of the position weight matrices after the training for uPBM and HT-SELEX for each TF.

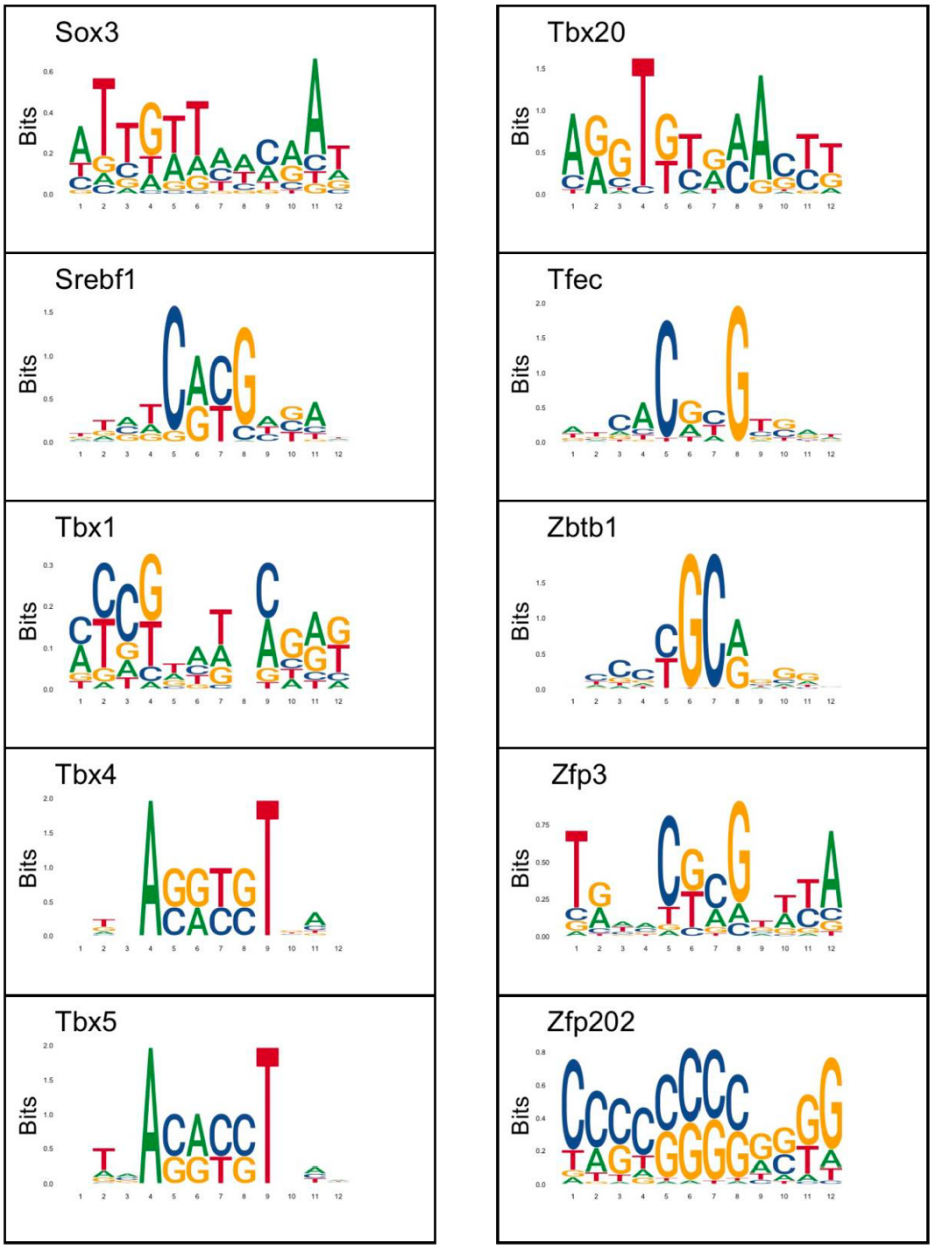


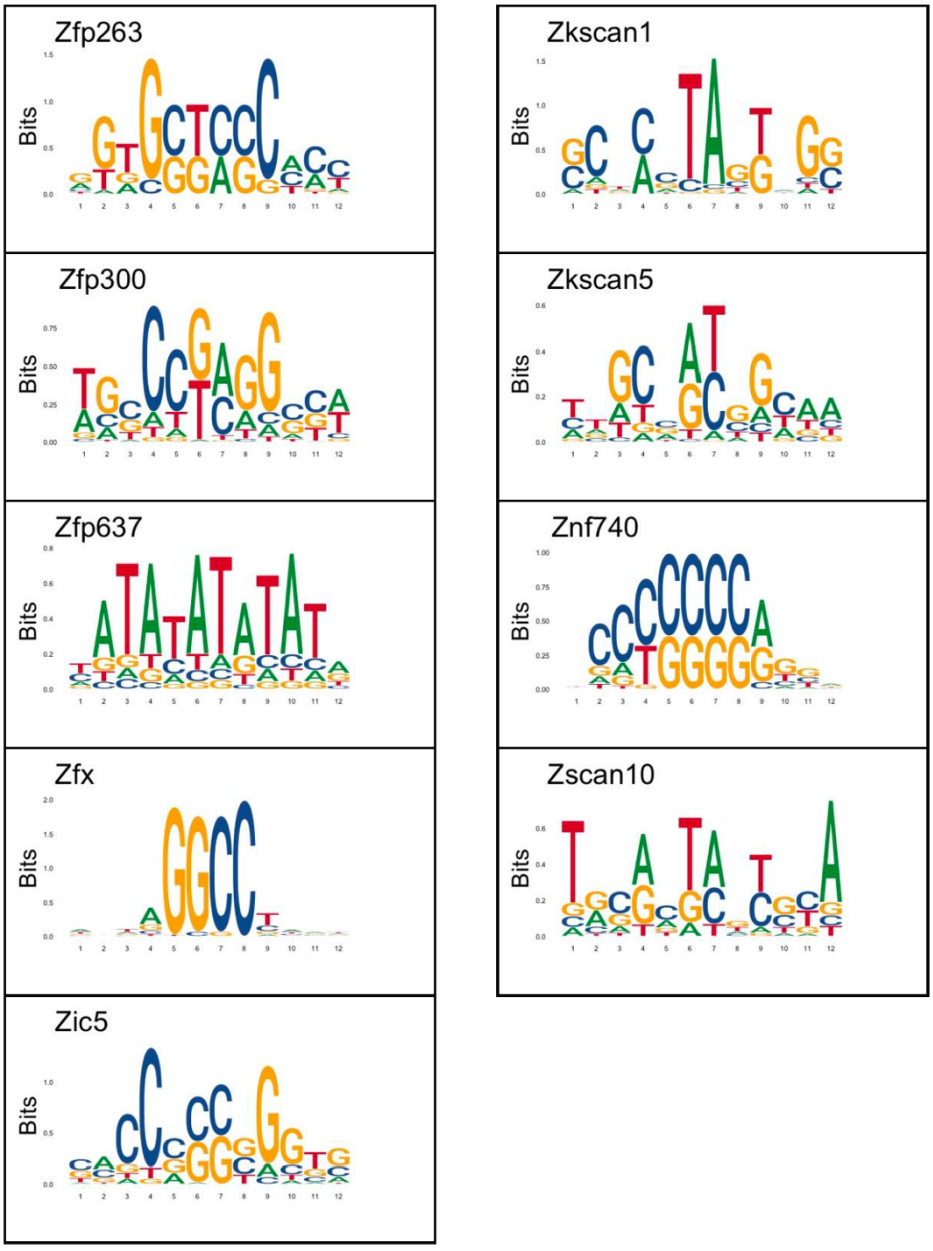

Egr3
Esrrb
Esrrg
Foxc2
Foxg1
Bits

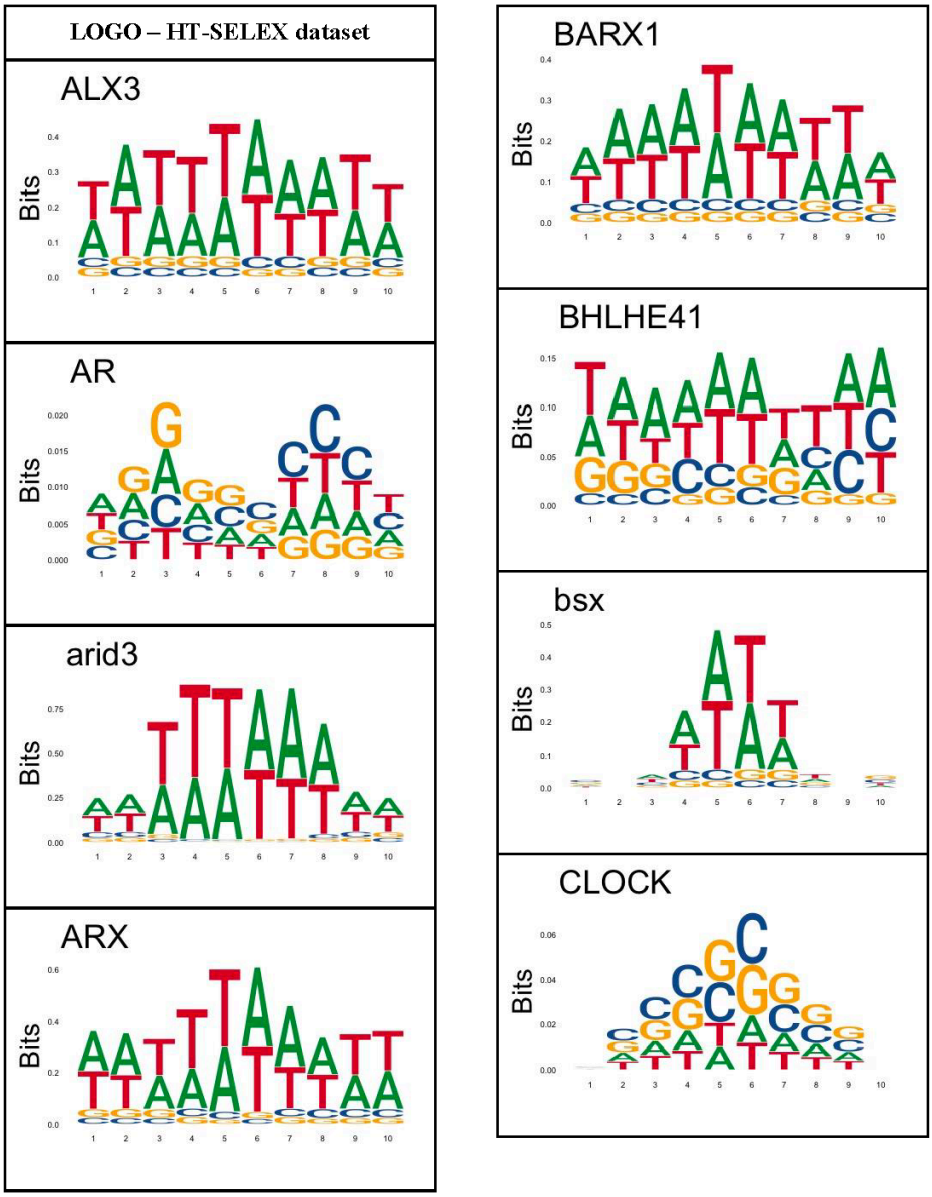

Foxo6
Klf8
Klf9
Klf12
Mlx
Bits

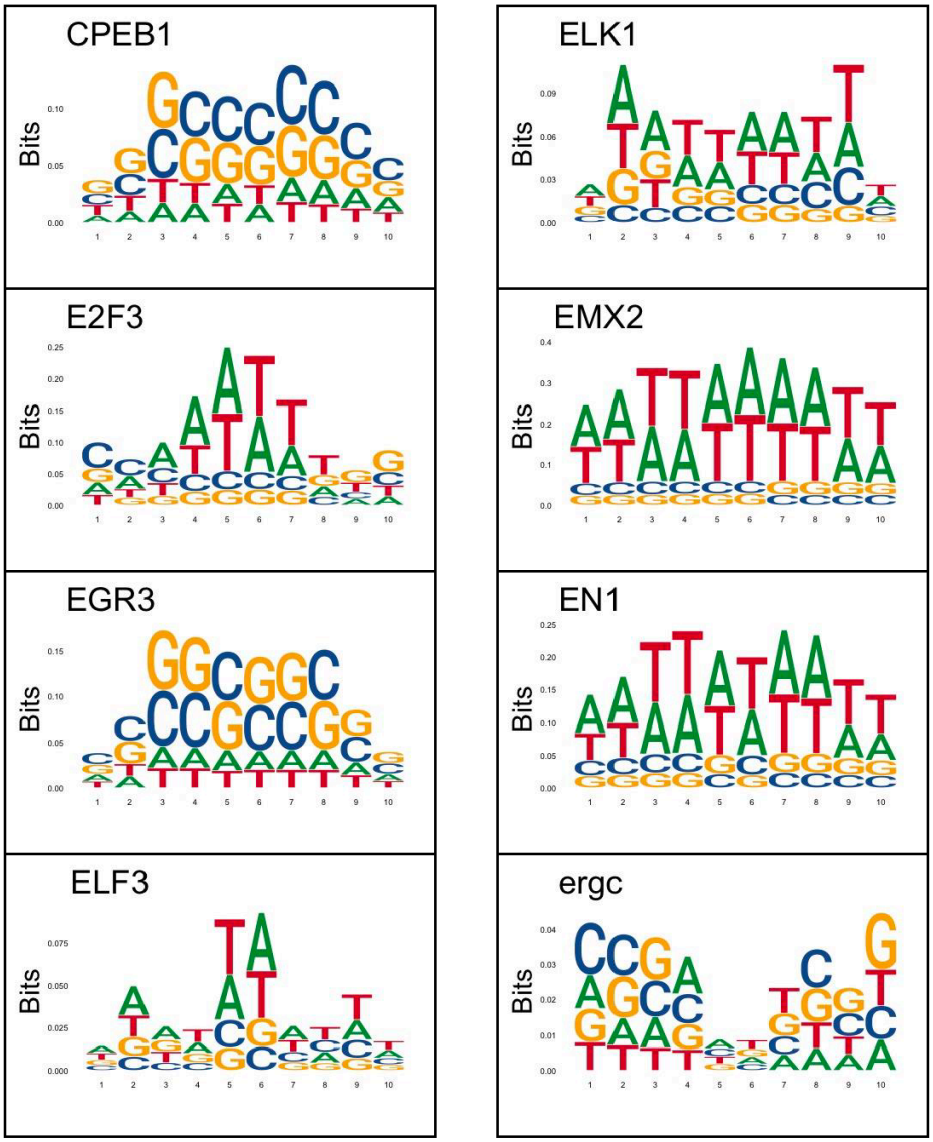

Mybl2
Nr2f1
Mzf1
Nr2f6
Nfil3
Nr4a2
Nhlh2
Nr5a2
Nkx2-9
P42pop
Bits

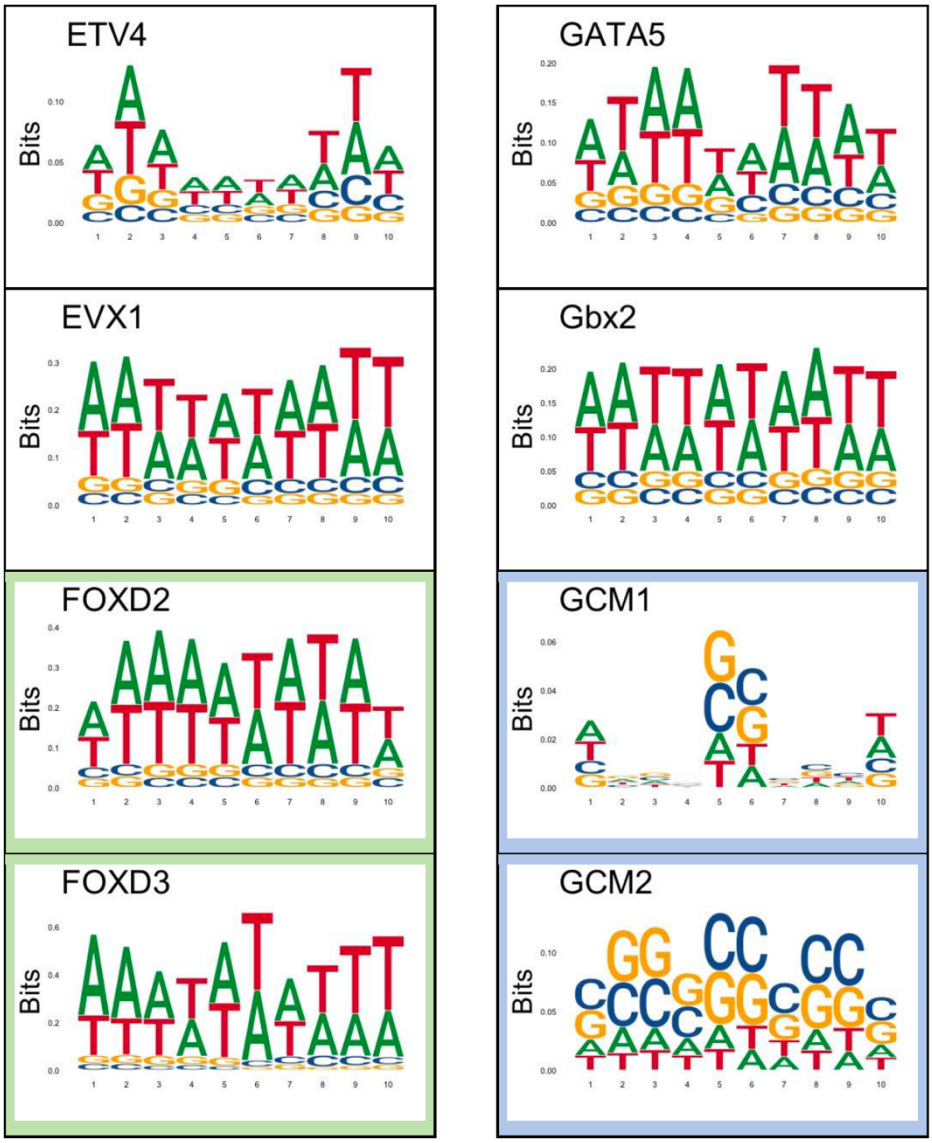

Pit1
Bits
Rora
Bits
Pou1f1
Bits
Rorb
Bits
Prdm11
Bits
Sdccag8
Bits
Rarg
Bits
Snai1
Bits
Rfx7
Bits
Snai1
Bits

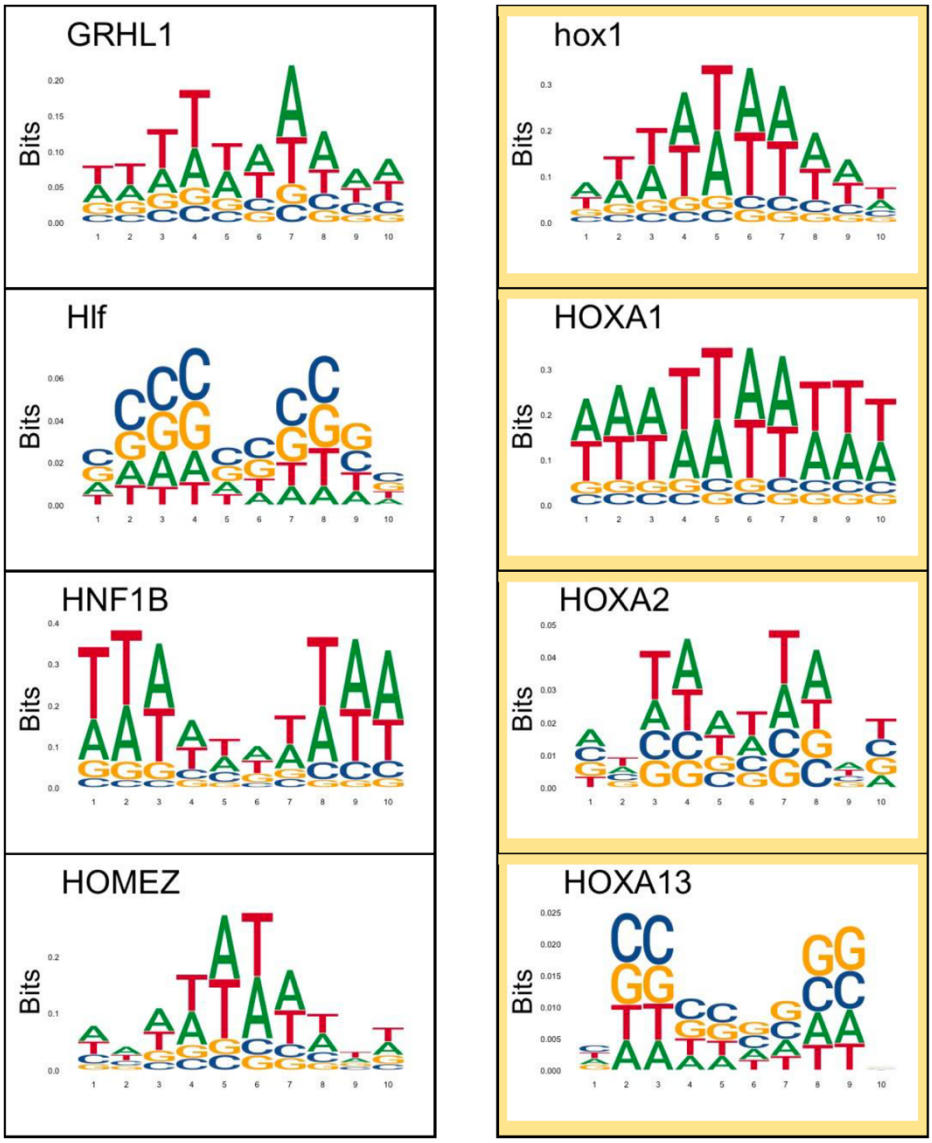

Sox3
Tbx20
Srebf1
Tfec
Tbx1
Zbtb1
Tbx4
Zfp3
Tbx5
Zfp202
Bits

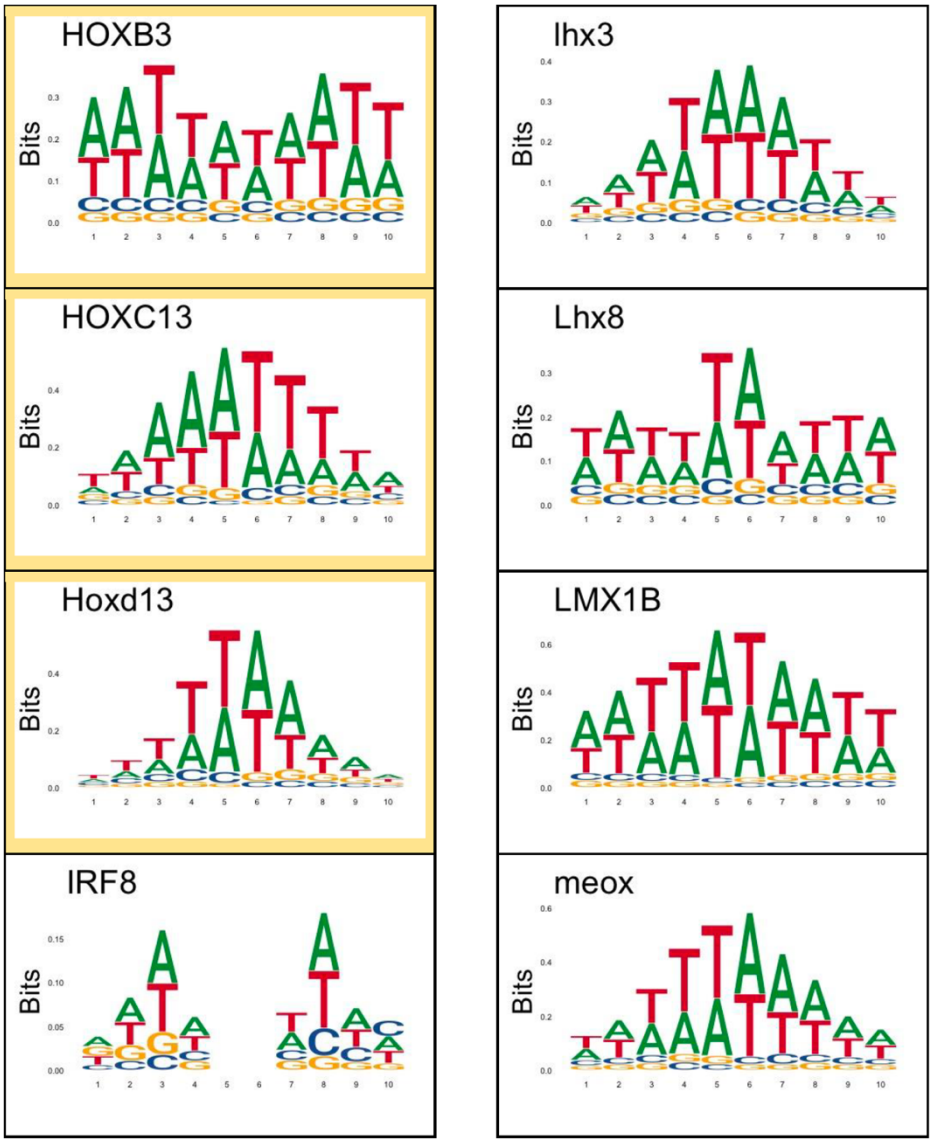

Zfp263
Zkscan1
Zfp300
Zkscan5
Zfp637
Znf740
Zfx
Zscan10
Zic5
Bits

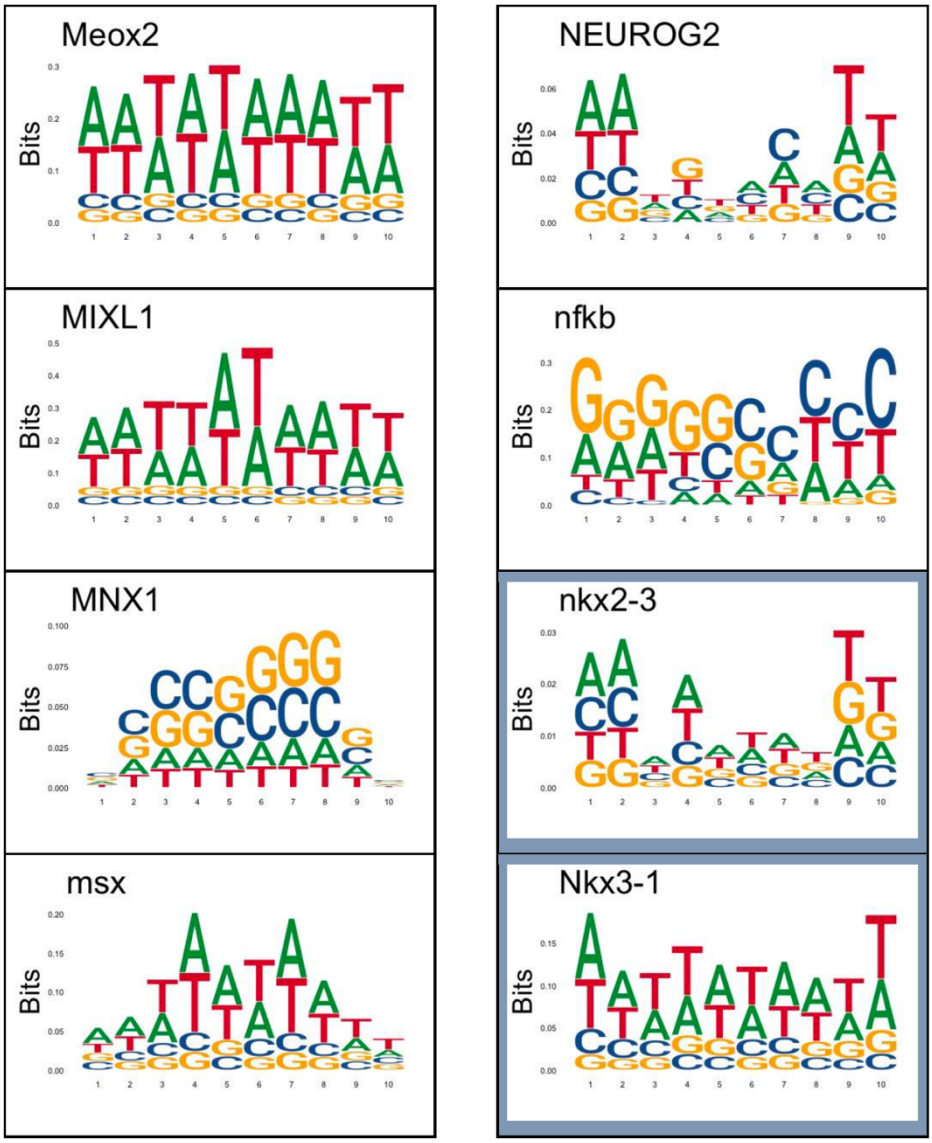

LOGO – HT-SELEX dataset
ALX3
AR
arid3
ARX
BARX1
BHLHE41
bsx
CLOCK
Bits

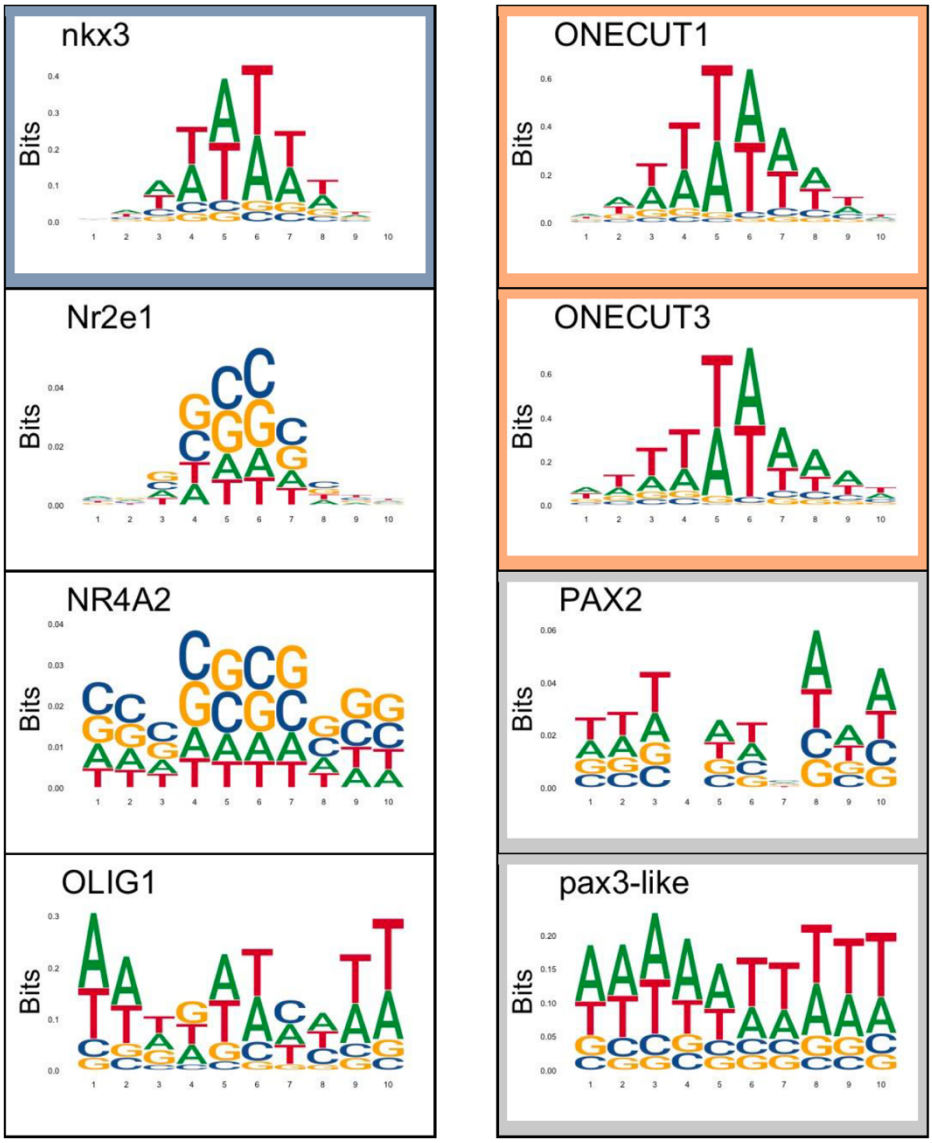

CPEB1
ELK1
E2F3
EMX2
EGR3
EN1
ELF3
ergc
Bits

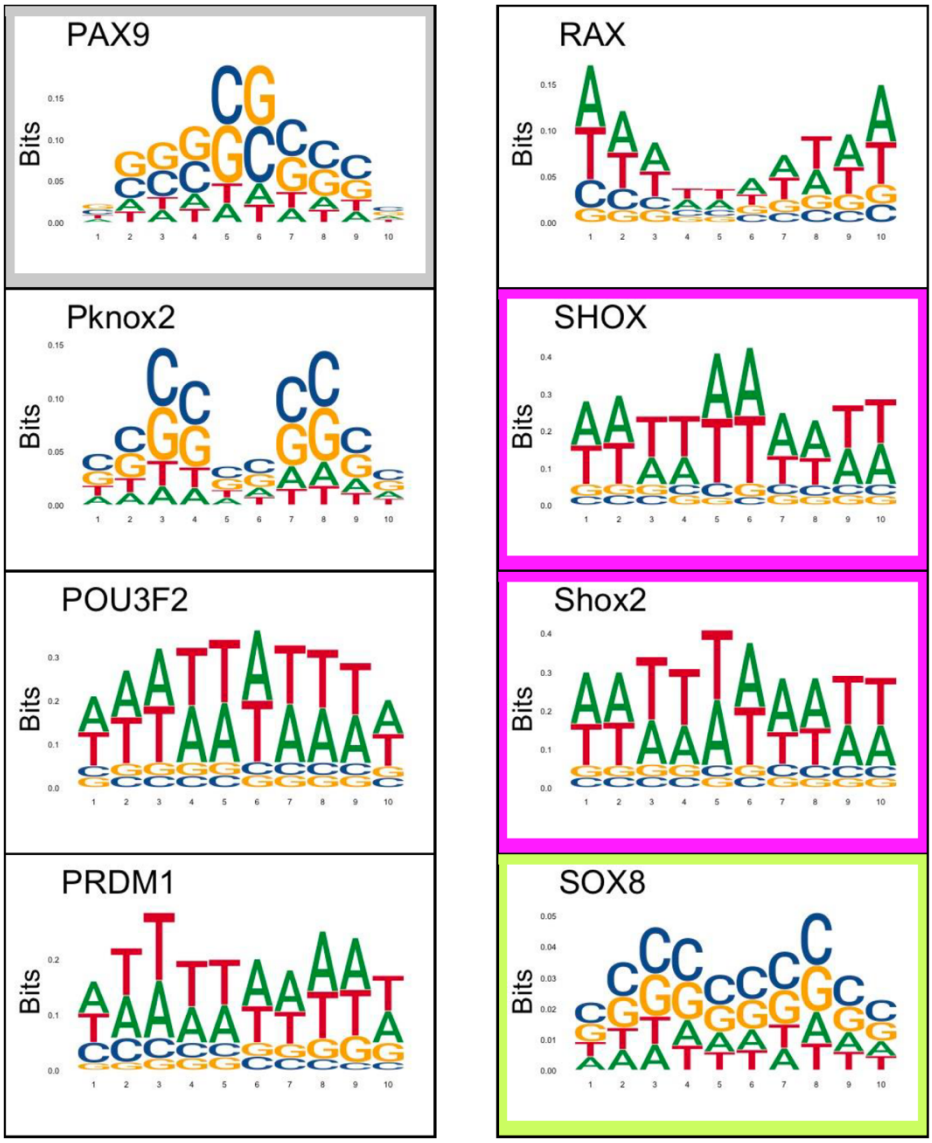

ETV4
GATA5
EVX1
Gbx2
FOXD2
GCM1
FOXD3
GCM2
Bits

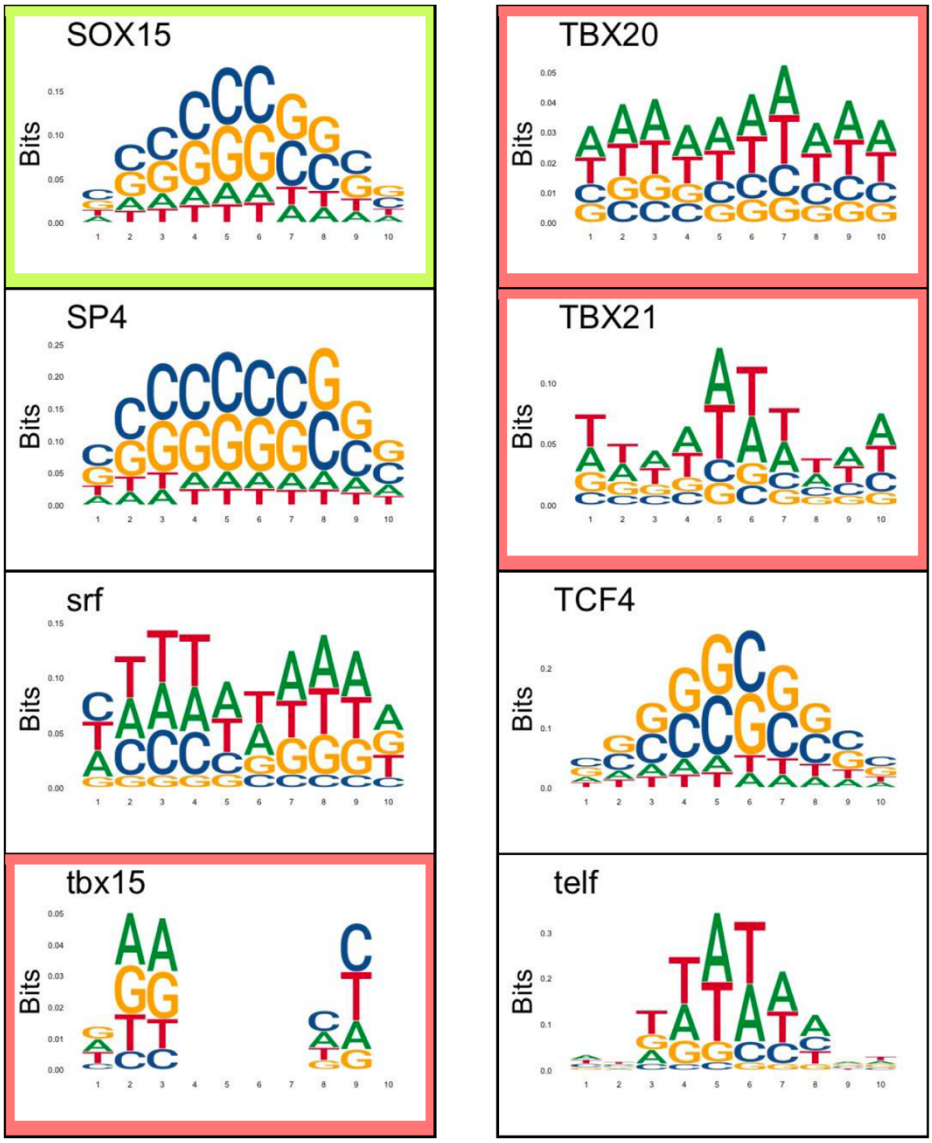

GRHL1
Bits
hox1
Bits
Hlf
Bits
HOXA1
Bits
HNF1B
Bits
HOXA2
Bits
HOMEZ
Bits
HOXA13
Bits

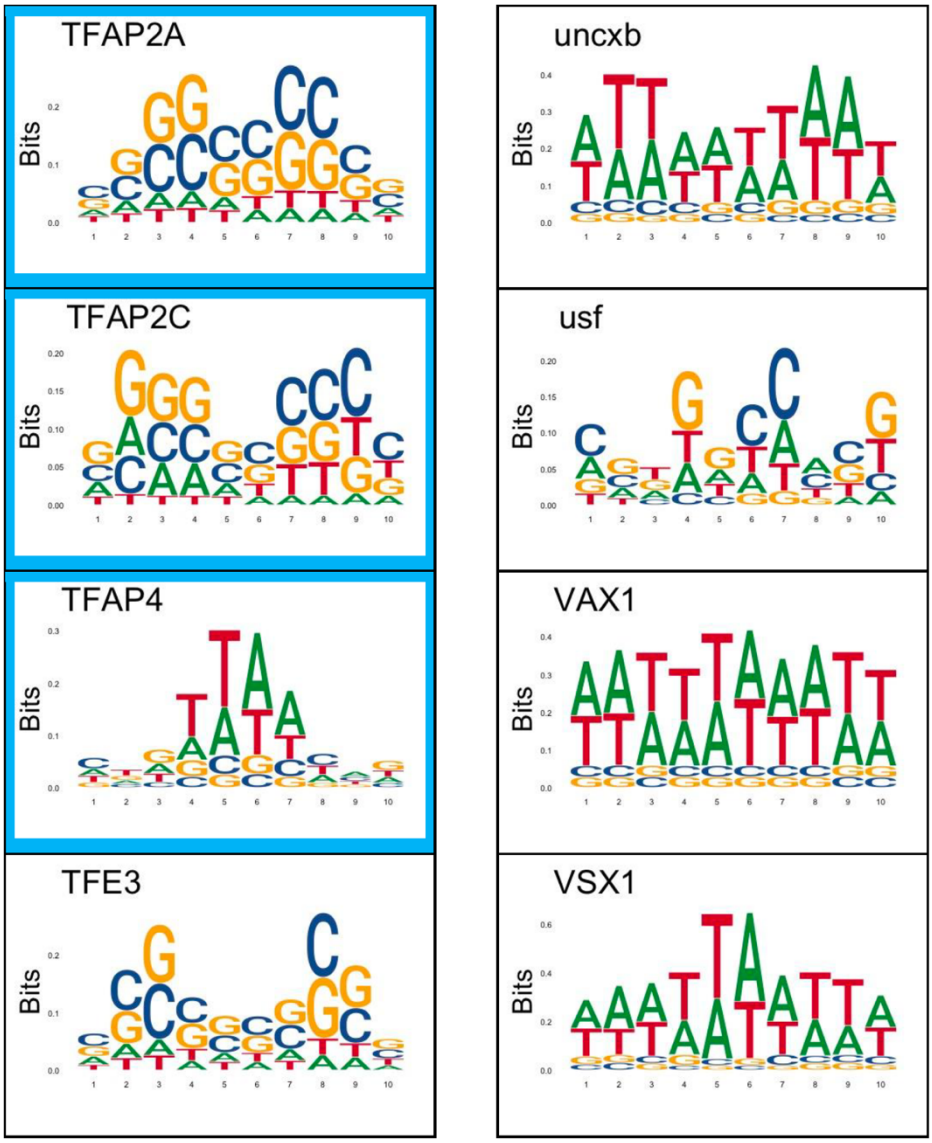

HOXB3
Bits
lhx3
Bits
HOXC13
Bits
Lhx8
Bits
Hoxd13
Bits
LMX1B
Bits
IRF8
Bits
meox
Bits

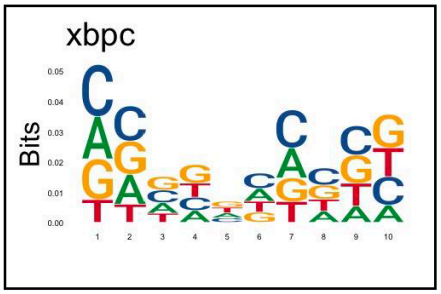

Meox2
Bits
NEUROG2
Bits
MIXL1
Bits
nfkb
Bits
MNX1
Bits
nkx2-3
Bits
msx
Bits
Nkx3-1
Bits

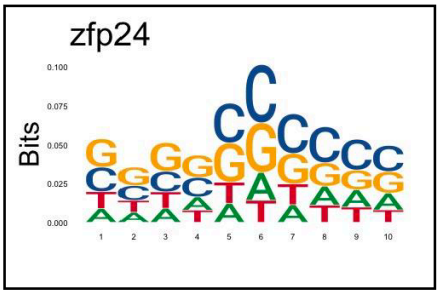

nkx3
Bits
ONECUT1
Bits
Nr2e1
Bits
ONECUT3
Bits
NR4A2
Bits
PAX2
Bits
OLIG1
Bits
pax3-like
Bits


PAX9
Bits
RAX
Bits
Pknox2
Bits
SHOX
Bits
POU3F2
Bits
Shox2
Bits
PRDM1
Bits
SOX8
Bits


SOX15
TBX20
SP4
TBX21
srf
TCF4
tbx15
telf
Bits


TFAP2A
TFAP2C
TFAP4
TFE3
uncxb
usf
VAX1
VSX1
Bits





Supplementary Table S4. MSE calculated for the results obtained using DNAaffinity, CRPTS/CRPT, DNAShapeR and DeepSELEX using the uPMB and HT-SELEX data respectively.

Method	DATA	MSE
DNAffinity	HT-SELEX	0.0005 ± 0.0002
DEEPSELEX	HT-SELEX	0.013 ± 0.008
DNAShapeR	HT-SELEX	0.034 ± 0.029
DNAffinity	uPBM	0.011 ± 0.004
CRPTS/CRPT	uPBM	0.16 ± 0.112
DNAShapeR	uPBM	0.008 ± 0.011

2. An integrated Machine-Learning model to predict nucleosome architecture.

Supplementary Information

AN INTEGRATED MACHINE-LEARNING MODEL TO PREDICT NUCLEOSOME ARCHITECTURE

Alba Sala[1,&], Mireia Labrador[1,&], Diana Buitrago[1], Pau De Jorge[1], Federica Battistini[1,2], Isabelle Brun Heath[1] and Modesto Orozco[1,2*]

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain

[2] Departament de Bioquímica i Biomedicina, Universitat de Barcelona, Barcelona, Spain

[&] These authors contributed equally to this work: Alba Sala, Mireia Labrador.

[*] Correspondence to M.Orozco: modesto.orozco@irbbarcelona.org

Name	Primer sequence (5'-3')	Amplicon size (bp)
UBX5_F	GACGACGACGAATATGAG	163
UBX5_R	CGAGTTTGGACATGATTG	
CKB2_F	GAGAATATGACACATGCC	130
CKB2_R	CCGCCTCTTTGTACTTAG	
PPT1_F	CAATGATCCGGCTGCTAC	171
PPT1_R	GGACCTTCATAATTGGCT	
TRP4_F	GTAGGTACTGGTGGTGAC	117
TRP4_R	GGATGTAGAAGCTTTACC	
BSP1_F	GGAAGAGCCGATATACCC	147
BSP1_R	CGGACTTTCTGTAACTTC	
DGK1_F	CCATTGCCCTTCCAAATA	182
DGK1_R	GCCGAACCATTCATGAGA	
SLM3_F	GATTGGAGAGATGTGAAC	123
SLM3_R	CGACCCTTCACTGTAGCC	
PAN5_F	GGGTACCGTTTTGGCAGT	178
PAN5_R	GCATTCGCATTTCTCCAC	
ALG9_F	CACGGATAGTGGCTTTGGT	149
ALG9_R	CAGCAGGAAAGAACTTGGG	
ACT1_F	GGTTGCTGCTTTGGTTATTGATAAC	271
ACT1_R	CAATTCGTTGTAGAAGGTATGATGCC	
RPA135_F	GAACAATGGCGAGGAGAAC	141
RPA135_R	CCACCTATTTCATCGGATTC	
NMD3_F	GGGTTGGATTTCTTCTATGC	164
NMD3_R	GGAACAATCTCGACAGAATATG	

Suppl. Table S1: Sequences of primers used in the gene expression analysis by qPCR, and the amplicon size (in base pairs) obtained using each forward (F) and reverse (R) pair of oligonucleotides.

	All genes	Our genes	W-open-W	W-close-W	Tandem (TTS)	Convergent	Highly Expressed	Lowly Expressed
Genes (TSS)	7071	5676	2749	644	-	-	713	669
Genes (TTS)			1134	942	2256	2496		

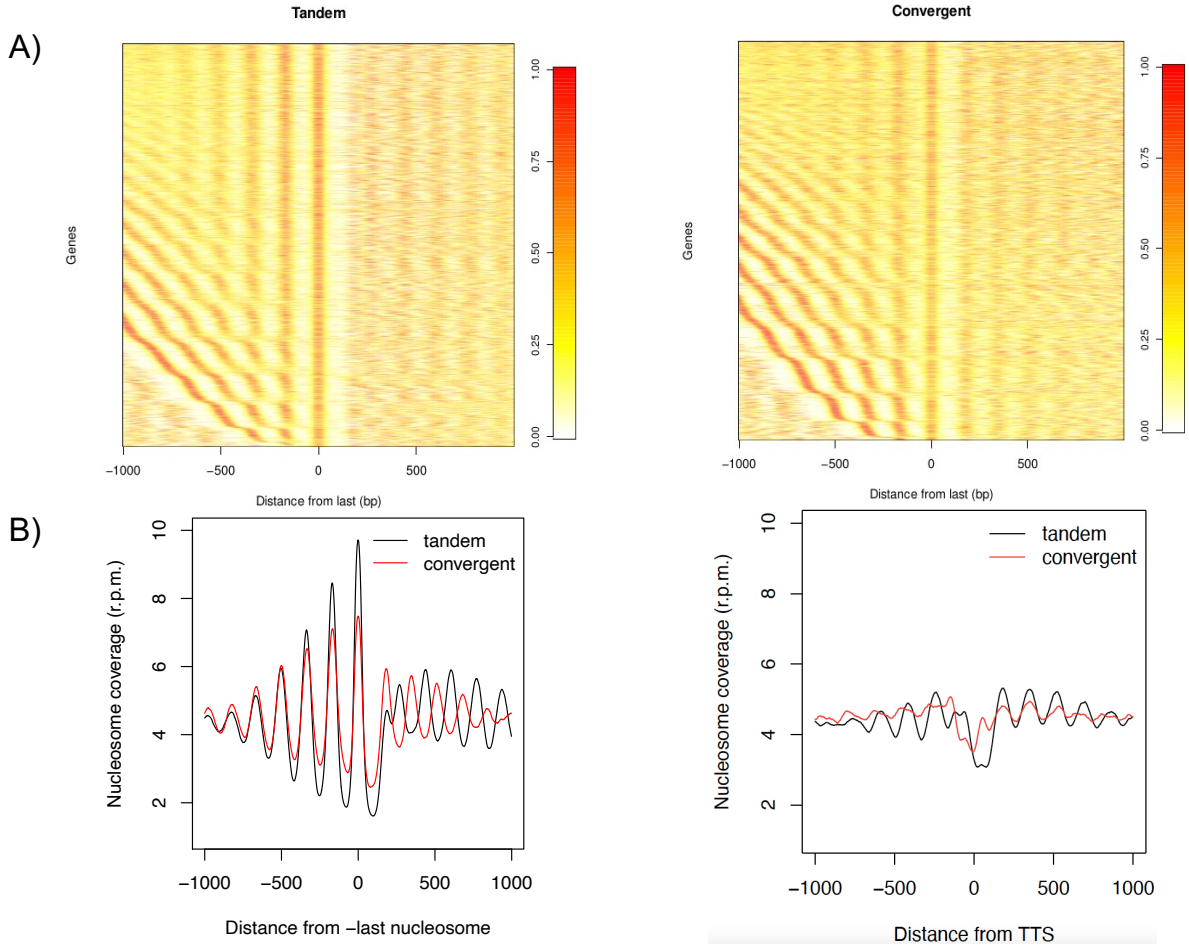
Suppl. Table S2: Number of genes corresponding to each analyzed category

Sample	Ctrl Mean	Ph Mean	P value from t.test
Strain 1	0.741	0.711	< 2.2e-16
Strain 2	0.743	0.722	< 2.2e-16
Strain 3	0.737	0.719	4.201e-11
Strain 4	0.744	0.732	2.846e-06

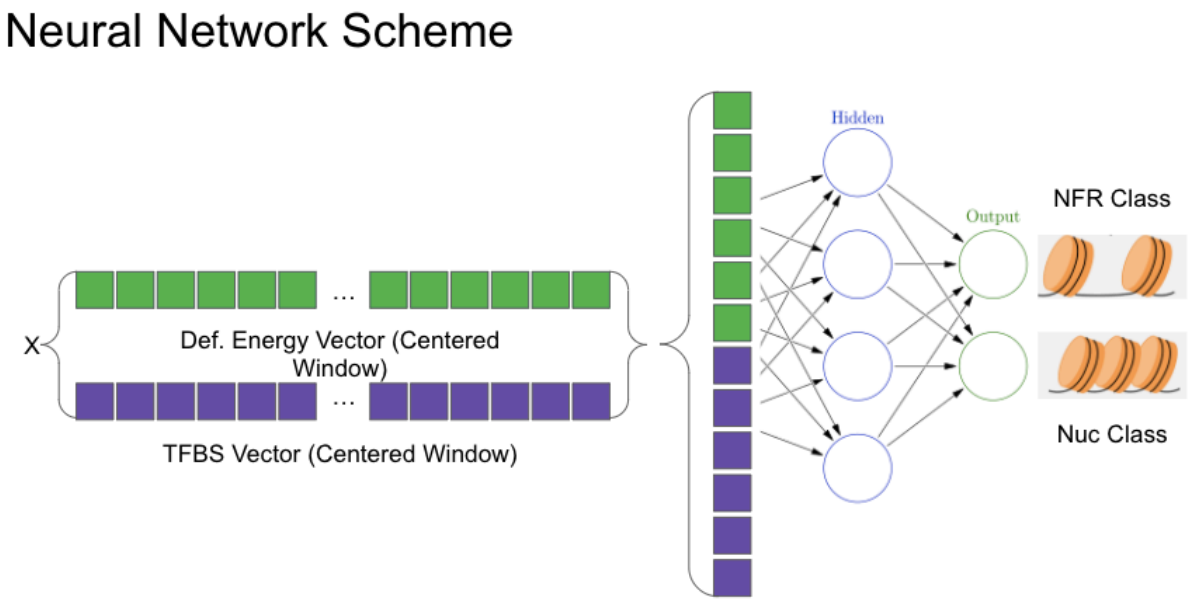
Suppl. Table S3: Autocorrelation (R) in the 4 control strains and 4 strains treated with phenanthroline

Strain	Sample name in GEO	Genes modified with the 81-nt insertion	Treatment	Data
Strain 1	GSM8081755	UBX5 and BSP1	Arrested in G1 with a factor	Mnase-seq
Strain 2	GSM8081756	CKB2 and DGK1	Arrested in G1 with a factor	Mnase-seq
Strain 3	GSM8081757	PPT1 and SLM3	Arrested in G1 with a factor	Mnase-seq
Strain 4	GSM8081758	TRP4 and PAN5	Arrested in G1 with a factor	Mnase-seq
Strain 1	GSM8081759	UBX5 and BSP1	Arrested in G1 with a factor + 30 min incubation with 1,10-phenanthroline	Mnase-seq
Strain 2	GSM8081760	CKB2 and DGK1	Arrested in G1 with a factor + 30 min incubation with 1,10-phenanthroline	Mnase-seq
Strain 3	GSM8081761	PPT1 and SLM3	Arrested in G1 with a factor + 30 min incubation with 1,10-phenanthroline	Mnase-seq
Strain 4	GSM8081762	TRP4 and PAN5	Arrested in G1 with a factor + 30 min incubation with 1,10-phenanthroline	Mnase-seq

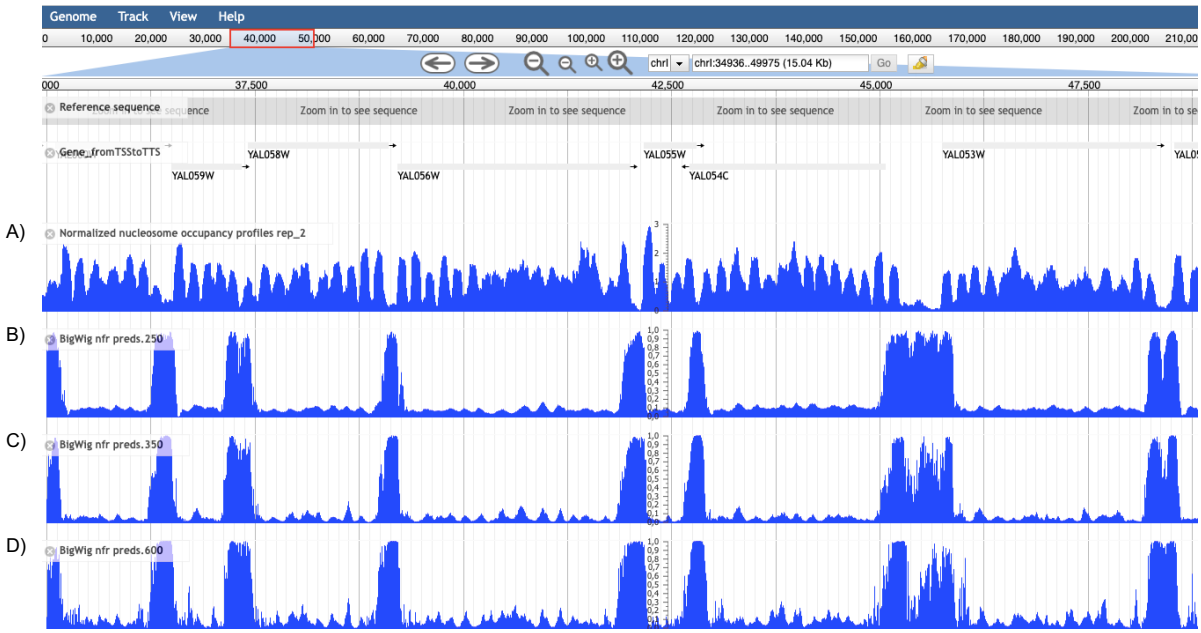
Suppl. Table S4: Generated datasets.



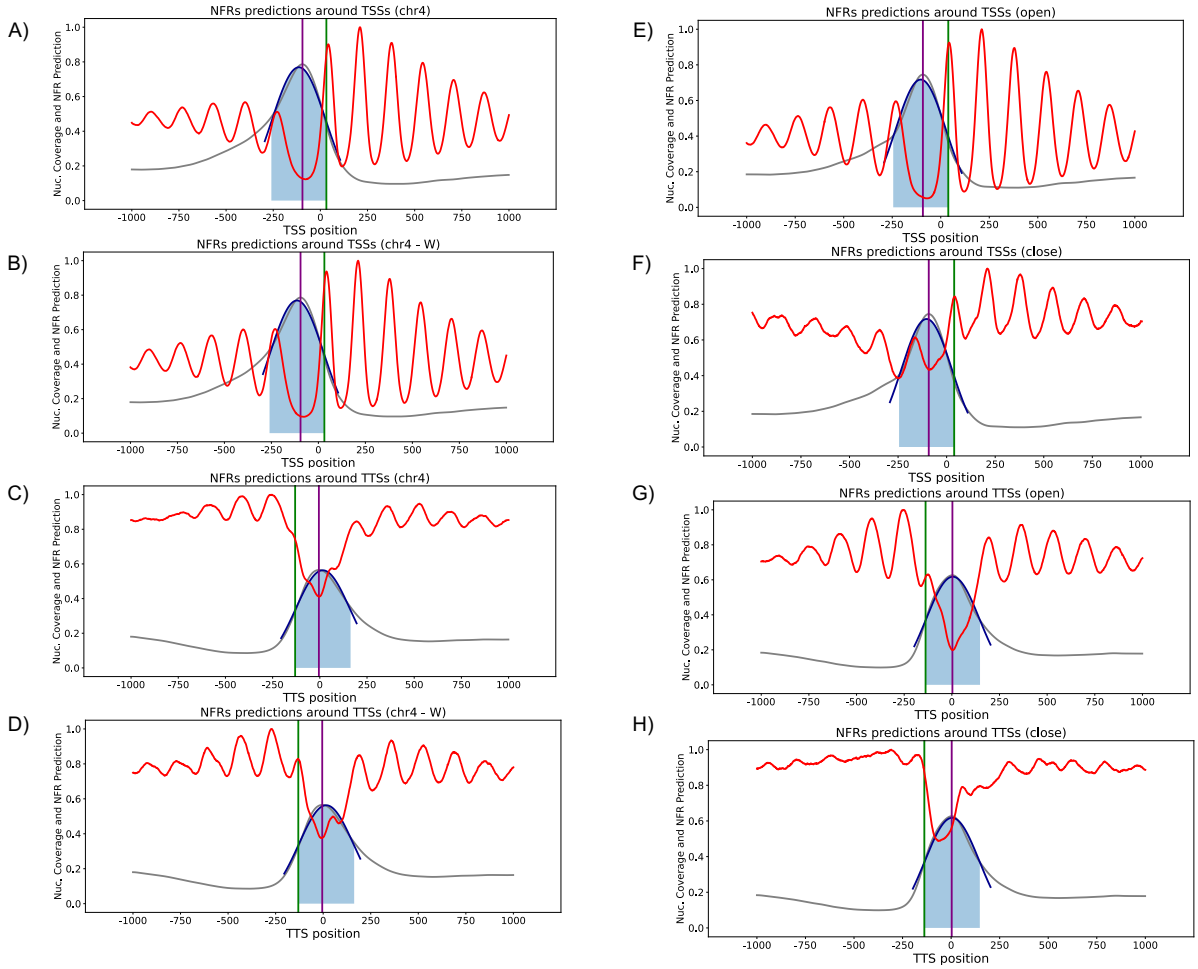
Suppl. Figure S1. A) Nucleosome coverage of tandem and convergent genes shown as a heatmap, centered at –last nucleosome. Genes are sorted by the distance between +1 and –last nucleosomes. B) Averaged nucleosome coverage among all genes of tandem (2256 genes) and convergent (2496 genes) genes centered at the –last nucleosome (left panel) and at the TTS (right panel). Separate curves are shown according to the orientation of the downstream gene.



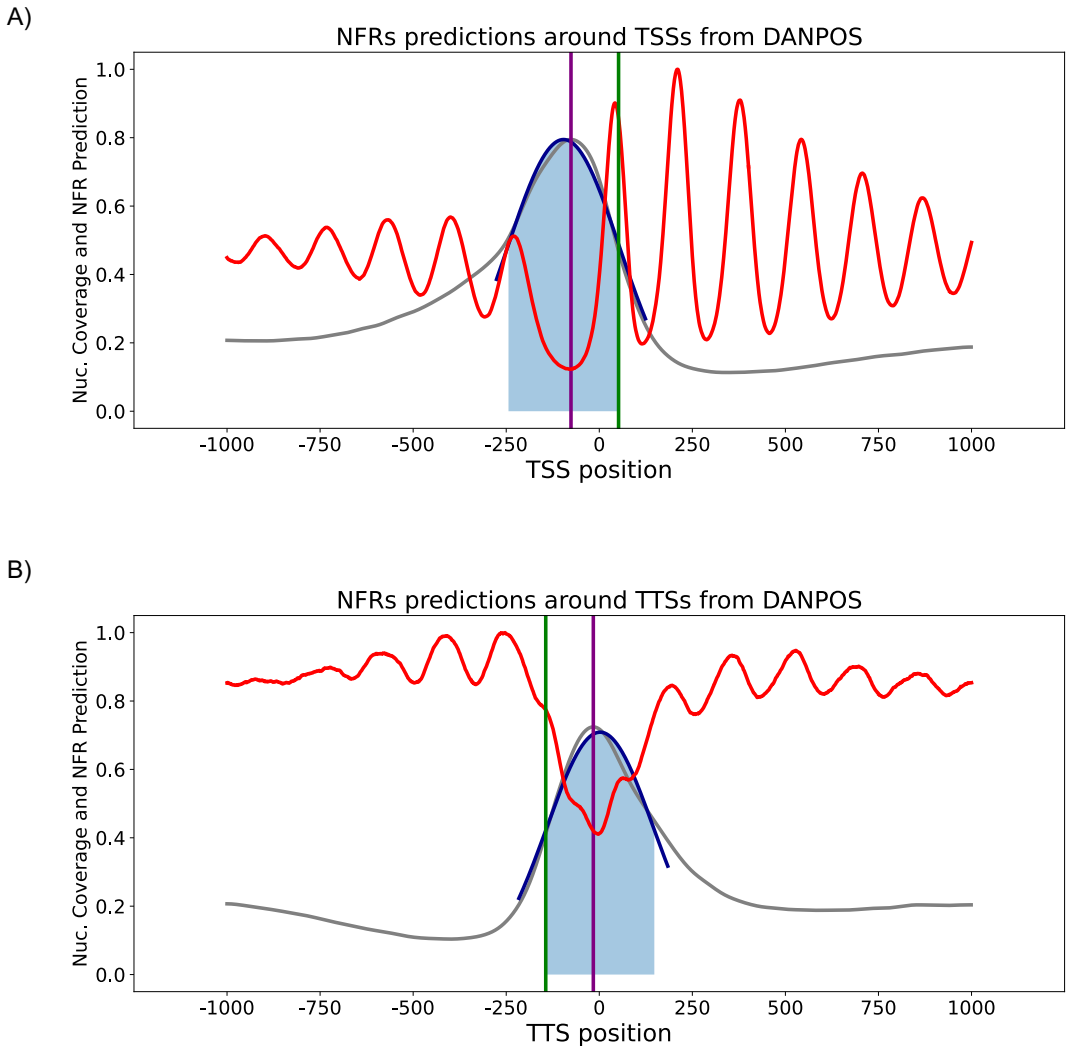
Suppl. Figure S2. Scheme of our Neural Network Classifier composed of three layers: the input which consists of two stacked vectors of size 350, a hidden layer defined by 30 neurons, and an output layer.



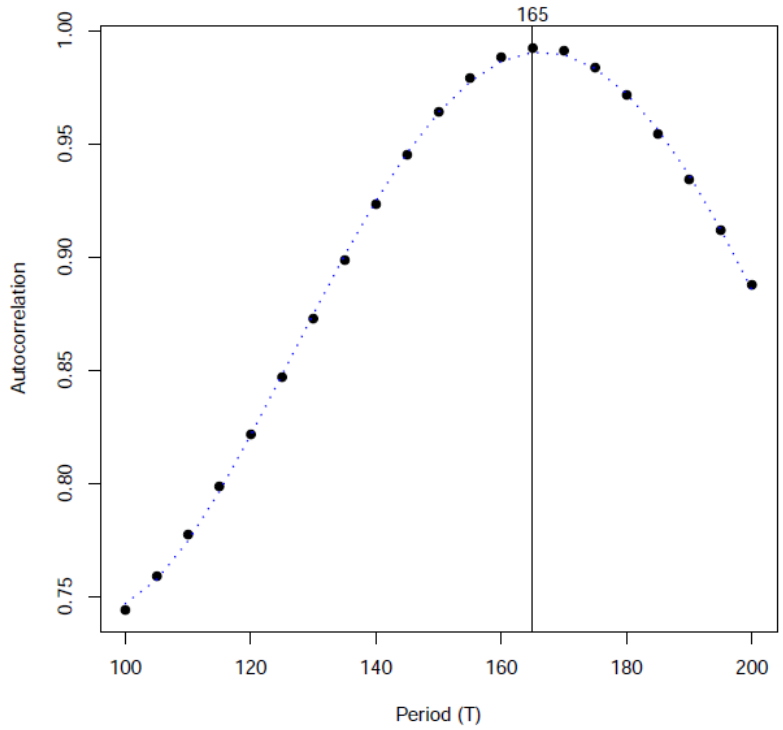
Suppl. Figure S3. Experimental nucleosome coverage (panel A) against the NFR probability predictions across a 15kb genomic region for different centered windows (B-D panels using 250bp, 350bp and 600bp windows respectively)



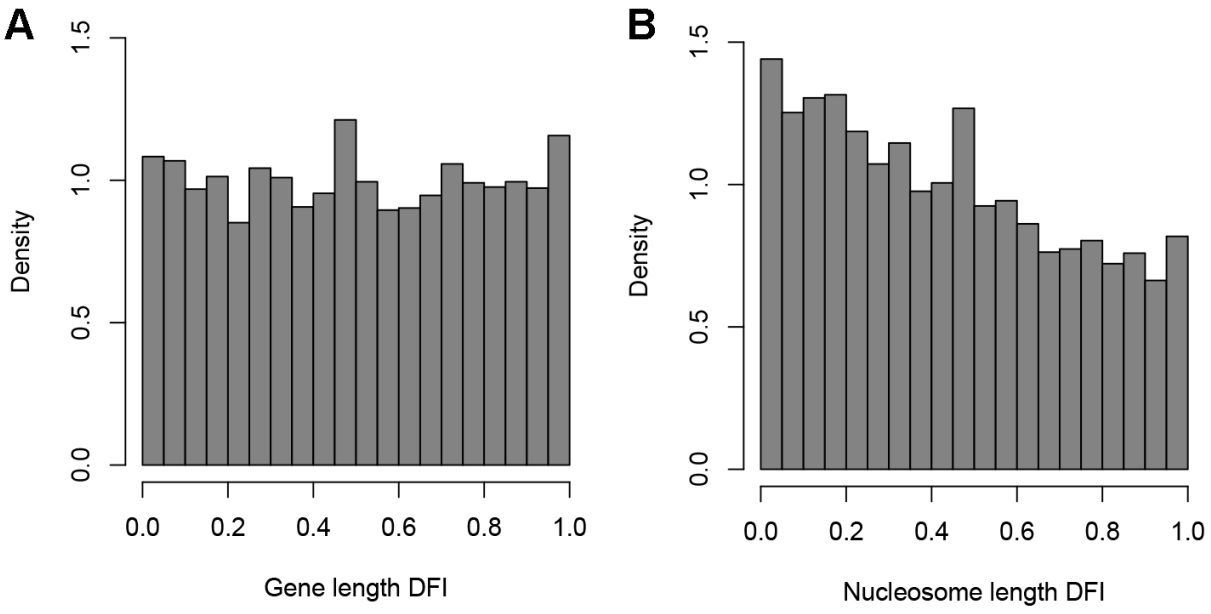
Suppl. Figure S4. NFR prediction (grey) against nucleosome experimental coverage (red) in our chrIV model for A) all TSSs, B) well positioned TSSs, C) all TTSs and D) well positioned TTSs. Green lines denote the average prediction of the +1 and -last nucleosomes 2stds from a fitted Gaussian distribution (dark blue). Similar plots are shown for our fully trained model for E) open TSSs, F) closed TSSs, G) open TTSs and H) closed TTSs.



Suppl. Figure S5. NFR prediction (grey) against nucleosome experimental coverage (red) in our DANPOS trained model for A) all TSSs, B) all TTSs. Green lines denote the average prediction of the +1 and -last nucleosomes 2stds from a fitted Gaussian distribution (dark blue).

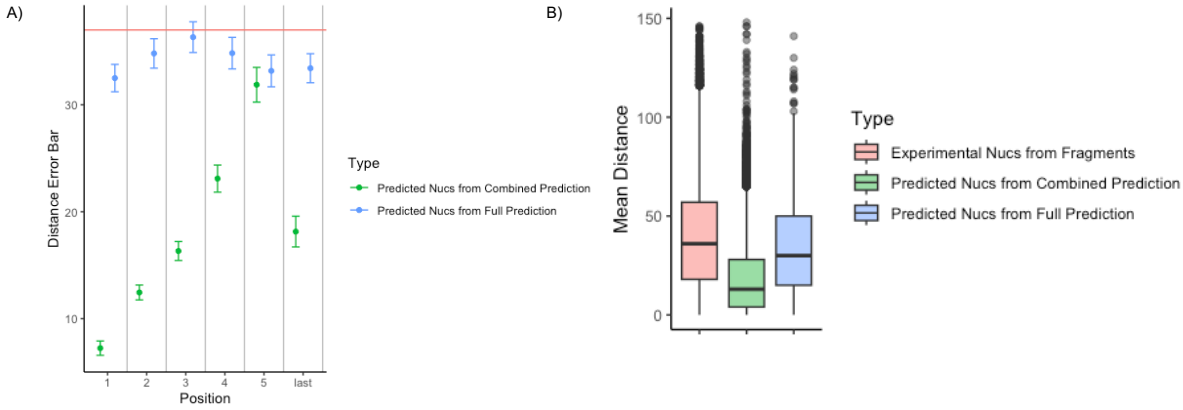


Suppl. Figure S6. Autocorrelation coefficient for the nucleosome coverage signal in different potential periods.

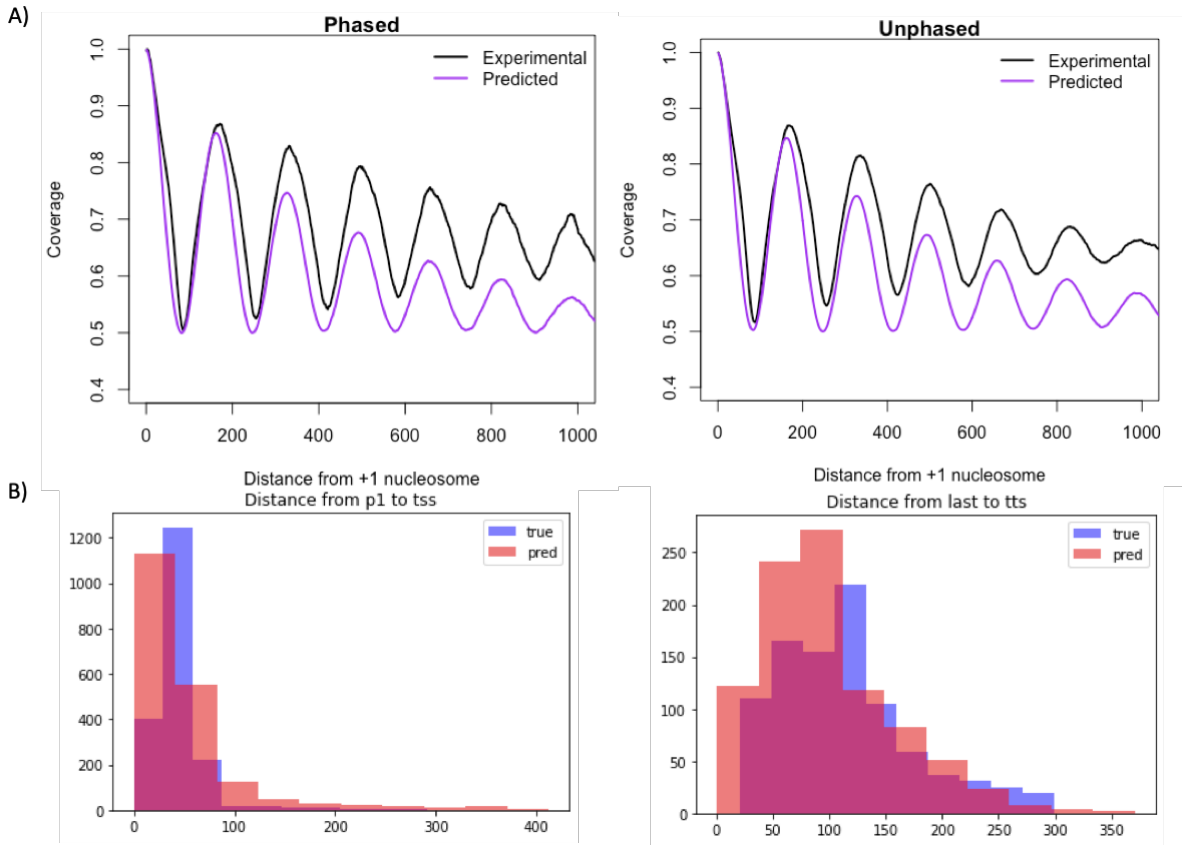


Suppl. Figure S7. Distance from integer (DFI, see Methods) score computed on (A) the gene length (distance between TSS and TTS) and (B) the nucleosome length (distance

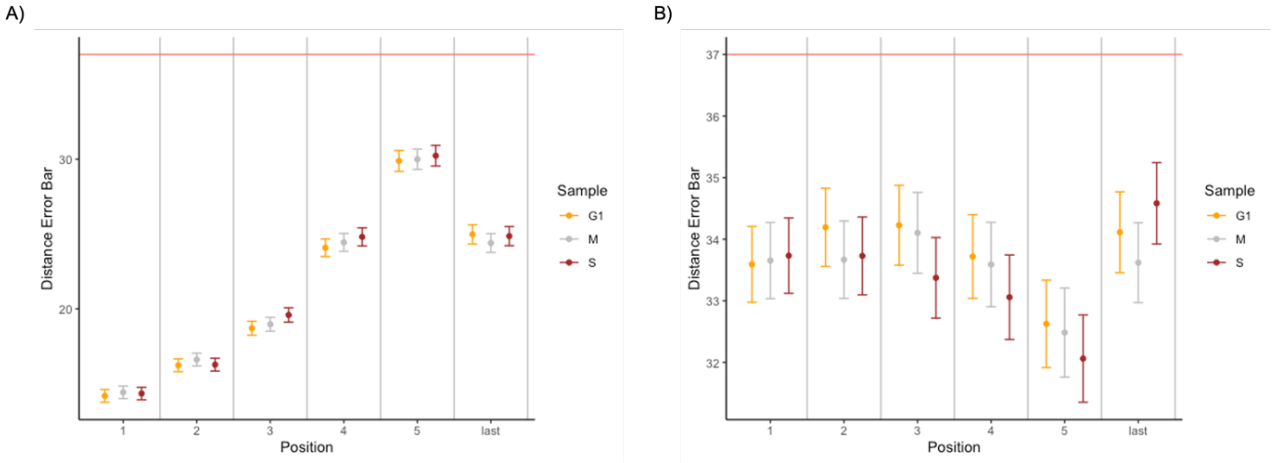
between +1 and –last nucleosome). DFI is normalized (between 0 and 1) dividing by T/2 and taking the absolute value.



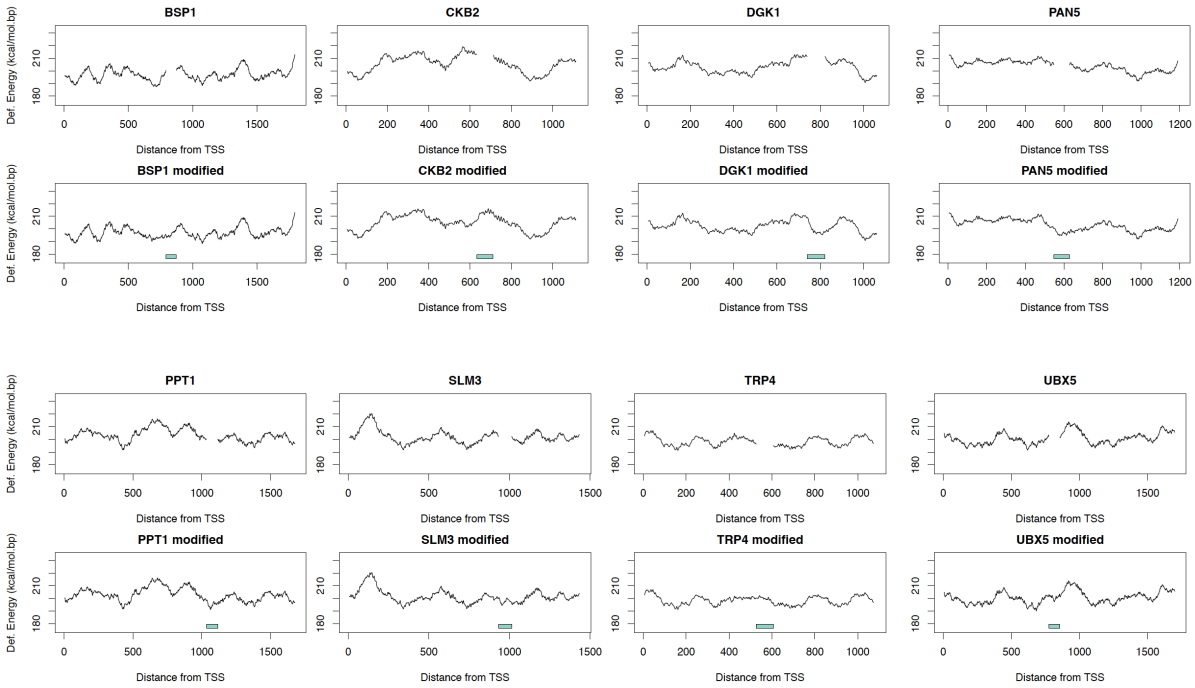
Suppl. Figure S8. A) Distance between the results of the combined prediction (green error bars) and the fully predicted (blue error bars) peaks through different position calls in comparison to the experimental mean (red line) for our starting testing data. B) Box plot distribution of all distances between the two scenarios (combined prediction in green, fully predicted in blue) and the experimental coverage peaks. Experimental fragments variability is shown in red.



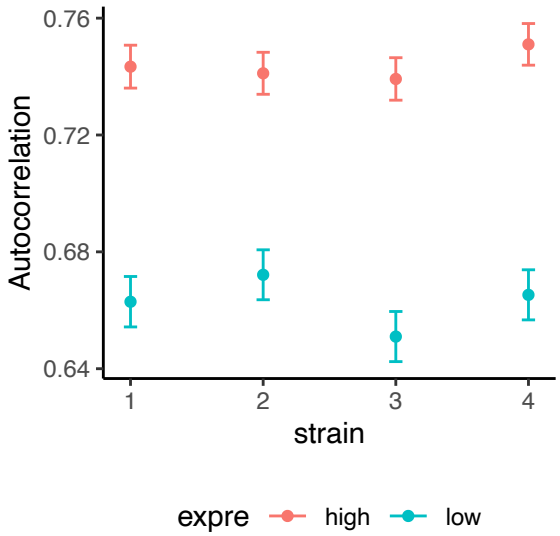
Suppl. Figure S9. A) Nucleosome coverage, experimental (black) and predicted (purple, see Methods) from +1 nucleosome, averaged across all genes. Genes are split into phased (left panel) or unphased (right panel) based on DFI<10 and DFI>40, respectively. B) Distances from predicted positions of the +1 and –last nucleosomes to the TSS (left panel) and TTS (right panel) respectively, of the predicted (red bars) and experimental (blue bars) positions.



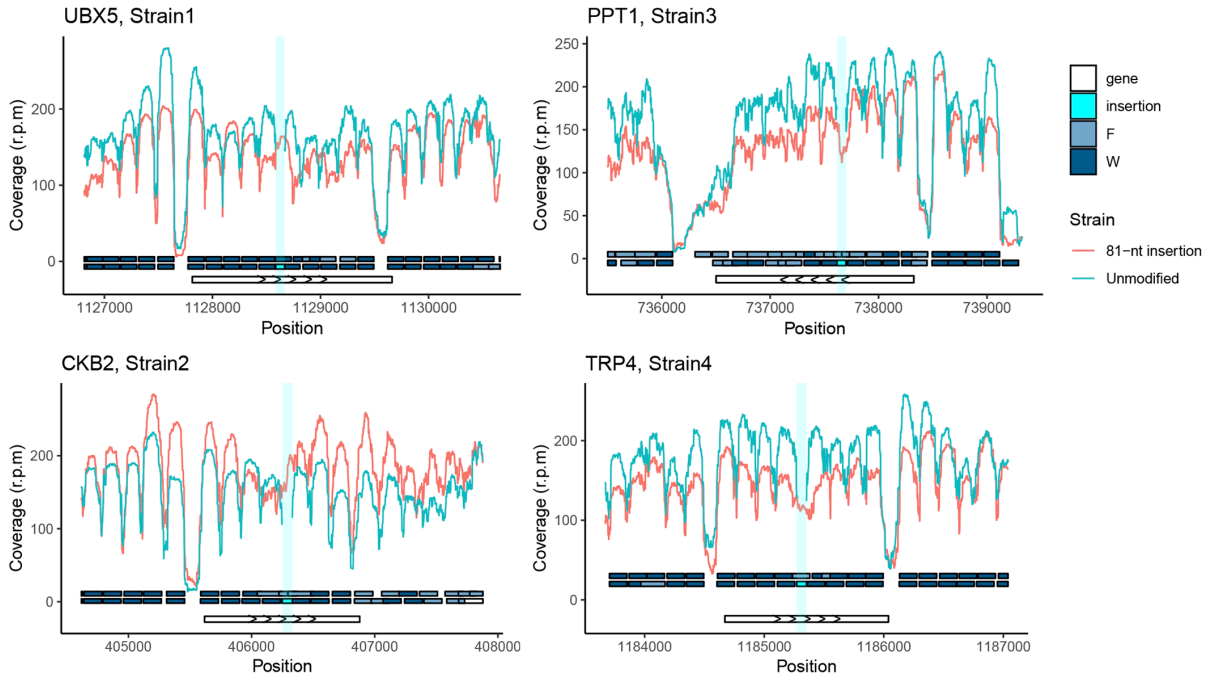
Suppl. Figure S10. Distance between experimental and predicted nucleosome peaks for G1 (orange error bars), M (grey error bars) and S (brown error bars) phase, using combined (panel A) and full (panel B) prediction methods. Red line represents the experimental mean variability.



Suppl. Figure S11. Deformation energy for the 8 selected genes without (top panels), or with (lower panels) the 81-nt insert (represented as a box coloured in cyan).

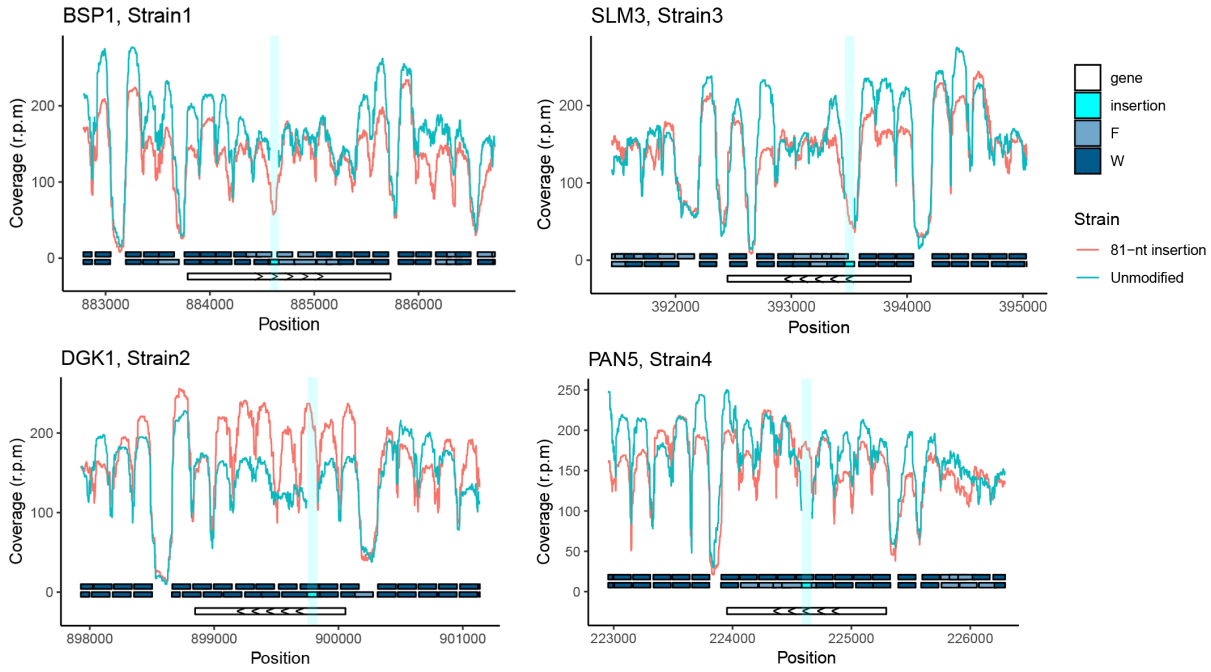


Suppl. Figure S12. Autocorrelation scores for highly (red) and lowly (blue) expressed genes derived from our four mutant strains (see Methods).

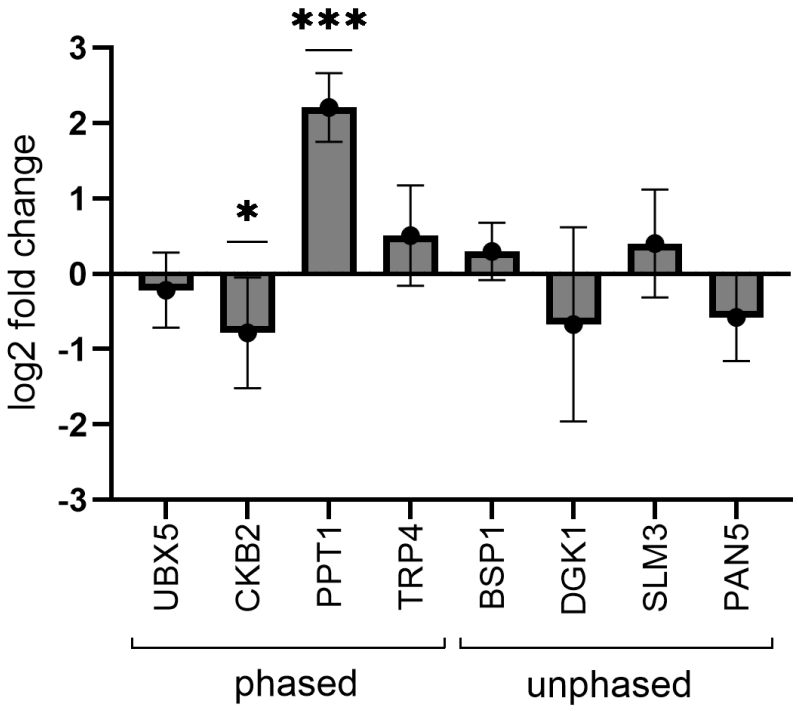


Suppl. Figure S13. Nucleosome coverage of the four phased genes in the unmodified strain (blue) and their nucleosome coverage in the strain with the 81-nt insertion (red). Nucleosome positions in the original and modified strains are shown colored by their

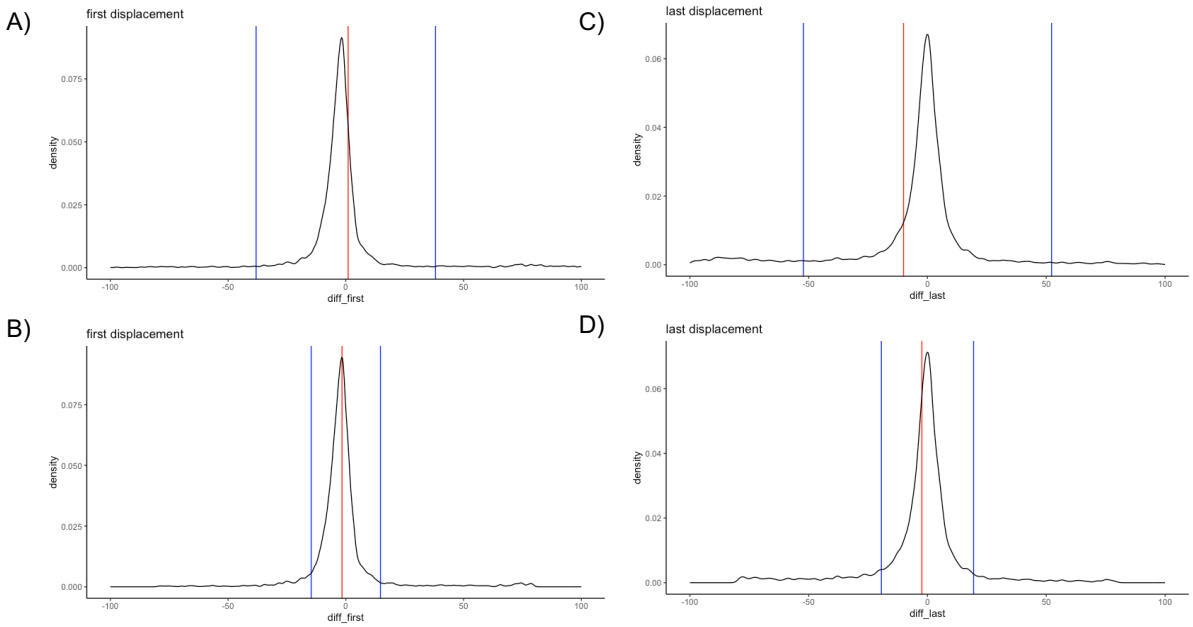
class: Well-positioned (W, dark blue boxes) and Fuzzy (F, light blue boxes). The gene body from TSS to TTS is depicted as a white box.



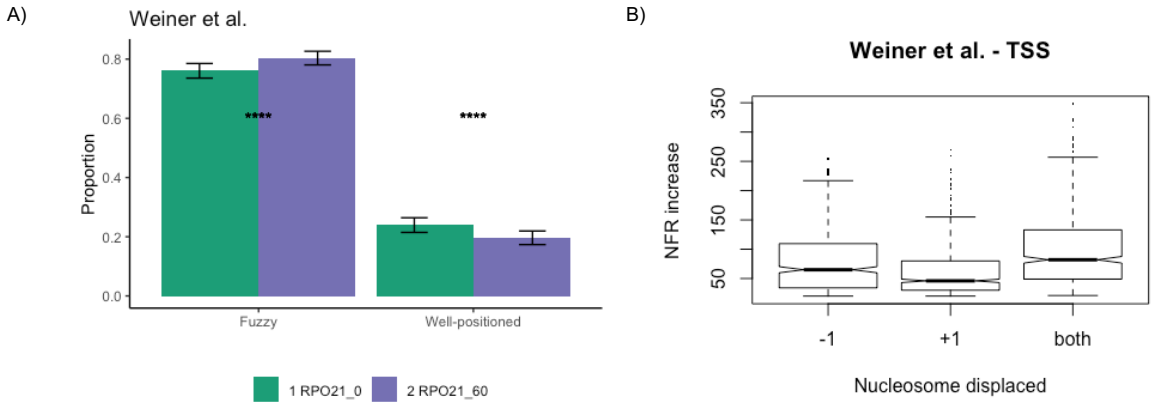
Suppl. Figure S14. Nucleosome coverage of the four not-phased genes in the unmodified strain (blue) and their nucleosome coverage in the strain with the 81-nt insertion (red). Nucleosome positions in the original and modified strains are shown colored by their class: Well-positioned (W, dark blue boxes) and Fuzzy (F, light blue boxes). The gene body from TSS to TTS is depicted as a white box.



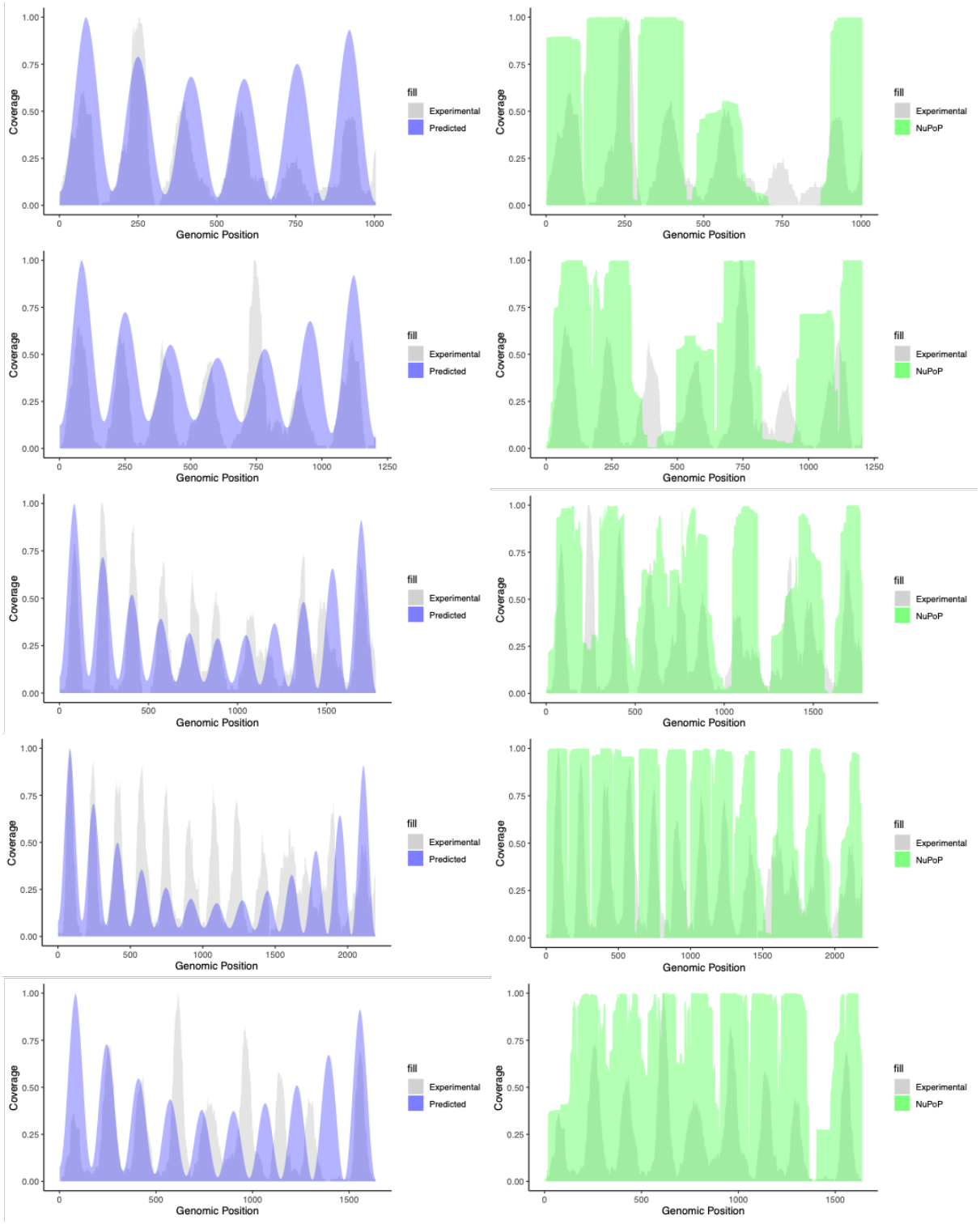
Suppl. Figure S15. Gene Expression fold change (qPCR) for each gene with or without the 81-nt sequence insertion for phased (green bars) and not-phased (red bars) genes.

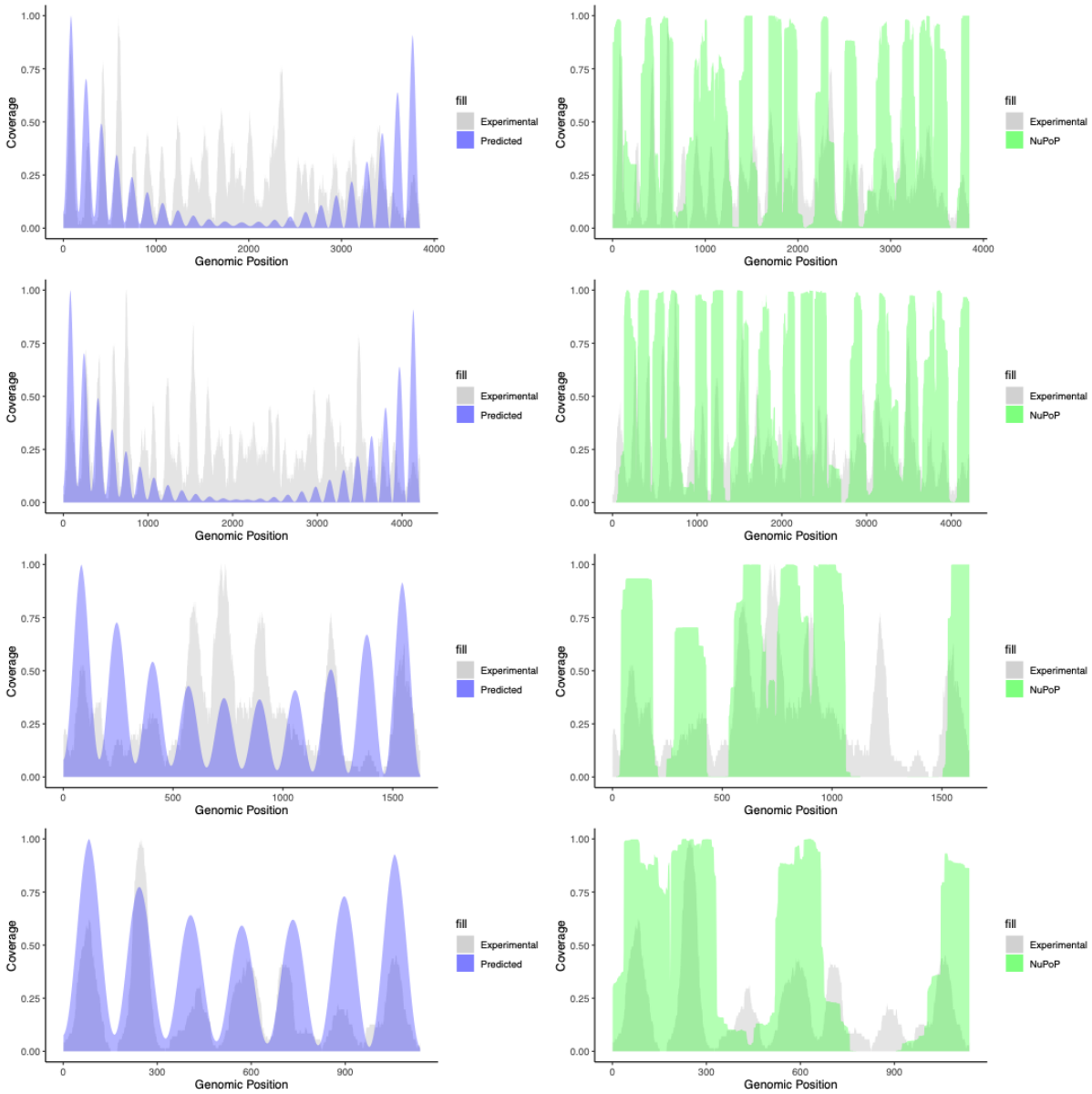


Suppl. Figure S16. Effect of phenanthroline on position of the +1 (A-B panels) and –last (C-D panels) nucleosome. Red lines denote the mean displacement, and blue lines denote the standard deviations from 0. Analysis performed on all the genes (A and C) or on the genes for which the displacement of the +1 or –last was inferior of 81bp (B-D).

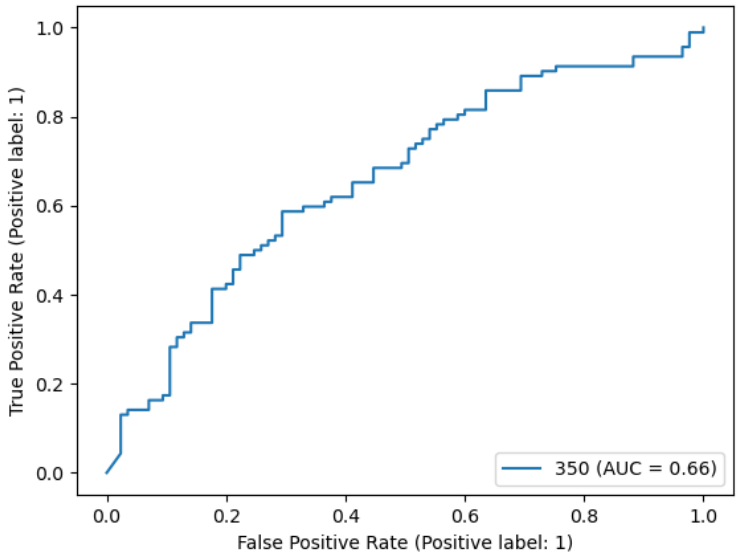


Suppl. Figure S17. Effect of transcription on nucleosome positioning from Weiner et al. (49). A) Change in the proportion of Fuzzy and Well-positioned nucleosomes upon transcription inhibition, with bars indicating relative standard error. B) Change in NFRs' width at the TSS (-1 to +1 nucleosome distance) upon transcription inhibition using a ts allele (only cases with significant displacements (> 20 bps) are considered in the box plots).


Coverage
Genomic Position
fill
Experimental
Predicted
NuPoP

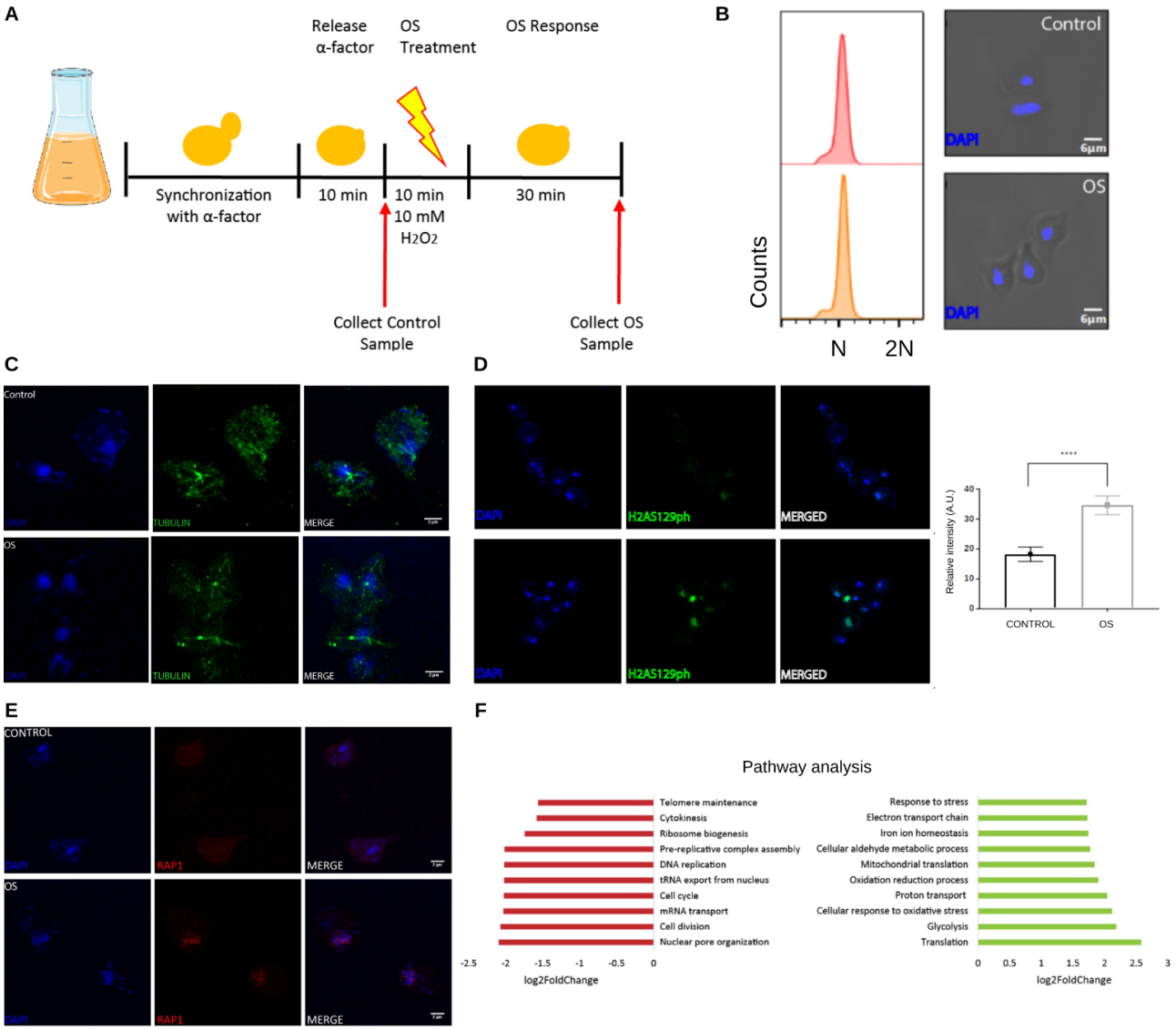


Suppl. Figure S18. Comparison between our method, full predictor (blue), and NuPoP (green) in the prediction of the nucleosome positioning for some genes along the genome respect to the experimental profile (grey).

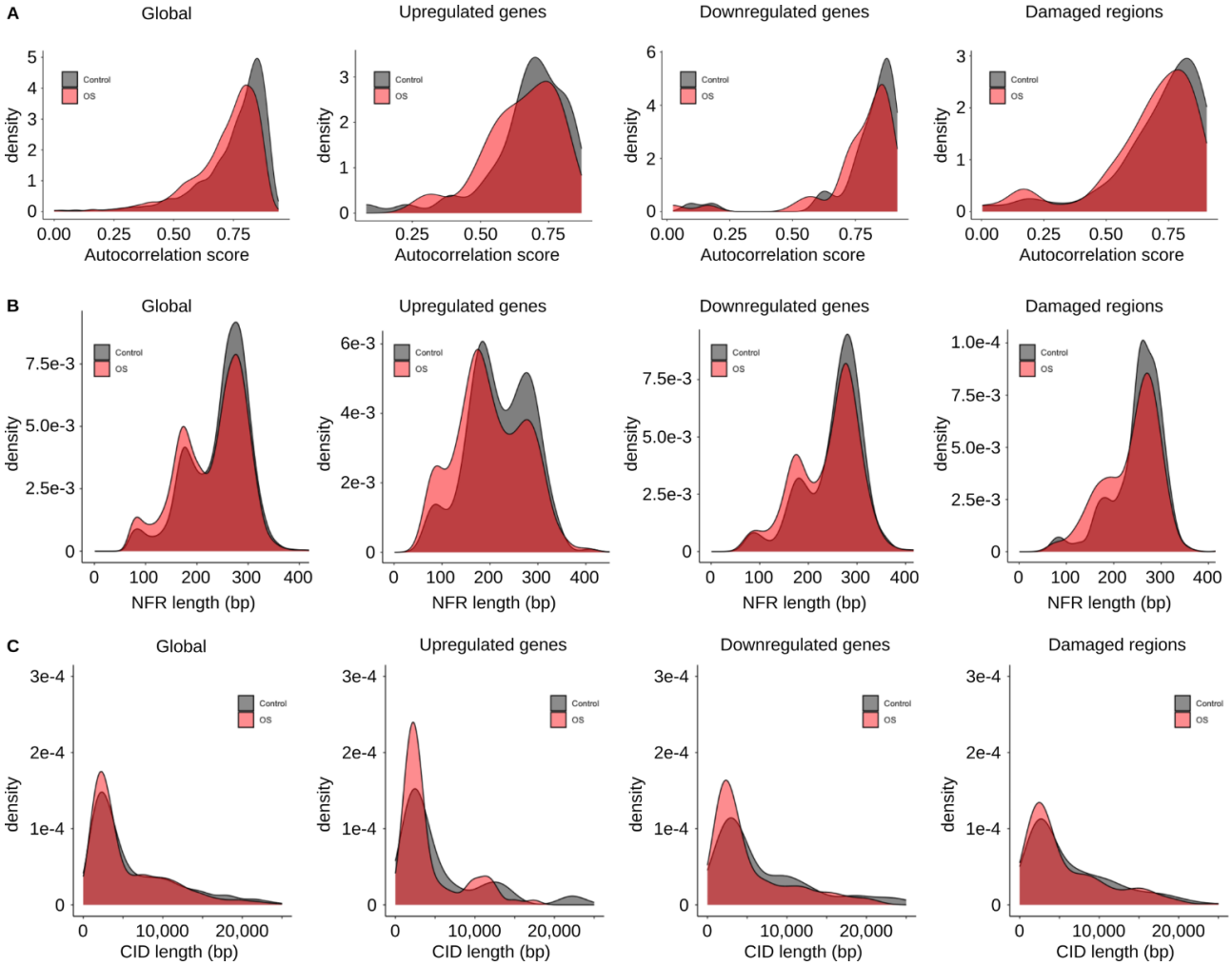


Suppl. Figure S19. Panel showing the receiver operating characteristic (ROC) curve results from our model trained on human nucleosome data for a 350bp window.

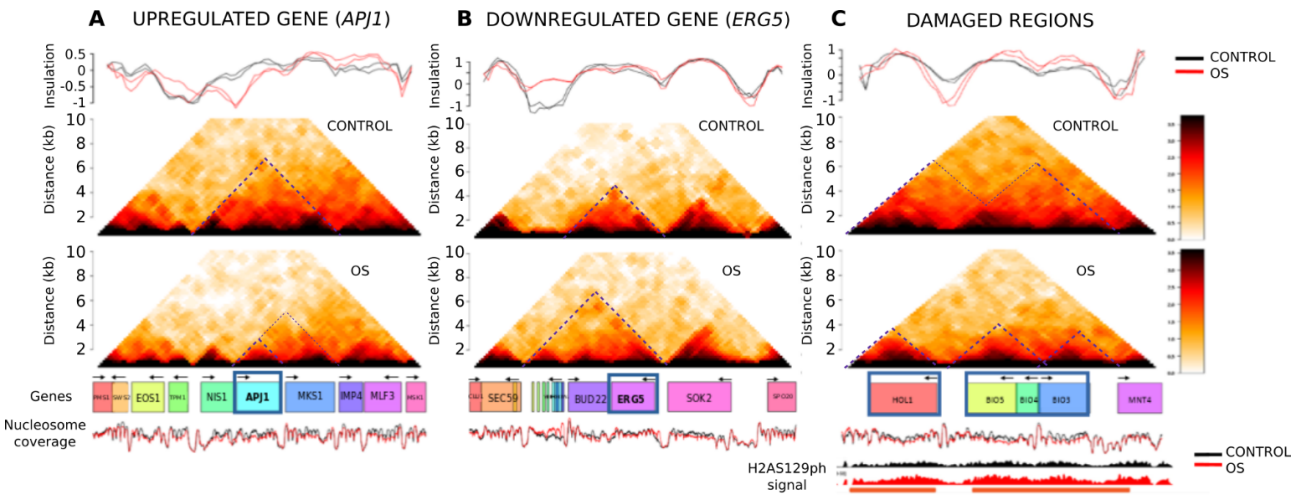
3. Effect of oxidative stress on 3D genome structure



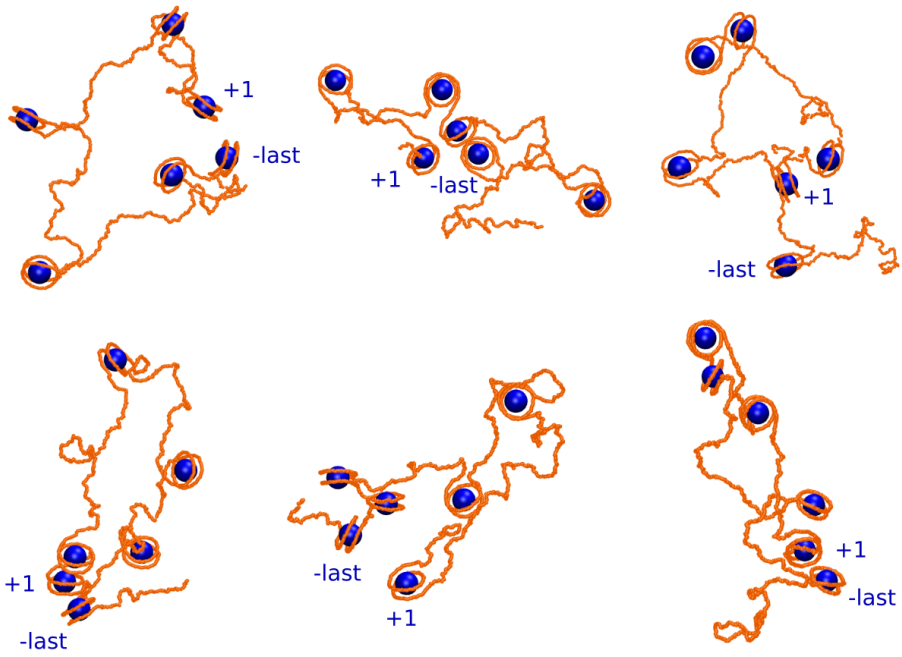
Supporting Figure S1. Experimental protocol, characterisation of cells and cellular response to oxidative stress. (A) Scheme of the experimental set up to analyse the effect of oxidative stress (OS) on 3D genome structure. (B) Fluorescence intensity after flow cytometry and fluorescence microscopy of control (top panels) and OS (bottom panels) cell samples. Nuclei were stained with DAPI. (C) Confocal microscopy imaging of control (top panels) and OS (bottom panels) cells labeled with DAPI (blue) and β-tubulin (green). (D) Confocal microscopy images showing H2AS129ph (green) an nucleus (blue) in control (top panels) and oxidative stressed (bottom panels) cells. Intensity of H2AS129ph was quantified in 132 cells per sample (mean +/- standard deviation is plotted, t-Student p = 0.00001). (E) Structured Illumination Microscopy (SIM) images showing RAP1 (red) and nucleus (blue) in control (top panels) and oxidative stressed (bottom panels) cells. (F) Gene set enrichment analysis under OS response. The ten most representative upregulated (green) and downregulated (red) pathways are shown (FDR < 0.05).



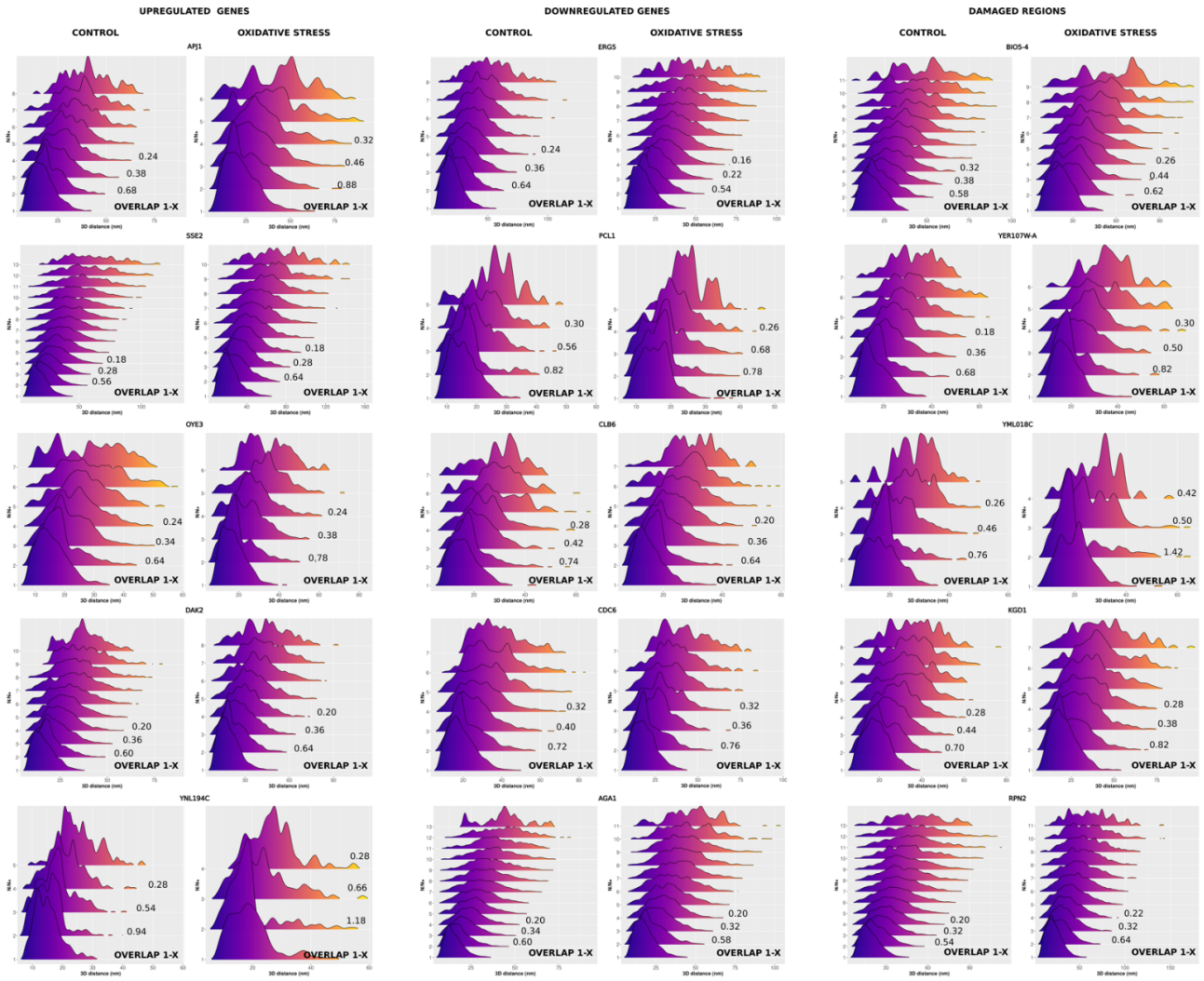
Supporting Figure S2. (A) Periodicity of nucleosome positions, (B) distribution of nucleosome free region (NFR) length around the transcription start site (TSS), and (C) distribution of chromatin interaction domain (CID) lengths in control (gray) and oxidative stressed (red) cells on the whole genome, upregulated genes, downregulated genes and damaged regions where H2AS129ph signal is enriched.



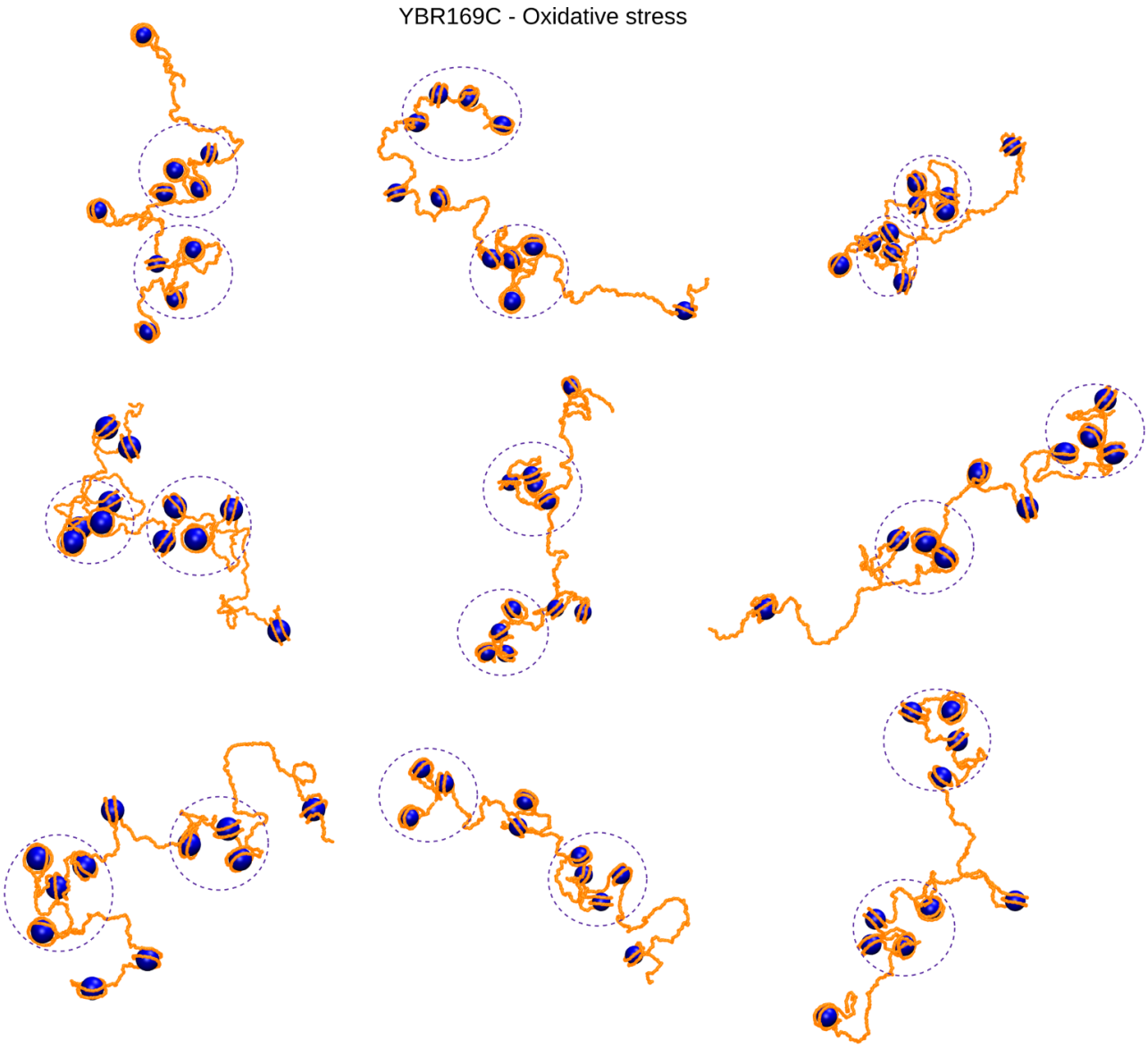
Supporting Figure S3. Effect of oxidative stress on chromatin structure. Micro-C contact frequency maps at 200 bp binning for control (top panel) and oxidative stress (bottom panel) in a 10kb region containing the upregulated gene *APJ1* (A), the downregulated gene *ERG5* (B), and the damaged region *HOL1-BIO3*. Insulation score track was plotted on top of the maps for control (black) and oxidative stress (red). Below the maps, sequence annotation and nucleosome position profile (black = control, red = oxidative stress) was added. Promoter orientation is indicated with black arrows. Genes of interest are highlighted with a blue box. In (C), H2AS129ph tracks are indicated for the control (black) and oxidative stress (red) samples, with differential H2AS129ph highlighted with an orange bar below the plot.



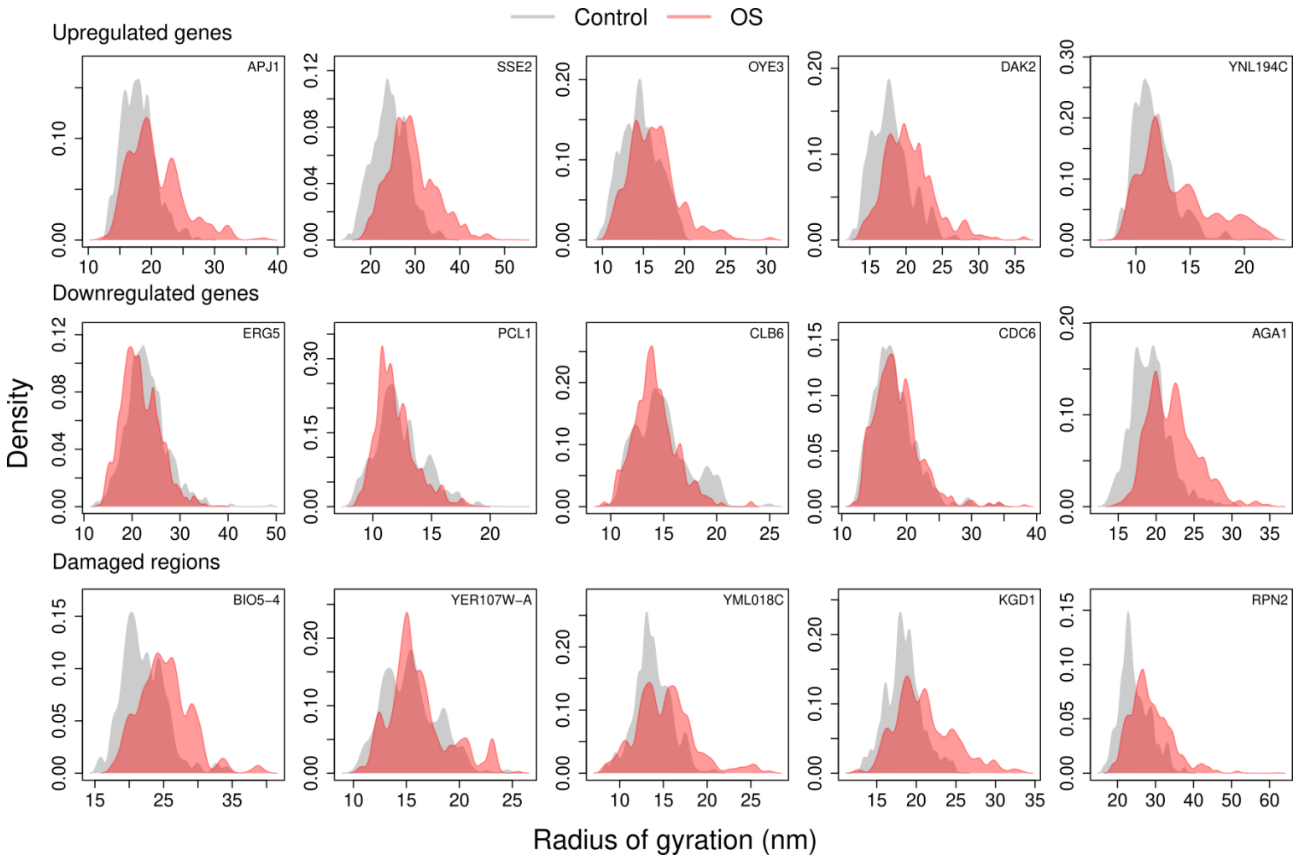
Supporting Figure SX. Chromatin fibre models for *APJ1* obtained with the coarse-grained approach at the nucleosome level. Representative structures extracted from the ensemble for oxidative stress sample. Extreme nucleosomes from the gene (+1 from the TSS and -last from TTS) are indicated.



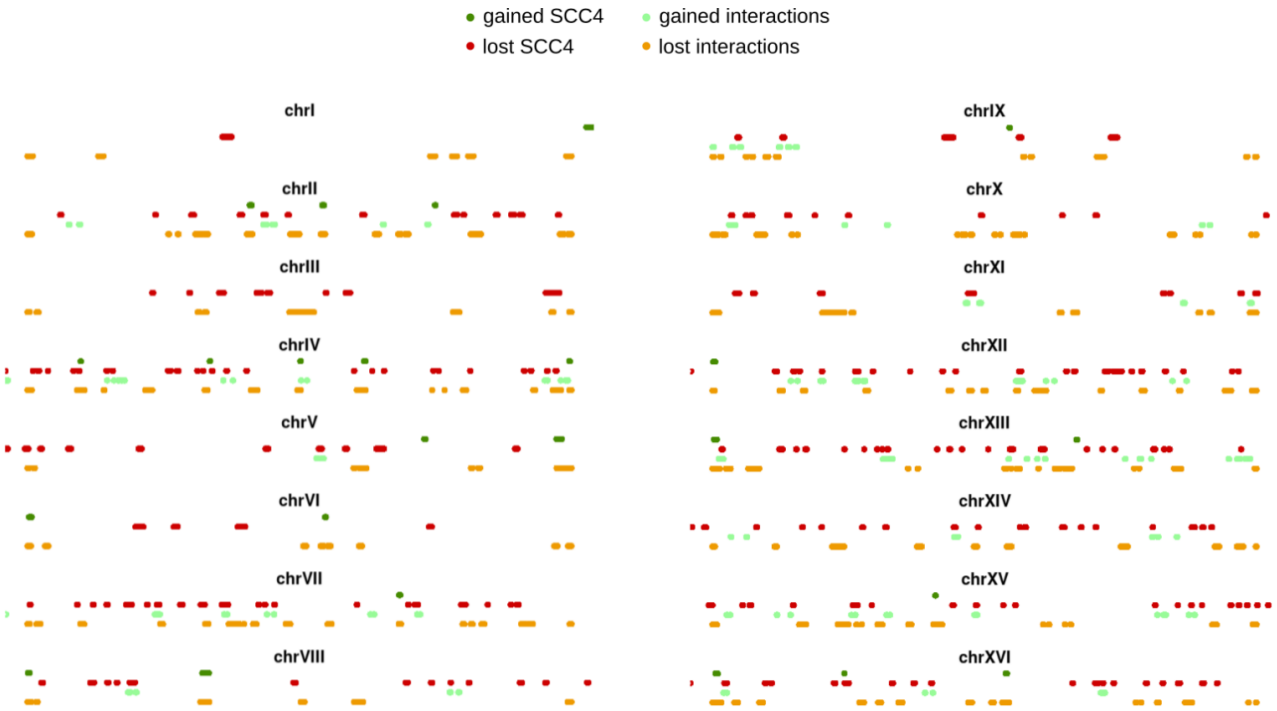
Supporting Figure S4. 3D distance distributions between nucleosome *N* and nucleosome *N+x* (*x* = 1, 2, … representing subsequent positions in the sequence) obtained for the ensemble structures from the coarse-grained model at the nucleosome level. Results are shown for different upregulated, downregulated genes and damaged regions in control and oxidative stress (OS) samples. Overlap between *N/N+1* and *N/N+x* distributions are shown (for *x* = 2, 3, 4). Overlap = normalized proportion of the unitary area under the *N/N+x* curve below the median of the *N/N+1* distribution (a value of 1 means both medians are coincident).



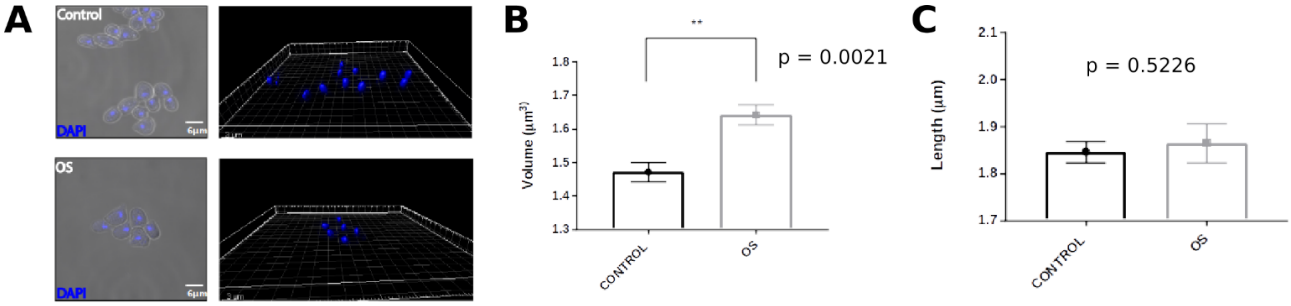
Supporting Figure S5. Chromatin fibre models for SSE2 obtained with the coarse-grained approach at the nucleosome level. Representative structures extracted from the ensemble for oxidative stress sample. Nucleosome clutches are highlighted by dashed circles.



Supporting Figure S6. Relative compaction of the modelled nucleosome fibers between control and stressed conditions, measured by radius of gyration (gray for control and red for oxidative stress). Top panel = upregulated genes, middle panel = downregulated genes, bottom panel = damaged regions.



Supporting Figure S7. Comparison between genomic locations with differential SCC4 ChIP-seq signal and genomic locations with differential chromatin interactions from Hi-C.



Supporting Figure S8. Nucleus morphological measures. (A) Left: Fluorescence microscopy (nuclei stained with DAPI) in control (top panel) and OS (bottom panel) cells. Right: Nuclei surface obtained with IMARIS® for control (top panel) and OS (bottom panel) cells. (B) Nuclei volume (mean +/- standard deviation). t-Student p-value = 0.0021. (D) Length of major principal axis (mean +/- standard deviation). t-Student p-value = 0.5226.

Supporting Table S1. Number of CIDs for control and OS conditions on the whole genome, upregulated genes, downregulated genes and damaged regions.

Condition	**No. CIDs**
whole genome	
CONTROL	1569
OS	1919
upregulated genes	
CONTROL	65
OS	80
downregulated genes	
CONTROL	203
OS	232
damaged regions	
CONTROL	168
OS	174

4. Systematic study of hybrid triplex topology and stability suggests a general triplex-mediated regulatory mechanism

Supplemental Information

Systematic study of hybrid triplex topology and stability suggests a general triplex-mediated regulatory mechanism

Vito Genna[1,2#], Guillem Portella[1,3#], Alba Sala[1#], Montserrat Terrazas [1#], Núria Villegas,[1] Lidia Mateo[1], Chiara Castellazzi[1], Mireia Labrador,[1] Anna Aviño[4], Adam Hospital[1], Albert Gandioso,[1] Patrick Aloy[1], Isabelle Brun Heath[1], Carlos Gonzalez[5], Ramon Eritja[4], and Modesto Orozco[1,6]*

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, E-08028 Barcelona, Spain.
[2]Nostrum Biodiscovery, SL. Barcelona, Spain
[3]Department of Chemistry, University of Cambridge, Cambridge, UK.
[4]Institute for Advanced Chemistry of Catalonia (IQAC), CSIC, Networking Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), E-08034 Barcelona, Spain
[5] Instituto de Química Física Blas Cabrera. CSIC. E-28006. Madrid
[6] Department of Biochemistry and Biomedicine, University of Barcelona, E-08028 Barcelona, Spain.

# Equally contributing authors

a

5' and 3' Deletion
5'-rCU CUC CUC C-3' j

3'-Deletions
5'-rCCU CUC CU-3' i
5'-rCCU CUC CUC -3' h
5'-rCCU CUC CUC C-3' g
5'-rCCU CUC CUC CC-3' f

5'-Deletions
5'-r UC CUC CCU-3' e
5'-r CUC CUC CCU-3' d
5'-rU CUC CUC CCU-3' c
5'-rCU CUC CUC CCU-3' b
5'-rCCU CUC CUC CCU-3' a

5'-dGGA GAG GAG GGA T T
3'-dCCT CTC CTC CCT T T B

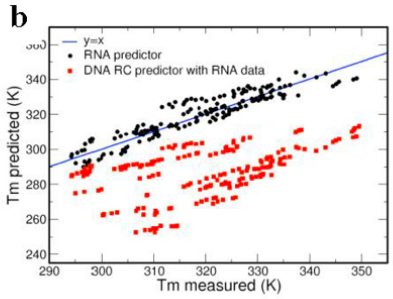
T T dGGA GAG GAG GGA-3'
T T dCCT CTC CTC CCT-5' A

5'-rUUC UCU UCU UUC-3' k
5'-dAAG AGA AGA AAG T T
3'-dTTC TCT TCT TTC T T C

5'-rUUC UUU UCU UUC-3' l
5'-dAAG AAA AGA AAG T T
3'-dTTC TTT TCT TTC T T D

5'-rUCC UCU CUC UCU-3' m
5'-dAGG AGA GAG AGA T T
3'-dTCC TCT CTC TCT T T E

b



c

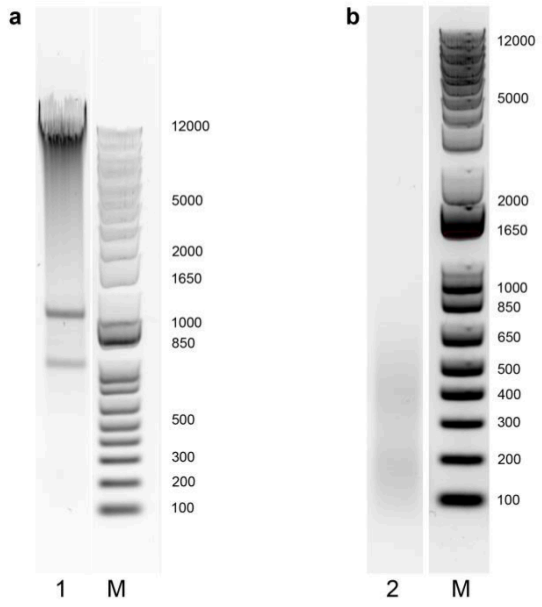
	T_m (K)		ΔH°		ΔG°	
Triplex	Measured	Predicted	Measured	Predicted	Measured	Predicted
A + a	335.2	336.2	-91.6	-83.9	-17.7	-16.8
A + b	334.7	331.3	-72.0	-72.9	-14.9	-14.6
A + c	323.3	324.2	-67.1	-67.2	-12.6	-12.7
A + d	321.5	322.6	-61.8	-61.5	-12.1	-12.0
A + e	314.3	313.7	-48.9	-55.8	-10.5	-9.1
A + f	331.0	335.7	-76.2	-78.2	-15.2	-16.1
A + g	330.5	330.3	-73.5	-67.2	-14.7	-13.9
A + h	324.5	322.9	-62.2	-56.3	-12.3	-11.6
A + i	313.0	313.3	-58.6	-50.5	-10.4	-9.7
A + j	330.9	330.3	-69.9	-67.2	-14.3	-13.9
B + a	331.9	336.3	-87.7	-83.9	-16.6	-16.8
B + b	335.1	331.3	-73.7	-72.9	-15.6	-14.6
B + c	326.8	324.3	-65.9	-67.2	-13.3	-12.7
B + d	325.2	322.5	-54.3	-61.5	-12.1	-12.0
B + e	313.5	313.7	-51.8	-55.8	-10.2	-10.1
B + f	328.1	335.7	-74.9	-78.2	-14.4	-16.1
B + g	322.8	330.2	-64.1	-67.2	-12.5	-13.9
B + h	317.1	322.9	-55.3	-56.3	-11.0	-11.6
B + i	307.0	313.2	-44.5	-50.5	-9.1	-9.7
B + j	332.7	330.3	-67.8	-67.2	-14.6	-13.9
C + k	318.2	315.8	-70.8	-75.9	-12.1	-11.0
D + l	306.1	309.9	-64.0	-67.3	-9.3	-9.8
E + m	316.4	328.3	-58.0	-68.2	-10.9	-13.6

Suppl. Figure S1. a. Sequences used in thermodynamic studies to generate the predictive rPy-dPu·dPy model. **b.** Plot comparing T_m values measured for TFO RNA sequences **a-m** [measured at different pH values (5.0, 5.38, 6.0, 6.5 and 7.0) and oligonucleotide concentration (5 μM, 8 μM, 12 μM, 18 μM and 22 μM)] vs. T_m values predicted for these sequences and conditions using our predictive rPy-dPu·dPy model (black dots) and the dPy-dPu·dPy model developed by Roberts & Crothers (ref 1; red dots). **c.** Comparison of measures and predicted $\Delta H°$ and $\Delta G°$ (≈ 37 °C) for the triplexes shown in panels a-g of this figure, examined at pH 5.0.

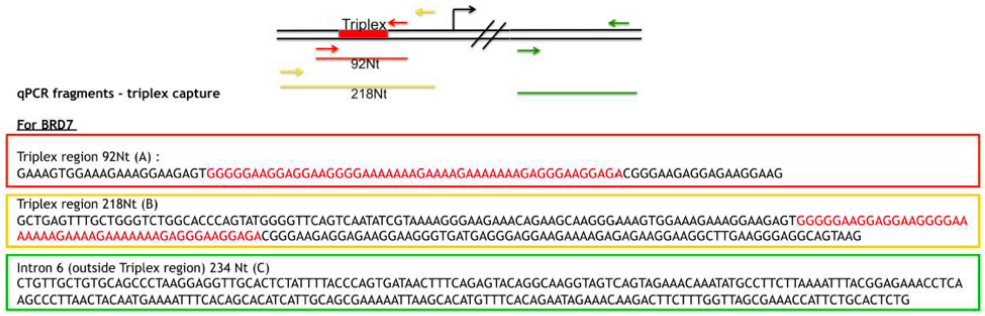
	pH	T_m (K) Measured	T_m (K) Predicted
5'-rUCU CUC UCU CUC-3'	5.6	335.8	329.1
5'-dAGA GAG AGA GAG $^{T}_{T}$	6.5	312.7	304.3
3'-dTCT CTC TCT CTC $_{T}^{T}$	7.0	304.1	292.0
5'-rUCU CCU CUC UCU-3'	5.6	335.2	328.3
5'-dAGA GGA GAG AGA $^{T}_{T}$	6.5	311.0	305.3
3'-dTCT CCT CTC TCT $_{T}^{T}$	7.0	298.7	293.9
5'-rCCU CUC UCU CUU-3'	5.6	330.9	327.9
5'-dGGA GAG AGA GAA $^{T}_{T}$	6.5	310.1	305.2
3'-dCCTCTCTCTCTT $_{T}^{T}$	7.0	299.8	293.9
5'-rCUU UUC UUU CUC-3'	5.6	311.4	315.5
5'-dGAA AAG AAA GAG $^{T}_{T}$	6.5	297.3	300.8
3'-dCTT TTC TTT CTC $_{T}^{T}$	7.0	290.4	293.2
5'-rCUC CUC UCC UCC-3'	5.6	336.6	337.6
5'-dGAG GAG AGG AGG $^{T}_{T}$	6.5	309.2	309.5
3'-dCTC CTC TCC TCC $_{T}^{T}$	7.0	295.1	295.9
5'-rUCC CUC CCU CUU-3'	5.6	326.5	330.9
5'-dAGG GAG GGA GAA $^{T}_{T}$	6.5	302.1	308.8
3'-dTCC CTC CCT CTT $_{T}^{T}$	7.0	291.7	297.7

	pH	T_m (K) Measured	T_m (K) Predicted
5'-rUUC CUU UCC UU-3'	5.6	313.9	314.0
5'-dAAG GAA AGG AA $^{T}_{T}$	6.5	298.9	300.2
3'-dTTC CTT TCC TT $_{T}^{T}$	7.0	291.4	293.0
5'-rUUC CUU UCU UU-3'	5.6	307.9	308.0
5'-dAAG GAA AGA AA $^{T}_{T}$	6.5	295.4	297.3
3'-dTTC CTT TCT TT $_{T}^{T}$	7.0	291.9	291.6
5'-rCUC UUC UCC UU-3'	5.6	321.1	320.0
5'-dGAG AAG AGG AA $^{T}_{T}$	6.5	301.8	300.4
3'-dCTC TTC TCC TT $_{T}^{T}$	7.0	nd	nd
5'-rCUU CUU CCU-3'	5.6	312.1	308.3
5'-dGAA GAA GGA $^{T}_{T}$	6.5	294.4	290.4
3'-dCTT CTT CCT $_{T}^{T}$	7.0	nd	nd
5'-rCUU CUU CCU UUU C-3'	5.6	319.6	322.6
5'-dGAA GAA GGA AAA G $^{T}_{T}$	6.5	301.6	306.2
3'-dCTT CTT CCT TTT C $_{T}^{T}$	7.0	292.4	297.8
5'-rCUC CUC CU-3'	5.6	313.5	313.2
5'-dGAG GAG GA $^{T}_{T}$	6.5	293.4	289.5
3'-dCTC CTC CT $_{T}^{T}$	7.0	nd	nd
5'-rCUC CUC CUC C-3'	5.6	329.4	330.3
5'-dGAG GAG GAG G $^{T}_{T}$	6.5	301.9	302.7
3'-dCTC CTC CTC C $_{T}^{T}$	7.0	292.7	289.3

Suppl. Figure S2. Sequences used in thermodynamic studies for cross-validation of our predictive rPy-dPu·dPy model (validation set), measured at different pH values (5.6, 6.5 and 7.0) and at 18 μM oligonucleotide concentration, and plot comparing measured vs. predicted T_m values. Black dots: comparison of measured T_m values vs. T_m values predicted for TFO RNA sequences shown in Suppl. Figure S1 (**a-m**) using our predictive rPy-dPu·dPy model. Red dots: comparison of measured T_m values vs. T_m values predicted for TFO RNA sequences shown in Figure S1 (**a-m**) using the dPy-dPu·dPy triplex developed by Roberts and Crothers (1).



Suppl. Figure S3. a. Lane 1: genomic DNA (isolated from HeLa cells) treated (i) with 0.5 unit of DNase I for 5 min at 37 °C and (ii) with 4 µL of 10% SDS and 4 µL of 20 µg/ µL of Proteinase K for 30 min at 37 °C. **b.** Lane 2: 10 µg of purified genomic DNA sonicated (6 cycles 30 sec ON/90 sec OFF) in 50 µL of buffer A [10 mM Tris-HCl (pH 7.4), 50 mM KCl, 5 mM $MgCl_2$]. M: DNA ladder.



Suppl. Figure S4. Schematic representation of the three different regions of BRD7 gene amplified by qPCR.

a

GO:BP		stats					
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$	T	Q	T∩Q	U
tissue development	GO:0009888	1.893×10^{-6}		2004	526	93	21010
regulation of developmental process	GO:0050793	6.366×10^{-6}		2465	526	106	21010
nervous system development	GO:0007399	1.499×10^{-5}		2513	526	106	21010
neurogenesis	GO:0022008	2.741×10^{-5}		1708	526	79	21010
skin development	GO:0043588	4.691×10^{-5}		318	526	26	21010
epidermis development	GO:0008544	5.799×10^{-5}		388	526	29	21010
epidermal cell differentiation	GO:0009913	7.793×10^{-5}		245	526	22	21010
keratinocyte differentiation	GO:0030216	1.190×10^{-4}		175	526	18	21010
regulation of cell differentiation	GO:0045595	1.799×10^{-4}		1562	526	71	21010
keratinization	GO:0031424	2.117×10^{-4}		82	526	12	21010
generation of neurons	GO:0048699	4.052×10^{-4}		1485	526	67	21010
negative regulation of developmental process	GO:0051093	6.427×10^{-4}		923	526	47	21010
neuron differentiation	GO:0030182	1.629×10^{-3}		1404	526	62	21010
positive regulation of multicellular organismal proc...	GO:0051240	1.911×10^{-3}		1629	526	69	21010
regulation of multicellular organismal development	GO:2000026	2.508×10^{-3}		1396	526	61	21010
potassium ion transmembrane transport	GO:0071805	4.807×10^{-3}		215	526	17	21010
potassium ion transport	GO:0006813	5.277×10^{-3}		239	526	18	21010
negative regulation of multicellular organismal pro...	GO:0051241	1.062×10^{-2}		1102	526	49	21010
adaptive immune response based on somatic reco...	GO:0002460	1.158×10^{-2}		302	526	20	21010
epithelium development	GO:0060429	1.223×10^{-2}		1233	526	53	21010

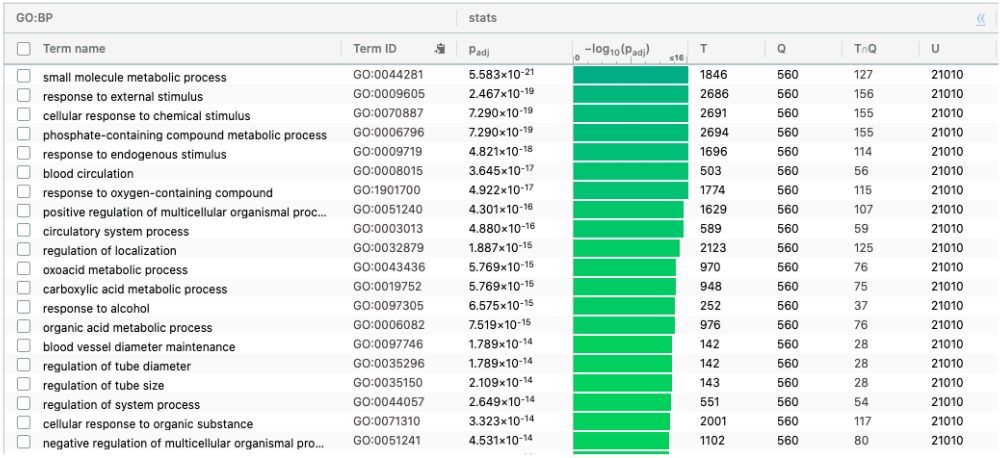
b

GO:BP		stats					
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$	T	Q	T∩Q	U
small molecule metabolic process	GO:0044281	1.494×10^{-23}		1846	2061	321	21010
cellular response to chemical stimulus	GO:0070887	8.235×10^{-21}		2691	2061	416	21010
response to external stimulus	GO:0009605	1.002×10^{-20}		2686	2061	415	21010
regulation of cell population proliferation	GO:0042127	1.002×10^{-20}		2184	2061	354	21010
positive regulation of multicellular organismal proc...	GO:0051240	2.045×10^{-20}		1629	2061	283	21010
regulation of developmental process	GO:0050793	2.580×10^{-19}		2465	2061	383	21010
homeostatic process	GO:0042592	2.043×10^{-18}		1704	2061	286	21010
tissue development	GO:0009888	5.240×10^{-18}		2004	2061	322	21010
response to oxygen-containing compound	GO:1901700	6.638×10^{-18}		1774	2061	293	21010
phosphate-containing compound metabolic process	GO:0006796	2.748×10^{-16}		2694	2061	398	21010
regulation of immune system process	GO:0002682	2.706×10^{-15}		1482	2061	247	21010
anatomical structure morphogenesis	GO:0009653	5.442×10^{-15}		2683	2061	391	21010
regulation of molecular function	GO:0065009	1.293×10^{-14}		2526	2061	371	21010
regulation of multicellular organismal development	GO:2000026	1.886×10^{-14}		1396	2061	233	21010
cell adhesion	GO:0007155	2.788×10^{-14}		1504	2061	246	21010
regulation of localization	GO:0032879	4.057×10^{-14}		2123	2061	321	21010
response to endogenous stimulus	GO:0009719	4.575×10^{-14}		1696	2061	269	21010
intracellular signal transduction	GO:0035556	2.164×10^{-13}		2609	2061	375	21010
negative regulation of multicellular organismal pro...	GO:0051241	3.193×10^{-13}		1102	2061	191	21010
regulation of transport	GO:0051049	5.031×10^{-13}		1762	2061	273	21010

c

GO:BP		stats					
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$	T	Q	T∩Q	U
regulation of developmental process	GO:0050793	1.456×10^{-7}		2465	806	154	21010
nervous system development	GO:0007399	4.971×10^{-6}		2513	806	150	21010
regulation of cell differentiation	GO:0045595	5.865×10^{-6}		1562	806	104	21010
multicellular organismal-level homeostasis	GO:0048871	5.877×10^{-6}		816	806	65	21010
positive regulation of multicellular organismal proc...	GO:0051240	1.265×10^{-5}		1629	806	106	21010
lipid metabolic process	GO:0006629	2.232×10^{-5}		1406	806	94	21010
homeostatic process	GO:0042592	5.788×10^{-5}		1704	806	107	21010
anatomical structure morphogenesis	GO:0009653	7.888×10^{-5}		2683	806	152	21010
cell migration	GO:0016477	8.607×10^{-5}		1475	806	95	21010
phosphate-containing compound metabolic process	GO:0006796	9.185×10^{-5}		2694	806	152	21010
regulation of cell population proliferation	GO:0042127	1.189×10^{-4}		2184	806	128	21010
cell motility	GO:0048870	1.189×10^{-4}		1676	806	104	21010
small molecule metabolic process	GO:0044281	1.232×10^{-4}		1846	806	112	21010
regulation of molecular function	GO:0065009	1.576×10^{-4}		2526	806	143	21010
tissue development	GO:0009888	2.472×10^{-4}		2004	806	118	21010
negative regulation of cell differentiation	GO:0045596	2.633×10^{-4}		676	806	52	21010
fatty acid metabolic process	GO:0006631	2.660×10^{-4}		396	806	36	21010
negative regulation of developmental process	GO:0051093	2.737×10^{-4}		923	806	65	21010
neurogenesis	GO:0022008	3.875×10^{-4}		1708	806	103	21010
cellular lipid metabolic process	GO:0044255	5.324×10^{-4}		1003	806	68	21010

d



GO:BP		stats					
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$	T	Q	TnQ	U
small molecule metabolic process	GO:0044281	5.583×10^{-21}		1846	560	127	21010
response to external stimulus	GO:0009605	2.467×10^{-19}		2686	560	156	21010
cellular response to chemical stimulus	GO:0070887	7.290×10^{-19}		2691	560	155	21010
phosphate-containing compound metabolic process	GO:0006796	7.290×10^{-19}		2694	560	155	21010
response to endogenous stimulus	GO:0009719	4.821×10^{-18}		1696	560	114	21010
blood circulation	GO:0008015	3.645×10^{-17}		503	560	56	21010
response to oxygen-containing compound	GO:1901700	4.922×10^{-17}		1774	560	115	21010
positive regulation of multicellular organismal proc...	GO:0051240	4.301×10^{-16}		1629	560	107	21010
circulatory system process	GO:0003013	4.880×10^{-16}		589	560	59	21010
regulation of localization	GO:0032879	1.887×10^{-15}		2123	560	125	21010
oxoacid metabolic process	GO:0043436	5.769×10^{-15}		970	560	76	21010
carboxylic acid metabolic process	GO:0019752	5.769×10^{-15}		948	560	75	21010
response to alcohol	GO:0097305	6.575×10^{-15}		252	560	37	21010
organic acid metabolic process	GO:0006082	7.519×10^{-15}		976	560	76	21010
blood vessel diameter maintenance	GO:0097746	1.789×10^{-14}		142	560	28	21010
regulation of tube diameter	GO:0035296	1.789×10^{-14}		142	560	28	21010
regulation of tube size	GO:0035150	2.109×10^{-14}		143	560	28	21010
regulation of system process	GO:0044057	2.649×10^{-14}		551	560	54	21010
cellular response to organic substance	GO:0071310	3.323×10^{-14}		2001	560	117	21010
negative regulation of multicellular organismal pro...	GO:0051241	4.531×10^{-14}		1102	560	80	21010

Suppl. Figure S5. Top 20 Gene Ontology IDs found with g:Profiler (2) from associated TFOs. Analysis is performed considering promoters from our miRNAs **(a)** and lncRNAs **(b)** candidates; UTRs in miRNAs **(c)** and lncRNAs **(d)** candidates. A Benjamin-Hochberg FDR index < 0.05 was set to assess significance. See Methods for details. Columns indicate Term size (T), Query size (Q), Overlap size (TnQ) and Domain size (U).

Supplementary Materials and Methods

Oligonucleotide synthesis, deprotection and purification

Oligonucleotides (DNA hairpins and RNA TFOs) used to generate and to validate our R*D·D predictor (sequences shown in Figures S1 and S2) were purchased from Sigma Aldrich. Hairpins **I-XII** (Figure 2, main text) were synthesized as previously described (3). TFOs 1-6 (Figure 2; main text) were purchased from Sigma Aldrich. Oligonucleotides **XIII**, **7-9** (Figure 7; main text) were synthesized via solid phase synthesis using standard phosphoramidite methods (4). Commercially available 5'-*O*-DMT-dC^{Ac}-, 5'-*O*-DMT-U-3'-succinyl-LCAA-CPG (Link Technologies) and 3'-protected biotin serinol CPG (Glen Research) were used as the solid supports. Phosphoramidite monomers of dA^{Bz}, dC^{Ac}, dG^{iBu}, T (used in the synthesis of oligonucleotide **XIII**) and 2'-*O*-TBDMS-protected phosphoramidite monomers of A^{Bz}, C^{Ac}, G^{dmf}, U (used in the synthesis of oligonucleotides **7-9**), 3'-protected biotin serinol CPG (used in the synthesis of oligonucleotides **8** and **9**), deblocking solution (3% TCA in CH_2Cl_2), activator solution (0.3 M 5-benzylthio-1-H-tetrazole in CH_3CN), CAP A solution (acetic anhydride/pyridine/THF), CAP B solution (THF/*N*-methylimidazole 84/16) and oxidizing solution (0.02 M iodine in THF/pyridine/water (7:2:1)) were obtained from commercial sources. All oligonucleotides were synthesized in DMT-ON mode.

Deprotection and purification of oligonucleotides **7-9**: After solid-phase synthesis, the solid support was incubated at 55 ºC for 2 h with 1.5 mL of NH_3 solution (33%) and 0.5 mL of ethanol. The supernatant was evaporated to dryness and the residue was treated with triethylamine (75 µL) and triethylamine trihydrofloride (60 µL) in DMSO (115 µL) at 65 ºC for 2.5 h. The oligonucleotides were purified using Glen-Pack Cartridges (Glen Research) following manufacturer's instructions.

Deprotection and purification of oligonucleotide **XIII**: After the solid-phase synthesis, the solid support was incubated at 55 ºC for 17 h with 1 mL of NH_3 solution (33%). The oligonucleotide was purified using Glen-Pack Cartridges (Glen Research).

Melting experiments

Melting curves were acquired on Varian-Cary-100 spectrophotometer equipped with a thermoprogrammer using the following buffers: 100 mM sodium acetate/acetic acid, 1 mM EDTA (experiments at pH 5.0, 5.38, 5.6) or 100 mM sodium cacodylate/cacodylic acid, 1 mM EDTA (experiments at pH 6.5, 7.0) (see Materials and Methods in the main text for details). Experiments were performed in 1 cm (for 5 μM and 8 μM oligonucleotide concentrations) and 1 mm path length quartz cells (for 12 μM, 18 μM and 22 μM oligonucleotide concentrations).

Obtaining thermodynamic parameters from UV melting curves

Following the original work of Robert and Crothers (1), we have modelled the equilibrium of triplex formation as arising from three states, Triplex <-> Hairpin + third strand <-> coil. We have used the T_m dependence on the oligonucleotide concentration to compute thermodynamic parameters for a bimolecular equilibrium between two non-self-complementary (e.g. see Marky and Breslauer (5) and eq. 3 in main text). Accordingly, at each concentration (5, 8, 12, 18 and 22 μM) we computed the T_m and the ΔH by fitting the derivative of the absorbance with respect to the temperature using van't Hoff equation and the bimolecular equilibria between two non-self-complementary sequences (6). For most cases, the melting curves of the duplex and the triplex are well separated, and therefore we used the equilibrium triplex ↔ duplex + third strand. When the triplex and duplex absorbance transitions overlap, which happens at pH ≤ 5.38 for sequence with large cytosine and cytosine dinucleotide contents, we used the equilibrium triplex ↔ duplex+oligo ↔ oligo.

For each triplex, we optimized the fit of all absorbance curves at different concentrations simultaneously, imposing that each share the same ΔH. The linear dependence the inverse of the melting temperature and the natural logarithm of the oligonucleotide concentration (RNA hairpin, in our case) was added as constraint for the overall optimization, which was solved using a non-linear least-square fitting. The selection of the boundaries around the T_m for each triplex-duplex transition can have an impact on the final result. We randomly varied these boundaries and after 2000 fitting iterations we selected the fitting that led to the highest overall coefficient of determination both for each individual melting curve and for the overall linear dependence of the melting temperature and the concentration.

Parameterization of the pH effect on ΔG

The ΔG component has the pH dependence included: it's a term that is zero at pH 5.6 and has a dependence on the number of C and CC (i.e. the number of cytosines and cytosine dinucleotides in the sequence). We extracted the ΔG at different pH values (see Figure S1) for the triplexes **A** + **a**, **D** + **l** and **C** + **k** shown in Figure S1. For each sequence composition and pH, we extracted the observed ΔG using the same methodology as described in section XXX. We then calculated the differences between the ΔG at a given pH value and the ΔG predicted by our parameterized model at pH 5.6. We finally fitted a and b in the expression,

$$\Delta G - \Delta G_{pH=5.6} = (C)(pG - 5.6)(a - b(CC))$$

using optimization by minimizing the sum of the squared errors. In this expression, a and b are the unknowns and $\Delta G_{pH=5.6}$ is the ΔG observed at pH 5.6. The RMSE of the fitted correction model was 0.5 kcal/mol.

NMR spectroscopy

NMR spectra were acquired in 30 mM phosphate (pH 6.0), 100 mM NaCl, 10 mM $MgCl_2$ buffer (9:1 H_2O/D_2O). The equimolar concentration of each oligonucleotide (hairpin **III** and TFO **2**) was 0.5 mM. The hairpin was first dissolved in the buffer and NMR spectra were acquired at 5 ºC, 15 ºC, 25 ºC, 35 ºC and 45 ºC. Then, TFO **2** (dry) was resuspended with the hairpin solution. The resulting solution was heated to 95 ºC, allowed to cool slowly to room temperature and stored at 4 ºC until NMR spectra (at 5 ºC, 15 ºC, 25 ºC, 35 ºC and 45 ºC) were acquired. Spectra were acquired in a Bruker spectrometer operating at 600 MHz, equipped with cryoprobe and processed with the TOPSPIN software. Water suppression was achieved by including an excitation sculpting module in the pulse sequence (7).

Bioinformatics scanning of potential RNA-DNA·DNA formation in humans.

The small RNA-seq data of lymphoblastic cells were taken from ArrayExpress under accession number E-MATB-8300 (Damien et al.)(8). Sequence information of the human genome was obtained from the UCSC database (version hg38; December 2013) (9). The reads were reversely mapped against the human genome using STAR (version 2.5.3a, Dobin et al.) (10) to detect the formation of potential parallel triplex cores with the following parameters: --runThreadN 10 --outFilterMultimapNmax 20 --quantMode TranscriptomeSAM GeneCounts --outSAMtype BAM SortedByCoordinate --outSAMattributes All.

Electrophoretic mobility shift assays to analyze triplex formation

TFO **21** was heated at 65 ºC for 10 min to prevent self-aggregation and then quickly cooled on ice. Triplex formation was initiated by addition of 3 µL of 3X triplex buffer [135 mM Tris-acetate (pH 5.5), and 30 mM $MgCl_2$], 2 µL ^{32}P-labeled hairpin DNA **XVII**, 2 µL H_2O containing KCl and 2 µL TFO in a final 9 µL reaction volume. To equilibrate triplex formation, the reaction mixture was incubated at 37 ºC for 6 h. Then, 2 µL 50% glycerol solution containing bromophenol blue was added and the sample was directly loaded onto a 15% native polyacrylamide gel, prepared in 50 mM Tris-acetate (pH 5.5) and 10 mM $MgCl_2$ buffer. Electrophoresis was performed at 8 V/cm for 16 h at 4 ºC in 50 mM Tris-acetate (pH 5.5) and 10 mM $MgCl_2$ buffer and the gel was analyzed by phosphorimaging.

Cell culture

HeLa cells were grown in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% Fetal Bovine Serum (FBS) and 1% penicillin/streptomycin.

Preparation of chromatin and genomic DNA purification

HeLa cells were grown in a T75 flask to 90% confluency. Cells were trypsinized by treatment with 1 mL of 0.25% Trypsin at 37 ºC for 5 min. Then, 10 mL of cold DMEM were added. The cell suspension was transferred into a 15 mL falcon tube and centrifuged for 5 min at 200 x g and 4 ºC. The pellet was resuspended in 10 mL PBS and centrifuged for 5 min at 200 x g. Then, the pellet was resuspended in 1 mL PBS, transferred to a 1.5 mL LoBind Eppendorf tube and centrifuged at 200 x g for 2 min. The pellet was treated

with 400µL of lysis buffer [10 mM Tris·HCl (pH 7.9), 100 mM KCl, 5 mM $MgCl_2$, 0.5% NP-40, 1 mM DTT] and the resulting suspension was gently pipetted up and down 3-5 times to resuspend the cells. The cell lysate was incubated on ice for 5 min. Meanwhile, a LoBind tube with 1 mL of cold sucrose buffer (10 mM Tris pH 7.4, 150 mM NaCl, 24% sucrose) was prepared. The cell lysate was gently overlaid on top of the sucrose buffer and centrifuged at 3500 x g for 10 min. The pellet was rinsed with 1 mL ice cold PBS-EDTA. Isolated nuclei were resuspended in 500 µL of glycerol/urea buffer [25% glycerol, 20 mM Tris-HCl (pH 7.4), 187.5 mM KCl, 0.5 M urea, 0.5% NP-40, 7.5 mM $MgCl_2$], mixed by vortexing 4 s and incubated on ice for 2 min. The lysate was centrifuged at 13,000 x g for 2 min to precipitate the chromatin-RNA complex. The pellet was briefly rinsed with PBS-EDTA, resuspended with 400 µL of DNAse I buffer [10 mM Tris-HCl (pH 7.5), 2.5 mM $MgCl_2$, 0.1 mM $CaCl_2$] and treated with 0.5 µL of 1 U/ µL DNase I at 37 ºC for 5 min. Then, 4 µL of 10% SDS and 4 µL of 20 µg/ µL of Proteinase K were successively added and the resulting mixture was incubated at 37 ºC for 30 min to yield DNA fragments with an average site of >10 Kb (Figure S3a). gDNA was isolated by phenol/chloroform extraction followed by ethanol precipitation and resuspension in 50 µL RNase free water.

***In vitro* triplex pull-down assay**

In two different 1.5ml Bioruptor pico microtubes, 10 µg of purified genomic DNA were suspended in 50 µL of buffer A [10 mM Tris-HCl (pH 7.4), 50 mM KCl, 5 mM $MgCl_2$] or buffer B [45 mM Tris-acetate (pH 5.5) 10 mM $MgCl_2$] and sonicated (6 cycles 30 sec ON/90 sec OFF) to yield DNA fragments with an average size of 200-300 bp (Fig. S2b). After sonication, both DNA mixtures were combined in an eppendorf tube and incubated with 20 pmol of biotinylated TFO at 4 ºC for 15 h. After incubation with MyOne Streptavidin C1 Dynabeads for 40 min at room temperature, beads were washed once with 2X B&W buffer (10 mM Tris-HCl (pH 7.5) 1 mM EDTA 2 M NaCl). Putative Rloops were digested by a 30mn incubation at 30ºC with 2.5u RNaseH in 50µl final. Beads were washed 2 times more with 2X B&W buffer. TFO-associated DNA were eluted by incubating the beads 30min at 37ºC with 25 ng/µL RNase A + 2.5u/µl RNase T1 in 50µl final.

Recovered DNA was analyzed by qPCR using 3µl of sample in 10µl reaction using the LightCycler 480 SYBR Green I Master Mix (Roche Diagnostics) and following manufacturer's instructions. Results were normalized to input DNA. Primers listed in Table S3 were used to amplify a 92bp or 218bp fragment containing the triplex region in the promoter of BRD7 or a 234bp fragment in BRD7 intron 6, 34kb downstream of the triplex region.

	MMG width	mMG width	mG width	H-bonds (WC)	H-bonds (H)	Twist	Roll	Inclination
r(Py)-d(Pu)·r(Py)	6.3 ± 2.1	4.6 ± 0.8	9.1 ± 0.9	18.3 ± 2.1	15.3 ± 1.8	29.2 ± 3	7.9 ± 5.3	14.6 ± 6.4
r(Py)-d(Pu)·d(Py)	9.8 ± 1.8	5.1 ± 0.9	7.8 ± 1.2	18.1 ± 1.7	15.1 ± 1.4	32.1 ± 2.8	3.3 ± 4.9	5.6 ± 4.2
d(Py)-d(Pu)·d(Py)	9.7 ± 1.7	4.7 ± 1.5	6.7 ± 1.4	17.6 ± 1.7	12 ± 1	29.8 ± 2.5	3.1 ± 4.5	6.6 ± 3.9
d(Py)-d(Pu)·r(Py)	8.6 ± 2	3.9 ± 1.2	9.3 ± 1.1	16.5 ± 1.8	11.4 ± 1.7	31.2 ± 3.3	4.4 ± 4.8	8.7 ± 4.3
r(Py)-r(Pu)·r(Py)	7.7 ± 1.3	6.6 ± 1.1	8.3 ± 0.8	16.9 ± 1.8	6.7 ± 2.3	29.7 ± 3.3	4.1 ± 3.7	8.4 ± 3.6
r(Py)-r(Pu)·d(Py)	8.4 ± 1.7	6.1 ± 0.9	8.6 ± 1.3	16.6 ± 1.8	8 ± 1.7	29.4 ± 5.3	4.6 ± 3.4	7.9 ± 3.2

Suppl. Table S1. Helical descriptors of the 6 triplexes considered here with the associated standard deviations. Maximum number of Watson Crick and Hoogsteen hydrogen bonds are 19 and 16 respectively Averages were done with the last 100 ns for all triplexes except the unstable r(Py)-r(Pu)·r(Py) and r(Py)-r(Pu)·d(Py), where they were obtained with the first 50 ns of trajectory.

	%puckering Pu (duplex)	%Puckering Py (duplex)	%Puckering (TFO)
r(Py)-d(Pu)·r(Py)	19% 01'n, 64% C2'n, 15% C1'x	74% C3'n, 23% C4'x	65% C3'n, 14% 01'n
r(Py)-d(Pu)·d(Py)	32% 01'n, 14% C2'n, 50% C1'x	42% 01'n, 4% C2'n, 53% C1'x	96% C3'n
d(Py)-d(Pu)·d(Py)	42% 01'n, 4% C2'n, 53% C1'x	50% 01'n, 4% C2'n, 36% C1'x	35% 01'n, 6% C2'n, 59% C1'x
d(Py)-d(Pu)·r(Py)	22% 01'n, 6% C2'n, 66% C1'x	61% C3'n, 29% C4'x	41% 01'n, 2% C2'n, 57% C1'x
r(Py)-r(Pu)·r(Py)	81% C3'n, 14% C4'x	76% C3'n, 19% C4'x	71% C3'n, 29% 01'n
r(Py)-r(Pu)·d(Py)	64% C3'n, 29% C4'x	73% C3'n, 18% C4'x	68% C3'n, 33% 01'n

Suppl. Table S2. Average puckering distributions for the 3 strands of the 6 triplexes considered here. The index "n" refers to "endo" and "x" to "exo" conformations. Averages were done with the last 100 ns for all triplexes except the unstable r(Py)-r(Pu)·r(Py) and r(Py)-r(Pu)·d(Py), where they were obtained with the first 50 ns of trajectory.

	Primer name	Sequence	Amplicon
BRD7_92bp	Primer F	GAAAGTGGAAAGAAAGGAAGAGTGG	92bp
	Primer R	CTTCCTTCTCCTCTTCCCGTCTC	
BRD7_218bp	Primer F	GCTGAGTTTGCTGGGTCTGG	218bp
	Primer R	CTTACTGCCTCCCTTCAAGCC	
BRD7_Intron6_234bp	Primer F	CTGTTGCTGTGCAGCCCTAAG	234bp
	Primer R	CAGAGTGCAGAATGGTTTCGC	

Suppl. Table S3. qPCR primers

References

1. Roberts, R.W. & Crothers, D.M. Prediction of the stability of DNA triplexes. *Proc. Natl. Acad. Sci. USA* **93**, 4320-4325 (1996).

2. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H. & Vilo, J. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists. *Nucleic Acids Res.* 47, 191-198 (2019).

3. Terrazas, M., Genna, V., Portella, G., Villegas, N., Sánchez, D., Arnan, C., Pulido-Quetglas, C., Johnson, R., Guigó, R., Brun-Heath, I., Aviñó, A., Eritja, R. & Orozco, M. The origins and the biological consequences of the Pur/Pyr DNA·RNA asymmetry. *Chem* **5**, 1619-1631 (2019).

4. Beaucage, S.L. & Caruthers, M.H. Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.* **22**, 1859-1862 (1981).

5. Marky, L.A.; Breslauer, K.J., Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves. *Biopolymers* **26**, 1601-1620 (1987).

6. Owczarzy, R. Melting temperatures of nucleic acids: discrepancies in analysis. *Biophys. Chem.*, **117**, 207-15 (2005).

7. Stott, K., Stonehouse, J., Keeler, J., Hwang, T.L. & Shaka, A.J. Excitation Sculpting in High-Resolution Nuclear Magnetic Resonance Spectroscopy: Application to Selective NOE Experiments. *J. Am. Chem. Soc.* **117**, 4199-4200 (1995).

8. Plassard D. (2019). Small RNA-seq comparing transcriptome (small RNAs) of lymphoblastic cells from FAME patients and control individuals, arrayexpress-repository, V1. https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8300.

9. The Genome Sequencing Consortium. Initial sequencing and analysis of the human genome (2001) *409*, 860-921.

10. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2012).

5. Sequence-Dependent properties of the RNA duplex

Supplementary Information

SEQUENCE DEPENDENT PROPERTIES OF RNA DUPLEX

Federica Battistini[1,2], Alba Sala[1], Adam Hospital[1] and Modesto Orozco[1,2]*

[1] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Baldiri Reixac 10, Barcelona 08028, Spain.

[2] Departament de Bioquímica i Biomedicina. Facultat de Biologia. Universitat de Barcelona, Avgda Diagonal 647, Barcelona 08028, Spain

* Correspondence to M.Orozco: modesto.orozco@irbbarcelona.org

Supplementary Methods

H-bond determination. Following Dans`s et al definition[1], h-bond between base pairs were calculated using cpptraj in Amber 18, a hydrogen bond is considered broken when the distance between the heavy atoms involved in the Watson-Crick interactions is greater than 3.5 Å. Solvent exchange refers to base openings where at least one donor-acceptor distance of WC h-bond is larger than 6 Å. These large separations allow water molecules to interact directly with the base, and eventually exchange protons with imino groups of the bases.

Supplementary Tables

SEQUENCE		Initial Structure	Final Structure (1 microsecond)
SEQ1	GCAACGUGCUAUGGAAGC	https://mmb.irbbarcelona.org/3dRS/s/vJ0vKh	https://mmb.irbbarcelona.org/3dRS/s/VGuGM1
SEQ2	GCAAUAAGUACCAGGAGC	https://mmb.irbbarcelona.org/3dRS/s/LInBaS	https://mmb.irbbarcelona.org/3dRS/s/HtkL5D
SEQ3	GCAGAAACAGCUCUGCGC	https://mmb.irbbarcelona.org/3dRS/s/Ix3B11	https://mmb.irbbarcelona.org/3dRS/s/Cn6UMn
SEQ4	GCAGGCGCAAGACUGAGC	https://mmb.irbbarcelona.org/3dRS/s/tM8XgE	https://mmb.irbbarcelona.org/3dRS/s/ji23N7
SEQ5	GCAUUGGGGACACUACGC	https://mmb.irbbarcelona.org/3dRS/s/vJnklE	https://mmb.irbbarcelona.org/3dRS/s/je0N1v
SEQ6	GCGAACUCAAAGGUUGGC	https://mmb.irbbarcelona.org/3dRS/s/CHtVwJ	https://mmb.irbbarcelona.org/3dRS/s/pBmj9w
SEQ7	GCGACCGAAUGUAAUUGC	https://mmb.irbbarcelona.org/3dRS/s/Ir0oNe	https://mmb.irbbarcelona.org/3dRS/s/xvifpa
SEQ8	GCGGAGGGCCGGGUGGGC	https://mmb.irbbarcelona.org/3dRS/s/JTUUj0	https://mmb.irbbarcelona.org/3dRS/s/70VeA7
SEQ9	GCGUUAGAUUAAAAUUGC	https://mmb.irbbarcelona.org/3dRS/s/4dPB7F	https://mmb.irbbarcelona.org/3dRS/s/Ols2Sj
SEQ10	GCUACGCGGAUCGAGAGC	https://mmb.irbbarcelona.org/3dRS/s/NShxZ9	https://mmb.irbbarcelona.org/3dRS/s/Airv7h
SEQ11	GCUGAUAUACGAUGCAGC	https://mmb.irbbarcelona.org/3dRS/s/mqidy0	https://mmb.irbbarcelona.org/3dRS/s/jfV3kP
SEQ12	GCUGGCAUGAAGCGACGC	https://mmb.irbbarcelona.org/3dRS/s/FSYpQJ	https://mmb.irbbarcelona.org/3dRS/s/uGAd01
SEQ13	GCUUGUGACGGCUAGGGC	https://mmb.irbbarcelona.org/3dRS/s/OKVBqX	https://mmb.irbbarcelona.org/3dRS/s/1WvO80

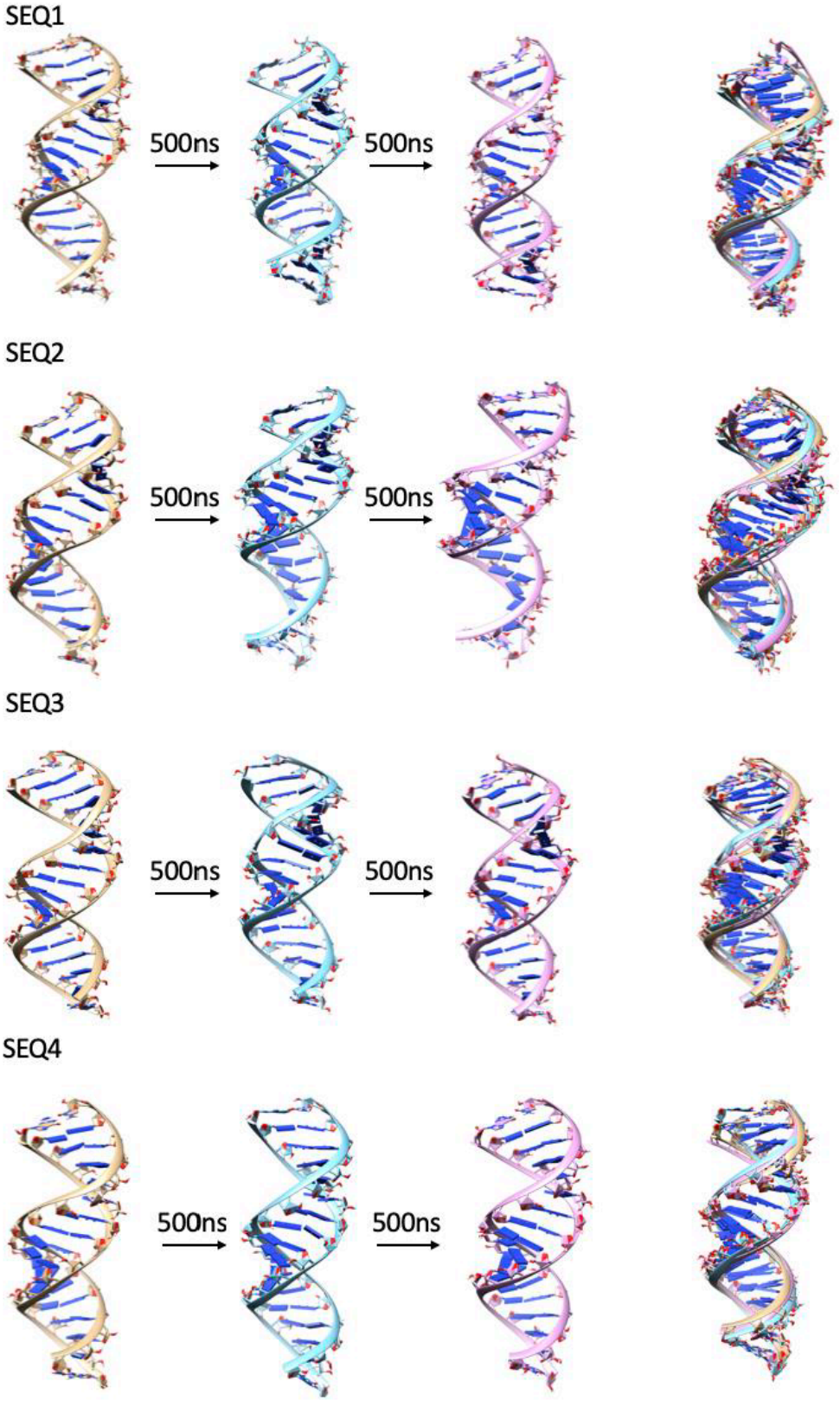
Supplementary Table S1. RNA sequences set and relative links of the starting and final conformations of the duplexes along the MD simulation

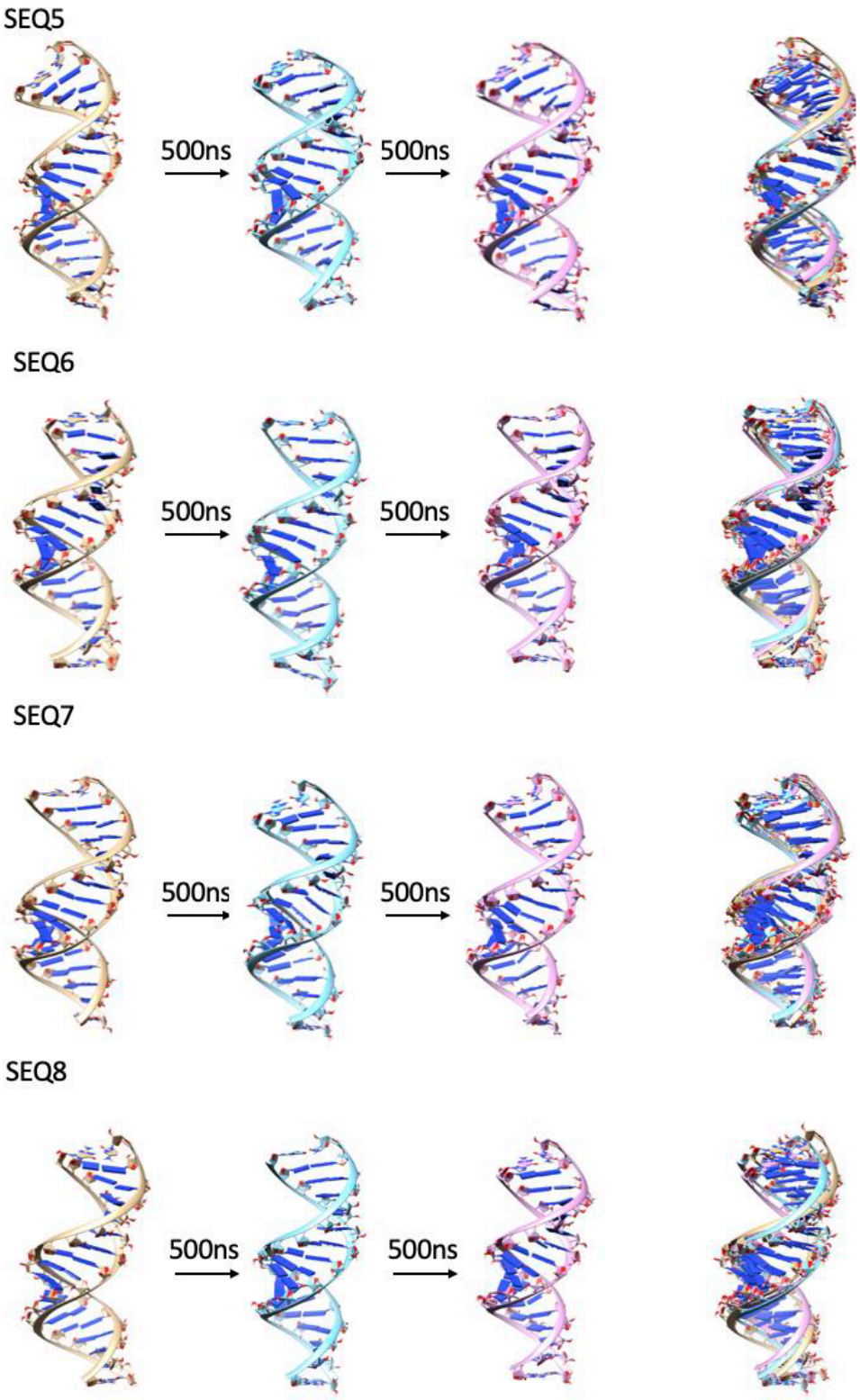
SEQUENCE	RMSD (A-RNA)	RMSD (average)
GCAACGUGCUAUGGAAGC	2.14±0.51	1.68±0.43
GCAAUAAGUACCAGGAGC	2.09±0.47	1.62±0.41
GCAGAAACAGCUCUGCGC	2.20±0.50	1.70±0.45
GCAGGCGCAAGACUGAGC	2.17±0.52	1.67±0.44
GCAUUGGGGACACUACGC	2.09±0.44	1.65±0.42
GCGAACUCAAAGGUUGGC	2.42±0.51	1.68±0.43
GCGACCGAAUGUAAUUGC	2.20±0.52	1.69±0.44
GCGGAGGGCCGGGUGGGC	2.27±0.47	1.65±0.43
GCGUUAGAUUAAAAUUGC	2.26±0.52	1.70±0.43
GCUACGCGGAUCGAGAGC	2.22±0.52	1.73±0.44
GCUGAUAUACGAUGCAGC	2.03±0.47	1.60±0.43
GCUGGCAUGAAGCGACGC	2.24±0.53	1.69±0.41
GCUUGUGACGGCUAGGGC	2.09±0.47	1.66±0.44

Supplementary Table S2. Average RMSd (in Å) deviation from the canonical A-form and for the MD-averaged structure.

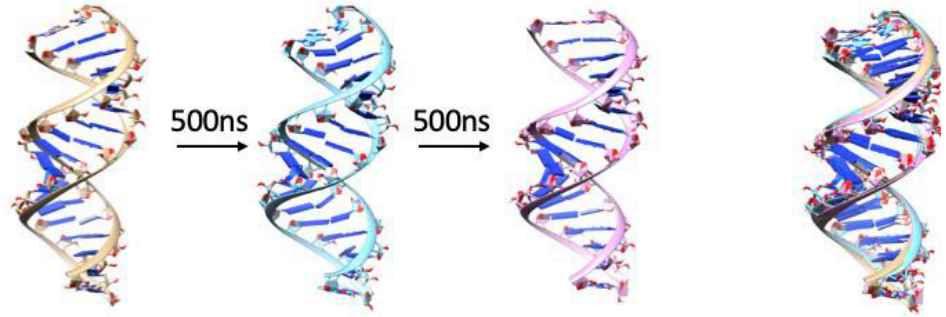
SEQ	RMSIP
1	0.99
2	0.98
3	0.99
4	0.99
5	0.98
6	0.99
7	0.99
8	0.96
9	0.99
10	0.99
11	0.99
12	0.99
13	0.99

Supplementary Table S3. RMSIP between the first (100-500ns) and the second (600-1000ns) part of each MD simulation for the 13 sequences studied.

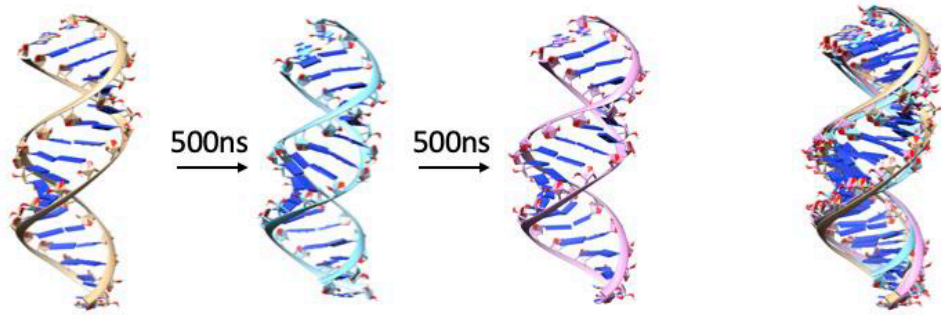

SEQ1
500ns
500ns
SEQ2
500ns
500ns
SEQ3
500ns
500ns
SEQ4
500ns
500ns


SEQ5
500ns
500ns
SEQ6
500ns
500ns
SEQ7
500ns
500ns
SEQ8
500ns
500ns

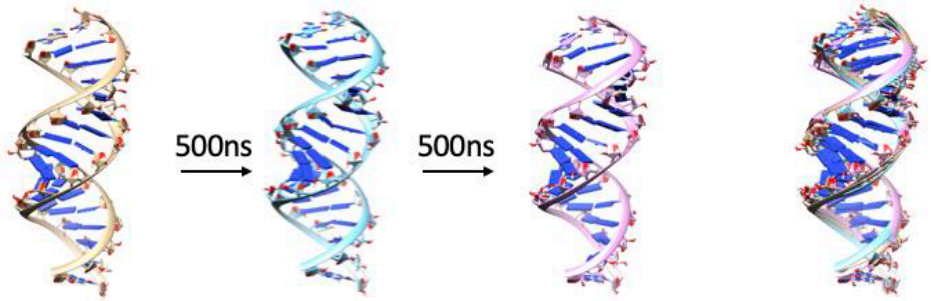
SEQ9



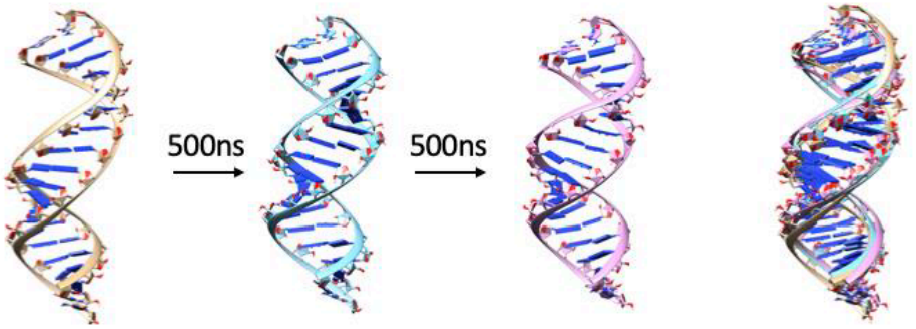
SEQ10

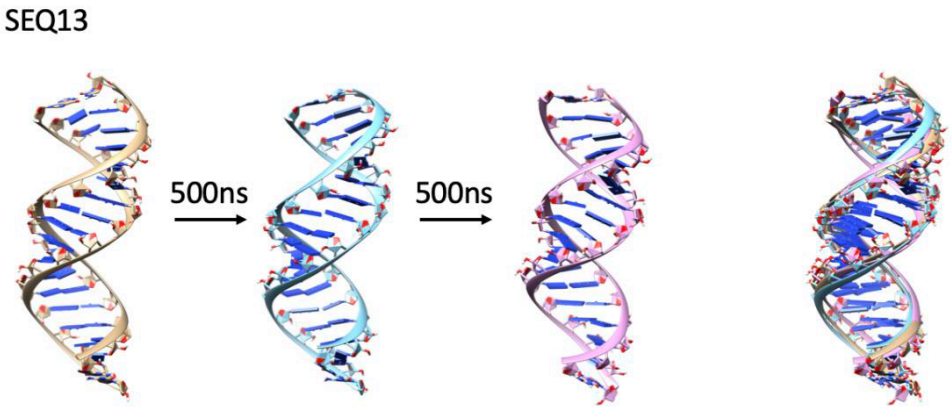


SEQ11

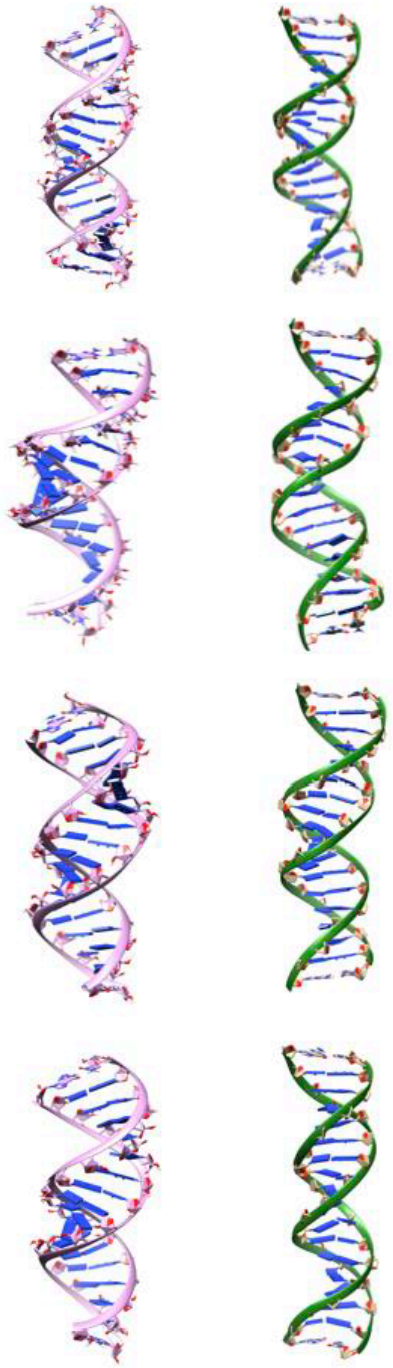


SEQ12

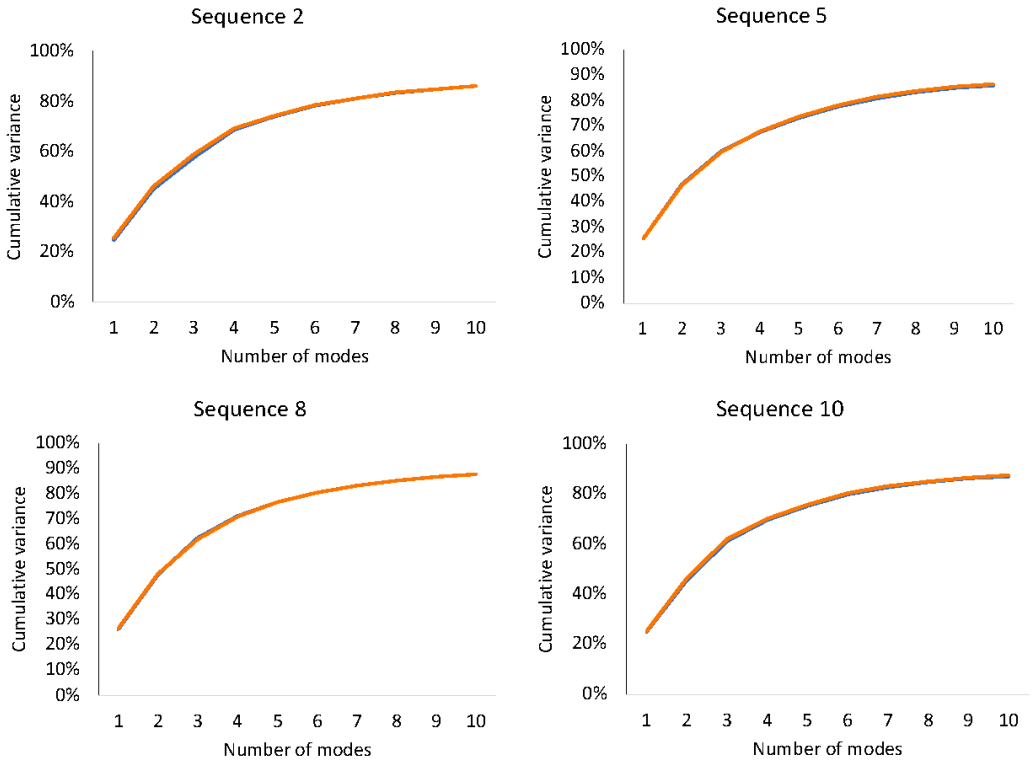




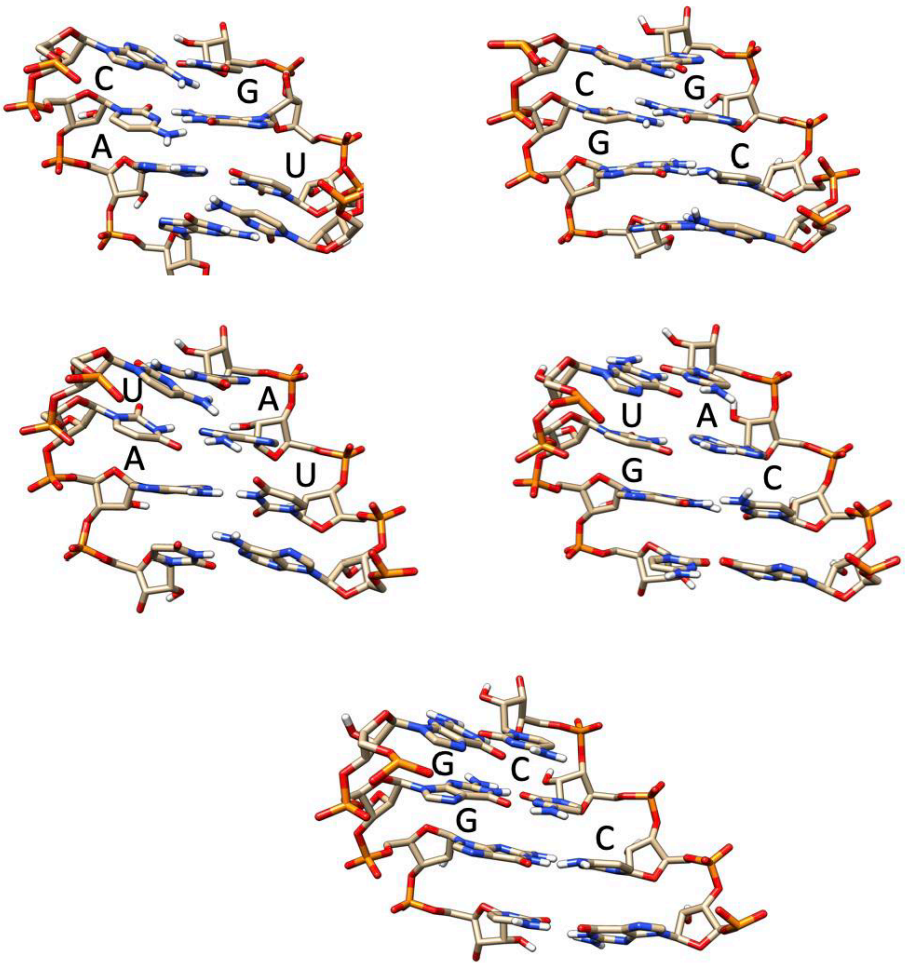
Supplementary Figure S1. Representation of each oligomer (SEQ1-13) along the MD simulation: after the equilibration (light brown ribbon, panel on the left), after 500ns (blue ribbon, central panel) and after 1μs (pink ribbon, right panel) of simulation. On the extreme right panel, the overlap of the three snapshots.



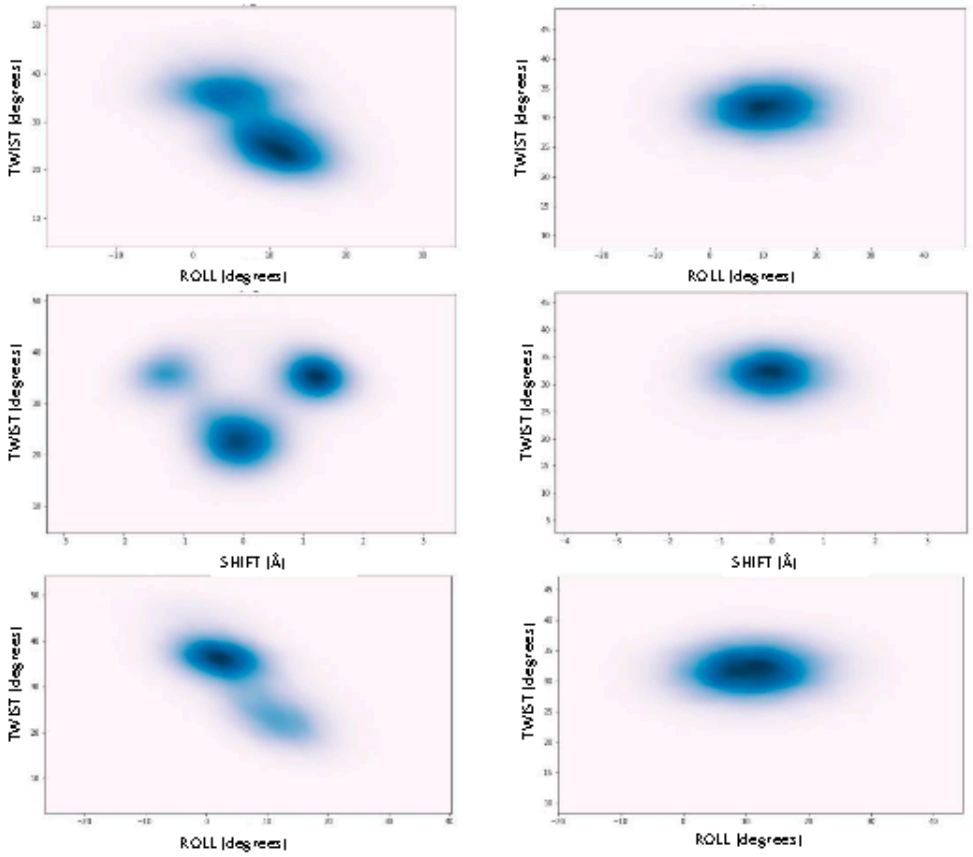
Supplementary Figure S2. Representation of some RNA and DNA duplexes at the end of the MD simulation: RNA (pink ribbon, panel on the left) and DNA (green ribbon, panel on the right).



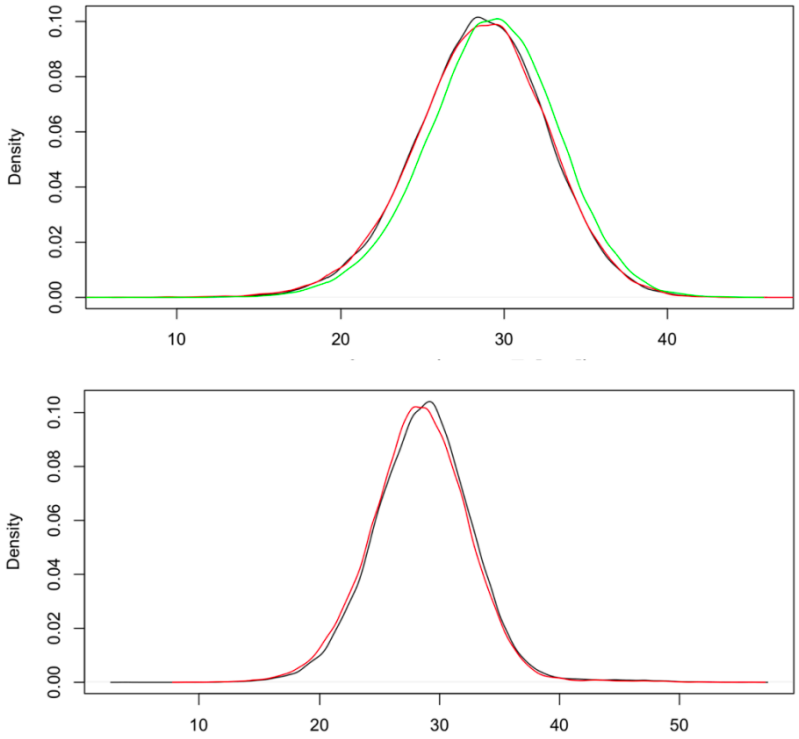
Supplementary Figure S3. Examples of profiles of cumulative variance (%) *vs* the essential deformation modes computed from first (blue) and second (orange) halves of the trajectories.



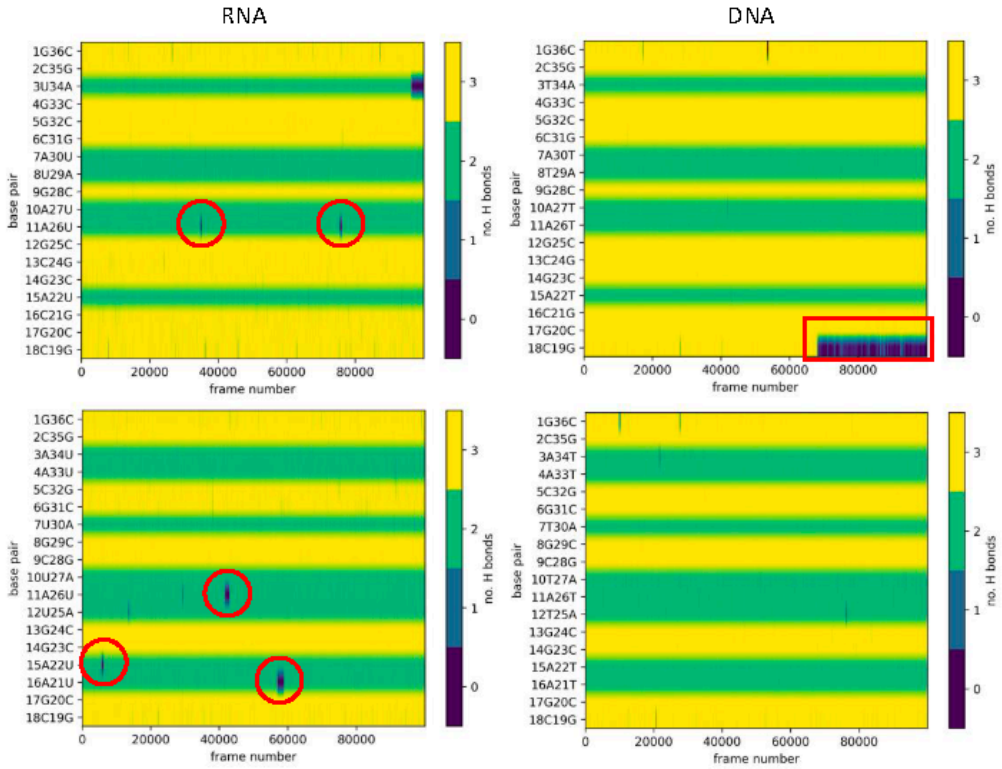
Supplementary Figure S4. Representations of base pairs conformation at the end of the MD simulation for the base pairs steps CpA, CpG, UpA, UpG and GpG.



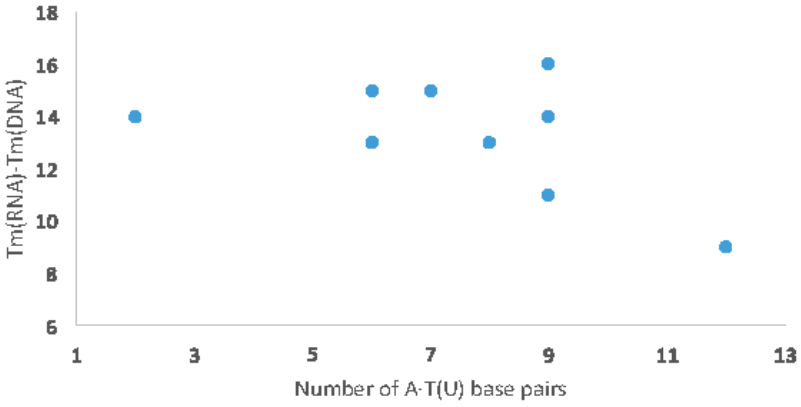
Supplementary Figure S5. Projection of the ensemble collected for CACG, CT(U)AG, T(U)CGA for DNA (left, data take from [1]) and RNA (right) for different pairs of helical parameters. All torsions (twist and roll) are in degrees and translation (shift) in Å.



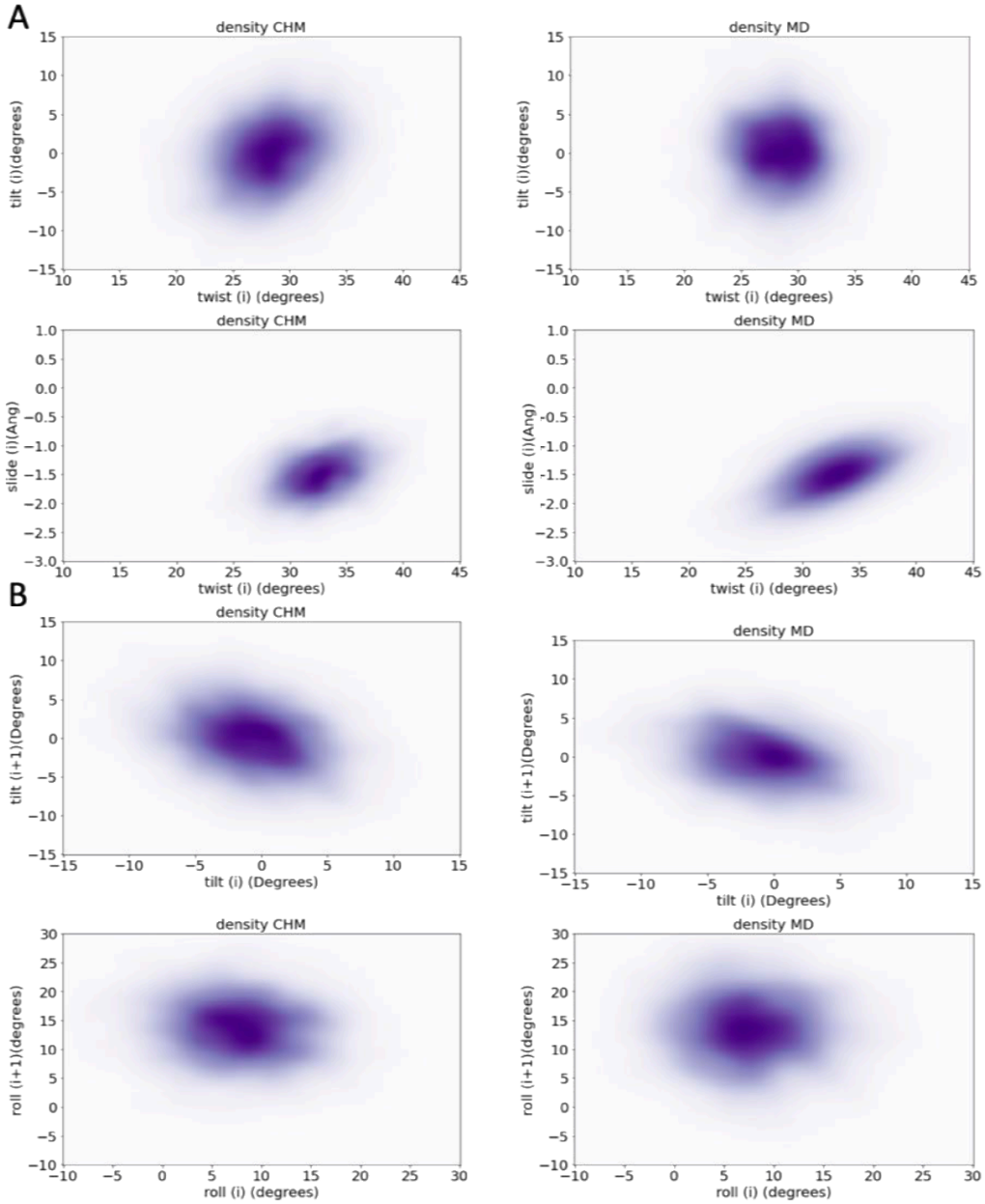
Supplementary Figure S6. Examples of Twist profile for the same tetramer in different sequences and hexamer environment for RNA: UACG in the same hexamer CUACGC (sequence 5, black line), CUACGC (sequence 10, red line) and in a different one AUACGA (sequence 11, green line) (top panel); and AAUU in UAAUUG (sequence 7, black line) and AAAUUG (sequence 9, red line) hexamers.



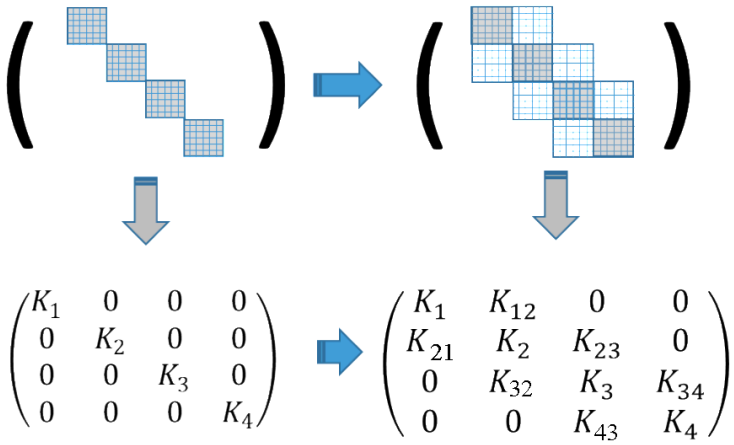
Supplementary Figure S7. Example of breathing events in RNA_2 (left panels) and DNA_2 (right panels, data taken from [1]) trajectories. Note the change from fraying in DNA_2 (highlighted in the red square) to openings in the middle of the helix for RNA_2 (highlighted in the red circles).



Supplementary Figure S8. Relative melting temperature, difference between the DNA and the RNA melting temperatures (from ref %), *vs* the number of A·T(U) base pairs in each sequence.



Supplementary Figure S9. A) Selected examples of the intra-base pairs ensembles (in helical space) using: correlated harmonic model (left panels, CG) and atomistic MD simulation (the reference, right panels). **B)** Selected examples of the inter-bps parameters for consecutive steps, i against the neighboring step i+1, using: correlated harmonic model (left panels) and atomistic MD simulation (the reference, right panels).



Supplementary Figure S10. Scheme to calculate the stiffness matrix. Scheme representing the covariance matrices and stiffness matrices for a segment of 4 bps in the normal harmonic (left panel) and correlated harmonic models (CHM, right panel). In the normal harmonic model, the lack of inter-bps correlations leads to block covariance matrices (on the left) with 6x6 blocks that can be inverted independently yielding to a block stiffness matrix (K). When inter-bps couplings are considered the $36N^2$ matrix (different for each sequence considered) should be diagonalized. Fortunately, the global stiffness matrix is very sparse [2,3] with zero elements out the nearest bps, i.e. it can be defined as a sparse banded matrix with 6x6 intra bps terms (in gray, panel top right) in the block diagonal and 6x6 cross-correlation terms out of the diagonal. This matrix can be easily diagonalized [2,3]; allowing a transferable method that can study any RNA/DNA sequence from pre-determined tetramer-deformations.

References

(1) Dans, P. D.; Balaceanu, A.; Pasi, M.; Patelli, A. S.; Petkevičiūtė, D.; Walther, J.; Hospital, A.; Bayarri, G.; Lavery, R.; Maddocks, J. H.; Orozco, M. The Static and Dynamic Structural Heterogeneities of B-DNA: Extending Calladine-Dickerson Rules. *Nucleic Acids Res.* **2019**, *47*, 11090–11102. https://doi.org/10.1093/nar/gkz905.

(2) Gonzalez, O.; Petkevičiūtė, D.; Maddocks, J. H. A Sequence-Dependent Rigid-Base Model of DNA. *J. Chem. Phys.* **2013**, *138*. https://doi.org/10.1063/1.4789411.

(3) Lopez-Güell, K.; Battistini, F.; Orozco, M. Correlated Motions in DNA: Beyond Base-Pair Step Models of DNA Flexibility. *Nucleic Acids Res.* **2023**, *51*, 2633–2640. https://doi.org/10.1093/NAR/GKAD136.